



HAL
open science

ETP4HPC's SRA 6 - Strategic Research Agenda for High Performance Computing in Europe

Sai Narasimhamurthy, Nico Mittenzwey, Fabrizio Magugliani, Marc Duranton, Craig Prunty, Pascale Rossé-Laurent, Manolis Marazakis, Paul Carpenter, Gabriel Antoniu, Sarah Neuwirth, et al.

► **To cite this version:**

Sai Narasimhamurthy, Nico Mittenzwey, Fabrizio Magugliani, Marc Duranton, Craig Prunty, et al.. ETP4HPC's SRA 6 - Strategic Research Agenda for High Performance Computing in Europe. Zenodo, 2024, ETP4HPC Strategic Research Agenda, 10.5281/zenodo.14268783 . hal-04846393

HAL Id: hal-04846393

<https://hal.science/hal-04846393v1>

Submitted on 18 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

ETP4HPC's SRA 6

Strategic Research Agenda for High-Performance Computing in Europe

European HPC Research Priorities for 2025 - 2029
December 2024



EUROPEAN TECHNOLOGY
PLATFORM FOR HIGH
PERFORMANCE COMPUTING

Table of contents

- Message from the ETP4HPC’s Chairman 5
- Glossary of Acronyms 6
- Executive Summary of Recommendations 8**
- 1. Goals and Objectives of SRA6 12**
- Why does the SRA exist and what are the primary objectives in SRA6?..... 13
- SRA6, A Community Perspective 13
- SRA6, A Policy Makers’ perspective..... 13
- Improvements in SRA6 14
- Better Structure and Organisation of SRA6 14
- SRA6 Release mechanism..... 14
- Seeking better linkages with EuroHPC and the EC 15
- Continued strong engagement with the ecosystem communities..... 15
- Strong focus on community feedback..... 15
- 2. HPC Policy Developments in Europe 16**
- Dealing with the “AI Explosion” 17
- Rise of Quantum Computing 17
- Advent of RISC-V..... 18
- The Race to Exascale 18
- Push towards energy efficiency & Green HPC..... 19
- Support for Destination Earth Flagship 19
- Other new notable Research and Innovation efforts 19
- 3. Advanced Computing: R&I Recommendations 20**
- Research Domain Recommendations..... 22
- Architecture and Hardware 22
- Software and Use 27
- Ecosystem technologies..... 34
- Thematic Trend Recommendations..... 35
- AI & Foundational Models..... 35
- Energy Efficiency & Sustainability..... 35
- 4. European HPC and AI Explosion..... 38**
- General Considerations 39
- Data 39
- AI for Science 39
- AI Services and Infrastructure 40
- Ecosystem Development 41

Top Technical considerations	42
Storage and Data Management.....	42
Lower precision	42
Opportunities for code generation/Optimization	43
Lessons from the Cloud dealing with AI	43
Other considerations.....	43
5. Post Exascale Vision & Challenges	44
Research Domains and Thematic Trends’ considerations	45
AI & Foundational Models’ considerations for Post Exascale.....	51
Scalability and Efficiency.....	51
Data Utilisation and Synthetic Data.....	51
Inference: Capacity Over Capability	51
Ethical Considerations and Explainability	51
6. Industrial HPC Usage in Europe.....	52
Introduction.....	53
Data Protection.....	53
Security considerations	54
Financial considerations and pricing	54
Legal considerations	54
Service level agreements.....	54
Interoperability with Cloud infrastructures.....	55
Software management	55
Potential for new pricing models	56
In support of European SME HPC Users.....	57
7. European HPC Technology SMEs.....	58
Introduction.....	59
Sharpening the tools	60
Conclusion	61
8. European Hardware Initiatives	62
European Processor Initiative.....	63
EUPILOT	64
EUPEX	65
RED SEA	66
eProcessor	67
Other activities	68
9. Conclusion and Acknowledgements	70
Research Domain White Papers	71
Thematic Trend White Papers	71

Table of Contents ■

Acknowledgements 72

APPENDIX: SRA6 Process & Structure 74

SRA6 Process 75

SRA6 Structure..... 76

 Research Domains & their Recommendations:..... 76

 Thematic Trends & their Recommendations:..... 76

Releasing “Federation” as an Agile white paper 78

Message from the ETP4HPC's Chairman

The publication of this new update of the Strategic Research Agenda (SRA) is an exciting event in the life of our association. As in previous editions, we stress the importance of continuing to develop new technologies for HPC. These efforts need to be intensified and extended to new application areas beyond traditional HPC. HPC has been an important tool for research and industry for decades, enabling major advances across the spectrum from basic research to industrial applications. By integrating new approaches (such as Artificial Intelligence, Quantum Computing), we will be able to push the frontiers of computing to new limits.

Thanks to the EuroHPC Joint Undertaking, Europe is currently deploying its first exascale system and preparing for the second. In addition, the first AI factories are being announced and will soon complete the impressive resource available to the European user community.

However, despite the encouraging results of the EuroHPC and Horizon Europe Research and Innovation programmes, much of the technology used still comes from outside Europe. This external dependency is a threat to Europe's sovereignty that needs to be addressed. This SRA provides a roadmap for the development of the technology that will power the systems of the post-exascale era by the end of this decade. These new developments are needed to address the major societal challenges we face in the various fields: Environment, Health, Energy...

ETP4HPC is indebted to the more than two hundred experts who contributed to this SRA. We are also particularly grateful to the working group leaders who helped to synthesise the contributions and to the editorial team who produced this document. This SRA will be completed with the forthcoming publication of white papers, where we'll focus on the specific topics developed by our expert groups. We expect the recommendations we make here to be reflected in future research and innovation actions.

ETP4HPC Chairman

Jean-Pierre Panziera

Glossary of Acronyms

CCI - Computing Continuum Initiative
CI/CD - Continuous Integration/Continuous Delivery
DDR - Double Data Rate (Memory)
DPU - Data processing Unit
EaaS - Exascale as a Service
EDA - Electronic Design Automation
EOSC - European Open Science Cloud
EMTS - Electromagnetic Transient Simulation
EPI - European Processor Initiative
FPA - Framework Partnership Agreement
HBM - High Bandwidth Memory
HCI - Hyper Converged Infrastructure
IMC - In Memory Computing
IPU - Infrastructure Processing Unit
SDV - Software Development Vehicle
SRA - Strategic Research Agenda
cPPP - Contractual Public Private Partnerships
RIAG - Research and Innovation Advisory Group
MSRIA - Multi Annual Strategic Research and Innovation Agenda
NIC - Network Interface Card
NMC - Near Memory Computing
NoC - Network On Chip
OoO - Out of Order
QCS - Quantum Computing System
R&I - Research and Innovation
RISC - Reduced Instruction Set Computing
RTL - Register Transfer Level
SMT - Single Instruction Multiple Threads
SNN - Spiking Neural Networks
SoC - System on Chip
SRAM - Static Random Access Memory
TRL - Technology Readiness Level
URLLC - Ultra Reliable Low Latency Communication
VC – Venture Capital

Executive Summary of Recommendations

The background is a deep blue with a subtle grid. It features glowing white and light blue circuit-like lines that meander across the page. There are several clusters of hexagons, some of which are filled with a lighter blue color. Small circles, some solid and some hollow, are scattered throughout, resembling data points or nodes in a network. The overall aesthetic is clean, modern, and technological.

The European Strategic Research Agenda 6, or the SRA6 reflects the state of the European Advanced Computing (AC) ecosystem¹, which includes High Performance Computing and associated areas such as Artificial Intelligence and Quantum Computing interactions - edited and published by the European Technology Platform for High Performance Computing (ETP4HPC). The publication of Strategic Research Agendas has been one of the primary goals of the ETP4HPC, which is the largest industry-led think tank for High Performance Computing in Europe, since its inception in 2012. The Strategic Research Agendas have been published approximately every two years, with the last SRA5 published towards the end of the COVID pandemic in September 2022. A lot has changed in the last couple of years, and that primarily also reflects the timing of the publication of these strategic research agendas, which captures the state of European HPC (& related) technologies at every 2 year “checkpoints”, which we think is a reasonable time window to see measurable changes in the HPC landscape.

The following is an executive summary of ETP4HPC’s key research and innovation recommendations for SRA6 in the broad areas of HPC technology, including in AI, ongoing programs focussed on EU HPC sovereignty and HPC technology SMEs.

- Development of a **strong data ecosystem for the sharing of data** across the EuroHPC federation of supercomputers used for AI models across various scientific domains, is urgently needed. There is also a need for the development of “ScienceGPTs” that can help cross disciplinary science and innovation in Europe. As part of this, a **multi-year plan and roadmap specifically in the area of AI for Science in Europe** needs to be developed including the possibility for multiple smaller projects if and as necessary. All this has to be done now if Europe doesn’t want to lag behind in AI for Science.
- Innovation in the area of software and middleware methodologies enabling **seamless integration of our upcoming Federated HPC/AI infrastructure**, with existing AI application and tooling ecosystem, and cloud infrastructures is very much needed. Innovations are also needed in quickly addressing increased energy requirements (& deeply studying the environmental and sustainability impacts) due to foundational models’ running on these federated EuroHPC/AI infrastructures. Even though the problem is well recognised, not much is happening yet.
- Developing and nurturing programs that support **targeted dissemination of EuroHPC AI initiatives** into the AI+HPC SME community with clearly laid out benefits is highly necessary. In similar vein, creation of a program similar to EUMasters4HPC, such as “**EUMaster4AI**” with focus not just on AI, but also on the usage and exploitation of HPC infrastructures for AI, utilising AI for Science, needs to be initiated urgently.
- We recommend **re-invigorated efforts in the area of data storage, data management and I/O** innovations in light of AI developments in the upcoming work programmes, which seems to be completely lost after the H2020 programs. This will be a critical centre piece in the AI revolution.

¹ Please note that we will use the term “HPC” in this document to include the whole of the Advanced computing landscape, that includes HPC and all associated areas, in this SRA

Executive Summary of Recommendations ■

- We recommend developing a **strategy for better tracking and adapting to the trend of lower precision arithmetic** both from the perspective of applications and algorithms and European hardware developments. More R&I in **exploiting AI assisted code generation methodologies** suitable in the context of HPC applications and systems software, and, the **reuse and adaptation of tools developed for the Cloud** for AI, for federated HPC infrastructures is also necessary.
- We recommend the **promotion of chiplet and interposer technology** ecosystem in Europe allowing industrials (incl. SMEs and startups) to propose innovative IPs that could end up in real HPC machines in a cost competitive fashion.
- Even though **FPGA technology** is fairly mature, we recommend addressing the challenges of **their large-scale deployment** and pushing to **support a more user-friendly software ecosystem**. AI focussed infrastructure needs have now presented a good opportunity for looking into these potentially more energy efficient architectures.
- For major hardware programs focussed on European sovereignty, we recommend better ways to **orchestrate national and EC funding** or provide solid mitigation measures for eventualities in national co-funding unavailability. **More flexibility in project timelines and budgets** is also needed to accommodate unforeseen delays and adjustments. Facilitating better access to world-class IP is also essential for these programmes.
- A very **thorough and continuous assessment of the supply chain is urgently needed** (more urgent after recent political changes globally) in the current unpredictable global socio-political environment, striking the correct balance between localising and diversifying the supply chain. For a 3-year project with multiple tape-outs, the socio-political situation changes make it hard to predict the price and turnaround time for tape-outs 3-4 years in the future! It will make it even harder now.
- We recommend a clearer definition of longer-term support for continued activities for the **EPI processor verticals** with a better indication of long term higher funding levels that would allow these activities to be executed in a timely way **with concrete commercial results**. In RISC-V, whilst the DARE FPA and the associated Specific Grant Agreements have been very ambitious pushing Europe in the direction of sovereignty, we recommend the development of a **well-structured seeding and deployment plan** in the European industrial and academic environments, to ensure the success of this ambitious project.
- We recommend further development (through further enabling adoption and R&I) of Europe's unique excellence in the development of the **Modular Supercomputing/Systems Architecture (MSA)** as a flexible and powerful way for the development of Supercomputers. We have now developed this expertise after more than a decade of intensive research. Projects around this concept need to be bolstered.
- Specifically for our SMEs, we recommend **financial management schemes with 100% reimbursement rates**, allowing for flexible contribution in R&I actions (not applying for example, minimum contribution rules) and creating special "SME Innovate" calls with fixed funding targeted towards SMEs supporting the development of disruptive technologies and an associated software ecosystem. Reservations can also be initiated for **European SME participation in some procurement calls**.

- Policy makers need to seriously start thinking about **developing more avenues for access to capital for HPC SMEs** (access to finance including and beyond VC, loans debt-financing, guarantees, & other financial instruments) as these HPC SMEs continue to be hit by cash crunches since its not typically an area where investors can expect a 10x return in 5 years! Europe will continue to be the backwaters of cutting-edge technology startups if not done so. We also recommend providing better support for SMEs (e.g. via tools and initiatives such as Horizon IP Scan) to handle the complex and costly procedures involved in **filing IPR applications across fragmented national systems**.



1. Goals and Objectives of SRA6



Why does the SRA exist and what are the primary objectives in SRA6?

The European HPC Strategic Research Agenda is intended to be *the* go-to document to capture the state of and provide a roadmap for Advanced Computing in Europe. The SRA6 is also intended to be fully vendor neutral and reflects the joint contributions of top Advanced Computing experts in Europe coming from large European HPC technology providers, international companies, European Small and Medium companies, European research organisations and European academic institutions.

The SRA6 is intended to be referred by the HPC community within Europe and globally to get a snapshot status of Advanced Computing technology and its ecosystem in Europe.

The SRA6 is also intended to be a guide for policy makers in Europe on getting a take on the Advanced Computing Research and Innovation (R&I) landscape and understanding the R&I priorities coming from top European HPC experts.

SRA6, A Community Perspective

The SRA6 is an agenda that presents a holistic view of the Advanced Computing ecosystem in Europe. The agenda also provides global visibility to “Advanced Computing in Europe” corresponding to the various Research and Innovation topics and the associated ecosystem.

Advanced computing in general and HPC in particular consists of multiple research areas, many of which are evolving very fast. It is very hard to infer the developments in all these research areas in one piece of writing, let alone keep pace with them all - be it programming models, hardware technologies, systems software, AI, Quantum, HPC data management, system architectures etc! The SRA6 aims to connect all these domains into a coherent picture in one set of documents. The SRA6 also provides a technology generalist view into the latest trends and themes in the area of Advanced Computing. The work aims to help the technology community understand and structure the Advanced Computing world.

HPC applications are evolving to keep pace and exploit the various new innovations in the area of HPC software and hardware infrastructures. There needs to be good synchronisation between the application communities and the technology developments and have a closed loop feedback mechanism between the two so as to make sure that the applications get the very best from what the technologies have to offer. The SRA6 is a means of connecting the Advanced Computing applications and use cases community and having them work closely with HPC technology experts.

Finally, HPC/AC is just one piece of the “Digital Continuum” which consists of various actors in the ecosystem such as those from the Edge, Cloud, Cybersecurity, AI/Robotics, wireless networking etc. There is a need not just to keep track of the developments in these areas but also to closely synchronise and work with them to come up with Research priorities in AI/HPC. For one, we don't want to reinvent in HPC the areas that are already well covered by these ecosystems. Dialogue and synchronisation with these communities help to develop coordinated efforts and action plans within Europe beneficial to society, industry and science. This is also one of the objectives of the SRA with a view toward community impacts.

SRA6, A Policy Makers' perspective

The SRAs in the past have influenced HPC Research and Innovation Work programmes under the Horizon 2020 Programme in the context of the cPPPs (contractual Public Private Partnerships). With the formation of the EuroHPC Joint Undertaking in 2018 and its RIAG (Research and Innovation

1. Goals and Objectives of SRA6 ■

Advisory Group), the SRAs have pivoted to accommodate this positive change and focus in the European HPC ecosystem.

One of the primary goals of SRA6 is to provide technical advice and inputs into EuroHPC on Research and Innovation programs and priorities. This will be done through the EuroHPC Research and Innovation Advisory Group (RIAG) with whom we have synchronised closely even while developing the SRA and we hope that this will help the upcoming EuroHPC Multi-Annual Strategic Research and Innovation Agenda (MSRIA). The SRA6 also aims to provide technical inputs and advice to the European Commission/DG-CNECT (Directorate-General for Communications Networks, Content and Technology) on Advanced Computing Research and Innovation status and priorities. The SRA6 is also expected to be referred to by other Joint Undertakings in Europe in the technology domain to develop their respective Strategic Research Agendas.

Improvements in SRA6

In view of the changing HPC ecosystem and based on some of the feedback we have received from the community during the SRAs, we intend to make some changes to the process and structure of the SRAs.

Better Structure and Organisation of SRA6

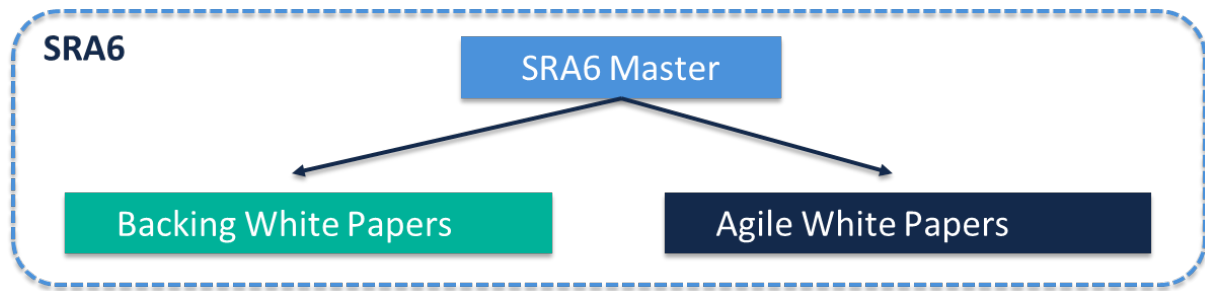
In the past we released the SRA as one big document. The previous SRA5² is an example. The SRA5 whilst being very detailed and exhaustive was also very long. This prompted us to listen to our readers' feedback that it would be even better if we could make it a bit more concise and pithier. However, the trick is to try and do that without losing any of the key points and the information that we needed the audience to be aware of. We hence have taken the approach of having this "Master" document which is concise and then the work and progress of each of the technical areas would be presented as a set of "Backing" White Papers that would then be summarised in the Master. These White Papers which will be released at almost the same time as the Master would then go into more detail in specific subject areas that can be referred by the reader based on his interest and area of expertise.

