A blueprint for the new Strategic Research Agenda for High Performance Computing

April 2019

Executive Summary

The objective of this document is to answer two key questions that arose as the European High-Performance Computing (HPC) ecosystem approached the end of the Horizon 2020 research programme, i.e.:

"What will be the role of High-Performance Computing in the 2021-2027 period?"

"How should one derive the main HPC research priorities for the next framework programme?"

The Authors intend to sketch the *big picture* of the major trends in the deployment of HPC and HPDA methods and systems, driven by the economic and societal needs of Europe, taking into account the changes expected in the underlying technologies and the overall architecture of the expanding underlying IT infrastructure.

Within the framework of the next long-term EU budget for 2021-2027, the EC has proposed substantial investments to consolidate Europe's digital capacity and infrastructure and support the digital single market. This Multiannual Financial framework (MFF) includes two major programmes: Digital Europe - to support R&D for the development of European HPC technology, and Horizon Europe – to procure new supercomputing systems.

The EC plans to develop and reinforce the European high-performance computing and data processing capabilities to achieve exascale capabilities by 2022-2023 and post-exascale facilities by 2026 or 2027. The EuroHPC Joint Undertaking[1] provides mechanisms to implement this strategy: a joint procurement framework, EU-level coordination and the pooling of financing, networking of national capacities and deployment of technology when it becomes available.

In order to develop the future work programmes addressing the challenges of the next generation IT infrastructure for HPC imposed by the strategy above, EuroHPC requires recommendations for research priorities from the European HPC community. By December 2019, ETP4HPC, the European Technology Platform for High Performance Computing, will deliver its fourth Strategic Research Agenda (SRA 4 – also called "Agenda" in the remaining part of the text), which will include those priorities. SRA 4 is a deliverable of EXDCI-2[2], a currently running coordination and support action (CSA) coordinating the European ecosystem.

This document presents a "Blueprint" for SRA 4. It outlines the "big themes" that will drive the selection of research priorities as potential building blocks of the future research calls in High Performance Computing. The methodology used in the process of preparing the Agenda and providing it as an input into EuroHPC's internal mechanisms is also explained.

This document and the upcoming SRA 4 is the work of experts associated with ETP4HPC in collaboration with:

- BDVA (Big Data Value Association), the other private-side member of EuroHPC's "Research and Innovation Advisory Group" (both ETP4HPC and BDVA will collaborate in the preparation of EuroHPC's Strategic Research and Innovation Agenda in the area of HPC and HPDA)
- HiPEAC and BDEC, two projects managing the European expertise in the area of system architectures and long-term big data and computing trends, respectively
- The HPC Centres of Excellence and PRACE, which have provided valuable input on the needs of application users
- AIOTI (Alliance for the Internet of Things Innovation) – we are developing a collaboration with this organisation in order to align the recommendations in the HPC/HPDA and IoT domains.
- The "European Processor Initiative" (EPI), as the cornerstone of the EuroHPC strategy, will also provide input to the upcoming SRA.

This Document and SRA-4 delineate the priorities for the next five years, up to the expected availability of exascale supercomputers. To achieve that goal, ambitious development efforts

[1] See report on the Digital Europe Programme: http://www.europarl.europa.eu/RegData/etudes/BRIE/2018/628231/EPRS_BRI(2018)628231_EN.pdf
[2] https://exdci.eu/about-exdci

are required - using the best-of-breed technology, novel approaches in system architecture and system-to-application co-design in order to master the challenges arising from scalability, robustness and power efficiency requirements.

Looking beyond 2020, HPC-related research will undergo a significant change as tomorrow's deployment scenarios for simulation and modelling will be extended to use cases outside of compute centres. We present, discuss and analyse this new situation. The term "co-design" in the future will not only mean an integrative, open, joint and focussed design approach between system and application specialist – it should be extended to include the domains of data analytics, artificial intelligence, the Internet of Things and cyber security. The challenges become much more complex and demanding.

The main objectives of this Blueprint paper and the upcoming SRA are twofold:

- To promote the development of HPC architectures and technologies as well as converged compute/data platforms tackling societally important problems, with a strong focus on commercial exploitation.
- To improve the cost performance, system efficiency, human productivity and predictability of the contributing fields (Simulation, Big Data, AI, IoT, etc.)

Contents

Scope and introduction

This paper outlines a conceptual framework for the next Strategic Research Agenda (SRA-4) for HPC due to be delivered by the end of 2019. It offers a structured approach to the identification of key research objectives in the 2021 – 2022 timeframe in the area of HPC and HPDA, including significant interactions with Internet of Things (IoT), Cyber Physical Systems (CPS) and Artificial Intelligence (AI). Wherever applicable, it also provides explanations and examples to better position the context of the next Agenda document.

The structure of this document follows a layered top-down approach as shown in Figure 1:

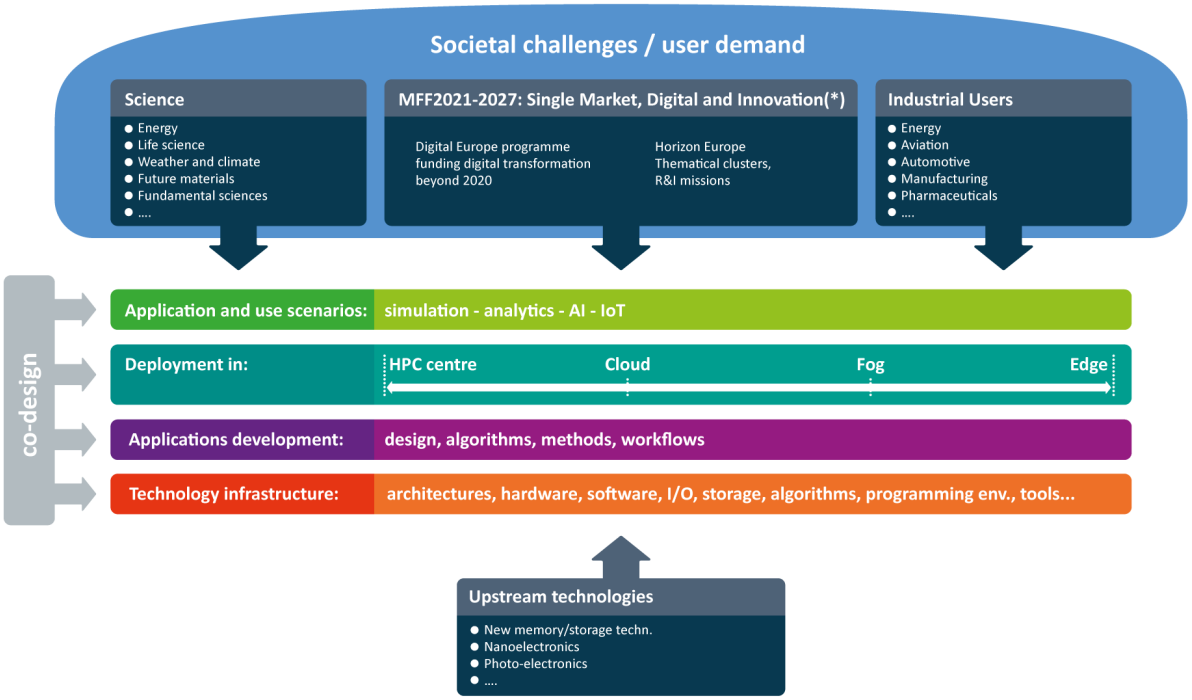


Figure 1: A structural approach to derive research priorities for HPC technology and the application of that technology
http://www.europarl.europa.eu/RegData/etudes/BRIE/2018/628231/EPRS_BRI(2018)628231_EN.pdf

- The top centre layer represents the political framework driving an extended use of HPC and innovation in technology provisioning in Europe in the forthcoming years. Under the "single digital market strategy", the next Multi Annual Funding Framework 2020-2027 (MFF) of the European Commission sets out, amongst others, the "Digital Europe programme[3]" to fund the digital transformation beyond 2020 and "Horizon Europe" with "Thematical Clusters" and "Missions" containing societal challenges stimulating 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 will establish a comprehensive portfolio of projects cutting across several clusters. The first few missions will be introduced in the first strategic planning phase for Horizon Europe[4]. The chapter on "Societal challenges and industrial competitiveness for Europe" on page 7 gives more details.
- The second source of drivers for future technology improvements is represented on the upper right side by commercial and industrial users of HPC. Especially in this category, new use-patterns for HPC are emerging in the context of new products and services (see chapter "Industrial and commercial users" on page 10).
- Science has a well-established role in providing major users and in driving the architectural development of HPC systems. Although several of the fields covered here also are at the

[3] See http://www.europarl.europa.eu/RegData/etudes/BRIE/2018/628231/EPRS_BRI(2018)628231_EN.pdf
[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

centre of the thematical clusters and missions referred to in the Horizon Europe framework, it is important to acknowledge the influence of all scientific domains.

- The next layer down ("Application and use scenarios") translates the use-cases defined in the clusters/missions into application and technology use scenarios across the domains of "simulation", "AI", "Analytics" and "Internet of things (IoT)". As argued below, these domains can no longer be handled separately as they are all required to implement solutions to the problems of today and tomorrow.

