ETP4HPC's Strategic Research Agenda for High-Performance Computing in Europe 4

Michael Malms, Marcin Ostasz, Maike Gilliot, Pascale Bernier-Bruna, Laurent Cargemel, Estela Suarez, Herbert Cornelius, Marc Duranton, Benny Koren, Pascale Rossé-Laurent, et al.

ETP 4 HPC

EUROPEAN TECHNOLOGY PLATFORM FOR HIGH PERFORMANCE COMPUTING

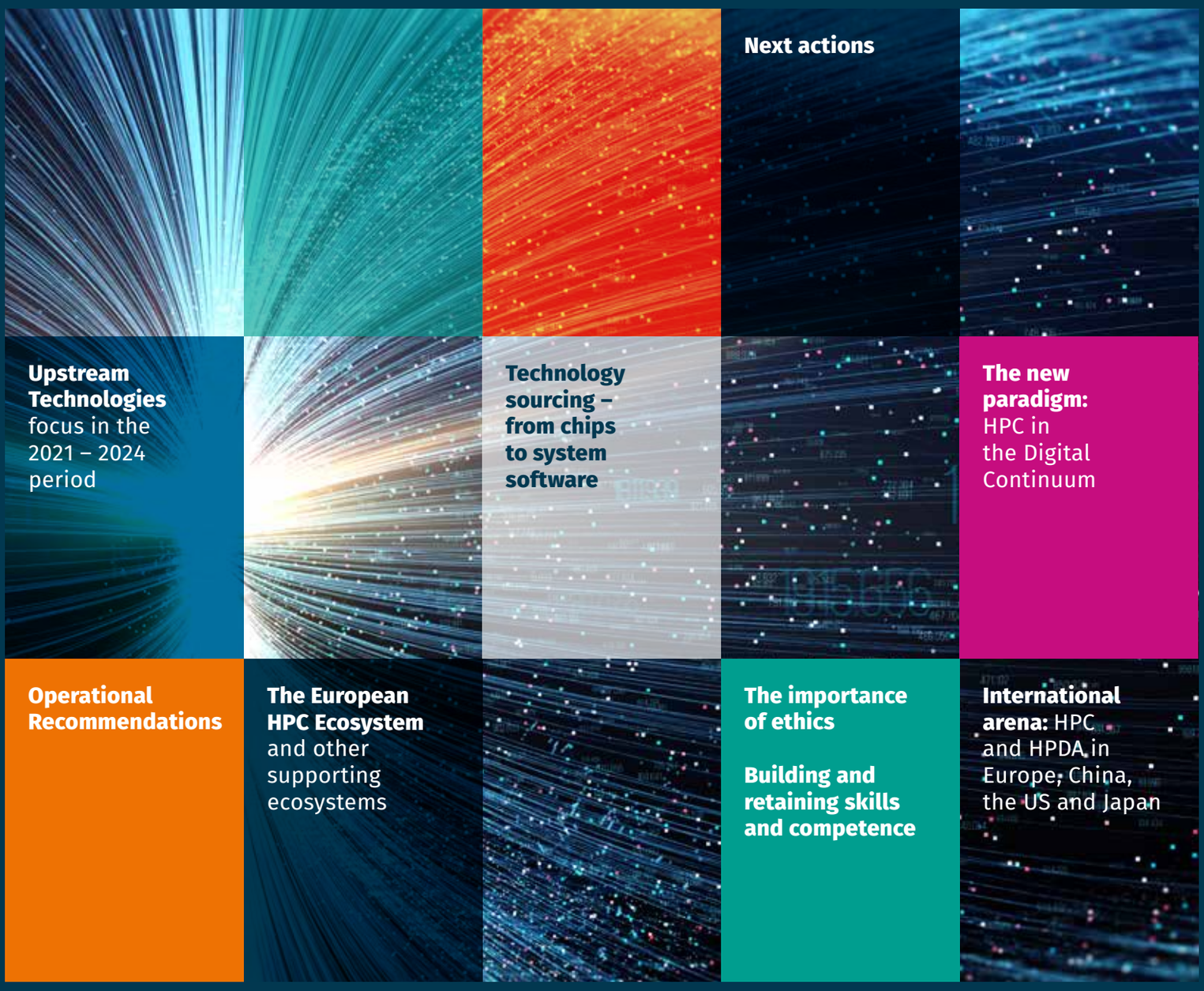
e dci European Extreme Data & Computing Initiative

2020 ETP4HPC's SRA 4

STRATEGIC RESEARCH AGENDA FOR HIGH-PERFORMANCE COMPUTING IN EUROPE

MARCH 2020

EUROPEAN HPC RESEARCH PRIORITIES 2021 - 2024



ETP4HPC Chairman's Message: The future role of HPC

This issue of our Strategic Research Agenda (SRA) is special as its role now extends beyond the role of the previous SRAs. Its findings will affect the distribution of an increased amount of funding dedicated within EuroHPC to the development of HPC technology and the mechanisms used to stimulate other areas **in the period from 2021 to 2024**. This SRA lays the foundation for this process by defining the European HPC technological priorities based on the input generated by the European HPC technology stakeholders: European HPC vendors, research organisations and users as well as the stakeholders of the related technologies.

This SRA approaches research priorities from a different angle. I urge the Reader to familiarise themselves with the concept of Research Clusters - an idea that I think we should continue - which represent the current challenges in the realm what we call **"The Digital Continuum"**. Thus, we provide flexibility in the definition of the research priorities of the future HPC research work programmes. These priorities could be derived from this SRA in three ways: by analysing the contents of the traditional Research Domains, by analysing the contents of the Research Clusters or by looking at where the two concepts intertwine - **the core of our future work is constituted by the HPC solutions needed to address the complex challenges of the Continuum**.

I would like to thank all ETP4HPC experts who took part in the writing of this SRA. A special mention is in order for our collaborators from outside of ETP4HPC – our partners in BDVA, HiPEAC, AIOTI, BDEC and other entities representing the areas that will complement HPC in the solutions of the future. We are facing the challenge of working together in order to design systems, the complexity of which cross the boundaries of a single technology. We all agree that we are facing a paradigm shift, but we still need to figure out *what that 'next big thing' is going to be*. For example, will the role of HPC in the systems of the future be to enable the workflows connecting various technologies? This SRA should also instigate a discussion with our partners on how to collectively implement the infrastructure for 'The Digital Continuum' as outlined in this document. As an example, the "European Green Deal" is an area where multiple digital technologies need to collaborate horizontally in R&D projects.

Besides my role as the Chairperson of ETP4HPC, I also preside over one of the two advisory bodies of EuroHPC – its Research and Innovation Advisory Group (RIAG), in which our Association has a leading role. The RIAG will use this SRA as a reference for its recommendations to the EuroHPC Governance Board in the definition of the 2021 - 2024 Work Programmes.

In particular, I would like to thank the SRA 4 working group leaders and the members of the working groups, the ETP4HPC Office, and other experts who joined this collaborative effort facilitated by Michael Malms.

I am looking forward to working with all our contributors again.

ETP4HPC Chairman
Jean-Pierre Panziera

Executive summary

This Strategic Research Agenda is the fourth High Performance Computing (HPC) technology roadmap developed and maintained by ETP4HPC, the European High-Performance Computing Platform with the support of the EXDCI-2 project. It continues the tradition of a structured approach to the identification of key research objectives. **The main objective of this SRA is to identify the European technology research priorities in the area of High-Performance Computing (HPC) and High-Performance Data Analytics (HPDA), which should be used by EuroHPC to build its 2021 – 2024 Work Programme**.

Over eighty HPC experts associated with member organisations of ETP4HPC created this document in collaboration with external technical leaders representing those areas of technology that together with HPC form what we have come to call **"The Digital Continuum"**. This new concept well reflects the main trend of this SRA – it is not only about developing HPC technology in order to build competitive European HPC systems but also about making our HPC solutions work together with other related technologies - the material included in this SRA is also a result of our interactions with Big Data, Internet of Things (IoT), and Artificial Intelligence (AI) and Cyber Physical Systems (CPS).

The targeted audiences of this document are:

- EuroHPC Joint Undertaking (EuroHPC JU) and in particular its Research and Innovation Advisory Group, which will use the research objectives identified in this SRA to build its Multi-Annual Strategic Plan,
- entities interested in forming project consortia in response to the EuroHPC (and related) calls,
- anyone interested in the development of HPC technology in Europe.

Apart from the well-developed eight technical focal areas of HPC related technology, this documents also presents an argument in favour of placing HPC within the context of "**The Digital Continuum** spectrum". The role HPC can play in this new concept is illustrated by four advanced use cases, which emphasise the need to master a multitude of new challenges presented by the complex workflows exemplifying this Digital Continuum.

We also introduce a new approach to defining the contents of the future research programmes. We believe that the most important challenge of European HPC now is to serve the development of the Digital Continuum, i.e. the unison of HPC and related technologies. We use the concept of Research Clusters to represent the main challenges of this Continuum. We argue that the future research should focus on how HPC can serve those areas. The priorities of the future Work Programmes could be extracted from this SRA by looking at the traditional Research Domains or Research Clusters, or by looking at their intersections, where – we believe – the core challenges lie.

The current state of technology for implementing the next generation of HPC infrastructure in Europe is analysed, the challenges for the upcoming four years are outlined, and the most important research priorities explained in detail.

The advancements in HPC and HPDA in Europe are placed in the context of those in the US, China and Japan. Also, upstream technologies with the potential to impact commercially available technology within the next 5-10 years, such as nanoelectronics and photonics, are discussed. Several operational suggestions in relations to future work programmes conclude the document.

Following the issue of this SRA, ETP4HPC's plan is to reach out to the other stakeholders of the European Digital Continuum. The entire ecosystem should jointly propose synchronised research actions aimed at tackling the multi-disciplinary technical challenges which facilitate the solutions to the problems European society will face in the next ten years. Over the six months prior to the publication of this document, ETP4HPC had actively participated in a multitude of conferences and horizontal collaborative work sessions aimed at uniting the critical forces needed to achieve that objective. ■

How to read this document

This Strategic Research Agenda is the most complex one to date and, depending on the needs of the Reader, it could be read in various ways.:

The **core of the SRA** is contained in the chapter ***5 Technical Research Priorities 2021 – 2024***, which identifies the European HPC Technology research priorities for the period from 2021 to 2024. This part provides material for the definition of the corresponding Research Work Programme. The SRA provides three dimensions from which these priorities could be extracted: the traditional Research Domains, the Research Clusters, which represent the challenges of the Digital Continuum, and the intersections of both, where the main challenges lie. Those interested in the areas that ETP4HPC recommends should be addressed by the upcoming calls for proposals and research projects should read the sections under *5.3 Research Domains*. Each of these concludes with a section titled *Intersection with Research Clusters*, which defines the cross-cutting topics which need to be researched within the given research domain and thus become part of the Research Work Programme. Additional technical expertise is included in chapter *6 Upstream technologies – focus in the 2021-2024 period*.

A quick overview of the SRA can be provided by the following parts: *Executive Summary, 1 Introduction, 5.1 The concept of Research Clusters and Research Domains, 6 Upstream technologies – focus in the 2021-2024 period, 3 The European HPC Ecosystem and other supporting ecosystems, 4 International arena: HPC and HPDA in Europe, China, the US and Japan and 11 Next actions.*

In order to gain **a more thorough understanding** of this SRA, the Readers should familiarise themselves with the contents of our Blueprint, which is the base of this SRA – reflected in chapter *2 The new paradigm: HPC in the Digital Continuum. This part includes: Application and use case scenarios and HPC use patterns: industrial and scientific use cases. Then, we encourage the Reader to select a few parts of chapters 5 Technical Research Priorities 2021 – 2024 and 6 Upstream technologies – focus in the 2021-2024 period and possibly 3 The European HPC Ecosystem and other supporting ecosystems, 4 International arena: HPC and HPDA in Europe, China, the US and Japan, 8 The importance of ethics, 9 Building and retaining skills and competence.*

Those interested in **applications and use cases** should read the section *2.3 Examples of industrial and scientific use cases* and these two Research Domains: *5.3.6 Mathematical Methods and Algorithms and 5.3.7 Application Co-design.*

The **recommendations** defined by ETP4HPC in relation to the **implementation** of the next Work Programme (the structure of the calls for proposals and the projects, the management and coordination of the projects) are in the chapters *10 Operational recommendations and 7 Technology sourcing – from chips to system software.* ■

Table of Contents

1 Introduction

1.1

EuroHPC – the driving force of European HPC

In order to position the value and purpose of this SRA, it is helpful to outline the political context it is created for. The EuroHPC Joint Undertaking, implemented in November 2018, is a joint initiative of the EU and European countries aimed at developing a World Class Supercomputing Ecosystem in Europe[1]. This partnership is intended to pool EU and national resources in High-Performance Computing with the initial objective of:

1. Acquiring and providing a world-class petascale and pre-Exascale supercomputing and data infrastructure for Europe's scientific, industrial and public users, matching their demanding application requirements by 2020. This would be widely available to users from the public and private sector, to be used primarily for research purposes.
2. Supporting an ambitious research and innovation agenda to develop and maintain in the EU a world-class High-Performance Computing ecosystem, Exascale and beyond, covering all scientific and industrial value chain segments, including low-power processor and middleware technologies, algorithms and code design, applications and systems, services and engineering, interconnections, know-how and skills for the next generation supercomputing era.

1.2

The role of this Strategic Research Agenda

The main role of this Strategic Research Agenda extends beyond that of the previous three SRAs: besides outlining research priorities in the area of technology – hardware and software throughout the entire stack of HPC IT infrastructure for the next 3-4 years, it also develops a sophisticated vision of the evolution of HPC in the next era of deployment. As outlined in Figure 1, the document feeds this information into EuroHPC's Research and Innovation Advisory Group as recommendations for the definition of the upcoming research calls to be launched in 2021 and 2022.

The major part of the document has been developed by SRA working groups, composed of technical experts recruited from ETP4HPC member organisations. The leaders and co-leaders of these working groups are well-recognised HPC specialists within the European and international HPC community (see *13.2 Acknowledgements* on page 104). In addition, a number of partner organisations have provided significant contributions e.g. the "Big Data and Extreme-Scale Computing" (BDEC-2) project, the "High Performance and Embedded Architecture and Compilation" (HiPEAC) project, the "Alliance for Internet Of Things innovation" (AIOTI), the "Big Data Value Association" (BDVA), the "Centres of Excellence for Computing Applications" (CoE) projects, the "Extreme Data and Computing Initiative 2" (EXDCI-2) project and the "European Organisation for Cyber Security" (ECSO) – see chapter *3 The European HPC Ecosystem and other supporting ecosystems* on page 28 for the complete list.

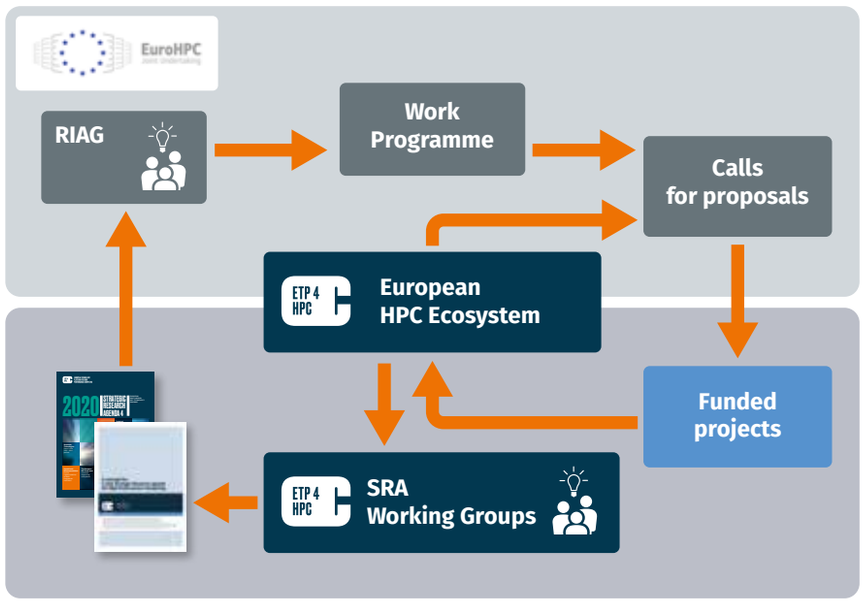


Figure 1: The role of SRA 4 - it reflects the principles defined in the ETP4HPC Blueprint (2019) and its main role is to feed the European HPC technology research priorities into the EuroHPC RIAG.

The SRA is meant to describe the major trends in the deployment of HPC and HPDA methods and systems, driven by economic and societal needs in Europe, taking into account the changes expected in the technologies and architectures of the expanding underlying IT infrastructure. The goal is to draw a complete picture of the state of the art and the challenges for the next 3-4 years rather than to focus on specific technologies, implementations or solutions. Any reference to products or solutions is intended as a reference in order to better explain the context and not as an implied promotion. The SRA thus remains completely agnostic in relation to brands and it maintains the diversity of implementation options. In this regard, it differs from the planning documents issued by EuroHPC and its RIAG, which delineate the implementation of research priorities in the form of work programmes and calls. These documents are driven by two factors: 1/ the political strategy agreed upon by the EC and the Participating States within EuroHPC and 2/ the relevant technical directions.

1. For more information on EuroHPC's mission, its organisational structure and the set-up of the RIAG, refer to the EuroHPC web site. https://eurohpc-ju.europa.eu/

1.3

The structure of this document

In April 2019, ETP4HPC issued a paper[2] titled "Blueprint for a new SRA" which outlines a layered, structured approach (Figure 2) to the identification of the research objectives in the 2021-2024 timeframe in the area of HPC and HPDA, covered by this SRA. This model includes significant collaborations with Internet of Things, Cyber Physical Systems and Artificial Intelligence. Its components are:

- The top-centre layer represents the political framework which aims to extend the use of HPC and innovation in technology provision in Europe. Being part of the "Single Digital Market Strategy", the next Multi-Annual Funding Framework 2020-2027 (MFF) of the European Commission includes the "Digital Europe programme"[3] to fund digital transformation beyond 2020 and "Horizon Europe" with "Thematic Clusters" and "Missions" which contain societal challenges whose resolution require investment in Research and Innovation (R&I) in HPC and HPDA. Five thematic clusters address the full spectrum of global challenges through top-down collaborative R&I activities. A small number of missions with specific goals leads to a comprehensive portfolio of projects, which cut across multiple clusters. The first missions are to be introduced in the first strategic planning phase of Horizon Europe[4].
- The second source of drivers of future technology improvements is represented in the upper right corner by commercial and industrial users of HPC. Especially in this category, new HPC use-patterns HPC are emerging in the context of new products and services (section *2.3 Examples of industrial and scientific use cases* on page 20).
- Science has a well-established role in providing major users and driving the architectural development of HPC systems. Although some of the scientific fields are also addressed by the thematic clusters and missions, it is important to acknowledge the influence of all scientific domains (section *2.3 Examples of industrial and scientific use cases* on page 20).
- The next layer down ("Application and use scenarios") translates the use-cases into application and technology use scenarios across the domains of "modelling and simulation",

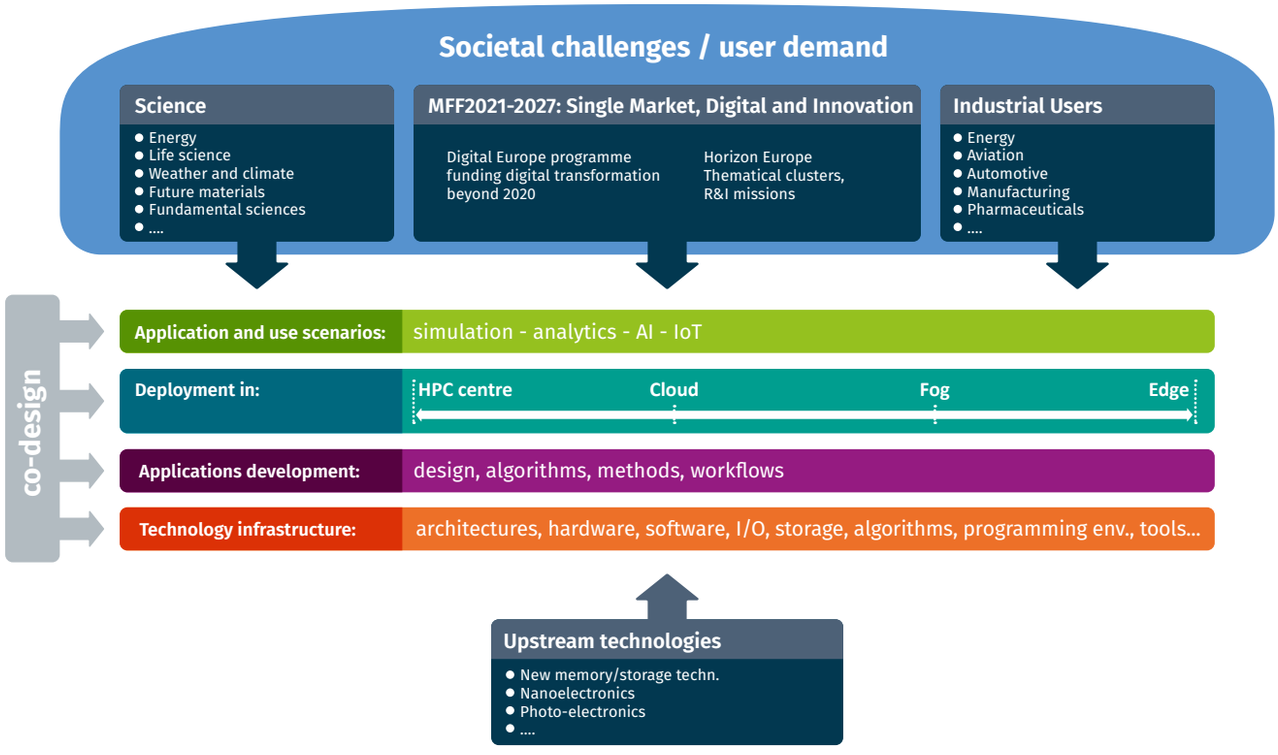


Figure 2: This SRA's structured approach to derive the research priorities for HPC technology and its application.

2. ETP4HPC, "A Blueprint for a new Strategic Research Agenda for High Performance Computing", 2019, https://www.etp4hpc.eu/pujades/files/Blueprint%20document_20190904.pdf or https://www.etp4hpc.eu/hpc-vision-018.html
3. https://ec.europa.eu/digital-single-market/en/policies/shaping-digital-single-market
4. The examples shown here are preliminary and taken from the Mazzacuto report available at https://ec.europa.eu/info/sites/info/files/mazzucato_report_2018.pdf

"AI", "Analytics" and "Internet of things". As argued below, these domains can no longer be handled separately as they are all required to implement solutions to the social challenges and other problems.

- HPC technology will not only be deployed in dedicated data centres in the future. A federation of systems and functions with a consistent communication and management mechanism across all participating systems will be required, thus creating a "continuum" of computing. The layer "Deployment" describes the challenges associated with this change whereas HPC functionality is now extended to Clouds, Fog computing and Edge computing (section *5.3.8 Centre-to-Edge Framework* on page 76).
- The next layer down ("Applications development: design, algorithms, methods, workflows") addresses the software development aspects of the application portfolio (section *5.3.7 Application Co-design* on page 70).
- The lower layer outlines the technologies used to implement the IT infrastructure discussed above. While most of the described components, functions and features will be deployed in data centres, local small-scale deployments (Edge/Fog) will also integrate the technology stack or a part of it. These technologies cover algorithms, programming languages and tools, system software, architectures, hardware components, I/O and storage as well as addressing critical features such as reliability and energy efficiency. **This is the core of the SRA** (chapter *5 Technical Research Priorities 2021 – 2024* on page 34).
- The emergence of upstream technologies which could be applied in future HPC system/component architectures constitutes another factor which influences the entire European technology domain. These upstream technologies are expected to facilitate novel and superior solutions. The related chapter outlines those candidate technologies which are most likely to be applicable within the timeframe of Horizon Europe (chapter 6 Upstream technologies – focus in the 2021-2024 period on page 82).

This document also provides **non-technical** analyses and priorities (as opposed to the technical contents listed above), which facilitates the understanding the context of the core technical chapters:

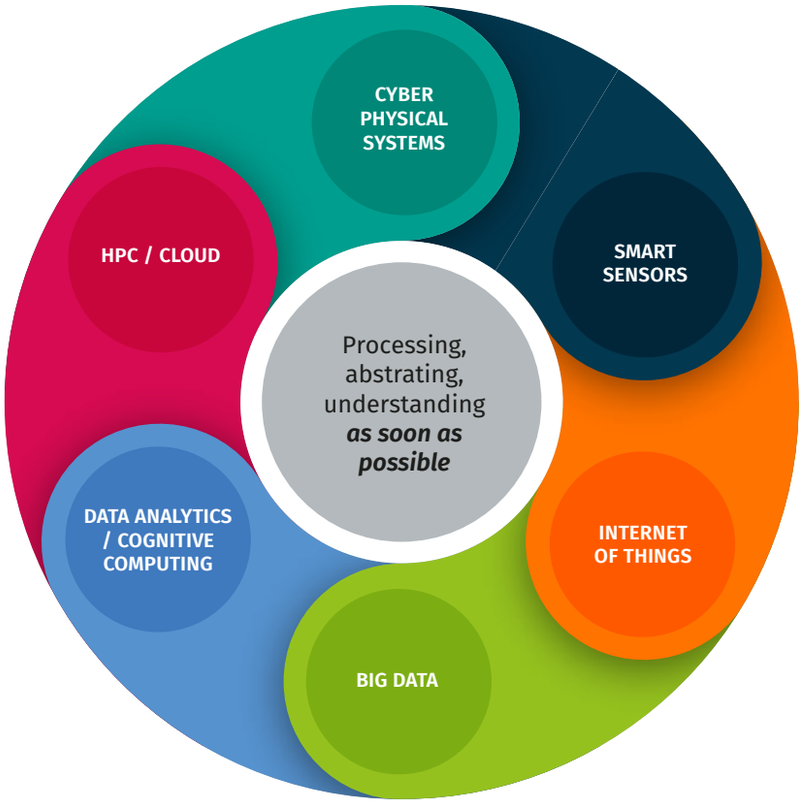
- A separate chapter is dedicated to the international HPC and HPDA arena. It presents an overview of the most prominent ecosystems, namely those of the US, Japan and China, and their strategies and achievements in order to understand how Europe can either benefit from their work or compete with them effectively (*4 International arena: HPC and HPDA in Europe, China, the US and Japan* on page 32).
- ETP4HPC also presents its view on the possibilities of implementing the research priorities outline in the previous chapter. Various elements of Open Source implementations well as the realisation of a European-sourced technology roadmap are outlined (*7 Technology sourcing – from chips to system software* on page 86).
- This SRA's operational recommendations for the implementation of Work Programme 2021/2022 are also presented. These include proposals for project types and support instruments not used in the previous HPC work programmes (*10 Operational recommendations* on page 92).
- Finally, the next objectives to be pursued in early 2020 are presented: ETP4HPC intends to engage with all stakeholders contributing their technologies to the "Digital Continuum" with an aim to propose large, high-TRL integration and demonstration R&I actions, which would validate the interoperability, robustness, and efficiency of the IT infrastructure supporting the Continuum (*11 Next actions* on page 108). ■

2

The new paradigm: HPC in the Digital Continuum

The rapid proliferation of digital data generators, the unprecedented growth in the volume and diversity of the data they generate, and the intense evolution of the methods for analysing and using that data are radically reshaping the landscape of scientific computing. The most critical problems involve the logistics of wide-area, multistage workflows that move back and forth across the computing continuum, between the multitude of distributed sensors, instruments and other devices at the network's Edge and the centralised resources of commercial clouds and HPC centres[5]. The objective of this SRA is to process this new paradigm of 'The Digital Continuum'.

5. BDEC, "Big data and extreme-scale computing: Pathways to Convergence - Toward a shaping strategy for a future software and data ecosystem for scientific inquiry", 2018, https://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/bdec_pathways.pdf



Enabling Intelligent data processing at the edge:

- Edge computing
- Fog computing
- Stream analytics

Transforming data into information as soon as possible

Collaboration between edge devices and the HPC/cloud ensuring:

- Data security and Privacy
- Lower bandwidth
- Better use of HPC/Cloud

> creating a continuous flow

Figure 3: HPC in the loop.

Figure 3 above illustrates High-Performance Computing as one element of a complex workflow ("HPC in the loop"), starting with data generated at smart sensors in an IoT environment. Data is being locally pre-processed at the Edge and relevant parts are forwarded to decentralised "Fog nodes" close to the Edge. A subset of data is then transferred for centralised Data Analytics or simulation and modelling in centralised HPC centres or clouds. In an increasing number of use scenarios based on the concept of the "Digital Twin", a "twin-copy" of a physical entity, is maintained and continuously updated on these central compute infrastructures (see section *5.3.8 Centre-to-Edge Framework* on page 76). It should be noted that in reality the dependencies between the segments of the loop shown in the Figure are not sequential in nature. The loop is not strictly repeating actions in a circular mode; the elements are cross-connected in rather complex, often fast changing event-driven flows.

The final outcome of the loop is a set of optimised actions in the "Cyber Physical Entanglement" representing physical systems (e.g. robots, vehicles, industrial processes) interconnected in complex intelligent networks.

2.1

Application and use case scenarios

In light of the rapid evolution of technology and use cases, the term "High-Performance Computing (HPC)" needs to be redefined: In the past, it used to be synonymous with "technical computing using supercomputers" applied in order to model or simulate complex scientific or technical phenomena. While HPC will still refer to systems facilitating scaling of applications to a larger number of nodes, the main change is that HPC systems will no longer be stand-alone systems but they will be part of a larger e-infrastructure to realise complex, efficiently managed and orchestrated workflows, including the interfaces of this structure with external devices (distributed and Edge devices), as indicated in Figure 3.

Tight integration of capabilities across individual system boundaries and between data centres and local small-scale HPC systems is expected. Each component in this integrated compute, communication and data infrastructure has different characteristics that can be summarised as:

- **Simulation**: relatively low amount of input data, large computation requirements (mostly in double or quadruple

precision floating point representation) with tight coupling between compute nodes (benefits from scale-up hardware and low-latency networks) and large amount of generated data (simulation results).

- **Big Data**: large amount of external input data, low-to-medium computational requirement with loose coupling between compute nodes (scale-out and "shared-nothing" models) and low amount of output data (information extracted from the input data).
- **Data stream processing**: streaming capabilities are becoming increasingly important for scientific and industrial HPC applications (e.g. CERN's Large Hadron Collider (LHC), Square Kilometre Array (SKA) project, astrophysics, physical simulations, digital twins, etc.), supporting important needs such as the ability to act on incoming data and computational steering. Coupling data streams produced by such experiments to computational HPC capabilities is an important challenge, and Big Data Computing's near real-time processing architectures and stream processing capabilities are able to rapidly analyse high-bandwidth, high-throughput streaming data.
- **AI** (for example, Machine Learning in the **training** phase): large input (local) database with very high access rate, large amount of computation (in low precision floating point representation) and relatively low amount of output data (the weights of the newly trained Neural Networks, typically a few hundreds of MBs).
- **AI** (for example, Machine Learning in the **inference** phase): medium input (depends on the application), low processing amount (reduced precision floating point or integer) and low amount of generated data.
- **AI** (**Reinforcement** learning) such as Alpha Zero system from DeepMind: the input is low (in volume, e.g. rules of a game, or physical laws or constraints), the output is also low (solution), but the system internally generates a large amount of data and computation to explore the various options and find a good solution. Simulation of the process to be optimised is in the loop to get an assessment of the quality of the solution found.

HPC has always advanced science by delivering results only made possible by the use of cutting-edge computer technologies. Throughout the last decade numerical computing has been growing rapidly in many directions: higher fidelity, coupled multi-physics and multi-scale models; a deluge of observational data from sensors and of simulated data; semi-automatic data analysis and post-processing; uncertainty quantification and newly AI-based models. Combining all these aspects will result in a highly complex application (software) architecture, which is becoming a research topic by itself.

In reference to Figure 1, this layer is driven by the thematic clusters and missions as well by industrial and scientific needs. The extraction of IT/HPC requirements out of representative and strategically important use case scenarios is necessary in order to drive HPC R&I in the right direction. They are key to assess new architectures or infrastructure as well as to provide testbeds to research and industrial teams.

In the context of promoting innovations for the HPC, HPDA and IoT ecosystem, the use cases identified must be such that we avoid alignment with technology "silos", which would strongly restrict the shaping capabilities for the R&I work program. Furthermore, fully addressing the societal challenges can only be achieved by considering end-to-end approaches where data production is integrated with data analytics, machine learning, numerical simulation, data archiving as well as the final use of the results. The use cases are based on applications which rely on complex workflows within which individual tasks are executed on a wide variety of systems and whereby the complete data management cycle is addressed.

However, many representative use case scenarios are difficult to analyse because they combine many heterogeneous components (e.g. which rely on different software stacks) as well as different resources or user governance strategies. For instance, the main challenge is presented by the existence of applications across a federation of systems - which includes HPC centres, Cloud facilities, Fog and Edge components and networks - while at the same time, preserving security and privacy from end-to-end. Furthermore, the economics aspects of the deployment of these applications must be considered.

To advance the state of the art, the supported uses cases must be able to demonstrate an application implemented over multiple entities while preserving security and privacy properties. Furthermore, their deployment should take place in an efficient manner (technically and economically).

Mapping the relationship between Simulation, Data Analytics and Machine Learning into a real environment, as illustrated in Figure 2, shows a loop of actions with HPC being one of the many elements besides Data Analytics, the Internet of Things and in many cases, cyber-physical entanglement (systems). Increasingly, computing systems act as direct controlling devices, which thus has an impact on the real world (Cyber-Physical Systems). HPC in the loop or Digital twin approaches add timing constraints so that the results of simulations can be directly used for choosing the adequate control of the system in due (real) time, and raise the stakes of validating and guaranteeing functional correctness, timing and security, because faults or breaches will have wide-ranging consequences in the real world.

2.1.1 Workflow and capabilities

Understanding the workflow and dataflows is of crucial importance for an analysis of real use cases.

Each use case (e.g. autonomous driving, personalised medicine, wind park operation, etc.) is built upon some basic "functional

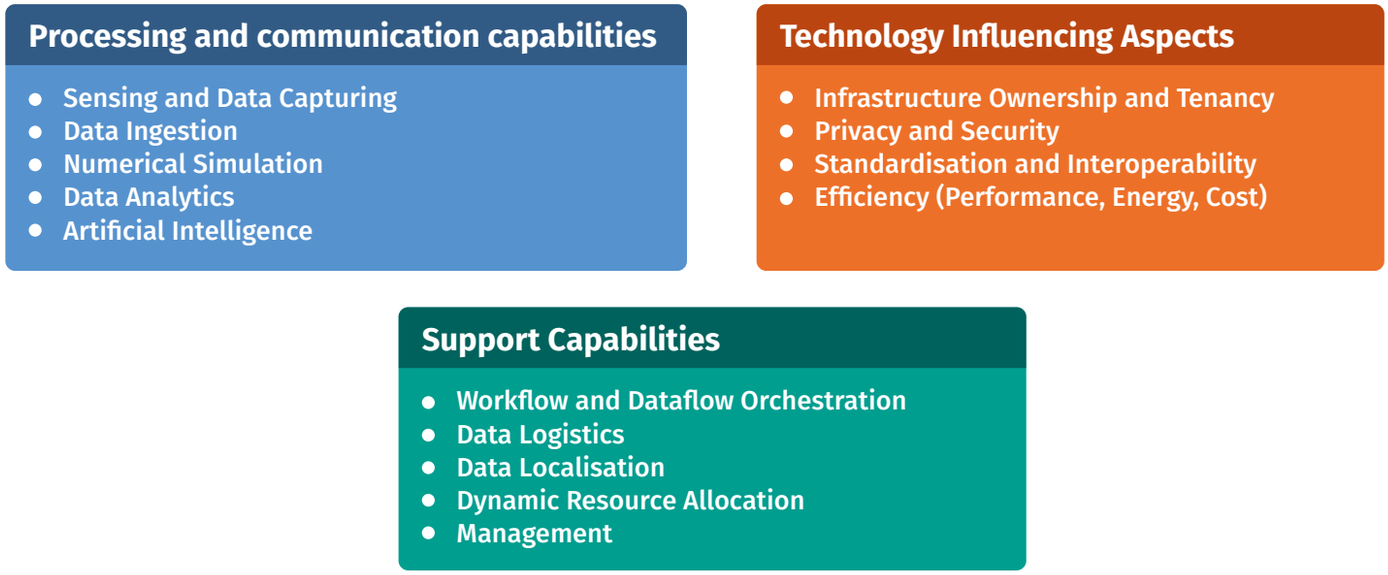


Figure 4: Categories of capabilities in mixed Simulation, Analytics, AI and IoT use scenarios

capabilities" (Figure 4), which are implemented in the form of similar structures (Figure 5).

- The "Processing and Communication Capabilities" listed in Figure 4 cover all areas which require compute capabilities, be it in a data centre, an Edge or Fog node or an IoT device – each of them with a different application scope. For a given workflow (use case), the individual processing capabilities are expected to be spread accordingly across locations and systems. We distinguish between data capture from devices, data ingestion into a compute environment, the typical HPC capability of numerical simulation, and the Big Data capabilities of data analysis and artificial intelligence. To address such new compute requirements, HPC capabilities must provide the processing capabilities for the Big Data environment, which includes interactive analytics as well as batch and real-time processing of data streams.
- The "Technology Influencing Aspects" are properties that have a large impact on the design, implementation and integration of the processing capabilities but do not directly provide any data processing capabilities. These properties must be provided by the processing infrastructure in ways that satisfy the end-user requirements to result in an effective and efficient solution. The governance of compute infrastructure and data imposes policies on data processing. In most use cases, security and privacy must be considered in such an environment to comply with regulatory and end-user needs. Interoperability and standards increase trust in developed workflows and accelerate the adoption by users. The efficiency of a solution is relevant insofar that the costs of a solution limit its adoption in use-cases with limited revenue. A well-performing, energy and cost-efficient system maximises industrial and commercial competence by enabling novel scenarios.
- "Support Capabilities" describe the crucial implementation aspects of a mixed scenario. As shown in Figure 4, the workflow reflects the interconnections of actions and data between the IoT devices, processing entities and data repositories. However, the identified capabilities for the environment discussed here are currently underdeveloped and require further R&D efforts.

The orchestration of workflows and automatic and efficient deployment across a complex hardware-landscape is required to exploit such systems. For instance, data must be placed and migrated intelligently to match the storage and processing capabilities of (IoT or Edge) systems. Alternatively, processing or computation can be distributed/mapped onto the physical infrastructure in a way that minimises data movements (moving the computation to the data instead of moving the data to the computation capability). Finally, workflows must adapt their processing capabilities dynamically depending on the input, or other external parameters such as the number of users or availability of processing capacity. This requires software layers that enable such dynamic, ad hoc changes.

We recognise that management procedures must be developed that deal with the distributed nature of computation, ownership, and conformance to standards while considering the efficiency aspects.

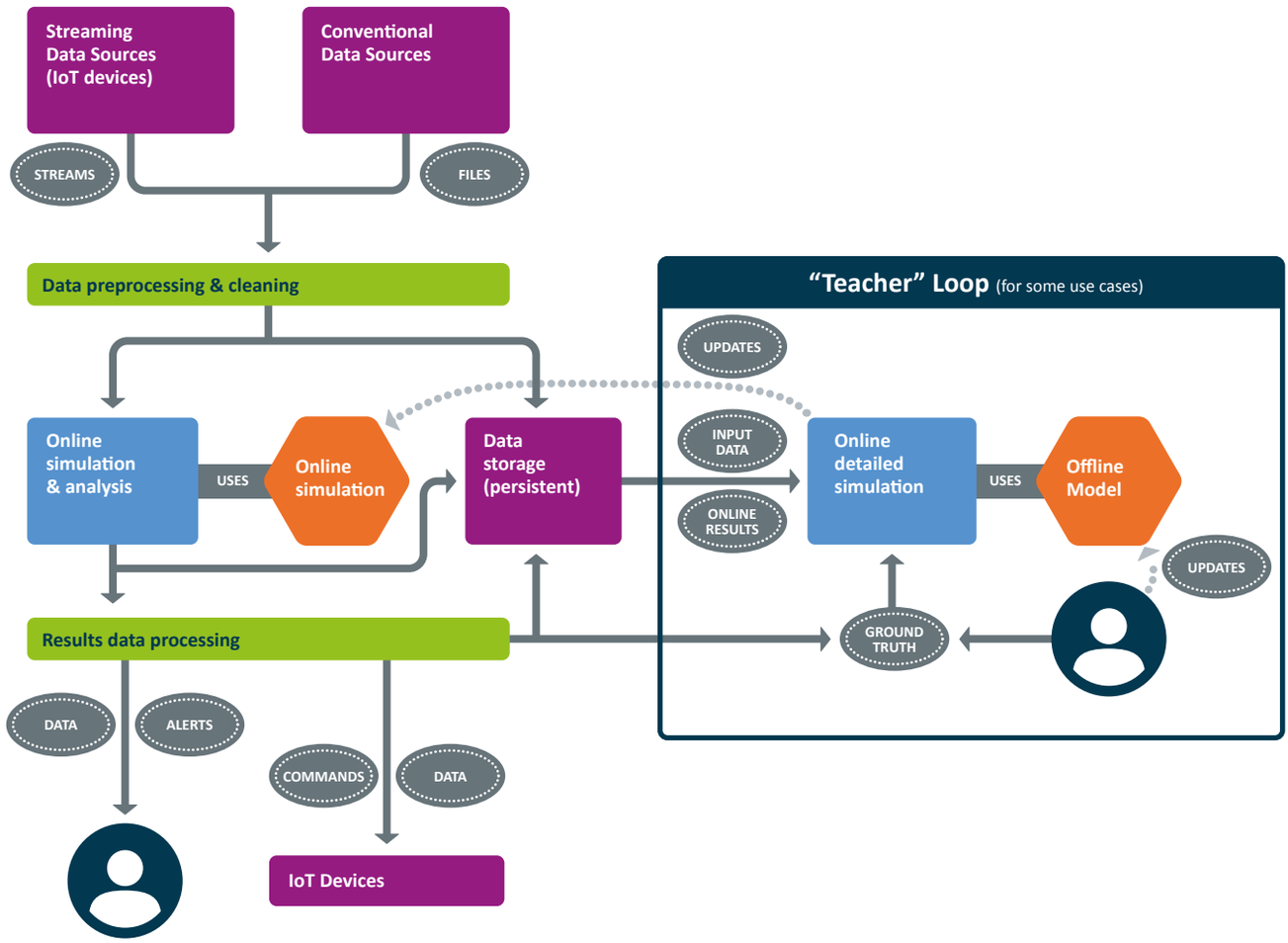


Figure 5: A typical mixed simulation and machine learning workflow

Figure 5 *unfolds the loop shown in* Figure 3 *and shows three steps that are common to the use cases discussed jointly by ETP4HPC and the BDVA: in the first step, data from a multitude of real-world sensors or conventional sources (e.g. databases) is ingested, pre-processed and cleaned. This can already involve significant processing, as in situations where the analysis of correlation between independent data streams is required. All or part of the resulting data is put into storage for documentation and for use in improving the analysis/simulation models.*

The second step consists of an in-depth analysis of the data from step 1 – this includes anything from image classification to computing the next status of a complex digital twin using multi-discipline simulation techniques. The online model encodes the analysis steps, and it can range from a simple rule set to a complex HPC simulation code. A part of the analysis results are again put into storage for later use.