The Master would have a summary of the key outcomes of the work of the different working groups involved in the SRA6, and would suffice for most of the readers, especially policy makers that would like a broader view before delving into details on the subjects of their choice.

SRA6 Release mechanism

The release mechanism would also entail periodic ETP4HPC "Agile" White Paper releases to keep pace with the rapid advances in the HPC technology and ecosystem *after* the release of the SRA6 Master and the first set of white papers - with adequate referencing back to the SRA6 Master. This will incorporate any new developments and discussions post SRA6 Master and Backing White Papers' release.

² https://etp4hpc.eu/download/60/sra/4321/etp4hpc_sra-5_2022.pdf



This SRA6 Master will be edited to take account of the most important new developments referenced in the Agile White Papers and will be a “living” document - with the most stable snapshot available on the ETP4HPC website.

This flexible modular release mechanism, we believe, will never make the SRA6 out of date at any time and make it have the right level of detail as needed by different categories of readers.

Seeking better linkages with EuroHPC and the EC

We started early work on SRA6 through agile interactions with the EuroHPC JU RIAG and the EC. We provided some early feedback into the MSRIA 2024 whilst we started developing the SRA6. We are now in close communication with key members of the RIAG who are included as observers in the process of developing the SRA6. We are also closely connected and had dialogues with the EC whilst we developed this SRA. We hence foresee an even closer connection between the SRA and some of the research agendas and associated activities from the EuroHPC JU. These strong interlocks and dialogues, we believe, will improve the quality and relevance of top-level research agendas and work programmes.

It is to be noted that the SRA6 is independent work and offers independent expert opinion from the SRA Core team.

Continued strong engagement with the ecosystem communities

The SRA5 covers some of the early work done with the ecosystems adjoining HPC through the “Transcontinuum Initiative” (TCI).³The SRA6 extends and expands this collaboration through the SRA6 teams’ engagement with new initiatives such as the Computing Continuum Initiative (CCI) spearheaded by HiPEAC. We also work closely with the TCI and CCI communities in developing the content for the SRA as we work towards synchronising our visions. We not only did seek inputs for SRA6 from the ecosystem, but also actively contributed to the discussions and agendas of the ecosystem.

Strong focus on community feedback

We will continue to gather feedback on the readership of the SRA through the newly updated ETP4HPC website, which will be used to continuously improve the quality and relevance of the SRA6 associated materials (e.g.: Agile White Papers) for the various communities and readers.

³ [TransContinuum Initiative: joint Vision document signed by 8 European associations – ETP4HPC](#)

2. HPC Policy Developments in Europe



The following are some of the key updates since the last SRA at the HPC policy makers level to be executed through the EuroHPC Joint Undertaking. Some of the key policy updates that we presented in SRA5 still remain relevant, especially the continued push towards digital autonomy and sovereignty. Just to reiterate, the EuroHPC JU was established in 2018 with the following objectives⁴:

The main objectives of EuroHPC Joint Undertaking are to develop, deploy, extend and maintain the EU super-computing, quantum computing and data infrastructure ecosystem; support the development of super-computing systems' components, technologies and knowledge; widen the use of that super-computing infrastructure; and support the development of key HPC skills for European science and industry.

There have been a few notable updates in the last couple of years driving some the policies against the backdrop of which this SRA has been developed.

Dealing with the “AI Explosion”

The objectives of the EuroHPC JU are now expanded to deal with AI based on the decision by the council of the European Union. EuroHPC will now help to develop and operate “AI Factories” to support the AI ecosystem in Europe that includes the support of SMEs⁵. This is a follow up of the state of the union address by President Ursula Von der Leyen⁶ which committed to helping AI innovation (& Foundation models developed by European organisations) in Europe through the EuroHPC supercomputing infrastructure. The competition to develop large scale models in Europe was also launched⁷ with support from EuroHPC JU machines.

EuroHPC is aiming to support industries, SMEs and public sector organisations who want to run ethical AI, Machine Learning and Data Intensive applications through access to the EuroHPC petascale (Vega, Carolina, MeluXina) and pre-exascale (LUMI, Leonardo and Marenostrum5) systems with 1-year allocations starting mid 2024.^{8,9}

There is also now an effort to support AI Applications on HPC infrastructure through “AI Support Centres” helping EuroHPC AI experts to scale up their AI workflows using EuroHPC infrastructure¹⁰.

Rise of Quantum Computing

The EuroHPC JU signed hosting agreements with six sites across Europe to host & operate EuroHPC quantum computers in June 2023, in Czechia, France, Germany, Italy, Poland and Spain¹¹ that will be tied to EuroHPC JU supercomputers. Calls for tender for the installation of the Polish Quantum

⁴ [Homepage - European Commission \(europa.eu\)](https://european-council.europa.eu/media/en/press-communications/inline-photos/attachment-data/file/attachment)

⁵ [EuroHPC Expands Objectives to Include AI Factories for Boosting EU's AI Ecosystem \(hpcwire.com\)](https://www.hpcwire.com/news/eurohpc-expands-objectives-to-include-ai-factories-for-boosting-eu-ai-ecosystem/)

⁶ [Commission opens access to EU supercomputers to speed up AI \(europa.eu\)](https://european-council.europa.eu/media/en/press-communications/inline-photos/attachment-data/file/attachment)

⁷ [Commission opens access to EU supercomputers to speed up AI \(europa.eu\)](https://european-council.europa.eu/media/en/press-communications/inline-photos/attachment-data/file/attachment)

⁸ [EuroHPC JU Access Call for AI and Data-Intensive Applications - European Commission \(europa.eu\)](https://european-council.europa.eu/media/en/press-communications/inline-photos/attachment-data/file/attachment)

⁹ [Our supercomputers - European Commission \(europa.eu\)](https://european-council.europa.eu/media/en/press-communications/inline-photos/attachment-data/file/attachment)

¹⁰ [Open Call to Support HPC-powered Artificial Intelligence \(AI\) Applications - European Commission \(europa.eu\)](https://european-council.europa.eu/media/en/press-communications/inline-photos/attachment-data/file/attachment)

¹¹ [One step closer to European quantum computing: The EuroHPC JU signs hosting agreements for six quantum computers - European Commission \(europa.eu\)](https://european-council.europa.eu/media/en/press-communications/inline-photos/attachment-data/file/attachment)

2. HPC Policy Developments in Europe ■

Computer¹² and the German Quantum Computer¹³ was launched in late 2023. The tender for the Czech and the French Quantum computers was launched in early 2024^{14,15}. Calls to establish two Quantum Centres of excellence for applications was also recently closed¹⁶.

Advent of RISC-V

There is a strong belief now in Europe that RISC-V provides an alternative to proprietary hardware for processors and accelerators developed outside of Europe. The European Chips Act indeed identified RISC-V as an area that Europe needs to invest in to achieve semiconductor sovereignty¹⁷. In view of the push towards RISC-V, the EuroHPC JU launched a call¹⁸ to support a Framework Partnership Agreement (FPA) to develop a large-scale HPC chip ecosystem based on RISC-V in 2023. The FPA will ensure the implementation of several complementary Research & Innovation (R&I) or Innovation actions to develop the various pieces of technology needed through the SGAs or Specific Grant Agreements¹⁹. The first of the specific grant agreements is now launched as of June 2024²⁰. The aim of this SGA is to design and deliver energy efficient tape-outs of a general-purpose processor and of two accelerators, an Artificial Intelligence (AI) Accelerator and a Vector Accelerator, for HPC based on RISC-V with advanced memory interfaces.

The Race to Exascale

In October 2023, the Franco-German consortium won the contract to build Europe's very first Exascale system at FZJ - which is called JUPITER²¹, ushering Europe into the Exascale era²². The machine plans to use European Processors developed by a European SME and which has been supported through the efforts of the Framework Partnership Agreement, the European Processor Initiative (EPI)²³. A module of JUPITER, JEDI is now ranked first in GREEN 500²⁴ as of May 2024.

The Jules-Verne consortium in France will host the second Exascale supercomputer²⁵ in Europe managed by GENCI and operated at CEA - which will be called "Alice Recoque".

¹² [EuroHPC JU Launches Procurement for EuroQCS-Poland - European Commission \(europa.eu\)](#)

¹³ [EuroHPC JU launches procurement for a new quantum computer in Germany - European Commission \(europa.eu\)](#)

¹⁴ [EuroHPC JU announced a tender for the supplier of the quantum computer of the LUMI-Q consortium - IT4Innovations](#)

¹⁵ [EuroHPC JU Launches Procurement for a New Quantum Computer in France \(hpcwire.com\)](#)

¹⁶ [European Quantum Excellence Centres \(QECs\) in applications for science and industry - European Commission \(europa.eu\)](#)

¹⁷ [European Chips Act - European Commission \(europa.eu\)](#)

¹⁸ [New call for developing an HPC ecosystem based on RISC-V - European Commission \(europa.eu\)](#)

¹⁹ Some more information is covered in Chapter 8 under "Other activities"

²⁰ [EU Funding & Tenders Portal \(europa.eu\)](#)

²¹ [JUPITER Technical Overview \(fz-juelich.de\)](#)

²² [Procurement contract for JUPITER, the first European exascale supercomputer, is signed - European Commission \(europa.eu\)](#)

²³ [Home - European Processor Initiative \(european-processor-initiative.eu\)](#)

²⁴ [JUPITER Exascale Supercomputer Claims 1st Place in GREEN500 \(hpcwire.com\)](#)

²⁵ [The Jules Verne Consortium Will Host the New EuroHPC Exascale Supercomputer in France - European Commission \(europa.eu\)](#)

EuroHPC JU at this time has thus procured 9 supercomputers across Europe.²⁶

Push towards energy efficiency & Green HPC

In line with supporting the European priority on the Green Deal, there is a need to address energy efficiency and environment sustainability in the design and operation of HPC systems. A call was announced in late 2023 and recently closed to develop energy efficient HPC software technologies suitable for Exascale and Post Exascale.²⁷

Support for Destination Earth Flagship

Destination Earth is a flagship initiative of the EC to develop a highly accurate digital model of the Earth supporting the European Green Deal and the European Digital Strategy²⁸. Access to a federation of EuroHPC Supercomputers is now planned for running the models and simulations of Destination Earth and satisfy its computing requirements through an interactive computing platform. ECMWF, which is responsible for creating the digital twins of the Earth in Destination Earth, has also been consulting with ETP4HPC and its ecosystem partners on various aspects such as Compute Federation, Data Management, etc.

Other new notable Research and Innovation efforts

Other new research and innovation efforts sought by EuroHPC as a backdrop of this SRA are:

1. Continued Innovation in high bandwidth networking for HPC²⁹
2. New Centres of Excellence driving continued innovations in applications seeking benefits from Exascale and Post Exascale Technologies³⁰
3. Seeking global partnerships (EU-India³¹ in 2024 following EU-Japan in 2023) to collaborate on HPC application areas and exchange of know-how

EuroHPC also has actions continuing to develop National Competence Centres³² and Support for SMEs³³.

The above thus summarises the policy context in which this SRA6 was authored.

²⁶ [Our supercomputers - European Commission \(europa.eu\)](#)

²⁷ [Energy Efficient Technologies in HPC - European Commission \(europa.eu\)](#)

²⁸ [Destination Earth | Shaping Europe's digital future \(europa.eu\)](#)

²⁹ [Innovation Action in Low Latency and High Bandwidth Interconnects - European Commission \(europa.eu\)](#)

³⁰ [Centres of Excellence for Exascale HPC Applications - European Commission \(europa.eu\)](#)

³¹ [EU-India Partnership - European Commission \(europa.eu\)](#)

³² [National Competence Centres for High Performance Computing - European Commission \(europa.eu\)](#)

³³ [Supporting competitiveness and innovation potential of SMEs - European Commission \(europa.eu\)](#)

3. Advanced Computing: R&I Recommendations

A futuristic server room with blue lighting and glass railings. The room is filled with server racks, each with numerous small lights and displays. The perspective is from a walkway with a glass railing, looking down a long aisle of server racks. The lighting is predominantly blue, creating a high-tech, digital atmosphere. The racks are filled with various components, and the overall scene conveys a sense of advanced computing and data processing.

The following sections summarise the R&I recommendations we have for the various Research Domains and Thematic Trends.³⁴

For the Research Domain recommendations, we group them into:

- **Architecture and Hardware:** Here we discuss the recommendations from System Architecture, System Hardware Components and Non-Conventional Architectures Research Domains. We include discussions of our perceived TRL (Technology Readiness Level) levels for these subjects as it's relevant for this topic. Low TRL levels indicate that the Architecture/Hardware subject is currently primarily in research. Mid TRL indicates that the Architecture/Hardware subject is in research with some active development. High TRL indicates that the subject is fairly mature and has started to be commercialised.
- **Software and usage:** We discuss here the recommendations from System Software, Programming Environments, Mathematics and Algorithms and I/O & Storage. We include I/O & Storage also in this category even though there are storage hardware aspects, as the predominant discussions and recommendations are around the usage and exploitation of such hardware through the I/O software stack & I/O tools.
- **Ecosystem:** This includes input from the ecosystem areas for which HPC is relevant.

After the above, we summarise the R&I Recommendations for the thematic trends.

We encourage the reader to refer to the respective Research Domain and Thematic Trend White Papers for more details which led to these recommendations.

³⁴ Please see the Appendix on definitions of the various Research Domains and Thematic Trends.

Research Domain Recommendations

Architecture and Hardware

System Architecture

Disaggregated Resources and Dynamic Resource Management (Mid TRL)

Dynamic resource management and disaggregated resources are essential for optimising HPC and AI systems' performance and efficiency. The increasing heterogeneity and dynamic nature of workloads necessitate dynamic, adaptive methods for resource allocation. Research in this area is vital to developing flexible, energy-efficient systems capable of meeting the diverse and evolving demands of exascale computing environments.

In federated computing infrastructures, this can include advanced fault-tolerance mechanisms (lower TRL), including automated recovery processes and resource reallocation, as well as re-routing of applications between centres. Additionally, it facilitates real-time resource allocation and dynamic management, crucial for urgent decision-making.

Energy and Resource Efficiency (High TRL)

Research should focus on optimising resource usage and developing highly efficient algorithms to ensure components operate effectively. Investing in advanced cooling technologies, particularly two-phase direct cooling, is crucial for managing high heat loads and reducing energy consumption. Implementing systems to capture and reuse heat for applications like district heating will enhance overall energy efficiency. Advanced power management techniques that dynamically adjust power usage based on workload demands and energy prices, reflecting renewable energy availability, are essential. Enhancing federated computing models to relocate jobs to data centres with lower energy costs will optimise costs and increase the use of renewable energy. Additionally, improving resource utilisation through disaggregated resources and scalable pooling solutions will further increase energy efficiency while reducing hardware requirements. Addressing these areas will significantly enhance the energy and resource efficiency of next-generation HPC and AI systems.

Further the role of programming models and libraries³⁵ to effectively support application developers in developing energy efficient applications becomes important in this context. Lastly the carbon footprint throughout the life cycle of the infrastructure (from procurement, running, re-use & decommissioning) has to be considered rather than just energy being a prime determinant³⁶.

Though the above System Architecture topics have been addressed to an extent in some of the H2020³⁷ and recent EuroHPC^{38,39} projects, we recommend that the above topics be more critically studied in light of AI hardware, federations and the need for sustainability, as a key action.

³⁵ Programming Models is a separate research domain in this SRA6 with separate recommendations

³⁶ Energy efficiency and sustainability is a separate research domain in this SRA6 with separate recommendations

³⁷ [DEEP-Projects](#)

³⁸ [Projects – DEEP-Projects](#)

³⁹ [REGALE – Open Architecture for Exascale Supercomputers](#)

System Hardware Components

Further enhancing “Byte per Flop” ratio [Mid TRL]

In the area of Memory evolution, diverse heterogeneous memory solutions are available now (MRDIMM⁴⁰, HBM⁴¹, DDR, etc). HBM is costly, difficult to buy currently and is power intensive. Alternatives and usage of such heterogeneous memories need to be further explored [Mid TRL]. Also, the capabilities of PIM⁴² are limited today for complex processing, and this has to be addressed to make PIM a viable proposition for HPC applications [Mid TRL]

We are also seeing the trend of coupling nodes into virtual “supernodes”. Appropriate Interconnects for such “super nodes” need to be looked into [Mid/High TRL]. Also, as AI models grow in complexity, the number of Network Interface Cards (NICs) per node is increasing for creating such “supernodes”, further emphasising the need for efficient data movement and memory coherence and orchestration. Existing concepts implemented by technology vendors and associated standards continue to be very promising in the realm of Unified Memory architectures and needs to be further explored [High TRL]. We need to consider the potential of Omni-path⁴³ evolving into Ultra Ethernet⁴⁴, a new high-performance networking standard aimed at surpassing InfiniBand's dominance in AI and HPC applications [Mid TRL]. Transparent management of memory and data movements in these virtual supernodes, thanks to specialised network controllers (Infrastructure Processing Units, Data Processing Units, etc.), needs to be further explored. In this context, software to support all hardware innovations in as transparent as possible fashion, leading to real hardware abstraction for the non-system programmer needs to be considered.

RISC-V [Mid-High TRL] & Accelerators [High TRL]

There is a need to continue European funding programs addressing RISC-V challenges for HPC applications. This should include the support of startups and innovative SMEs in this field, by easing the pain of very costly leading-edge design and silicon required for HPC applications⁴⁵.