- HPC technology will not only be deployed in dedicated data centres in the future. "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 requiring new small-scale deployment environments for HPC, as shown in Figure 2. 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. The layer "Deployment" describes the challenges associated with this change, where HPC functionality is now extended to clouds, fog computing and edge computing:

  - Edge computing is a distributed computing paradigm largely or completely based on multiple compute capabilities positioned close to the end-user or IoT (Internet of Things devices. This is in contrast to having all compute capability in a few, centralised Cloud data centres, with potentially long lines of communication to end users and devices. Edge computing relies on more ubiquitous use and deployment of wireless communication and is meant to alleviate the communication burden (in particular latency) by processing data close to data sources or consumers. Edge computing also supports effective geo-fencing of data that should not leave regulatory domains.

  - Fog computing serves as decentralised intelligence to further reduce the volume of data in transit between the edge and centralised centres and enabling actions close to the edge. With an increasing demand for "near to real-time decisions", it is important to put this local intelligence in place thus reducing the load on centralized data cloud/HPC centres, where data analytics and modelling take place. Fog computing is a layered model for enabling ubiquitous access to a shared continuum of scalable computing resources.

  Both fog and edge computing will profit from power-efficient compute solutions ("embedded HPC"). Privacy requirements must be observed, and they will pose restrictions to data propagating through the network. E.g. hospitals processing patient data would probably use what may be considered fog nodes as they are at the edge of the cloud network, while mobile users use devices at the edge of the mobile network.

- The next layer down ("Applications development: design, algorithms, methods, workflows") addresses the software development aspects of the application portfolio. See chapter "Application development challenges" on page 20 for more details.

- The next 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 integrate the technology stack as well. The range of technologies covers 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.

- The emergence of upstream technologies for future HPC system/component architectures complements the influence of the societal challenges and is 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. See the chapter on Upstream technologies on page 29.

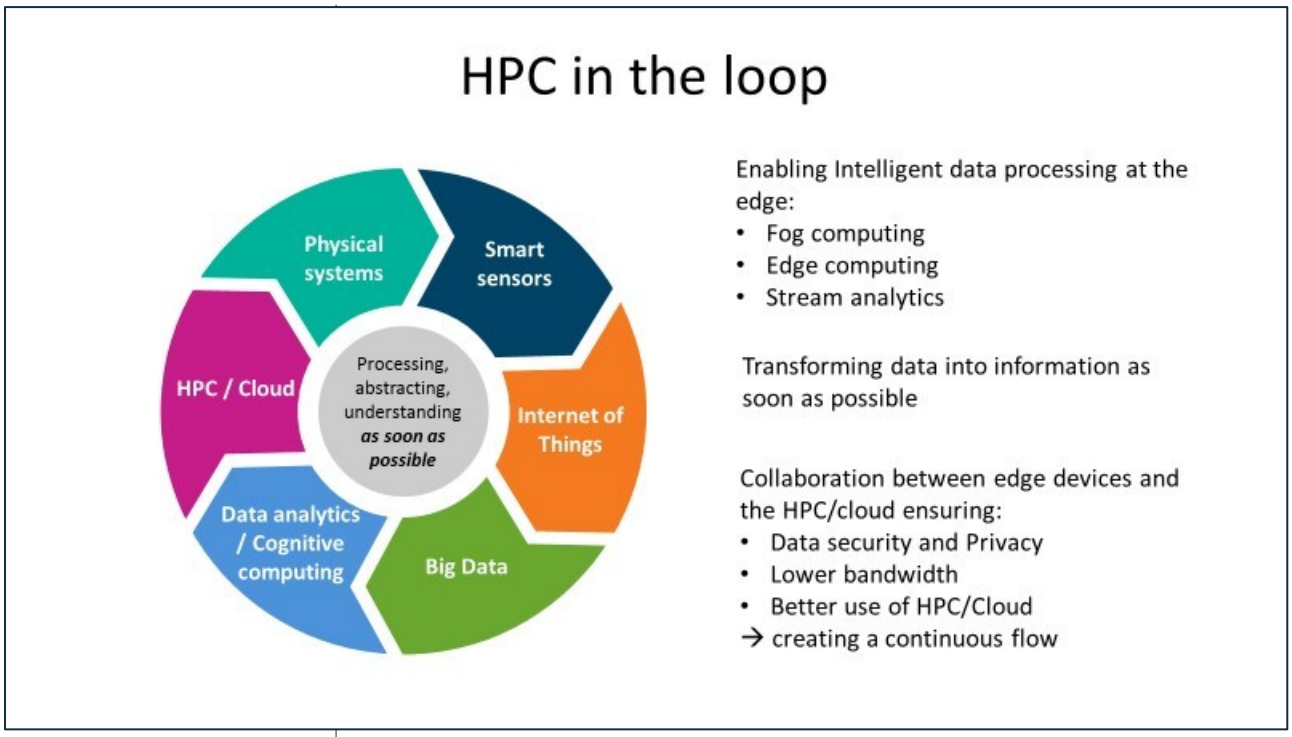


Figure 2: "HPC in the loop" (source HiPEAC)

Figure 2 illustrates High Performance Computing as one element of a circular 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, relevant parts are forwarded to decentralised "fog nodes" close to the edge. A subset of data is then transferred for centralized Data Analytics in clouds or simulation and modelling in centralized HPC centres. In an increasing number of use scenarios based on the concept of the "Digital Twin"[5] a "twin-copy" of a physical entity is held and continuously updated on these central compute infrastructures. The final outcome of the loop is a set of optimized actions in the "Cyber Physical Entanglement" representing physical systems (e.g. robots, vehicles, industrial processes) interconnected in complex intelligent networks.

[5] It is based on the idea that a digital informational construct about a physical system could be created as an entity on its own. This digital information would be a "twin" of the information that was embedded within the physical system itself and be linked with that physical system through the entire lifecycle of the system.

Societal challenges and industrial competitiveness for Europe

Technology development needs - first and foremost – to enable addressing grand societal challenges and enhancing industrial competitiveness. Both concepts have global implications beyond Europe. However, Europe can play a leading role in researching their solution, implementing sustainable solutions and achieving technology leadership.

The current Horizon-2020 programme[6] had already been defined with reference to the UN's 2030 agenda for sustainable development. The UN agenda has led to the formulation of 17 sustainable development goals (SDG) for our society. Horizon Europe's novel mission-based approach[7] aims to achieve these goals with actionable development projects – at which level the specific role of technology can be defined for each mission individually. Successful missions will be critical for enhancing competitiveness through more focused innovation.

The thematic clusters on health, inclusive and secure society, digital and industry, climate, energy and mobility, food and natural resources form the central pillar on global challenges and industrial competitiveness in the new framework programme, and HPC-technologies will be at the heart of many fundamental developments and an important enabler for fulfilling missions addressing the grand societal challenges. On the technology side, this ambition is supported by the Digital Europe programme that aims to deploy technology options and solutions for achieving both global challenges and European industrial competitiveness.

The key research sectors of societal relevance where HPC-technologies play a role are Earth-system science, food-, bio- and life sciences, astrophysics, physics, chemistry and materials science, as well as engineering, transport and communication.

Examples of specific challenges are:

- How advanced bio-medical simulations and personalised medicine can contribute to the well-being of our society given the apparent demographic change;
- How combining advanced Earth-system models with the vast amount of environmental data can prepare society for dealing with climate and environmental change and mitigate the impact of extremes on health, energy, water and food resource management through urgent computing and data handling;
- How materials and systems can be engineered to guarantee availability of secure, clean and efficient energy and smart, green and integrated transport including autonomous vehicles;
- How society can evolve towards being inclusive, innovative and secure exploiting IoT data in support of multi-variable decision support.

[6] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges
[7] https://ec.europa.eu/info/sites/info/files/mazzucato_report_2018.pdf

Break-through requirements

The evolution of Individual technologies and – perhaps even more so - the possibilities for combining those technologies offers many opportunities for addressing societal challenges and creating economic growth in new ways. The technologies we are referring to are: "pure" HPC; cloud/fog/edge computing; user-driven generation and exploration of information available from IoT; powerful methodologies based on artificial intelligence.

Combining diverse technological options require a more specific analysis of what the break-through requirements are for addressing societal challenges and spawning commercial opportunities. These requirements need to include the definition of a generic framework of metrics for success.

Examples for such requirements have been formulated by the PRACE Scientific Steering Committee for key science areas, however with a main focus on computing and output data handling:

- **Fundamental sciences:** Enhanced simulations of the dynamics of black holes and neutron stars.
- **Climate, weather and Earth-system sciences:** Much enhanced simulations to resolve critical small-scale processes driving global atmospheric and oceanic circulation and to allow full-waveform seismic simulations down to frequencies affecting the stability of construction and infrastructure.
- **Life sciences:** Enhanced genetic sequencing across species and for large samples, much better resolution of the structural complexity of proteins, and automated image analysis and visualization.
- **Energy:** Full-waveform seismic simulations down to wavelengths resolving oil and gas reservoirs, and turbulence-resolving simulations of plasmas for magnetic fusion.
- **Infrastructure and manufacturing:** High-resolution structural simulations of aircraft components under stress, and the design of smart cities and an optimized transport and renewable energy provision.
- **Material science:** Design of new materials supported by simulations of the dynamics, thermodynamics, heterogeneity, chemical processes and response to external factors between electronic and continuum scales.

It is presently estimated that these science areas require between 10-1000 more computing power to fulfil their scientific ambition and to effectively support the global challenges in the next decade. Since technology alone cannot be expected to deliver such growth factors, the applications themselves need to become an integral part of a science-technology co-design process.