The third step is the processing of the analysis results, and communication with human users or IoT devices/Cyber- Physical Systems (CPSs). Depending on the nature of the problem, the loop can be closed by the commands passed to a CPS affecting its sensor readings, which requires the update of the analysis in step 2. In the Digital Twin case, the analysis in step 2 keeps its own state and runs in "streaming mode", receiving updates from the real world, reconciling them with the CPS's virtual model, and sending out commands to the CPS.

The role of the "teacher loop" is very apparent for Deep Learning based analysis approaches – the online model at the heart of step 2 is created in a separate training phase and then made "live". For reinforcement learning, the online model is improved by assessing its performance and rewarding/punishing certain aspects. Taken to the next step, the online model could represent a simplified version of a car (for example), which is updated and extended/improved by a full, physically correct car model. The key idea behind splitting off the teacher loop is that it can be disconnected after a while (analogously to real life with teachers and pupils, once a certain proficiency has been achieved). The online model can then be made significantly simpler than e.g. a fully physically correct six degrees of freedom driving model, reducing the amount of processing needed per instance and consequently reducing energy requirements.

2.1.2

Data life cycle and dataflow in a scientific environment: an example

Understanding the necessity for a dataflow orchestration in mixed Simulation and Big Data use scenarios is important. The capacity of storage infrastructure, the increased sophistication and deployment of sensors, the ubiquitous availability of computer clusters, allow the development of new analysis techniques and real time capabilities to ingest "fresh" data during simulation.

There are multiple scenarios:

- Input data coming from experimentations is injected into simulation to enhance it. In this case, the improvement of the simulation will depend on the availability and the quality of this new data set.
- Data is produced by sensors in a streaming mode and local (close to the Edge) or remote HPC resources are used to train the model. The model-training frequency will depend on data source obsolescence.
- Output or step-by-step data can be extracted from simulation for new in situ processing, visualisation and simulation context modification (computational steering). in situ processing could be performed using AI for inferring on the fly pertinent structures (and by consequence reduce the amount of data to be saved, saving also energy) and performing a continuous but light training of the existing models

For these new scenarios, we observe the need for different levels of curation (sensors producing non-curated data versus use of databases with curated data):

- Unstructured data issued e.g. by major scientific instruments or experimental facilities, which may be residing outside of supercomputer centralised facilities, will require non-trivial transformations before an ingestion could be realised by a simulation or Machine Learning or other HPDA steps. Depending on the real-time availability and quality of this data, the transformation and availability for simulation need a strong coordination effort (near real time data preparation).
- Qualified structured data resources shared by the communities through archives, databases or any specifics formats accessible through the internet have a well-known preparation process to enable their use in a simulation.

The challenge here is to add and to coordinate the integration of these new data resource types in end-to-end application workflows without drastically increasing storage space dedicated to data availability. A well-balanced architecture will mostly depend on the efficiency of the dataflow and on the capability to reduce, filter, pre-process data close to the source (on Edge computing devices or Fog nodes). The objective is to limit network and global storage congestion.

This distributed data transformation must be integrated in a Big Data life cycle model that includes activities aimed at combining data curation with the research life cycle more closely. The activities address planning, acquiring, preparing, analysing, preserving, and discovering data, describing the data and assuring its quality.

The relationship between scientific community data repositories and new distributed data workflows as well as the reproducibility in computational science need to be understood. Documenting data sources, experimental conditions, instruments and sensors, simulation scripts, processing of datasets, analysis parameters, thresholds, and analysis methods ensures not only a much-needed transparency of the research, but also data discovery and future data use in science.

Important as well is the notion of where data is stored and how/where data is accessed for computation. In a federated scenario, data could be stored across distributed Edge, Fog and possibly multiple centralised "data-centre-like" systems, e.g. reflecting the data production sites or specific access policies. Solutions to allow simulation, analytics or AI applications access data across federated and heterogeneous sites must be designed and built to strike the proper balance between data access performance, cost and consistency, while at the same time satisfying access control and privacy constraints.

In conclusion, the design of a global infrastructure allowing the combination of external Edge- or peripheral environments with a central, shared infrastructure will require the analysis of the entire software environment, the identification of new data sources and of the quality of data.

2.2

HPC use patterns: industrial and scientific use cases

Ten to fifteen years ago, only a number of selected specific domains used HPC. Today's widespread use of HPC is a result of its expansion into a wide range and scientific and industrial fields, for example:

- In **engineering** (e.g. automotive/aero-spatial industries), HPC is now used widely to simulate complex multi-physics systems, such as the ones used in the analysis of combustion engines, aerodynamic properties, or vehicle safety at high precision.
- In the **domain of natural resources** (the oil & gas industry, in particular) HPC is traditionally widely used in resource or production management (e.g. oil search).
- HPC is used in **industrial production** (e.g. in the pharmaceutical industry in drug design) and in t**he design and testing of complex technical processes** used.
- The **financial sector** too relies heavily on HPC which is applied in real-time simulations.

An average HPC user was thus a large company, operating its own

HPC centre on its own premises, having at its disposal experts who operated the system and ran their compute intensive applications. The applications were either developed by specialised ISVs (e.g. for the automotive sector), or developed jointly with a mostly academic open source community or, in some cases, they were developed in-house and closely protected.

The change we are observing today is mainly driven by two factors:

- **HPC as a service**: HPC resources being available today "as a service" (typically proposed by Cloud service providers), make simulation of (multi-) physical systems available to a much wider range of users, often in a multi-tenanted set-up. In this set-up, neither ownership of the computing resources is required, nor highly specialised in-house competences in HPC[6]. It should be noted that sharing an HPC infrastructure between different industrial users will not only add new requirements in terms of security for providing high levels of data protection and guaranteed isolation between users but also bring new challenges for scheduling and orchestration.
- **Data driven applications**: Furthermore, we see an increase in a wide range of new "data driven applications" deploying functionalities such as analysis of very large data sets and machine learning (relying on large data sets for the training phase), which have given rise to a wide range of new applications. The following two examples attract a lot of attention: e-mobility, including autonomous vehicles, and customising medication and drug consumption to the personal needs of a patient. A number of sectors (e.g. designing/operating wind turbines and industrial production processes) have started to deploy the "Digital Twin" concept: digital twins are software representations of assets and processes that are used to understand, predict, and optimise performance in order to achieve improved business outcomes. Digital twins consist of three components: a data model, a set of analytics or algorithms, and knowledge[7]. As shown in Figure 2, HPC simulation has moved "into the loop", and become an indispensable part of a product. This trend fundamentally changes the requirements on the HPC software, systems and integration/management. The HPC systems needs to facilitate the connection of external sensors/Edge computing without compromising security and HPC systems protection.

As a consequence of this transformation, the needs of the industrial users have evolved for a number of reasons:

- First, today, users rely on the provision of HPC resources for all scales of computations and flavours (Data oriented, HPC oriented). This applies not only to small users without in-house resources but is also true in the case of large companies who need a seamless integration of in-house capacities with external secured Cloud-based capacities.
- Second, as reflected by the growing use of PRACE resources by European industries, as the use of HPC by European industry grows, the use of HPC must be available to a variety of players, even without highly specialised in-house resources. These new users would need to rely on services and support to guide them how to use HPC effectively in their businesses.
- Third, the European industry needs increased support in application development: to develop effective HPC applications is intrinsically difficult – and the adoption of such codes to new hardware (for example, to accelerators such as GPUs) requires detailed expertise. Access to novel and experimental system architectures is needed to allow users and application developers to prepare their codes for the next generation of machines.
- In the definition of the strategic research topics for the upcoming Horizon Europe framework, the input from industrial users is crucial in order to (I) address their technical needs by taking into account the key requirements of future industry-relevant applications, and (II) support the European industry at large in the uptake of numerical simulations and data driven applications for their businesses.

2.3

Examples of industrial and scientific use cases

The following **five USE CASES** illustrate how diverse and highly complex modern workflows can be. In all examples, the HPC infrastructure is used to either simulate or model, or it serves as an efficient, high performing infrastructure for Deep Learning applications. Also, besides the multi-player interdependencies at a technical level, many non-technical aspects such as security and privacy, multi-tenancy of communication and compute infrastructure and resilience are outlined.

The following are the names and the originators of the five use cases:

- USE CASE 1: **Extremes prediction in the Digital Continuum** by Peter Bauer, ECWMF
- USE CASE 2: **Autonomous Driving** by Ovidiu Vermesan, SINTEF and AIOTI
- USE CASE 3: **AI Automation on premise** by Cristiano Malossi, IBM Research
- USE CASE 4: **AQMO1,2: An Edge to HPC Digital Continuum for Air Quality** by Francois Bodin, IRISA and University of Rennes
- USE CASE 5: **FTRT – Faster Than Real Time for seismic, volcanic or tsunami events** by Stephane Requena (GENCI)

6. The projects SHAPE, Fortissimo and Fortissimo-2 demonstrated that SMEs could benefit greatly from access to such re-sources and support to solve business problems. The Fortissimo marketplace has been developed by the projects to offer such services.
7. https://www.ge.com/digital/applications/digital-twin

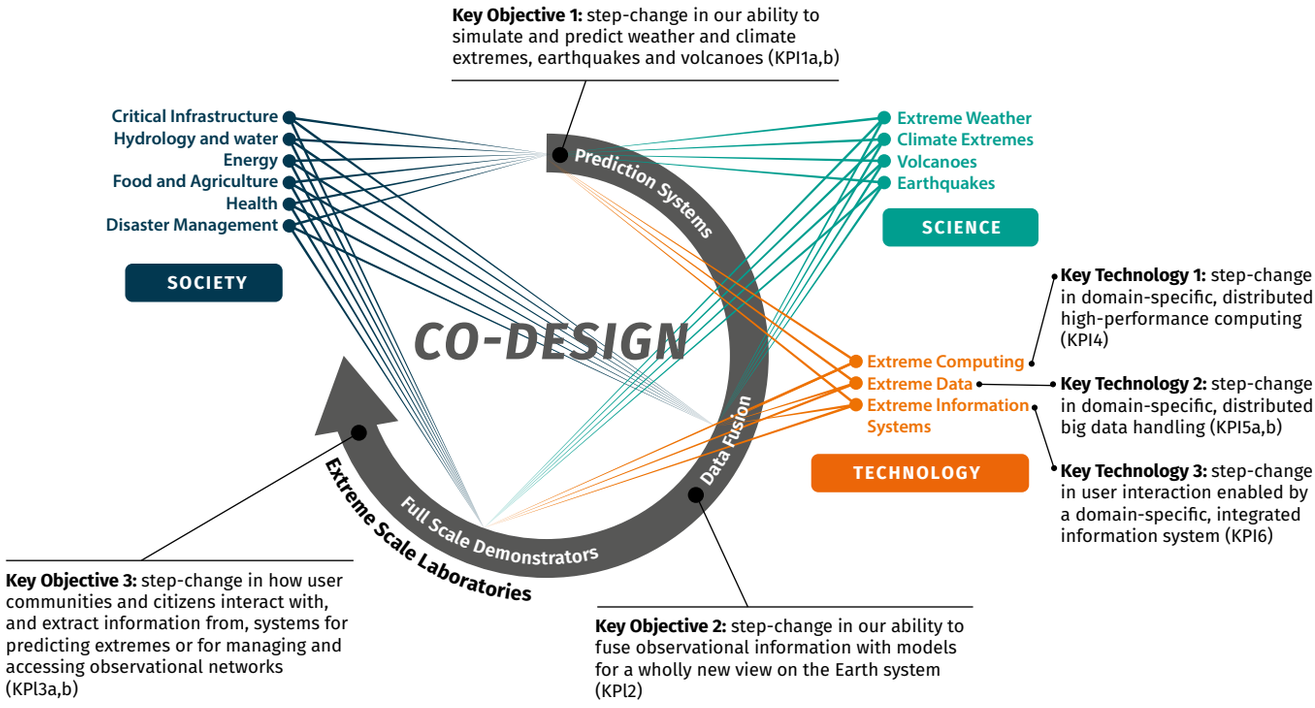


Figure 6: Extremes prediction loop (prediction of extreme climate conditions) from science to impact objectives, enabled by new technologies.

2.3.1

USE CASE 1: Extremes' prediction in the Digital Continuum

Natural hazards represent some of the most important socio-economic challenges our society is facing in the next decades. Natural hazards have caused over 1 million fatalities and over 3 trillion Euros economic loss world-wide in the last 20 years, and this trend is increasing given drastically rising resource demands and population growth. Apart from the impact of natural hazards on Europe itself, the increasing stress on global resources will enhance the political pressure on Europe through yet unprecedented levels of migration.

Dealing responsibly with extreme events does not only require a drastic change in the ways our society is solving its energy and population crises. It also requires a new capability of using present and future information on the Earth-system to reliably predict the occurrence and impact of such events. A breakthrough of Europe's prediction capability can be made manifest through science-technology solutions delivering yet unprecedented levels of predictive accuracy with real value for a society. This is the objective of the *ExtremeEarth* project proposal[8].

2.3.1.1

The relevance for Digital Europe – and beyond

The science-technology solution includes the entire loop from: (1) basic Earth-system science, established in (2) enhanced prediction models significantly, combined with (3) the vast range of Earth observation data ranging from advanced satellite instruments to commodity devices available through the internet of things, exploiting (4) extreme-scale computing, Big Data handling, high-performance data analytics and artificial intelligence technologies, for feeding (5) impact models that translate scientific data into information close to the real users responsible for critical infrastructures, hydrology and water, energy, food and agriculture, health and disaster management. Only through the full loop can the investment in breakthrough science-technology solutions be turned into value for society. This loop is shown together with key application objectives and key enabling technologies in Figure 6. How the elements of the Digital Continuum of this SRA map onto this application use case is explained in the following.

2.3.1.2

Physical systems

The main objective is to create a digital twin of the Earth-system that employs all available observations and high-definition simulations of the past, the present and the future in order to interactively extract information that can drive institutional and industrial processes. Near real-time data processing, on-demand high-performance computing, and integrated science-application workflows combined with high-performance data analytics capabilities form the technology backbone.

8. Full proposal for download - https://extremeearth.eu/sites/default/files/2019-07/ExtremeEarth_FullProposal_public.pdf

The institutional and industrial processes will allow the protection of critical infrastructures and the management of catastrophic consequence of extremes, the future design of resilient renewable energy sourcing, securing water and food supply, and the protection of human health and European society's response to political pressure caused by environmental change.

2.3.1.3

Smart sensors and Internet of Things

At present, Earth-system observations already comprise hundreds of millions of observations collected daily to monitor atmosphere, oceans, cryosphere, biosphere and the solid Earth, the largest data volumes being provided by hundreds of satellite instruments. This volume is expected to increase by several orders of magnitude in the next decade, with a need to ingest such observations in digital twin systems within hours. Smart-sensor technology is highly relevant for satellite-based observations and dedicated station networks, but also for observations from commodity devices deployed on e.g. phones, car sensors and specialised industrial devices monitoring agriculture, renewable energy sources and infrastructures – made available through the internet of things. Such technology will allow outsourcing data pre-processing to Edge and Fog computing, and thus implement fully agile data management and information extraction.

2.3.1.4

Big data and Data Analytics

The daily volume of Earth-system observation and simulation data already exceeds petabytes today, prohibiting effective and timely information extraction, critical for proactive and reactive response for anticipating and mitigating the effects of extremes. Both simulations and observations need to be generated and assimilated in the Earth-system's digital twin within minutes to hours of time-critical workflows towards near-real-time decision-making. Overcoming the data-transfer bottlenecks between the digital twin and downstream applications is crucial and future workflow management needs to make such applications an integral part of the observation and prediction infrastructure. Powerful data analytics technology and methodologies offer the only option to make the effective transfer between raw data and information tailored to those sectors needing to prepare and respond to extremes, namely water, food, energy, health, finance and civil protection.

2.3.1.5

HPC and Cloud

Today, experimental and operational Earth-system simulations use petascale HPC infrastructures, and the expectation is that future systems will require about 1000 times more computational power for producing reliable predictions of Earth-system extremes with lead times that are sufficient for society and industry to respond. This need translates into a new software paradigm to gain full and sustainable access to low-energy processing capabilities, dense memory hierarchies as well as post-processing and data dissemination pipelines that are optimally configured across centralised and Cloud-based facilities. European leadership in this software domain offers a unique opportunity to turn the European investment in HPC digital technology into real value.

2.3.2

USE CASE 2: Autonomous driving

The entire traditional transportation ecosystem is undergoing significant changes with five main trends accelerating this transformation:

- New mobility modes and behaviours,

Autonomous Vehicle View Source Sintef

- **Sensors, actuators, vision, maps**
- **Connectivity**
- **Vehicle architecture**
- **High-performance vehicle computing platform, flexible, and programmable**
- **Infrastructure platform**
- **AI design and implementation platform**
- **Edge computing platform and solution**
- **Data Center solution for fleet simulation and testing**
- **Pervasive security, safety trust program**

Figure 7: Critical technologies and platforms for autonomous vehicle driving

- A rise in autonomous/automated driving technologies,
- Development and use of digital features impacting industry and consumers,
- The electrification of powertrains and introduction of AI techniques,
- Methods for implementing intelligent solutions and components.

For the autonomous driving capabilities at level 4 and 5, new capabilities of autonomous vehicles need to be addressed in terms of computing, control, cognition, connectivity (4C attributes). Through sensing, detection, perception, processing and decision functions vehicles need to "see" (sense/locate) the surroundings, perceive obstacles and act safely in accordance with the vehicles' perceptions.

In addition, the **integration of multipurpose in-vehicle platforms** as shown in Figure 7, and the distribution of functions between automated vehicles, other vehicles, infrastructure, Edge/Cloud platforms and HPC centres must accommodate solutions for over-the-air (OTA) updates, predictive maintenance and vehicle-to-everything (V2E) connectivity.

The **convergence of several ecosystems** and standardisation activities means that intelligent electric, connected, autonomous/automated (ECA) vehicles will become ubiquitous devices among Internet of Vehicles (IoV) applications and services.

Vehicles will access, consume, create, enrich, direct and share digital information between businesses, people, organisations, infrastructures, vehicles, other elements of the Internet of Things (IoT) and Industrial Internet of Things (IIoT) applications. Mobile, Edge, Cloud and high-performance computing technologies together form a computing continuum for new, disruptive IoV and IoT/IIoT applications, providing an information pipeline for safety and mission-critical workflows.

These applications need an **Edge-to-Cloud Digital Continuum** where data/information flows are stored, processed and analysed, while using HPC centres for extensive traffic simulation, weather forecasting, virtual energy grid simulation and energy flow, and virtual fleet performance validation using digital twin representations of the vehicles etc. New end-to-end IoV and IoT/IIoT applications in this computing continuum require a scalable, composable and automated intelligent infrastructure.

To illustrate **the role of HPC**, for a safe and efficient journey of many connected and autonomously vehicles driving at the same time, an emphasis needs to be put on providing optimised environmental simulation at large scale to model their movement along millions of kilometres. In addition, simulation of potentially complex scenarios and the behaviour of vehicles is required to optimise their handling in different driving conditions (sun, rain, snow, twilight, night, backlit, etc.). Exascale HPC facilities can support the use of autonomous/automated vehicles in fleets, evaluating and optimising thousands of possible impediments that vehicles might encounter (bikes, motorcycles, trains, pedestrians, stop signs, potholes, black ice, traffic circles) and the need to provide scalable computing and storage systems necessary for safety and mission-critical applications.

A challenge will be the orchestration and optimisation of processing power across HPC-, Cloud- and Edge-computing infrastructures to balance the workload, and optimise and distribute the storage capabilities, creating real-time computing environments that respond to different usage needs.

At the Edge, i.e. in the autonomous vehicles, urgent decisions based on real-time data will have to be taken locally. In this context, **advanced AI technologies** must be distributed across the computing and information continuum (optimised based on processing capabilities and latency requirements). Autonomous/automated vehicles must be able to continuously learn and make optimal decisions under all circumstances.

One of the major challenges for future autonomous/automated vehicles and IoV applications is the dynamic orchestration and optimisation to access multiple intelligent infrastructures in the Digital Continuum and to cost-effectively manage the terabytes of information required at the right time and in the right location by users who need it for reliable, safe decision-making.

2.3.3
USE CASE 3: AI Automation on premise[9]

Today, AI models are widely used in many applications. Major industries have become interested in this technology and aim to use it in their development and production environments in the immediate future. However, the know-how required to build very accurate, compact and not too computationally expensive models based on ML or DL approaches is very demanding and users lack the expertise necessary to do that on their own.

For example, in the literature the phase of construction of neural network architectures is never accounted in the actual cost and performance of the entire process, and it is assumed as a given grail. However, as a matter of fact, this is the most expensive task. For instance, despite the many years of research and the huge amount of literature published so far, no neural network architecture today is able to predict correctly more than 90% of CIFAR-100[10] images; in other words, the design time of a perfect architecture for CIFAR-100 is currently infinity. Moreover, most of the experience acquired by the data scientists while developing these models is not shared and becomes lost over time. This includes all the experiments (models) built and discarded during the optimisation phase. "Automation of ML" is a platform that will open the possibility to apply state-of-the-art ML/DL metho-

9. Florian Scheidegger, Luca Benini, Costas Bekas, A. Cristiano I. Malossi , NeurIPS, 2019 and Roxana Istrate, Florian Scheidegger, Giovanni Mariani, Dimitrios S. Nikolopoulos, Costas Bekas, A. Cristiano I. Malossi, *AAAI Conference on Artificial Intelligence*, 2019
10. The CIFAR-10 and CIFAR-100 datasets (Canadian Institute for Advanced Research) are a collection of images that are com-monly used to train and algorithms. It is one of the most widely used datasets for machine learning research.

On-prem Cloud + HPC architecture solution

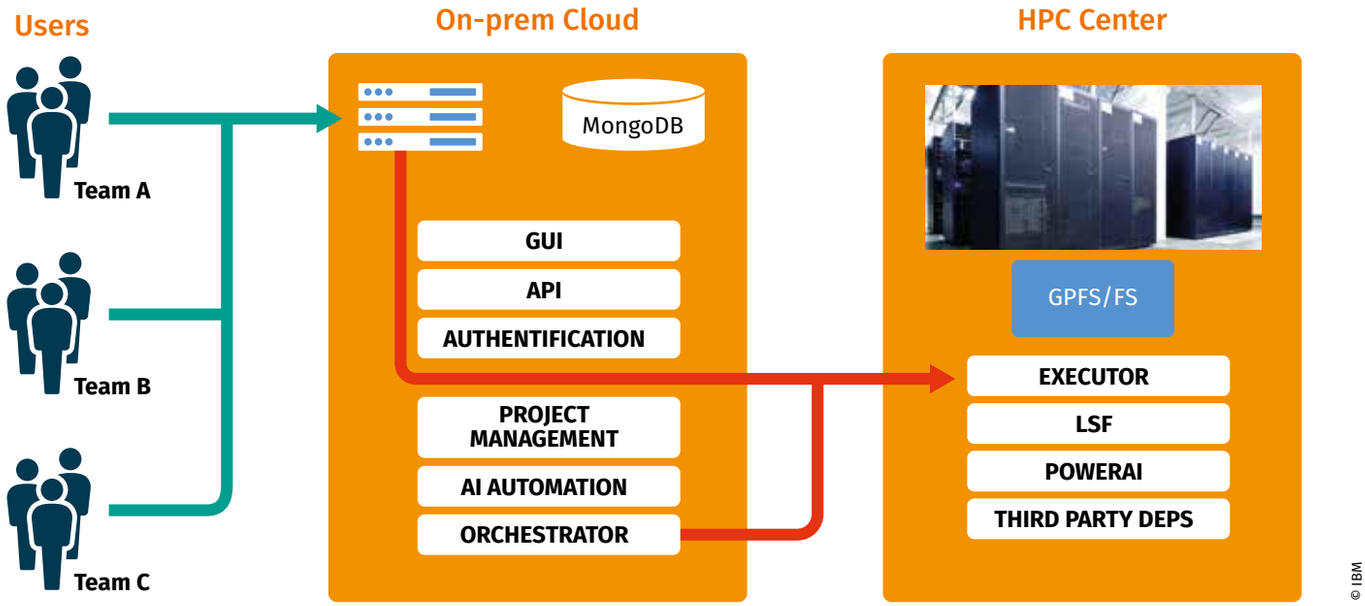


Figure 8: Example of an infrastructure for AI Automation

dologies, without having any expertise concerning data-science, optimisation, data preparation, hyper-parameter selection and training analysis.

The project is rooted around the idea that every time a user generates manually or semi-automatically an AI model, the system can learn from this work and improves over time. In this setting, HPC centres could become large incubators where the Automation of AI framework learns automatically from all the users developing AI models and gradually is enabled to automatically generate new models for a similar future workload. This is achieved by leveraging meta-learning information acquired during the execution of the user workloads.

The platform through which the users submit jobs must be intuitive and as simple as any tool that requires at most a few clicks of a mouse. All user-based decisions will be described by high level constraints, such as accuracy, reliability, size of model, time for inference and latency. Based on the data, and the user provided constraints the automated framework will synthesise the best model and serve it to the user.

The infrastructure used (see Figure 8) consists of:

- a front-end/Cloud-similar system, to manage UI, user access, projects, and AI Automation logic. This does not require GPUs; however, it must be scalable to be used by many users at same time. It can also be containerised and should be supported by classical DBs.
- a back-nd HPC system, with many nodes equipped with GPUs or FPGAs. This system is where all the compute intensive workloads occurs (e.g. model synthesis and training).

The HPC infrastructure, typically made of many high-performance nodes and one or more submission front-end servers (with ssh access) needs to be tightly coupled with the on-premise Cloud- similar nodes, where permanent services can be run for authentication, UI, DBs, backup and orchestrators instances. Those services cannot run on the classical HPC server front-end, because they are too heavy and would compromise performance of the front-end where users expect to do ssh access for the non-AI workloads.

The coupling between Cloud-based nodes & classical HPC nodes should be made in a way that:

- There is a fast connection (e.g. Infiniband type) between the Cloud-HPC nodes
- DBs instantiated on the on-premise Cloud nodes should be visible and fast accessible from all nodes.
- The shared file system on the HPC nodes should be visible from the on-Cloud nodes.
- The on-premise Cloud nodes will run services containerised.

AQMP Project Overview

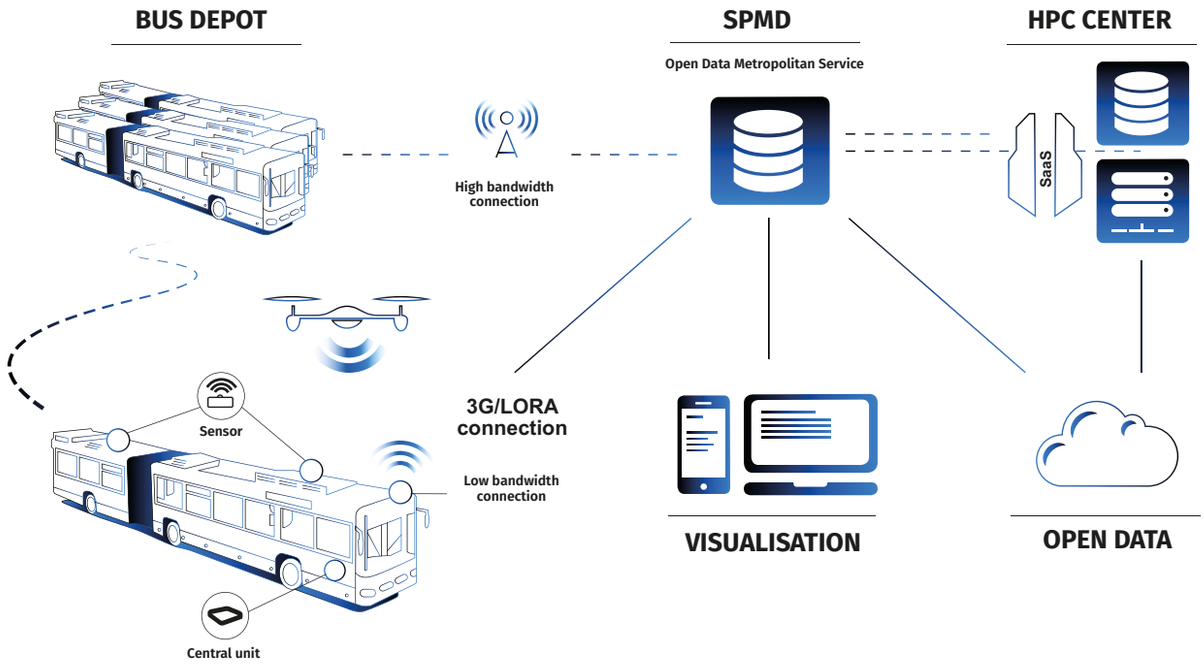


Figure 9: AQMO continuum overview

2.3.4

USE CASE 4: AQMO[11,12]: An Edge to HPC Digital Continuum for Air Quality

Air quality improvement is a major challenge for most metropolises. Proposing efficient policies to address this challenge requires solving two issues: 1/ performing air quality measurement with a thorough temporal and spatial coverage and 2/ understanding the dispersion of the pollution as well as being able to analyse "what-if" scenarios.

The measurement issue is addressed using multiple sensors while the second one is related to the use of HPC numerical simulations. Of course, the two issues are intimately entangled. The measurements provide the basis for elaborating the model inputs and validation while the numerical dispersion model (currently the SIRANE model[13]) gives an insight into how the pollution reaches the citizens.

The AQMO project provides an end-to-end urban platform that extends current practices in air quality measurements. Figure 9 shows an overview of the Digital Continuum designed to implement the platform. This continuum integrates Edge technology, Cloud facilities and supercomputers. It is intended to provide citizens, local authorities, scientific organisations and private companies with new Open-Data and innovative services based on computing simulation (High-Performance Computing - HPC - and Edge Computing / IoT). These new services are a pilot implementation of a new business model called "HPC as a service" which will be analysed as a new way to access the future European Exascale HPC facilities.

To implement an air quality analysis in a cost-effective manner in a wide area, a local transportation bus network is embedded with mobile sensors. In the case of measurements for catastrophic events, the use of drones is explored (in connection with the UAV-Retina project supported by the EIT-Digital). This strategy allows us to use fewer but more accurate sensors. The Edge computing part of the continuum performs two main functions: 1/ the storage of the collected data in order to manage the intermittent communications issue and 2/ respecting citizen privacy. The platform makes use of cameras to detect if a bus is stuck behind a truck or another vehicle (which tamper with the pollution measurement). An IA engine is implemented at the Edge and only the image analysis is sent back, and no images are stored.

11. The AQMO project is co-financed by the European Union through its Connection Europe Facility (CEF) program 2017-FR-IA-0176, http://aqmo.irisa.fr.
12. The partners of this project are: AmpliSIM, the University of Rennes, GENCI, the Rennes metropolis, AIR BREIZH (organisation in charge of air quality monitoring in region Brittany), KEOLIS (a bus operator), CNRS-IDRIS (supercomputing centre), NEOVIA, UCit and RYAX Technologies.
13. Soulhac, L., Salizzoni, P., Cierco, F. X., & Perkins, R. (2011). "The model SIRANE for atmospheric urban pollutant dispersion; part I, presentation of the model", Atmospheric environment, 45(39), 7379-7395.

The resource continuum is logically assembled using a global workflow management technology and is designed to support new sensors as well as complex numerical models for routine use in smart-city solutions.

2.3.5

USE CASE 5: FTRT – Faster Than Real Time for seismic, volcanic or tsunami events.

Tsunami Service Providers (TSP) (which provides tsunami warnings in the framework of the systems coordinated by IOC/UNESCO worldwide) and other national tsunami warning centres strive to complement, or replace, decision matrices and pre-calculated tsunami scenario databases with FTRT (Faster Than Real Time) tsunami simulations. The aim is to increase the accuracy of tsunami forecasts by assimilating the largest possible amount of data in quasi real time and performing simulations in a few minutes of wall-clock time, possibly including the coastal inundation stage. This strategy of direct real time computation, which would have seemed unfeasible a decade ago, is now feasible due to the astonishing recent increase in the computational power and bandwidth evolution of modern GPUs. In the field of such urgent computing access modes, combining end to end workflows of data coming from network of sensors to HPC resources, the ChEESE CoE[14] (bridging together researchers from NGI in Norway, INGV in Italy and UMA in Spain) aims to demonstrate using the example of Urgent Seismic Simulations, Faster Than Real-Time Tsunami Simulations and High-Resolution Volcanic Ash Dispersal Forecast that technology is available to design Exascale Urgent Computing workflows that support contingency plans for seismic, volcanic and tsunami events.

Moreover, natural catastrophes often occur with very little anticipation and precursors that are difficult to detect. The earlier the warning, the more effective the mitigation of the impact of the dangerous phenomena.

This is **Early Warning** and to implement it, ChEESE will make use of the most efficient High Performance Computing architectures, software and workflows to demonstrate prototype Early Warning systems for tsunamis.

- How big will the next volcanic eruption be at Jan Mayen in Norway? How will that potentially impact the population and the air traffic in Europe, considering the statistical variability of wind intensity and directions?

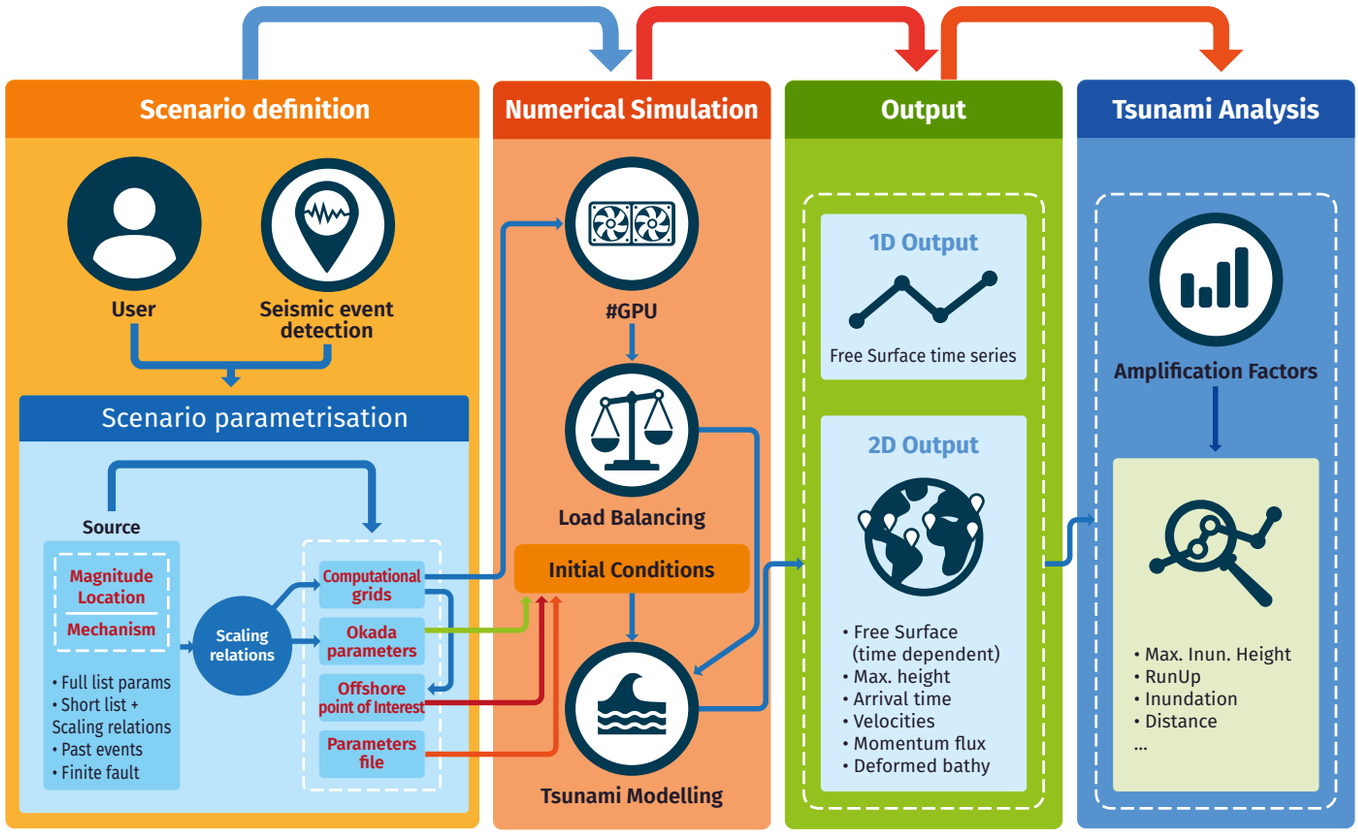


Figure 10: Example of a complete end-to-end workflow of real time tsunami alert system

14. https://cheese-coe.eu/

- Can scientific research, together with advanced computing technology, help reduce volcanic risk for the population and for productive activities in Southern Italy and optimise land use, taking into account the potential impact of future volcanic eruptions at Campi Flegrei and Vesuvius?
- What is the probability of ground acceleration exceeding a given threshold for a nuclear plant site or for another critical infrastructure in Europe? What is the probability that a large earthquake in the Mediterranean produces a tsunami wave higher than one metre in the Marseilles harbour?

Building a prepared society resilient to natural catastrophes requires the capability of managing the complexity of the natural phenomena and the large uncertainty associated with their development in a probabilistic framework. This requires performing large *ensembles* of accurate scenario simulations to reproduce the complex physics of the natural systems and the wide variability of initial and boundary conditions.

This is **Probabilistic Hazard Assessment,** which ChEESE will demonstrate using Pilot Demonstrators on Physics-Based Probabilistic Seismic Hazard Assessment, the capability of Exascale Computing to perform large ensemble, physics-based simulations in order to quantitatively assess natural hazards and their related uncertainties. ■

3

The European HPC Ecosystem and other supporting ecosystems

Many of the topics developed in this SRA are also tackled by various other associations, organisations, projects or initiatives in Europe. ETP4HPC has established collaborations with these organisations and works jointly with them on the definition of the corresponding research roadmaps (including this Research Agenda). What follows is a summary of the roadmaps, vision documents and uses cases issued by the other parts of the European Digital Continuum which complement this SRA effort. We focus on how the work of these organisations has impacted the findings of the SRA.

According to **HiPEAC's (High Performance and Embedded Architecture and Compilation in Europe)** vision roadmap[15], ICT solutions should not be considered as silos but rather "a continuum ranging from deep-Edge (microcontrollers linked to sensors or actuators), to Edge, concentrators, micro-servers, servers and Cloud or HPC."[16]). Besides the purely technical challenges of this integrated vision, organisational issues are also expected to arise: due to the size and the heterogeneity of the systems and their providers, interoperability will be key and de facto standards (or standards) will be necessary to allow this integration to happen. This paradigm has given rise to a number of new characteristics of this SRA – e.g. the establishment of the **Research Clusters** concept, the inclusion of the "Centre-to-Edge Framework" research domain and the "Trustworthy Computing" cluster (see *5.1 The concept of Research Clusters and Research Domains* on page 35).

The **BDVA** (the European **Big Data Value Association)** is an industry-driven organisation of around 200 members, with a well-balanced composition of large, small, and medium-sized industries as well as research and user organisations. It has the objective of developing the European Big Data Value Chain. BDVA is the private counterpart of the EC in implementing the Big Data Value PPP program. BDVA and the Big Data Value PPP pursue the common shared vision of positioning Europe as the world leader in the creation of Big Data Value. BDVA is also the other private member of the EuroHPC Joint Undertaking, besides ETP4HPC.

The integration of Big Data priorities into the Digital Continuum is driven by the requirements of "data driven applications". Jointly with the Big Data Value Association, ETP4HPC had compiled a repository of uses cases use cases which represent this new paradigm: Big Data Computing (BDC) workloads increase in computational intensity (traditionally an HPC trait) and some High-Performance Computing (HPC) workloads increase data intensity (traditionally a BDC trait) – which was followed by an analysis of these uses cases. The management of this discrepancy in one area requires adopting certain techniques from the other area. What the convergence of Big Data and HPC could look like at the software stack level is detailed in the ETP4HPC-BDVA joint white paper titled "The technology stacks of High-Performance Computing and Big Data"[17].

The pattern of this integration is also noticeable in the work of the **Centres of Excellence for Computing Applications (CoEs), including the FocusCoE coordination action**[18] (the task of which is too facilitate cooperation among the CoEs). The CoEs consolidate the European application expertise (the main actors and the main codes of various application domains present in Europe) into projects. Each of the CoEs helps strengthen Europe's leadership in HPC applications by developing, optimising (including re-design) and scaling HPC application codes towards Exascale computing. They offer services and consultancy to industry and SMEs, conduct research in HPC applications and address the skills gap in computational sciences. As of the time of writing this SRA, ten application domains are covered, e.g. weather and climate, material science, molecular biology, medicine, and engineering. As an example of the work of the CoEs, in the domain of weather and climate developed by the Esiwace[19] CoE, the nowcasting of precipitation amount can be improved by taking into consideration not only data from local weather stations but also mobile sensors in windscreen wipers[20]. The Cheese CoE (in the domain of solid earth – see USE CASE 5 in the previous chapter) requires that the integration of HPC and HPDA be understood by the European institutions in charge of operational geophysical monitoring networks. The contribution of the experts associated with the CoE to this SRA is in the area of applications and related uses cases.

15. https://www.hipeac.net/vision/2019/
16. https://www.hipeac.net/vision/2019/, p.10
17. https://www.etp4hpc.eu/pujades/files/bigdata_and_hpc_FINAL_20Nov18.pdf
18. https://ec.europa.eu/digital-single-market/en/news/ten-new-centres-excellence-hpc-applications
19. https://www.esiwace.eu/
20. http://www.lebensraumwasser.com/big-data-von-scheibenwischern-sollen-vor-lokalem-starkregen-warnen/ (in German)

BDEC (Big Data and Extreme-Scale Computing)[21] fulfils the important task of providing a forum to exchange visions, ideas and new concepts at the international level, bringing together US, European, Chinese and Japanese world-class experts ("gurus") in a series of joint workshops. BDEC focuses on the convergence of HPC, Big Data, AI and other technologies. In the first round of BDEC, its work focused on developing 'Pathways to Convergence' - a series of white papers documenting the discussions taking place[22] - and also a repository of use cases in order to illustrate the paradigm of convergence. In its second phase, which is running now, BDEC is working on broadening the convergence paradigm in order to cover the Digital Continuum, the concept of which, together with an analysis of related elements, is BDEC's main contribution to this SRA.