Usage of more specialised accelerators such as Infrastructure & Data Processing Units (IPUs and DPUs), FPGAs, etc to support AI (and other tasks) should be further addressed. The main roadblock is the software to efficiently use them.

Integration of computing, memories and interconnects within the same package [Mid TRL]

A chiplet and interposer technology ecosystem should be created (interoperability between providers of chiplets, common HW interfaces, etc) with some efforts already in this direction⁴⁶. This is an action that should be promoted by Europe, allowing its diversity of SMEs, companies, start-ups to propose innovative IPs that could end-up in real HPC machines at minimum cost using this approach of interposers supporting a variety of IPs from different providers. The difficulty will be to convince non-European providers (High performance cores, memories) to provide their IPs in chiplet compatible form. Development of active interposers (Power conversion, health monitoring, IOs) and photonic

⁴⁰ [DDR5 MRDIMM | Multiplexed Rank DIMM | Micron Technology Inc.](#)

⁴¹ High Bandwidth Memory

⁴² Processing in Memory

⁴³ [Intel® Omni-path Architecture: Enabling Scalable, High Performance Fabrics | IEEE Conference Publication | IEEE Xplore](#)

⁴⁴ [Ultra Ethernet Consortium](#)

⁴⁵ Separately, a different chapter on ongoing European Hardware Initiatives in this SRA6 master covers perspectives on the ARM based European Processor Initiative

⁴⁶ <https://www.uciexpress.org/>

3. Advanced Computing: R&I Recommendations ■

interposers to improve Bandwidth Per Watt is also needed. In the context of Network on Chips (NoC), a 2D mesh is used in many existing NoCs. However, 3D integration should be considered for large future systems.

Reliability [High TRL] & Security [High TRL]

There is a continuing need to develop robust real-time health monitoring systems (due to the increase of components in post-exascale systems) to detect corruptions, and there is now an opportunity to look into various methodologies (incl. those driven by AI) to handle them.

Continued exploration of DPUs & IPU's to offload tasks (data management, compute and inter-node communication) and security is also necessary, whilst noting that their interoperability with existing hardware/software remains a concern.

Some of these hardware topics have received some focus in light of continuing efforts on European hardware sovereignty (see the chapter on European Hardware Initiatives). *However with the changing geopolitical environment, we recommend a thorough assessment of supply chain considerations to strike the right balance between localising and diversifying the supply chain, including trusted alternatives outside of Europe - as a key action.* The topics that need to be analysed in this context include, but not limited to, Advanced Packaging Technology (which is high impact, with increasing relevance over time, and low-cost relative to advanced node semiconductor manufacturing), memory technology developed in Europe, EDA (Electronic Design Automation) tools which are a US duopoly at the moment, and European Manufacturing capability (already supported by the European Chips Act⁴⁷).

Non-Conventional Architectures

FPGA & dataflow Architectures

FPGAs represent a mature technology with a Technology Readiness Level (TRL) set to the highest value when considering single devices; more challenges still exist when they have to be massively deployed (at-scale). Large availability of FPGAs requires promoting their adoption in large experimental testbeds (TRLs 5-6) e.g., dedicated experimental partitions on supercomputers, including by offering sufficient training and support for application developers.

There is a need to push the development of a more user-friendly SW ecosystem in a fashion similar to CUDA/ROCm⁴⁸ ecosystems for GPUs, developing libraries and SW abstractions that ease the programming task. Creating and spreading out open-source (soft) cores to enable end users to develop applications around this technology is also needed.

The growing pressure on creating AI-focused HPC infrastructures agencies, as a means to support the development of large AI models (e.g., Foundation Models) is a good chance for experimenting with more energy efficient architectures (static dataflow style, FPGA-based, etc.) moving to middle-high TRL. There is a need to enhance the SW ecosystem at scale, including middleware. Large deployments of enhanced SmartNICs and Data Processing Units should also be supported.

⁴⁷ [European Chips Act - European Commission \(europa.eu\)](https://european-council.europa.eu/media/en/press-communications/inline-photos/attachment-data/file/attachment)

⁴⁸ [ROCm Software \(amd.com\)](https://www.amd.com/en/rocm)

In-memory and near-memory (IMC and NMC) computing

There is a need to support the development of novel hardware addressing the spectrum of needs (from edge to datacentre), extending beyond the current or soon to be commercially available (TRLs 7-9) approaches targeting edge applications. These need to evolve towards higher-performance data centre/HPC applications with the exploitation of non-volatile memories for IMC (In Memory Computing) with increased density and performance, including 3D memory and heterogeneous integration (TRLs 4-6). There is a need to support the development of software, including programming models, compilers, etc, to optimally exploit and integrate the IMC and NMC (Near Memory Computing) hardware capabilities with prototype demonstrations, including integration in open ISAs such as RISC-V (TRLs 4-6). IMC and NMC technology can be promoted by creation and access to larger-scale deployments demonstrating HPC/AI applications acceleration, fostering a European ecosystem of partners to drive innovation at all levels, from memory technology companies, accelerator technology companies, to academic and scientific users.

Neuromorphic Computing

The development of intermediate representations and compiler standards for the configuration and deployment of SNN (Spiking Neural Networks) models on heterogeneous neuromorphic platforms.⁴⁹ (TRLs 4-6) needs to be supported. Multi-platform benchmarks that can be used to make fair comparisons of available neuromorphic solutions should be developed and maintained. There is also a need to stimulate the development of unified and standardised programming models and software frameworks to enhance interoperability and facilitate the porting of SNN models across different neuromorphic platforms. Further, the creation of environments/programming models and tools which support the porting of applications written for conventional architectures on neuromorphic platforms needs to be supported. It should be noted that, currently, there is a lack of accessibility to neuromorphic HW, which is primarily available to a limited number of research teams. Only a few companies, such as SpinCloud, Brainchip, and Synsense, offer TRLs 8-9 neuromorphic platforms. To address this challenge, we have identified **three strategies**:

- Promote the development of open source HW architectures deployable on FPGAs.
- Support the development of cloud-based prototyping platforms hosting the major available neuromorphic architectures accessible remotely by developers.
- Engage leading players in the silicon industry in the development of neuromorphic SoC architectures.

Massive data archival and storage

Given the exploding needs of HPC and AI in terms of data, whether as input, as intermediate results, or as outputs in their workflows, there is a need to ensure a sustainable and seamless integration of current and future massive archival storage solutions, as well as promoting the research for improved and/or new storage hardware technologies spanning TRLs 4-9. Work previously done on hierarchical storage management for Exascale compute in the SAGE, SAGE2 and IO-SEA EU research projects should be extended to account for the criticality to keep an affordable, secure, and sustainable record of quickly expanding datasets.

⁴⁹ [Exploring Neuromorphic Computing Based on Spiking Neural Networks: Algorithms to Hardware | ACM Computing Surveys](#)

3. Advanced Computing: R&I Recommendations ■

DNA Computing

DNA computing, currently at TRLs 2-3, still remains in early research stages. To advance this technology beyond its experimental phase, key actions include the following points:

- **Error Correction:** Develop robust techniques to enhance DNA-based calculation reliability through improved sequencing and molecular encoding.
- **Scalability:** Address DNA computing scalability by developing strategies to control molecular interactions without causing degradation and enhance reaction specificity and efficiency.
- **Standardisation:** Establish standards and integrate DNA computing with existing frameworks, fostering collaboration between biologists, computer scientists, and engineers.
- **Applications:** Focus on high impact use cases like data storage or cryptography to demonstrate DNA computing's practicality and attract funding.
- **Gene-editing:** Explore CRISPR-Cas9 applications in bioinformatics to advance DNA computing technology.

We thus strongly recommend taking FPGAs and In Memory Computing to the next step for exploitation by HPC/AI applications. We also recommend that lower TRL areas (Neuromorphic and DNA computing) that are highlighted above should not be ignored whilst we pursue exploitation and use of higher TRL research by-products. These are key actions in this area.



Software and Use

System Software & Management

The following are the R&I recommendations in the different recognised subcategories of System Software and Management.

Addressing AI

AI users have different practices compared to traditional HPC users. AI workloads now pose new requirements on HPC environments. On the system software side, there is now a need for a ready-to-use AI toolbox designed for deployment on HPC infrastructures. This needs to be compatible, for example, with workload scheduling via Slurm, delegated SSO via Keycloak and containerized workloads. End-users can then potentially easily customise their environments and lower operational burden on administrators. There is also a need to support better productivity of data scientists by enabling effective sharing of data, models and code. Co-scheduling is uncommon in classic HPC environments, but the presence of coarse-grained power-hungry accelerators motivate increased attention to (i) supported units of allocation (e.g. processor socket associated with one or more GPUs, instead of an entire multi-socket node), and (ii) OS and runtime tuning to mitigate, or at least reduce, performance interference.

Software integration processes and portability guidelines

Globally, integration processes and portability guidelines need to be stringently adopted for HPC systems software. It requires hardware-software co-verification, prototyping, and CI/CD methodologies. Essential software infrastructure components (e.g. EasyBuild⁵⁰, Spack⁵¹, GitLab, Podman⁵²) are already in wide use and have matured significantly but putting together a robust and scalable CI/CD infrastructure remains an error-prone and maintenance-heavy commitment in the long-term. The EuroHPC JU has been funding CI/CD activities (e.g. with projects like DEEP-SEA, HiDALGO2, and Castiel2), to establish the foundations of CI/CD infrastructure; however, mid-/long-term investment will go a long way towards making such infrastructure sustainable over time, particularly in lieu of increased complexity in software stacks for HPC as well as HPDA and AI.

Facilities and Automation rules

It is critical to put in place facilities and automation rules, in the form of “software factories”, i.e. a tool-assisted regimented process for building software with verifiable provenance, to allow a modular but real integration of different software and hardware components to eventually work together and jointly form the European HPC ecosystem. Effective integration of increasingly complex software stacks on supercomputing environments builds on top of CI/CD infrastructure, but further entails a long-term development strategy for managing the life cycle of software stacks and presumes strong coordination among the various actors.

Integrated software stacks

It is recommended that concentrated effort be invested towards raising the TRL level of integrated software stacks, particularly in support of large codebases with a wide range of software component

⁵⁰ [EasyBuild: building software with ease](#)

⁵¹ [Spack - Spack](#)

⁵² [Podman](#)

3. Advanced Computing: R&I Recommendations ■

dependencies and a complex set of configuration settings. To do this, it will be necessary to provide communities of infrastructure beneficiaries with:

- Seamless access to continuous integration and deployment resources tailored to HPC, HPDA, and AI projects, at both small/medium and large scales;
- Turnkey software stacks targeting specific workloads, built continuously;
- Adaptations to the underlying heterogeneous hardware architectures;
- Performance regression testing and benchmarking;
- Standardised libraries and APIs to facilitate ecosystem composability.

Note that initiatives such as UXL⁵³, Kokkos⁵⁴, oneAPI⁵⁵ demonstrate that software development communities are trying to address such problems with regards to development environments. At the software stack level, initiatives such as HPSF from Linux Foundation could be an opportunity to federate these efforts. Today it is mainly driven by the US ECP community (cf. HPSF - High Performance Software Foundation, launched at the ISC'24 conference in May 2024, with a charter akin to that of the Linux Foundation). Taking the viewpoint of digital sovereignty, a European ecosystem could be nurtured along similar lines, but it would still benefit from establishing efficient tie-ins with existing initiatives.

HPC and Cloud/Big Data (merges)

Certain software factories initiate CI/CD processes where it could be useful to merge practices in between HPC and the Cloud/Big Data domains. Still, this merging remains a partial solution. Investment must be continued to enhance control of software integration dependencies and conflicts.

Live Orchestration

New complex application workflows need live orchestration, which is in conflict with current resource management practices on supercomputers. Concepts from initiatives towards the computing continuum (i.e. a computing and data processing environment where edge devices, heterogeneous nodes and both cloud and HPC resources are seamlessly integrated - cf. HiPEAC Vision 2024) are directly linked with these topics.

We strongly recommend further research and exploration of these System Software topics in light of its adaptations and changes needed for accommodating AI, most urgently.

Programming Environments

AI adaptations

There is a strong need to support the integration of AI techniques into HPC applications. This includes support for container orchestration, as well as efficient interoperability and resource sharing between HPC and AI stacks.

Tools to help move from 64-bit to lower/mixed precision and innovative formats, which are expected to be well supported on future platforms, by helping the developer to choose the best precision for

⁵³ [UXL Foundation: Unified Acceleration](#)

⁵⁴ [Kokkos · GitHub](#)

⁵⁵ [oneAPI Programming Model](#)

■ 3. Advanced Computing: R&I Recommendations

each part of the application should be researched and supported. This needs to be further supported by investments into numerical analysis of codes & code modification.

There is also a strong need to support the development of AI-based tools for all stages of HPC software development, including optimization, compiler code generation, identification of code sections suitable for optimization, runtime optimizations, performance analysis and correctness. Tools need to take into account HPC's focus on extreme parallelism, as well as heavy use of Fortran.

The process of building a consistent stack of software tools, layers and computing capacities dedicated to AI4HPC, leveraging existing efforts in Europe, should be immediately supported.

Adoption of high productivity languages/Co-existence

HPC applications must be continuously encouraged to adopt modern high productivity languages and models that facilitate high performance across a range of hardware and execution conditions. This requires standardisation, training and funding for code modernization, as well as continued development of more effective and understandable languages, language features and runtime support (for example, for fine-grained, dynamic tasks).

Programming environments, parallel languages and runtimes, to facilitate composability by exposing adequate semantics and interfaces to enable and manage co-existence with other languages/runtimes should be encouraged.

Data Placement and the Digital Continuum

APIs and runtime techniques for optimised data placement in heterogeneous memory systems, e.g. HBM and DRAM needs to be further researched. Machine learning techniques have already started to show very good promise in this direction.

Environments and tools for data-centric workflow composition across hybrid HPC-Cloud-Edge infrastructures should be further developed as not much has been done in this area. Unified data abstractions to enable interoperable data storage and processing across such a continuum should be carefully considered.

Malleability

Interfaces and programming models that simplify application support for malleability (dynamic use of resources) have been studied, for example, in the context of DEEP projects. This area needs to be further explored in light of new use cases.

Measurements and Metrics

There is a need to develop metrics for quantifying HPC coding productivity, which is an essential part of the HPC stack. Metrics exist for general purpose software development, but they are not suitable for HPC.

The programming environment must ensure that existing codebases, many developed over decades in languages like Fortran, are compatible with diverse hardware architectures, such as CPUs, GPUs, AI accelerators, and more - as well as having to adapt to different and heterogeneous operational conditions in the computing continuum. This must be done with increased urgency.

3. Advanced Computing: R&I Recommendations ■

I/O & Storage

Characterising I/O System Pressure

There is a need to conduct comprehensive I/O pattern studies to understand how HPC systems are impacted. Common I/O patterns to characterise the “pressure” on I/O systems should be investigated. Initiatives like the *I/O Trace Initiative*⁵⁶ from projects like IO-SEA and ADMIRE should be expanded to inform system optimizations.

There is now a very good opportunity to leverage AI and machine learning tools to monitor and analyse I/O and storage system data, providing real-time optimization, fault detection, and predictive maintenance in complex HPC environments. This problem of storage system analysis and monitoring has continued to be highlighted by us since the very early SRAs.

Object Storage and Distributed File Systems

Bridges between Object Stores and applications should be developed. There is a need to Invest in tools to seamlessly integrate Object Stores with HPC applications. This includes building file systems on top of Object Stores to overcome semantic gaps. In the context of H2020 and EuroHPC, Object Stores have only just started to be explored by projects such as SAGE and IO-SEA. There is also a need to align HPC Object Store usage with cloud technologies, ensuring interoperability of tools and protocols between environments.

Data Placement, Data Transfers and Energy Efficiency

Research is needed in smart data placement across storage tiers, from NVMe SSDs to tapes, to minimise data movement and reduce energy consumption. Development of advanced metadata cataloguing systems such as EUDAT, B2SAFE⁵⁷ etc, to track data and support automated staging between HPC systems and external repositories is also necessary. This will ensure efficient life cycle management.

The I/O middleware should continue to support scalable architectures, managing data across multiple storage layers (e.g., NVM, SSD, HDD). Improving cross-centre data transfer protocols, including peer-to-peer protocols like BitTorrent for distributed datasets, and enhancing user interfaces for managing large transfers is extremely essential considering the exploding volumes, for example, of scientific and instrument data. There is also a need to automate workflow-aware data management that optimises resource usage across storage tiers.

Focus on reducing energy costs associated with data movement, employing data aggregation techniques to optimise bandwidth usage and reducing power consumption is also necessary. This could also include compression and energy aware management of data transfers. This area has received very scarce treatment until now in EuroHPC and H2020 projects.

Fault Tolerance and Data Protection

Development of mechanisms to protect data from hardware failures, such as erasure coding and multiple version management to ensure data redundancy whilst having an eye on performance implications should continue to be looked at in light of novel system architectures and hardware. HPC

⁵⁶ Nafiseh Moti, André Brinkmann, Marc-André Vef, Philippe Deniel, Jesus Carretero, Philip Carns, Jean-Thomas Acquaviva, and Reza Salkhordeh. 2023. The I/O Trace Initiative: Building a Collaborative I/O Archive to Advance HPC. In Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis (SC-W '23). <https://doi.org/10.1145/3624062.3624192>

⁵⁷ [B2SAFE | EUDAT](#)

■ 3. Advanced Computing: R&I Recommendations

Storage systems should continue to be better protected against cyber threats like ransomware, integrating offline backups (e.g., tapes) and snapshot capabilities for rapid recovery.

Computational Storage

Exploring computational storage paradigms to process data near where it is stored, reducing data movement costs is even more critical now, having started to be addressed in SAGE and Sage2 H2020 projects. This includes optimising task-based programming, developing middleware and workflow integration, focusing on integration with AI and machine learning workloads for processing efficiency.