Success metrics

These examples need to be mapped to the full breadth of IT technologies so that the added benefits from the co-development of new technologies can be quantified. An important element is the time-to-solution constraint: for example, the urgency for predictive analyses in response to environmental and demographic change (necessitating extreme computational capabilities) or the urgency for developing disease diagnostics and treatment options compared with the less strict time constraints for fundamental research. Again, technology refers to the combination of hardware, software and scientific methods with the strong impetus from AI methods.

As a success metrics, the ability to solve the specific problems posed by the challenge addressed, and the application-specific performance and energy used to achieve this are clearly important. Therefore, the applications addressing grand challenges need to be translated into representative benchmarks at full IT-technology scale and take into account trade-offs between the abovementioned realisation-time requirements and technology-readiness roadmaps. Successful technology development has to be demonstrated for each application with application-specific metrics rather than relying on basic technology metrics, as done in the past (an example being Linpack performance measurements).

Necessary evolution of scientific methodologies and economy of scale

Addressing societal challenges with IT technologies is not a mere engineering challenge because the term "technology" comprises hardware, software, industry practices and scientific methodologies. As hardware options change, software and science need to adapt, and as science requirements change, hardware and software design need to adapt. Finding a near-optimum outcome for this mutual adaptation process is a hard problem. The best-known method is to adopt a strict co-design process involving algorithms, software, and system and component architecture. New technologies like AI methods and AI-optimized hardware and IoT data streaming offer new potential for scientific methodologies and workflows. Their adaptation can be intrusive and deviate from the classic first-principles-driven science and linear-workflow approach. A more flexible science-technology co-design with fast turn-around of innovation at the interface between science applications, engineering and computational science is clearly required, and it will breed new areas of knowledge and expertise for Europe. Hence, applications need to evolve as well to obtain the best trade-off between scientific accuracy and computational efficiency, also with a view at which information end-users actually require.

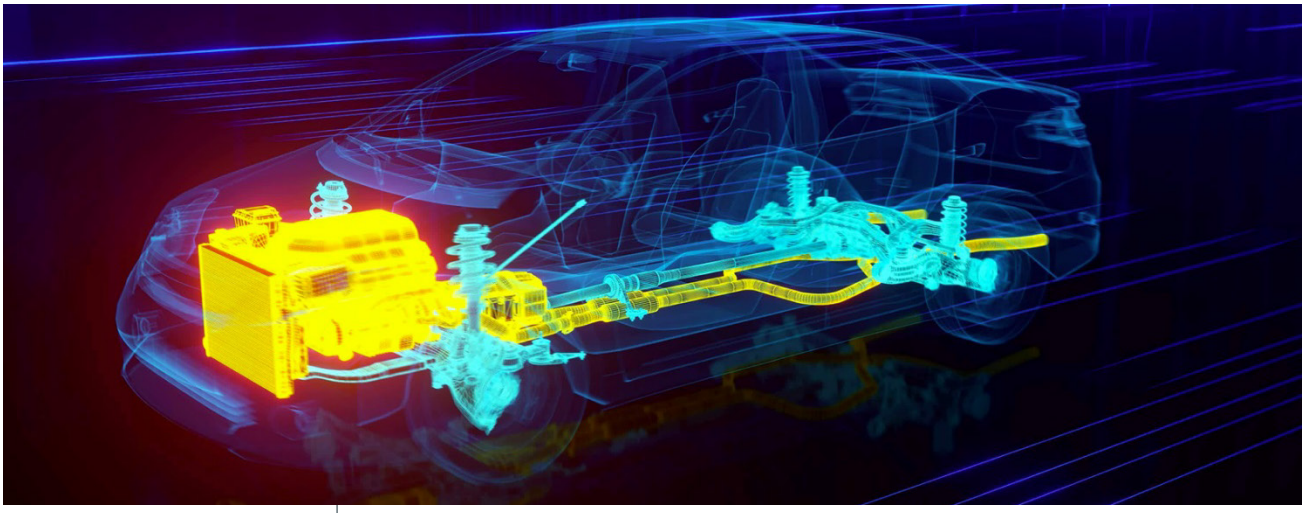
Significant potential for new knowledge and expertise exists between disciplines. It is considered very important to assess overlaps between application areas, and across scientific methodologies and the science-enabling technologies to exploit synergies such as injecting advanced developments from one discipline into another. This will also allow the improved characterisation of the boundary between general-purpose and domain-specific methodologies and technologies, and eventually produce a much more effective approach for the investment in large-scale technology infrastructures in Europe.

The importance of ethics

The tremendous efforts in research and development of AI technologies and solutions and the resulting rapid development of AI capabilities has been one of 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 (such as for military purposes).

Ethical topics impact the SRA in different ways. To avoid these not being discussed until something goes wrong requires ethical aspects 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.



Industrial and commercial users

Over the past 10 to 15 years, HPC was used by a selected few and only in some specific domains. In engineering (e.g., automotive/aero-spatial industries) HPC is used widely to simulate at high precision complex (multi-)physics systems, such as systems for the analysis of combustion engines, aerodynamic properties, or vehicle crashworthiness. The domain of natural resources (oil & gas industry in particular), is another example where HPC is traditionally widely used, as well as for the design and testing of complex technical artefacts used in industrial production (as for example in the pharmaceutical industry for drug design). The financial sector too relies heavily on HPC for its real-time simulations.

The typical HPC user was thus a large company, operating its own HPC centre on its premises, having at its disposal experts for operating the system and for running 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 developed in-house and closely guarded.

The change we observe 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 setup. In this set-up, neither ownership of the computing resources is required, nor highly specialised in-house competences in HPC[8]. Note 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: Moreover, we see a raise in a wide range of new "data driven applications", such as analysis for very large data sets and machine learning (relying on large data sets for the training phase), giving rise to a wide range of new applications. Two much discussed examples are e-mobility with autonomous vehicles and customizing medication and drug consumption to the personal needs of the patient. A number of sectors, examples being designing and operating wind turbines as well as 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 optimize 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 [9]. As shown in Figure

[8] The projects Fortissimo and Fortissimo-2 demonstrated that SMEs could benefit greatly from access to such resources and support to solve business problems. The Fortissimo marketplace was developed by the projects to offer such services.

[9] https://www.ge.com/digital/applications/digital-twin

2, HPC simulation has moved “into the loop”, and becomes 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:

- Users today 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, this is also true for large companies who need a seamless integration of in-house capacities with external cloud-based capacities.
- Second, a broader use of HPC across European industries means that the use of HPC must be available to many, even without highly specialised in-house resources. They would rely on services and support to guide them in how to use HPC effectively for their business and their purposes.

Third, the European industry needs support in application development: to develop effective HPC applications is intrinsically difficult – and the adoption of such codes to new hardware (as for example for GPGPUs) requires deep expertise. And last but not least, 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.

For defining the strategic research topics for the upcoming Horizon Europe framework, the input from industrial users is crucial (i) for addressing their technical needs by taking into account the key requirements of future industry-relevant applications, and (ii) for supporting European industry at large in the uptake of numerical simulations and data driven applications for their businesses.

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 was synonymous with "technical computing using supercomputers" 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 not be any longer stand-alone systems but part of a larger e-infrastructure to release complex, efficiently managed and orchestrated workflows. It will concern also the interfaces of this structure with external devices (distributed and edge devices), as indicated in Figure 2.

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 (over-) simplified as:

- Simulation: relatively low amount of input data, large computation requirements (mostly in high 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, medium computational requirement with loose coupling between compute nodes (scale-out and share 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 hold promise i.e. to rapidly analyse high-bandwidth, high-throughput streaming data.
- AI (for example, Machine Learning in 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, few hundreds of MB.)
- AI (for example, Machine Learning in 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 (e.g. rules of a game, or physical laws or constraints), the output is also low (solution), but the system internally generates a lot of data and computation to explore the various options and find a good solution. Simulation of the process to be optimized 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, multi-physics models; deluge of observational data from sensors and of simulated data; semi-automatic data analysis and post-processing; uncertainty qualification and AI-based models. Combining all these aspects will result in a highly complex application (software) architecture, currently a focus area in related research.

In reference to Figure 1, this layer is driven by the thematic clusters and missions detailed in the previous chapter 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 provide testbeds to research & 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 challenge can only be achieved when considering end-to-end approaches where data production is integrated with data analytics, machine learning, numerical simulation, data archiving as well as final use of the results. Underlying the use cases are applications relying 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 since they combine many heterogeneous components (e.g. relying on different software stacks) as well as different resources or user governance strategies. For instance, this is about applications across a federation of systems - that includes HPC centres, cloud facilities, fog and edge components, networks - while at the same time preserving security & privacy from end-to-end. Furthermore, the economics aspects of the deployment of these applications must be considered.

As a consequence, this means facing extreme scale heterogeneity where, in the worst case, the common denominator may be a common governance and resource allocation policy. At a high level, the main technical challenges are how to achieve interoperability between the application workflow components, their orchestration as well as reproducibility of execution in order to allow debugging and ease of deployment. In addition, infrastructure management and resource allocation policies are also strong roadblocks to overcome. For instance, supercomputers today are typically deployed in a way that they become 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 software stack provisioning. It is difficult to make them part of an application workflow that would include components deployed in the Cloud, handle streaming of data, for example.

To advance the state of the art, supported uses cases must be able to show an application implemented over multiple entities while preserving security and privacy properties. Furthermore, efficient deployment should be demonstrated (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 just one element besides Data Analytics, the Internet of Things and in many cases a cyber physical entanglement. Computing systems are more and more directly controlling devices having 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, since faults or breaches will have wide-ranging consequences in the real world.