The **AIOTI (the Alliance for Internet of Things Innovation)** aims to create a dynamic European ecosystem of the Internet of Things players and thus speed up the take up of IoT. Our collaboration with AIOTI is more recent (it began in 2018) and it has taken the form of joint workshops – our objective is to develop it further to reach a level of maturity on par with other organisations. The expertise of AIOTI in Edge technology contributed to the use case repository included in this SRA (*2.3.2 USE CASE 2: Autonomous driving* on page 22).

The mission of **PRACE (Partnership for Advanced Computing in Europe)** is to enable high-impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness for the benefit of society. PRACE seeks to realise this mission by offering world class computing and data management resources and services through a shared HPC infrastructure. PRACE maintains its PRACE Scientific Case[23], which defines the requirements of scientific and industrial research in relation to the European HPC infrastructure, applications and technology. Along with the Centres of Excellence (above), these use cases and the resulting user requirements have been valuable input to this SRA for understanding the needs and expectations from research communities.

21. https://www.Exascale.org/bdec/
22. http://www.Exascale.org/bdec/sites/www.Exascale.org.bdec/files/whitepapers/bdec2017pathways14Nov17.pdf
23. http://www.prace-ri.eu/third-scientific-case/

Eurolab-4-HPC is a 2020 funded support and coordination action, which has the overall goal to strengthen academic research excellence and innovation in HPC in Europe. The project has formed a network of hundreds of domain experts in all technologies related to future HPC systems, from applications, software and hardware, giving rise to a long-term roadmap which identifies technological opportunities and hurdles with a 10 to 15-year perspective[24]. The "Eurolab-4-HPC Long Term Vision on High Performance Computing" was published in August 2017. This vision presented an assessment of potential changes for HPC in the period 2023 to 2030[25]. It foresees radical changes in computing over the decade to 2030, in particular with the potential advent of new, disruptive hardware such as for example Quantum Computing; because of the long-term perspective and unavoidably speculative nature, the authors started with an assessment of future computing technologies that would – if mature - strongly influence HPC hardware and software. As such, Eurolab-4-HPC's roadmap is complementary to this SRA: the SRA focuses on development needs at a time frame of 5 to 7 years, whereas Eurolab-4-HPC looks at a 10 to 15 years perspective. Nevertheless, some of the topics mentioned in Eurolab-4-HPC's long term vision are already tackled and foreseen in today's SRA. In particular with respect to the emerging applications, such as Industry 4.0, smart cities and autonomous cars, which require real-time and interactive analysis of data (based on machine learning). Since the publication of the 2017 vision, there has been significant evolution in all the above topics and a growing momentum in the area of open hardware. These topics will be revisited and expanded in the updated Eurolab-4-HPC vision to be published in January 2020 and will be fed into the material for our following SRA.

Over the past years, numerous policy makers world-wide have articulated a clear and consistent vision of global Open Science as a key factor, which could enable the new paradigm of transparent, data-driven science and accelerate innovation. In Europe, this vision is being realised through an ambitious programme within the **EOSC (European Open Science Cloud).** The goal of the EOSC is to offer 1.7 million European researchers and 70 million professionals in the areas of science, technology, the humanities and social sciences a virtual environment with open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines. This will be achieved by federating existing scientific data infrastructures, which is currently dispersed across disciplines and the EU Member States. In this SRA, EOSC is a source of information which helps anticipate upcoming requirements in the area of academic applications, in particular with respect to data management (data logistics, storage, in situ data analysis, interoperability) and the ease-of-use of HPC infrastructure services (such as authentication and authorisation mechanisms, portals for accessing computing/data resources, transfer of data and remote visualisation).

The **ECSO (European Cyber Security Organisation)** is the contractual partner of the European Commission in the implementation of the Cyber Security contractual Public-Private Partnership (cPPP). ECSO members include a wide variety of stakeholders such as large companies, SMEs and start-ups, research centres, universities, end-users, operators, clusters and association as well as European Member States' local, regional and national administrations. In the past, HPC systems were mainly installed on secured sites, with restricted physical and remote access to the outer world. The set of users was known (and authenticated) in advance, with predefined access rights. Nowadays, with HPC becoming part of "the loop", HPC systems will be increasingly exposed to users, who may not have been identified ex-ante. HPC systems will also be exposed to attacks and malicious attempts trying to break down the system or to misuse its resources or data. In this context, cyber security gains utmost importance for HPC in order to provide seamless, but at the same time secure and trustworthy HPC computing resources to the user, which is jointly discussed and tackled by the ECSO and ETP4HPC in this SRA.

24. http://eurolab4hpc.eu
25. https://www.eurolab4hpc.eu/vision/

4

International arena: HPC and HPDA in Europe, China, the US and Japan[26]

The recent period has seen remarkable international efforts to upgrade supercomputing capabilities aimed at attaining Exascale performance. At the same time, a data-saturated and AI-enabled world is emerging at an extremely rapid rate. This emergence is beginning to have a strong impact on the new generation of cyberinfrastructure that needs to combine floating point, machine learning and data IO capacities at (exa)scale and beyond.

The major regions (US, Europe and Asia) have already mapped out their paths to Exascale performance – see below. However, the installation of very large scientific instruments, notably SKA and LHC2, will change the game and also broaden the geographical basis of the large cyberinfrastructure. Furthermore, the reliance of many machine learning algorithms on lower precision arithmetic is making exa-ops (i.e. Exascale operations), and possibly, in the future, zetta-ops performance, available at a far lower cost thanks to dedicated, specialised processor architectures.

After some initial hesitation, the US has now a firm and aggressive calendar for the implementation of Exascale capabilities. For the period of 2020-2025: ▶

- two pre-Exascale machines will be installed, one in 2020 and one in 2021;
- two Exascale machines are planned in 2022;
- one or two Exascale machines are planned in 2023;
- two Exascale machines are planned in 2024;
- one Exascale machine is planned in 2025.

This gives a total of 6 or 7 Exascale machines over the period from 2020 to 2025. Each Exascale system is budgeted at $500M-$600M, plus R&D investments (ECP, AI initiative, etc.). Total investment in R&D will be $1B-$2B per year by the government and the vendors involved. The Processors and the vendors will be American. The first three DOE machines will all be Cray Shasta systems with fat CPU-GPU nodes.

China's program is similar to the US. For the period 2020-2025:

- two pre-Exascale machines will be installed, one in 2020 and one in 2021;
- one or two Exascale machines are planned in 2022;
- one Exascale machine is planned in 2023;
- one Exascale machines is planned in 2024;
- two Exascale machines are planned in 2025.

This gives a total of five or six Exascale machines over the period. Each Exascale system is budgeted at $350M-$500M, plus R&D investments. Total investment in R&D will be over $1B per year, by the government and the vendors. Due to the recent US embargo policies, the processors and the vendors will be Chinese. One machine each will be prepared by Sunway (Shenwei), Sugon (x86 with AMD) and NUDT (Chinese ARM processor).

The backbone of **Japan**'s program is the "Post K" national integrated project. For the period 2020-2025:

- one Exascale machine will be installed in 2020 (Fugaku, a non-heterogenous machine, the first outcome of Post K project);
- Then, in the 2021-2025 timeframe, other Exascale systems are expected, significantly supporting next generation AI data intensive applications.

The Post K flagship project has been budgeted at $1B, including R&D investments. Many smaller systems, $100M-$150M are planned. The processors (Fujitsu with ARM-based A64FX) and vendors will be Japanese.

In the **EU**[27], the EuroHPC deployment program has recently been announced by the EuroHPC Joint Undertaking. For the period 2020-2025:

- three pre-Exascale machines will be installed in Finland, Italy and Spain in 2020-2021;
- two Exascale machines are planned for 2023-2025;

Each Exascale system is expected to be budgeted at over $300M, plus R&D investments. The investment in infrastructure and R&D is expected to be around $1B per year, shared among EU, the Participating Member States and other stakeholders.

In addition to the above, **the emerging regions** are increasing present in this arena. India is leading with a $1B investment programme for the purchase of 70 machines over a period of seven years. Russia has exhibited petaflop machines (the Lomonosov series) during the last five years. On the African continent, South Africa has invested heavily in a national HPC infrastructure. In South America, Brazil is the only country appearing in the Top500 list.

4.1

International arena: looking ahead

Deep Learning is where today's and especially tomorrow's supercomputers can be applied. It has different requirements in terms of how it uses data, how it produces outputs, and how it scales to use many processors from the more traditional HPC applications based on large-scale simulations. An edifying example is the DOE's project that defines the design of future Exascale systems[28] to ensure they are tailored for Deep Learning. As this project exemplifies, the combination of Exascale systems and Deep Learning will let scientists solve problems for leading Edge cancer research in entirely new ways. Also, in new material design, we will increasingly find that simulation, AI and robotics will be tightly linked together.

As far as hardware is concerned, machines built around fat nodes with GPUs, large memory, reasonable network, connected to a large amount of non-volatile memory are suitable for machine learning algorithms, though these algorithms require much lower precision (32-, 16-bit or even lower), compared to 64-bit for traditional computational science. However, if we look ahead to 100 exaflops or zettaflop scales, it may not be the case that the best architecture for AI problems is also a reasonable architecture for simulation and vice versa. AI algorithms need different kinds of sparsity from numerical simulations and their demands in terms of storage, I/O and memory bandwidth are different. The US Exascale machines that are planned in the next few years are going to be reasonable platforms for doing both simulation and AI in the *short term*.

In the longer term, the scale of investment in future, traditional HPC systems would seem to be daunting. Furthermore, future AI architectures (for scientific research, at least) should probably be driven by standard benchmarks that still need to be developed. All the above arguments plead in favour of a reappraisal of future, post-Exascale cyberinfrastructure investment and development strategies. ■

26. This section is based on the following references: Hyperion Research HPC Market Update, ISC 2019; Rick Stevens, HPC Wire interview, July 2019'; Asch, et al. Pathways to Convergence. IJHPCA, 2018.
27. EuroHPC involves some non-EU countries which are associated with the EU.
28. https://candle.cels.anl.gov/

5

Technical Research Priorities 2021 – 2024

This chapter represents the core of the SRA. The model we have used in this SRA to define the research priorities is based on Figure 2, which defines the high-level EC framework for determining research directions. Our model covers the "Applications development" and "Technology infrastructure" layers of that high-level framework.

In our model, the research areas identified are grouped according to two overlapping layer dimensions: "Research Domains" and "Research Clusters", as presented in Figure 11 below. The priorities to be addressed by the future Work Programme are also to be found on the intersection of the Research Clusters and Research Domains – each of the Research Domain chapters highlights those common elements. This process is described in detail below.

5.1

The concept of Research Clusters and Research Domains

The concepts of "Research Domains" and "Research Clusters" is an evolution of the four-dimensional research model used in previous SRAs, see Figure 11:

- **Research Clusters** is the new concept developed in this SRA. They represent cross-cutting "themes", which **capture the research priorities** for the next generation of HPC infrastructure. They are shown as vertical elements in Figure 11 as the impact of most of them cuts through the majority or all of the research domains. Under the heading of each "Cluster", several aspects with a high degree of similarity are grouped (or 'clustered') together. Some of them are traditional themes such as "Energy efficiency" or "Resilience" (already present in the previous SRAs) and some others are new and originate from outside the traditional HPC e.g. "Data everywhere" or "AI everywhere".

  The clusters are described below, listing the following five characteristics (not all apply to all clusters):

  1. "**Intro**: what topics are covered by this cluster": describes the technical field associated with the name of the cluster.
  2. "**Maturity** (time to market)": illustrates the distance of the components of the cluster to market exploitation and commercialisation.
  3. "**Relevance** and impact (why chosen)": explains why we think the technical aspects are worth being bundled under one name and topic.
  4. "**Hurdles** to overcome" mention the challenges (technical and other) in making significant progress in the field addressed by the cluster.
  5. "**Driving competence**": identifies the most active stakeholders in Europe with contributions to the discussed subject.
  6. "**Cost** of research to gain significant uptake": attempts to quantify the investments required to bring the addressed technology to a state where commercialisation is possible within a few years.

- **Research Domains** are needed to describe the essential layers and elements of a HPC functional stack. While "System Architecture" applies a holistic view of all elements of the stack, "System Hardware Components" covers the lowest hardware level. "System Software and Management" focusses on all aspects of operating and managing the underlying hardware facilities. "Programming Environment" adds a user's (programmers) view to using the HPC hardware and system software infrastructure. "I/O and Storage" covers the

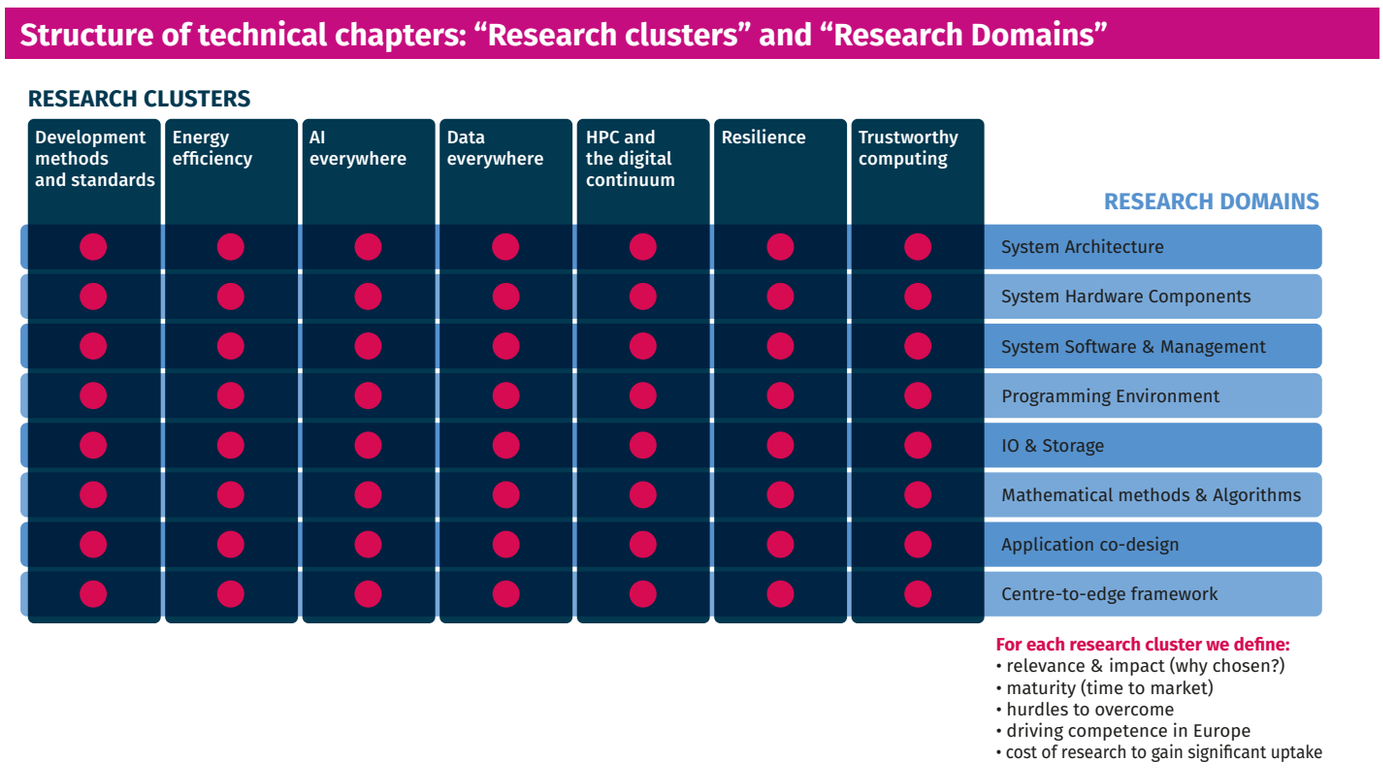


Figure 11: The sources of research priorities: the concept of Research Clusters and Research Domains

space of feeding the compute nodes with data and storing data. "Mathematical Methods and Algorithms" look at the backbone of any level of software used in HPC systems (not to be confused with the algorithms used in the application layer). "Application Co-design" offers an application writer's input into the needs for the implementation of next generation HPC infrastructure all together. Finally, given the new trends of associating HPC with the other parts of the Digital Continuum, the "Centre-to-Edge Framework" Domain covers all aspects of HPC functionalities used in complex multi-tiered workflows.

Each domain description is divided into three parts:

1. "**Research trends** and current state of the art": introduces the domain by laying out the status quo and today's research areas and priorities.
2. "**Challenges for 2021 – 2024**": outline the most important problems and limitations that need R&I-attention in that period.
3. "**Intersections with Research Clusters**" describe the specific technical characteristics common to both the research domain and a given research cluster. It illustrates the areas where the domain needs to feed solutions to the problems and challenges mentioned in the cluster. These intersections implicitly show where the research domains have a "common ground", i.e. where they overlap with the same research cluster. It is these intersections that could constitute the core of the calls for proposals in the next Work Programme(s), beside the Domains and the Clusters.

5.2

Research Clusters

5.2.1

Development methods and standards[29]

5.2.1.1

Intro: What topics are covered by this cluster?

In the recent decades, application performance has been increasing significantly due to successive generations of hardware improvement. Immeasurable advances in science and industry have been obtained primarily by exploiting corresponding increases in computational performance or throughput. The landscape has profoundly evolved due to data related technologies (e.g. ML) and the end of Moore's Law. On the one hand, many new applications are based on complex workflows to be deployed over a continuum of devices/systems spanning from the Edge to HPC/Cloud infrastructures. On the other end, the heterogeneity of the hardware and software is increasing due to the increasing availability of accelerator technologies (GPU, FPGA, etc.) and the integration of HPDA software stacks. Development complexity has reached a new level.

The long-term consequence of these effects is that enormous and increasing amounts of effort are spent on developing codes and software which is wasted in future generations. With the increased complexity and cost of systems, we seek a more sustainable set of development practices that are i) focused on longer-term insulation against hardware changes ii) help combat complexity and heterogeneity in systems iii) consider a broader set of technical objectives in their design iv) enable inter-disciplinary and cross-cutting approaches v) seek holistic solutions to broadly stated requirements.

This research cluster considers approaches to development that are forward-looking, broad, holistic and sustainable:

- **Future-proofing and application portability.** Enabling applications to perform across varying and heterogeneous architectures is essential for ensuring sustainable performance on emerging Exascale computing systems, and preventing development investments from binding applications into a specific solution path. Frameworks that consist of domain-specific languages, libraries, programming and abstraction frameworks (spanning from single task codes to complex workflow-based applications), models, toolchains, virtualisation (a.k.a. containers) have proven to provide a good practical approach.
- **Performance analysis and workflow monitoring and profiling.** Understanding the performance of applications as a whole becomes more challenging due to the heterogeneity of systems and the diversity of languages, DSLs and frameworks used in those applications. Python, DSLs and AI frameworks all abstract the machine performance from the program but by doing so make profiling challenging. Advancements in tools are required to understand performance as well as data logistic bottlenecks.
- **Development best practices.** Relaxed approaches to programming conventions and the large variety of application domains have meant that HPC does not generate a set of common best practises. Such a body of knowledge is increasingly necessary in order to offset the growth in complexity and interoperability of frameworks/languages.

5.2.1.2

Relevance & impact (why chosen?)

As infrastructures become more complex, with increased parallelism, increased heterogeneity, novel architectural accelerators (e.g. dataflow, FPGA) and deep software stacks, solutions based on the traditional practices will cease to be effective and (therefore) commercially viable. Considering future-looking, broad, multi-disciplinary and sustainable solutions will therefore become the new normal. The regions that are able to invest in sustainable approaches most convincingly will reap the largest longer-term scientific and productivity gains as a result.

5.2.1.3

Hurdles to overcome

Multidisciplinary approaches spanning the hardware-software stack, from vendors to application developers, toward portable

29. Rodrigues, Arun F., et al. «The structural simulation toolkit.» *SIGMETRICS Performance Evaluation Review* 38.4 (2011): 37-42.

solutions are needed. This requires both research, political and community organisation strong and focused involvement in a clear and shared goal.

5.2.1.4

Driving competence in Europe

The Centres of Excellence represent an important community effort targeting specific application domains, and could mobilise co-design activities well. They should lead the activities toward best practices specification and their promotion.

Standards are by definition a global activity and require pan-national collaboration across multiple domains and industries. Increased involvement of the EU community in the specification of standards is required to put EU in the leading seat.

A number of strong performance analysis toolsets are developed in Europe and this could be considered a core competence. Some of the leading future-proofing and application portability activities are European in origin.

5.2.2

Energy efficiency

5.2.2.1

Intro: What topics are covered by this cluster?

The main objective of the energy efficiency cluster for 2021 – 2024 can be formulated as "building a usable Exascale system within an affordable energy consumption and power envelope". Primarily, power- consumption has to be optimised for the most power-demanding components, but at the same time building a power efficient architecture remains important: even if single components have a modest demand for energy, the overall system's energy efficiency might be constrained if these components become a performance bottleneck. As such, this cluster encompasses a wide range of aspects related to energy/power/performance efficiency, including optimising the system-Power Usage Effectiveness (PUE), the technology (transistor) energy efficiency, the compute architecture performance-per-watt (including heterogeneous acceleration), the memory architecture, infrastructure and cooling. Moreover, it includes software techniques required to enable effective and efficient use of processors, accelerators, memory and storage hierarchy and network in the programming environment and resource management for energy, power and performance.

5.2.2.1.1

Optimising PUE

The power consumption of high-end HPC systems alone is expected to continue to be in the 5MW to 50 MW range in the foreseeable future. To make the situation worse, the total energy consumption and power envelope has to include the additional energy needed to provide the electricity distribution in the data centre, the Uninterrupted Power Supplies (UPS), air conditioning and cooling systems. The PUE ratio has been introduced to monitor the total AC power input of the entire data centre compared to the system power consumption alone. Ten years ago, it was possible to achieve a PUE as bad as 1.7. HPC systems have been an important driver in improving the PUE-factor since data centres and systems are enhanced and deployed at the same time. As a result, an optimised HPC data centre of today shows a PUE close to 1 mainly by using free or very efficient cooling techniques, limiting the usage of UPS or implementing a High DC voltage distribution.

The next objective would be targeting a PUE-factor below 1. Doing so is possible if the energy dissipated by the (Exascale)- system is reused for another purpose (waste heat recovery), e.g. as a heat source for central heating facilities. Alternatively, new research directions try to reuse dissipated heat for providing free air conditioning, or, even more challengingly, producing electricity. To increase the Carnot efficiency[30], the temperature difference between the cool and hot spots (heat generated by the computing cores and associated memories and communication links) should also increase, but increasing the temperature of the components has a drastic drawback on reliability, given the currently used technologies.

At the system level, power supplies, the power distribution and voltage regulator modules are specifically tuned for each larger system component, such as CPUs, GPUs and other accelerators. Thereby, 10% to 20% of the energy supplied to the compute engine is lost in the power conversion stages required to yield the multitude of processor and accelerator voltage levels. Research in new power components, new power system architectures and topologies is currently underway to improve this situation, e.g. initiatives such as US Energy Star[31]. New designs tend to place the voltage converters as near as possible to the cores (e.g. using interposers and heterogeneous technologies for chiplets), allowing to increase the voltage delivered at the package level.

5.2.2.1.1

Usable FLOPs per Watt

Hardware

In electronic components, the energy efficiency at the transistor level is key: reducing transistor geometry is an important driver for improving power efficiency. For example, gate power consumption is reduced by about 25% when moving from 10 nm to 7 nm operating at the same frequency. It means that the same component ported onto the next silicon technology level will result in less energy consumption.

Today, the HPC compute component -whether CPU socket or accelerator - is one of the most important contributors to power consumption (80% of a 2-socket baseboard). Its architecture is also influencing peak power consumption: today's frequency range has levelled at 1GHz to 4 GHz, but the number of cores or SIMD engines and the use of accelerators continues to grow. Since power consumption does not scale linearly with regards to frequency, different design points are envisioned. As a result, Flops per watt could vary from 5 GFlops/W to 100 GFlops/W among the different

30. The Carnot efficiency is the theoretical maximum efficiency one can get when the heat engine is operating between two temperatures.
31. https://www.energystar.gov/

solutions available today and in the near future. The major factor of energy consumption in a computing core are not really the ALUs, but the Vector Processing Unit and the energy to move data around the chip and to/from outside the compute node.

Memory architecture, I/O devices and network consume less energy than the components above. Today, they represent less than 20% of the power consumption of a 2-socket baseboard for HPC. Here efficiency is achieved by providing an adequate configuration for the customer requirements in order to exploit maximum compute performance. In-package memory, HBM, NV-DIMMs and processing in memory are promising technologies in this regard.

The storage subsystem is not directly significant in the context of energy/power efficiency as a disk consumes a few watts whereas CPUs, GPUs and accelerators can reach up to 300W today. This statement may be revisited for new use patterns such as High-Performance Data Analytics and Artificial Intelligence if storage subsystems might evolve significantly.

Finally, it is important to define a balanced system architecture that meets the needs of the targeted applications and workloads. For example, if the application is memory bound, a system optimised for peak compute performance would not be power-efficient. This emphasises the importance of comprehensive co-design approaches.

Software

For real applications, the power efficiency (energy- and time-to-solution) will depend on the application and system capability to work together. Today's hybrid architectures tend to provide better power efficiency by allowing the execution of application parts on different types of compute components. Such heterogeneity requires evolution, adaption and standardisation of the programming environment, middleware, resource manager, performance/power/energy models and profiling tools to ensure effective and efficient use in existing and future applications. However, today's tools still do not provide a direct feedback to the programmer on how energy-efficient the code is.

Replaying customers applications on different systems remains difficult and expensive. Therefore, HPC system power efficiency levels continue to be measured and compared using benchmarks. Today, LINPACK still is the most widely used benchmark but it is increasingly unrepresentative. The High-Performance Conjugate Gradients (HPCG) Benchmark suite is gaining popularity but there is a need for other types of benchmarks.

Limiting the energy waste is the first optimisation that is handled by operating systems, including switching off unused resources. Modern components can have different behaviours depending on their operation conditions (temperature, power consumption and silicon variability). For example, they run faster at low temperatures and slower at high temperatures. New software tools optimise energy efficiency through hardware controls, as this performance variability is a major problem for typical tightly coupled HPC applications. Hence, developing software that optimises power consumption using the set of different metrics and controls offered by the hardware is an important research area.

Dynamically detecting worse behaviour or abnormal energy consumption is also key for energy optimisation. New software tools are emerging for measuring, tuning and predicting energy consumption of applications at the HPC system level. Pushing instrumentation deeper, generating a fine grain profile of an application is also a way for optimising power efficiency of applications since it will allow to identify where precisely energy is spent.

Machine Learning is now deployed in order to optimise system power efficiency: ML can be applied to the system parameters that are captured and then it could be used to correlate them with power consumption. Various experiments of this kind on infrastructure controls are already being carried out.

5.2.2.2 Relevance and impact (why chosen?)

The energy consumption, peak power demand and thermal dissipation are important constraints on overall performance. This is true across the whole computing continuum from embedded/IoT to HPC.

Energy consumption is an important contributor to TCO. Related to this, the availability of enough peak power supply is an important constraint on the maximum size of system that can be installed in a given location. The environmental impact of energy consumption is of great interest as well, especially in the context of climate change and Europe's goal to reduce carbon emissions by 40% by 2030.

5.2.2.3 Maturity (time to market)

Energy efficiency has been a challenge in HPC for many years, with a broad recognition driven by the Green500 list, announced in November 2007. Since then, the LINPACK energy efficiency of the number one system has increased from 0.2 GF/W to 16.9 GF/W in November 2019. Nevertheless, it is expected that energy efficiency will continue to improve over time for allowing one Exascale system within a 20 MW range in the 2022-24 time period[32].

5.2.2.4 Hurdles to overcome

There is a need to develop all the elements required to allow portable and productive use of heterogeneous acceleration: system architecture, resource management, standard APIs for integrating heterogeneous accelerators and tools allowing to give relevant advices to the programmers in term of data placement, data transfer and use of the various heterogeneous resources.

It is also necessary to pursue research to deal with performance variability from energy saving mechanisms.

32. Assumption: in 2022, an accelerator will provide around 50TF in a 500W envelop. Based on peak performance and for specif-ic workloads, 1 exaflop could then be fed with 15Mwatt (covering compute nodes and interconnect).

5.2.2.5

Driving competence in Europe

In the hardware area, Europe has been able to innovate in processor architecture (e.g. ARM) and in system design (e.g. Atos) and system integration (e.g. E4, MEGWARE). EPI provides an opportunity for Europe to continue to innovate in processor and accelerator design. Europe has also several leading research groups in programming environments (e.g. Fraunhofer, BSC, Inria).

5.2.2.6

Cost of research to gain significant uptake

Energy efficiency is a vertical challenge that covers almost all aspects of HPC systems and its infrastructure. It can be advanced further by a combination of small research projects, each of them in the 2-4 M€ range, targeting individual challenges together with larger co-design projects as for example the effective support of hybrid architectures. Tight integration of the various hardware and software approaches is crucial to achieve global and sustained energy efficient systems.

5.2.3

AI everywhere

5.2.3.1

Intro: What topics are covered by this cluster?

Artificial Intelligence (AI) is living a second youth and is here to stay. Unlike in the past, the AI models and techniques are now feasible due to (I) the existence of a large amount of data that represents a high-potential source of valuable insights and (II) advances in the underlying hardware and software ecosystem, which provide the computational performance to train the models associated to these AI systems.

The high computational demand of fields such as Deep Learning (DL)[33], have contributed to emphasise the need of high-performance hardware and software (e.g. GPGPUs) and therefore put HPC technologies in the foreground.

Nowadays, many scientific and industrial applications generate and use huge volumes of different kind of data (static data, real-time data, etc) and combine data analytics techniques with simulations. As a paradigmatic example, the autonomous car requires dealing with simulations of hypothetical situations as well as analysing extremely high volumes of data, coming from sensors, databases, etc. This type of applications benefits from architectures with thousands of cores and distributed storage systems, but they also need tailored solutions in order to be able to run Machine Learning (ML) algorithms in order to organise and cluster data and to speed up the queries and the algorithms that use this input data.

The convergence between Big Data, data-driven AI and HPC demands a new software and hardware ecosystem, which, at the same time, should fulfil the requirements of the Exascale era. In relation to this, it is important to cover two different angles: "HPC for AI", i.e. HPC supporting the efficient execution of AI approaches, and "AI for HPC", i.e. AI improving and enabling new HPC solutions.

In the context of this cluster, the following topics will be addressed:

- Use of AI in the context of HPC hardware, such as neuromorphic architectures, which mimic the human brain[34].
- Scalable and high-performance AI solutions. Heterogeneous HPC architectures can contribute to increase the scalability of these solutions, as well as the application of emerging technologies such as quantum or neuromorphic computing. All these alternatives have to be researched in depth.
- Distributed DL Networking Acceleration - DL model training time becomes a critical piece in the overall productivity and adoption of DL in the field. Training times are becoming shorter due to various optimisations in ASICs and software but there is a constant need to improve them. Fast training time can improve data scientists' work dramatically by reducing DL model training time from days to minutes. This fundamentally changes the way data scientists operate.
- Learning across the Digital Continuum by means of the emergence of Distributed AI and Edge Analytics[35] and its integration with HPC-based approaches. In particular, to run inference models at the Edge will contribute to reduce the energy consumption, avoiding some data movement (saving bandwidth) and enabling the use of low energy HPC devices.
- AI software should be easy to deploy, be customisable, run anywhere (across the continuum) and be able to include human in the loop, if needed. In this sense, explainable models will contribute to their acceptance. ML techniques can also be applied to software engineering, enhancing and accelerating the process of creating software.
- Application of AI and HPC not only to scientific scenarios but also to industrial applications, especially driven by the impact of AI in industry today. HPC tools and infrastructure should fulfil the needs of these industrial use cases.

5.2.3.2

Relevance & impact (why chosen?)

AI is one of the significant pillars of the upcoming fourth industrial revolution. AI will proliferate into ALL aspects of the technology stacks and across the whole Digital Continuum: from Edge devices, cell phones and IoT devices all the way to data centres. This presents a great opportunity as a significant inflection point in technology. AI will disrupt not only the IT space but

33. Kenneally, Jim, and Hoppe, Hans-Christian, editors, "The technology stacks of High-Performance Computing and Big Data Computing: What they can learn from each other", 2018. A joint publication of ETP4HPC and BDVA, https://www.etp4hpc.eu/pujades/files/bigdata_and_hpc_FINAL_20Nov18.pdf
34. ETP4HPC, "A blueprint for the new Strategic Research Agenda for High Performance Computing", 2019, https://www.etp4hpc.eu/blueprint.html
35. Bisset, D., Curry, E., García-Robles, A., Hahn, T., Lafrenz, R., Liepert, B. and Zillner, S. (eds). "Strategic Research, Innovation and Deployment Agenda (SRIDA) for an AI PPP: A focal point for collaboration on Artificial Intelligence, Data and Robotics", Brus-sels. 2019, BDVA – euRobotics

most of other industries (practically across most of the domains and verticals: manufacturing, automotive, finance, communication) and will probably create new ones.

AI will also change the way software and hardware developments are done today, by employing AI techniques to create SW algorithms (that is, software writing software) and build new hardware. Finally, AI workloads will also increase the network bandwidth demands for Edge to data centre and inside the data centre.

5.2.3.3

Maturity (time to market)

Although AI is a reality, many aspects of AI are not yet solved. For instance, there are many ethical aspects that have to be solved in the making-decision process. The liability of the AI system has to be clearly defined. On the other hand, at the technical level, it is necessary to find a trade-off between two parameters essential in AI scenarios: time-to-model and the model accuracy. HPC can contribute to improve these technical issues (e.g. by improving the scalability of AI algorithms and systems). The interoperability and compatibility of existing tools in AI and HPC have also to be improved.

5.2.3.4

Hurdles to overcome

- Scalability of AI systems and algorithms: in many current applications, AI has to be applied to a huge amount of data, and often Big Data. Therefore, scalable data management techniques and scalable AI techniques are needed. Furthermore, the underlying infrastructures and frameworks should also exhibit this behaviour. Otherwise, these demanding applications will not be able to take advantage of the advances in AI.
- Performance and energy efficiency of AI methods: there is a huge ground of collaboration between the HPC and the AI community for optimising, scaling out and reducing the memory/energy footprint of AI applications.
- Interoperability of tools and software stack: Although there have been many efforts to make the convergence between HPC and Big Data possible[36], the software stack of both disciplines is completely different, and the interoperability is still an issue. In the case of AI, the scenario is similar. The interoperability of tools and frameworks will make easier the appropriate combination of HPC and AI approaches.
- Ethical aspects: AI systems should follow a human-centric design[37], i.e. oriented at improving human welfare and freedom. This design has to exhibit three main characteristics: lawful (compliant with laws and regulations), ethical (ensuring ethical values) and robust (with good intentions).
- Liability of AI systems: If certain decisions are made by AI engines instead of humans, it is necessary to set up the liability of these decisions. This is difficult to do, particularly when the decision-making process is complex and involves different parties.
- Explainable AI: The omnipresence of AI in our lives, and particularly in critical or sensitive aspects of them, will demand the use of explainable AI techniques, which increase and guarantee the trust in the different applications as well as their acceptance.

5.2.3.5

Driving competence in Europe

AI technology usually will be combination of AI acceleration hardware (dedicated AI accelerators, GPUs, CPU), significant software stack that runs on top of this and the data that feeds the AI process. The SW stack is a place where competence can be gained in Europe, from developing new AI algorithms to tuning, customising and optimising existing ones. Thus, Europe should invest largely and primarily in the development of software solutions. At the hardware level, the competition with other giants from USA, Japan or China is harder. However, Europe is also contributing to this race, mainly through the EPI project and in industries such as car manufacturing. Finally, Europe is strong in industrial data, due to the production happening in Europe.

5.2.4

Data everywhere

5.2.4.1

Intro: What topics are covered by this cluster?

Science and innovation have typically been driven by observation and experimentation over the course of centuries. Advancements in mathematics have given rise to the ability to predict the behaviour of complex systems through mathematical models, for example, Newton's laws of motion paved the way for many innovations during the time of the industrial revolution. Over the past few decades, advancements in computing have led to the use of computational methods which could handle ever more complex systems, leading to the rise of compute-driven simulation. However, this dynamic has been further altered wherein scientific and industrial innovation is now driven increasingly through the ability to capture, store, analyse and learn from vast amounts of data. We have thus entered the era of data-centric computing. Exploding volumes and complexity of data have demanded a sea change in the methodologies used to derive scientific insights and enable industrial innovation which is tremendously accelerated by tapping into the vast amounts of data that can now be captured, stored and analysed. The amount of data that will be generated by various sources will reach 175ZB (ZettaBytes or 10^{21} Bytes) by 2025, growing from 33ZB in 2018, according to IDC[38].

36. BDEC, "Big data and extreme-scale computing: Pathways to Convergence-Toward ashaping strategy for a future softwareand data ecosystem for scientific inquiry", 2018, https://www.Exascale.org/bdec/sites/www.Exascale.org.bdec/files/whitepapers/bdec_pathways.pdf
37. According to: European Commission, "High-Level Expert Group on Artificial Intelligence, Ethics guidelines for Trustworthy AI", 2019
38. IDC, "Digitization of the world from Core to Edge", IDC White Paper, https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Typical workflows now start with data that is generated by instruments, sensors and the Internet of Things. The learnings obtained from this data form drivers for simulations which give us much better predictions and the ability to take actions for the future. The example of the classic use case of driverless cars involves the full data infrastructure chain of Edge, Fog and centre (typically the supercomputing centre). In that, huge volumes of data generated by cars every day are temporarily stored and processed within the Edge locations (which could be the cars themselves). Some of the time-critical data analytics and AI inferences are performed within the Edge, for example for traffic situation analysis, navigation and monitoring, in order to take shorter-term decisions and provide excitation responses for the behaviour of the car. This data at the Edge locations coming from the cars could be fed to another tier of Fog nodes, which are smaller data centres, where further data processing, analytics and learning activities are performed. This could give a full picture of the status of the traffic in a specific geographic region, for instance. This data then moves to a centre or a HPC service provider where longer-term models of the evolution of traffic over time is performed and then outputs are fed back to the Edge nodes through the Fog nodes. To barely scratch the surface of such examples, data-centric workflows can also be obtained from personalised medicine (data collection at Edge locations within the hospitals), weather and climate (data collected by weather and climate sensors, satellites and also users themselves) and radio astronomy. For instance, the Square Kilometre Array is expected to generate around 1 ExaByte (10^{18}) of raw data at several remote locations every single day when it becomes operational in the mid-2020s! Apart from the sheer data volumes, there is a need to satisfy the "quality" aspects of data such as completeness and validity, timeliness of availability and data consistency.

5.2.4.2

Relevance & impact (why chosen?)

Due to the ever-growing digitalisation of the everyday life, massive amounts of data start to be accumulated in Cloud data centres and in HPC clusters, providing larger and larger volumes of (past) data on more and more monitored systems. Simulations can generate huge amounts of data for a virtually infinite number of scenarios to be processed with extreme velocity.

However, the most explosive proliferation in data generation today is taking place across the network, away from these data centres, at its edge. Such new data sources include major scientific experiments and instruments (e.g. the Square Kilometre Array telescope) and a deluge of distributed sensors from the Internet of Things.

Today's data analytics systems need to correlate different types of data (past and present) coming from all sources (Edge, Fog, Cloud, HPC) to understand the status of the system and to predict the its future evolution in a unified fashion. The challenge is to combine Big Data processing techniques (e.g. stream-based processing) with in situ and in-transit data processing techniques inspired by the HPC area in order to simultaneously support processing of such heterogeneous data at extreme scales across the Digital Continuum.

5.2.4.3

Maturity (time to market)

As of today, there is no full understanding of base technologies to implement the end-to-end data storage infrastructure for the Edge-Fog-centre continuum at very large scale. Let us take the prime example of storage and memory technologies. There have been many innovations in recent times: new Persistent Memory devices that now form part of the repertoire of data storage building blocks alongside memory, hard disk and existing flash/solid state technology. These are starting to become available in vendor roadmaps and also have some experimental usage with Enterprise customers. However, the best way to use them is still unclear. For example, there is the question of whether to address data in Persistent Memory as bytes using load/store instructions or as blocks in a storage system using I/O interfaces. This has implications in the design of memory addressing capabilities across large diverse and distributed data spaces. A software ecosystem which effectively uses such newer technologies has still to evolve and the usage of such technologies in the system architectures, as well as system and application software needs to be better understood. This argument also applies to the networking plane which connects the various data pools. Further, the programming models and methods that exploit these new data storage paradigms need to further mature. Systematic and detailed co-design is the key to successfully address these challenges.

5.2.4.4

Hurdles to overcome

For the new Edge-Fog-centre infrastructure paradigm, the speed of making decisions based on data coming in becomes very important in many use cases. However, ubiquitous system infrastructures for rapid data transfer between these different entities are not fully developed today and there is a need for substantial increases in network capacity and bandwidths. Options such as 5G wireless networks, TeraByte Ethernet and InfiniBand XDR are appearing on the horizon. However, the pace of data generation and growth is expected to exceed what the networks can offer, judging by recent and historical growth patterns. There is hence a need to think about new highly distributed systems and multi-level data concentrators or caches over very wide geographies. Adding Fog nodes between the Edge and central data centres can be a solution. Of course, physically hauling the data between the different locations always remains a crude back-stop option until the network issues are sorted out.

There is also the problem of data logistics, or data life cycle management, which has to deal with how long the data needs to be retained in different portions of the workflow, how and when it needs to be moved to the different pools across the infrastructure and how it needs to be shared. Policies need to be carefully developed on a use case by use case basis, determining how long the data needs to be archived and when it has to be purged. Since data can be sourced at any place across the continuum (Edge, Fog

and centre), it becomes crucial to be able to collect and exploit provenance of data by the users (e.g. exposed through programming models).