Security and Authentication Mechanisms

Integrating multi-factor authentication (e.g., Kerberos, OpenID Connect) into HPC systems, ensuring secure and user-friendly access control and a focus on developing efficient encryption mechanisms for both data in transit and at rest, ensuring metadata security and compliance with modern standards should continue to be addressed as architectures and workflows continue to change.

The key action here is to continue R&I in Computational Storage and Object Storage paradigms whilst looking into the continued exploitation of deeper storage hierarchies respecting the requirements of storage observability, fault tolerance, security and performance - continuing the early works of SAGE, Sage2, NextGenIO and IO-SEA projects into the era of HPC/AI.

Mathematics & Algorithms

Scalable Algorithms

Further research on communication avoiding and hiding algorithms, where new algorithms must be designed to avoid and hide communication beyond linear solvers, needs to be looked at. This entails exploiting multiple levels of parallelism - algorithms have to become more flexible in leveraging different levels of parallelism and support dynamic load balancing. For increasing the level of parallelism, Parallel-in-Time algorithms need to be integrated with spatial parallelisation into applications.

Approximate Computing and Correctness

In the area of low-precision arithmetic, there is a need to derive rigorous error bounds for mixed-precision algorithms and pursue efficient implementation of such algorithms. Lossy data compression techniques need to be looked at, expanding the range of applications where compression techniques can be applied and address side-effects like workloads becoming irregular and unpredictable. Further, establishment of tight error bounds for randomization techniques and better integration in numerical software libraries should be pursued.

Resilience and Correctness

There is a need to minimise or avoid check-pointing. Algorithms should be designed where its strategy for resilience does not rely heavily on periodically saving the program state, support scaling on HPC systems, and can cope with heterogeneous architectures. Development and implementation of communication-avoiding algorithms to enhance resiliency and energy efficiency for HPC should be pursued. Further, research on correctness challenges that are specific to HPC and its entire software stack as well as on tools that address these challenges should be looked into.

3. Advanced Computing: R&I Recommendations ■

Algorithms and Methods for Hybrid Mechanistic and Data-Driven Modelling

Methods and algorithms for hybrid modelling should be studied. A future is to develop mainstream algorithms and methods that support the implementation of hybrid modelling workflows. Another area in this context is the dynamical-systems-based deep learning which Improves the understanding of the properties of neural network architectures and training methods.

Mathematical Methods and Algorithms for HPC Technologies

Development of new scheduling algorithms and scheduling solutions that support increasingly important challenges including co-scheduling of complex workflows or resource allocation in the context of federated and compute continuum infrastructures is necessary. Also the development and implementation of efficient autotuning mechanisms for key numerical libraries for upcoming HPC solutions becomes very important in this context.

Quantum and Hybrid Algorithms

The three main areas for R&I needed in this subject are (a) Problem embedding, quantum-state preparation, error mitigation/correction (b) Algorithms on hybrid HPC/quantum architectures and (c) Simulation of quantum circuits.

Advancing and introducing new mathematical methods and algorithms is critical to ensure efficient use of current and future HPC architectures and technologies. The key action here is to continue to explore all the above methods in upcoming R&I programs as they become crucial for exploiting and exploring future Non-Von-Neuman computer architectures like quantum computers.

HPC Use Cases

We provide a brief summary of the R&I recommendations for HPC Use cases, even though the focus of this chapter is primarily on HPC technology.

AI for Science

There is a need to strengthen the role AI plays in scientific workflows (“AI for Science”/ “AI4HPC”). This entails building models from data and using these models as part or alongside simulations for better science. There is a lot that can be learnt from the DoE’s AI4Science townhall discussions⁵⁸ in the US.

Application Development and Maintenance

Sustained long-term efforts for maintenance, adaptation, and development of scientific applications is very much needed. There also needs to be better adoption of software engineering best practices for HPC application development. Adoption of performance portability frameworks, and high productivity programming approaches also needs to be critically looked at.

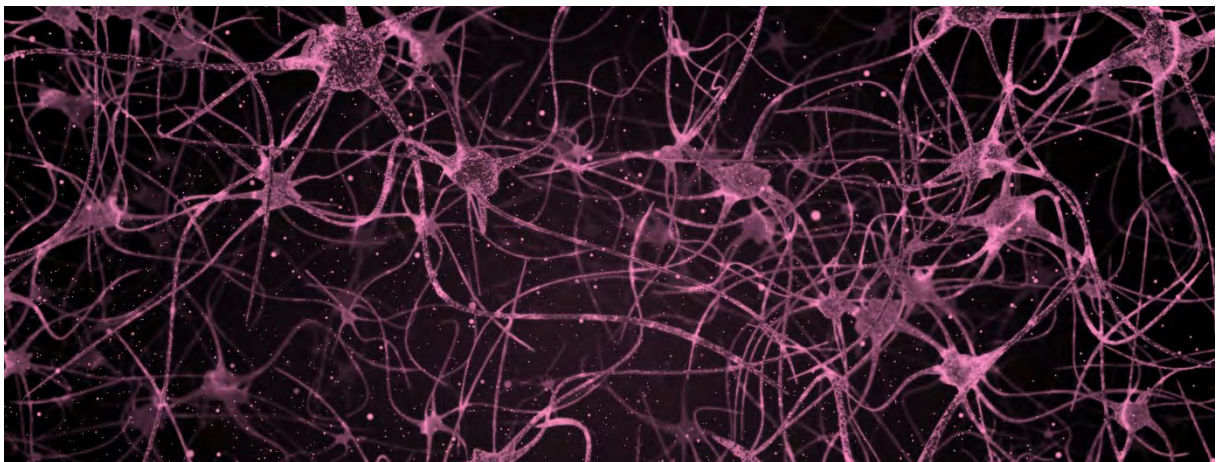
Lower precision

Development of novel methods and algorithms that can work with lower precision, which are able to exploit modern hardware better is needed, considering the trend towards lower precision arithmetic hardware developments to support AI.

Disruptive Technologies

Further exploration of which use cases could profit from disruptive technologies (quantum, neuromorphic) is needed as the ecosystem for them slowly starts to develop.

The key action here for policy makers is to help kick start initiatives and efforts that enable better liaisons between R&I projects and the new CoEs⁵⁹ that are looking into application innovations to exploit Exascale and Post-Exascale.



⁵⁸ [158802_0.pdf](#)

⁵⁹ [Kick-off of 10 Centres of Excellence in HPC to support the transition towards exascale - EuroHPC JU](#)

Ecosystem technologies

Integrating HPC in the Continuum

HPC will be one element in a continuum of computing, in charge of simulating very large tasks (like a digital twin of the Earth, or of industrial plants and processes) but fed with data coming from the real world. This will have a lot of implications for HPC technologies: e.g. computing loads with hard or soft real-time requirements for obtaining results or streaming processing “on the fly”, communication of data in and out with all the security related aspects, authentication of the systems connected to the HPC engine, etc. (see the section on Destination Earth requirements in the Backing White paper on this topic). Real effective work on the integration of HPC in a continuum of computing needs to be promoted to evaluate in practice the changes required, looking at specific examples such as the Destination Earth project.

Addressing AI in the continuum

New types of workloads, more data intensive (like AI) that will be linked with more “classical” HPC loads will be introduced in the context of the digital continuum. That could have an impact on the hardware architecture as the requirements are not the same as what existed until now:

- AI needs a lot of fast memory,
- “Aggregation” of resources or coupling nodes into virtual “supernodes” - similar to what Nvidia is doing by virtualizing several GPUs together to form a super-GPU with few PB of memory accessible from the nodes⁶⁰ (kind of “remote memory pooling”) with all the implications in term of number of NICs, latency, coherency and bandwidth, as has been discussed also in the Research Domain on System Hardware Components.
- Simulation needs FP64 (or more), while AI could work with far lower precision, down to FP4 for example,
- The move from “data storage” to “data lakes”.

A key action for policy makers is to help break silos and enable better interdisciplinary R&D in the IT-space at large. This is absolutely necessary. A necessary, but not sufficient, step forward could be taken by a series of new CSAs discussed for 2025 work programmes.

⁶⁰ See the section “System Hardware Components” of this SRA

Thematic Trend Recommendations

AI & Foundational Models

We dedicate a separate chapter on AI Explosion in the context of HPC in this SRA6 Master that covers detailed recommendations here.

Energy Efficiency & Sustainability

The following are a summary of the R&I recommendations for the Energy Efficiency and Sustainability Thematic trend.

Looking at the Big Picture

There is a need to look at the big picture when it comes to this topic. Energy efficiency needs to be looked at not just from the perspective of hardware elements (which is addressed in some of the other Research Domains) but also the software/ middleware. There are already some good steps in this direction in the EU⁶¹. Sustainability however is extremely hard to quantify as the parameters are innumerable (data centre constituents, the hardware, along with their associated carbon footprints throughout their life cycle - manufacture, use, and end-of-life). There are lessons to be learnt from some of the other actors in the European IT ecosystem in this area⁶². Energy Efficiency and Sustainability should be looked at holistically and programs continuing to address that needs to be supported, including the development of sustainability modelling tools.

Cooling

Improvements on efficient DC cooling infrastructure are needed (warm water cooling, heat re-use, elimination of water evaporation, etc). This will have a big impact in the HPC data centres.

Monitoring

Power monitoring systems should be harmonised, a standard metric must be defined that reliably reflects the complete “real” power consumption that includes CPU, accelerators, memory, network, storage etc. Continuous monitoring of the efficient usage of the system resources should be established and reasonable power management capabilities should be made available.

⁶¹ [Energy Efficient Technologies in HPC - EuroHPC JU](#)

⁶² [Sustainable computer systems | HiPEAC Vision](#)

3. Advanced Computing: R&I Recommendations ■

Exploiting Digital Twins

An HPC system digital twin could allow for more realistic prediction of power consumption, which should also be aided by reliable ways to characterise applications. In the context of HPC and sustainability, at the very minimum there needs to be reliable predictions of the energy consumption of the HPC resource as a whole (including its share on the surrounding data centre infrastructure) in comparison to the achieved performance. The desired type of HPC digital twin would be one which adds up the energy usage of the different components and reproduces a result which compares well to actually measured typical consumption values over time. The digital twin can also give a better idea of the efficiency of the HPC system beyond pure energy usage.

We very much welcome the new call on Energy Efficiency which tackles this problem from a software standpoint, referenced earlier. Key actions here include helping to establish a path towards holistic energy/CO₂ accounting, developing In-depth analysis of all energy/sustainability related effects in the overall life cycle of HPC/AI systems and development of the aforementioned digital twin for HPC/AI systems (in combination with prediction of application behaviour).



Quantum Computing & HPC

Applications & Benchmarking

There is a need to make sure that applications are considered in hardware-software co-design as well as any testing of middleware approaches in order to check that the requirements from the applications are taken into account. Further, application-oriented benchmarking is important to be able to evaluate the quality of quantum systems, both for internal comparisons and for comparison with classical algorithms.

Middleware & Software

There is a need to facilitate a thriving open-source software stack to make sure that European hardware approaches and applications with quantum advantage are supported. Authentication, accounting and scheduling mechanisms that are well-proven in HPC environments need to be adapted and integrated in the HPC-QC stack. For the design of the middleware, a holistic approach needs to be taken into account - considering the application and the hardware requirements as well as the interfaces between system software, applications, and user related software.

Emulators

Emulators, supported by HPC infrastructures, are essential for validating and refining quantum algorithms and technologies for both today's noisy quantum computers and future advancements, as they ensure accurate simulations of the state vector evolution independent of quantum computer capabilities.

European Quantum Flagship SRIA recommendations

In addition to the more specific technical recommendations, there is a need to make sure that the more general recommendations of the European Quantum Flagship SRIA are supported.

In the area of HPC-QC middleware, there is a need to start to integrate and/or extend conventional HPC software stack components; support the development of HPC-QCS integration technology (connectivity, middleware, and libraries); initiate developments to create open-source middleware, schedulers, compilers, and other open-source software tools; where necessary think about software standards or contributing to them; and; support the development of software components, tools, runtimes and environments.

Further, development of scientific software applications should be supported. Finally, there should be better engagement between the HPC and QCS communities - developing training mechanisms, contributing to the definition and emergence of global standards.

In general, we recommend that care should be taken to adopt a holistic approach in the early stages of quantum computer hardware and software stack development, so that requirements of all layers of the stack are taken into account. Some parts of HPC-QCS integration can take advantage of previous developments in the HPC domain and start therefore from a higher TRL level than other parts.

4. European HPC and AI Explosion



We summarise in this chapter the R&I trends and challenges for European HPC as AI based tools, methods and techniques have proliferated in the last few years with the advent and usage of GPGPUs and the continued development of the AI software stacks with exploitation of increased volumes of data. More details corresponding to each of these R&I trends and challenges can be obtained in the individual White Papers for the Research Domains and the Thematic Trends. These need to be recognised and addressed especially from a policy perspective.

General Considerations

Data

There is a strong need to share large, open scientific datasets for training and testing. Open datasets are data that are freely available for use and have open licences associated with them for access, modification and sharing. The FASST.⁶³ program in the US is taking measures for data sharing across HPC research centres, academic and industrial partners. The DoE has a vast repository of scientific data and this data will be made available to partners across government, industry, and the scientific community to train, test, and validate the next generation of scientific AI models. In Europe there are initiatives such as data.europa.eu.⁶⁴ portal which provides access to data sets from multiple EU & International portals. However, interoperability data will become increasingly important across various scientific domains that exploit HPC. It is reasonable to expect that the data produced by scientists and researchers will only grow exponentially in the coming years. There will be a strong need for standardised processes for collating and organising all this data for AI models running in HPC infrastructure including within the upcoming AI factories.

The federation of EuroHPC supercomputers and AI Factories will also generate enormous volumes of synthetic data that should be organised and shared. Trustworthy access needs to be setup through the data institutions in Europe such as data.europa.eu. Sharing bigger volumes of data sets across the federation of EuroHPC supercomputers may however continue to pose problems.

We recommend the development of a strong data ecosystem for the sharing of data across the EuroHPC federation of supercomputers needed for AI models across various scientific domains

AI for Science

There is a need to develop and train AI models suited for scientific applications. This could be similar to large language models such as ChatGPT but is trained on data from specific scientific domains and responding to queries corresponding to that scientific domain. Further the models from various scientific domains can be combined into a “meta” model that enables cross disciplinary research. In the US, Argonne National Laboratory (ANL) is creating a generative AI model called AuroraGPT which combines text, codes, specific scientific results, papers, into the model that science can use to speed up research. The Aurora supercomputer is being used for that which in late 2023 delivered around half an Exaflop. Similarly, there are also efforts in Japan that indicates that a ScienceGPT (FugakuGPT) is on the horizon. Such a ScienceGPT is needed in Europe now that the infrastructure elements a.k.a. EuroHPC Exascale Supercomputers and plans around AI Factories have started to come into play. This will have massive implications for European science and innovation. Plans for such needs to be put in place.

⁶³ [Frontiers in Artificial Intelligence for Science, Security and Technology \(FASST\) | Department of Energy](#)

⁶⁴ [About data.europa.eu | data.europa.eu](#)

4. European HPC and AI Explosion ■

We recommend the research and development of a “ScienceGPT” that can help cross disciplinary science and innovation in Europe.

There is a need to strengthen the role of AI in scientific workflows and classic simulations. AI based hybrid approaches and re-designing numerical codes in various fields requires knowledge of the physical models. There are huge opportunities for AI to enable most applications and use cases in areas such as biology, chemistry, material science, high energy physics etc. Thus, large scale simulations need to be refactored to allow for AI to enable better science and faster insights. Broad stroke visions are in place, for example, in the US in the area of AI for science which we reiterate below from [AI4Science-ASCAC.pdf \(osti.gov\)](#)⁶⁵, which outlines the 10 year vision in the area of AI for Science.

- Learned models begin to replace data
 - Queryable, portable, pluggable, chainable, secure
- Experimental discovery processes dramatically refactored
 - Models replace experiments, experiments improve models
- Many questions pursued semi-autonomously at scale
 - Searching for materials, molecules and pathways; new physics
- Simulation and AI approaches merge
 - Deep integration of ML; numerical simulation and UQ
- Theory becomes data for next-generation AI
 - AI begins to contribute to advancing theory
- AI becomes a common part of scientific laboratory activities
 - AI is integrated into science, engineering and operations

We recommend the need for developing a multi-year plan and roadmap specifically in the area of AI for Science in Europe including the possibility for multiple smaller projects

AI Services and Infrastructure

Following up from the plan for AI Factories, there is a need to develop, deploy and maintain a natively European ecosystem for AI. The EuroHPC federated HPC infrastructure services needs to provide integration with AI applications and Tools, which also include access to cloud infrastructures in Europe. This includes liaisons with efforts such as EOSC⁶⁶ (Scientific), GAIA-X⁶⁷ (Industrial) and hyperscalers with operations in Europe.

We recommend innovation in the area of software and middleware methodologies enabling seamless integration of our upcoming Federated HPC/AI infrastructure, with existing AI application and tooling ecosystem, and Cloud infrastructures.

The AI Explosion is indeed a massively energy intensive phenomenon, with vast amounts of power needed for our upcoming AI Factories. Training Foundational Models on such an infrastructure is bound to be very energy intensive going by the lessons learnt from Large Language Models such as ChatGPT. GPT-4 took nearly 50GW-hours⁶⁸ and was almost 50 times as intensive as GPT-3. It is conceivable that training multi-modal scientific foundational models on EuroHPC infrastructure & AI Factories in the coming years will consume even more power. Various aspects need to be thought through including

⁶⁵ <https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/202004/AI4Science-ASCAC.pdf>

⁶⁶ [European Open Science Cloud \(EOSC\) - European Commission](#)

⁶⁷ [Home - Gaia-X: A Federated Secure Data Infrastructure](#)

⁶⁸ [Data centres improved greatly in energy efficiency as they grew massively larger \(economist.com\)](#)

the possible advanced modelling of data centres to maximise power usage (incl. the position of racks, servers, etc). Advanced cooling designs to deal with the heat generated by these AI chips also needs to be addressed and supported. There is also a need to critically look at the sustainability and environmental impacts of providing AI services and infrastructure.