Work flow and capabilities

Understanding the workflow and dataflows is of crucial importance for an analysis of real use cases. Each case (e.g. autonomous driving, personalised medicine, wind park operation, etc.) has its unique composition of basic "functional capabilities" (see Figure 3) composed into similar structures as shown in Figure 4.

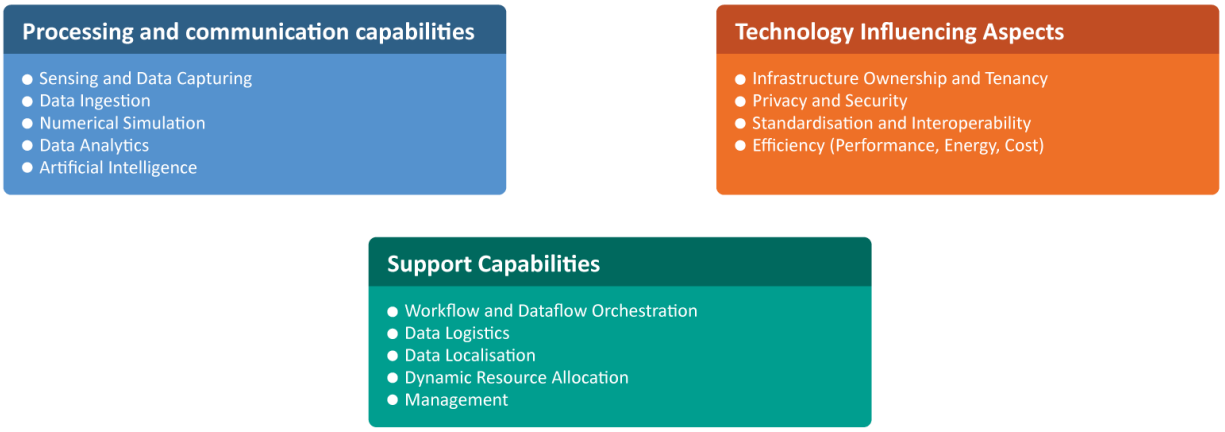


Figure 3: Categories of capabilities in mixed Simulation, Analytics, AI and IoT use scenarios.

- The "Processing and Communication Capabilities" listed in Figure 3 cover all areas which require compute capabilities, be it in a datacentre, 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 meaningfully 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 very much impact the design, implementation and integration of the processing capabilities but do not directly provide any data processing capabilities. They 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 the data processing. Security and privacy must be considered in such an environment for most use cases to comply with regulatory and end-user needs. Interoperability and standards increase the trust in developed workflows and accelerate the adoption by users. The efficiency of a solution is relevant insofar that the costs of a solution limits the adoption in use-cases that yield limited revenue. A performant, energy and cost-efficient system maximizes industrial and commercial competence by enabling novel scenarios.

- "Support Capabilities" describe 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. The identified capabilities are currently underdeveloped for the environment discussed here and require further R&D efforts.

The orchestration of workflows and automatic and efficient deployment across the complex hardware-landscape provided, 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. Finally, workflows must adapt their processing capabilities dynamically depending on the input, or other external parameters like 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.

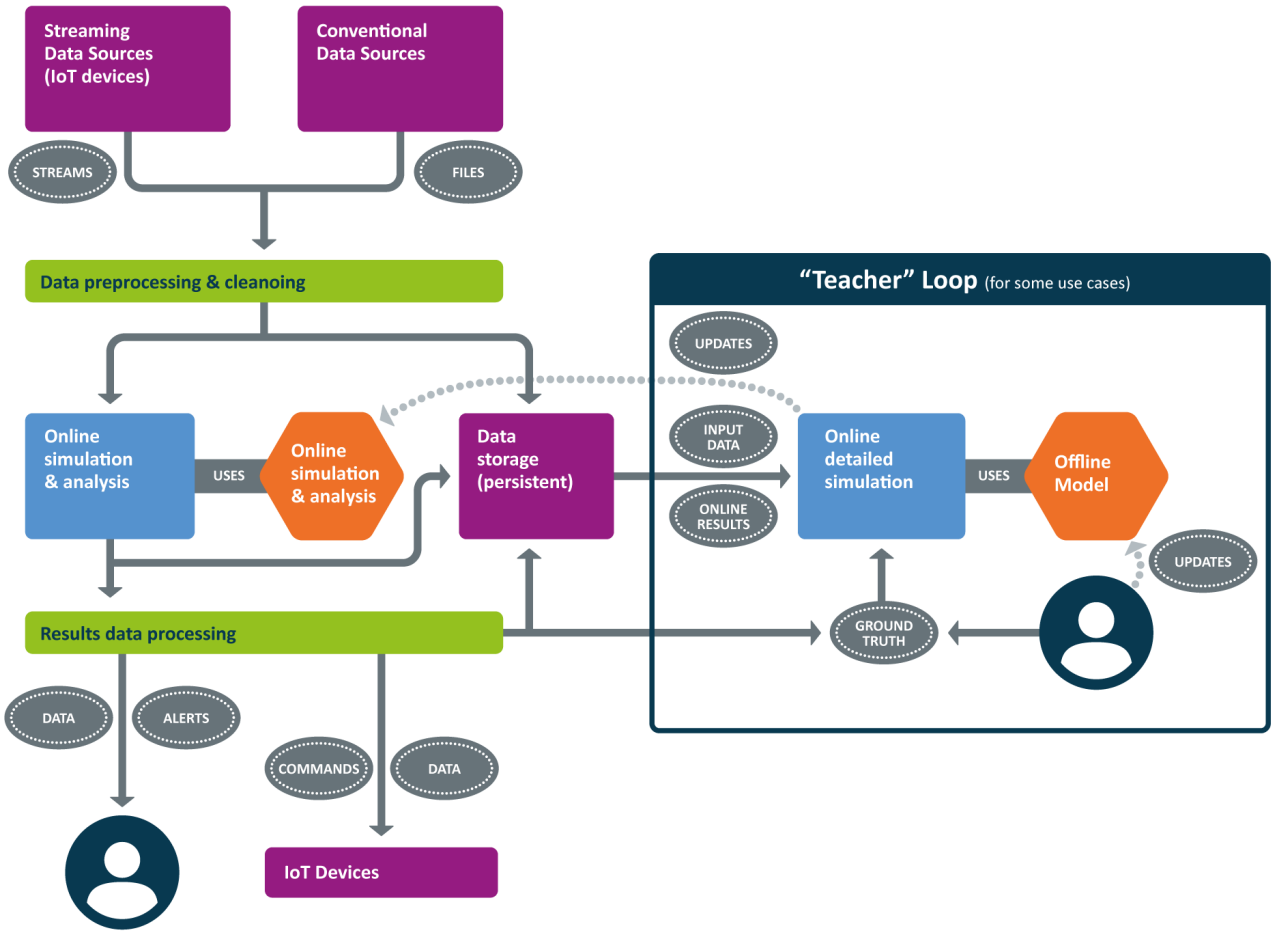


Figure 4
A typical mixed simulation and machine learning workflow

Figure 4 unfolds the loop shown in Figure 2 and shows three steps that are common to the use cases discussed between ETP4HPC and BDVA: in a first step, data from a multitude of real-world sensors or conventional sources (e.g. databases) is ingested, pre-processed and cleaned. This already can 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 technical 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. Part of the analysis results are again put into storage for later use.

The third step is the processing of the analysis results, and the 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 effecting 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", and 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, for instance, a fully physically correct six degrees of freedom driving model, reducing the amount of processing needed per instance and consequently reducing energy requirements.

Data life cycle and dataflow: an example

Understanding the necessity for a dataflow orchestration in mixed Simulation & 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 quality of the simulation will depend on the availability of this new set of data on the HPC system
- Data is produced by sensors in a streaming mode and 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.

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

- Unstructured data for instance issued from 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 data transfer, compression process, encryption process could slow down the simulation workflow and could require some provisioning or concurrent data and simulation processing.

Then the challenge 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.

For data driven application workflows, 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. The pre-emption of HPC resources depends on data workflow optimisation.

This distributed data transformation must be integrated in a Big Data life cycle model that includes activities to more closely combine data curation with the research life cycle. These 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 are to be studied. 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.

Orthogonally, new HPC systems will have to consider, at the same time, where data is stored and how/where the same is accessed for computation. In a federated scenario, data could be stored across distributed Edge, Fog and possibly multiple centralized “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. For example, AI-assisted intelligent caching systems could be designed and deployed to take advantage of the read-only nature of many workloads (e.g., Deep Learning training or Big Data input), either using high performance node-local storage like NVM disks or leveraging existing NAS infrastructure.

The design of a global infrastructure allowing one to combine external edge- or peripheral environments with a central, shared infrastructure will require the analysis of the heterogeneity of the entire software environment, the identification of new data sources and the quality of the data. The diversity of application requirements, in terms of workflow patterns and data distribution will define the rules for new combined data use solutions.

Deployment structures

As mentioned in previous chapters, mixed simulation, data analytics and AI workflows often include a large ensemble of heterogeneous resources (from HPC centres to cloud facilities as well as edge/fog components). The BDEC project issued a survey paper[10] introducing the concept of "continuum computing" to define a new paradigm of converged sensor-data-compute needs in the era of Big Data and the Internet of Things.

By the end of this decade, the world's store of digital data is projected to reach 40 zettabytes (10^{21} bytes), while the number of network-connected devices (sensors, actuators, instruments, computers, and data stores) is expected to reach 20 billion. While these devices vary dramatically in their capabilities and number, taken collectively they represent a vast "digital continuum" of computing power and prolific data generators that scientific, economic, social and governmental concerns of all kinds will want and need to utilize. The diverse set of powerful forces propelling the growth of this digital continuum are prompting calls from various quarters for a next generation network computing platform—a digital continuum platform (DCP)— for creating distributed services in a world permeated by devices and saturated by digital data.