Data federation is also a major issue as part of data logistics. With data generated in different parts of the infrastructure and possibly in different geographic regions, there is a need to federate and combine all these different data pools in various ways. Different pieces of this federated data will be of interest to different "actors" within the scientific workflow. Related relevant problems are data sharing, data security and data privacy issues across geographies with different data protection regimes. Appropriate handling of "sensitive data" is becoming increasingly important in this context. We cannot assume data can simply be collected and aggregated from the various sources that generate them and be used. Even if the data is not connected to a person, data may be connected to products or to systems, where the provider of the product or the specific system might want data to only be used for very specific purposes and thus needs to be protected accordingly. Further, the diverse data generated are processed through dedicated, specific APIs and programming models (e.g. MPI for HPC, MapReduce or Scala in the Cloud, Edge). There is a lack of unified APIs able to deal with globally federated data. These unified APIs should allow to efficiently integrate simulations and data analytics through extremely scalable data processing architecture combining traditional Big Data processing (batch- and stream-based) with HPC-inspired data processing (in situ, in transit).

In this new, unified data model continuum, data generates models and models generate data. There is a need to move away from unidirectional approaches to be able to support this continuous loop of data and model updates. This ultimately enables a better understanding of a system.

Apart from data that is generated by the users, system telemetry data will be available for analysis; machine learning techniques can be used to optimise infrastructures, enable them to "intelligently" react to changing load or conditions and increase resiliency. For this, telemetry data must be collected, gathered and analysed from all over the infrastructure and decisions and learnings must be fed back into the infrastructure (for example, when is the network adapter going to fail?) Such predictive mechanisms will become extremely important as systems scale to higher and higher complexity and telemetry data analytics becomes very important.

Data consistency is a hard problem to solve because of the different choices made by the Big Data and HPC communities: Big Data developers typically rely on the storage system to coordinate data access, while on HPC platforms, developers use application-level tools to handle data consistency issues. The challenge is to reconcile ACID (Atomicity, Consistency, Isolation, Durability) strict consistency with weaker consistency models, which are used conversely across the network (from the Edge to the Clouds and HPC, according to the processing place of data). This could reduce program development complexity and potentially speed up the processing.

5.2.4.5

Driving competence in Europe & Cost of research to gain significant uptake

Within Europe, many of the FETHPC H2020 initiatives have started to address some of the above problems. Also, many of the organisations (e.g. ETP4HC and BDVA) have also structured their collaboration plans to continue to develop the European ecosystem towards a better collaboration among the Big Data, HPC and AI/ML communities to address the data problems. However, more investment is needed in software frameworks as well as system architectures to enable them to handle the projected "data explosion". Investments are needed for the ground-up training in data management related methodologies and the support of the SME ecosystem around the development of tools, methods and techniques for better management of data. Recently, the SME segment in Europe involved in HPC or Big Data has been shrinking. The promotion of Open Source adoption of many of the data-oriented software frameworks will also be important to lower the hurdles of entry and enable more academic and commercial players to come up with applications, tools and key software components.

5.2.5

HPC and the Digital Continuum[39,40]

5.2.5.1

Intro: What topics are covered by this cluster?

"HPC technology will not only be deployed in dedicated data centres in the future"[41]. "Embedded HPC", "HPC in the box", "HPC in the loop", "HPC in the Cloud", "HPC as a service", "near-to-real-time simulation" are concepts which require new small-scale deployment environments for HPC. A federation of systems and functions with a consistent communication and management mechanism across all participating systems will be required creating a "continuum of computing."

HPC systems will be part of this "continuum" (see Figure 3 and Figure 12), where data will be generated and pre-processed by IoT and Edge devices. An example of a scenario can be: Edge intelligence will extract information from the raw data and will contribute to satisfy the privacy, lower bandwidth and lower latency constraints. Then, if required, information will be processed and transmitted to a hierarchy of devices and servers until it reaches more traditional, larger-scale HPC systems such as supercomputers or HPC-enabled clouds. In addition, new scenarios combining simulation and analytics are emerging (e.g. in the context of the increasing usage of data-enhanced digital twins), where synthetic data generated by simulations run on

39. HiPEAC, "HiPEAC vision 2017", 2019, https://www.hipeac.net/vision/#/
40. ECS, "Strategic Research Agenda 2019, executive summary», 2019, https://aeneas-office.org/wp-content/uploads/2019/02/ECS-SRA-2019.pdf
41. ETP4HPC, "A blueprint for the new Strategic Research Agenda for High Performance Computing", 2019, https://www.etp4hpc.eu/hpc-vision-018.html

EDGE — FOG — HPC/Cloud

Size	Nano	Micro	Milli	Server	Fog	Campus	Facility
Example	Adafruit Trinket	Particle.io Boron	Array of Things	Linux Box	Co-located Blades	1000-node cluster	Datacenter & Exascale
Memory	0.5 KB	256 KB	8 GB	32 GB	256 GB	32 TB	16 PB
Network	BLE	WiFi/LTE	WiFi/LTE	1 GigE	10 GigE	40 GigE	N*100 GigE
Cost	$5	$30	$600	$3K	$50K	$2M	$1000M

Count = 10^9
Size = 10^1

Count = 10^1
Size = 10^9

Count x Complexity = ~Constant

Stateful vs. Stateless

Source: Beckman, Beck, Dongarra, Ferrier, Reed, and Taylor / University of Utah

Figure 12: The Digital Continuum paradigm seen from a count-complexity perspective[42]

supercomputers are jointly analysed with real-time data monitored on Edge systems as part of a global process allowing for continuous refinement and self-improvement of simulation models. Such joint analytics could be performed on traditional HPC supercomputers but also, for instance, on medium-scale, HPC-enabled Cloud-based infrastructures.

These newly envisioned HPC systems (which will carry compute-intensive processing that could be done on HPC centres or on HPC-enabled clouds) will be used to further refine processing according to the incoming information:

- In case of use of *Artificial Intelligence* related tasks, they will support the heavy compute tasks required for learning and finding the best network topology (the metaparameters) of the AI system (the so-called "Auto-ML" approach).
- In the "*digital twin*" approach, the new HPC systems will be used to run and update the simulated digital twin according to the new information captured on the Edge. This digital twin can be used to test new solutions (e.g. personalised medicine), to forecast faults or deficiencies or, in a more time constrained approach, to generate parameters to control in return to the physical systems (e.g. a factory). In this later case of "*cyber-physical system*", the HPC system is effectively "*in the loop*" of sense-compute-react and should compute efficiently to provide "*near to real-time decisions*". In particular cases, the High-Performance System could be a (very) high-end embedded system.
- More classically, the information collected at the Edge on real devices can help to fine-tune the numerical simulations.

The following topics are therefore relevant for this theme of HPC and the continuum:

- Artificial Intelligence related workloads for HPC:
  - Acceleration of computation for IA: current Deep-Learning approaches require heavy computations in two cases: 1/ during the learning phase or 2/when the metaparameters (such as the neural network topology) are automatically generated. The basic operations are relatively simple (multiply-accumulate mainly) but performed on a very large set of data.
  - Efficient communication between computing and storage: during the learning stage of Deep Learning approach, a high amount of data needs to be processed with rather simple operations, therefore the efficient access of data is key for good performances.
  - Support workflows on heterogeneous systems: because hardware will become increasingly heterogeneous in order to increase efficiency, this rise of complexity should be supported in an easy to use manner in the workflows.
- Convergence of Big Data Analytics (High-Throughput Computing - HTC) and HPC to support hybrid scenarios combining HPC simulations and analytics:
  - Unified data storage abstractions and systems enabling efficient data sharing across the Digital Continuum: data have to be exchanged from Edge devices to HPC-class machines, therefore the data should be presented in a coherent and easy to use form for all machines in the "continuum".
  - High-performance data analytics will require HPC/data architectures providing high level of IOPS using tiered memory/storage systems, scalable file systems and high-speed interconnects.

42. Beckman, Beck, Ferries and Taylor (University of Utah)

- Unified real-time data processing techniques favouring the joint use of HPC-originated approaches such as in situ/in transit processing with stream-based processing techniques now common in Big Data analytics frameworks.
- Interoperability and composability of programming models and of the underlying software frameworks and tools used for computation, analytics and learning: As new applications will not only run on one machine but they will also distributed on different computing infrastructures which are different (Edge to HPC), the programming approach should be coherent and encompass all the various forms where computation will be done (and taking care of the communication and distributed form of the systems).
- Interoperability of the data exchange formats (see above).
- Seamless usage of heterogeneous architectures (and of the corresponding low-level software)

- HPC for Cyber Physical Systems (e.g. in the context of digital twins):
  - (Near to) real-time computation to generate near to real-time decisions: as most Cyber Physical Systems will have to close the loop between sensing-computing-acting by reacting on the physical world, the time constraints of the physical world will drive the maximum execution time of the part sensing-computing-acting.
  - Access in real-time to incoming streams of data and streaming out of the results of computation (see above).
  - Cohabitation of processing in stream mode together with classical batch access mode: the real world generates data continuously and due to the expected real-time reaction (see above). This implies continuous processing of streams of data.
  - Enforced privacy and data security: as HPC systems will be open to "untrusted" data and accesses from outside, the requirements of security and ethics such as privacy should be enforced.
- HPC as a service or HPC in the Cloud:
  - Interactive access to HPC resources: new users (such as IA scientists) are becoming accustomed to interactive accesses and not to the classical use of HPC machines with "jobs" and "batches".
  - Elastic resource allocation, orchestration and mediation on resource and workflow: due to their experience with Cloud, users would like to have resources that scale according to their needs in a dynamic mode.
  - Enforced security: similar requirement as for HPC for CPS.
  - Efficient and user-friendly tools for visualisation of large data sets and interactive modification of the parameters of the computation.
- HPC in the box or Embedded HPC:
  - Very high efficiency systems (low volume, power, cost) per TOPS (such as for being embedded in vehicles).
  - Data security

5.2.5.2

Relevance & impact (why chosen?)

The applications are becoming less "in silos" and they are also becoming less "localised" in a single machine or even in a single source code: they are becoming increasingly distributed, collecting information on the fields, processing data on its way to central computing resources and then decisions or new parameters generated by the computations are used to control real-time physical devices. Predicting the behaviour of systems and anticipating what could happen is key for a lot of industries.

Typical use cases are "digital twins": now on the top of the hype curve, they correspond however to real problems in industry (control of factories, anticipation of failures), medicine ("personalised medicine), aeronautics, etc.

5.2.5.3

Maturity (time to market)

The Digital Continuum vision is currently in its early phase of implementation. Therefore, it will take approximately five years to reach maturity.

5.2.5.4

Hurdles to overcome

- The current HPC hardware and software are essentially fine-tuned in order to fit numerical simulations with high precision. To cope with the new workload requirements, the type of computation should be extended (lower precision) with a high-access rate to data. Heterogeneity will be key in improving performance without exploding the power budget.
- The way of using HPC systems will also change: they will be increasingly dealing with (real-time) streams of data than processing data in batches. The data should flow to the HPC system from outside (and vice-versa), with all the consequences in terms of security, access and interconnectivity. Consequently, orchestration and mediation on resource and workflow will be essential, together with interoperability. For example, containerisation might solve some of these constraints.
- So far, the data models and data processing techniques implemented in state-of-the-art software stacks for HPC systems, Big Data analytics and AI-based systems have generally developed separately from each other. New approaches and supporting environments for seamless data storage, sharing and processing across the Digital Continuum (including HPC, clouds and Edge devices) are needed.
- Increased interaction between the HPC, data analytics and AI communities is critical.

These challenges will also have impact on the mindset and practice of people operating these new generation of HPC systems.

5.2.5.5
Driving competence in Europe

- Good knowledge in systems and systems of systems
- Good software developments
- Relevant problems directly impacting the European industry
- All ingredients seem to exist in Europe, albeit in distinct communities which do not interact with one another.

5.2.5.6
Cost of research to gain significant uptake

The research should be interdisciplinary and should link together different communities (HPC, Big Data, real-time systems, simulation, AI, middleware and applications).

Interoperability and standards allowing to seamlessly share data and run computing tasks across the Digital Continuum is still a challenge. These new requirements will lead to a change in the structure of the machines in terms of compute, storage and communication and these requirements might not be fulfilled by off the self-components designed for different purpose. It is important to keep hardware and software knowledge and industry in Europe in order to implement effective co-design, leading to efficient system solutions coping with European requirements.

5.2.6
Resilience[43]

Resilience is widely recognised as a critical challenge for HPC systems because of their increasing complexity at several levels: the data centre level (with critical support infrastructure, such as power distribution, UPS, and cooling), individual hardware and software components, subsystems and complete heterogeneous system configurations. At scale, and particularly with HPC jobs requiring a large number of heterogeneous resources, we can no longer assume HPC system failures to be uncommon events. Moreover, even more challenging failure modes have emerged beyond the assumptions of the commonly used fail-stop model, raising concerns about the integrity of computations as well as data at-rest and in-transit. In spite of frequent failure, application-based correctness and execution efficiency is therefore crucial to ensure the success of the extreme-scale HPC systems and, in a wider context, for data centre-scale systems such as Cloud infrastructure. Further challenges arise from the interplay between resiliency and energy consumption: improving resilience often relies on redundancy (replication and/or checkpointing, rollback and recovery), which consumes extra energy.

Resilience in HPC systems encompasses a wide spectrum of fundamental and applied research and development, including theoretical foundations, failure detection and prediction, monitoring and control, end-to-end data integrity, enabling infrastructure and resilient computational algorithms. Moreover, facility operations and cost management concerns need also to be weighed in, in the context of a systematic risk management framework.

5.2.6.1
Relevance and Impact (why chosen?)

System resilience is one of the hardest Exascale requirements, particularly due to its cross-layer nature. The European HPC community, however, currently lacks a strong concentrated research effort in resilience, which makes resilience one of the greatest challenges of the EU HPC initiative[44]. Ensuring the resiliency of large-scale HPC systems is complex and requires research and engineering effort for the analysis, development and evaluation of reliability features. Additionally, resilience is a vertical problem that needs holistic solutions. For all these reasons, we advocate that HPC system resilience is properly represented in future research and innovation programmes.

5.2.6.2
Maturity (time to market)

Before proposing any mitigation technique, we need to quantify the cost of the failures in the field (i.e. how many server-hours are lost due to the failures) by considering the likelihood and severity of component failures (data errors, hardware failures and temperature-dependent faults) and quantifying the costs of failures (whether rebooting servers, restarting jobs, replacing components, testing failed servers), while taking into account typical HPC job sizes. A cost factor that needs to be considered is the potential increase of recovery time and the associated energy cost. This analysis needs to be done at the supercomputing centres that own the system failure logs and also know the distribution of the sizes of jobs typically executed in production. This analysis will lead to a cost-benefit study of the potential approaches for resiliency, including (but not limited to) stronger ECC, checkpoint and restart, pre-failure alerts, and software-supported RAS.

Specific HPC resilience features could be implemented with a relatively short-term effort, e.g. between one and two years. Some examples of such features are:

- Failure logging and analysis (Tier-0 HPC systems) with provisions for anonymisation and reduction of sensitive operational data to support the development and validation of fault prediction models
- Algorithms for log anonymisation and sharing between European HPC service providers
- Quantifying the cost of HPC system failures
- Pre-failure alerts (HPC system monitoring software), including considerations regarding the expected lifespan and durability of memory and storage devices
- HPC job placement and migration (based on awareness of fault conditions and triggered by pre-failure alerts)
- Application-based checkpoint/restart (potentially triggered

43. Petar Radojkovic, Paul Carpenter, Olly Perks, Reiley Jeyapaul, Manolis Marazakis, and Will Toms: «Towards Resilient EU HPC Systems» (Extended Abstract), November 2018. Public deliverable 2.3 from the EuroEXA H2020 project (FET Proactive, Grant Nr 754337).
44. ETP4HPC, "Strategic Research Agenda: Achieving HPC Leadership in Europe", 2013, page 42, www.etp4hpc.eu/sra

by pre-failure alerts), including adaptation to runtime and storage system load conditions

- Cost-benefit study of various resiliency approaches.

Holistic solutions would require substantially more effort in the order of three to five years.

5.2.6.3

Hurdles to overcome

Technical challenges in improving resilience of emerging HPC systems originate from the complexity of having to consider architecture and mechanism design in a cross-layer manner as resilience is a "vertical" problem, i.e. it cannot be isolated in one of multiple layers but it rather requires consideration of layer interfaces and interactions.

The lack of publicly available failure logs from large-scale clusters, stemming from data confidentiality concerns (including restrictions imposed in vendor-provider contracts), is a limiting factor for substantial investment in improved resilience. There is no openly accessible evidence of the effectiveness and/or limitations of resilience measures, particularly RAS-centred proposals, in large-scale HPC infrastructures. In a similar vein, HPC Resilience efforts also need to overcome non-technical challenges, centred on the justification of resilience-mandated redundancy and cross-layer complexity. We need to consider resilience in an overall cost and risk assessment framework, particularly for long-running resource-intensive HPC jobs processing high-value datasets. The evolution towards converged HPC/HPDA/AI infrastructures will intensify requirements from increased resilience.

5.2.6.4

Driving competence in Europe

Resilience insights and experience come from diverse sources: device manufacturers, system developers and integrators, HPC infrastructure facilities and research centres, HPC and data processing algorithm experts. HPC resilience is a "vertical" problem that requires competencies from different fields:

- Device-level resilience, e.g. for CPU, memory, accelerators, interconnects, etc.
- HPC system-level and node-level integration (including fault sensors and logging mechanisms in all electronic subassemblies)
- System software at the node level: e.g. failure detection, reporting and failure impact mitigation.
- System software, HPC system monitoring: e.g. failure logging and reporting, pre-failure alerts.
- Data analysis: quantitative analysis of system failures and their impact, failure prediction
- Building resilient HPC algorithms and applications.

5.2.6.5

Cost of research to gain significant uptake

The research effort has to be cross-layer and cover both technical and non-technical challenges. Application correctness and execution efficiency in the presence of failures is essential to ensure the success of upcoming extreme-scale HPC systems.

5.2.7

Trustworthy computing

5.2.7.1

Intro: What topics are covered by this cluster?

Core security in HPC is based on establishing a secure perimeter, focusing on user registration, authentication and permission management for access to the infrastructure and access to files via LDAP and Linux file permissions. The lack of mechanisms was motivated by the less critical nature of the workloads e.g. simulations of natural phenomena and closed (trusted) environments.

With the increasing amounts of data produced in many emerging domains, HPC has shifted from its classical use in protected centres towards embedded HPC devices in the Edge and Cloud environments. Typical HPC domains such as parallel processing are adopted by a wider user base such as data scientists in a diversity of domains.

The type of processing extends from classic simulations to processing of a multitude of sensor data, including personal (e.g. medical) and business critical (e.g. in financial analysis) data. There, one of the main challenges for organisations will be to maintain data confidentiality/privacy, integrity and security. In addition, legal and business requirements such as regulatory compliance in service-oriented environments and assured fulfilment of security-related service level agreements have to be satisfied through technical means.

Hence, future developments and HPC related research must consider cases other than the one of the heavily protected, central data centre and factor in threats not usually encountered in scientific HPC.

We will focus on the technical aspects of providing suitable hardware/software platforms. This includes HPC chip and system design, runtime environments, storage solutions all the way up to algorithm and application design and monitoring. Only if the complete stack has consistent security measures in place, one can assume a trustworthy computing environment.

Moving HPC technology to the Cloud means virtualisation of hardware and concurrent use of systems or at least interconnects by different users (multi-tenancy). Each of the users must be guaranteed to own and work in a secure environment.

Starting with the trust into **hardware (and associated firmware)**, the semiconductor system components such as processors, memories and networks, etc. must fulfil basic security requirements and minimum standards. In processor design, Spectre and Meltdown[45] vulnerabilities have shown that even on the chip le-

45. https://meltdownattack.com/

vel security leaks are possible. The HPC chip design must have this in focus.

Whole compute nodes need to implement a secure boot procedure to check whether hardware or firmware components has been changed and then whether to introduce mutual authentication procedures.

In a shared environment, special hardware (e.g. FPGAs or ASICs) might be used for on-the-fly encryption of data without reducing bandwidth to network attached storage devices for I/O, or as Root-of-Trust such as the Google Titan chip. Moving to open instruction-set-architectures and open source hardware could additionally build trust with potential customers.

With the increasing number of components in future Exascale supercomputers, resilience must be implemented in hardware to provide basic fail-over mechanism in case of attacks or hardware failures. This includes a certain degree of redundancy, monitoring and new approaches for bug-free HPC hardware design which might be adopted from the embedded space where microcontrollers take over tasks in critical systems. More details are covered within the *Resilience* research cluster on page 45. Of course, such provisions will increase costs and energy use and their gain has to be balanced against these costs.

Trustworthy hardware must be complemented by **secure operating system and runtimes**. The user must know that the service or infrastructure acquired is sufficiently protected to handle sensitive data.

HPC data centres are used to provide their resources to customers as "bare metal", meaning that customers receive full control over the hardware and therefore can achieve full performance. However, if HPC centres are moving towards interactive and dynamic service provisioning, resource sharing is a potential way to improve utilisation and decrease costs for the customers. Resource sharing has the downside compared to bare metal machines in that the actions of one party could influence another. A possible approach is to provide hardware and services using virtual machines or containers. Virtualisation techniques enable to securely compartmentalise operating systems and applications of different criticality or owners and to retain a level of control in case of an attack by accessing the host and isolating the infected virtual machine. Thus, virtualisation architectures should enable full security/performance isolation at all levels. However, using additional layers can decrease the absolute performance for the customers, who must implement bullet-proof mechanism to avoid any access across virtualisation boundaries.

In addition, use of container technologies or hypervisors is a common trend, especially in constrained environments as it is the case in embedded systems. Container services and hypervisors are usually not developed with specific HPC requirements in mind and offer a varying degree of security; initial use of containers for HPC applications, for instance in Life Sciences, has shown ways of overcoming performance limitations.

Adapting Cloud-like business models to HPC systems must include developments and research on secure hypervisors and container services by also offering some bare metal methodologies such as RDMA, PGAS, etc. Data access is one of the most critical events when it comes to security. Security must be ensured across operating system file systems, object stores, etc.

For especially critical applications, features such as logical or even physical separation into user specific sub-clusters should be considered (secure partitioning).

A trustworthy environment critically relies on trustworthy **algorithm and application design**. HPC applications and programming languages have been developed with performance and fine-grained control in mind. However, the more control the language offers to programmers, the more vulnerable the resulting code might be. The effort put into a secure application design obviously depends on the task performed. For example, handling personalised or other critical data in HPC should require certain standards in application design methodology and application implementation review. New programming paradigms for HPC could help preventing security leaks through bugs in the application implementation. This could include formal methods or new programming languages.

Algorithm research could include applying HPC technologies to deal with e.g. encryption (homomorphic or not) and privacy preserving transformations - e.g. for artificial neural networks. Trustworthy environments must also include a high degree of fault tolerance to be able to recover from software and hardware failures.

As described, to build a trustworthy environment the whole **software and hardware stack** must be considered. However, such environments must also be monitored to verify security and identify potential intruders (applications, scripts, hardware levels).

Moving towards new usage models such as "HPC in the loop" requires **external interfaces** and predefined **data centre APIs**. To foster a common European HPC landscape and lower the barriers for the on-boarding of new customers, such interfaces should implement standardised security features and authentication methods as well as a common API where a certain endpoint could define the targeted data centre.

The demand for such interfaces will increase in the future with the growth of data and processing in the Edge.

Enabling a more dynamic HPC data centres **user registration and permissions** must be streamlined and standardised across data centres. A possible feature of **user registration and permissions** is to bypass Linux users and file permissions and introduce e.g. roles, sub-groups, user key management, which is common in Cloud environments. "Super-Users" of organisations should be able to organise and register their own user groups. User specific keys would be used for data encryption.

Data centres should build trust through certificates such as ISO 27001. Continuous monitoring mechanisms should be imple-

mented by the data centre provider as well as to the customer to quickly react to critical events.

There is an increasing demand for regulated data usage in HPC data centres with respect to GDPR regulations and alike. Data centres should offer support to customers with such demands. These regulations currently mainly implemented in Europe will also likely be adopted in other parts of the world. Having HPC solutions able to cope with such requirements can become a competitive advantage for Europe.

5.2.7.2

Relevance and impact (why chosen?)

Trustworthy computing is important in several dimensions. First, HPC centres are faced with a new customer base coming from data science, dealing with the analysis of heterogeneous data sources and often dealing with critical data such as personal data or business-related data. Such a customer base is used to different usage patterns such as interactive computing, which introduces challenges for the existing HPC data centres. Further dealing with critical data, the new HPC customers have different requirements in terms of SLAs such as the privacy or security of the infrastructure.

The second dimension is the application of HPC technology in new domains such as the CPS and embedded market where security and reliability are key topics. The advantages of taking up HPC here can only be realised after the security challenges are solved.

Finally, European HPC technology developments must ensure strong security measures to be able to deter a multitude of attack scenarios and international threats. Currently, the value of HPC to science and business makes HPC systems a target of security attacks (following the trends in other computing areas).

5.2.7.3

Maturity (time to market)

Most of the techniques required are known and deployed in other domains such as Cloud and Enterprise computing and sections of embedded computing. Deployment and use in the HPC sector will mainly require adaptation to meet HPC specific needs and integration with the prevalent system and software infrastructure in the field.

5.2.7.4

Hurdles to overcome

A significant challenge in adopting existing trustworthy computing mechanisms for HPC is the necessity to minimise the overhead of such mechanisms with regards to delivered compute performance and energy efficiency. Secondary challenges arise from the inevitable change in established usage procedures and the integration with HPC-specific system and software elements (such as high-performance fabrics and parallel programming environments).

5.2.7.5

competence in Europe

Research on this topic must follow the co-design principle. Current developments that take place in silos must complement each other to form a trustworthy hardware and software stack as well as cross-disciplinary including customer domains and Edge, Fog, Cloud and the HPC community.

The starting point of this research could be to drive cross-disciplinary projects not just spanning different application sectors such as medicine, agriculture or oil & gas but also different communities (e.g. embedded, Edge, HPC and AI). The collaboration of expert groups such as ETH4HPC, AIOTI, Big Data Value, etc. should explicitly be motivated to generate such joined efforts. Cross-community conferences, education and training might help as well.

5.3

Research Domains

The Research Domain sections below have the following structure: first, the state-of-the art of the given area is described; then, the main challenges for the period from 2021 to 2024 are identified; and finally, intersections (or “overlaps”) with any of the research clusters identified above are shown - these crossing points constitute the areas which should constitute the future R&I Work Programmes.

5.3.1

System Architecture

5.3.1.1

Research trends and current state of the art

The design and configuration of HPC systems has changed over time, continuously striving to increase the computing performance while reducing procurement and operational costs. After hitting the power wall in 2004, the core frequency has stagnated, while the number of cores per node and the number of nodes per systems has steadily increased and with it the overall core-count in HPC systems. This has consequences for failure rates and raises resilience concerns. Power consumption has become a major concern and a clear trend towards heterogeneous systems has been established: combining general purpose CPUs with different kinds of acceleration devices such as GPUs, many-core processors, VPUs[46] and application specific accelerators. Accelerators deliver a high Flop/Watt ratio but come at a price of increasing programming complexity, with application developers often required to rewrite significant parts of their codes to be executed efficiently on these devices. The orchestration of tasks on different accelerators also requires efficient communication and adequate programming models.

The trend towards heterogeneity is not only noticeable on the compute part of HPC systems but also in their memory and storage architectures. Deeper memory and storage hierarchies are built using a variety of technologies, allowing to optimise for either higher bandwidth (on the upper layers of the hierarchy) or higher capacity and lower price (on the lower layers). Relative newcomers are high-bandwidth memories (HBM) within the node (both for CPUs and accelerators - GPUs or FPGAs) and non-volatile

46. https://en.wikipedia.org/wiki/Vision_processing_unit

storage class memories (SCM) used both as memory (NVDIMM) or as a middle layer storage in front of solid state/NVMe drives or traditional hard disks.

The majority of modern HPC systems continues to utilise a scale-out (clustering) architecture. A critical aspect of the HPC system design is the high-speed interconnect and its topology. Tree topologies are currently the preferred approach, with new designs (e.g. butterfly, dragonfly, dragonfly+, etc.) deviating from the traditional fat-tree topology in order to improve TCO by using less switches and cables for a given number of nodes, decrease latency by reducing the number of network hops and decrease congestion by offering alternate routing for heavy traffic, while also improving extensibility. Network components (NIC, switch, cables) continue to scale with node performance by increasing network capabilities from one generation to the next. In addition, upcoming optical technologies such as Silicon Photonics (SiPh) are expected to enhance future networking and interconnect fabric capabilities and capacities. Regarding network technologies, high-speed interconnects (and in particular InfiniBand and Cray fabric) are the most frequent choice for high-end HPC systems. However, Ethernet – commonly used in large scale Cloud-computing and hyper-scale computing centres – is gaining popularity and latency-optimised implementations of the protocol are arising, which might strongly influence the interconnect landscape in the HPC market segment. Optical interconnect, even the today very efficient blade to blade interconnect, can migrate to on-board interconnect and, at the package level, it could work with photonic interposers in the future.

The high computing density of today's machines and the pressure to reduce power consumption requires an increased focus on the packaging, cooling, monitoring and power aspects of the system. Direct liquid cooling, immersion cooling techniques and free cooling are seen as major trends for saving power. In addition, a more flexible reuse of the waste heat generated by HPC systems can be further developed in order to reduce the overall Power Usage Effectiveness (PUE) and gain better sustainability.

For better modularity of HPC system architectures, more (hardware) disaggregation technologies based on standardised interfaces should be explored on the sub-system level such as compute acceleration, processing, memory, networking, I/O and storage. This will allow each sub-system to evolve and improve at its own pace without the need to compromise with other components that otherwise are too tightly connected with and dependent on each other.

5.3.1.2

Challenges for 2021-2024

System architecture design faces a constant challenge: delivering the maximum compute performance at the best possible energy and cost efficiency. What changes with time is the order of magnitude in performance and the specific technological challenges to be addressed in order to reach it. Additionally, the convergence of HPC with HPDA and AI, as well as new usage models for HPC (e.g., HPC everywhere) brings a far more diverse set of requirements. Properly addressing them is a new challenge for component and system architects.

5.3.1.2.1

Integration of heterogeneous resources

Numerous and diverse low-energy and more well-performing compute technologies (CPU, GPU or other accelerators, application-tuned programmable logic, etc.) are appearing, increasing the performance per watt ratio for a given set of applications and workloads. Also, revolutionary non von-Neumann computing approaches bring new hope in overcoming the end of the Moore's Law era. For instance, neuromorphic and quantum computing are able to deliver performance levels unreachable by any standard computer when solving specific classes of problems, such as pattern recognition or problem optimisation. They benefit from new silicon node technologies and micro-packaging techniques, providing larger silicon area per component. System architects will need to seriously consider how these new types of computing components can be integrated and adapted to the traditional HPC environment both from a hardware and software point of view.

The expanding diversity of computing elements calls for new system architectures which should be able to orchestrate them and share them in an efficient and flexible way between applications. Several approaches exist that take advantage of the evolution of both standard and proprietary interfaces and protocols (e.g. PCIe, CXL, CCIX, GenZ, NVlink, OpenCAPI, RDMA, RoCE) in order to interconnect the system components. The traditional host-device approach, in which, within a node, one or more accelerators are attached to a host CPU that takes over booting, orchestration and network communication capabilities, has evolved in the meantime towards "island-constructions" where accelerators (in particular GPGPUs) are interconnected with each other building "very fat nodes". On the other hand, composable node-designs are best suited for creating heterogeneous nodes with the right mix of components (CPU, memories, accelerators, network) for specific problems. An example of the alternative or complementary European approaches to the heterogeneous node approach is the "modular supercomputing architecture"[47](MSA), which interconnects a series of potentially large clusters – so called "compute modules" – at the system level instead of at the node level. The individual modules (clusters) are tailored to the needs of specific parts of applications and workflows and a common software stack enables codes to run orchestrated across different modules. Similar approaches are being applied internationally (e.g. the Chinese Tianhe-3 system[48]). The variety of computing resources opens new possibilities for "reconfigurable computing"

47. E. Suarez, N. Eicker, Th. Lippert, «Modular Supercomputing Architecture: from idea to production», Chapter 9 in Contemporary High Performance Computing: from Petascale toward Exascale, Volume 3, pp 223-251, Ed. Jeffrey S. Vetterm, CRC Press. (2019) [ISBN 9781138487079] https://juser.fz-juelich.de/record/862856
48. "Heterogeneous Flexible Architecture", Slide 40 in https://www.r-ccs.riken.jp/R-CCS-Symposium/2019/slides/Wang.pdf

which offers better application efficiency by adapting the system to the needs of each individual user. Virtualisation and containerisation approaches provide each user with a "virtual cluster" according to specific needs. This is applied at the system level by allocating the right mix of compute components and at the node level by using programmable devices such as FPGAs or data-flow engines. Such approaches also open HPC to Cloud usage models, enabling higher requirements in terms of security and isolation between users. In this new context, research is required in order to optimise scheduling, resource allocation and sharing of multiple (potentially heterogeneous) resources as well as to minimise the overhead of the resource abstraction and virtualisation. Ideally, enough intelligence should be available in the system software for it to decide on which hardware to run each part of an application without transferring this burden to the user. Furthermore, standard programming interfaces, adaptive libraries and APIs are needed to facilitate programming the new devices and improve performance portability.

5.3.1.2.2

Memory and storage hierarchy

New memory technologies shall enable increasing the memory capacity inside the node, with new memory models employed to exploit it. Deeper memory hierarchies shall increase the effective bandwidth and reduce the effective latency. Additionally, non-volatile, byte-addressable "storage-class memory" can improve I/O bandwidth, IOPS and scalability, speeding up data transfers between workflow steps and addressing extremely memory intensive (capacity-wise) applications. On-volatile, network-attached memories are expected to optimise data movement by storing data in the network, accessible to all networked computing elements, e.g. via NVMeoF. With intrinsic computation capabilities, such devices can optimise collective operations and further improve application performance. Memory hierarchies and I/O node architectures within HPC systems remain highly active topics in view of expanding access methods (parallel file systems, object stores, etc.) to data at Exascale levels and beyond. New persistent memories could also trigger a move towards more direct object storage, where access is not any more by addresses but rather by "keys" (key-value storage).

5.3.1.2.3

Network

The increasing system size and the pursuit of energy efficiency require new means to ensure scalability, serviceability, manageability, reliability and resiliency. A key element in this context is the system network, which must become more efficient and adaptable. First of all, better network performance is needed both within the package (on-chip), within the node (processor to accelerator) and also between nodes (connecting nodes to other nodes and computing to storage infrastructure). Improved connection to the network can be achieved through: faster links (higher SerDes speed); wider interfaces, higher message rates; improved reliability (error correction feature); multi-path and adaptive routing; more efficient communication with compute elements through more feature rich interfaces (improved caching mechanism or tighter integration with compute element); and also offloading parts of network-relative computing to the fabric when it reduces bottlenecks on compute side (collectives, messages/tag matching, PGAS effective support, etc.).

Additionally, cabling technologies need to be further improved in order to reach acceptable bit error rate, better TCO and good serviceability for long and short connections. Optical connections and optical co-packaging will become more common both to bridge large distances between cabinets and to improve bandwidth with in-node optical links. Fault-tolerant routing, network virtualisation and software-defined networks will be necessary to recover from node/fabric failures, to reduce jitter and performance interferences (through isolation of jobs when driven by intelligent scheduling strategies), and to enable network reconfiguration ("smart networks").

A fabric is shared by many users, consequently quality of service and congestion management are important features, such as CRAY's Slingshot or Alibaba's HPCC High Precision Congestion Control approach. Fabric management shall include mechanisms to control node behaviour, alternate route selection, different traffic class allocation, etc. Cloud usage or sharing of infrastructure between different tenants/users requires new security mechanisms (fabric partitioning, encryption, etc.) and new access style for HPC machines (interactive login, "front-end machine filtering access and ensuring security, etc.).

An additional level of complexity arises with the federation of computing resources between HPC systems and the Edge, comprising the computing centres, Cloud-computing services, local clusters, data-generation instruments and IoT. Smart network approaches with some computation performed at the fabric-level could be valuable to perform on-the-fly computation and reduce the need for expensive data transport. The network could take over more tasks than merely the transport of information from one place to another within a system. In particular, on-network, acceleration technologies could help reduce data movement and speed-up collective operations. Network-attached memory and on-network computing techniques are clearly beneficial technologies worth exploring further.

5.3.1.2.4

System-level integration and Sustainability

Handling the evolution of individual components is amongst the first targets for system integration. The maximum possible power envelope interacts directly with component maximum performance because it is one of the major bottlenecks. The cooling techniques that are key for reliability, TCO and density become attractive for pushing up performance limits. Pushing the limits on liquid cooling such as increasing the maximum power for free cooling is an example of the development path. Energy reuse and a better controlled energy supply (capping, energy consumption optimisation, monitoring) are areas where active research improves HPC carbon footprint. Additionally, monitoring with access to diverse sensors should enable identifying the power-hungry components and determine how and where to execute different

parts of applications and workflows in order to achieve the best overall energy-to-solution. Applying AI techniques on monitoring data promises progress in this area.

Interposer techniques, connectors, copper and optical cables, optical co-packaging, printed circuit board materials and form factors, and physical 3D implementation must be improved continuously to meet design constraints. Main objectives in these fields are allowing faster operation frequency, improving density to allow large scale systems, providing serviceability and reliability allowing an effective maintenance of very large data-centres.

HPC system deployment, management and maintenance is heavily based on a platform management infrastructure (out-of-band microcontrollers, software stack, management network, management nodes, etc.). Targeting larger systems, adding new security constraints and adopting new usage models such as Cloud requires a modernisation of the software management stack. The main outcomes could be system management through restful interfaces.

Promoting open and standardised system integration frameworks seems a promising initiative to support development by different suppliers of the numerous heterogeneous nodes described in previous sections while offering an effective HPC system integration umbrella.

Furthermore, a more sustainable approach using less raw materials would be desirable. It should holistically encompass production phases and system lifetime. As an example, the extension of existing machines over time with the highest possible reuse of system and infrastructure parts over various generations should be anticipated since their initial deployment.

5.3.1.2.5

Co-design

All the above improvements on system architectures have to be implemented taking into account the evolving needs and characteristics of applications. The only way to make it possible is applying stringent co-design approaches in a coordinated development of hardware, middleware and applications. As modern and future HPC systems will be used for a wide variety of fields, co-design must include realistic use-cases and datasets from all relevant fields: HPC, HTC, HPDA, AI, ML, DL, etc. These have to be supported not only individually but also in combination with each other in complex, orchestrated application workflow scenarios that can also be executed concurrently.

Accordingly, computer systems should be adaptable to very diverse requirements. The decision on which parts of the heterogeneous system a given code is executed should be supported by system modelling and simulators that enable forecasting application performance on different system configurations. The deployment of codes onto the system should guarantee the security of the sources and their associated data e.g. through application containerisation and isolation. Additionally, methods to manage huge amounts of data from the system to the Edge, including large experiments (e.g. LHC[49], SKA[50]) are required. Independently of the chosen system architecture, component balance within the overall system to achieve the maximum efficiency for real-world applications should be considered as the most important goal. Adequate benchmarks and associated evaluation metrics should be chosen to characterise real application behaviours. In addition, the delivered performances and energy consumption have to be estimated for the aforementioned relevant application classes.

5.3.1.3

Intersections with Research Clusters

5.3.1.3.1

Development methods & standards

Performance portability across varying and heterogeneous Performance portability across varying and heterogeneous architectures: standard interfaces needed, inclusion of discovery and adaptation logic in common support libraries (mathematics, etc.) and middleware.

- Co-design of applications, runtimes and architectures: standard techniques and processes are needed in this area, which currently is not well defined (the term is understood differently by different people)
- Full system architectures, with all elements designed together as an ensemble with a given goal: methods needed to cover “hardware-software co-design” area.
- Modelling system components and whole system: modelling methods based on standard, vendor-agnostic criteria are needed.
- Standard APIs to integrate heterogeneous accelerators and new computing elements (e.g. quantum, neuromorphic).
- Trade-off between market-specific designs and cross-sector all-component system design

5.3.1.3.2

Energy efficiency

System architecture choices have impact on power consumption.

- Required ability to power-off parts of system (e.g. cores in the node, accelerators) when these are not in use, in order to reduce consumption when idling. Ensure a secured share of resources in order to avoid unused resources when applicable.
- Heterogeneous accelerated architectures for better power efficiency: systems are becoming increasingly heterogeneous.
- Power system architecture: a holistic approach to power design and distribution in order to reduce losses through power-conversion.
- Cooling: direct liquid cooling technologies should be used

49. https://home.cern/science/accelerators/large-hadron-collider
50. https://www.skatelescope.org/

for better energy efficiency and sustainability, also to reduce TCO and operational noise pollution levels.

- Develop the software to estimate and visualise energy consumption both for operations staff and users. Metrics and precision shall be clearly known and exposed.
- Optical co-packaging can reduce network power by up to 30-40%.

5.3.1.3.3

AI everywhere

- Use of AI to improve system utilisation, power consumption, efficiency. HPC for AI and AI for HPC.
- Use of new devices, e.g. neuromorphic architectures and quantum annealing: there is a need for ways to integrate them into the system architecture.
- Integration of systems into the Digital Continuum with good connectivity between HPC system and the Edge: networking aspect of system architecture.
- Heterogeneous HPC architectures can contribute to increasing the scalability of scalable and HPC-AI solutions: system architecture needs to accommodate the needs of AI.
- Distributed AI and in-network computing acceleration: training times shortened with ASICs or other specific acceleration devices.

5.3.1.3.4

Data everywhere

- There is a need for interconnectivity and on-the-fly computing (from Edge to HPC centre) to perform critical calculations and reduce data close to its location: i.e. the network aspect of system architecture and also of overall HPC- and data-centre architecture.
- Extremely scalable data processing architecture combining traditional Big Data processing (batch and stream based) with HPC-inspired data processing (in situ, in transit): there is a need for unified APIs to integrate simulations and data analytics ways of using HPC.
- Reconciling different architectures with respect to data addressing: there is a stronger need for byte-addressability instead of block access.
- Extremely scalable, flexible object-based data storage (allowing metadata for tracking provenance and accesses, complex security policies, high availability through easy replication) instead of hierarchical (POSIX) file systems.

5.3.1.3.5

HPC and the Digital Continuum

- Small-scale HPC deployments and computer federation: specific system architectures might be needed including solutions to the connectivity and federation between them.
- Embedded HPC: eventually problem-specific system architectures will arise.
- Seamless use of heterogeneous architectures (connection also to system software and programming environment).
- Extended lower precision calculations with high-access data rate: hardware optimisations for low/mixed precision and data exchange, completed/in coordination with research on algorithms and/or mathematics. New number representations (e.g. POSIT) are being proposed, which need to be supported by the hardware/software system.
- Real-time streams of data need to be processed: architectures must be able to very quickly and efficiently transfer data between storage, memory and computing entities, network connectivity and imperative scheduling become key. Performing operations in-memory, in-storage and/or in-network could offer advantages, for the implementation of which such functionalities need to be further developed.