We recommend innovations in addressing increased energy requirements (& studying carefully the environmental and sustainability impacts) due to foundational models' running on federated EuroHPC/AI infrastructure.

It is also to be noted that appropriate isolation of AI data, model weights etc from federated HPC infrastructure operator needs to be provided. There is also a need to protect the HPC users' highly sensitive data as they are subjected to AI workflows.

Ecosystem Development

Support for SMEs in the HPC space in their use of AI methods, tools and techniques is critical for Europe. Use of AI is currently skewed towards larger companies (51% vs. 31% of SMEs), a trend that needs to be addressed if SMEs are to continue to be a driving force for prosperity across Europe. It is important to make it easy for SME HPC/AI users to use EuroHPC infrastructures. Programs such as FortissimoPlus⁶⁹ are good vehicles to enable that. Better incentives for SMEs to participate in upcoming AI calls (EuroHPC, Horizon Europe) should be provided including through better promotion and dissemination of these programs into these organisations. Further, the regulatory regime must support AI innovation and experimentation in SMEs as legal and compliance requirements are holding SMEs back including those in the HPC area to adopt AI. Finally, Europe is generally lagging behind in the access of private capital for upcoming technology startups, including in the AI+HPC area. Well defined programs to help SMEs have access to private/VC capital for growth to compliment the seeding through public funding is needed.

We recommend targeted dissemination of EuroHPC AI initiatives into the AI+HPC SME community with clearly laid out benefits.

Education and Training in the area of AI within the HPC ecosystem is indeed part of the wider digital skills gap facing Europe. In that, 61% of businesses in the EU said not being able to find staff with the right digital skills is slowing down their business performance, and over a quarter say it is preventing them from adopting AI. But over a third (37%) of employees say they do not have enough time to learn new skills, and nearly half (45%) say the cost of training programmes is prohibitive⁷⁰. There is an absolute strong need to grow talent and develop the right skills in AI within the HPC community in particular considering the differing foundations for HPC and AI - which have now started to come together.

Projects successful in answering to the call on developing Support Centres in Europe for HPC powered AI applications⁷¹ - which we believe is a very timely initiative - should begin to start addressing some of these issues.

There are some lessons and insights also to be gleaned from the EUMasters4HPC program. The EUMasters4HPC is tasked with developing a Masters' program HPC curriculum in Europe in close collaboration with universities, research centres and the industry. This is now leading to the first pan-

⁶⁹ [Fortissimo Plus \(FFplus\) - EuroHPC JU \(europa.eu\)](https://europa.eu/europa/en/fortissimo-plus)

⁷⁰ <https://aws.amazon.com/blogs/training-and-certification/new-european-study-ai-skills-will-significantly-boost-productivity-and-salaries/>
<https://assets.aboutamazon.com/bb/2e/9077b9f44a2898c01fcc7f35440d/aws-ai-europe.pdf>

⁷¹ [Support Centre for HPC-powered Artificial Intelligence \(AI\) Applications - EuroHPC JU \(europa.eu\)](https://europa.eu/europa/en/support-centre-for-hpc-powered-artificial-intelligence-ai-applications)

4. European HPC and AI Explosion ■

European HPC Masters' program.⁷² The EUMasters4HPC program have also led to internships and opportunities for students.

We recommend the creation of a program similar to EUMasters4HPC, such as "EUMaster4AI" with focus not just on AI but also usage and exploitation of HPC infrastructure for AI & utilising AI for Science.

Top Technical considerations

Storage and Data Management

The following are a summary of the technical recommendations in the area of AI driven HPC, obtained from the various Research Domains and Thematic Trends. More details can be found in those respective white papers.

There had been an emphasis on Storage and I/O in the Horizon 2020 program with the funding of many storage- and I/O-related projects (SAGE, NextGENIO, etc) and also in the early phase of the EuroHPC program (IO-SEA, ADMIRE, etc). However, that has been de-emphasised. The pathway from SAGE, Sage2, NextGENIO and IO-SEA is now unclear though there have been very interesting developments that came out of these projects in the area of Object Storage, Hierarchical Storage Management and Non-Volatile Memories. Even though Intel focused efforts on 3DXPoint the support of which is now pulled out, such technologies are highly applicable for any multi-tiered storage systems independent of technology type. AI/ML has put a need for renewed emphasis on data and storage, and I/O and data management will play a very critical part in the HPC/AI ecosystem.

We recommend re-invigorated efforts in the area of data storage, data management and I/O innovations in light of AI developments in the upcoming work programmes.

Lower precision

The new generation of AI hardware from the leading vendors have forsaken double precision floating point operations, which had been the mainstay of simulation science in the last decades. AI training might drive the direction of hardware evolution for the next several years. This means that the algorithms need to evolve and select parts of the HPC codes will have to move to lower precision to maintain their performance. The processor industry now has a much stronger motivation to optimise products for AI, with increasing silicon area dedicated to reduced precision arithmetic. Further following this AI trend, the hyperscalers are also developing their own processors to suit the lower precision needs of AI workflows. What should the European HPC community do? Should we develop more AI specific specialised hardware accelerators? What should our longer-term strategy be? The answers to these questions are not very clear.

We recommend developing a strategy for better tracking and adapting to the trend of lower precision arithmetic both from the perspective of Applications and Algorithms and European hardware developments.

⁷² [Home - EUMaster4HPC \(uni.lu\)](#)

Opportunities for code generation/Optimization

LLMs are continuing to prove very effective in generating code (e.g.: copilot). More than 40% of the code on GitHub is claimed to be AI generated.^{73,74} ML generated code generation/optimization could now be introduced in mainstream compilers. There is now a possibility to identify sections of code that can be optimised. LLMs could identify the purpose of a code section and possibly suggest a better/faster/more secure open-source implementation. However, there is a clear open challenge in better understanding how LLMs can generate better parallel code in the context of HPC. Combining LLMs with structured reinforcement learning approaches could lead to interesting results for HPC. In the area of DevAI there is a need to smartly combine LLMs and traditional developer tools like compilers and static code analysers.

We recommend more R&I in exploiting AI assisted code generation methodologies suitable in the context of HPC applications and systems software

Lessons from the Cloud dealing with AI

HPC can start to adopt and adapt some of the tools, methods and techniques developed in the cloud ecosystem - for AI. It is important to be able to reconfigure HPC data centre resources, building new virtual infrastructures out of existing building blocks dynamically for AI workflows. There is also a need to combine federated HPC infrastructures with the Cloud to provide increased flexibility to schedule HPC/AI tasks in large workflows - which also needs to take into consideration location of data sources. For example, can Slurm and Kubermettes work together? Also, as a lesson from the Cloud, a well-defined and ready to use “AI tool box” is needed for federated HPC environments - which considers, data security, scheduling, identify and access management and well thought out considerations for isolation/containerizations. Also many of the tools for understanding abnormal and malicious behaviour available in the Cloud can be adapted and reused for AI workflows on HPC infrastructure.

We recommend the reuse and adaptation of tools developed for the Cloud for AI, for federated HPC infrastructures.

Other considerations

Exascale and Post Exascale systems are going to be extremely complex infrastructures where failures and abnormal behaviours will be a norm. Detecting anomalies due to malfunctions or due to malicious behaviours on such large-scale systems will be extremely difficult to accomplish using the tools and methods that the HPC community has at its disposal today. AI presents an enormous opportunity here by drastically cutting down the time and energy needed by quickly getting to the root cause of any such abnormal or malicious behaviours, by leveraging, for example, Large Language Models for massive on-line log analysis. AI models can help to mitigate and protect against malicious system behaviours that could also be caused by rogue actors leveraging AI with malicious intent. AI methods can also help to continually optimise energy efficiency in such large-scale systems - with energy consumption becoming such a critical facet for Exascale and Post Exascale.

⁷³ [What is AI code generation? · GitHub](#)

⁷⁴ [Stability AI CEO: There Will Be No \(Human\) Programmers in Five Years - Decrypt](#)

5. Post Exascale Vision & Challenges



This section summarises the Post Exascale vision and related challenges in the different Research Domains and Thematic Trends and highlights the general considerations of AI and Use Cases for Post Exascale.

Post Exascale can indeed be very broad, and the scope is very wide. It is hard to predict how Post Exascale will look like in the long run and how it might evolve. However, we can look at what exists today and extrapolate our vision for Post Exascale in the near term. Please note that this is by no means an exhaustive and accurate list of challenges (which is impossible to assess), but general indicators and possible paths. We should hope that such an incremental and near-term assessment of Post Exascale is more realistic and useful to have.

Research Domains and Thematic Trends' considerations

We envision a Post Exascale where AI moves from a proof-of-concept to full scientific validation and large-scale implementation.

The Exascale era will introduce new HPC uses, posing methodological and software engineering challenges such as AI model learning, physics-based AI, hybrid simulations, data reduction, distributed AI, and intelligent in-stream/in-situ data analysis. This ecosystem, rich in data, computing power, and expertise, offers opportunities to leverage AI for complex scientific questions. Recent AI advances highlight the need for these elements to develop powerful models.

Key challenges in this context:

- Developing a European HPC and AI software stack with highly interoperable software components and tools.
- AI-centric co-development of these software components with use-case sharing.
- Investigating AI-assisted tools for scientific code generation, robustness, and efficiency.

We envision a Post Exascale where the infrastructure ecosystem consists of (ideally) seamless federation of HPC centres interfacing with multiple Cloud services.

The Digital Continuum is currently dominated by large Cloud providers, with HPC centres striving to offer alternative solutions for managing and processing large datasets. A major problem for HPC centres is proving their value compared to Cloud providers, particularly for real-time data processing. Additionally, the multi-tenant nature of the Digital Continuum, where data is used for various purposes and infrastructure is shared, necessitates building trust among entities and prioritising cybersecurity.

Key challenges in this context include:

- Develop software supporting a Post Exascale vision built on a Cloud/federation of Exascale systems
- Developing environments and tools supporting data-centric workflow composition across hybrid HPC-Cloud-Edge infrastructures
- Developing programming environments ensuring sustainable interoperability across the HPC-Cloud-Edge continuum (requiring middleware able to interoperate and execute in hybrid scenarios)
- Develop the concept of EaaS (Exascale as a Service), for Tier-0 European systems

5. Post Exascale Vision & Challenges ■

We envision a Post Exascale where the Cloud Service providers offer increasing capabilities and support for HPC.

Cloud services already deliver Exaflops of compute power for private customers and also researchers. The classical division between High Performance Computing (running large-scale, tightly coupled applications), and High Throughput Computing (many small applications executed independently from each other) in data centres and the cloud, is becoming less clear. Cloud service providers are deploying very large-scale computing centres that not only provide compute capacity, but also include high speed networks and HPC software layers. Applications in Artificial Intelligence (AI), e.g. Large Language Models, are driving this trend. The question to whether a user seeks compute time on an HPC centre or at a Cloud service boils frequently just to the question of whether they can rather afford to pay for compute time on the Cloud, or wait for the next compute time application cycle on an HPC centre.

A Key challenge in this context:

The HPC community needs to strongly consider and study reducing the entry barrier for HPC, both in terms of time and required expertise.

We envision a Post Exascale where AI drives hardware evolution and its diversity.

Also, the technology landscape is strongly influenced by the economic drive of AI. The processor industry has a much stronger motivation to optimise a product for AI, than it has for HPC. A clear consequence already visible now is the increasing silicon area dedicated to reduced-precision arithmetic, with respect to the double precision that HPC users are so keen on. Application developers in HPC should seriously rethink their approaches and manage as much as possible the algorithms requiring double precision arithmetic, since this will automatically increase compute performance and energy efficiency. Another interesting development is the fact that hyperscalers are developing their own processors for their cloud services but are not selling them as products. This means, a number of recent processor technologies cannot be integrated in HPC centres.



A key challenge in this context is the assessment of the path forward and options for the HPC community: to collaborate and adapt as to benefit from the technology advances done by hyperscalers, and/or to develop its own technology tailored to HPC.

We envision a Post Exascale where Quantum technologies mesh with AI.

Quantum Computing + AI are highly synergistic, as Quantum Computing has the potential to significantly enhance AI's capabilities⁷⁵. Quantum + AI combines the power of Quantum Computing to create new algorithms, machine learning techniques, search procedures and data processing techniques that are impossible to achieve with classical computers.

A key challenge in this context is that scientific computing will need to look into exploiting this development as Quantum + Classical HPC mesh into hybrid systems.

⁷⁵ More details on Quantum + HPC can be found in corresponding Research Domain White Paper

We envision a Post Exascale where there could be opportunities for the evolution of DNA and Neuromorphic non-conventional computing.

DNA computing is still in the early stages of Research and Development, according to its Technology Readiness Level (TRL), which is now between 2-3. In order to move DNA computing from an experimental and conceptual stage to a more developed technology, it is important to emphasise certain critical actions.

Key challenges in this context are:

- Improved Error Correction Techniques: Improving the reliability of DNA-based calculations requires the development of strong error correction techniques.
- Scalability Solutions: It's critical to address the scalability issues that come with DNA computing.
- Standardisation and Integration: DNA computing will be easier to accept and use if a standard is established and it is integrated with already-existing computational frameworks.
- Application Development: The practicality of DNA computing will be illustrated by concentrating on certain, high-impact applications, including data storage or cryptography.
- Gene-editing Technologies: Given the enormous potential for technological advancements in bioinformatics, it is recommended to investigate and test the CRISPR-Cas9 technique and its applications in this field.

There have been some interesting new developments in Neuromorphic computing from the vendors' side.⁷⁶ including those in Europe.⁷⁷

Key challenges in this context include:

- Promoting the development and maintenance of multi-platform benchmarks that can be used to make fair comparisons of available Neuromorphic solutions.
- Stimulating the development of unified and standardised programming models and software frameworks to enhance interoperability and facilitate the porting of SNN models across different neuromorphic platforms.
- There is currently limited access to neuromorphic HW, which is primarily available for a limited number of research teams. To overcome this issue, we have identified three strategies:
 - Promote the development of open source HW architectures deployable on FPGAs.
 - Support the development of Cloud-based prototyping platforms hosting the major available neuromorphic architectures accessible via remote by developers.
 - Engage leading players in the silicon industry in the development of Neuromorphic SoC architectures.

We envision a Post Exascale where there are Hyperconverged Infrastructures in HPC centres with seamless integration of computing, storage and networking technologies.

As we move beyond Exascale computing, the vision for hyper converged infrastructures (HCI) in HPC depends on unprecedented integration, efficiency, and flexibility. Post Exascale HCI will seamlessly combine compute, storage, and networking resources into a unified, software-defined environment. This integration will streamline management, reduce latency, and optimise resource utilisation, crucial for handling the massive data and complex workloads typical of Post Exascale systems.

⁷⁶ [Intel unveils largest-ever AI 'neuromorphic computer' that mimics the human brain | Live Science](#)

⁷⁷ [SpiNNcloud Systems Launches SpiNNaker2 to Advance Neuromorphic Computing for AI Systems \(hpcwire.com\)](#)

5. Post Exascale Vision & Challenges ■

Future HCI in HPC will leverage advancements in Artificial Intelligence and Machine Learning to enhance automation and predictive maintenance. AI-driven resource management will dynamically allocate resources based on real-time workload demands, ensuring optimal performance and energy efficiency. Enhanced fault tolerance and self-healing capabilities will minimise downtime and maintain system reliability.

The Post Exascale era will also see increased adoption of heterogeneous architectures within HCI, incorporating CPUs, GPUs, FPGAs, and specialised accelerators. This diversity will support a wide range of applications, from scientific simulations to AI research. Standardised, open-source middleware will play a pivotal role in achieving interoperability across these varied components, simplifying integration and scaling.

Ultimately, Post Exascale HCIs will empower researchers and organisations to tackle even more complex scientific and engineering challenges, driving innovation and discovery in the next frontier of HPC.

We envision a Post Exascale where monitoring of large complex systems will become a very critical issue.

Systems involved in data centres are becoming larger and more complex as we move towards Terascale and Petascale eras and this issue will be exacerbated at Exascale and beyond. The need for monitoring and analysis is more critical than at any time before because of the large number of subsystems involved. The number of probes needed and metrics to be gathered start to become unimaginably large. Including applications and simulation codes metrics to be considered alongside the supercomputing infrastructure makes the telemetry gathering and analysis problem ever harder.



The path forward at this time would be to gather the whole monitoring data in an unstructured/semi-structured database, potentially based on the NOSQL type paradigms, and using the AI tools that are now available at our disposal. The ML/DL frameworks that the AI tools bring forth will be very useful in identifying correlations between the various components, leading to optimizations and early identifications of hardware issues or misconfigurations.

Key challenges here are the management of these enormous amounts of telemetry data and picking the right AI tools to come up with meaningful understandings and correlations.

We envision a Post Exascale where power consumption will continue to pose very major challenges.

We have already discussed in the previous chapter the power consumption implications of new AI driven hardware. Whilst efforts continue to be pursued in optimising power consumption in upcoming heterogeneous hardware and associated systems, the power consumption will continue to grow to new levels never seen before. AI queries are an order of magnitude more power hungry than internet searches and power demand in data centres is predicted to grow by 160% by 2030⁷⁸ globally. This has deep ramifications for European AI/HPC infrastructures that continues to raise debates on cost and sustainability.

⁷⁸ [AI is poised to drive 160% increase in data center power demand | Goldman Sachs](#)

Apart from the compute subsystems, the data infrastructures (including storage and I/O) will become increasingly power hungry. There are hence further opportunities for smarter data placement and continued usage of older generation technologies such as Tape. Modular disaggregated architectures present opportunities for better managing and optimising power. Efforts are already underway in this direction in Europe.⁷⁹

Key challenges here are better approaches to manage power consumption in light of increasing AI applications and associated infrastructure, and the associated management and movement of data.