But experience shows how challenging the creation of such a future-defining platform is likely to be, especially if the goal is to maximize its acceptance and use, and thereby the size of the community of interoperability it supports. Also, the ability to integrate a large variety of technologies in a single workflow where the individual technology components may be deployed in completely different control domains needs to be further developed.

[10] M. Asch, T. Moore, et al. Big data and extreme-scale computing: Pathways to Convergence - Toward a shaping strategy for a future software and data ecosystem for scientific inquiry. *International Journal of High Performance Computing Applications*, 32 (4), pp. 435-479. (2018).

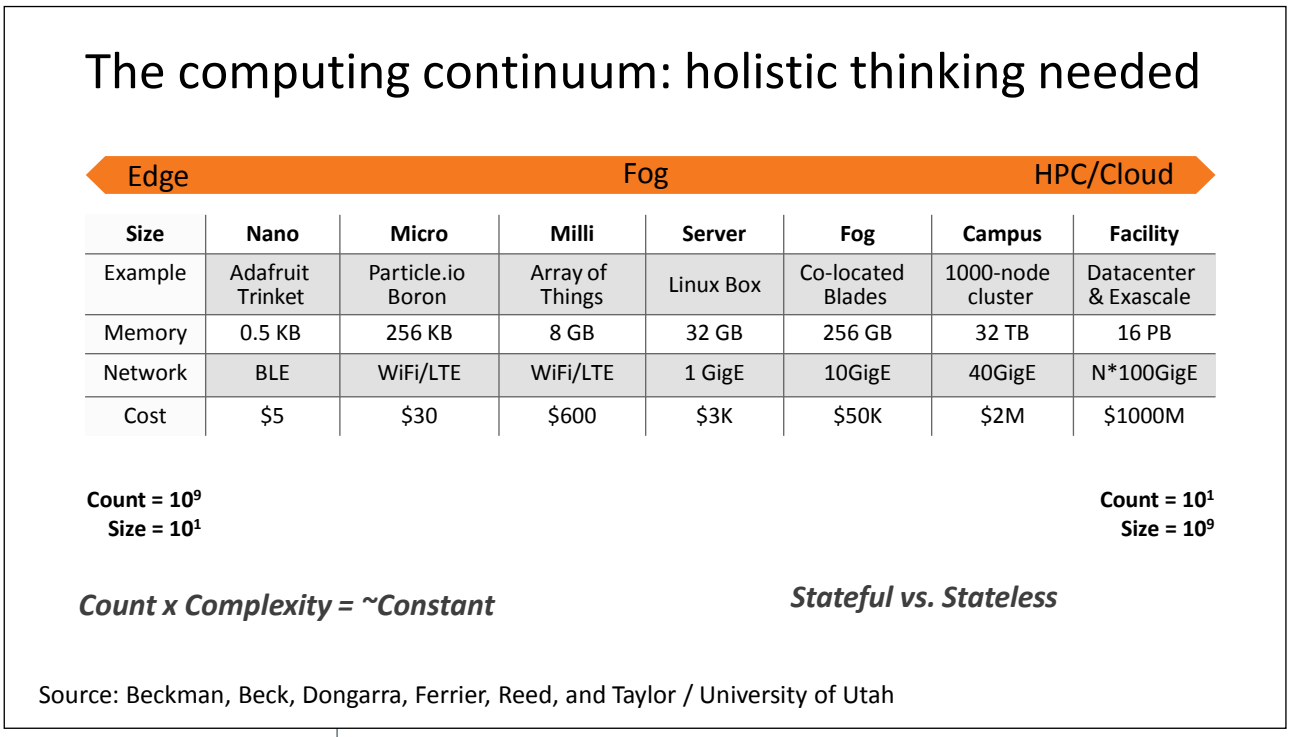


Figure 5:
BDEC's view of the Digital Continuum Paradigm

Application development challenges

Next generation applications need to efficiently exploit new IT infrastructures, they will likely be much more integrated between simulation and data analytics capabilities, supported by AI and they will have to satisfy more demanding user requirements in terms of response time, flexibility and ease of use. This chapter illustrates some of the most important developments, new features and characteristics expected of applications in the timeframe of 2021-2017.

The use of AI in HPC

Combining HPC and AI techniques to produce applications that are superior in their capabilities and performance to "pure-breed" applications is a very hot and rapidly advancing research field. From humble beginnings (such as finding hidden and very intricate patterns in simulated and observed data), the first generation of combined applications is now emerging. At SC18, no less than five of six applications nominated for the highly prestigious Gordon Bell award combined AI and HPC techniques, with the spectrum ranging from AI-driven preconditioning of linear systems resulting from earthquake prediction, to the use of HPC techniques to optimize complex neural networks (NN) topologies. In addition, the DoE Summit and Sierra systems exhibited very high efficiency in both the HPC and AI realm for these applications.

Still, careful study is required to understand which applications will potentially benefit from AI techniques and which will not. AI techniques and frameworks have to be improved to best fit into HPC applications and run on HPC systems. Finally, the combination of AI and HPC will most likely result in the need to support dynamic and interactive execution modes.

Higher Level Abstractions

Computing systems are becoming more complex and are also typically based on heterogeneous architectures, making them increasingly difficult to program efficiently while also maintaining portability. At the same time, applications are getting more complex, combining different modules in an often dynamic way. Higher level abstractions, e.g. in the form of domain specific languages (DSLs), are needed to increase programmer efficiency, hide the hardware complexity from the scientist and allow hardware independence. These high-level abstractions need to be efficient and optimized, including actual HW/SW co-design and allow expert users to optimize on lower levels.

Tolerate Latency and Exploit More Parallelism

Many applications have reached the end of strong and even weak scaling and are thus not able to exploit the increased hardware parallelism levels available at the Exascale. At the same time, many algorithms require tight integration and cannot tolerate latency, a fundamental limitation on Exascale systems where the deepened memory hierarchy and the increased physical system size will introduce new levels of latency. Current latency-hiding techniques might be insufficient

to deal with the latency induced at the Exascale. To boost the parallelism available in applications approaches like ensemble parallelism, speculative execution, parallel-in-time iterations, relaxed/eventual consistency models. which have been successfully employed in some domains, need to be developed for a wide variety of domains and algorithmic formulations need to increase their tolerance towards latency. This will also require appropriate resource management that can deal with large ensembles of applications and allow for dynamic steering.

Dynamic Execution Modes

Only very few applications will use Exascale systems to run a single, large simulation. Instead, ensembles of smaller runs, workflows, and speculative execution will become more common on these systems. This not only requires the appropriate resource management that can deal with large ensembles, but also increased interactivity to allow for dynamic steering. This interactivity will also be needed for the convergence of HPC, data analytics and AI. Already today, in-situ data analysis is being used to reduce the enormous data size these large-scale applications have to deal with or to perform online visualization of (partial) results. Increasingly, this in-situ analysis is also being used to steer the further execution of the application, either automatically or with user intervention. The increased use of AI techniques will further accelerate this.

New algorithms, solvers and methods: FP precision, data locality

We need to revisit the algorithms, solvers, and methods used in HPC applications for Exascale. Many HPC applications have been using double precision floating point (FP64) by default, but this is particularly expensive. The selective use of DP(double=FP64)/SP(Single=FP32)/HP(half=FP16)/BP(Google bfloat16 FP[11]) or some other newly proposed floating point format should be considered to reduce the system cost/consumption with equivalent overall precision of the results obtained.

While it is comparatively easy to design hardware for reduced precision datatypes, the problem lies in validating which algorithms or applications can actually use these and still produce useful results. The numerical stability and error propagation of key HPC kernels needs to be studied, and automatic tools that assist developers with this menial task will be needed. The recent work on uncertainty quantification can be a good starting point.

In addition, algorithms have traditionally been designed with the primary goal of reducing the number of operations. In Exascale hardware, however, data movement will be much more costly than floating point operations. A more data centric approach at all levels is thus required, and the existing research on communication-avoiding solvers has to be accelerated. Also, the current numerical schemes need to be revised and also drastically new methods such as switching from traditional numerical methods to empirical schemes should be considered.

Democratization of HPC

HPC is still seen as a domain of an elite of highly skilled scientists and engineers. To make the benefits of HPC available to a much larger circle of scientific and commercial users, applications have to become friendlier to users lacking specific HPC skills. This in particular includes the development of high-level interfaces and portals that hide the underlying complexity for average users and providing HPC applications as a service (e.g. through HPC-enabled clouds). In addition, much increased training and education efforts are needed to increase the pool of highly skilled HPC people.

Progress with the convergence of HPC, Big data and AI will to a large degree depend on availability of experts with competency in HPC and at least Big Data or AI; it will be imperative to break the “silo walls” and offer targeted education and training to the existing circle of experts in each of the fields, and integrated programs for students.

[11] Bfloat 16 uses the 8-bit exponent of FP32 and reduces the mantissa to 7 bits.

Code base modernization and maintenance

With the fast-moving hardware landscape and the development of new algorithmic formulations, the maintenance of an efficient and productive application base becomes an ever more challenging task. Constant effort is needed to port and tune applications on new hardware (particularly SIMD, accelerators, novel memory/storage technologies), adopt them to new concepts (e.g. object storage, novel parallelism schemes), and include revised or changed algorithms and solvers. As many current HPC applications are legacy applications whose origins trace back several decades, code modernization will be a necessity to integrate novel concepts.