5.3.1.3.6

Resilience

- Heterogeneous systems with increasing complexity bring more failures. Hardware and software approaches (including AI-enabled predictive involvement) are required in order to reduce application crashes and automatize application re-starts with minimum time-loss.
- Failure detection, prediction, monitoring, recovery: all need to be integral part of the system architecture.
- Log-sharing between sites: maybe not really a system-architecture topic, more a “political” one. Establish common metrics between sites.

5.3.1.3.7

Trustworthy computing

- Sharing an HPC system in a trustworthy manner requires secure support for virtual machines or containers. As a result, any definition or choice of software and hardware components must take into account its trustworthy behaviour. This applies to processors, accelerators, interfaces, operating system, I/O drivers, resource managers, deployment tools and more.
- HPC in the Cloud means that data privacy has to be provided between users but also between users and system administrators. Data encryption and a secure key management system are the natural answer. In the HPC world, performance and power efficiency constraints may modulate the chosen architecture depending on the HPC system usage.
- Seamlessly connecting external sensors or scientific equipment to the HPC Data centre is a new feature that the HPC system architecture must meet.

5.3.2

System Hardware Components

5.3.2.1

Research trends, current state of the art and future evolutions

The system hardware elements can be decomposed into:

- Computing elements (including accelerators)
- Memories and near processing storage
- Short distance interconnect
- Integration of computing, memories and interconnect, packaging.

For each domain, the current state of the art and the future evolutions will be summarised with the short-term priorities being ranked in the following section "Challenges for 2021-2024".

5.3.2.1.1

Computing elements

Accelerators such as GPUs have been introduced in HPC based on their high energy efficiency and huge peak performance compared to general purpose processors. Eight of the ten top systems in the Green500 of November 2019 use GPUs, while the Top 1 system uses the Fujitsu A64FX processor and the Top 2 system is based on a massively parallel PEZY-SC2 associated with a Xeon D-1571 16C 1.3GHz CPU. Also, the Top10 of the fastest supercomputers includes accelerator-based designs and purely CPU-based supercomputers. This trend shows that the underlying processing technologies for HPC become very heterogeneous and many supercomputers will include a mix of accelerators and classical CPUs. While today more than three quarters of the accelerators are based on Nvidia GPU technologies, it is expected that AMD will gain significant shares of the HPC GPU market, e.g. by Frontier - the first Exaflop system in the USA to be deployed in 2021.

Today, GPUs are a mature technology and are a good match for vector or SIMD-kind operations, while CPUs execute most of the scalar and sequential operations. Including GPUs enables executing most of the compute power at a high energy efficiency level, while this comes at the cost of more complex programming. This is a challenge for the programming tools, which should hide the underlying hardware characteristics as much as possible. Autovectorisation is available on compilers, while pragma-based programming models, such as OpenMP (and derivatives) and OpenACC help portability of programs and help the programmers, which still should have knowledge about the underlying architecture to design efficient codes.

The heterogeneity of accelerators might also include FPGAs, dedicated vector processors such as NEC's SX-Aurora and specific machine-learning accelerators in the future. FPGAs are very efficient when dealing with workloads that can be executed in a spatial (parallel) way, with a relatively limited complexity in data representation (operating rather at the bit level than using double precision floating-point representation). CGRA (Coarse Grain Reconfigurable Arrays) are like FPGA but with reconfigurability at coarser grain than for FPGA. Dedicated Artificial Intelligence accelerators are developed mainly to deal with the matrix operations with relatively low precision done in most of Deep Learning schemes.

It should also be considered that CPUs are becoming more energy-efficient by moving accelerator-instructions into the CPU, for example, by adding large vector instructions. The Fujitsu ARM-based A64FX processor for the Fugaku (Post-K) machine has 48 compute cores plus 4 assistant cores and SVE SIMD vector extension with 512-bit width. Intel's processors have advanced vector extensions of also 512 bits width elements and new encodings are being added, such as bfloat16 for deep-learning acceleration in Intel's Cooper Lake.

For short term products, the targeted technology is 7 nm (Fujitsu's A64FX, AMD's Milan, Intel's Sapphire Rapids) and the number of cores is increasing to 64 cores. We can also observe a diversity of ISAs and CPUs: Intel is still dominant in the systems' share of the TOP 500 and it is expected that AMD will be able to take over a significant share of new x86-based installations. The number of systems being based on IBM Power seems to be very small, while it is powering part of the Number 1 and 2 systems Summit and Sierra. The ARM64 architecture is slowing gaining its place in the HPC domain: the Riken's Fugaku machine will be using ARM64+SVE extension architecture, without accelerators.

The RISC-V architecture has also to be considered but rather for designing accelerators: in the short term, its ecosystem seems not to be mature enough and work is needed so that the architecture can be used in the future also as the main processor in an HPC machine.

Due to the new usages of HPC (data intensive, AI), the systems might need to be more open to the outside world, with the related increasing problems of security, access control, etc. These considerations will have to be taken into account also at the hardware level, similarly to what is done for Cloud based systems.

5.3.2.1.2

Memories

Local memories are becoming heterogeneous, with a mixture of HBM or HBM2 for high bandwidth and DDR for volume storage. This is driven, for example, by Big Data analytics and Artificial Intelligence workloads which need more data accesses (and more bandwidth) and other classical memory-bound HPC workloads such as QCD. Many new workloads are memory-bound on traditional hardware, with a byte/flop ratio of 10 or higher.

We also observe a blurring of the lines between memory and storage, especially with the introduction of NVRAM and persistent memories. This might also change the classical memory hierarchy with several levels of caches, main memory and storage. How caches are distributed and intertwined with communication between cores and accelerators is also a very active system architecture topic. Byte addressable persistent memories allow new abstraction and, for example, more structures using key-value than address-data. Byte-addressable NVRAM (such as Intel's 3D XPoint) will contribute to the performance increase if used

with an optimal data placement into the different memories according to their specificities.

5.3.2.1.3

Interconnect

On half of the Top10 of the TOP500 systems InfiniBand interconnect is used (generally provided by Mellanox in EDR or HDR versions). Cray has its Aries and now Slingshot interconnects technology and Intel Omni-path - OPA – is also used but Intel will not further develop its OPA200 technology and it will rather develop Ethernet-based technologies following the acquisition of Barefoot Networks. Fujitsu systems use the Tofu interconnect. The Enhanced Data Rate (HDR) InfiniBand frameworks is evolving towards a 400 Gbps version (XDR), an upgrade over the current 200 Gbps, utilising 100Gbps serdes technology. Europe has its own developments with Atos-Bull's BXI and EXTOLL.

PCI Express bandwidth is increasing more rapidly than ever, while Gen 4 is already being used; Gen 5 (32Gbps per lane) and Gen 6 (64Gbps per lane) are in development.

Disaggregation over the network of Compute and storage resource drives more bandwidth requirement from the network and requires disaggregation protocols (e.g. Non-Volatile Memory Express over Fabrics, NVMeoF).

Interposers will allow having high bandwidth between chiplets, with low latency. Photonic interposers are under development in research organisations and can improve further the bandwidth between chips.

"In-network" computing (see Figure 13) is a new method to accelerate HPC and AI application by moving portions of the HPC application to be run "in the network". This approach was introduced in InfiniBand EDR and HDR generations and today it is used to accelerate MPI collective operations and distributed AI training workload. The potential to process the data on the fly in the network will enable new use cases and new HPC acceleration frontiers.

5.3.2.2

Integration and packaging

One of the important parts of the energy budget is in communication: moving data takes time and energy (orders of magnitude difference between an on-die transfer and transfers between chips on boards, or even worse between boards). By reducing the distance between compute nodes (general purpose processors and accelerators), networking, memory and storage (persistent and non-persistent), it is expected to further increase efficiency of nodes. The ultimate option is the emerging field of "computing near or in memory" architectures but it is not mature enough to be used in short term production systems.

Modularity and composability are also important requirements of current machines: composable nodes (comprising CPU, accelerators, DDR or HBM memories, persistent storage, network interface, interlinked by (a) (coherent) switch(es)) allow to tune the efficiency towards the different workload; composable racks (changing the ratio between compute and storage) are also emerging.

3D stacking, and the approach using chiplets and interposers is increasingly used, together with 2D and 3D packaging.

For example, Intel has its EMIB (Embedded Multi-Die Interconnect Bridge) technology allowing to have several dies (e.g. processor and memories) in the same package and it is announcing its Foveros technology which will allow to mix chiplets from different technologies on the same interposer. For its Rome generation of Epyc processors, AMD uses 8 7nm chiplets (for a 64 cores processor) arranged around a 14nm interconnect and memory interface die. The European Processor Initiative (EPI) is also following this approach using the interposer and chiplets approach in their Common Platform. (cf. "Technology sourcing").

Having the possibility to integrate different dies (chiplets) in the same package allows to reduce the global cost (several small chiplets have a higher yield than a big chip) and facilitate diversity (the same chiplets can be used with different combinations, e.g.

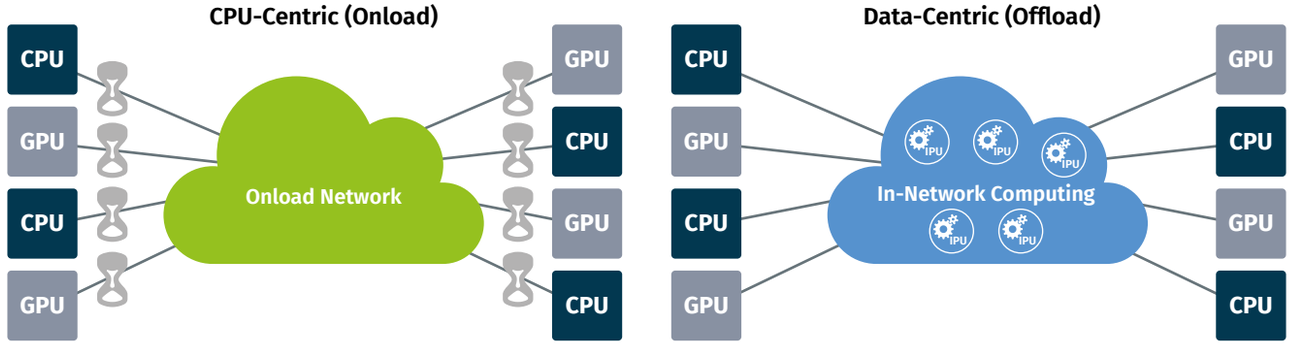


Figure 13: Onload Network vs. In-Network Computing

to adapt to different markets by changing the computing/memory ratio for example). It also reduces the length of interconnect, and therefore also the power lost in connection, and increases the number of wires compared to a PCB, therefore also increasing the local bandwidth. Interposers help combining dies or chiplets made from different technologies optimised for their purpose (e.g. logic, memories and analogue devices for power converters).

Lowering overall interconnect power consumption drives the industry towards packaging the optics close to the electronic interconnect, which is called “Optical co-packaging”. This new way to build system can be done with either VCSEL technology or Silicon photonic.

Reducing power consumption is also key to reduce the cost of cooling, which is mainly taken into account at the system integration level. Most packaging solutions of chips allow freedom to use air cooling, cold plates or liquid cooling (or even total immersion).

5.3.2.3

Challenges for 2021-2024

We have seen an evolution from increasing the clock speed through increasing the number of cores (due to the end of Dennard’s scaling) to now adding more specialised and more efficient accelerators. Perhaps tomorrow the architectures will increasingly avoid data displacements. This evolution is schematised in the following Figure 14:

Following the evolutions of computing observed in the last section, the two main challenges for the forthcoming period in the area of computing hardware components will be (energy) efficiency and supporting a broader range of workloads (convergence of HPC, Big Data, and AI).

Processor architecture has always followed the semiconductor technology evolution. In the 2000s, as transistors became smaller, the performance of processors was improved due to higher fre-quency achieved at each new technology node. The resulting power consumption was more than compensated with the reduction of the supply voltage (Dennard’s scaling). Around 2006, the end of Den-nard scaling and the resulting inability to increase clock frequencies significantly caused most processor architectures to rely on many-core as an alternative way to improve performance combined with low power design techniques to keep dynamic power consumption reasonable. Around 2016, the processors reached a power con-sumption envelope limit due to the increasing current leakage. Since it was still possible to improve the transistor density due to Moore’s law, the only way to improve the processor’s computing performance was to improve the energy efficiency of the compute node with the usage of dedicated hardware (such as GPU) optimised for a range of applications. In the coming years, the major increase of data amounts to be processed will make this heterogeneous architecture non-efficient, due to too many data transfers between heterogeneous cores. A possible solution to cope with data transfer power consumption is to put the computation in the memory. This technique is called In-Memory Computing. In addition, the Moore’s law slow-down will favour new devices such as Non-Volatile Memo-ries, which could be used for new computing paradigms e.g. neu-romorphic computing.

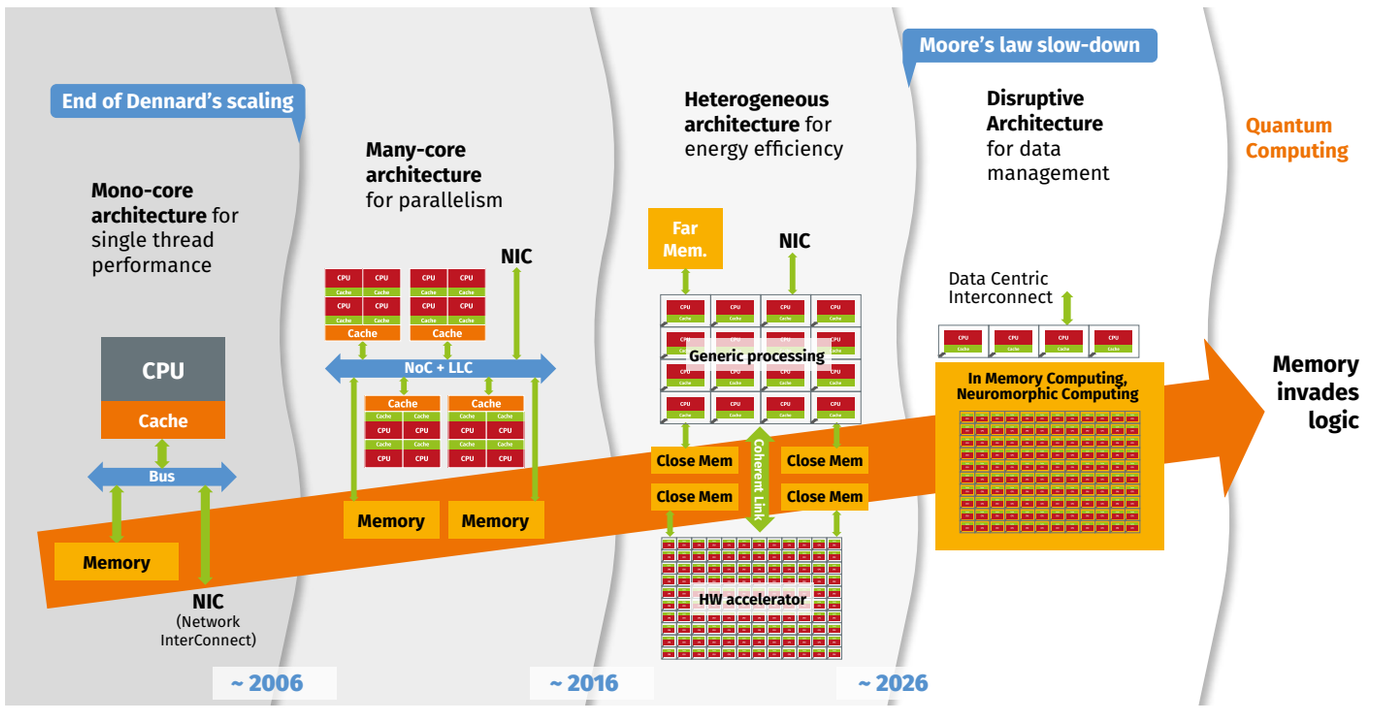


Figure 14: Processor Architecture Evolution

Convergence, i.e. systems able to support traditional HPC workloads (simulation), Big Data analytics and Deep Learning loads will be mandatory. That will have several consequences such as increasing the communication/computing ratio, moving processing near data, more modularity and composability, diversification of accelerators techniques, dealing with more diverse types of data and opening to the "continuum" will involve more security concerns for HPC.

Indeed, the peak communication/computing ratio (bytes per flop) should be improved. The objective should be to provide the maximum possible byte/flop ratio from the hardware side and (since it is difficult and expensive) modify the applications to be able to live with a lower byte/flop ratio that they ideally would like to have. Applications should also avoid moving data and keep computation near or in memory.

The move towards distributed systems and locating processing closer to where the data is generated is required to improve performance and efficiency. Communication bandwidth is up to 10 TB/s on a node, while it falls to several hundred of Gb/s at the interconnect level. That requirement will push towards architectures that have accelerators or systems where "Processing In Memory" or "Processing Near Memory" will be realised.

Modularity and composability are also important topics to be more efficient for the various workloads, which need various compute/memory/communication ratios. Dynamicity in this domain is a must but hard to achieve in an efficient way.

There will also be an increasing diversity of accelerators, from GPU to Deep Learning, graph processing and other specialised functions-oriented ones, including reconfigurable architectures based on FPGAs or "Coarse Grained Reconfigurable Architectures" (CGRA). This will also have impact on the data types: the data representations will be more diverse, from bit, to bytes, to integer, to bfloat16, to float, to double precision, and should be dynamically adapted to the current workload. For example, we need to consider completely different floating-point representations such as "POSIT" or "Universal Numbers" (Unums)[51]. In the longer term, accelerators and systems should also support new computing paradigms, where information is not necessarily binary coded, such as neuromorphic accelerators ("spike" coding) and quantum ("qbits").

AI and HPC workloads are blending together and "in network" acceleration of these workloads is gaining more momentum and will become yet another way to accelerate computing, in spite of the end of Dennard's scaling and the forthcoming end of Moore's law.

HPC systems will be more open to the outside. New uses (such as interactive accesses) will lead to new cybersecurity threats which will also deeply affect the HPC computer industry by creating a requirement for more robust computer systems. Securing various firmware components inside the platform which ensure secure boot and secure firmware updates will become a major requirement. Moreover, "encrypt everything" approaches to both system memory and network traffic are emerging.

Increasing computing performance and energy efficiency will require a complete system (hardware and software) view and will be achieved by combining several axes at the same time, i.e.:

- Increasing the number of cores per chip
- Using interposers and chiplets to increase diversity in designs, reduce costs and increase efficiency by closely coupling chiplets made of technologies optimised for each purpose
- Adding more efficient accelerators and smoothly integrating them in the programming environment
- Developing 3D stacking to have efficient interconnection between processing and memories, therefore increasing the bandwidth between processing and memory and decreasing the "distance" of the data movements
- Reducing the length of interconnection, e.g. using interposers to have processing, accelerators, memory, storage and interconnect in a single package. Dynamic power management and the use of energy efficient techniques and technologies to reduce the energy consumption.
- Increasing the (local) bandwidth and having an efficient memory hierarchy (HBM2, DRAM, NVM, etc.). Again, the goal for these new memory structures will be to reduce the data movements.
- Support new addressing schemes such as byte addressing and key-value (associative access).

In parallel to these hardware-oriented challenges, the corresponding software challenges will be to make these increases of parallelism, heterogeneity, new memory hierarchy increasingly transparent to the users and efficiently and automatically used by tools (compilers, software stack).

5.3.2.4 Intersection with Research Clusters

5.3.2.4.1 *Energy efficiency*

Energy efficiency is a major objective for system hardware components, such as computing engines (i.e. reducing the power dissipation and having a better performance/watt) and communication (i.e. reducing the energy cost of it).

5.3.2.4.2 *AI Everywhere*

AI workloads will have an impact on system hardware components in the following ways:

- For compute engines, AI workloads should support a diversity of computation objects (small matrices, "tensors") and data types (e.g. 16 bit floating point, fp16, bfloat16). Accelerators are currently developed to improve the efficiency of Deep Learning applications (both during learning stage – mainly done today on GPUs – and during inference stage). Neuromorphic accelerators are still in the research phase.

51. For the definition of Universal Numbers, see: https://en.wikipedia.org/wiki/Unum_(number_format)#POSIT

- The memory per node should be increased to reduce the communication overhead due to access to data (storage) during learning phases (Deep Learning).

5.3.2.4.3

Data everywhere

As in the case of "AI everywhere", the growing amount of data will require an increase in the ratio of communication to processing. This will have an impact on the memory hierarchy and the local storage size.

5.3.2.4.4

HPC and the Digital Continuum

The system hardware components of a large spectrum of devices should be efficient, from large HPC machines to small low power Edge devices. Interoperability is important (at least on data exchange format), together with security concerns (involving encryption and other security means at the hardware level).

5.3.2.4.5

Resilience

Resilience is a major requirement for system hardware components. Hot swap and monitoring of health of components are important features to ensure resilient operation, together with architecture features, e.g. redundancy and task migration.

5.3.2.4.6

Trustworthy computing

Security devices, such as integrity monitoring, should be encrypted (in hardware, to ensure efficiency) at least at the interface with the HPC centres. Secure boot, and other trusted solutions might also migrate in the core of HPC centres.

5.3.3

System Software and Management

5.3.3.1

Research trends and current state of the art

Power efficiency, scalability, and heterogeneity support still drive the system software landscape at Exascale for 2025-2027 but the large scope of new applications forces us to rethink the architectural solutions in order to meet those challenges. System software solutions need to:

- Support applications diversity and an expanding computing scope, in particular meeting the requirements of the convergence of Simulation (HPC), Big Data (HPDA) and AI in the same IT continuum.
- Master the complexity of optimised or specialised hardware and software combinations, with environments and tools supporting dynamic and flexible execution models.
- Offer smart tools to assist in the development, optimisation and control of the efficiency of application workflows over heterogeneous hardware architectures.

Simulations often require large amount of computations, so they are often run on a general-applicability HPC infrastructure, built as a cluster of powerful high-end machines, interlinked with high-bandwidth low-latency networks. The compute cluster is commonly augmented with hardware accelerators (co-processors, GPUs or FPGAs) and a large-capacity and fast parallel file system. All equipment and services are set up and tuned by systems administrators. In big-data analytics, the focus is rather on the storage and access to data and data processing is often performed on Big Data infrastructure customised for the problem at hand. Those infrastructures offer specific data stores and are often installed in a more or less self-service way on a public or private Cloud, typically built on top of commodity hardware. It is our understanding that the HPC world realises that there is more to data storage than just files and that self-service ideas i.e. minimising system administrator intervention for resource provisioning (e.g. by offering advanced policy-driven resource provisioning functionality through a HPC gateway/portal) are attractive to users. In the meantime, the big-data world realises that co-processors and fast networks can really speed up analytics. All Cloud providers now offer HPC services and currently HPC centres are looking to add Cloud-based technologies to their offerings. In this setting, we are considering convergence of HPC/HPDA/AI workloads as a major milestone in the evolution of system software infrastructure and tools.

5.3.3.2

Challenges for 2021-2024

5.3.3.2.1

Convergence of Simulation (HPC), Big Data (HPDA) and AI in the same IT continuum

Converged HPC/HPDA/AI workloads have different characteristics from pure HPC workloads and therefore demand additional features on the systems they run on. With compute intensive workloads currently running on HPC clusters, primary concerns are close-to-the-metal performance and efficient use of high-end/dedicated hardware in an environment where users are granted exclusive, albeit time-limited, access to resources. With data intensive (Big Data, AI) workloads, currently running mostly on Cloud systems, primary concerns are the instant and elastic availability of resources and fault tolerance, in a multi-tenant environment and with sufficient flexibility to select between a "self-service" operating mode or rely on a ready-made software stack. Converged systems should be able to cope with wide-ranging workload diversity and AI-inspired solutions could be used for efficiently managing the complexity introduced. It is important to be able to reconfigure the data centre dynamically: elastic reconfiguration and efficient scheduling are potential solutions. New technologies such as AI methods and AI-optimised hardware or commodity-device observations and IoT data streaming offer new potential for scientific methodologies and workflows. Some of the toughest technical challenges will depend on understanding and modelling the data and workflows in the underlying multi-owner, and multi-tenant IT infrastructure. The data and computing continuum is a disruption for present application development putting a major emphasis on a data-aware execution flow and security across the full application workflow.

At the architecture and system levels, computing together with data concerns will drive supercomputer design. **In memory computing** integrated at different locations of the architecture will

have a great impact. **Kernel-bypass** networking is a promising technique to address the throughput and system stack scalability issues but we still lack standard interfaces and protocol implementations for applications in order to take advantage of them**. Programmable packet processing** is an emerging technique for applications that can offload their request processing in part or in full to the NIC. OS disaggregation will offer new capabilities for applications and runtime-specific optimisations.

Even if the operating system and hardware and smart I/O services are expected to enhance data access and data sharing, major efficiency improvements will depend on application redesign to minimise data movement over the multiple steps of a computation or the multiple steps of the workflow. The real challenge of convergence appears to lie in integrating flexibility with heterogeneity. System architects and application programmers need to re-think the way that information is accessed, shared and stored. A more flexible science-technology co-design flow with fast turn-around of innovation at the interface between science applications, engineering and computational science is clearly needed. Adaptation can be very intrusive, influencing the way in which applications are designed and deviating from the classic first principles-driven science and linear-workflow approach.

Support capabilities such as **workflow and dataflow deployment and orchestration, data location** and logistics and dynamic resource allocation (compute, network, storage) are complex issues that will require progressive enhancements in order to address major challenges such as security, efficiency, programmability and reproducibility.

5.3.3.2.2

Efficiency of the combination of IT infrastructure and applications execution environments

To minimise power consumption, the hardware will be increasingly heterogeneous, and it will use processors for computation and orchestration of the dataflows and diverse accelerators, such as FPGAs and GPUs or their derivatives for Deep Learning. At the global level, electricity power sizing and its associated cost ultimately limit the size of the machine: higher energy efficiency allows a more powerful system for the same cost. There are several options to increase the efficiency of the machine and in practice they should be combined:

- "Adequate/ appropriate" computing - The idea is to adapt the accuracy of the operation to the needs. For example, the learning process in Deep Learning does not really need double precision floating point operations and GPUs are directly supporting half-precision (float16), which is enough while decreasing the size and energy required. Some operations do not even need to be exact, so operators can be simplified while being "good enough" for the requirements. On the other hand, floating point representation can induce errors in iterative computing and new formats (e.g. UNUM) can help in solving effects such as numerical instability.
- Application specific hardware is more efficient in terms of FLOPS/Watt or Ops/Watt than the general-purpose option because computing resources are tuned to the application class and their control is more efficient. For example, in terms of throughput, GPUs are more efficient than general purpose processors, yet their compute capabilities are more limited (SIMD instead of MIMD execution) and, as a result, programming can be more difficult in the general case. Reprogramming capabilities or dedicated software optimisation will need to be designed. It will be a challenge not only from the programmer's point of view but also from the point of view of system managers to combine different accelerators into a **unified programming model** supported by a dynamic and elastic resource management infrastructure.
- In situ/in transit processing - In situ processing is a more efficient alternative, allowing data visualisation, curation, structuring or analysis to happen online as data is generated by the simulations, thus reducing the volume of refined data to be stored and in consequence saving energy. Big Data management approaches include in situ processing capabilities that are of particular interest for addressing this challenge, i.e. by bringing the computation to where data is located.

5.3.3.2.3

Availability of tools for dynamic and flexible execution models

New software stack integration and compliance capabilities will be necessary to support applications portability over heterogeneous infrastructures. Software-defined infrastructure solutions offer advanced policy-driven data and resource management capabilities for managing storage and compute resources respectively. Regardless of whether the computing is HPC or Big Data, launching jobs with high resource requirements will require efficient support to reduce job launch latency, monitor job progress and resource consumption in real time and handle runtime node and other failures.

Matching hardware resource capabilities with applications-oriented environments is a great challenge which requires the evolution of multiple tools. Even if virtualisation abstractions help to deploy applications over multiple architectures, significant tools evolution will be required to cover variability of applications and hardware resources. Application development for this level of complex architecture will require new programming APIs, new run time combinations and tools that offer an abstraction layer which hides a part of this complexity and guarantee application portability. Moreover, we need tools to analyse, profile, trace and predict efficiency of flexible execution environments. Embedded AI and analytics methods will be helpful to master the complexity of development and deployment of that new style of applications.

Efficient integration of virtualisation or container approaches would improve the ease of use, efficiency and resilience of systems. To allow arbitration between different users and applications in the current resource management tools, some features will need to be re-thought in terms of the global workflow: the

allocation rules, data provisioning and dataflow management policy or engine. This will raise new challenges in terms of resiliency, security and reproducibility of simulation.

Reproducibility will be a major challenge in the next decade. Integration in applications workflow of capabilities to capture contextual information and provenance information during application execution with limited scalability or efficiency impact is therefore an important topic. In order to enable the reproducibility, the use of virtualisation will be again essential to provide a portable environment where experiments can be controlled and mimicked.

Research should target mechanisms for adaptive and dynamic scheduling, management and use of heterogeneous system components to achieve energy efficiency and resilience, while meeting application performance requirements. New practices such as HPC as a Service will be impacting the management solutions of the Exascale supercomputers. Supercomputers must become accessible from the Cloud and be compliant with the Cloud in terms of system management criteria and practices, while still keeping high performance and scalability as major objectives. Significant challenges lie in the coordination of orchestration of applications workflows, resource management and data management cycle by using the combination of HPC and the Cloud. The development of some kind of meta-orchestration approach, which manages HPC and Cloud orchestrators and seamlessly enables the use of both HPC and Cloud resources, will pave the way to get the required compliance.

Finally, it is worth stressing that there are still many differences in terms of tools, protocols and philosophy of usage between the HPC and Cloud communities. This creates a gap which will have to be bridged for the desired convergence between both disciplines to become a reality.

5.3.3.3

Intersection with Research Clusters

5.3.3.3.1

Development methods & standards

With an increasing evolution of HPC/HPDA/AI infrastructure towards more dynamic and elastic resource provisioning and management, infrastructure owners/operators are increasingly expected to pay attention to alignment with development tools and standards. Examples include:

- System software low-level APIs for task offload to accelerators, over emerging system interconnects (e.g. CCIX)
- Applications model (e.g. microservices, workflows)
- Tools to monitor and predict application behaviour/performance on different architectures and system configurations
- Applications-centric containers
- DSL/runtime support for domain-specific optimisation
- New generic APIs over different runtimes.

5.3.3.3.2

Energy efficiency

Efficient and timely metrics collection and low-level resource monitoring APIs will continue to be crucial for overall infrastructure effectiveness and efficiency. Examples include:

- Augmenting job-level accounting with profiling of power consumption
- Power-aware job resource allocation via extensions to workload managers
- Power control and power saving for increasingly heterogeneous resources.

5.3.3.3.3

AI everywhere

AI integration in HPC/HPDA application workflows (including specific libraries, tensor data types, data ingestion, visualisation, continuum processing) increases the pressure for flexible and effective support for integration of diverse platform capabilities in application workflows. Examples include:

- System software integration packages
- System software mixed precision support
- Workflow and orchestration capabilities support
- Integration of workflow control with resource management
- Front-end persistent integration services (e.g. Spark, TensorFlow).

5.3.3.3.4

Data everywhere

With ever increasing emphasis on data processing in converged HPC/HPDA/AI infrastructures, there is a pressing need to improve data streaming support at the network and OS levels. Examples include:

- Data and workflow coordination, taking into consideration the data lifecycle
- Security features (e.g. encrypted datasets, compliance with prescribed assurance levels)
- Memory/Storage sharing over network links (e.g. via NVMeoF)
- In situ processing /in memory processing.

5.3.3.3.5

HPC and the Digital Continuum

With the emerging integration of HPC systems into a Digital Continuum, infrastructure providers need to improve support for data-driven applications, particularly for data coming from the Edge. Example areas include:

- Front-end services and related orchestration
- Increasingly critical dependence on reliable connectivity
- Hierarchical resource allocation and management to determine where to deploy the individual parts of application systems
- Security features (e.g. isolation, privacy preservation).

5.3.3.3.6

Resilience

An increased focus to resilience necessitates more effective interaction with the runtime system to improve robustness as experienced by application owners. Examples include:

- Exposure of low-level fault and recovery events (e.g. ECC) and error containment actions.
- Visibility to platform-level event handling (e.g. PMIx)
- Checkpoint/restart, with application/runtime guidance, to account for increasingly heterogeneous platforms and transient system states.
- Dynamic routing and congestion control, to bypass network regions exhibiting increased contention or high probability of data loss.

5.3.3.3.7

Trustworthy Computing

HPC systems and particularly converged HPC/HPDA/AI systems will need to provide system software support for advanced security mechanisms to satisfy increasingly pressing requirements for trustworthy processing of sensitive datasets. This support includes assistance from OS for putting in place several essential security mechanisms and to set up secure execution environments (e.g. containers). Examples of mechanisms needed include:

- Mechanisms to support appropriate handling of sensitive data sets – particularly, Personally Identifiable Information (PII) – e.g. encryption and additional access controls to prevent loss of confidentiality
- Protection against loss of integrity for sensitive data sets
- Protection against loss of confidentiality with respect to intellectual property.

5.3.4

Programming Environment

5.3.4.1

Research trends and current state of the art

The overriding theme of this area is to effectively support the productive development of scalable, efficient and effective high-performance applications across the extended HPC landscape (the Digital Continuum, from Edge-level, embedded HPC to HPC clouds and Exascale computing and including HPC-in-the-loop). Expanding on that theme results in the following strategic research directions:

- Enabling effective application development and deployment at extreme scales requires high-productivity and performance-oriented programming environments. This is also true when targeting the expansion of HPC workloads to applications from Cloud, analytics or AI arenas while maintaining high efficiency.
- The adaptation of existing programming models (or the creation of new ones) and of their associated runtime systems and compilers requires co-design and efficient interaction with aspects covered by the other SRA domains, such as the system software, scalable underlying data I/O, storage and processing. These programming models should fit the needs of highly relevant applications.
- There should be strong interoperability throughout the programming environment across the Digital Continuum, including compiler tools and runtime systems, debuggers and performance tools, linking information from these tools to the programming model and source code.

An aspect of key importance is to establish the acceptance and adoption of the programming models and application programming interfaces (APIs) in industrial and scientific production codes. This requires an emphasis on long-term reliable and robust support of the programming models, APIs and the related runtime system functions and tools, including via formal and de facto standardisation.

Modelling and simulation (well established in industrial and scientific computing) benefit from a relatively long past history of HPC use. Big Data Analytics ("BDA", which includes learning-based analytics) and Artificial Intelligence (AI, in particular the Deep Learning variant of Machine Learning) exhibit a trend for very rapid development and deployment of programming frameworks and associated languages, tools and software systems. A convergence of HPC and BDA/AI offers great opportunities and there is a significant potential in the identification of commonalities in the software stacks (e.g. common storage and compute abstractions) and in the possible provision of "cross-area" programming environment components (e.g. unified data processing techniques, common models for tensors). While the definition of unified APIs may appear as a long-term goal, to address this convergence for a shorter term at the programming level, possible approaches include composability and the definition of interoperability-oriented APIs.

Dynamic workflow management systems adapted for high-performance execution are required in order to support the development of complex workflows of applications, including those expected to run across the HPC-related Digital Continuum, in particular to support "HPC in the loop" scenarios, in situ data analysis and visualisation. Such systems will enable the coupling of simulation, databases and data streams, data analytics and visualisation that interact together in real time. For instance, results from intermediate data analysis steps performed while the simulation is running should be able to trigger detailed (or refined) further simulation steps. In other scenarios, to improve the quality of decision making based on data analytics, on-demand HPC simulations or AI predictions can be necessary. Such scenarios require support for dynamic resource management of hybrid workflows combining simulations, analytics and AI/ML training or inference, ultimately across the Digital Continuum.

5.3.4.2

Challenges for 2021-2024

As discussed above, the use of HPC technologies is evolving from dedicated data centres to encompass the Digital Continuum, spanning Edge-to-Fog-to-Cloud/HPC centre deployment and workflows involving modelling and simulation, data analytics, and machine-learning/AI components. The HPC programming environment necessarily needs to follow that evolution. Nevertheless, the key programming environment challenges presented in the previous ETP4HPC SRA (reflected in the following four subsections) essentially retain their relevance and importance; those key challenges permeate through the digital computing continuum.

Irrespective of the target deployment of an application in this new HPC scope, there is a crucial need to support the evolution of HPC applications by providing high-productivity and performance-oriented programming environments. Improved productivity for application developers can be addressed by the reduction of programming complexity through advancements throughout the programming model and system software stack. An approach is to explore the convergence of the (different) programming models and languages traditionally used by the areas of HPC, Big Data analytics and ML/AI. Potential approaches for this include increased intelligence throughout the programming environment and higher-level abstractions allowing separation of core algorithmic issues from implementation and optimisation concerns.

While scalability is an attribute that is obviously needed for Exascale computing, it is actually a technology attribute required across the whole Digital Continuum. It requires that the development of the programming environment be carried out in a co-design activity with the developers of the computing systems and Digital Continuum infrastructure, together with applications of high relevance for the societal challenges driving the technology developments covered within the SRA.

Energy efficiency and the support for resilience are two technology characteristics, represented by specific research clusters in the SRA, which require close collaboration with the system software level in order that the application programmer can realise optimal deployment of production codes. The successful adoption by industrial and scientific production codes necessitates the establishment of formal or de-facto standardisation (including interoperability and composability across the communities and technical areas arising in the foreseen Digital Continuum).

5.3.4.2.1

Innovative and higher-productivity parallel programming

This topic covers approaches targeting increased productivity of application development, including legacy codes, through complexity reduction and includes in particular research into application/domain frameworks as well as dynamic workflow systems.

The separation of algorithmic expression vs. implementation concerns is a key approach to reduce programming complexity for the application developer. The aim is to provide an abstraction of the underlying computational algorithms (which are typically hardware-neutral) separated from the actual data structures and parallel programming/runtime system implementation (which must be adapted to a specific target hardware system). An intrinsic aspect of the high-productivity programming approach is that specific hardware (and related runtime software) features such as accelerators and near-memory/near-storage processing would be supported in a way that is transparent to the application developer. Approaches include the use of meta-programming, high-productivity languages (such as Python) and domain-specific languages (DSLs), particularly those built upon a general-purpose framework for new applications. Similarly, the approach should facilitate the use of auto- and self-tuning libraries by applications (especially for legacy applications).

At the workflow level, there is a need for application-independent dynamic workflow systems, adapted for high-performance execution, that enable the integration of simulation and modelling with data analytics and AI. Such workflows are expected to be composed of HPC simulations, data analytics (at the input, interleaved with computation, or at the output), AI/ML training or inference steps, and visualisation and output to persistent storage/databases or to data streams. Existing workflow models and environments from Cloud and general data centre computing were not developed with effective and efficient support for parallel, HPC-style applications and supercomputers. It will be important to find ways of integrating such applications and systems with existing, commonly accepted ways to define and run workflows. The handling of persistent objects, potentially distributed across storage and computational systems, is needed to support the future modes of use of application workflows, e.g. object re-use across “work sessions”.

Potential approaches for high-productivity and performance-oriented programming environments include increased intelligence based on learning throughout the programming environment and its underlying runtime and system software, to support smart and efficient resource usage.

5.3.4.2.2

Effective interaction with the run-time system

The effective interaction with the run-time system requires use of appropriate APIs by the applications (or high-level application environments) in order to transfer information (application metadata) between the application and the computing system in order to realise the computational schemes that best exploit the system. For the latter, key aspects include data layout, data movement, dynamic load balancing, resilience and the ability to dynamically adapt to, and request, changing resources and application needs, thus including pro-active resilience methods. For the targeted distributed, heterogeneous compute and data systems, effective workload forecasting, and scheduling of the dynamic workloads will be essential.

The design of suitable abstractions at the application level and programming model development to support those abstractions is needed in order to enable the runtime system to realise optimi-

sations. The information transfer (from API through to the runtime system) needs to support flexible runtime hierarchies occurring in dynamic and heterogeneous systems, supported by malleable resource management approaches and systems for storage and computation. In addition, favouring the collection and processing of provenance metadata through the API exposed to the users should facilitate effective interaction with the runtime system.

5.3.4.2.3

Interoperability, composability and standardisation

Efforts made to achieve interoperability between programming models and resulting frameworks or environments should be continued with a particular emphasis on establishing migration paths for legacy codes in established industrial and scientific HPC application communities. The related support from performance analysis and debugging tools is required; the tools must understand the programming model abstractions, for example, but not limited to, mapping performance/correctness issues to the original source code and observing, manipulating and debugging tasks or parallel loops.

To efficiently integrate simulations and data analytics, ensuring a high interoperability for data processing is an important step, which should ultimately lead to the definition of unified APIs for managing data globally across the continuum. Such unified APIs should facilitate the design and implementation of extremely scalable data processing architectures combining traditional Big Data processing (batch- and stream-based) with HPC-inspired data processing (in situ, in transit). Data access raises similar interoperability challenges, as data models and data access APIs are currently very heterogeneous (byte-level, object-level, structured data APIs, etc.). As stream processing gains momentum, byte level access to storage is needed increasingly to support manipulating data items with fine granularity. In addition, AI/ML training or inference steps will need to be integrated.

Composability is the ability to use multiple programming models for a single application with defined rules. Single applications could then combine the use of different robust programming models to enhance usability and achieved efficiency. Where composability involves multiple "components" (including the runtime system), they must cooperate among themselves and with the system software to efficiently exploit the shared physical resources. Composability between programming models/languages and higher-order frameworks is a particular challenge, as is the ability to handle applications aimed for deployment across the Edge–Fog–Cloud–data centre continuum. One relevant aspect for composability and interoperability across such a large Digital Continuum regards the semantics and the management level of data consistency. For instance, data consistency is often managed at the storage level on Cloud-based Big Data storage systems but substantial improvements in performance and energy efficiency are available if data consistency is exposed at the programming level in HPC systems. An important challenge is to reconcile these aspects as HPC becomes a piece in a larger Digital Continuum.

While existing standards need to be upheld by new developments, filtering innovations into those standards, new standards (formal or de facto) addressing in particular composability and the broadened HPC targets are important to ensure take-up by industrial and scientific production applications.

Finally, complex dynamic workflows combining HPC simulations and analytics are expected to be deployed in a possibly broad hybrid environment across the Digital Continuum (including Edge/Fog devices, clouds and supercomputers or a subset of them). In this new context, interoperability (including storage abstractions and processing techniques), composability and standardisation become critical for the design of programming frameworks and of their supporting tools for data storage and processing, computation and analytics across such hybrid infrastructures. This includes unified real-time data processing techniques favouring the joint use of HPC-originated approaches, such as in situ/in transit processing, with stream-based processing techniques now common in Big Data analytics frameworks. In addition, security and data privacy concerns will now be posed in a very wide distributed environment, including the necessity to comply with laws and regulations.