We envision a Post Exascale where data movements across and between infrastructures will continue to pose major problems.

Large amounts of data remain co-located to data generation instances or are preserved in long-term archives. While significant network capacity has been added globally, both the mismatch between available bandwidth and the to-be-moved data volumes make it increasingly unfeasible to download entire datasets before accessing it for analysis. It is interesting to note that the energy efficiency of moving data information across wires is not reducing whilst the energy efficiency of transistors continue to improve which means that at the micro and macro scale, data movement will be a major problem.⁸⁰ in relation to computation, as long we use conventional wires and cabling for data transfers. If the distance involved in data movements increases, this problem exacerbates. Making the most of shared/federated infrastructures and reducing financial and environmental/energy per use requires research into workflow and data locality aware resource orchestration.

This means that the whole datacentre and computer centre must have been designed to reduce the distance between the different actors. The data infrastructure and the applications themselves must have been optimised as well, to reduce the data movements. As data is transmitted through networks, they put a heavy pressure on them, making networks and wiring/cabling/topologies the next barrier/obstacle Post Exascale. This is a **key challenge** in this context to be addressed.

We envision a Post Exascale for HPC in the Digital Continuum with HPC working with other ecosystem technologies generating highly integrated solutions.

We defined the digital continuum as the Ecosystem technologies around HPC such as IoT, Edge, Cloud, Cybersecurity, etc. The focus on building such large complex integrated solutions will shift towards establishing sustainability and anticipating new technological advancements and evolving user needs.

The **key challenges** to consider here are:

- Use the latest developments in Artificial Intelligence optimally in such hybrid solutions.
- Fostering the development of next-generation federated systems that not only connect but also intelligently manage resources across a wide range of computational environments. This will include implementing fully automated (AI-driven) service deployment, orchestration, and scheduling systems that take into account data location and transfer and dynamically adapt to changing computational loads and priorities across global networks.
- Establishing robust, decentralised identity and access management systems (for example using distributed ledger technologies) to enhance security and privacy while ensuring seamless user experiences across multiple platforms.

⁷⁹ [Energy Efficient Technologies in HPC - EuroHPC JU \(europa.eu\)](https://www.europa.eu)

⁸⁰ [The future of computing beyond Moore's Law | Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences \(royalsocietypublishing.org\)](https://royalsocietypublishing.org)

5. Post Exascale Vision & Challenges ■

- Promoting the adoption of next-generation communication technologies, such as 6G and beyond, which will facilitate ultra-reliable low latency communications (URLLC) and massive machine-type communications.
- Developing and integrating quantum-resistant cryptographic solutions to secure streaming data against future threats posed by quantum computing.
- Advancing the development of AI-enhanced security operations centres that can pre-emptively respond to cyber threats before they impact system operations.



AI & Foundational Models' considerations for Post Exascale

The transition to Post Exascale supercomputing marks a significant milestone in the evolution of AI and foundational models, offering unparalleled computational power and efficiency. We summarise below the key considerations at Post Exascale:

Scalability and Efficiency

As we move into the Post Exascale era, the ability to train AI models with billions or even trillions of parameters will become increasingly feasible. Training these vast models necessitates the use of capability clusters, specifically designed to handle the immense computational loads. Unlike traditional numerical simulations, AI training does not require high floating-point precision, continuing the shift to lower precision in advanced micro-architectures optimising the silicon area and enhancing the efficiency of AI training processes.

To fully leverage the capabilities of Post Exascale supercomputers, it is crucial to integrate advanced parallelism techniques, including data, model, operation, pipeline, and sequence parallelism. Memory-efficient strategies like checkpointing, offloading, and mixed-precision arithmetic will further enhance the scalability and efficiency of AI training on these next-generation systems.

Data Utilisation and Synthetic Data

The availability and quality of training data remain pivotal for developing high-performing AI models. In the coming years, traditional data sources will be thoroughly exploited, necessitating the generation of new and synthetic datasets. Post Exascale supercomputers will play a critical role in creating these synthetic datasets through precise simulations, providing fresh and high-quality data essential for continuous AI model improvement. This approach will ensure the sustained growth and advancement of AI technologies despite the eventual exhaustion of existing data sources.

Inference: Capacity Over Capability

Inference in AI, particularly for large models, is more about capacity than capability. The transition from numerical kernels to neuronal approaches will require handling millions of parallel model calls efficiently. Each call, being independent, necessitates a system designed for high-throughput, low-latency processing. This requirement will likely lead to the development of specialised partitions, akin to the GPU clusters of the 2000s, optimised for the unique demands of AI inference.

Ethical Considerations and Explainability

With the increased power and complexity of AI models, ensuring ethical use and explainability is more important than ever. Post Exascale systems must include robust mechanisms for monitoring and mitigating biases, and tools for explaining AI decision-making processes. These measures will help build trust in AI technologies, ensuring they are used responsibly and ethically.

6. Industrial HPC Usage in Europe



EMTg-3310-007

Characteristics:
Rated Output 310
Rated Torque 1.337
Input 150-230
Current 2.3
Speed 4300
Weight 0.91

02AG

02AG

Input Torque

RPM

Date	Developer	No.	Agency
	John M.	EMTg-1337	GRDNLT

DR89

3/4 view

BA31

RT31

RT83

EM-1

RT81

EM-5

This chapter discusses the requirements and priorities of European Industrial users of HPC in light of the availability of EuroHPC computing infrastructure. EuroHPC machines will predominantly be used by scientific users. Bringing in industrial users into the same infrastructure is bound to introduce unforeseen challenges both for the users as well as for the infrastructure providers/EuroHPC. Having computing infrastructure that is capable and adaptable for both scientific and industrials will help to propel Europe's societal and technological advancement. We summarise below some of the perspectives from the industrials.

The ETP4HPC has put together an Industrial User Working Group consisting of some of the major industries that have been leveraging HPC resources for industrial innovation for the last many years/decades - that have provided inputs for this chapter. Many of the HPC systems used to be in-house and there has been a lack of availability of public infrastructures that cater to their specific needs. However, the new EuroHPC machines offer the possibility for incumbent industrial users to expand their computing capabilities, as well as for new industries to explore and exploit the power of Advanced Computing.

Introduction

Industrial HPC end-users rely on a large set of core applications, which perform simulations, optimisations, complex data analysis, and increasingly, AI/ML processing. Such core applications continue to cover a huge scope - such as simulation of internal and external flows (CFD), combustion and other chemical reactions, structural analysis (often using finite element codes), materials sciences, molecular dynamics, quantum effects, circuit simulations, financial risk analysis, docking and protein folding, *-omics, traffic and crowd simulations, weather, large-scale graph analytics, discrete optimisation and many others. While the algorithms used in these do not differ widely from scientific applications, differences are caused by the different scale and complexity of problems, the frequent need for ensemble scaling to drive optimisations, and the role of closed-source, licensed applications. They can also influence the choice of system technology & design.

The use of HPC for emerging Machine Learning (ML) and Artificial Intelligence (AI) applications in the various fields of engineering allow, among other things, to develop Digital Twins and model-based design methodologies. The combined power of new methods and new computing architectures facilitate innovation, increase the efficiency and safety in many critical engineering areas, ultimately contributing to a sustainable digital transition.

The following are some of the areas that needs attention.

Data Protection

Commercial users store and control significant amounts of data with high commercial value and highly sensitive data, and personal data in light of GDPR considerations. Such data commonly resides in the protected IT infrastructures of the customer, and it has to be transferred onto the IT infrastructure of a HPC provider, accessed and modified via HPC applications and workflows there, and transferred back (results) or purged (input data which is no longer useful). It is critical that technical measures to protect data from unauthorised access and modification are in place throughout the complete data processing chain. In addition, effort and costs for data movement must be kept under control. Data storage and appropriate data segregation should be thoroughly considered.

6. Industrial HPC Usage in Europe ■

Security considerations

Industrial companies operate their own IT infrastructure. The industrial work environment relies on high-level frameworks and workflows, which match the business or engineering processes in place, and provide end-users with easy-to-use interfaces. Such frameworks and the corresponding high-level workflows have to be supported by any HPC usage model. Authentication, authorisation and accounting support to non-personal accounts may be required, and strong monitoring and security mechanisms are required for any operations that touches confidential data. The EuroHPC machines (& the personnel operating these) must thus ensure that industry standards for Cyber Security and Data Security are met. This has to be explained and potentially audited from/to the industry users including mandatory company conducted risk assessments.



Financial considerations and pricing

From a contractual point of view, the financial engagements linked to the requested capacity should be clear from the beginning, for example, the rental costs by Core/CPU/Node hour, etc. Larger upfront investment is difficult to handle from industry side. Pricing options should reconcile with the use of public funds in the TCO of these systems.

Accounting and billing must comply with the business requirements of the end users and accommodate paying per use. Corporate rules such as payment terms, for example, might be non-negotiable. Pricing models should consider dedicated queues or pseudo-idle resources ready to be used by industry.

Legal considerations

Contracts tend to become very complicated due to existing corporate and legal obligations (SLAs, DPA, IP, liabilities, etc). Also, corporate rules such as payment terms may be non-negotiable. Forming legal consortia to access the machines might thus be very involved and complicated. Such legal issues need to be considered.

Service level agreements

Due to (very) expensive licences, queuing times and time-to-access need to be reduced to a minimum. To become (more) attractive, public infrastructure needs to outperform existing Cloud offerings. There should be a consistent ecosystem across all EuroHPC sites and “paperwork” should be kept minimum and simple. The big question is whether SLAs for scientific users can be adapted for meeting those of industry specific needs?

It is to be noted that there could be a need for a tailored architecture for different workflows. In the industry there are typically two kinds of user groups - “computing intensive group” that needs high memory bandwidth processing where the filesystem use is not very intensive (eg: CFD users) and high

throughout users that have small workloads and need fewer core and small wall clock times but they stress a lot, the filesystem (typically, users are computational mechanics, AI, and Big Data). These two different classes require different architectural setup, configuration and tuning that often are in opposition. The SLAs need to reconcile with these kinds of contradictions.

Experienced centres such as HLRS in Europe, catering to industrial users can inform and advice.

Interoperability with Cloud infrastructures

In order to enhance the capabilities of organisations to “follow the data”, it is important to establish a common technological framework including a HPC for Edge laboratory facility, and software platforms and services to enable products and services to evolve toward the Edge, and leverage Cloud technology.

As digital ecosystems expand, the ability of systems, devices, and applications to work together becomes critical. Edge, Fog, and Cloud computing can support interoperability by standardising protocols and interfaces, facilitating data exchange and process coordination, targeting both HPC and Cloud resources in a distributed federated approach, via user-friendly tools and interfaces.



Software management

It's essential to implement a robust and efficient management system to simplify the use of software licences on private Clouds and to work together with the software vendor to optimise licences usage in HPC environment. It should also include features for tracking and monitoring the usage of licences, which can help in optimising resource allocation and reducing costs.

6. Industrial HPC Usage in Europe ■

Additionally, integrating automated tools for deploying and configuring HPC applications can significantly simplify the process and reduce the time required to get an application up and running.

Potential for new pricing models

Finally, we explore the possibility of novel pricing models for European industrial users.

Manufacturers of tangible goods measure their efficiency and competitiveness against their competitors using metrics (key performance indicators) linked to the goods they produce. For example, a car manufacturer measures the efficiency of its factories according to three fundamental criteria: cost per car produced, number of cars produced per day and production time per car.

This method can also be applied to intangible assets. For example, a company that designs molecules, such as a pharmaceutical laboratory, would like to know in advance the calculation cost per molecule, the number of molecules simulated per 24 hours and the calculation time per molecule, in a predictable and fixed manner. A company that relies on inference queries wants to know in advance the cost of responding to a query, the number of queries processed per second and the latency to obtain the response. This enables them to build their business model on stable production and financial bases, giving them greater visibility over the evolution of their sales and profits.

All manufacturers and companies that use or will use "digital factories" for their business have switched all or part of their production to hyperscalers. Their operational (COO) and financial (CFO) managements are rapidly facing two critical problems: (1) the unpredictability of production costs (with the big hyperscalers, the number of parameters involved in the billing model is very large, they are linked to each other, and are difficult to understand and anticipate, making the billing model unreadable and unpredictable from one month to the next); (2) the monthly cost, very often much higher than before the switch of production to the Cloud.

Based on this analysis, for companies that switch their production to the Cloud or a HPC service model, it is necessary to develop an offering based on a commitment to results, also known as OaaS ("Outcome as a Service"), rather than a simple commitment to means, however sophisticated, offered by traditional hyperscalers. The OaaS model will become the preferred choice of financial and operational managers, as it provides a simple answer to both these problems, offering a predictable billing model, guaranteed per unit of production and aligned with business needs.

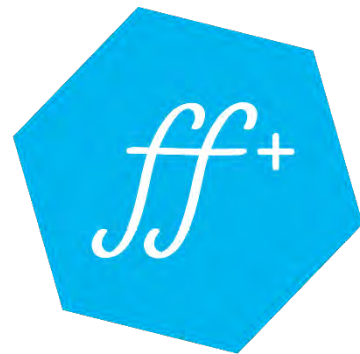
The OaaS approach requires a great deal of research, industrialization and production work to meet all needs across all verticals.

Furthermore, the OaaS approach encourages sobriety, as all stakeholders benefit from optimization. If the plant is optimised, production costs come down and production criteria (lead times, etc.) are improved. Thus, the integration of issues related to energy savings, eco-responsibility, carbon footprint calculation, water, ozone, etc as required by the European green taxonomy naturally gets addressed.

In support of European SME HPC Users

Europe is rightly encouraging programs that support the adoption of HPC by European Startups and SMEs. The Fortissimo projects⁸¹ and the FF4EuroHPC⁸² have successfully executed more than 130 experiments that have resulted in success stories from over 20 EU countries where new products and services were created exploiting HPC/AI. FortissimoPlus⁸³ is now launched as a continuation of the Fortissimo projects, where Open Calls will be published for the next few years (2024 - 2028) to acquire business experiments and innovation studies targeting SMEs. Proposals are expected to address business challenges faced by European SMEs.

We continue to encourage the support of this program helping to catalyse the European economy through Advanced Computing at this grassroots level, whilst we continue to gather further inputs on the issues faced by European SME industrial users in the era of AI.



FORTISSIMO
PLUS

⁸¹ [Fortissimo project](#)

⁸² [FF4EuroHPC High-performance computing technologies for SMEs](#)

⁸³ [Fortissimo Plus](#)

7. European HPC Technology SMEs

Source of innovation - opportunity for achieving leadership

SME



This chapter discusses the current role and the support needed for European HPC Technology Small and Medium (SME) businesses. We have now introduced this subject in SRA6. Even though this chapter is brief and provides a summary of the state of affairs and requirements of European HPC Technology SMEs, we will follow up with a more detailed SRA6 Agile White Paper on this, which is warranted considering the extremely critical role HPC technology SMEs will play in the European HPC value chain.

Introduction

European SMEs involved in the design, development and deployment of leading-edge technologies (in HPC, AI and QC) play a pivotal role in enhancing the competitiveness of the European industrial ecosystem across several critical sectors. By developing original solutions and infrastructures leveraging European staff, these HPC SMEs not only foster innovation in the end users, by providing solutions for the computational power, but enable the creation of a native European know-how ecosystem addressing the wide range of technical skills, optimised design and production processes, business organisation required to bring to the market the most advanced and cost-effective solutions.

Because of the innovativeness of its HPC SMEs, the EU has secured a strong international position in high-performance computing in the past years – a unique advantage to exploit in areas such as AI, and to stimulate further private investment. HPC SMEs serve a vital role both directly as stakeholders in the innovation processes (R&D, system integration, resilient supply chains, public-private partnerships, etc.), as well as indispensable sources of technical and “boots on the ground” expertise for end users and other European actors.

Following the launch of the EuroHPC Joint Undertaking in 2018, which created a large public infrastructure for computing capacity located across Member States, HPC SMEs have collaborated, and often have led, programmes and projects focused on further development of innovative architectures, technologies and industrial HPC use cases. Already today, European SMEs contribute substantially to the design, deployment and operation of the EuroHPC systems across all sizes. Moreover, the EU AI Innovation Package is progressively opening opportunities for SMEs and start-ups to develop European native infrastructures for the booming market of AI models and applications.

Despite these successes, however, European HPC SMEs continue to operate in a market that remains widely dominated by large non-European global companies. In particular, SMEs continue to face fierce competition by their larger counterparts when it comes to attracting and retaining talent and skilled staff necessary to stimulate innovation. To avoid or minimise potential negative scenarios, we cannot continue the “business-as-usual”. The EU needs to take decisive action and leverage the strengths of its HPC SMEs to make use of the huge untapped potential of European SMEs in every region of Europe. It is not only a matter of prosperity but of European and global sovereignty for the years to come.

Sharpening the tools

While the European HPC SME community has benefited greatly from EuroHPC to this date, some actions may further enhance the ability to compete vis-à-vis against the large global vendors and help even more HPC SMEs to successfully contribute to the European HPC ecosystem.