HPC & HPDA Systems: Architecture and technology

Two orthogonal and complementary approaches laid out in this chapter can enhance the architecture and technology of future HPC systems and bring solutions to the new challenges and applications scenarios described above:

- Adaptation of the technologies to new applications requirements
- Integration of new specific architectures optimized for sub-classes of applications in general purpose HPC systems

Convergence of Simulation, Big Data and AI in the same IT continuum

Converged HPC/HPDA/AI workloads have different characteristics from pure HPC loads and therefore demand additional features of the systems they run on. The systems should be able to cope with workload diversities, AI inspired solutions could be used for efficiently managing the complexity introduced. It is likely that the hardware of the system will be heterogeneous, using processors for computation and orchestration of the dataflows, and various accelerators, such as FPGAs and GPUs or their derivatives for Deep Learning. It is important to be able to reconfigure the data centre dynamically: elastic reconfiguration and efficient scheduling are potential solutions.

To progress in this direction, some advances in four axes are needed:

- workflow management
- efficiency of the deployed infrastructure
- efficiency in using this infrastructure, especially in programming it
- features with growing importance

Emergence of workflow management

As seen in the previous sections, some of the toughest technical challenges are centred around understanding and modelling the data and workflows in the underlying multi-owner IT infrastructure. Support capabilities such as workflow and dataflow deployment and orchestration, data localisation and logistics and dynamic resource allocation (compute, network, storage) need to be developed and integrated.

While there will still be a clear distinction between HPC/servers and edge/embedded computing on the hardware level, software layers should allow to use different devices in a seamless manner, i.e. the distinction becomes somewhat opaque from a user's perspective and allow to form a continuum of processing and storage: data needs to be processed where it has to be for safety, privacy, cost or efficiency reasons, and the complete hierarchy should collaborate and exchange capabilities when it will be needed.

In advanced systems, it might be necessary to migrate the processing from and to different parts of the continuum, an easy migration of efficient code shall be required (even if dynamic migration is not required in most of the cases).

Efficiency of the system

The main roadblock for the development of new HPC infrastructure is its efficiency in terms of energy since its power budget is constrained. Even though reliability is also an important part of the cost of ownership (cost to replace the faulty parts and to restart operations if intermediate states are not properly saved), the main problem remains the efficiency in terms of power consumption.

Closing the gap between real applications and benchmarks shall insure that the performance per watt ratio is relevant to evaluate, compare and tune HPC/HPDA architectures.

Better energy efficiency will reduce thermal footprint and cooling cost at node level. At global level ultimately, the electricity power sizing and its associated cost limits the size of the machine; thus, 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 supporting directly float16 which is enough while decreasing the size and energy required. Some operations don't even need to be exact, so operators can be simplified while being "good enough" for the requirements. On the other side, floating point representation can induce errors in iterative computing, and new formats, like UNUM, can help solving the effects like numerical instability.

- **Application specific hardware** is more efficient in terms of FLOPS/Watt or Ops/Watt than general purpose because computing resources are tuned to the application class and their control is more efficient. For example, for throughput, GPUs are more efficient than general purpose processors, yet their compute capabilities are more limited (SIMD instead of MIMD execution), and programming can as a result be more difficult in the general case. This approach is described blow in the "new architecture" section.

- To increase energy efficiency, not only processing, memory and communication should be optimized, but the **complete supporting infrastructure** as well: power supply, cooling etc. For example, bringing the final DC to DC converters as near as possible to the ICs will improve efficiency by decreasing the path of high intensity current.

- A challenge will be to **combine different accelerators** into a unified programming model. In terms of realization, some physical phenomena might be more adequate for simulating or processing data. For example, optics has an inherent large parallelism and can be used to accelerate processing like matrix operation, random generation or operations in the Fourier domain. This will be more detailed in the "upstream technologies part.

- **Storage:** computation is not the only part that needs to be pushed to its limit: to be efficient, all elements such as communication and storage need to be improved. We see now the emergence of new storage, based on advances in upstream technologies (see this section below). Communication is also a major challenge due to its cost in energy and the increasing bandwidth required for new applications. Photonics was used for rack interconnect, it might become more interesting at board and even at chip (interposer) level.

- **In situ/in transit processing:** Traditionally, in the HPC area, datasets resulting from scientific simulations are typically shipped to some auxiliary post-processing platforms for offline visualisation, processing and analysis, which becomes more and more costly in terms of storage requirements as data volumes grow. In situ processing is a more efficient alternative, allowing data visualisation and 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.

Productivity of Application Developers

An equally important area of efficiency concerns productivity of application developers (human productivity): the effort to program complex systems must be commensurate with the benefits gained in using those systems; in the extreme case, the most powerful computing infrastructure is of no use if it is "un-programmable", in terms of complexity and/or non-standard and bespoke programming models or APIs. Whereas a convergence of infrastructure technologies and systems will potentially support modelling and simulation, big-data, AI and IoT communities on infrastructures spanning HPC-Cloud-Fog-Edge resources, the convergence of applications would be expected to happen at a higher level in the sense that new applications and workflows would encompass the methodologies, algorithms and programming paradigms of those areas. It is expected that the specialised applications from the different communities are unlikely to migrate to a common programming language: while numerical simulation codes typically use compiled languages like Fortran, C and C++ with communication libraries such as MPI, the other fields have very different practices, such as "R" for statistics, Tensorflow or PyTorch for Deep Learning and some communities have moved to interpreted languages such as python). However, the growth in more complex applications and application workflows calls for programming environments that facilitate a combination of approaches or indeed higher-level abstractions/systems that provide interoperability, composability and support for automatic deployment of the appropriate programming paradigms and languages for application components. Therefore, it is important to promote solutions and programming practices allowing to "orchestrate" a large variety of code, e.g. by exposing API outside of the codebase.

Managing heterogeneity in an efficient way is another important factor: As explained above, there will be heterogeneity in the hardware, which will be a major challenge to efficiently exploit the variety of hardware resources without explicit mapping and communication done by the programmer (who don't want to know the hardware). There will be also heterogeneity in programming environments and programming models due to the heterogeneity of programmers and applications. Perhaps here also AI related techniques can be used, to extract the requirements from the programmers or users, and remap them efficiently onto the available hardware.

Key efficiency metrics for the programming environment continue to be ease of programming; ease of migration for legacy application codes; code efficiency (performance); standardisation; portability across heterogeneous hardware resources including accelerators. Application developer productivity challenges related to this are: providing the appropriate programming languages and models for components (including streaming, on-the-fly processing and not always read from memory, process and store); providing software stacks below the high-level, workflow or specialised application programming interfaces that exploit the best practices (performance, scalability, reliability, etc.) currently.

Features with growing importance

There are at least three noticeable topics that will become more and more important in future HPC systems.

Predictability of the execution time

Predictability is also a major challenge for new applications of HPC. Of course, the results should be reproducible, but with applications like HPC in the loop, where HPC systems will be directly in the loop to control real-life processes, predictability in time is important, i.e. when the results will be available. They should not appear when it is too late. Current approaches and methods are not very well designed to ensure this time predictability. However, it is clear that the requirements are for HPC softer than for hard real-time systems.

Guaranteed QoS

Another aspect is to guarantee the Quality of Service, including availability and reliability. Here also, the requirements are different depending on the applications: for simulations, the results have to be exact, even if there is a failure of some parts of the system, and execution time deadline is not essential, while for real-time application, approximate results are often better

than missing the deadline. AI solutions often give approximate (or statistical) results, and without 100% accuracy, but they can be useful to bootstrap more accurate simulations. Depending on the characteristics of the applications and their main criteria for defining the QoS, the system should be able to adapt and ensure than the "contract" of the QoS is ensured, by relaxing some constraints less essential.

Security and privacy

As HPC is becoming more and more prevalent in use cases involving personal data and critical systems (such as banks and hospitals), security and privacy are playing an increasingly important role. On the one hand, we need to consider all research aspects related to the security of HPC hardware and software technologies, on the other hand HPC technologies could be used to offer security and privacy services. We mainly focus on the first area, focusing on the security and privacy requirements of emerging use cases of "HPC in the loop" and "HPC in the cloud / as a service".

Today, the main challenge in in winning additional HPC users is to convince them about the value that HPC methods can bring to their science or commercial results. Trust issues, such as concerns about maintaining data privacy, integrity and security, handling regulatory compliance in a service-oriented environment, and enforcing service level agreements will become more important in the future. Concerns voiced so far are mainly related to the effectiveness and efficiency of traditional governance and protection mechanisms, for example the collection of events by security event and information management tools or forensics in the cloud and maintaining the security and integrity of data retained in the cloud, potentially where retained over many years.

Traditional HPC cybersecurity relies on strict authentication and access control to HPC installations (based on passwords, SSL certificates and for particularly sensitive installations, two-factor authentication), on restricted usage rights (like f.i. not allowing instantiation of Internet connections from the HPC systems), and on monitoring suspicious or abnormal activities (e.g. f.i. catching crypto mining applications). Data protection does mainly rely on storage and file system access rights, plus in some cases encrypted storage. Data operated upon by HPC workloads is typically un-encrypted. Rights and responsibilities of users are laid down in a signed user agreement.