5.3.2.4.4

Performance analytics, debugging & program correctness

With the evolution of computing systems expected to encompass the whole continuum from embedded HPC to HPC clouds and Exascale computing, it is clear that the amount of data produced will be enormous, while only few useful tool sets are available to make sense of the data, scalability of the tools thus being a specific concern. While significant progress has been made in the research area of automated data centre monitoring/surveillance/maintenance techniques, there is a need to develop things further for the HPC context (and for HPC within the Digital Continuum). Other performance analysis themes that have been concerns in the past have increased importance in the context of extreme scale computing and dynamic application workflows deployed in the Digital Continuum:

- Intelligent performance tools including energy analysis
- Mapping of information to the (multiple layers of) source codes, e.g. original application source code in a high-level language, numerical libraries, and runtime systems.
- Integrated and user-friendly tools allowing users to collect, analyse and trace the information for both system- and application-levels.
- Provision of information relating data access in the application source code to run-time data layout and transfers.

Moreover, debugger technology is needed which can support applications that have been developed on and for dynamic, heterogeneous computing systems, using both current and non-conventional programming models, languages and APIs, and deployed on the full range of target systems within the digital computing continuum.

In addition, open challenges remain for program correctness. Current compilers and runtimes already perform many analyses and checks to warn application users about (potential) issues on sequential applications. This kind of support should be extended and enhanced to also cover issues related to the parallel/distributed execution of an application (e.g. data-race detection in parallel applications).

5.3.4.3

Intersection with Research Clusters

5.3.4.3.1

Development methods and standards

- Research Priority: Interoperability, Composability and standardisation
  - Programming model/API standardisation addressing performance portability, composability and interoperability.
- Research Priority: Performance analytics, debugging and program correctness
  - The use of advanced and intelligent performance analysis and debugging tools are integral to the software development process.

5.3.4.3.2

Energy efficiency

- Research Priority: Effective Interaction with the runtime system
  - The programming environment needs bi-directional information transfer and the ability to both demand and adapt to dynamically changing situations regarding energy consumption by applications. One specific aspect would be the identification and subsequent optimisation of bad or abnormal energy consumption.
- Research Priority: Performance Analytics
  - The need for intelligent performance tools also extends to the area of energy/power consumption by applications.

5.3.4.3.3

AI everywhere

- Research Priority: Innovative & higher-productivity parallel programming
  - The high-productivity environment also addresses AI (or AI-oriented) applications and incorporates many generic features, including the transparency of hardware system for the application developer and support for DSLs (built on general-purpose frameworks).
  - Dynamic workflow support also would be applicable to AI applications, in particular through the provision of an end-to-end coordination layer that supports streaming inputs and outputs.
- Research Priority: Effective interaction with the run-time system
  - The target to enable performance optimisation via close interaction with the run-time system is equally applicable to modelling, simulation and AI applications.
- Research priority: Interoperability, Composability and standardisation
  - This priority targets both complex, dynamic workflows as well as the composability needed to support applications aimed for deployment across the Edge–Fog–Cloud–data centre continuum and thus supports multiple objectives of this cluster.

5.3.4.3.4

Data everywhere

- Research priority: Interoperability, Composability and standardisation
  - Unified APIs in programming environments: Currently the data generated across the network are processed through dedicated, specific APIs and programming models (e.g. MPI for HPC, MapReduce / Scala in the Cloud, Edge). There is a lack of unified APIs able to deal with global data. These unified APIs would allow to efficiently integrate simulations and data analytics through extremely-scalable data processing architecture combining traditional Big Data processing (batch- and stream-based) with HPC-inspired data processing (in situ, in transit).
  - Incorporating the Digital Continuum: In the new Digital Continuum, data generates models and models generate data. The programming environments should move from the unidirectional approach to be able to support this continuous loop of data and model updates. This ultimately enables a better understanding of a system.
  - Data consistency: the challenge is to reconcile strong ACID consistency with weaker consistency models that are used conversely across the network (from the Edge to the Cloud and HPC, according to the processing place of data). Overall, weaker consistency potentially speeds up the processing.
  - Transparency in data addressing: The data generation and processing continuum raises the challenge of reconciling different architectures with respect to data addressing: some architectures support accessing individual bytes of data rather than only larger units (i.e., blocks), typically used by the object storage backends in the Cloud, for instance. With the advent of stream processing, however, byte level access is more needed in order to gain access to data items with a fine granularity.
- Research Priority: Effective interaction with the run-time system
  - Provenance data: since data can be sourced at any place across the network (Edge, Fog, HPC, Cloud) it becomes crucial to be able to collect and exploit provenance data directly via the programming models exposed to users.

5.3.4.3.5

HPC and the Digital Continuum

- Research priority: Innovative & higher-productivity parallel programming
  - There is a need of application-independent dynamic workflow systems that enable the integration of HPC simulation and modelling with data analytics.
- Research priority: Interoperability, Composability and standardisation
  - Composability between programming models/languages and higher-order frameworks is a particular challenge, as is the ability to handle applications aimed for deployment across the Edge–Fog–Cloud–data centre continuum.
  - In the context of complex workflows deployed across the Digital Continuum, interoperability, composability and standardisation become critical for the design of programming frameworks and of their supporting tools.
- Research priority: Effective interaction with the runtime system
  - Abstractions at the application level (e.g. for unified data storage abstractions and efficient data sharing across the Digital Continuum) require the support (realisation) by the programming environment and the interaction with the runtime system.

5.3.4.3.6

Resilience

- Research Priority - Effective Interaction with the runtime system
  - The programming environment needs bi-directional information transfer and the ability to both demand and adapt to dynamically changing resources, thus including pro-active resilience methods, is foreseen as one of the outcomes for new APIs to the system software/runtime system stack. Similarly, information on the system could guide application checkpointing.
  - Ability for storage-mapping of error-tolerant data sources/outputs would align with resilience-energy optimisation aims.
  - In addition to the telemetry efforts to perform automatic data analytics on the I/O system state, interfaces should be designed to interact between the runtime and telemetry and data analytics system. This would allow for resilience software to use this information and autotune their techniques to optimise and increase efficiency. For instance, checkpointing libraries could regularly monitor the condition of the file system and automatically reschedule checkpoints to less congested periods, according to the data gathered and analysed by the system.

5.3.4.3.7

Trustworthy computing

- New programming paradigms could favour trustworthiness by leveraging formal methods.
- Trustworthy programming environments require a high degree of fault tolerance to support recovery from failures.

5.3.5

I/O and Storage

5.3.5.1

Research trends and current state of the art

The convergence of HPC, Big Data, and AI is significantly increasing the demands, in terms of diversity of workloads and complexity of access patterns, from HPC storage systems, while the still growing number of cores per processor, and the increasing number of accelerators, are further widening the gap between compute and storage performance. The problem is further exacerbated by the heterogeneity available in the compute nodes, with the continuously increasing computing power of accelerators and co-processors, which further increases the potential for stressing the I/O subsystem. These do not show any signs of abating since the time the previous SRA was written.

While the previous SRA focused on the convergence between HPC and Big Data, AI (and Deep/Machine Learning) are now also starting to become a very integral part of modern HPC solutions. Traditional HPC storage has been optimised to serve small to medium input sizes and huge outputs which occur in short bursts. Big data applications, in contrast, are typically constantly reading huge data streams, while also producing big outputs when shuffling intermediate results, while AI programs additionally require random accesses, especially during the training phase. Detailed I/O requirements for many different classes of AI applications confound storage and I/O architects, especially when designing solutions for scale. Building unified, robust storage frameworks able to efficiently serve all three scenarios will be one of the main tasks within the time frame of SRA 4.

The first step to incorporate the new usage scenarios is to further deepen the storage hierarchy. Flash/NAND-based SSDs have become an important component of modern HPC clusters, either as dedicated and external burst buffers or as node-internal storage. Flash storage is faster than magnetic disks (HDD) and has also much better random-access properties. The last couple of SRAs looked at the evolution of Non-Volatile Memories beyond Flash. Non-volatile main memory (NVM) is now finally becoming available commercially and will help to build another hierarchy level even closer to the CPUs, which is significantly faster than flash, while also being persistent, byte-addressable, and less expensive than DRAM. It will be important to integrate NVM into the storage stack by building deeper storage hierarchy frameworks and by developing faster software storage stacks.

Application developers and scientists fear that a deeper storage hierarchy also increases the complexity to use it. It is therefore important to provide semi-automatic mechanisms to move da-

ta between the different storage hierarchy levels, while at the same time offering a unified namespace. More of these trends have become visible since the time of the previous SRA with more innovation in data management techniques and programming models that aid in providing such unified name spaces, strengthening the "Memory/Data Centric Computing" trend which was appearing on the horizon when the previous SRA was drafted. The first steps in this direction have been performed by developing ad hoc file systems which can be deployed during the runtime of applications, while widespread parallel file systems such as Lustre, GPFS and BeeGFS now also extend the caching layer to node-local storage. These schemes have to be coupled with the underlying resource manager and batch scheduler to place programs closer to their data (in situ processing) as well as with Quality of Service (QoS) mechanisms to ensure that huge I/O streams do not interfere with each other. Indeed, in situ processing becomes even more important in highly distributed storage infrastructures explained below.

With the proliferation of the Internet of Things (IoT), coupled with the need for AI/Deep Learning for taking decisions in near-real time, the system infrastructure assumptions continue to change. The "full" HPC system now consists of the HPC data centre, the intermediate caching "Fog" nodes and the storage infrastructure very close to the devices themselves ("Edge"). The previous SRA mentioned that such a change was happening, but with the increased number of use cases exemplified by autonomous driving, this is starting to become a reality. The software infrastructure to store and manage data in the workflows that get deployed on such systems are starting to come into focus. Adaptations required in object storage software and parallel file system software to holistically cater to these workflows is becoming increasingly relevant, so are the data management frameworks to federate data across these distributed infrastructures. For example, new I/O APIs such as ADIOS are evolving (with its own internal file format) to provide more precise control of application data pipelines while leveraging existing underlying parallel file systems. There is also a need to rethink aspects of storage system resiliency (and telemetry analytics) and the need for new storage and I/O benchmarks in such a regime - considering that both the use cases as well as the data infrastructure needs are evolving. Indeed, these questions continue to be of relevance just for large central HPC data centres themselves as they scale up towards Exascale.

5.3.5.2

Challenges for 2021 - 2024

5.3.5.2.1

Non-Volatile Main Memories in the I/O Stack

The emergence of fast storage devices such as the NVMe protocol (block addressable, PCIe-based SSDs) and NVM (byte addressable, memory-connected) devices creates enormous opportunities and challenges. These technologies reduce access latencies to microsecond-level scales and can scale throughput to memory-interconnect levels. Such devices are projected to have two uses: either as persistent storage to improve I/O performance or as persistent memory to allow applications to keep larger datasets closer and cheaper to the processor. Both directions help future systems to keep up with the required data capacities, transfer rates, IOPS and latencies, especially given the continuous growth of dataset sizes and complexity.

In addition, exploring the convergence of persistent memory and persistent I/O where data does not need to change representation when moving between memory and storage and even reducing data movement can open new pathways for improving data processing efficiency. However, the storage stack is far from delivering the required performance levels to applications and enabling new uses, such as extending the application runtime heap beyond DRAM and over NVM. Adapting HPC storage to these new device technologies will therefore require tackling the following challenges.

Reduce the costs from the user to the device

The current I/O stack in Linux has been optimised for relatively slow devices. New storage technologies are pushing this stack above its limits and I/O stack latencies become an impending point to obtain their peak performance. One of the reasons for that is that the current I/O path needs to go through the system call interface and many kernel functions on its common path.

The clearest example is the FUSE interface that is even insufficient for newer PCIe SSD devices. Several solutions have been proposed to change this interface but they need further investigation and also have to be moved to the mainstream. These approaches either try to reduce the number of kernel context switches or move I/O processes from the kernel to the user space. Approaches that allow faster access to storage devices by reducing the processing overhead are important as they will allow future systems to process more data with the same energy budget.

Moving control to user space to improve access performance nevertheless complicates the provision of security guarantees. The first demonstrations to date to combine data protection with access speed require severe modifications to the underlying operating system and/or hardware. HPC on the other hand (partially) assumes that nodes are accessed by a single application under a single authority so that relaxed security guarantees might be sufficient in certain scenarios and should be investigated.

Memory style addressing of persistent storage

Non-volatile main memories (and in general Storage Class Memory (SCM) technologies) incorporated as part of the I/O stack will expand the memory addressing capabilities of applications. This raises the prospect of unified memory style access to all persistent storage resources as data in the lower tiers can be mapped to the higher NVM tiers.

Nevertheless, blurring the boundaries between memory and storage has an impact on how HPC applications have to be designed. Traditional HPC applications write their data in phases and data is always consistent at the end of an I/O phase. Memory style addressing requires that data structures are valid at the same time in their memory and in their storage representation to be able to overcome the (de-)serialisation overhead of traditional

storage, while writing to memory-like storage will most probably also require supporting transactional-like updates from the applications programs.

5.3.5.2.2

Hierarchical storage and middleware environments

New storage technologies including burst buffers, node-local SSDs, NVM, SCM and various types of emerging disk drive technologies have made the storage hierarchy deeper, while new HPC applications from the fields of IoT, AI, and Big Data also made it much more complex. It is therefore necessary to support end-users and remove the burden of manually staging data between the different storage layers from them, and help them use the new storage technologies more efficiently.

Ad hoc file systems

Burst buffers and node-local storage can provide temporary storage space that caches metadata and data during the execution of a parallel job, workflow or even for simulation campaigns. ad hoc file systems providing distributed and unified namespaces can be tailor suited to these jobs' requirements, while they do not have to be universal like today's parallel file systems. Examples developed during the SRA period show that the metadata performance of ad hoc file systems can improve by many orders of magnitude compared to the state of the art with such approaches.

It is now important to include such file systems in standard HPC workflows in order to adapt them to new usage domains by extending their functionality and increasing their reliability and to co-design them with HPC/HPDA applications. Relaxed semantics, as being used in the Cloud context, have to be jointly investigated with application developers, e.g. to understand whether difficult-to-scale commands such as rename or move are really required during HPC application runs.

Middleware to provide Quality of Service and ML-interfaces

Data centric computing requires that data is delivered according to Service Level Agreements (SLAs) at predefined I/O rates. In the past, HPC only delivered predictable compute performance by exclusively assigning nodes to individual jobs, while the storage resources were shared between all jobs running on a cluster.

Basic building blocks to extend Quality of Service (QoS) to HPC storage have now been developed by the European and international HPC community. It now becomes necessary to connect the individual components including the resource manager and batch scheduler, the parallel backend file system and the ad hoc file systems, while also securely interfacing machine learning frameworks. A better coordination between the storage system and the scheduler including I/O-aware scheduling ensures less contention in the I/O stack, resulting in improved job runtimes. It becomes necessary to provide a control plane that allows both schedulers and applications to define their storage requirements, e.g. during stage-in or before a checkpoint. This control plane's API should be able to define IOPS requirements and reliability demands, while it should not be too fine-grained to not produce too much metadata overhead.

Automatic data placement

New storage hierarchies, starting at tape archives, including the parallel file system, burst buffers, and node-local storage as well as NVM require that the right data is at the right place to be used efficiently. Straining users with placing data at the right level inside these deeper and more complex storage hierarchies would reduce their productivity. The increase of hardware complexity should be addressed by broader middleware or software agents being able to shift data from one level to the other or to predict the access pattern in order to tolerate access latency. Depending on the storage semantic (block, file or object) these software agents will analyse, classify and anticipate I/O requests. Such improvements will shield end-users from complexity and improve performance and portability of applications, hence protecting early-adopters' software investment against future hardware evolution. Early work has appeared in this area since the previous SRA significantly more work is required.

In-storage processing and serverless computing

Movement of data continues to be a huge energy and performance challenge. This was introduced in the previous SRA and there has been a start of activity in this area. Including processing power in the storage path can reduce the limitations imposed by I/O bandwidth or latency, including unnecessary data movements. In the last few years, some hardware has presented solutions that allow including key-value databases within the storage devices. This idea could be explored further and in several layers: from storage devices, co-attached FPGA or processors or via software using data movements to other layers to process the data and transform it, e.g. via function shipping and serverless computing. For these options to work, research in protocols, algorithms and data format definitions is needed.

Reducing data movements even becomes more important for object-stores, especially when the objects stored are big. Solving this problem requires extending the capabilities of object stores by enabling the execution of user-defined behaviour (e.g. filtering, pre-processing, etc.) before data leaves the store so that only necessary data is transferred to the application space. Data format transformations between the objects in the store and the data structures managed by applications is also a source of performance losses. Adding semantics and structure to objects in such a way that applications can directly process the data held by the store and also store the generated data structures afterwards would help to reduce both application development and execution time.

The reduction of data sizes using e.g. hardware-assisted compression/decompression can also be beneficial for addressing the challenges of data-centric computing.

Highly Distributed storage resources (at the Edge, Fog and the HPC/Cloud)

As indicated in the research clusters the rise of the Edge, the Fog and the HPC/Cloud complicates the storage architectures. With workflows distributed across these infrastructure components,

data pools will be geographically distributed. There will be a need for understanding data life cycle management in such an environment. Access latencies get into the picture and these parameters need to be included in data life cycle management assumptions. Also, data needs to be intelligently cached at the right location in the geographically distributed workflow. Software/system infrastructure to enable better federation of data across geographies needs to be further developed. Existing solutions in this area are quite rudimentary and more suited to archives and not real-time workflows. Further pre-processing and machine learning operations need to be deployed to appropriate data pools, quite often closer to the Edge locations. Hence there is a need to understand the I/O implications of AI/ML components in workflows for highly distributed storage infrastructure. The storage pools with appropriate performance characteristics could also be distributed geographically as per the needs of the workflow. This raises the possibility of these different storage pools to be a part of a single geographically distributed storage system with appropriate storage infrastructure software (e.g. Object Storage).

5.3.5.2.3

Object Storage Systems in HPC and distributed environments

Parallel file systems such as Lustre or Spectrum Scale (e.g. GPFS) have been the incumbent high-performance storage solution for HPC. Parallel file systems and the associated POSIX semantics have to be continuously adapted to the new hardware ecosystem and the new architectural assumptions such as hierarchical storage.

Object storage continues to present an interesting alternative for the future as parallel file systems and associated POSIX semantics reach scaling limits. Object stores present potentially unlimited flat name space for unstructured data, with the possibility of having any associated rich metadata defined through key-value type infrastructure. It is also possible to easily "layer" different access semantics on top of the object storage foundation (e.g. data formats, data analytics plugins and POSIX) with flexible APIs, evidenced by the storage research in the last couple of years. Object stores have also now proven to work as a "scratch" for applications accessing hierarchical storage including NVM but only at very smaller scales. Flexible API's could also make object storage amenable for rich feature extensions and improved usage in various modes at different points in the Edge-Fog-data centre HPC infrastructure. Indeed, lessons learned from object store implementations in the Cloud (for example, performance implications of using key-value stores) can be applied for HPC based object store deployments. Designs that can combine both worlds (HPC and the Cloud) has the potential to improve efficiency of I/O as HPC moves towards Cloud deployments, while at the same time bringing new benefits from a unified underlying data storage infrastructure, especially with the projected heterogeneity in storage devices and processing units.

5.3.5.2.4

Understandability of storage systems

Failures at Exascale are going to be a norm rather than an exception. It will be important for data centre operators to obtain detailed insights on what is going on in the storage and I/O subsystem and take preventive actions in case of issues. The telemetry subsystem needs to help provide a systematic and detailed view of the state of the system, which is still not possible with existing mechanisms of manually going through unstructured log messages. This is particularly not a suitable method at scale. At scale, automated telemetry data analytics should be able to automatically predict infrastructure component failures and it should be possible to take preventive action such as predictive failure maintenance. There is a role for AI algorithms here as well. System telemetry should hence be provided in such a way as to be suitable for data analytics and AI algorithms. Standardised telemetry infrastructure is often lacking at the storage device level and current hardware event counters, etc. are extremely vendor dependent. Telemetry data also helps to provide inputs to storage system simulators that is still not available in the community to understand at-scale behaviour of storage subsystems. Runtime interfaces could use telemetry information and autotune their techniques to optimise and increase efficiency. For instance, checkpointing libraries could regularly monitor the condition of the file system and automatically reschedule checkpoints to less congested periods, according to the data gathered and analysed by the system. Collecting and publishing telemetry data on storage system failures could be part of an open science effort at the EU level to improve HPC data centre reliability in the EU as a whole - leading to better design of future EU based HPC systems.

As part of understanding storage systems, there continues to be a need to develop better I/O benchmarks that are representative of changing real world workloads. For example, latency and IOPS oriented benchmarks (not just batch/throughput oriented I/O benchmarks) may need to be developed to address real-time constraints imposed by AI/Deep Learning, etc.

5.3.5.3

Intersection with Research Clusters

5.3.5.3.1

AI Everywhere

- Storage APIs (e.g. Object Storage) need to have the necessary support for efficiently querying and indexing multi-dimensional data needed for AI, as described in the middleware interfaces section above.
- Fast and timely data access is necessary for model training to reduce the time to insight. It is also important to additionally better support random access patterns (compared to parallel file systems that are optimised for sequential accesses).
- Storage co-design with AI oriented applications and workflows will become increasingly important.

5.3.5.3.2
Data Everywhere

- It is necessary for all the aspects of storage and I/O research described above to look at (a) designing, (b) deploying, (c) managing and (d) understanding highly scalable storage and I/O infrastructures suitable for the era of exploding volumes of data.
  - Object storage exemplifies deployment models of such systems – leveraging software
  - Quality of service and middleware aspects (such as AI) address the management aspects
  - Storage Telemetry and Simulation address the understandability aspects
- It will be extremely important to reduce the cost of data storage as a very small percentage (~5% or less as per IDC) of useful data that is generated is actually stored. The rest of it (95%) is discarded due to capacity and cost limitations. The loss of this "dark data" eliminates any possibility to extract value in the future, e.g. with improved analysis and simulation methods.

5.3.5.3.3
Development Methods

- Storage software and APIs need to be future proofed for emerging storage device paradigms such as persistent memory
- Storage and I/O simulation will be an integral part of modelling of full systems.
- Standardised storage and I/O benchmarks and I/O performance analysis methods needs to be addressed.

5.3.5.3.4
Energy Efficiency

- Data movements between compute and I/O continue to be a major obstacle for building energy efficient systems. This problem is exacerbated with the increasing amounts of data volumes. In-Storage computing is a critical mechanism to reduce the energy footprint of data moving between the compute and storage systems.
- Emerging storage devices need to have well defined metrics on energy footprint per byte/block of data accessed.

5.3.5.3.5
HPC and the Digital Continuum

- Storage and I/O systems for HPC need to holistically address the data generated at the Edge and how to efficiently manage geographically distributed data pools and their real-time processing requirements (with Edge AI, etc.). There will be a big role to play for solutions such as in-storage computing.
- HPC storage and I/O systems need to be Cloud-enabled for better federation with distributed data and computing sources that will be part of the scientific workflow.

5.3.5.3.6
Resilience

- The storage system stack needs to develop well defined telemetry mechanisms to understand the causes of failures and the usage of telemetry information to predict upcoming failures, considering that failures of storage components (disk, flash, etc.) will be a norm at Exascale.
- Furthermore, end-end data integrity needs to be addressed by the storage stack and the storage device components to handle corrupt/lost data due to component and node level failures at extreme scale.

5.3.5.3.7
Trustworthy computing

- New storage software stacks and emerging storage devices need to have the ability to prevent malicious users from gaining access to sensitive data. Security needs to be an in-built feature in NVMe and Object stores, etc.
- Security mechanisms (e.g. authentication) need to be in-built in emerging Application I/O interfaces.

5.3.6
Mathematical Methods and Algorithms

5.3.6.1
Research trends and current state of the art

The development of future HPC architectures strongly depends on the evolution of underlying technologies, which, for example, lead to more and more parallelism or deeper memory hierarchies. New mathematical methods and algorithms are important ingredients in ensuring efficient usage of future architectures and technologies and in exploiting parallelism at multiple levels. Mathematics and algorithms are employed to handle increasingly wide and deep parallelism in order to exploit mixed, variable and reduced precision as well as new, non-standard floating-point number representation formats such as bfloat16, and in the management of increasingly complex and heterogeneous memory systems. With power consumption now a major factor in the total cost of ownership at almost any location, unconventional hardware architectures are becoming a competitive option with specific important application domains. Novel architectures of interest include data-flow systems, less reliable processing devices and approaches such as in-memory or in-network processing, which operate on data in place and thus enable extremely energy-efficient, application-specific acceleration. Developing new or enabling and adapting existing algorithms and mathematical methods will play a critical role in facilitating efficient exploitation of these new architectures.

The report *Mathematics for Europe*[52] outlines the importance of mathematics and stressing the need of scalable and robust algorithms for applications in HPC, data analytics and artificial intelligence. Mathematical methods and algorithms are to tackle societal, industrial and scientific challenges such as energy and

52. https://ec.europa.eu/futurium/en/system/files/ged/finalreport_maths.pdf

energy efficiency, sustainability, poverty, climate change adaptation and weather extremes, water shortage, transport, health and well-being. These problems must be addressed with advanced technological solutions based on efficient, robust and scalable mathematical methods and algorithms. In the following, we discuss specific challenges and opportunities, where research on mathematical methods and algorithms is particularly important.

5.3.6.2

Challenges for 2021-2024

5.3.6.2.1

Robust methods and algorithms enabling extreme scalability

To exploit the performance of future massively parallel architectures, many problems require new algorithms which express increased levels of concurrency for fixed problem size exploiting multiple levels of parallelism and/or specific hardware accelerators.

New mathematical methods may lead to innovative computing approaches that generate new levels or degrees of concurrency. Algorithms must be hierarchical to reduce both communication and synchronisation and to simplify task scheduling as well as closely match emerging computer architectures. In addition, the performance of system components is expected to vary across the different future Exascale systems. Global communication-hiding/avoiding algorithms such as pipelined Krylov solvers as well as hierarchical and hybrid stochastic/deterministic algorithms are particularly relevant and important. Stochastic and hybrid approaches not only introduce extra parallelism but also allow treating uncertainties at scale. Communication lower bounds for specific algorithmic approaches can be expressed and new algorithmic approaches should seek to be provably optimal in this respect or at least to explicitly declare their communications in addition to computational complexity. Moreover, task-based parallelisation strategies in mathematical algorithms are required, which efficiently promote critical path computations and handle asymmetric workloads. Algorithms enabling more extreme parallelism are of interest for time-critical applications as well as for enabling simulation of more complex and precise models. Here research is required to develop, for example, novel time integration methods that do not trade dispersion errors of fast propagating features with time-to-solution efficiency in multi-scale flows (e.g. exponential time integration). Closely related, additional levels of parallelism may be extracted from the time domain, for which parallel-in-time methods need to be developed that are robust in real world applications.

In light of massively increased concurrency, the robustness of algorithms is becoming more important. Fault-resilient algorithms that can run on large and error-prone systems will be increasingly required. Reproducibility of numerical results within given error bounds is important in some domains such as Numerical Weather Prediction and may lead to significant efforts for validating new methods. The traditional field of numerical algorithm stability remains critically relevant, but it also must address new areas such as Deep Learning as well as the effects of using reduced precision arithmetic. Addressing robustness will reveal many opportunities to trade performance for accuracy, making it possible to consider approximate computing scenarios where the precision of some computations can be reduced along with the number of floating-point or memory operations while keeping the quality of the results within user defined margins.

Beyond communication, novel numerical algorithms should exploit memory hierarchies and the availability of large non-volatile memory that might well make new or previously discarded mathematical algorithms attractive.

5.3.6.2.2

Methods for (scalable) data analytics and artificial intelligence

Significant challenges arise from extreme data problems, which require a paradigm shift from a compute-centric view to a more data-centric view. Algorithms for data discovery, and in particular those that discover global properties of data, such as graph analytics, require highly scalable compute resources and are non-trivial to parallelise. They often also lead to different data access and communication patterns. Other methods require analysis and visualisation to be embedded in highly parallel simulations or real-time processing of massive data streams for event detection and support of ad hoc decision making.

The central algorithmic approaches in these areas may not have been designed with parallelism in mind. Thus, core algorithmic developments may be required before the methods are well suited to execution on supercomputers.

5.3.6.2.3

Algorithms reducing energy-to-solution

The potential of reducing energy-to-solution has been demonstrated for a number of cases where different methods for solving the same problem revealed significant differences in terms of an energy-to-solution metric. Minimisation of data movement is an important aspect in this context as (off-chip) data transport is the most expensive in terms of energy consumption. New opportunities arise from the introduction of instructions supporting reduced-precision arithmetical operations leading to less energy per floating-point operation and energy savings due to the reduced amount of communicated bits. Additional aspects must also be addressed, including effective exploitation of a given hardware architecture and the design of crucial algorithms (e.g. FFTs) suitable for architectures based on particular power-efficient technologies.

5.3.6.2.4

Vertical integration and validation of mathematical methods and algorithms

Efforts are required to ensure that any of the developed mathematical methods and algorithms can be efficiently implemented on different suitable types of HPC architectures and can easily be used by a broad user community. Vertical integrations should ensure scalability at all levels, i.e. at mathematical models/methods level, through the algorithmic level, down to systems and architecture levels. The objective is to establish an integrated approach to the development of mathematical methods and

algorithms that leads to efficient and scalable programming models, tools and high-performance libraries optimised for specific architectures. Scalability at all levels and efficient use of the available hardware resources has to be assured up to Exascale.

In particular, vertical integration needs to address the following:

- **Vertical integration and validation**: Algorithms and mathematical methods are to be tested and validated with respect to scalability at all levels as well as ease of implementation, tuning and optimisation on various architectures. This must involve an exploration of full vertical integration from lower levels of system hardware to upper levels of system software. Performance portability across hardware potentially requires mathematical algorithms to be formulated in a higher-level representation from which optimisations can be derived.
- **Tuning of algorithmic parameters for Exascale**: The parameter space for tuning algorithms to maximise their scalability, performance and energy efficiency on current, emerging, and future Exascale architectures should be investigated.
- **Integration into diverse software stacks**: Convergence of HPC, extreme-scale data analysis and artificial intelligence will progress and a computing continuum from HPC to Edge computing is expected to emerge. This requires mathematical methods and algorithms to be integrated into a diverse set of software stacks.

5.3.6.3

Intersections with Research Clusters

5.3.6.3.1

AI everywhere

Mathematical methods and algorithms developed in the context of the research priority "Methods for (scalable) data analytics and artificial intelligence" are not limited to large-scale HPC systems but will also allow to develop AI applications that, for example, will be executed at the Edge. Furthermore, there is overlap with the research priority "Robust methods and algorithms enabling extreme scalability" as it encompasses research on the use of reduced precision, which is of particular interest when only a very small power envelope is available.

5.3.6.3.2

Data everywhere

This research cluster addresses in particular the challenges related to the processing of extreme-scale data volumes and progress of data analytics and artificial intelligence methods to derive information and knowledge. This overlaps with the research priority "Methods for (scalable) data analytics and artificial intelligence".

5.3.6.3.3

Development methods

This research cluster does not have a specific overlap with any of the proposed research priorities. However, various aspects of this research cluster will depend on mathematical methods and algorithms. This concerns e.g. strategies for enabling performance portability or the creation of (mathematical) models for understanding performance.

5.3.6.3.4

Energy efficiency

The research priority "Algorithms reducing energy-to-solution" will directly contribute to this research cluster. Furthermore, indirectly also other research priorities are concerned. Research on the use of reduced-precision calculations within the research priorities "Robust methods and algorithms enabling extreme scalability" and "Algorithms reducing energy-to-solution" can be expected to have a major impact on energy consumption. For energy-savings to materialise deployment of algorithms with improved energy efficiency properties is key. This means that there is also overlap with the research priority "Vertical integration and validation of mathematical methods and algorithms".

5.3.6.3.5

HPC and the Digital Continuum

This research cluster aims for enabling integration of vastly different types of systems used for data processing ranging from IoT devices to HPC systems. While mathematical methods and algorithms are not a major factor in achieving such an integration, they will be important in enabling individual components of workflows that can leverage such a continuum of data processing systems.

5.3.6.3.6

Resilience

The challenge of enabling resilience in an efficient way will typically strongly depend on progress in mathematical methods and algorithms. The research priority "Robust methods and algorithms enabling extreme scalability" explicitly addresses strategies for achieving resilience such as development of fault-resilient algorithms or the question of reproducibility of numerical results in the presence of less reliable systems.

5.3.6.3.7

Trustworthy computing

This research cluster does not have a specific overlap with any of the proposed research priorities. However, various aspects of this research cluster will depend on mathematical methods and algorithms. This concerns, e.g. efficient algorithms for establishing trust relations or protecting sensitive information.

5.3.7

Application Co-design

5.3.7.1

Research trends and current state of the art

Sustaining excellence and European world-leadership in HPC applications is key for European science, industry (including SMEs) and the public sector. There is a breadth of applications in the fundamental, applied and social sciences, where computing plays a pivotal role[53]:

53. E.g. PRACE, "The Scientific Case for Computing in Europe (period 2018-2026)", http://www.prace-ri.eu/third-scientific-case/

- A world class European computational infrastructure will expand the frontiers of fundamental sciences such as physics and astronomy, supporting and complementing experiments. Researchers will be able to simulate the formation of galaxies, neutron stars and black holes, predict how solar eruptions influence electronics and model properties of elementary particles. This will explain the source of gamma-ray bursts in the universe, advance our understanding of general relativity and help us advance understanding of the fundamental structure of matter by means of simulating the theory of strong interactions called quantum chromodynamics. This fundamental research itself leads to advances in the state-of-the-art of scientific computing and helps attract new generations to science, technology, engineering and mathematics.
- Simulations are critical in Climate, Weather, and Earth Sciences. Exascale resources will enable sub-kilometre resolution instead of the current 10km resolution, a more realistic representation of all Earth-system components, better mathematical models, and ensembles of simulations for uncertainty quantification. This will extend the reliability of forecasts to the extent needed for the mitigation and adaptation to climate change at the regional and national level, in particular with respect to extreme events. In analogy to weather and climate prediction, much enhanced simulation capabilities of solid Earth physics from higher spatial resolution and seismic frequencies down to 10Hz will enable a break-through in the detection and prediction of the precursors of volcanic eruptions and earthquakes, and their impact on infrastructures. A prediction capability at this level of detail is crucial for a wide range of societal impact sectors for food and agriculture, energy, water management, natural hazard response and mitigation and finance and insurance.
- High-end computing capabilities are becoming increasingly important for life sciences, medicine, and bioinformatics and will have tremendous impact, e.g. for enabling personalised medicine. Researchers are already able to rapidly identify genetic disease variants and it will become possible to identify diseases that are caused by combinations of variants, with treatments tailored both to the patient and state of the disease. Structural biology will increasingly rely on computational tools, allowing researchers to predict how the flexibility and motion of molecules influence function and disease. Deep Learning techniques will provide more specific diagnosis and treatment plans than human doctors, making medical imaging one of the largest future computing users.
- For Energy applications, the oil and gas industries are moving to full waveform inversion combined with neural networks for accurate detection. Exascale resources will make this technique feasible, allowing more accurate predictions of reservoirs and, while still fossil fuels, oil and gas have much reduced CO_2 emissions and air pollution compared to that produced by coal, which is still the dominant source of energy in the world. Simulations are of importance to improve the efficiency of nuclear power, hydropower, wind turbines and, not least, batteries and high-voltage cables to enable transmission and storage. Likewise, simulations are essential for the discovery and optimisation of renewable forms of energy supplies, their storage and distribution. In the longer term, magnetohydrodynamic simulations of plasma are critical for fusion energy, as in the ITER project.
- Computing is already used widely in Engineering and Manufacturing. Engineering applications based on fluid dynamics, combined with orders-of-magnitude-faster resources, will enable direct numerical simulations of the Navier-Stokes equations for better accuracy leading to improved designs and thus significant fuel savings e.g. in cars and airplanes, while also helping us understand phenomena such as cavitation. New data-driven approaches will enable scientists in academia and industry to integrate all aspects of design in models, use information from internet-of-things sensors, include uncertainty quantification in predictions and consider the entire life cycle of a product rather than merely its manufacturing.
- Chemistry and Materials Science will remain one of the largest users of computing, with industry increasingly relying on simulation to design, for example, catalysts, lubricants, polymers, liquid crystals and materials for solar cells and batteries. Electronic structure-based methods and molecular dynamics will access systems, properties and processes of increased complexity. At the same time, these methods are becoming more accurate. These are being complemented both with multi-scale models and data-driven approaches using high-throughput and Deep Learning to predict properties of materials and accelerate discovery. This will enable researchers to fulfil the grand challenge of designing and manufacturing all aspects of a new material from scratch, which will usher in a new era of targeted manufacturing.
- Advanced computing will increasingly be applied to handle complex data. Existing fields are starting to use Deep Learning to generate knowledge directly from data instead of first formulating models of the process, while next-generation infrastructure must be able to handle these applications with drastically increased data storage and I/O bandwidth capabilities. This will be needed for e.g. autonomous driving, Industry 4.0 and the Internet of Things. This will also enable computing to be applied in a whole range of non-traditional areas such as the humanities, social sciences, epidemiology, finance, promoting healthy living, determining return-on-investments for infrastructure by considering behavioural patterns and, not least, in helping to develop society and secure democracy, of which Cyber Security is an important aspect.
- As the previous topics show, industry can be found almost everywhere when it comes to large scale simulation. While in many cases, in particular in the case of the use of simulation, data analytics and AI by SMEs, industry is not requiring yet the fastest systems available, it is absolutely crucial that industry have a direct link to those who use these systems and to their

technologies in order a) to prepare for their own use as soon as they have the need for this kind of performance and b) to be able to run exceptionally complex tasks not every day but for particularly large challenges, where this compute power is needed to open up new possibilities. If industrial companies are not able to seamlessly connect to the highest end simulations, data analytics, and AI technologies, they will suffer severe disadvantages compared to those that do.

5.3.7.2

Challenges for 2021-2024

The computing requirement of these scientific and industrial applications drive technological development and applications must respond in turn by porting, adapting (including the use of new methods and algorithms) and optimising application codes for the new technologies. An effective co-design process between application and technology developers is crucial for a successful HPC ecosystem to ensure technology is relevant for applications and can be fully exploited when available. Moreover, the co-design approach fosters an ecosystem of skills and expertise in Europe by retaining scientific talent and enhancing the competitiveness of industries on the global market.

Access to increased computational power and to applications exploiting the features of the system remains crucial to enable more detailed and large-scale (compute intensive) modelling and simulations. At the same time, new approaches such as data-driven computing, High Performance Data Analytics, AI and their convergence with classical HPC enable new opportunities and require new capabilities, including efficient access to large amounts of data with low latencies and high-bandwidth and support for new and large workflows and ensembles encompassing orders of magnitude more active jobs than today.

Many applications are severely limited by memory bandwidth or communication latency and as the throughput of floating-point operations have increased faster than the data transport capabilities; even many traditionally floating-point bound applications are now highly memory sensitive. Intra-node and inter-node communication require drastically reduced latencies and high-speed networks, particularly for algorithms based on fast iterations of short tasks so that they can achieve significantly-improved performance and strong scaling.

Storage and I/O requirements are expected to grow even faster than compute needs, with much larger data sets being used e.g. for data-driven research and machine learning. This is not limited to the amount of storage but data-heavy applications will also need exceptionally high-bandwidth parallel file systems, and/or advanced data caching solutions on each node. This increase in storage and I/O resources must be coupled with provisioning of a large-scale end-to-end data e-infrastructure to collect, handle, analyse, visualise, and disseminate petabytes to exabytes of data.

Long-term maintenance and portability of codes (both in terms of performance portability as well as compiling/building the codes on new architectures) are other important factors, requiring standardised, open, and supported programming environments and APIs, supporting a wide range of different hardware technologies. In addition, the software engineering practices need to improve and a Europe-wide effort in training and education needs to be organised.

While in the longer-term computing leadership will require the development of alternatives to the current technologies, including perhaps quantum, data-flow, neuromorphic or RNA computing, there is consensus that even today the fundamental mathematical and computer science algorithms needed to meet the requirement of leadership science and industrial competitiveness are not in place. The energy-efficient, application-oriented next-generation computing platforms therefore require an ambitious programme of algorithm development integrated with the co-design/co-development of the overall infrastructure and sustained over longer timescales than the usual 3-5 years funding cycles.

All these approaches require co-design activities involving architectures, OS, communication libraries, workload management and end-user applications to achieve the intended results.

5.3.7.3

Intersections with Research Clusters

Clearly, all the research themes defined in *5.2 Research Clusters* on page 36 are of great importance to the various application domains; however, not all themes are equally important to all applications domains. In the following table, we sketch a rough "heatmap" of the relative relevance of a given research cluster to the various domains of applications. Blue means "relevant", Yellow - "very relevant" and Red - "highly relevant". Of course, this is just a rough categorisation and there will be subdomains or individual codes in an application area where this heatmap does not fully cover every aspect.

In the following paragraphs, we discuss generic application aspects of the research clusters and we highlight application domains for which a specific research theme is highly relevant.