- 1) **Improve processes to support SMEs' participation in the EuroHPC R&I programmes** by:
 - a) reducing the overall administrative burden for SME participants through the implementation of a central financial management scheme with a 100% reimbursement rate for SMEs,
 - b) allowing SMEs' participants to contribute to R&I actions, independent from the total share of their contribution (i.e. not applying the minimum contribution rules to the SMEs).
 - c) EuroHPC opens twice per year 'SMEs Innovate' calls for a fixed amount of funding (e.g. 1,5M euros) and targeted for consortia of European SMEs addressing Innovative/Disruptive technologies as well as establishing a vibrant software ecosystem around them. The 'SMEs Innovate' call could then comprise the development and implementation of Innovative/Disruptive new hardware/software technologies based on upcoming new technologies to advantageously position European SMEs for the future - which are summarised in this SRA6 master and discussed in detail in the backing White Papers.
 - d) establishing regular meetings/videoconferences between SMEs and the EuroHPC JU to identify the development subjects where SMEs active in HPC could bring value.
- 2) **Design procurement calls (at all levels) in a way that encourages the participation of HPC SMEs, emerging players and start-ups.**
 - a) Reserving to European SMEs the participation to specific programmes.
- 3) **Since the investments required to develop new products, deploy a large and complex infrastructure or to support growth and scaling of HPC start-ups may be beyond the financial envelope of the SME, new financial concepts should be developed for**
 - a) facilitating the interactions with Venture Capital and private funding opportunities,
 - b) simplifying and improving the access to finance (beyond venture capital),
 - c) providing improved access to traditional forms of lending (loans, debt-financing, and guarantees),
 - d) developing financial instruments tailored to the size of investment and the needs of HPC SMEs and start-ups.
- 4) **Facilitate the transformation of research results towards globally competitive commercial products and solutions and improve SME's ability to protect their IP by**
 - a) working towards reducing the multiple regulatory, legal, fiscal and other bureaucratic differences across EU member states to enable SMEs to grow out of their local markets and to fully leverage the advantages of the EU single market instead,
 - b) establishing an interconnect framework between universities, research and technology institutions and SMEs to achieve a better exploitation of the existing knowledge⁸⁴,

⁸⁴ According to the European Patent Office (EPO), much of the knowledge generated in research institutions remains commercially unexploited. Data show that only about one-third of the patented inventions registered by European universities or RTOs are commercially exploited.

- c) raising awareness and providing better support for SMEs (e.g via tools and initiatives such as Horizon IP Scan) to handle the complex and costly procedures involved the filing IPR applications across fragmented national systems.
- 5) Further expand the academic programmes such as EUMaster4HPC to nurture and train talents who may be immediately productive when entering the industry.

Conclusion

The successes of European SMEs in HPC are a direct result of decades of EU framework programmes that emphasised the importance of SMEs as drivers of innovation in the European ecosystem and economy. With ongoing support and the removal of bureaucratic hurdles, the European single market provides great potential for its SME innovators to grow to become globally competitive technology leaders. Overall, moving from an SME-friendly to an SME-focused approach demands a firm commitment from all the stakeholders to preserve European sovereignty and be prepared for today's and tomorrow's challenges.



8. European Hardware Initiatives

The background of the slide is a complex, abstract geometric pattern in shades of blue. It consists of numerous overlapping, semi-transparent squares and rectangles that create a sense of depth and movement. The pattern is reminiscent of a digital or architectural structure, with lines and planes intersecting to form a three-dimensional effect. The overall color palette is monochromatic, ranging from deep navy blues to bright, glowing cyan highlights.

Multiple hardware initiatives had already been underway when we drafted the last SRA5 in 2021, with the main goal of achieving sovereignty in European HPC hardware infrastructures (including processors and networking technologies). It is critically important to take note of what was learnt from these projects as we pave the way forward towards European sovereignty and reduce dependency on non-European players. We indicate the lessons learnt and the feedback gleaned from each of these projects, which we hope are valuable inputs to policy and decision makers.

European Processor Initiative⁸⁵

Current Status

The European Processor Initiative (EPI) is a project currently implemented under the second stage of a Framework Partnership Agreement (FPA) between the Consortium and the EC under the regime of EuroHPC. The first stage of the EPI ran between 2018 and 2021 during the time of the Horizon2020 programme. The basic aim of EPI is to design and implement a roadmap for a new family of low-power European processors for extreme scale computing, high performance Big Data and many emerging applications.



At this time, an emulation infrastructure is running at SiPearl⁸⁶ for performance validation and benchmarking. The Software Development Vehicles (SDVs) for the European Processor Accelerator (EPAC) and Arm have now enabled various applications to be tested.

For the General Purpose processor, the Electromagnetic Transient Simulation (EMTS) for SiPearl's Rhea1 processor is now complete. The design of Rhea1 was finished in May 2024. Final steps prior to tape out are now underway. The board design from Eviden is also under way. The microarchitecture design of the Rhea2 chip is now undertaken.

For the accelerators, the test versions of the EPAC chip are brought up and are now running well. The SDV is setup and is now continuously enriched. The developed IPs for the VEC (a self-hosted RISC-V CPU with a Vector Processing Unit) and STX (Deep Learning and stencil specific accelerator) are now transferred to the EUPILLOT project - advancing further collaboration between these family of projects.

Lessons learnt

High-end HPC processor and accelerators design in Europe is creating a completely new industry and R&D community. There has been a realisation that during these first attempts, plans could not be precisely defined upfront especially for such complex devices. It is expected that with the maturity of all involved in the process, this will significantly improve. Sustained efforts are needed for such longer-term developments especially as European silicon proven IPs created by EPI are a foundation for many potentially upcoming projects developing European processors and accelerators needed for High Performance Computing.

The decision of EU policy makers to initiate those actions has proven to be visionary but it should be followed with the definition of longer-term support for continued activities for both EPI processor

⁸⁵ [Home - European Processor Initiative \(european-processor-initiative.eu\)](https://european-processor-initiative.eu)

⁸⁶ <https://sipearl.com/joint-projects-european-processor-initiative>

8. European Hardware Initiatives ■

verticals and with long term higher funding levels that would allow such activities to be executed in a timely way with concrete commercial results.

Supercomputing is broader than ever, embracing classical modelling and simulation that are increasingly boosted by Machine Learning models. The instability of the worldwide geopolitical environment is stressing how it is important to master high end chip technology for resilient and sovereign European research and industry.

EUPILOT⁸⁷



Current Status

The European Pilot aims to deliver the first All-European open source and open standards based software and hardware integrated HPC system by creating a set of accelerators for computationally intensive applications designed, implemented, manufactured and deployed in Europe. Two complimentary RISC-V accelerator chips are being developed, a VEC (Vector/Matrix accelerator) for scientific computing and MLS (AI Machine Learning and STX Stencil Tensor Accelerator) for AI & Machine Learning.

The project has recently completed the tape-out and received the initial test-chip for validating the 12-nanometer technology node. Its structure and interfaces are being evaluated. Simultaneously, preparations are underway for the tape-out of both the VEC and MLS chips while continuous enhancement of FPGA emulation of various components are aiming to ensure correctness and optimising performance. The software stack is being prepared—from firmware to tools, libraries, and applications—by porting and optimising them for optimal deployment on the hardware. Strategies are being developed, for the development of accelerator boards and assemblies, networking, power configurations, immersion cooling tanks, and all logistical requirements essential for the successful deployment of EUPILOT once all components are integrated into systems.

EUPILOT is about two-thirds complete at the time of drafting this SRA6 Master. The chips will be sent to fabrication that will power the HPC & AI accelerator systems that will in turn lead to the deployment of those systems. There has also been an ongoing collaboration with EUPEX project to work on a demonstration at the end of the project showing interoperability between the pilots.

Lessons learnt

Key lessons learned thus far in EUPILOT underscore the critical importance of tight coordination and integration across different technical areas to closely track project timelines. Timely adjustments and strategic resource reallocation have been vital for managing project timelines, especially since internal restructuring within partner organisations could cause significant delays for such a complex project. EUPILOT recommends policymakers to:

- Have better ways to orchestrate national and EC funding, or provide solid mitigation measures for eventualities in national funding
- Support flexibility in project timelines and budgets to accommodate unforeseen delays and adjustments

⁸⁷ [home - The EuPilot Project](#)

- Streamline processes for amendments to grant agreements and consortium agreements
- Facilitate access to world-class IP (e.g. like Europractice⁸⁸, but for common semiconductor IP blocks like DDR, PCIe, USB, UCIe, UALink, etc)

EUPEX



Current Status

The objective of EUPEX is to build a modular Exascale pilot system with the pilot hardware and software integrating European technology. The project aims to demonstrate the readiness and the scalability of the pilot technology in general and the Modular Supercomputing Architecture in particular, towards Exascale. The project also aims to prepare applications and European users to efficiently exploit future Exascale machines.

The project has now completed the workloads and workflow analysis. There is ongoing optimisation of applications for EUPEX architecture features. The Modular platform and compute hardware definitions are now complete. In terms of the software ecosystem, the Management software stack, execution environment, performance & energy efficiency tools, and storage & I/O components are identified, and their integration is ongoing. An Alpha test system (based on the ARM architecture provided in kind) is open to project users. An early access programme⁸⁹ has been launched to open the access to select external users, and one Centre of Excellence has already accepted to join this.

The project is collaborating with EUPILLOT to develop synergies through a demonstrator aiming at ensuring the interoperability between the EUPEX modular supercomputer and the EUPILLOT accelerator systems. There will also be work on optimising applications on the pilot system embedding GPPs & GPUs based on SiPearl Rhea1 & Nvidia chips.

Lessons learnt

EUPEX has a strong external dependency with SiPearl to get Rhea1 specification & chips. The uncertainties associated with the delivery dates might imply the review of some of the initial objectives, notably evaluating the full capabilities of the target system in terms of performance. A future action call could be used to fulfil such an objective. EUPEX project is also key to preparing the “Jupiter” system (Europe’s first Exascale machine at Julich Supercomputing centre) blades with Rhea 1 chips, and there is a belief that this effort has to be sustained beyond EUPEX project, especially to prepare systems with next chips generation (Rhea 2 – targeted for Alice Recoque, Europe’s next Exascale system hosted by the Jules Verne consortium⁹⁰).

Furthermore, the SEA Projects⁹¹ and the DEEP-Projects⁹² along with EUPEX have indeed proven Europe’s leadership in the Modular Systems Architecture (MSA) methodology for a very flexible and powerful way for the development of Supercomputers - evidenced by its usage in Europe’s first Exascale

⁸⁸ <https://europractice-ic.com/>

⁸⁹ [Eupex - Early Access Programme](#)

⁹⁰ [Signature of the Hosting Agreement for the Second European Exascale Supercomputer, Alice Recoque - EuroHPC JU \(europa.eu\)](#)

⁹¹ [SEA-Projects](#)

⁹² [DEEP-Projects](#)

8. European Hardware Initiatives ■

supercomputer, JUPITER. This architecture needs to be further promoted for adoption in HPC centres and further developed as part of a broader R&I strategy.

RED SEA⁹³



Current Status

Three years after its start in April 2021, the RED-SEA project concluded in March 2024.

Within Exascale systems, as well as emerging AI systems, interconnection networks serve as the backbone and play a crucial role in overall performance. These networks need to support high node counts, parallel processing systems, efficient connection to the data-centre, and emerging data-centric and AI-related applications. Additionally, they need to incorporate features such as efficient network resource management, in-network computing, and power-efficient support for accelerators and compute units. The goal of the RED-SEA project was to address these challenges by providing an innovative, low-latency, scalable, and reliable European interconnection network. In doing so, the project sought to contribute to the overall development of European exascale systems.

RED-SEA has successfully advanced interconnect network technologies. One significant outcome lies in the advancement of the European Interconnect BXI, with a focus on enhancing the current version (BXIv2⁹⁴) and laying the groundwork for its next generation (BXIv3). Another key achievement has been the exploration of new, efficient network resource management. For example, RED-SEA has enhanced network features such as collective operations by offloading collaborative work from compute resources (CPU, GPUs) to the network components, and improved congestion control reducing global latencies.

Moving forward, the vision for RED-SEA is to continue advancing European technological capabilities in high-performance computing, AI, and interconnect solutions. This involves 1) further development and commercialisation of BXIv3, incorporating the latest advancements in hardware and software in the future funded project NET4EXA⁹⁵, 2) exploring new network technologies such as new protocols, photonics, 3) contributing to standardisation bodies and consortiums to ensure compatibility and interoperability across systems, fostering a more open and collaborative ecosystem, especially within the UltraEthernet Consortium⁹⁶.

Lessons learnt

Overall, the RED-SEA project has contributed to advancing European capabilities in high-performance computing, AI, and interconnect solutions, laying the foundation for future innovations and collaborations in the field. The following are some of the lessons learnt from RED-SEA and policy aspects to be considered going forward:

Enhanced Funding: There is a need to prioritise funding for components of HPC and AI infrastructure, e.g. network interconnection, to ensure that it remains at the cutting edge and supports complex simulations, data analysis, and research.

⁹³ [RED-SEA project – A EuroHPC project \(redsea-project.eu\)](https://redsea-project.eu)

⁹⁴ [BXI V2 - Atos](#)

⁹⁵ [European Interconnect for HPC and AI - NET4EXA](#)

⁹⁶ [Ultra Ethernet Consortium](#)

Industry Adoption: Encouraging the adoption of HPC and AI technologies including network interconnection technologies across various industries, including healthcare, finance, and manufacturing, to improve efficiency and competitiveness is increasingly necessary.

Standardisation Efforts: More participation and efforts are needed to standardise HPC and AI technologies, facilitating interoperability and collaboration across different platforms and systems.

Encouraging Sustainable Technology Development: Last but not the least, policymakers should prioritise sustainable technology development to address environmental challenges and support a greener future. By investing in energy-efficient computing and promoting sustainable practices, we can align technological advancements with environmental goals.

eProcessor⁹⁷

Current Status



The eProcessor project aims to build a new open Out of Order (OoO) processor based on RISC-V and deliver the first open European full-stack ecosystem based on this CPU. The eProcessor technology will be suitable for HPC and embedded applications.

This 4-year project (1/4/21 - 31/3/25) is currently in its final stages.

On the hardware side the following are the developments; (i) the ASIC containing the single-core design has been taped-out and is currently being packaged, (ii) the PCB and the carrier board for the ASIC have been developed and fabricated, and (iii) the RTL of the multi-core design is frozen and being tested on the FPGA SDV.

On the software side, (i) the Linux operating system is working successfully on the FPGA SDV and is ready to be tested on the ASIC, (ii) the runtime libraries and performance tools are ported to RISC-V and are ready to be ported to the ASIC and to the FPGA SDV, and (iii) the application use cases have been ported to the FPGA SDV and are ready to be ported to the ASIC.

One of the main objectives of the project is to contribute to the portfolio of European IP blocks that can be used in high-performance processors. To this end, the eProcessor project has developed new IP blocks (such as the out-of-order core) and has improved and/or extended existing IP blocks coming from other projects (such as the vector accelerator, the home node, the last level cache and the network-on-chip coming from EPI). These IP blocks can be leveraged in upcoming European projects such as DARE. In addition, the owners of each IP block have defined their respective exploitation plans.

Lessons Learnt

The project has been able to successfully develop competitive IP blocks that can be used in high-performance processors (including the CPU core, the vector accelerator, the AI accelerator, the caches, the on-chip interconnection network, etc) and a complete software stack to run the applications of interest for the project. Along the way, the project encountered different issues and challenges which they have been able to successfully resolve providing some important lessons. The two most important

⁹⁷ [Deliverables – eProcessor](#)

8. European Hardware Initiatives ■

lessons learned follow. One important lesson learnt is that doing two tape-outs in a 3-year project is extremely challenging. Fixing issues in the RTL development and verification is very hard and time consuming, so every small issue can cause non-negligible delays within the whole project. On top of this, the current socio-political situation makes it hard to predict the price and the turnaround time of tape-outs 3/4 years into the future (from writing the project proposal to doing the actual tape-out). On the technical side, it has been learned that the frequency of the hardware designs is severely limited by the latency of the memories (SRAMs). Off-the-shelf SRAM libraries included in the memory compilers are significantly slower than the heavily optimised SRAM libraries used in the cutting-edge high-performance processors of industry-leading manufacturers.

It would be good to find a way to overcome the issues explained above in future projects, from a policy perspective.

Other activities

The Framework Partnership Agreement (FPA)⁹⁸ for European hardware and software technologies, based on RISC-V in order for delivering high-end processors, accelerators and systems, and implementing the associated demonstrators - was awarded to the DARE consortium in 2024. The DARE consortium consists of partners across Supercomputing centres, research centres, Industrials (incl. SMEs) & universities. The FPA aims to develop a strong open RISC-V ecosystem in Europe suitable for Exascale and Post Exascale supercomputers.