The fast evolution of new use cases and deployment models is posing significant new challenges. Use of HPC in the Cloud or of HPC services requires full use of strong AAA (Authentication/Authorization/Auditing) techniques, such as trust certificates. The introduction of identity as a service (IaaS) with identity data potentially spread across multiple, different trust silos further complicates the HPC service chain. As a consequence, security can no longer rely solely on a set of static system configurations defined by a human administrator -- an ongoing adaptive process in which policy-based techniques are used to provide automated configurations to dynamically handle security requests and events is now required.

HPC services will likely be based on a range of trust agreements, identity management options, and compliance mechanisms to ensure that other parties are adequately enforcing privacy and security. The security management will follow the feedback loop already in use for network and systems management, which includes monitoring, analysis, planning, and execution steps.

One of the main problems in the future Cloud and Edge use cases is not assurance of individual services one by one, but rather end-to-end (E2E) assurance, where we have to deal with services that offer their security assurances as well as assess the security of their sub-services (including storage or computing services). We need common frameworks which enable providers to advertise their security rules and events and allows customers to continuously monitor the actual security of a service.

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 it can be costly in terms of transistor, data paths and energy compared to what is needed for an application. It implies a lot of

movement and duplications of data, which is not efficient (bringing data from external memory is 3 orders of magnitude more energy demanding than a floating-point operation on those data). There is a research path to propose architectures that will be more efficient for some class 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 chapter "Upstream technologies" on page 35). Some concepts of new architectures are generic (see below dataflow or IMC) or targeted specific class of algorithms (see below neuromorphic, graph and simulated annealing).

The integration of new architectures with standard ones into heterogenous systems is facilitated by the emergence of "open processor interconnects"[12] which allows high performance and coherent communication between processors, accelerators and memory subsystems.

Dataflow

In dataflow architectures, data moves between modules that perform the computation on the data. You do not have any program counter that controls the execution of the instructions as in 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 linked to FPGA (Field Programmable Gate Array) as most of the ideas have not led to the tape out of specific circuits but have been tested and implemented with FPGA.

With the slowdown of standard processors performance increase, development of data flow architectures can provide an alternative to deliver this performance increase. The development of reconfigurable architectures (e.g. Intel CSA Configurable Spatial Accelerator) and progress toward flexible reconfigurable FPGA will be an asset for implementing data flow architectures.

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

These architectures couple the storage with some computing capabilities. The idea is that bringing the computation to the memory will be cheaper in resources than moving data to the computing units. Most the time this approach is mixed with a standard architecture to allow computation on several data.

The architecture is also related to the development of Non-Volatile Memory (see also next chapter) and appealing as long as the cost of the in-memory computation is low.

Neuromorphic

The development of AI, and especially applications using Deep Learning techniques, has led to a huge interest for neuromorphic architectures that are inspired by a theoretical model of a neuron. This architecture can be used for AI tasks but can also be viewed as a generic classification function or a function approximation.

As more and more applications (or a part of an application) are mapped to this paradigm, it is worth to develop specific circuits that implement only the operations and data paths mandatory for this architecture. Several examples already exist as Google's Tensor Processing Unit chip (TPU) or Fujitsu's Deep Learning Unit chip (DLU). These efforts have not exploited all the possible options and have not developed all the interested features of the architecture, so research in this area is still valuable.

We can distinguish various kind of possibilities:

1. Using classical digital arithmetic, but designing more specialized architectures (examples: TPU and DLU)
2. Using another way of coding information, like "spikes" as used by human brain
3. Using « physics » to make computation (e.g. Ohms law for products and Kirchhoff law for summation; see section "Analog computing" on page 31).

[12] Examples are the Gen-Z and CCIX consortia or the OpenCAPI, NVlink, AMBA and Intel CXL (Compute Express Link) interconnects.

Of course, the approaches can be combined. Typically, most people call “neuromorphic” the approaches using 2), because it is closer to the way the nervous system communicates.

An important aspect is that this architecture is a good candidate to introduce alternatives to CMOS.

Graph computing

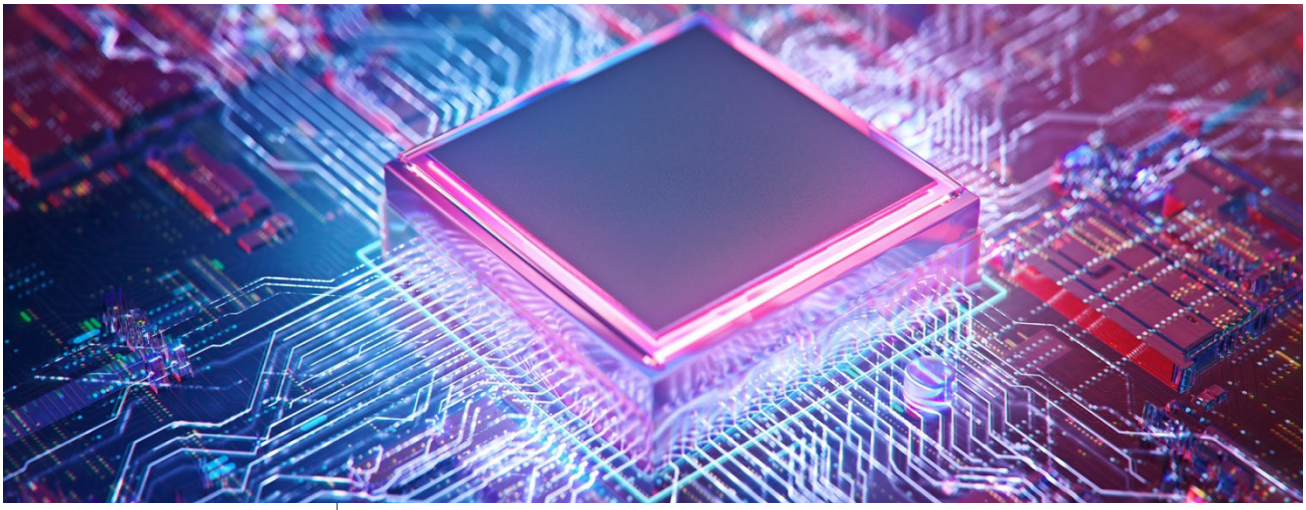
Graphs play an important role in data representation and in some AI or optimisation problems. As standard processors have poor performance due the non-regular access to data, developing a specific architecture can be relevant.

Simulated annealing

Simulated annealing is a method to solve complex optimization problems. It can be implemented by software on classical Von Neumann processors, but you 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.

This approach has been implemented by Fujitsu with its “Digital Annealing” processor. This project has developed a standard CMOS ASIC and a software stack to map the optimization problem to the circuit.

Other efforts use quantum devices (see Upstream technologies section) to target the same class of problems (this approach requires cryogenic operation).



Upstream technologies

Significant investments are done in R&D on next- or even further technology generations not directly linked to HPC, but to a broader use in computing and data processing. For some, a clear deployment within the next years can be foreseen, like components for neuromorphic computing or "in memory computing" (covered in chapter "New architectures" on page 26). For others options their commercial exploitation is not yet obvious, but they deserve attention today, as a path for their utilization (e.g. quantum accelerators) needs to be addressed now. In this chapter, the options are presented starting from the ones in continuity with the current state of the art to the most disruptive ones.

Enhancements of current CMOS technologies

CMOS scaling

Even if we are close to the limit of CMOS scaling, there is still room for improvement in this domain. The leading foundries (TSMC, Intel, Samsung) are investigating for at least two more technology nodes compared to their current technologies. This could provide a way to put roughly about 4 times more transistor in the same surface of silicon as of today. However, this scaling is at the cost of very expensive equipment (e.g. Extreme ultraviolet lithography - *EUV* or *EUVL*), and the power density of those technologies is still not known, perhaps limiting the number of devices active at the same moment on the die. It should also be noticed that even labelled with the same name (e.g. 7nm), all technology nodes are not equivalent.

CMOS scaling is also related to the evolution of the structure of the transistor. After FDSOI (Fully Depleted Silicon on Insulator) and FinFet, the structure of the transistor could be based on silicon nanowires.

2.5D/3D stacking

2.5D/3D stacking provide a way to reduce latency and energy and to avoid package bandwidth bottleneck when we want several chips to communicate together. 2.5 D stacking is the concept of small dies (called chiplets) integrated on a common substrate (the interposer) that can be organic, passive silicon, active silicon or using photonic technologies. 3D stacking is the stacking of layers of integrated circuits on top of each other. It can be done either by wafer to wafer, chip to wafer stacking, or by monolithic 3D which allow a finer granularity (down to the level of transistors). HPC is already benefiting from this technology with the first HPC systems using high bandwidth memory (HBM or HMC) and processor manufacturers (e.g. AMD, Fujitsu and others) have 2.5D in their roadmap. The boost in memory bandwidth is a great improvement for memory bound applications and a must for architectures with accelerators that require this kind of bandwidth to deliver their performance. 2.5D also allows to mix chiplets with various technologies, and for example, with active interposers, having the power conversions integrated "in the chip", providing a global better energy efficiency.

It can also be a path for production of hybrid packages mixing chips of different architectures or even chips manufactured with different technologies. Nevertheless stacking "compute" chips with a higher heat dissipation than memory chips leads to thermal problems that today limit the number of chips that could be put in a package.

"High end" computing is more and more important for the automotive market (Advanced Driver Assistance Systems (ADS) and self-driving) and might be a drive for having European actors in 2.5 D and integration of complex systems on dies or interposers.

The European EPI (European Processor Initiative) plans to use 2.5 D technology.

Hybrid of CMOS and other technologies: NVMs, silicon photonics

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. The developments in this domain have several impacts for HPC. The energy to retrieve data is decreased, the latency to read the data is reduced and the density can be increased (especially with solutions implementing multi-states 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.