5.3.7.3.1

Development methods and standards

Extreme-scale computing resources offer great capabilities, but this comes at the expense of increasingly complex hardware architectures. The efficient exploitation of current heterogeneous systems with mixed general purpose and accelerator processors, variable memory hierarchy and network topologies require considerable efforts by software developers. Thus, productivity of researchers will be greatly increased through suitable programming models and runtime systems. The systems should allow a higher level of abstraction for orchestration of both large HPC compute tasks as well as massively parallel, high throughput execution cases. Adequate workflow platforms will also play an important role in the process. It should be noted that a large number of such development methods already exist (i.e. programming models, runtime systems and workflow platforms) but that in itself creates additional obstacles to wider adoption, provisioning and maintenance. Duplication of effort and "reinvention

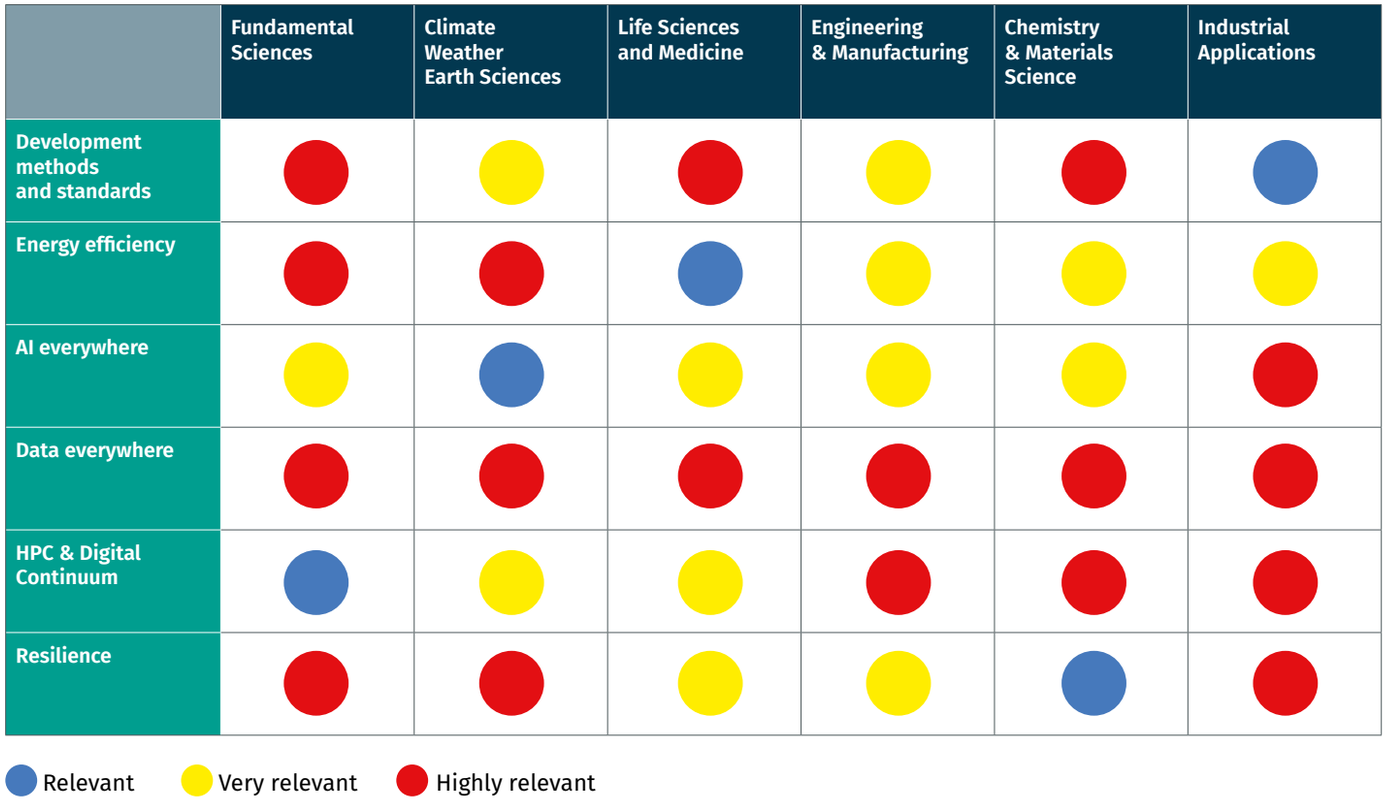


Figure 15: The relative relevance of the Research Clusters to application co-design.

of the wheel" should be avoided as much as possible. Standardisation will play an important role in this process. At the same time, long-term support and sustainability of programming environments is a must for widespread adoption.

In addition, most solutions will not be able to exploit full bi-sectional bandwidth making systems with islands of hyper-connectivity more appropriate but requiring more intelligent queue systems and APIs. Any increase in ensemble, sampling and parameter scanning approaches to combine hundreds/thousands of simulations (which themselves are highly parallel) mean that Exascale queue systems must be flexible and agile.

Several applications require the execution of very large ensembles of simulations which can be relatively short (e.g. a few minutes) or long in duration, taking many hours and even days. Very efficient mechanisms to build and execute such ensembles are needed. Such ensembles will also play an increasingly important role in material science and will increase the stress on related services, such as databases.

Applications require development methods that flexibly exploit (heterogeneous) processor and memory hierarchy architectures and maximise sustained performance given evolving architectures and standards. Domain specific languages (DSL) assume a key role as the interface between user-friendly languages and intermediate representations that can be mapped onto specialised kernels with hardware dependent optimisations. These programming/runtime systems need to be highly efficient, particularly in the case of individual simulations with short execution times which are sometimes found in ensembles. They also need to minimise the overhead they introduce.

Many applications have widely adopted accelerators and significant effort is spent to port and optimise code for these architectures. Programming environments that ensure performance portability are required to avoid new porting/optimisation efforts for each target architecture.

5.3.7.3.1 *Energy efficiency*

The power consumption of supercomputers is a major challenge for system owners, users, and society, with many current high-end systems drawing over 10 MW. Only few future Exascale systems will get close to the very ambitious 20 MW limit. These levels limit the capacity of system installations, require significant cooling infrastructures and create a large carbon footprint. Reducing power during application execution without changing the application source code or increasing time-to completion is highly desirable in real-life high-performance computing scenarios.

However, the power management run-time frameworks proposed in the last decade are based on the assumption that the duration of communication and application phases in an MPI

application can be predicted and used at run-time to trade-off communication slack with power consumption. This assumption is an over-simplification and leads to poor predictions, slowing down applications and thereby jeopardising the claimed benefits. New approaches need to be developed, which will make it possible to achieve performance-neutral power savings in MPI applications without requiring labour-intensive and risky application source code modifications.

5.3.7.3.3

AI everywhere

The recent success of AI applications has been a driving force in the design and development of AI-optimised processors or accelerators. Examples include Intel Nervana, Google tensor processing unit or the Cerebras Wafer Scale Engine with 400 000 cores, Graphcore chip. However, the trend of introducing many new AI-optimised processors poses challenges to software development for AI applications. Major research efforts will be required in the addressing performance, productivity and portability of AI software both for Edge devices and large-scale computing systems.

In addition, AI technologies are increasingly used in HPC system in several application domains such as biology and healthcare systems. In the former, new ML applications that are processing demanding have been proposed; one representative example is ML-based genotype imputation which is utilised for the diagnosis of several genetic diseases. In healthcare systems, new highly parallel ML applications have been developed for the analysis of vital signals (e.g. EEG, ECG) to predict several disorders and medical conditions including brain haemorrhage, depression, stroke, Alzheimer's disease and epileptic seizures. The healthcare applications have an additional characteristic compared with the more conventional HPC applications: they require low latency since they have to perform enormous amounts of operations in a matter of minutes.

Also in the fundamental sciences, data-centric needs driven by observation and experiment (e.g. LIGO, SKA and the LHC at CERN) encompass HTC computing and increasingly exploit ML and AI paradigms.

5.3.7.3.4

Data everywhere

Data is a key aspect for applications and efficient handling of data is paramount for the overall application performance. This starts from efficient data movement from various data sources, e.g. scientific instruments (e.g. LHC, SKA, etc.), satellites, data collections, and Edge devices (e.g. sensors in cars), staging of data in the every more complex storage and memory hierarchies of HPC systems, to efficient processing in simulations and data analysis applications. Tailored data management plans need to be developed to support these complex workflows and efficient data tools need to be available. In addition, data often needs to be compressed or even reduced at various stages of the workflow to keep things manageable. This includes data filtering and transformation at the source (Edge), at data federation points (Fog), as well as in situ approaches in running simulations. Moreover, the increasing usage of ensemble techniques in large scale systems increase the amount of (potentially small) files that need to be managed efficiently.

For instance, in Medicine and the Life Sciences, dynamic studies of biomolecular systems through molecular simulations generate large amounts of trajectory data, which must be analysed (i.e. in order to extract relevant statistics) and which is also often correlated with data obtained through other means such as experimental studies or data mining of external resources. As an example, high-throughput screening of large drug target libraries requires analysis combining both, simulation and experimental data. Visualisation of the data also plays a crucial role in structural biology research. The aforementioned trajectory data is routinely visualised as well as structural biology data originating from microscopes, such as Cryo-EM. Data privacy and security also plays an important role in these domains as personalised data needs to be appropriately protected.

Weather prediction employs very complex workflows for executing observational data collection, initial condition generation, ensemble forecast production and model output post-processing and data dissemination. While weather forecasts run these workflows in burst mode along daily schedules, climate prediction executes workflows with stable workloads for longer periods of time. Weather and climate prediction comes with two categories of data concerns: observational input data volumes and diversity will increase through more sophisticated satellite observations but mostly through IoT observation with unprecedented data management and privacy issues. Forecast output will grow by three orders of magnitude, which implies that downstream applications must move near the data source enhancing the need for a more flexible and open, data-centric workflow management.

Across the fundamental sciences, data (simulation – with latency/bandwidth limitations, movement, storage and analysis) is rapidly becoming a limiting factor. Code and algorithm development is needed to adapt to new programming paradigms if storing and transmitting data is more expensive than recomputing. Already, latency/bandwidth bottlenecks are becoming (and in some case already are) the limiting factors. Storage and I/O requirements will be 100x greater than today, outpacing the growth in computing needs and with related requirements for high-bandwidth file systems and/or advanced cache solutions on-node. This increase in storage and I/O must be coupled with a large-scale end-to-end data e-infrastructure to collect, handle, analyse, visualise and disseminate petabytes to exabytes of data via e.g. high-bandwidth gateways to future EOSC hubs.

5.3.7.3.5

HPC and the Digital Continuum

The Digital Continuum has these two aspects: hardware/infrastructure and software/application. On the hardware side, the continuum ranges from local to remote/Cloud. While the local infrastructure has always been present and the Cloud has been established over the last years, the "in-betweens" are fairly new: When local processing is too slow but the data is too large to be

communicated via network into the Cloud, the so-called Edge-computing can be applied to bridge the gap and to establish the continuum. For industry, this kind of continuum is important in order to be able to provide the best performance, regardless of the current environment (e.g. autonomous driving in a dead spot where Cloud access is not available but the local power in the control unit might not be sufficient).

On the software side, the simulation programme can be a part of the Digital Continuum in that it allows users to handle the whole function of a product in the virtual environment. The so-called digital twin mimics the behaviour of the real-world part in the computer and thus allows a full valuation of a scenario without the need to have the physical set-up available. For industry, this is an extremely important part in order to be able to cost-effectively develop new products as well as set up new plants and be sure that it works right from the start. HPC performance is crucial for today's increasingly complex products and plants, which require significant compute resources.

5.3.7.3.6
Resilience

In large-scale parallel systems, faults are not an exception but rather an undesired but unavoidable norm. Usually, faults may arise from application failures, operating system failures, system uptime and service failures. While hardware and software techniques have been developed to make systems more resilient to failures (checkpoint/restart, component redundancy, algorithm-based fault tolerance), a reliable solution is not yet available. Future Exascale systems are projected to suffer from an increase in the frequency of the faults, mainly because of the increase in the number and complexity of components and equipment, and recent studies show that up to 20% more circuitry and energy could be required to counter the increase.

Because of the design and manufacturing costs, hardware vendors tend to deliver systems for general-purpose customers, without paying too much attention to the requirements of Exascale customers. As a result, the future Exascale systems will most likely be built with off-the-shelf components with the responsibility for resiliency delegated as much as possible to the software layers and techniques. Beyond the increase in components in Exascale-class systems, the tightly coupled programming model may result in a fast fault propagation across the whole system following the fault on just one node. While resilience schemes can be developed for a specific application, a numerical algorithm or a set of similar applications, this approach does not address or solve the vulnerability of the whole system and leaves ample room for vulnerability that can jeopardize the usability/reliability of the system. Therefore, resilience should be addressed within an end-to-end framework so that the range of vulnerabilities remains as much as possible limited and localised and does not propagate through the whole system. These frameworks need to improve on the current approaches involving redundant hardware and application checkpoint/restart that are both costly and time consuming and thus likely not scalable towards Exascale systems.

Resilience is an important aspect in weather prediction, for example, as all aspects of our lives closely depend on evolving weather patterns. A number of key decisions are taken according to hourly-changing weather forecast and associated meteorological predictions. Moving forward, Exascale-class of systems will be capable of resolving processes on sub-km scales, enabling an explicit representation of the storms that are important to the weather forecast in the mid-latitudes as well as equator to pole energy transport which is crucial for the climate. The current approach (adopted at e.g. ECMWF) is building an environment comprising two independent systems located in separate halls. Each system has separate resilient power and cooling systems to protect against a wide range of possible environmental failures. The approach of physical redundancy of the systems is untenable moving towards Exascale-class systems because of the huge infrastructural needs and intrinsic fragility of the approach at the Exascale. System resiliency will have a key role in enabling Exascale-class systems to provide timely forecasts implementing intelligent and predictive early failure detection sensors, techniques for the localisation and limitation of the fault (preventing a localised fault to become a system-wide faults) and algorithms recovering from a failure (hardware, operating system, libraries, I/O).

Industrial applications usually depend on ISVs (Independent Software Vendors), who typically do not develop their codes with Exascale-class systems as their first priority and rely on the system's hardware and software environments to manage faults. To address the requirement for maximising the reliability of Exascale-class systems running industrial applications, a co-design approach between system designers and ISVs is needed to create a new class of applications showing fault-tolerant features and intrinsic resilience to system's faults.

5.3.7.3.7
Trustworthy Computing

All applications require a basic level of security and trust, particularly when computing on the Cloud or other shared environments. As there is usually a trade-off between performance and trustworthiness, many applications accept basic levels of security allowing to focus more on performance. However, trustworthy computing is of particular importance for applications that deal with personalised data (e.g. in the Life Sciences and Medicine) and also for industrial applications, where business secrets need to be protected. However, data integrity is important for still other domains, particularly when data is coming from different sensors and instruments, as in weather and climate or astronomy. For those domains, appropriate mechanisms on both, the hardware and software levels are required to be able to build applications at the required level of security and trust while allowing as much performance as possible.

5.3.8

Centre-to-Edge Framework

5.3.8.1

Research trends and current state of the art

Future HPC use will increasingly transcend the classic model of running scientific simulation and modelling applications within a single HPC centre environment. Instead, HPC use will be characterised by a complementary mix of HPC centre and Cloud approaches and integration with a multi-tier device, Fog, Edge and central infrastructure.

Considering the rapid evolution of technology and use cases, the term "High Performance Computing" or "HPC" needs to be redefined: in the past, it was synonymous with "technical computing using supercomputers" to model or simulate complex scientific or technical phenomena. While HPC will still include systems specifically designed for application scaling (up to many thousands of nodes), the larger share of HPC systems will become part of a wider digital infrastructure which runs complex, efficiently managed and orchestrated workflows. One example of a workflow is the application of machine learning models on Edge devices such as cars or mobile phones (inference), with the model training taking place within an HPC centre. Figure 16 shows a simplified dataflow from centre-to-edge for this scenario.

- Our understanding of "Edge computing" refers to the distributed computing paradigm largely or completely based on placing compute capabilities close to the end-user or to IoT (Internet of Things) devices. This avoids high latencies caused by long lines of communication between central computer systems and end users and devices, and can facilitate compliance to local regulatory or business requirements (data privacy and geo-fencing come to mind). Edge computing relies on ubiquitous use and deployment of wireless communication, in particular 5G protocols with ultra-low latencies.
- "Fog computing" refers to layers between the Edge and the central compute infrastructure which reduce the volume of data in transit between these and enable complex actions and decisions close to the Edge. The increasing demand for "near real-time decisions" drives the need for a shared continuum of scalable compute and data resources which consists of a combination of Edge, Fog and central HPC or Cloud data centres.
- "Cloud computing" rather refers to a usage model than to a specific architecture or parts of it. Instead of running computation and data analytics purely on systems placed on premise, Cloud computing leverages remote resources in a flexible and (almost) transparent way and enables billing according to actual usage. Today, Clouds are implemented as a collection of large data centres (with providers often referred to as "hyperscalers") and WAN networks, with elaborate system and application software layers ensuring highly efficient and reliable operation and fulfilment of customer SLAs for highly complex applications and workflows.

5.3.8.1.1

Current state of the art

Today, the different domains of Edge, Cloud and, in particular, HPC are developed and maintained by almost completely distinct communities. However, the Edge domain starts to be leveraging Cloud centres for data processing and storage and Cloud providers start investing in bringing HPC applications into their premises.

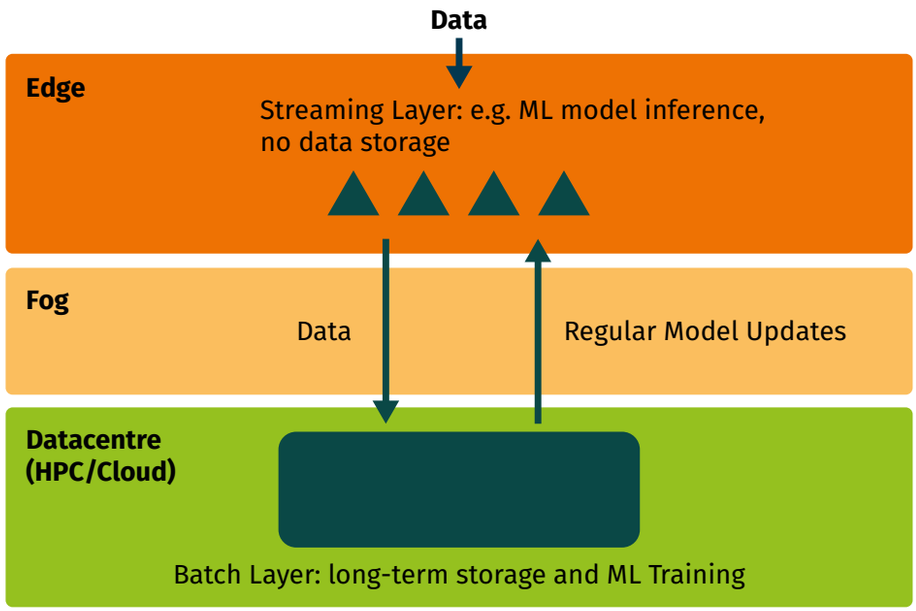


Figure 16: Centre-to-Edge workflow for ML (example).

Despite leveraging Cloud computing the embedded or Edge domain is moving from many small, embedded devices towards (virtualised) high performance CPUs/accelerators for Cyber-Physical Systems use cases (CPS, such as automated or autonomous cars and shop-floor robots). This enables them to cope with increasing amounts of data produced by sensors and the complex computation and analysis operations required. An alternative is to send data into small computing clusters or Cloud environments for analysis. Controlling and optimising operations of large number of such CPS systems will require a tiered infrastructure fully embracing the Fog/Edge approach. Network providers are moving towards new high throughput communication standards such as 5G. As a result, 5G base stations and general network infrastructure are trending fast from using dedicated, single-purpose devices to leveraging general-purpose CPU plus accelerator combinations for efficiency, which creates an entry point for adding HPC capabilities at the Fog and Edge levels.

Cloud service providers (CSP) are focused on providing general-purpose resources for commercial IT tasks, data storage and single-node compute intensive applications. Compute, data and network resources are highly optimised to meet the customer demands (e.g. small-scale parallel execution, micro-services) and maximise profitability.

Cloud user interfaces provide easy user access and elastic processing in terms of computational resources as well as storage. User on-boarding, scheduling, data replication, etc. are fully automated so that no manual interaction needs to take place. Services are offered to the end-user as Infrastructure-as-a-Service (IaaS) and serverless solutions such as Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) or Function-as-a-Service (FaaS). All services are tightly coupled to create on demand, elastic (serverless) workflows for data processing.

Cloud centre scale surpasses even the largest HPC centres. For that reason, hardware vendors consider the CSPs as priority customers. Top CSPs have a huge influence on technology providers and integrators or may even develop their own hardware in order to meet the security and privacy demands as well as to provide performance guarantees to customers. Examples here are FPGA usage for on-the-fly encryption for network based I/O or for creating virtual private networks and Root-of-Trust chips for hardware verification.

For many years, one of the main arguments against using CSPs has been that Cloud environments are not secure enough for sensitive data. However, CSPs have invested heavily in building trust and by necessity have become experts in meeting IT and data security as well as privacy requirements in order to be able to support the handling of sensitive business and possibly personal data.

Since they are at the heart of their business model, CSPs have built an underlying composable services infrastructure which gives them the capability to dynamically and automatically define prices and matching SLAs. These tend to be proprietary, thus raising the problem of how to integrate true “Multi-Cloud” environments.

Increasingly, large CSPs such as Microsoft Azure and Amazon Web Services (AWS) are entering the market of providing closely coupled, parallel HPC systems and applications. They are facing an economic challenge: closely coupled, strongly scaled HPC applications require costly high-performance low-latency interconnects and top-bin mixed CPU and GPU types, and these investments will not pay off running the classic CSP applications. Finding the right prices and maximising the use of this high-cost infrastructure by high-value HPC applications is a key requirement.

HPC applications typically use and produce very large amounts of data which need to be transferred back and forth across a WAN infrastructure, which can become very expensive. It is highly advantageous to store this data close to the compute sites.

All major nations are investing heavily in Exascale HPC computing following the classical HPC centre model. Such HPC centres are designed and built to serve scientific communities or commercial entities to run medium to large-scale numerical simulation jobs which require tightly coupled computing nodes due to the need for low-latency, high-bandwidth inter-process communication characteristic for HPC applications. They provide centralised computation and storage resources that are accessible by command-line based techniques or special-purpose portals. The authorisation and accounting are based on organisational membership and mostly slow review processes for user identification, which increases the barriers for new potential customers.

Customers need to use a batch scheduling system to queue their applications for execution. Batch scheduling is preferred by HPC centre providers, as it leads to high system utilisation, which again leads to energy and cost efficiency. In some cases, however, it may not support fast-turnaround, on-demand and pay-as-you-go provision models. Nevertheless, advances in offering flexible access methods and interfaces are noticeable in industrial HPC and in large and ambitious scientific projects (such as the Human Brain project).

Besides providing highly optimised hardware, parallel storage and software, HPC centres can offer their customers unique expertise in application optimisation and high-touch support and training for large-scale HPC workflows.

Deep Learning (DL) is rapidly emerging as a main driver for the design and use of HPC capabilities in service-oriented environments, besides the classical simulation codes. DL relies on heavy use of dense linear algebra operations by high-level software frameworks which totally isolate the end user from any HPC specifics.

5.3.8.1.2 *Emerging challenges*

The increasing amount of data collected and produced especially by IoT/Edge devices and the quickly increasing use of compute-intensive Deep Learning and Data Science workloads means that architectures emphasising high performance and scalability will become increasingly important across the Edge devices them-

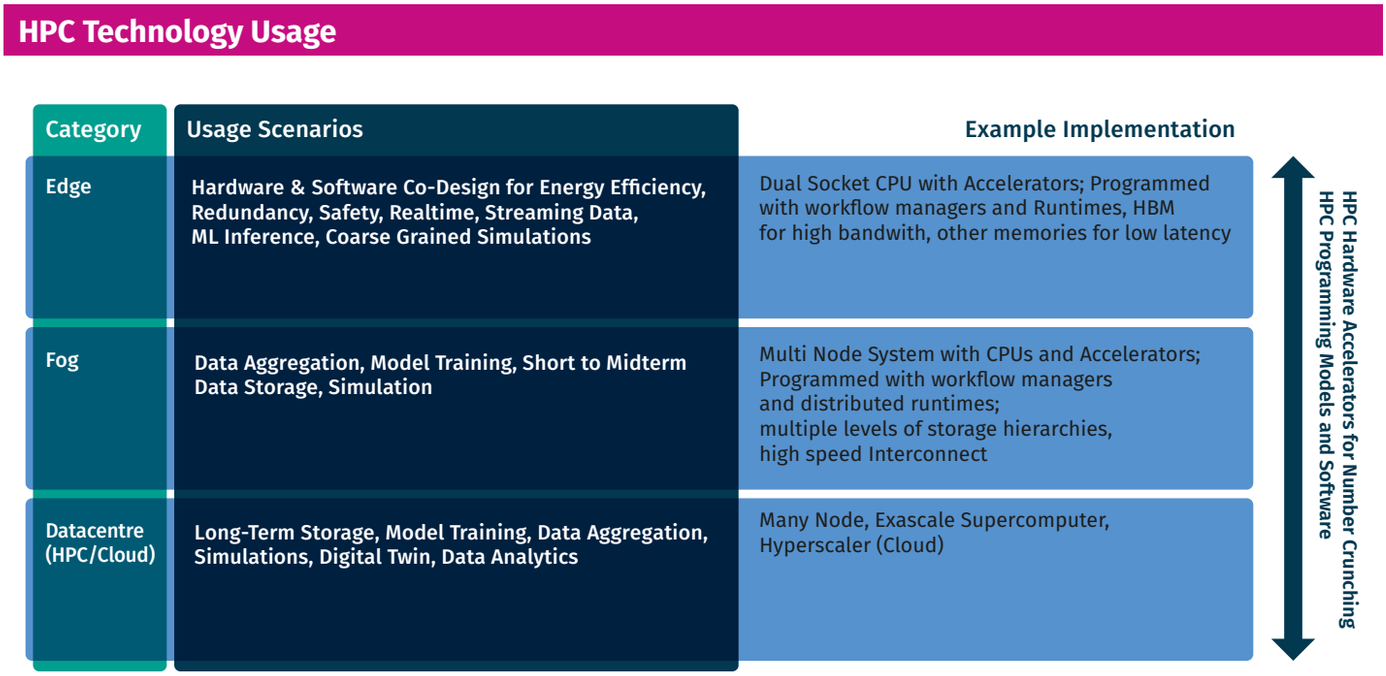


Figure 17: Emerging Application of HPC technology and methods

selves, the Fog aggregation layers and the Cloud. Hence, emerging HPC technology usage scenarios will not be restricted to tightly coupled, single data centre applications but extend to distributed usage domains (see Figure 17). Edge devices will leverage parallel and specialised hardware solutions such as accelerators and multi-core systems, Fog and Cloud providers will implement HPC hardware and tools for larger scale processing and simulations which makes the integration of HPC programming models and runtimes necessary. In order to foster this adaptation, HPC technology needs to be optimised to fulfil the energy-efficiency, security and reliability demands, e.g. in the domain of autonomous cars, mobile devices, etc.

Additionally, HPC will be considered as a service and thus it will become a part of the overall application workflow across the continuum of domains. Such workflows must be able to define a complete data flow all the way from the Edge devices down to the data centres.

This requires effective management of extreme scale heterogeneity, where, in the worst case, the common denominator between systems might be common governance and resource management interfaces and policies. At a high level, the main technical challenges are interoperability between the application workflow components, efficient orchestration and scheduling plus reproducibility of execution in order to allow debugging and ease of deployment. Fractured infrastructure management and resource allocation policies are strong roadblocks which must be overcome. For instance, supercomputers today are typically deployed in silos with limited external connectivity, proprietary access processes, relatively rigid operational models that expect users to submit batch jobs, and limited flexibility in terms of provisioning application-specific software stacks. It is difficult to host application workflows which include components deployed in the Cloud, handle streaming of data and react dynamically to the way the execution is unfolding or to external data and events. Tight integration of capabilities across individual systems and between large data centres and local (on-premises / Fog) small-scale HPC systems will be required.

We expect that for the most part HPC and Cloud solutions will complement each other. However, future HPC centres will be faced by tough competition from CSPs in the areas of such application workflows. On the one hand, CSPs enjoy huge economies of scale in their investments and operations due to their large or even global customer base, multitude of supported applications (from business IT tasks and micro services to highly complex and compute-intensive workflows, such as in for life sciences) and unprecedented size and capacity of their data centres. By deploying more tightly coupled computing resources in their environment, the CSP will become a competitor at least in the area of small to medium scale HPC applications.

On the other hand, we expect that in the case of large-scale HPC workflows, classical HPC data centres will continue to be more cost effective, reliable and predictable in terms of performance and costs for the end users. For specific parts of their workflow, customers will be using the offerings from the provider

which delivers the best cost-performance or cost-productivity ratio. Some of the current HPC customers will trend towards using Cloud environments and, on the other hand, new customers will be leveraging the capabilities of HPC centres.

The new HPC customer base, from data analytics and machine learning areas, is used to the service-oriented, seamless and automated usage models that CSPs or integrated desktop systems have been offering for a long time and they will expect a similar service and interfaces from future HPC centres. More specifically, there will be requirements for:

- interactive use of HPC resources, with the expectation of getting results back with minimal delays or even under real-time constraints
- availability of optimised support for high-productivity languages and customisable environments such as Python, Scala, R and platforms such as Apache Spark for data analytics and Deep Learning frameworks such as Tensorflow and Pytorch
- integration of external, and potentially streaming data sources from Edge devices, which would need support to push and pull data in and out of data centres, targeting data centre storage or running, potentially highly parallel HPC applications
- modern, data-driven science applications which will need large volumes of storage, data access performance commensurate with evolving CPU and accelerator speeds (best provided through having multiple levels of memory and storage), non-POSIX storage interfaces (such as object stores, key/value stores, byte-addressable NVM) and potentially in-memory or in-storage compute capabilities
- monitoring of resources via graphical user interfaces (data interfaces bandwidth usage, resource usage, etc.)
- IT and data security and privacy mechanisms protecting the compute and storage infrastructure and data stored or processed therein. These must comply with legal requirements in the protection of personal data (such as the European GDPR and member state privacy laws) and commercial best practices.

The implications of this wider user base are that the current business models of the HPC centres need to be re-thought to reflect the new usage models. For example, rather than copying what a standalone CSP could do, we believe that HPC centres could provide those specialised services as a federation of resources and unified access to publish the HPC services (IaaS/PaaS/SaaS or a combination) that they agree to share with the community along with the associated SLAs (price/automation/any other terms). Such mechanism coupled with a robust software stack, industrial support and efficient operational processes could allow integration of smaller centres or CSPs where they would best fit, as well as allowing HPC centres to be positioned with differentiated offerings. An important factor is the competency of HPC centres in supporting application developers to achieve best performance and efficiency.

Having a federated European HPC landscape would enable geographically distributed data storage and processing, with the objective of simple load balancing or ensuring redundancy or managing data which is not allowed to leave a certain regulatory domain.

5.3.8.2

Challenges for 2021 – 2024

HPC centres must offer streamlined and straightforward access to HPC resources:

- There should be a unified way across HPC centres to allow user registration to be performed within minutes without manual interaction.
- There should be flexible user and permission management in order to accommodate company accounts to consolidated billing, the creation of sub accounts to restrict permissions and possibly quotas for specific users, and to allow easier administration without manual involvement of the HPC centre provider.
- One of the main advantages of the large Cloud providers is that they offer one SLA for all their datacentres. In the case of HPC, many different providers offer their services to the customer under a variety of SLAs. The HPC landscape needs a more standardised way of specifying services, prices and SLAs, support for data brokers and application and data marketplaces.
- New accounting and billing methods are necessary to cope with the upcoming new usage scenarios, where data is fed into an HPC centre and exported, dynamic workflows are executed with varying demand in computation. This includes more fine-grained billing based on consumption (use) e.g. pay-as-you-go way, through short-, mid- and long-term reservations or subscriptions to services.
- Data Scientists are used to interactive usage of computing resources. These should be accessible through low-threshold graphical interfaces such as Jupyter Notebooks, including effective job monitoring.
- To be able to integrate the HPC centres in continuum (Edge, Fog, Cloud) workflows, it is a necessity to implement standard interfaces for data I/O (even for streaming data), monitoring, remote access and resource allocation. Such interfaces and APIs should be standardised across all HPC centres.

New programming paradigms and environments need to be natively supported in order to include Deep Learning and Machine Learning into application workflows. This includes HPC-optimised versions of APIs and high-productivity programming environments and SW stacks such as Python, Julia, R, Tensorflow and Pytorch as well as tools from the Hadoop or Apache ecosystem. One of the challenges is coping with such changing programming

environments - e.g. Python environments should be easily customisable by the end user via containerisation or other methods. Such changes must be made persistent and selectable for long running batch jobs.

Big Data and HPC middleware solutions (schedulers, orchestrators, file systems, parallel runtimes, monitoring and control) must be adapted. New frameworks and runtimes need to be developed to ease the development of cross-continuum workflows for batch and streaming data.

New (non-Posix) storage solutions for (un-)structured data should be offered, such as object stores and databases (SQL and NOSQL).

Integrating new domains into the HPC centres will require higher security and privacy standards which must be guaranteed according to laws and SLAs. This includes authorisation, authentication, encryption for data in-transfer or at rest. To build trust towards the customer these techniques must be clearly documented and communicated.

In order to adapt HPC technologies to the Edge and Fog markets, extensive co-design is needed to enhance energy efficiency and meet (soft) real time constraints and security requirements. The co-design process will not be limited to new hardware architectures but needs to include new operating systems, runtimes, compilers and programming languages.

5.3.8.3
Intersections with Research Clusters

5.3.8.3.1
Development methods and standards

New frameworks such as domain-specific languages, libraries, programming and abstraction frameworks must be developed or integrated to account for highly heterogeneous systems in data centres and for cross-continuum centre-to-edge workflows. Standardised interfaces and APIs should be created to ensure portability between different data centres and layers of the Cloud/Fog/Edge continuum.

5.3.8.3.2
Energy efficiency

Energy and space efficiency are important for the adaptation of HPC technology to power restricted domains such as Edge and Fog environment as well as in data centres. This includes the use of heterogeneous computing platforms to improve the efficiency of computation and storage.

5.3.8.3.3
AI everywhere

Machine Learning and Deep Learning will be one of the main drivers for centre-to-edge workflows where compute intense part of model training will be done within the HPC centre. For training and inference, high-performance and energy-efficient architectures such as special accelerators or even non von-Neumann architecture (neuromorphic, quantum, etc.) will be used on the Edge and within the centres. This requires common software stacks and programming methodologies and standards.

Developing centre-to-edge workflows integrating IoT, Edge and Fog devices will require data handling at every level. This includes data life cycle management, data security and privacy, scalable storage solutions for structured and unstructured, streaming and at rest data. Interfaces and data formats should be standardised in order to ease processing. Distributed runtimes should be able to support workflows running on multiple, centres distributed geographically.

5.3.8.3.4
HPC and the Digital Continuum

The topics from the Digital Continuum research cluster are reflecting the challenges identified in this working group. This covers the spectrum from the deployment of small scale HPC cluster and computer federation to embedded HPC in Fog and Edge devices and the seamless integration of heterogeneous architectures in data centres. The coding and managing complex workflows for streaming and batch data needs to be supported. The new usage scenarios need interactive access to resources and provide external APIs and interfaces for data I/O and monitoring.

5.3.8.3.5
Resilience

Deploying HPC technology in critical environments in the continuum in particular requires resilience to recover from hardware and software failures as well as monitoring and logging for reporting. These topics do extend to the applications running in the HPC centre, especially when the number of computing devices and the amount of workloads are growing, failures must be considered and properly handled.

5.3.8.3.6
Trustworthy computing

Introducing new application domains for HPC technology as well as integrating HPC centres in centre-to-edge workflows will create new threat models and legal requirements. To cope with these, security must be rethought and introduced at every level - from hardware development to software stacks, applications design, processes and data handling. Also, the HPC focus on achieving utmost performance has to be reconsidered in the light of the overhead required for security methods/processing, such as end-to-end encryption. ■

6

Upstream technologies – focus in the 2021–2024 period

6.1

Context

This part of the SRA focuses on the upstream technologies which have the potential to contribute to the European HPC effort in the context of this SRA (in the period 2021 – 2024).

Today's HPC system architecture is dominated by standard CPU+GPU solution. This architecture has been effective to propose the performance increase requested by HPC users while challenging them to exploit the massive parallelism and heterogeneity offered by this solution.

We expect few changes in the 2020-2023 timeframe, with the first Exascale systems based on this approach. Thereafter, in order to sustain the growth of the number of operations per watt, new solutions will have to be found as the Moore's law will be fading and Dennard's scaling is not applicable any more. Progress can be made along three axes:

- New architectures
- New data representation schemes
- New technologies (compared to CMOS)

Most of the new approaches are a combination of at least of two of the three elements above but it is important to understand that we have the following three degrees of freedom that can be played with:

- Switching from a compute-centric execution used by processors and GPU (akin Von Neumann architecture) to a data-centric paradigm in order to reduce the overhead introduced by the data movement.
- Changing operand precision or introducing multi-bits or analogue coding or other ways of encoding information (e.g. Quantum),
- Introducing new materials that will deliver more efficient ways (in terms of timing and/or energy) to store, switch and/or process information.

This gives a very broad set of options but only few will emerge due to economic constraints, critical mass issues, industrialisation aspects, legacy and usability problems. The following sections present some of the most promising paths even if most of the impact of these technologies will be visible after 2025.

6.2

Progress of current technologies

Even if the end of the Moore's law is approaching, there are still ways to improve the technologies used for current chips[54]. As the time range of this document is only up to 2025, we expect progress in the VLSI technology with technology nodes that will propose 7nm, 5nm and perhaps 3.5nm process. The transistors will be further optimised to reduce leakage, e.g. using nanowires or nanosheets. Besides these new technology nodes, the use of 2.5D, 3D (and monolithic) stacking will continue to expand. This will deliver improvements in communication between the different components of the architecture with advantages in term of latency, bandwidth and energy/transfer.

6.3

New architectures

Today, standard processors and GPU accelerators are based on a Von Neumann architecture where a controlled execution applies operations onto data that are stored in registers (fed by caches, fed by memory). This architecture is very flexible but can be costly in terms of transistors, data paths and energy compared to what is just needed for an application. It implies a lot of moves and duplications of data, which is not efficient (bringing data from external memory is three orders of magnitude more energy demanding than a floating-point operation on those data). There is a research path which proposes architectures that will be more efficient for some classes of problems. Some of these new architectures can be implemented using standard CMOS technology or providing opportunities to introduce new technologies that will be more efficient than CMOS (see next sections). Some concepts of new architectures are generic (see below dataflow or IMC) or target a specific class of algorithms (see below neuromorphic, graph and simulated annealing).

6.3.1

Dataflow

In dataflow architectures, data move between modules that perform the computation on the data. There is no program counter that controls the execution of the instructions such as in a Von Neumann architecture. Deep Learning architecture (see below: neuromorphic architecture) can be implemented as a specific dataflow architecture (the main operations are matrix based). The investigation of dataflow architectures is often linked to FPGA (Field Programmable Gate Array) as most of the ideas have not yet led to taping out specific circuits but have been tested and implemented with FPGAs.

6.3.2

IMC/PIM (In Memory Computing; Processing In Memory)

These architectures couple the storage with some computing capabilities. They are based on a premise that bringing the computation to the memory will be cheaper in resources than moving data to the computing units. These architectures are also related to the development of Non-Volatile Memory (see next section) and appealing as long as the cost of the in-memory computation is low (in terms of energy and chip volume). Nevertheless, as this approach needs a complete re-thinking of the applications, its emergence will take more time and is not expected in mainstream systems in the 2025 horizon.

54. https://www.anandtech.com/show/15217/intels-manufacturing-roadmap-from-2019-to-2029

6.3.3
Deep Learning and Neuromorphic

The development of AI and especially applications using Deep Learning techniques has led to a huge interest in specific architectures that are inspired by a theoretical, simplified model of a neuron. This kind of architecture can be used for AI tasks but it can also be viewed as a generic classification function or a function approximation. As an increasing number of applications (or parts of applications) are mapped onto this paradigm, it is worthwhile to develop specific circuits that implement only the operations and data paths mandatory for this architecture. Several examples already exist such as Google's TPU chip or the Habana AI chip. These efforts have not exploited all the possible options yet, e.g. more complete models of the neuron and using spiking encoding, they and have not developed the features needed to create potential architectures, so more efficient accelerator solutions are expected in the near future.

6.3.4
Graph computing

Graphs play an important role in data representation and in some AI or optimisation problems. As standard processors have poor performance due the non-regular access to data, developing a specific architecture can be relevant. Nevertheless, there are less efforts in this field and it is still uncertain whether accelerators of this kind will be developed.

6.3.5
Simulated annealing

Simulated annealing is a method to solve complex optimisation problems. It can be implemented by software on classical Von Neumann processors, but one can also design an ASIC that will significantly speed-up the computation by mapping directly the variables and their interactions and by providing a hardware based random generator. We expect to see more accelerators in this domain that will be useful for optimisation class of problems.

6.4
Integration of new technologies with CMOS

6.4.1
NVMs

Different technologies are being developed to propose Non-Volatile Memory. Besides the existing NAND, resistive memory (memristor), phase change memory (PCM), metal oxide resistive random-access memory (RRAM or ReRAM), conductive bridge random access memory (CBRAM) and Spin-transfer torque magnetic random access memory (STT-RAM) are interesting technologies, with the aim to reach the speed of RAM with persistence of data and high integration density. The developments in this domain (not all at the same level of maturity and readiness) may influence HPC in a number of ways. The energy to retrieve data is lower, the latency to read the data is reduced and the density can be increased (especially with solutions implementing multi-state storage for each cell). NVM also play a role in providing easy implementation of the IMC/PIM architecture when compute elements can be associated as in Memristive Computing. NVM technologies will continue to impact HPC system design in the SRA timeframe. For example, it could drastically change the current memory hierarchies or allow to move to key-data access types.

6.4.2
Silicon photonics

Silicon photonics can be used either to compute or to provide interconnection between computing elements. The properties of light can be used to perform computations. For example, the interaction of light rays the phase of which has been modulated according to inputs can produce operations over these inputs. This idea can be used to implement neuromorphic architectures where the main operation is a scalar product.

Photonics is already used for medium to long distance communication in HPC systems. The technology is also appealing for rack level communication. Perhaps the most interesting aspect will be at package level, with the development of either chip integrated photonics links or active interposer with embedded silicon photonics network between chips or chiplets. The bandwidth and the energy efficiency can be increased to surpass current CMOS solutions. These solutions are expected to be used in HPC systems before 2025.

6.5
New coding schemes

In the past, HPC architecture evolved to provide an increasing amount of precision for operations as HPC was focused on simulations where stability and convergence of algorithms were depending on this feature. Today, in the case of new applications coming mainly from neural networks, high precision is not mandatory (the "learning phase" of Deep Neural Networks can be done in most cases with 16-bit floating point operations and for inference, 4 up to 16 bits are usually sufficient) and switching to low precision can save space (i.e. transistors and data paths) and energy. This trend is already underway and we expect more HPC systems tailored to the precision needed by applications.

An even more radical switch is to use analogue computing (i.e. a physical (or chemical) process being used to perform calculation). We see the emergence of optical systems used to compute some functions due to light properties and optical devices such as lenses. This approach is an extremely energy efficient way compared to traditional computers and can offer massive parallelism. This technology cannot suit every application but a number of algorithms such as scalar products and convolution-like computations (e.g. FFT, derivatives and correlation pattern matching) are naturally compatible. Other options are possible such as using electrical or thermal systems to find solutions of

some differential equation problems. We expect these technologies to be used as accelerators even if their domain of application might be limited.