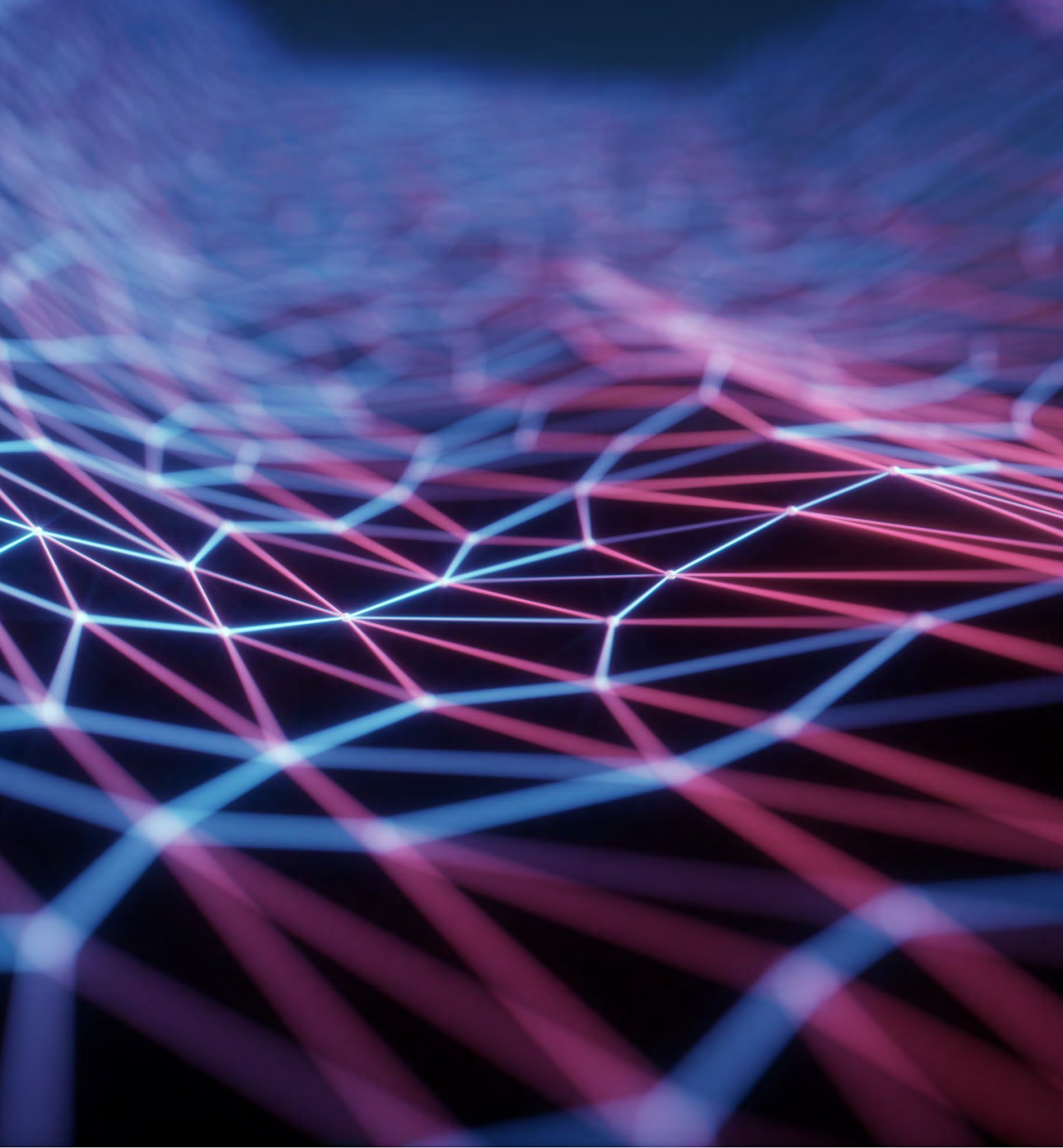
The preparation of the Specific Grant Agreements is now in process. For the first phase⁹⁹ the goal is to deliver tape outs of RISC-V based General Purpose Processors and two accelerators - an AI accelerator and a Vectorial accelerator. This phase will also develop the RISC-V software stack that includes programming models, runtimes, tools, libraries, etc. Appropriate performance levels will be targeted, and hardware/software co-design will be employed.

Whilst these are indeed very positive initiatives, we recommend the development of a well-structured seeding and deployment plan in the European industrial and academic environments, to ensure the success of this ambitious project.

⁹⁸ [EU Funding & Tenders Portal \(europa.eu\)](https://europea.eu)

⁹⁹ [Specific Grant Agreement \(SGA\) for the development of a large-scale European initiative for HPC ecosystem based on RISC-V - EuroHPC JU \(europa.eu\)](https://europea.eu)

9. Conclusion and Acknowledgements



In this SRA6 master (released late 2024) we have summarised the key R&I recommendations from the different areas of HPC and R&I policy priorities needed in Europe at this time to bolster Research and Innovation in HPC and develop the HPC ecosystem in general. The SRA Master will be supported by the following Backing White Papers, which will cover more details on each of the different technical topics. We refer the reader to these papers that will be available at [Strategic Research Agenda – ETP4HPC](#) within the coming weeks.

Research Domain White Papers

- System Architecture
- System Hardware Components
- System Software and Management
- Programming Environment
- I/O & Storage
- Mathematics & Algorithms
- HPC Usage
- Non-Conventional Architectures
- Ecosystem technologies

Thematic Trend White Papers

- Quantum Computing & HPC
- Energy Efficiency & Sustainability
- AI and Foundational Models

Once we release these white papers, we will continue discussions with the HPC community and the ecosystem, and release “Agile” White papers based on upcoming trends and endeavour to keep the SRA6 Master up to date to reflect these changes in the next two years.

Acknowledgements

The ETP4HPC Office would like to thank the SRA6 Work Group leaders for providing valuable inputs into this Master. Each of these Work Group leaders have actively worked with a team of experts, providing invaluable contributions over many months.

- Nico Mittenzwey, MEGWARE
- Fabrizio Magugliani, E4
- Marc Durantou, CEA
- Craig Prunty, SiPearl
- Pascale Rossé-Laurent, Eviden
- Manolis Marazakis, FORTH
- Paul Carpenter, BSC
- Gabriel Antoniu, Inria (BDVA)
- Sarah Neuwirth, University of Mainz
- Philippe Deniel, CEA
- Dirk Pleiter, KTH
- Utz-Uwe Haus, HPE
- Erwin Laure, MPCDF
- Andreas Wierse, SICOS
- Tobias Becker, MAXELER
- Robert Haas, IBM RESEARCH
- Michael Malms
- Hans-Christian Hoppe, FZJ (& ParTec)
- Valeria Bartsch, ITWM
- Sagar Dolas, SURF
- Ondřej Vysocký, VSB
- Maria Perez, UPM
- Andy Forrester, HypeAccelerator
- Kristel Michielsen, FZJ
- Estela Suarez, FZJ

The ETP4HPC Office would also like to thank the ETP4HPC SME Working Group, the ETP4HPC Industrial User Working Group, and the project coordinators of European Hardware Initiatives for their valuable inputs.

ETP4HPC Office SRA6 Contributors

- Dr. Sai Narasimhamurthy - Chief Editor
- Marcin Ostacz
- Gabriella Povero
- Pascale Bernier-Bruna
- Jean-Pierre Panziera

The background is a dark blue, abstract composition of overlapping, semi-transparent geometric shapes and lines, creating a sense of depth and movement. Scattered throughout are small, glowing blue squares and lines, some of which appear to be part of a larger, faint grid or network structure. The overall aesthetic is futuristic and digital.

APPENDIX: SRA6 Process & Structure

The SRA6 is the work of more than 100 European experts in the different domains of High Performance Computing.

SRA6 Process

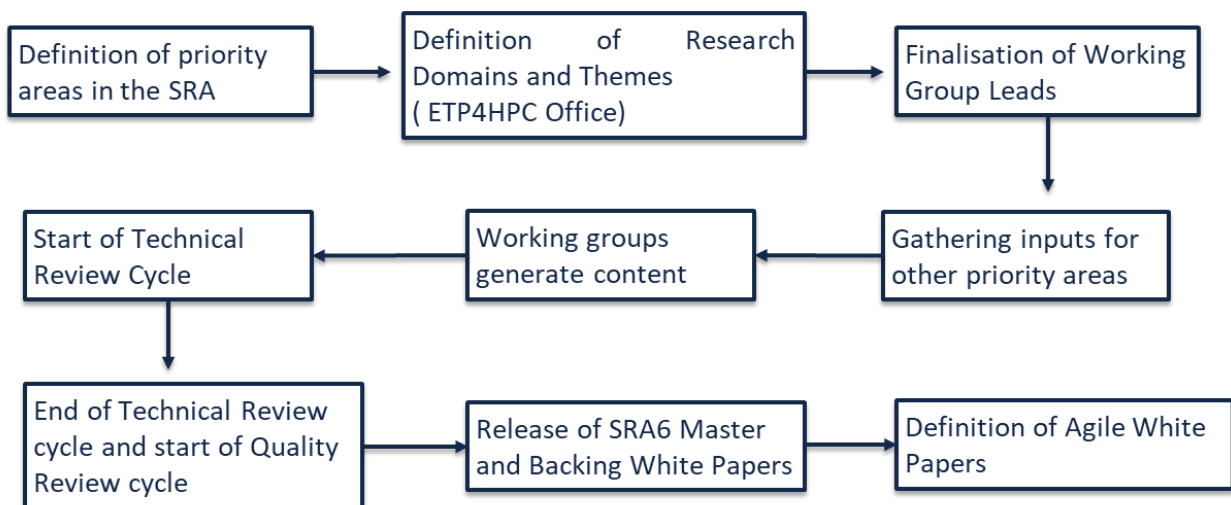
The SRA6 process starts with the definition of top-level priorities that need to be defined in the SRA by the ETP4HPC Office with presentations/feedback from the ETP4HPC’s General Assembly which consists of about 110 organisation representations from Europe at the time of drafting the SRA6. The different Research Domains and Themes are also defined and finalised at that point.

The different working groups corresponding to the Research Domains and Thematic trends are also formed. The SRA6 working groups are typically led by two individuals with established track record in each of the Research Domains and Thematic Trends. The ETP4HPC office also solicits voluntary contributions into the different Research Domains and Thematic Trends from the ETP4HPC member base. The members who are interested to contribute to any area are vetted by the ETP4HPC Office and the Work Group leads - where only individuals with established background in a certain area are chosen to contribute to the Working Group. The Research Domains and the Thematic Trend Working Groups build up the content with regular synchronizations with other working groups and the ETP4HPC Office. The members of the ETP4HPC RIAG are also included in the conversations on these topics.

Inputs for other priority areas are also sought through separate discussions, for example, with ETP4HPC Industrial User Working Group (for the inputs on Industrial Usage) which also includes discussions at BoFs (Birds of a Feather Sessions) at events such as ISC, & ETP4HPC SMEs (for the inputs on European SMEs).

The draft content that is generated now undergoes the technical due diligence through the “Technical Review Cycle” where each of the sections are reviewed by at least 2 chosen experts.

The end of the Technical Review cycle then begins the start of the Quality Review cycle where the final editions are made by the ETP4HPC office and editorial teams. The final SRA6 is then released along with the backing White Papers after including the design elements. At this point we also start to define the upcoming Agile White Papers that will be released post SRA6 Master and Backing White Paper releases.



SRA6 Structure

The SRA6 structure is driven by the key priority areas and the decision to keep separate the SRA6 Master and the backing White Papers. The key priority areas are defined in the following categories:

Research Domains & their Recommendations:

The Research Domains discuss progress and research priorities and recommendations in the following technical topics:

- **System Architecture** deals with the evolution of overall HPC system architecture including the integration of the different hardware and software components, including processors, memory, network, storage and associated system administration aspects.
- **System Hardware** discusses the evolution of the above-mentioned hardware components including but not limited to processors and accelerators
- **Non-conventional architectures** explores technologies that are not very common place in HPC today. We explored FPGAs etc in this context previously. This could also include (but not limited to) for example Neuromorphic computing
- **I/O and storage** looks into HPC application I/O and data storage aspects including data management related issues
- **HPC in the Digital Continuum** looks into synergies with domains neighbouring HPC (IoT, Cloud, Cybersecurity etc)
- **System software and management** explores the evolution of the infrastructure software that binds, connects and manages the various system and hardware entities within the HPC infrastructure
- **Programming environments** explores the interfacing and interactions of applications and workflows with the underlying HPC systems
- **HPC Use** explores the various cutting edge HPC applications and use cases that will exploit the HPC infrastructure
- **Mathematics and algorithms** looks deeper into how the underlying mathematics and algorithms associated with use cases have evolved to keep pace with HPC technologies

Thematic Trends & their Recommendations:

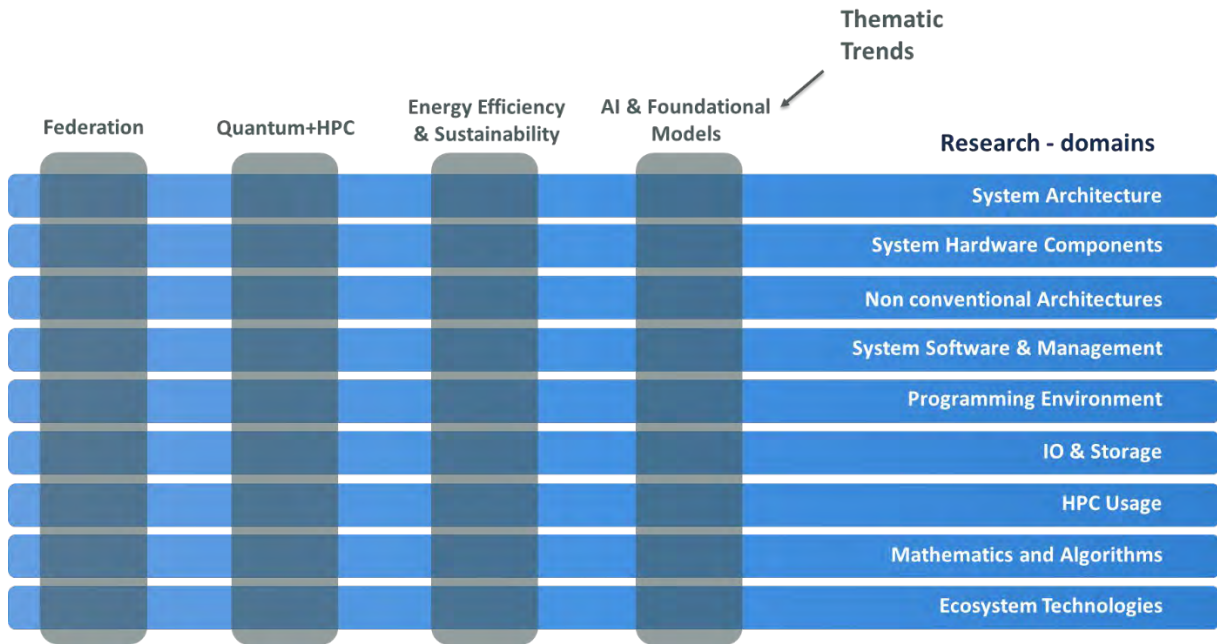
The thematic trends discuss the latest trends in HPC in the European context that cut cross the various Research Domains and provide their recommendations:

- **Federation** looks at how the various distributed HPC sites/infrastructures, and associated data infrastructures in Europe can be smartly combined to provide value for distributed users and applications
- **Quantum & HPC interactions** looks at the evolution of Quantum Computing¹⁰⁰ in the light of HPC and how they can be hybridised and used with classic HPC infrastructures

¹⁰⁰ Please note that Quantum Computing, per se, is not a subject of this SRA - but we focus mostly on interactions and interfaces with classical HPC.

- **Energy Efficiency and sustainability** will look at how evolving HPC technologies eventually needs to keep pace with global climate and sustainability goals
- **AI and Foundational models** look at how these evolving class of problems will use and get value from HPC

The following figure depicts the various research Domains and thematic Trends, each of which will be described in their respective Backing white paper released immediately following the Master, with the exception of Federation, which will be delayed and released as an Agile White Paper for reasons which we will explain.



AI Explosion & HPC:

Proliferation of AI applications and use cases have been the mainstay of global technology in the last couple of years. In this SRA, we cover the European/ETP4HPC perspective of how HPC technologies need to adapt to the explosion of AI applications and use cases, in addition to a separate thematic trend discussing the AI research topic.

European Exascale Hardware Initiatives (& RISC-V):

We discussed the progress in European Exascale hardware (and related system) initiatives considering the push for European sovereignty in this area. There has been some exciting and promising efforts here during the last 2 years since the last SRA.

Post Exascale:

We believe this is an appropriate time to discuss Post Exascale trends and topics now that we are at the cusp of achieving Exascale in Europe¹⁰¹. We discussed Post Exascale from the perspective of each of the Research Domains and Thematic trends and then summarise.

European Industrial Users and SMEs:

We presented the latest perspectives from the European Industrial users and European SMEs on the opportunities and challenges they face.

Releasing “Federation” as an Agile white paper

Achieving federation of HPC resources (computing and data) has been the objective of a multitude of collaborative projects and initiatives, starting with the Unicore and Globus systems in the late 1990s, and leading to European federation systems like for instance FENIX¹⁰² and EOSC¹⁰³ that are used by a multitude of scientific research communities. In 2023, the EuroHPC JU has started the procurement of a SW solution for the federation of its supercomputers¹⁰⁴. This solution will need to achieve wide-ranging and challenging objectives (see for instance https://eurohpc-ju.europa.eu/eurohpc-federation-platform_en). At the time of writing the SRA6, the procurement process was still ongoing, and no public information as divulged by EuroHPC JU or the tendering parties about the specific planned approach, architecture and capabilities. Acknowledging the size and import of that EuroHPC JU investment, it does not seem appropriate to formalise a Federation research agenda at this time.

Once sufficient information becomes available about this EuroHPC JU federation system, we will plan to release an Agile White Paper on advantageous and feasible additional steps to further improving federation of European HPC resources, including on integrating existing, research driven infrastructures for compute and data resources and leveraging the latest concepts created in the Cloud sphere. This paper will also look at significant data federation efforts (E.g.: EU commission driven SIMPL¹⁰⁵ initiative), and at extending federation for example to scientific instruments and novel potentially disruptive computing architectures.


¹⁰¹ [JUPITER - Exascale for Europe \(fz-juelich.de\)](https://www.fz-juelich.de/en/jupiter)

¹⁰² [Home | FENIX \(fenix-ri.eu\)](https://www.fenix-ri.eu/)

¹⁰³ [The European Open Science Cloud - EOSC Association](https://www.eosc.eu/)

¹⁰⁴ [EU Funding & Tenders Portal \(europa.eu\)](https://ec.europa.eu/eu-funding-tenders-portal/)

¹⁰⁵ [Simpl: Cloud-to-edge federations empowering EU data spaces | Shaping Europe's digital future \(europa.eu\)](https://ec.europa.eu/digital-single-market/en/simpl-cloud-to-edge-federations-empowering-eu-data-spaces-shaping-europes-digital-future)

A photograph of a server room with rows of server racks. The image is overlaid with a complex digital graphic consisting of blue and orange lines, squares, and dots, creating a sense of data flow and connectivity. The server racks are illuminated with blue light, and the overall atmosphere is futuristic and high-tech.

Cite as: Narasimhamurthy S. et al. ETP4HPC's SRA 6 – Strategic
Agenda for High-Performance Computing in Europe – 2024. Zenodo
<https://doi.org/10.5281/zenodo.14268784>

DOI: 10.5281/zenodo.14268784

© ETP4HPC 2024