Silicon photonics

Silicon photonics can be used either to compute or to provide interconnect between computing elements.

Compute

The properties of light can be used to perform computation. For example, the interaction of lights of which the phase has been modulated according to inputs can produce operation over these inputs. This idea can be used to implement neuromorphic architecture where the main operation is a scalar product.

This approach is promising but several steps are still to be achieved: assessment of the value proposal in term of energy efficiency and industrialization path of the technology.

Another path is to use the massive parallelism of optics to perform complex operation (typically where the complexity is not a linear increase versus the size of the problem). An example is the system proposed by the start-up LightOn, integrated in an OVH cloud server (see section "Analog computing" on page 31).

Interconnect

Photonics is already used for long distance communication in HPC systems (electrons are easy to create and interface, they have attenuation with the distance (Ohm's law), while photons are energy demanding for creation and interfacing but have low attenuation with the distance). The technology is also appealing for rack level communication. But perhaps the most interesting aspect will be at package level with the development of active interposer with embedded silicon photonics network between chips or chiplets. The bandwidth and the energy efficiency can be increased compared to current CMOS solutions.

New solutions more efficient than CMOS

CMOS has been such an industrial success story that it has reduced the effort on alternative solutions to implement transistor 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 they will be able to deliver more computing performance than CMOS.

Superconducting

With the use of superconducting material, the expectation, based on the zero resistivity of the interconnects, is that power consumption could be up to two orders of magnitude lower than that of classical CMOS based supercomputers.

Nevertheless, superconducting circuits have still to overcome several drawbacks as density, switching time, interface with external systems or noise reduction to be seen as a potential solution for HPC. Most of the time the implementation uses Josephson junctions and so has the same disadvantages as analogic computing.

Magnetoelectric and spin-orbit MESO

Researchers from Intel and the University of California, Berkeley have proposed a new category of logic and memory devices based on magnetoelectric and spin-orbit materials. These so-called "MESO" devices will be able to support five times the number of logic circuits in the same space as CMOS transistors. Compared to CMOS the switching energy is better (by a factor of 10 to 30), switching voltage is lower (by a factor of 5) and logic density is enhanced (by a factor of 5). In addition, its non-volatility enables ultralow standby power.

This path is promising even if the roadblocks for industrialization are still difficult to assess.

Memristive devices

Besides the uses of the resistive memory for NVM and analog neuromorphic architectures, memristive devices can be interesting to implement 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 reprogram 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.

Other materials

There is some research 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. Nevertheless, at this stage of the research, it is too early to assess whether these options will propose a valuable solution for HPC systems.

Analog computing

We call analog computing when a physical (or chemical) process is used to perform calculation. (An analog computer or analogue computer is a type of computer that uses the continuously changeable aspects of physical phenomena such as electrical, mechanical, or hydraulic quantities to model the problem being solved. – Wikipedia)

Optical systems

Optical systems can be used to compute some functions thank to light properties and optical devices like lens. This approach is an extremely energy efficient way compared to traditional computers. This technology cannot suit every application but a number of algorithms as scalar products, convolution-like computations (e.g. FFT, derivatives and correlation pattern matching), are naturally compatible. Some demonstration has been made by the EsCAPE project with the computation of spectral transforms by an optical system. The precision of the results can be a problem if the spectral transform is the input of a subsequent algorithm needing high resolution. Nevertheless, this method is well suited for correlation detection, sequence alignment test or pattern matching applications.

Optical system has also been used to implement reservoir computing. Reservoir computing and Liquid State Machines are models to solve classification problems and can be seen as "part" of neuromorphic architecture. Nevertheless, as this approach is often coupled with research to implement this model with analogic optical computing, it is integrated in this section.

Other options

Other options are possible as using electrical or thermal systems to find solutions of some differential equation problems.

New computing paradigm: quantum computing

Quantum computing is a new paradigm where quantum properties are used to provide a system with computing capacity. Today research in this field can be split in two categories:

- The "universal" quantum computers based on qubit and gates performing operation on these qubits. It uses two quantum properties, superposition (capacity to be at the same time in a superposition of 2 states) and entanglement (capacity to link the state of an element to the measure made on another element). From these properties, a mathematical model of universal quantum computer has been developed. In this model a system of qubits can be put in a state that represents the superposition of all the values of the computed function (i.e. this system has in "parallel" computed the values of a function for all the 2N inputs).

- The quantum annealers, or quantum simulator, mainly represented by the Dwave machine, which use quantum physics to escape from local minima in optimization functions using quantum fluctuations. This class of machines is limited to problems that can be modelized as minimization of function, like the travelling salesman, flow optimization, molecular simulation etc.

Most of the efforts target the first approach. Nevertheless, developing a physical system that behaves like the "universal" model is at the level of research and will need to solve hard problems such as the decoherence of the qubits, a reliable measurement system, error correction and the NxN interconnection between the qubits.

Conclusions and outlook

The deployment of “High-Performance Computing” is undergoing a significant change and the term ‘HPC’ no longer applies to only supercomputers in large datacentres but also a compute infrastructure supporting simulation, modelling and data analysis in a 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 as well as on HPC technology will expand from the current fields deploying HPC solutions to adjacent fields to address AI, Data Analytics and IoT-related challenges. This will influence the selection and definition of research priorities in the next SRA and this can only be effective and meaningful as the result of a true interdisciplinary effort.

Several workshops and collaborative sessions are planned throughout 2019. The analysis of a diverse set of “digital continuum use scenarios” will play a significant role in determining the research focal points for the next SRA.

Ultimately, the recommendations given in the next Agenda will serve as input to the research and innovation advisory process assisting the EuroHPC Joint Undertaking in the definition of its R&D&I work programme for the period 2021-2024 (and beyond).

Appendix

Glossary

AI	Artificial Intelligence
AIOTI	Alliance for the Internet of Things Innovation
AMBA	Advanced Microcontroller Bus Architecture
BDEC	Big Data and Extreme-scale Computing
BDVA	Big Data Value Association
CBRAM	conductive bridge random access memory
CMOS	Complementary Metal-Oxide-Semiconductor
CoE	Centre of Excellence (for Computing Applications)
CPS	Cyber- Physical System
CSA	Configurable Spatial Accelerator
CXL	Compute Express Link
DCP	digital continuum platform
DoE	Department of Energy
DPU	data processing unit
DSL	domain-specific language
E2E	end-to-end
EC	European Commission
EPI	European Processor Initiative
ETP4HPC	European Technology Platform for High Performance Computing
EU	European Union
EXDCI	European Extreme Data and Computing Initiative
FFT	Fast Fourier Transformation
FLOP	floating point operations
FP	Framework Programme / floating point
FPGA	Field Programmable Gate Array
GDPR	EU General Data Protection Regulation
GPGPU	general-purpose GPU - a graphics processing unit (GPU)
HBM	high bandwidth memory
HiPEAC	High Performance Embedded Architectures and Compilers
HMC	Hybrid Memory Cube
HPC	High-Performance Computing
HPDA	High-Performance Data Analytics
HW	hardware
IMC/PIM	In Memory Computing; Processor In Memory)
ISV	Independent Software Vendor
IT	Information Technology
MB	Mega Byte
MFF	Multiannual Financial Framework
MIMD	multiple instruction, multiple data
MRAM	Magnetic RAMs
MW	megawatt
NAND	resistive memory (memristor)
NN	neural network
NVM	Non-Volatile Memory
OpenCAPI	Open Coherent Accelerator Processor Interface
OxRAM	oxide based resistive memory
PCM	phase change memory,
PCRAM	Phase Change RAM
PRACE	Partnership for Advanced Computing in Europe
QoS	Quality of Service
R&D	Research and Development
R&I	Research and Innovation
RAM	Random-Access Memory
RRAM or ReRAM	metal oxide resistive random-access memory
SC	Supercomputing Conference
SDG	sustainable development goals
SIMD	single instruction, multiple data
SME	Small and Medium-Sized Enterprise
SRA	Strategic Research Agenda
STT-RAM	Spin-transfer torque magnetic random-access memory
SW	software
TCO	total cost of ownership
TPU	tensor processing unit
TSMC	Taiwan Semiconductor Manufacturing Company
UN	United Nations
YE	Year End

Acknowledgements

This document is a collective product, and it is a pleasure to thank the many people involved in producing it. We would like to give special thanks to all authors who provided invaluable insights and to the reviewers who helped us improve both its form and content.

Gabriel Antoniu, INRIA (BDVA)

Marc Asch, U-PICARDIE (BDEC-2)

Peter Bauer, ECMWF

Costas Bekas, IBM

Pascale Bernier-Bruna, ETP4HPC

Francois Bodin, IRISA

Laurent Cargemel, Atos

Paul Carpenter, BSC

Marc Duranton, CEA (HiPEAC)

Maike Gilliot, ETP4HPC

Hans-Christian Hoppe, INTEL

Jens Krueger, ITWM-FRAUNHOFER

Julian Kunkel, University of Reading

Erwin Laure, KTH

Jean-Francois Lavignon, TECHNOLOGY-STRATEGY

Guy Lonsdale, SCAPOS

Michael Malms, ETP4HPC

Fabio Martinelli, CNR (ECSO)

Sai Narasimhamurthy, SEGATE

Marcin Ostasz, BSC

Maria Perez, UPM (BDVA)

Dirk Pleiter, JSC

Andrea Reale, IBM (BDVA)

Pascale Rosse-Laurent, Atos