6.6

New technologies

CMOS has been such an industrial success story that it has reduced the effort on alternative solutions to implement transistors or computing elements. With the end of CMOS progress, more emphasis will be put on these other options, even if it is still to prove that they will be able to deliver more computing performance than CMOS. The industrialisation will also take time, so any impact is rather expected after 2025. No winning technology is clearly emerging at the moment.

6.6.1

Superconducting

With the use of superconducting material, based on the zero resistivity of the interconnects, the expectation is that power consumption could be up to two orders of magnitude lower than that of classical CMOS-based supercomputers. Nevertheless, superconducting circuits still need to overcome several drawbacks such as density, switching time, operating temperatures, interfacing with external systems or noise reduction to be seen as a potential solution for HPC. Most of the time, their implementation uses Josephson junctions and so it has the same disadvantages as analogue computing.

6.6.2

Memristive devices

Besides the uses of the resistive memory for NVM and analogue neuromorphic architectures (see the sections above), memristive devices can be interesting as a method of implementing logic gates and to compute. Even if the switching time may be slower than CMOS, they can provide a better energy efficiency. The integration of memory into logic allows to re-program the logic, providing low power reconfigurable components and can reduce energy and area constraints in principle due to the possibility of computing and storing in the same device (computing in memory). Memristive devices can also be arranged in parallel networks to enable massively parallel computing.

6.6.3

Other materials

There is some research being done on new materials that could lead to new ways to compute. To name some of those we have carbon nanotubes, graphene or diamond transistors and spintronic devices. Nevertheless, it is unlikely that this research will impact HPC during the SRA period.

6.6.4

Quantum computing

Quantum computing (QC) is a new paradigm where quantum properties are used to provide a system with computing capacity. Breakthroughs are required and underway to achieve stable devices with long entanglement time and minimum error correction. A lot of basic research in Europe in various areas related to quantum computing (e.g. optics, gate-controlled quantum dots circuits or approaches in trapped ions) is currently being carried out. Also, simulators are available such as e.g. ATOS QLM to teach to "think quantum" and program future quantum computers.

Nevertheless, before 2025, we do not expect the emergence of solutions that will be useable in the HPC context. QC will most probably be implemented as accelerators for dedicated workloads in chemistry and optimisation first.

6.7

Summary

There are plenty of research approaches potentially capable of delivering more processing capability per energy unit and to meet the users' permanent need for more powerful HPC systems. Nevertheless, the HPC market by itself is not large enough to justify the huge investments that will be needed to validate and industrialise some of these technologies. This is why we will only see the short or mid-term emergence of solutions that are either cheap to develop or adaptations of technologies developed for bigger markets (AI, powerful Edge systems and perhaps some new best-of-class applications not foreseen yet).

The main trend for the period covered by this SRA will be the emergence of more types of accelerators. Besides today's GPUs, Deep Learning and neuromorphic accelerators of different technologies will appear in HPC systems, from standard architectures to analogue systems. Accelerators for dataflow processing will also be proposed either as FPGAs or as ASICs and perhaps for simulated annealing or graph computing. The flexibility of FPGAs to map computation "in space" and the regularity of their architecture might be an enabler to validate more specific architectures. The integration of these accelerators will benefit from the 2.5/3D integration processes and will be facilitated by the adoption of chip interconnection standards. Solutions having part of the computation done "in" or "near" memory will emerge.

We also expect the emergence of photonics and memristive technologies integrated with CMOS. Silicon photonics could potentially be used for chip-to-chip connections to improve latency, bandwidth and energy. Accelerators could also be based on a mix of photonics and CMOS. Memristive technologies could also be integrated in HPC systems and change the way data and computing are organised today.

In summary, over the next 5 years, we predict an evolution of the HPC systems pushed by accelerators and integration of new technologies. More drastic changes can be anticipated for the 2030 HPC systems. To focus on the more promising paths for HPC, the research on these new computing solutions must adopt a co-design approach. ■

7

Technology sourcing – from chips to system software

While the above chapters focused on "WHAT" is important in the coming years in the context of HPC related research, this chapter tackles the implementation aspects for the technology discussed – i.e. the "HOW" of this effort.

HPC/Supercomputing is critical for the European economy and society and the well-being of its people regarding all aspects of life. Therefore, Europe should pursue a comprehensive, coherent, well-coordinated and holistic approach to future HPC/Supercomputing towards Exascale and beyond, including a robust roadmap and plan for hardware and software technologies and products, applications and infrastructure, together with the development of a vibrant ecosystem. It should stimulate and support close cooperation and partnership between science, research, academics and industry to facilitate strong and synergetic co-design approaches. It must also foster balanced mutual international collaboration with (non-European) partners and entities regarding respective technologies. The guiding principles should be based sustainable high performance and high energy efficiency.

7.1

Open source vs. proprietary sourcing

7.1.1

Motivation

The main motivation for the use of open source is reduced costs, increased independence and control, increased flexibility, better products, faster innovations, faster and better uptake and access by users and industry and building communities with "crowd intelligence".

Several examples exist where a mix of open source and closed source (proprietary) have provided good solutions. In addition, there are often pros and cons for the choice of open source solutions. A decision on whether to use open source elements should take into account a wide range of factors.

7.1.2

Definitions[55]

The term "Open Source" refers to solutions that can modified and shared because their design and implementation are publicly accessible. Open Source projects, products or initiatives embrace and celebrate principles of open exchange, collaborative participation, transparency, meritocracy, and community-oriented development.

*"**Open Source Software (OSS)** refers to software that is designed, developed, tested, and can be inspected, modified or enhanced through public collaboration, and distributed with the idea that it may be shared with others, ensuring an open (future) collaboration."*

*"**Open Source Hardware (OSH)** refers to the design specifications of a physical object (e.g. electronic or computer hardware) which are licensed in such a way that said object can be built from design information, created, studied, modified, improved and distributed by anyone. Such information, made available for public use, can include documentation, schematic diagrams, construction details, parts lists and logic designs."*

Software/Hardware designs and inventions are subject to copyright, patent and trademark law. Open Source uses these intellectual property laws creatively to make the respective designs publicly accessible.

7.1.3

Different attributes

The choice for technology implementation via open source or commercial/proprietary solutions is complicated and it depends on many aspects. The situation is not "black-and-white", neither is it a case of "one size fits all"! The attributes detailed in the table on the next page characterise some of the key considerations.

7.1.4

Discussion

An example of the potential outcome of the aforementioned activities could be the design, development and implementation of a federated and orchestrated "European HPC Solution Stack" based on open and standardised interfaces, which would allow the interoperability of OSH/OSS and commercial offerings which would run on European HPC platforms and infrastructures, supported by its European HPC ecosystem. The result could be a mixed model of open source and proprietary products, which provides a complete, functioning and performing solution stack for its HPC users, applications and workloads.

Commercial success and sustainability of the European-based HPC solution stack are important and closely related goals. An open source policy is not a pre-requisite for achieving these goals and the benefits need to be established on a case-by-case analysis. It is more crucial to ensure the interoperability of the components of the European HPC solution stack, supported by a broad ecosystem which involves all key, active players in Europe, which requires open and standardised interfaces. The result could be a mix of open source and proprietary solutions providing a complete, functioning and performing solution stack for its HPC users, applications and workloads.

55. Further material at: opensource.com/resources/what-open-source and opensource.com/resources/what-open-hardware

Open Source	Proprietary
Product Offering & Support	
Effort of a community or a single organisation with a public offering, but no commitment to provide support (unless there is an SLA for paid added-value service). Specifically, for industry there are often no liabilities, no guarantees, no accountability	Specific vendor commercial offerings and support; vendor has multiple obligations
Costs & Dependency	
It is affordable with higher level of independence	It could be affordable… but completely dependent on the specific vendor
Flexibility & Extensibility	
Full public community control and flexibility, but potential danger of fragmentation and bifurcation of versions and features and possibly lacking support of extensions	Specific proprietary vendor control, only vendor specific flexibility, but normally standardised versions only
Examples	
RISC-V, OCP[56] are 2 approaches for OSH, Linux for OSS, gcc, LLVM and GNU libc;	e.g. x86/Windows, MacOS are proprietary, widely used vendor specific compilers and libraries
Business & Support Model	
Main business via paid value-add services and components, enhancements or support. Support may also come from community volunteers.	Commercial sale, copyrighted and licensed, source is not available commercial service and support or troubleshooting depending on SLA
Improvement Model	
Community public enhancements or value-add for charge	Vendor specific (proprietary) enhancements for charge might be possible, otherwise depending on specific vendor product strategy
Benefits & Usage	
Helps educating, training and developing skills for software/hardware development.	Focused on using ready-to-go software/hardware for users
Security	
Potentially more secure because everyone can publicly see, modify and correct it	Need to trust the vendor/provider
Liability, Responsibility & Accountability	
Often unclear to not present, in particular in the case of absent commercially backed services	Provided by the vendor
Stability & Reliability	
Potentially more stable/reliable and adhering to open standards due to public community	Need to rely fully on the vendor/provider
Sustainability	
Long-term sustainability possible due to open access for the community but project dependent and not guaranteed	Fully controlled by the producer: Solution may be sustainable in case of commercial success, but solution may stop being available and accessible when producer goes out of business
Monetary Investment	
Involves sometimes monetary investment (mainly labour costs and some materials), might need additional funding for needed support of production environments. OSH involves (almost) always monetary investment for physical materials, normally not financially feasible for OSH projects' physical components to be offered for free	Involves always monetary investment, it is paid for.
Licensing	
Free and open source licensing Free of charge? No, but in many cases, it is free of charge, which generates a (sub-) ecosystem and community	Commercial licensing, for charge

7.2

European vs. global sourcing

ETP4HPC welcomes and supports initiatives for the European HPC technology ecosystem as part of the European digital autonomy strategy. To reduce the European dependencies on technologies and solutions that are developed and created outside of Europe while making a European-based HPC solution stack commercially successful and sustainable, a careful balance between autonomy and involvement in internationally developed technologies and solutions needs to be achieved. Autonomy is particularly relevant in the context of export-controlled technologies, while dependencies could also be reduced through suitable global sourcing strategies.

These generic considerations can be exemplified by the case of the "European Processor Initiative (EPI)[57], which will be a key element of the European-based HPC solution stack as it concerns export-controlled technologies. To maximise the impact of this investment and to make the effort sustainable and commercially successful, EPI needs to open up and enable cooperation with European and global technology providers.

The proposed introduction of the "EPI Common Platform" could be an opportunity for enabling such a cooperation. The platform concept is based on EPI's integration approach where multiple chiplets are based on an interposer. Some part of the chiplets could come from technology providers outside of EPI and provide e.g. specific accelerator capabilities. Application-specific accelerators have become an interesting option in the market of Deep Learning and artificial intelligence solutions. Open and standardised interfaces are a key pre-requisite to allow various European and global technology providers to make investments in innovative technologies to be integrated in customised versions with the EPI processor.

One step further, on a system level, collaboration between European and international players can be foreseen on a much larger scale when it comes to creating a complete ecosystem as a basis for the above mentioned "European HPC solution stack".

The main motivation for any level of collaboration between EPI and projects outside the EPI consortium should be to generate sustained business value based on dependable and stable long-term technology, tools and support roadmaps. ■

56. https://www.opencompute.org
57. https://www.european-processor-initiative.eu/

8 The importance of ethics

9 Building and retaining skills and competence

The tremendous efforts in research and development of AI technologies and solutions and the resulting rapid development of AI capabilities has been one of the multiple factors leading to an increased awareness of ethical topics in the context of the development of IT technologies. Policy makers react to this by creating new standards that need to be respected. The EU's General Data Protection Regulation (GDPR) is a very prominent example.

Ethical topics need to be considered in many different dimensions including: foresight ethics (identifying the potential impact of tools and methods); governance ethics (how to ethically manage the governance within projects and businesses); stakeholder ethics (the ethical approach to engagement with the wider stakeholder community); data ethics (in case of personal data); testing ethics (in case of research involving animals/humans); dual-use of HPC technologies (for civil and for military purposes). In all these areas, reflections on limits and scope of use are important. Ethical topics impact the SRA in different ways. Ethical aspects need to be anticipated for any development of technology. Openly addressing such topics will also help to improve acceptance of new technologies amid significant scepticism as regards such new technologies, e.g. in the area of AI. While ethical considerations may on the one hand limit research and development, the latter can also help to address ethical topics. For instance, new technologies may allow for better handling of personal data and help to improve protection of data. ■

Training and education are crucial for the European human capital in HPC. Without building the necessary skills in both, using and developing HPC technologies and applications, Europe will not be able to capitalise on its investment in HPC. Building these skills needs to happen at all stages: from training established HPC practitioners in novel technologies, training existing IT and computational science people in HPC to educating the next generation of users as well as application and technology developers.

Education in HPC technologies and applications needs to start early in the education process, ideally already at high-school level. HPC needs to move out from its current niche of being considered a very specialised technology, to mainstream education, in order to ensure the necessary skills are available to use and develop HPC technologies in all domains and at the required breadth. Currently, HPC is only considered as a highly specialised subject, while it will be required to include HPC knowledge in all science education and increasingly also in the humanities. Appropriate curricula in HPC, data science, and related computational science subjects need to be developed and introduced across the European higher education sector.

A special emphasis needs to be put on gender aspects to ensure a balanced supply of talent from the ground up. When it comes to gender balance, HPC is currently a domain where women are highly underrepresented, and this can only change if perceived barriers are removed early in the education process. Presenting success stories of females in HPC and show-casing role-models can help overcoming these barriers[58].

In addition to education, constant training efforts are needed to provide established HPC practitioners with the necessary skills in new technologies as well as attracting new talent to HPC. As with education, training is needed in both HPC usage as well as application and technology development. Particularly when new technologies are being introduced (such as GPUs, NVRAM, etc.) there is a tremendous training need among application developers so they can port and optimise their codes to the new hardware.

Currently, much of this training is provided by academia, via universities or HPC centres, e.g. through PRACE or the HPC Centres of Excellence. However, to enable more widespread use of HPC in industry and SMEs, these will require dedicated training. Particularly, SMEs need a dedicated effort as in the majority of the cases they cannot afford to have their developers spend significant time on learning on their own how to develop and use efficient codes for a (heterogeneous) HPC machine. Moreover, training is expected to be vital in the future since, based on the diversity of the future HPC systems, it is very likely that each machine (or set of machines) will require different development approaches (even through a common development environment) so as to expose their full capabilities. Additional factors increasing the attractiveness of HPC to SMEs should also be considered, such as training courses tailored in local language and customised to specific application areas (e.g. civil engineering, mechanical manufacturing, etc.), online courses, and support offerings such as staff exchange programs. The upcoming HPC Competence Centres are expected to play a pivotal role here, building on the expertise available in academia.

In addition to training and education measures, retaining skilled personnel is of high importance for the European HPC ecosystem. Highly skilled IT personnel are needed in many sectors and there is a high risk that excellent HPC personnel moves elsewhere. This is a particular challenge for Universities, where HPC personnel often does not follow the classical tenure track and thus has little to no opportunities for their career development. Universities need to define appropriate career paths for those people that gives enough recognition to be able to retain them. ■

58. https://womeninhpc.org/

10

Operational recommendations

While the chapters above provide a detailed overview of the research priorities, this chapter provides recommendations for the implementation of these priorities in the form of suggestions for R&I structures and instruments.

Within H2020, most HPC calls were either launched within the Future Emerging Technologies (FET) section of the "Excellent in Science" pillar or as part of the ICT work programmes, which belongs to the "Industrial Leadership" pillar. The usual duration of projects was 36 months (with some projects extended to 48 months) and the projects were single events. The possibility of continuing a project in any subsequent calls depended on how they met the requirements of those calls. Thus, the overall process was a bottom-up approach resulting in over 48 projects with individual results and very limited collaboration between projects. The overriding objective of the H2020 FETHPC and related preceding work programmes was to achieve the capability to develop Exascale systems by 2023, which still acts as an appropriate qualitative and quantitative objective for this effort.

The following two sections are intended to make a few operational recommendations to EuroHPC on various implementation options for future work programmes.

10.1

Research projects implementation options

The R&I processes deployed in industries and some national R&D initiatives[59] could mirror the following structured end-to-end approach. It contains a staged sequence of projects with a specific end-goal and intermediate checkpoints governed by sets of KPIs (Figure 18):

Mapped onto the next 7 years Multi-annual Financial Framework, Horizon Europe, this approach could work as follows:

1. Proof-of-concept projects

- Their scope is to validate high potential proposals, quantify their technical and business potential and risks, and engage critical skills, partners and stakeholders. Also, this early validation of new concepts is the main intent of the 'working agile' style of conducting projects in many R&D organisations.
- A potentially large number of small (budget) projects with a relative short duration (ca. 12 months).
- The expected Technology Readiness Level (TRL) should be low to medium.

All projects will be evaluated against the first set of KPIs and only those (few) passing the test will have a chance to enter the next stage:

2. Research and Innovation projects

- Their scope is to produce solutions to specific problems, generate IP and demonstrate the solution on in a well-defined, limited context. As a reference, the R&I scope and expectations would be similar to the ones used to select projects within the H2020 FET-HPC work programmes launched in 2014/15, 2016 and 2017.
- A medium number of medium-size (budget) projects with a duration long enough to produce tangible results.
- In most cases medium and in some cases high TRL. While most projects will only be able to reach a medium TRL, specific research projects (most likely in the software area) solving important problems can be expected to exit with a high TRL. Wherever advisable, projects need to be set up with a high degree of collaboration to avoid the generation of isolated "result-islands".

All projects will be evaluated against the second set of KPIs focussing on the technical feasibility, robustness, efficiency and usability will have a chance to enter the next stage:

3. Large scale integration/demonstration projects

- The scope is the integration of high TRL technology (hardware to software) into complete solutions in a pre-commercial, problem solving and application-driven environment.
- The number of these projects is low, the KPI-driven entry criteria are very demanding, the funding envelope is larger than in the case of the other two project types.
- A high TRL is outcome of such integration projects. A limited effort to bring up the TRL level of sub-components to meet the required level should be implemented.

4. Incubator projects

- Independently of the cascade of three project types mentioned above, the forming of start-up companies introducing new technology, methods, tools or business ideas involving HPC should be supported in the same way as strengthening "young" SMEs in getting their businesses off the ground.

The EC has instruments in place[60] which could be utilised or adopted.

This staged approach from level 1 to 3 in Figure 18 applies mainly to projects pursuing new ideas or concepts with no preceding work in earlier projects. Given that there are numerous project results available today, future R&D projects using any of these results (potentially targeting a higher TRL level) could enter at level 2 or 3 of Figure 18, depending on the alignment with the entry criteria.

59. https://www.entreprises.gouv.fr/innovation-2030/the-challenge-the-3-phases?language=en-gb
60. https://ec.europa.eu/easme/en

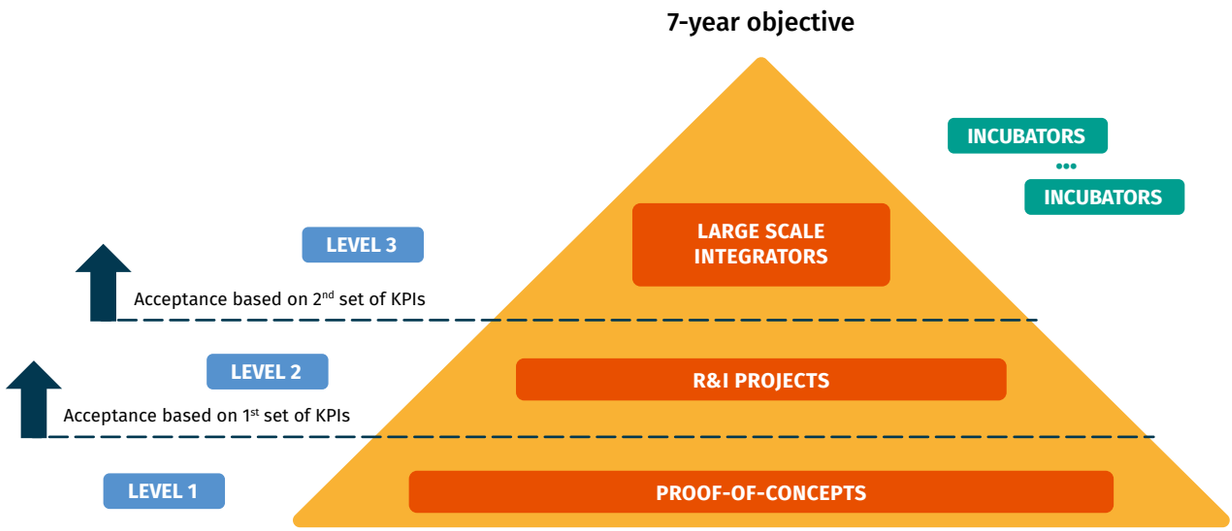


Figure 18: Conceptual structure of project types flow

10.2

Managing the next 7-year work plan, work programmes and calls

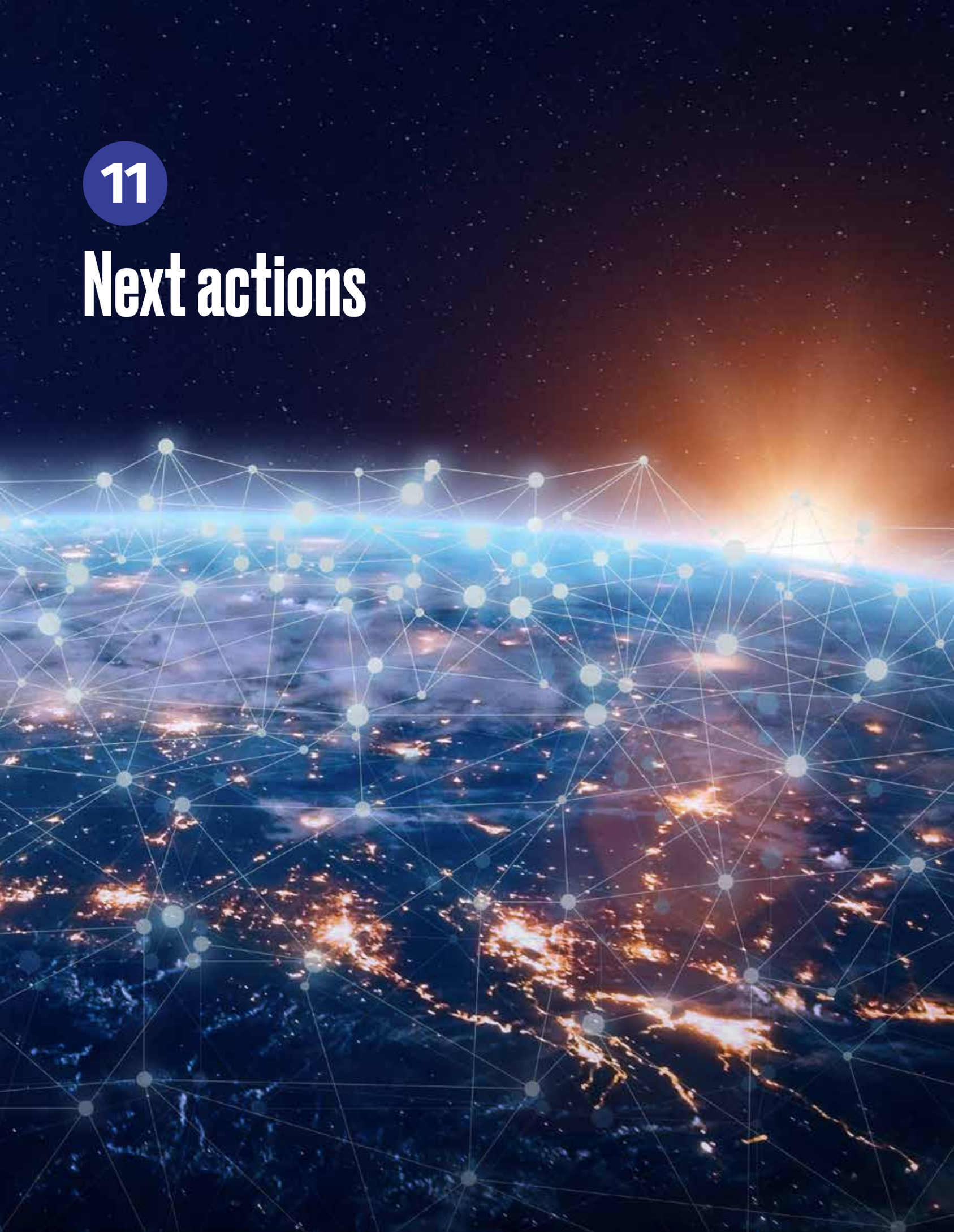
As pointed out in the previous section, a different approach would be required to increase the effectiveness the work programmes and calls. Instead of the bottom-up method, a programmatic, top-down, active program-management should be instantiated:

- First, a new long-term objective for HPC related R&D in the years following the availability of Exascale systems is required. What is the one, major goal all work programmes and calls should be aligned to? The theme of "HPC in the Digital Continuum" can be helpful in formulating an ambitious target, in both a qualitative and quantitative dimension.
- Second, the full spectrum of projects outlined above should be exploited, including a strict execution of the stages outlined in Figure 18.
- Active management of the entire portfolio-coverage of R&D topics and priorities should be enforced. In case certain areas of the priorities do not find enough attention and interest, these "white spots" must be taken care of. There are multiple options for "filling the holes" (e.g. through special incentives or relaxed acceptance criteria for projects).
- Facilitate the transition to commercialisation. Most European HPC hardware or software vendors in Europe are SMEs. In line with EuroHPC's objective to strengthen the indigenous industry, supporting start-ups and SME should be of crucial importance. It is needs to be carefully evaluated whether existing instruments could be utilised or new ones (tailored to the growing needs of HPC functionality in the Digital Continuum) should be put in place.

The suggestions made in this chapter should be considered as a possible implementation scenarios that would address the fragmentation seen in the past and achieve the integration needed to establish the European HPC stack mentioned in chapter *7 Technology sourcing – from chips to system software* on page 86. ■

11

Next actions



ETP4HPC intends to engage with key technology stakeholders in a horizontal collaborative effort in order to implement the idea of synchronised research calls from 2021 onward (the term "synchronised calls" means that each stakeholder funds its contributions separately from the other stakeholders and there is no merged (joint) overall funding mechanism). The stakeholders of this engagement are the providers of key co-technologies required to implement complete end-to-end solutions in the Digital Continuum. Among the players should be:

- **BDVA**, Big Data Value Association
- **AIOTI**, the Alliance for Internet- of-Things
- **EU Robotics**, the association for all stakeholders in European robotics
- **ECSO**, the European Cyber Security Organisation
- **5G IA**, the 5G Infrastructure Association

An important aspect of such a large horizontal action is the definition of the set of problems to be solved. As shown in Figure 2, Horizon Europe missions[61] offer a wealth of options for connecting societal needs to such a collaborative effort.

In addition, the new European Commission has recently announced its guidelines for 2019 to 2024 with "European Green Deal"[62] as a key priority. This includes a broad range of technical challenges, which could well be addressed by a collaborative R&D effort as described in the next section.

11.1

A large-scale collaborative effort: Transcontinuum Extreme-Scale Infrastructures

Recent hardware and software advances have motivated the development of a *transcontinuum digital infrastructures* concept to account for the convergence of data and compute capabilities. This concept is not in a straight line from the past efforts and a paradigm change is needed: we will have to design systems encompassing hundreds of billions of cores distributed over scientific instruments, IoT, supercomputers and Cloud systems through LAN, WLAN and 5G networks.

Pushed by massive deployments of compute and storage capabilities at the *Edge*, we require new system design to accommodate the ecosystem change to be expected in the coming decades (environmental and technological) and horizontally integrate the different actors. The new demands and challenges that combine data and compute, distributed across the continuum, and the maintenance and resource efficiencies, are pushing for drastically increased software and hardware *sustainability*. Furthermore, the need to provide high-level *cybersecurity* is profoundly changing the game. Efficiency and resilience will have to reach levels never achieved so far, while taking into account the intrinsic distributed and heterogeneous nature of the continuum. In addition, the question of dealing with such high volumes of data needs to be faced, and quality versus quantity will become unavoidable. These considerations will spread over all components. Long-lifetime hardware devices will have to be reconfigurable, modular, and self-aware in order to be operational over extended periods. Algorithm efficiency will need to be drastically pushed up (e.g. more efficient AI). Management and deployment of large-scale application workflows will have to be adapted or invented. Network protocols will have to offer better control over the data logistics, etc.

Furthermore, it is widely recognised that AI will play a central role in these extreme-scale, continuum infrastructures. This will occur at three levels:

- AI for Digital Infrastructure,
- Digital Infrastructure for AI, and
- AI for Science, Industry and Societal Challenges.

The first addresses how AI can pilot and monitor the continuum and in so doing provide solutions to the points listed in the previous paragraph. The second treats the question of re-designing the e-infrastructure to efficiently deal with data analysis and machine learning, which means tuning of data access, I/O, and low precision arithmetic. The last deals with the ever-increasing needs to exploit AI techniques for extreme-scale, combining Data and Compute through the interpretation and coupling of computing results, measurements and observations (e.g. Digital Twins in extreme earth modelling, combining climate models with satellite data and on-ground sensors).

The overall objective is to target high TRL solutions (7 and more), based on horizontal synergies between all the concerned digital infrastructure technologies: HPC, Big Data, Machine Learning, IoT, 5G, cybersecurity, processor technology (EPI) and robotics. All of these components of the digital infrastructure will *together* be able to address the critical societal challenges and sustainable development goals by mobilising their amazing potential all the way across the continuum. ■

61. https://ec.europa.eu/info/news/commission-announces-top-experts-shape-horizon-europe-missions-2019-jul-30_en
62. https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf

12

Conclusions and Outlook

13

Appendix

The deployment of "High-Performance Computing" is undergoing a significant change and the term "HPC" no longer ap-plies to only supercomputers in large data centres but also to a compute infrastructure supporting simulation, modelling and data analysis in the digital computing continuum, as outlined in the chapter on "Deployment structures". Furthermore, core HPC technologies and methodologies are being used to enable concurrent processing to permeate all levels of that digital computing continuum.

Research on both HPC applications and HPC technology will expand from the current fields which deploy HPC solutions to adjacent fields, in order to address AI, Data Analytics and IoT-related challenges. This has influenced the selection and definition of the research priorities in this SRA. A truly interdisciplinary effort is required to address this paradigm shift. Fostering collaborative cross-domain research involving the technology providing stakeholders in Europe should be among the highest priorities for the period 2021 – 2024. This is a unique opportunity for Europe to take the lead! ■

13.1

Glossary

ACID: Atomicity, Consistency, Isolation, Durability

AI: Artificial Intelligence

AIOTI: Alliance for the Internet of Things Innovation

AISBL: Association Internationale Sans But Lucratif (International Non-for-Profit Association)

ALU: Arithmetic Logic Unit

AMBA: Advanced Microcontroller Bus Architecture

API: Application Programming Interface

AQMO: Air Quality and Mobility Project

ASIC: Application-Specific Integrated Circuit

AWS: Amazon Web Services

BD: Big Data

BDA: Big Data Analytics

BDC: Backup Domain Controller

BDEC: Big Data and Extreme-Scale Computing

BDV: Big Data Value

BDVA: Big Data Value Association

BSC: Barcelona Supercomputing Center

CBRAM: Conductive Bridge RAM (Random Access Memory)

CCIX: Cache Coherent Interconnect for Accelerators

CEA: Commissariat à l'énergie atomique et aux énergies alternatives

CERN: Conseil européen pour la recherche nucléaire

CGRA: Coarse Grain Reconfigurable Arrays or Coarse-Grained Reconfigurable Architectures

ChEESE: Center of Excellence In Solid Earth

CIFAR: Canadian Institute For Advanced Research

CMOS: Complementary Metal-Oxide-Semiconductor

CoE: Centre of Excellence (for Computing Applications)

cPPP: contractual Public-Private Partnership

CPS: Cyber- Physical System

CPU: Central Processing Unit

CSA: Configurable Spatial Accelerator

CSA: Coordination and Support Action

CSP: Cloud Service Providers

CXL: Compute Express Link

D: Deliverable

DB: Database

DC: Direct Current

APPENDIX

DCP: Digital Continuum Platform

DDR: Double Data Rate

DG: Directorate General

DL: Deep Learning

DOE: Department of Energy

DoW: Description of Work

DPU: Data Processing Unit

DSL: Domain Specific Language

E2E: end-to-end

EC: European Commission

EC: European Commission

ECA: electric, connected, autonomous/automated

ECC: Error-correcting code

ECG: electrocardiogram

ECMWF: European Centre for Medium-range Weather Forecasts

ECSO: European Organisation for Cyber Security

EESI: European Exascale Software Initiative

EIT: European Institute of Innovation & Technology

EMIB: Embedded Multi-Die Interconnect Bridge

ENES: European Network for Earth System modelling

EOSC: European Open Science Cloud

EPI: European Processor Initiative

EPOS: European Plate Observing System

EsD: Extreme-Scale Demonstrators

EU: European Union

EXDCI: European Extreme Data and Computing Initiative

FaaS: Function-as-a-Service

FET: Future and Emerging Technologies

FFT: Fast Fourier Transformation

FLOP: floating point operations

FORTH-ICS: Foundation for Research and Technology - Hellas

FP: Floating Point

FP: Framework Programme

FPGA: Field Programmable Gate Array

FTRT: Faster Than Real Time

FUSE: Filesystem in Userspace

GDP: Growth Domestic Product

GDPR: (EU) General Data Protection Regulation

GENCI: Grand équipement national de calcul intensif

GNU: GNU's Not Unix

GPFS: General Parallel File System

GPGPU: general-purpose GPU

GPU: Graphics Processing Unit

H2020: Horizon 2020 - The EC Research and Innovation Programme in Europe

HBM: High-Bandwidth Memory

HDD: Hard Disk Drive

HDR: Enhanced Data Rate

HiPEAC: High Performance Embedded Architectures and Compilers

HMC: Hybrid Memory Cube

HPC: High-Performance Computing

HPCG: High-Performance Conjugate Gradients

HPDA: High-Performance Data Analytics

HTC: High-Throughput Computing

HW: hardware

I/O: Input/Output

ICT: Information and communications technology

IDC: International Data Corporation

IESP: International Exascale Software Project

IIoT: Industrial Internet of Things

IMC: In-Memory Computing

IMC/PIM: In Memory Computing; Processor In Memory

INVG: Istituto Nazionale di Geofisica e Vulcanologia (Italian National Institute of Geophysics and Volcanology)

IOC/UNESCO: Intergovernmental Oceanographic Commission of UNESCO (The United Nations Educational, Scientific and Cultural Organization)

IoT: Internet of Things

IoV: Internet of Vehicles

IRISA: Research Institute of Computer Science and Random Systems

ISA: Instruction Set Architecture

ISV: Independent Software Vendor

IT: Information Technology

JGU: Johannes Gutenberg-Universität Mainz

JU: Joint Undertaking

KPI: Key Performance Indicator

KTH: Kungliga Tekniska högskolan

LAN: Local Area Network

LDAP: Lightweight Directory Access Protocol

LDAP: Lightweight Directory Access Protocol

LHC: Large Hadron Collider

LIGO: Laser Interferometer Gravitational-Wave Observatory

LLVM: Low Level Virtual Machine (please note this acronym has officially been removed to avoid confusion)

M: Month

MB: Mega Byte

MC/PIM: In Memory Computing; Processing In Memory

MFF: Multiannual Financial Framework

MIMD: Multiple Instruction, Multiple Data

ML: Machine Learning

MPI: Message Passing Interface

MRAM: Magnetic RAM

MSA: Modular Supercomputing Architecture

MW: megawatt

NAG: Numerical Algorithms Group

NAND: NOT-AND

NGI: Next Generation Internet

NIC: Network Interface Controller

NN: Neural Network

NOSQL: non SQL

NUDT: National University of Defense Technology

NV-DIMM: non-volatile dual in-line memory module

Nvlink: a wire-based communications protocol serial multi-lane near-range communication link developed by Nvidia

NVM: Non-Volatile Memory

NVMe: NVM Express

NVMeoF: NVMe over Fabrics

OCP: Open Compute Project

APPENDIX

OPA: Omni-path Architecture

OpenACC: Open Accelerators

OpenCAPI: Open Coherent Accelerator Processor Interface

OpenMP: Open Multi-Processing

OS: Operating System

OSH: Open Source Hardware

OSS: Open Source Software

OTA: over-the-air

OxRAM: Oxide-Based Resistive Memory

PaaS: Platform-as-a-Service

PB: petabyte

PCIe: Peripheral Component Interconnect Express

PCM: Phase Change Memory,

PCRAM: Phase Change RAM

PGA: Program Global Area

PII: Personally Identifiable Information

PM: Person Month

PMIx: Process Management for Exascale environments

POSIT: a hardware friendly version of Unums

POSIX: Portable Operating System Interface

PRACE: Partnership for Advanced Computing in Europe

PUE: Power Usage Effectiveness

Q: Quarter

QLM: Quantum Learning Machine

QoS: Quality of Service

R&D: Research and Development

R&I: Research and Innovation

RAM: Random-Access Memory

RAS: Reliability, availability and serviceability

RDMA: Remote Direct Memory Access

RFP: Request for Proposal

RIAG: Research and Innovation Advisory Group

RISC-V: Reduced Instruction Set Computer - five

RNA computing : RiboNucleic Acid computing

RoCE: RDMA over Converged Ethernet

ROI: Return On Investment

RRAM or ReRAM: metal oxide Resistive Random-Access Memory

SaaS: Software-as-a-Service

SC: Supercomputing Conference

SCM: Storage Class Memory

SDG: sustainable development goals

SerDes: Serialiser/Deserialiser

SHAPE: SME HPC Adoption Programme in Europe

SHS: Social and Historical Sciences

SICOS: SImulation, COmputing and Storage

SIMD: single instruction, multiple data

SINTEF: Stiftelsen for industriell og teknisk forskning

SiPh: Silicon Photonics

SKA: Square Kilometre Array

SLA: Service Level Agreement

SME: Small and Medium-sized Enterprise

SQL: Structured Query Language

SRA: Strategic Research Agenda

SSD: Solid-State Drive

SSH: Secure Shell

STT-RAM: Spin-transfer torque magnetic random-access memory

SW: software

TCO: Total Cost of Ownership

TOPS: tera operations per second

TPU: tensor processing unit

TRL: Technology Readiness Level

TSMC: Taiwan Semiconductor Manufacturing Company

TSP: Tsunami Service Providers

UAV: Unmanned Aerial Vehicle

UI: User Interface

UMA: University of Málaga

UN: United Nations

Unums: Universal Numbers

UPS: Uninterrupted Power Supply

US: The United States of America

V2E: vehicle-to-everything

VCSEL: Vertical-Cavity Surface-Emitting Laser

VPU: Vision Processing Unit

vs.: versus

WG: Working Group

WLAN: Wireless LAN

WP: Work Package

XDR: External Data Representation

YE: Year End

ZB: zettabyte

13.2

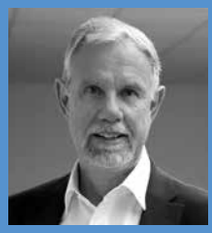
Acknowledgements

ETP4HPC would like to express its gratitude to everyone involved in the writing of this Strategic Research Agenda. This document is the collective work of ETP4HPC Members and experts representing other organisations and projects, whose input and advice allow this document to tackle all components of the Digital Continuum. The writing of this SRA was supervised by our SRA editorial team: Michael Malms (IBM & ETP4HPC), Marcin Ostasz (ETP4HPC), Maike Gilliot (TERATEC & ETP4HPC) and Pascale Bernier-Bruna (Atos & ETP4HPC). Special thanks are in order for the Leaders of the Working Groups of this SRA as most of the work that built this document has taken place within the following teams:

- System Architecture led by Laurent Cargemel (*Atos*), Estela Suarez (*Juelich Supercomputing Centre*) and Herbert Cornelius (*Megware*)
- System Hardware Components led by Marc Duranton (*CEA & HiPEAC*) and Benny Koren (*Mellanox*)
- System Software & Management led by Pascale Rosse-Laurent (*Atos*), María S. Pérez-Hernández (*Universidad Politécnica de Madrid & BDVA*) and Manolis Marazakis (*FORTH-ICS*)
- Programming Environment led by Guy Lonsdale (*Scapos*), Paul Carpenter (*BSC*) and Gabriel Antoniu (*Inria & BDVA*)
- I/O & Storage led by Sai Narasimhamurthy (*Seagate*) and André Brinkman (*Universität Mainz – JGU*)
- Mathematical Methods & Algorithms led by Dirk Pleiter (*Juelich SC*) and Adrian Tate (*Cray - NAG*)
- Centre-to-Edge Framework led by Jens Krueger (*Fraunhofer*) and Hans-Christian Hoppe (*Intel*)
- Application co-design led by Erwin Laure (*KTH*) and Andreas Wierse (*SICOS*)


Gabriel Antoniu


André Brinkman


Herbert Cornelius


Marc Duranton


Jens Krueger


Erwin Laure


Sai Narasimhamurthy


María Pérez-Hernández


Estela Suarez


Adrian Tate


Laurent Cargemel


Paul Carpenter


Hans-Christian Hoppe


Benny Koren


Guy Lonsdale


Manolis Marazakis


Dirk Pleiter


Pascale Rosse-Laurent


Andreas Wierse

We also owe a special mention to the following external experts, whose input, comments and reviews represented a great value in the writing of this SRA:

- Cristiano Malossi, *IBM*
- Francois Bodin, *IRISA*
- Jean-Francois Lavignon, *TECHNOLOGY-STRATEGY*
- Jean-Philippe Nominé, *CEA (ETP4HPC)*
- Mark Asch, *U-PICARDIE (BDEC-2)*
- Ovidiu Vermesan, *SINTEF (AIOTI)*
- Peter Bauer, *ECWMF*
- Stephane Requena, *GENCI*

The core and heart of this SRA, the technical contents of the chapters of the Working Groups, was collaboratively written by the members of ETP4HPC. We would like to express our gratitude to those members of ETP4HPC who took part in the arduous process of defining the respective research priorities within our SRA Working Groups. ■

ETP 4 HPC

EUROPEAN TECHNOLOGY PLATFORM FOR HIGH PERFORMANCE COMPUTING

e dci
European Extreme Data & Computing Initiative

Contact ETP4HPC
contact@etp4hpc.eu
www.etp4hpc.eu

@etp4h

Text
ETP4HPC

ETP4HPC Chairman
Jean-Pierre Panziera

Editorial Team
Michael Malms (Coordinator)
Marcin Ostasz
Maike Gilliot
Pascale Bernier-Bruna

Graphic design and layout
Antoine Maiffret (www.maiffret.net)

Printed and bound by
SNEL. Belgium

March 2020

The EXDCI-2 project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 800957.

Contact ETP4HPC
contact@etp4hpc.eu
www.etp4hpc.eu

@etp4h