WHITE PAPER

HPC Usage

February 2025
An SRA 6 White Paper

EUROPEAN TECHNOLOGY
PLATFORM FOR HIGH
PERFORMANCE COMPUTING

Co-leaders

Erwin Laure (MPCDF)

Andreas Wierse (SICOS)

Table of contents

Executive Summary

This is a white paper released as part of the ETP4HPC's Strategic Research Agenda 6.

This covers the Research trends, Challenges, Post Exascale Vision and R&I priorities in the area of HPC Use Cases.

Glossary of acronyms

CFD - Computational Fluid Dynamics

LLMs - Large Language Models

Research trends and current state of the art

Sustaining excellence and European world-leadership in HPC applications is key for European science, industry (incl. SMEs) and the public sector. There is a breadth of applications in the fundamental, applied and social sciences as well as in industrial application fields where computing plays a pivotal role[1]. And computing here not only applies to the well-known HPC field, but increasingly includes AI, especially the area where HPC and AI overlap and can benefit from each other; e.g. AI can help reducing the search space, allowing to run fewer full scale simulations while at the same time more detailed simulations can be run, thus improving the overall outcome:

Climate, Weather, and Earth Sciences

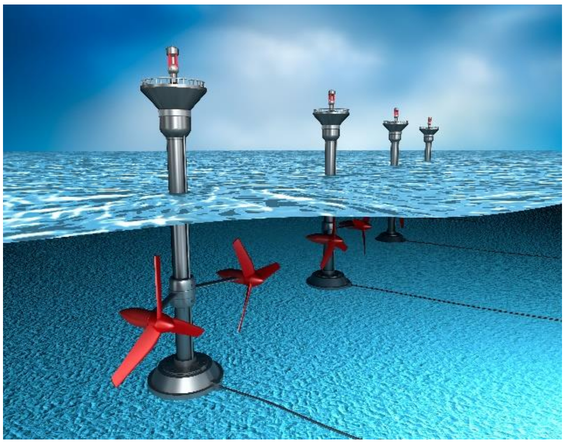
Simulations are critical in Climate, Weather, and Earth Sciences. Exascale resources will enable sub-kilometre scale resolutions, a more realistic representation of all Earth-system components, better mathematical models, and larger ensembles of simulations for uncertainty quantification. This will extend the reliability of forecasts to the extent needed for the mitigation and adaptation to climate change at global, regional and national levels, in particular with respect to extreme events. In addition, coupling of mesoscale and microscale models will enhance the accuracy of pollutant propagation in cities, mid-term power generation in wind farms, etc as well as “nowcasting” predictions requiring short term HPC access. In analogy to weather and climate prediction, much enhanced simulation capabilities of solid Earth physics from higher spatial resolution and seismic frequencies down to 10Hz will enable a break-through in the detection and prediction of the precursors of volcanic eruptions and earthquakes, and their impact on infrastructures. High resolution interferometry from satellite data can also help detecting small displacements of natural as well as artificial objects. A prediction capability at this level of detail is crucial for a wide range of societal impact sectors for food and agriculture, energy, water management, natural hazard response and mitigation, and finance and insurance. Recently, AI-powered models have gained importance in the field and are expected to have significant impact in future.

Life sciences, medicine, and bioinformatics

High-end computing capabilities are becoming increasingly important for life sciences, medicine, and bioinformatics and will have tremendous impact, e.g. for enabling personalised medicine. Researchers are already able to rapidly identify genetic disease variants, and it will become possible to identify diseases that are caused by combinations of variants and design treatments tailored both to the patient and state of the disease. Structural biology will increasingly rely on computational tools, allowing researchers to predict how the flexibility and motion of molecules influences function and disease. Deep learning techniques will provide more specific diagnoses and treatment plans than human doctors, making medical imaging one of the largest future computing users. At anatomical scales, organ level simulations present both a challenge and an opportunity: the Virtual Patient modelling for precision medicine. It is the ultimate example of supercomputer usage for multiphysics/multiscale modelling problems. Tumours can be simulated by following the fate of thousands, millions or even billions of cells as their internal circuitry of signalling molecules respond to each other and their environment. This approach can be used to predict the effect of targeted therapies,

[1] See for instance: The Scientific Case for Computing in Europe (period 2018-2026), PRACE

which is particularly useful to triage potential combination therapies based on the patient's molecular profile. At organ level, modelling is done by tightly coupling different physics (e.g. fluid, tissue, electrophysiology, chemical reactions, heat, transport of large bodies, particles or species) with contributions from different temporal and spatial scales (cells, tissue, organ, system). To make things more complex, these problems present issues such as patient variability and comorbidities in complex geometries, with extremely difficult validation. A strategy to address these two issues is to run problems in a virtual patient population. All these things together make the use of supercomputers a decisive factor.



Energy

For Energy applications, multiscale simulations, coupling ab initio calculations, all atom and coarse-grained molecular dynamics and continuum models (based for example on CFD) are of importance to improve the efficiency of hydropower, wind turbines, nuclear power, photovoltaics and, not least, batteries and high-voltage cables to enable transmission and storage. Likewise, large scale simulations are essential for the discovery and optimisation, on real scales, of renewable forms of energy, their storage and distribution. The oil and gas industries are moving to full waveform inversion combined with neural networks for accurate detection. Exascale resources will make this technique feasible, allowing more accurate predictions of reservoirs and, although still fossil fuels, oil and gas have much reduced CO_2 emissions and air pollution compared to those produced by coal, which is still the dominant source of energy in the world. Accurate magnetohydrodynamic simulations of plasma are critical for fusion energy, as in the ITER project.

Engineering & Manufacturing

Computing is already used widely in Engineering & Manufacturing. Engineering applications based on computational fluid dynamics, combined with orders-of-magnitude-faster resources, will enable direct numerical simulations of the governing equations of fluid mechanics, including the effect of chemical reactions (e.g. combustion) and the presence of multiphase systems, with better accuracy, leading to improved designs and thus significantly better fuel efficiency e.g. for cars and airplanes, while also helping us understand phenomena such as cavitation, flow separation and pollutant formation. New data-driven approaches will enable scientists in academia and industry to integrate all aspects of design in computational models, use information from internet-of-things sensors, include uncertainty quantification in predictions, and consider the entire life cycle of a product rather than merely its manufacture (digital twins). In addition, the increased available computational power will allow increasing the resolution as well as improving physical models. This in turn will improve the quality of simulation results especially for complex phenomena.

Chemistry & Materials Science

Chemistry & Materials Science will remain one of the largest users of computing, with industry increasingly relying on simulation to design, for example, catalysts, lubricants, polymers, liquid crystals, and also materials for solar cells and batteries. Electronic structure-based methods, coupled with all atom and coarse-grained molecular dynamics, and with continuum models (e.g. CFD) will handle systems, properties and processes of increased complexity, and drive towards extreme accuracy. These methods are being complemented both with multi-scale models and data-driven approaches using high-throughput and deep learning to predict properties of materials, accelerate discovery and to design innovative more efficient processes to produce these novel materials. This will enable researchers to address the grand challenge of designing and manufacturing all aspects of new materials from scratch, ushering in a new era of targeted manufacturing, that fully addresses also the issues of recycling of critical raw materials. Moreover, electronic structure methods can further enable machine learning and AI approaches by providing large sets of computed and curated data for training.

Electronics Engineering

Computational models are fundamental to the advancements in Electronics Engineering, particularly due to the progression of nanoscale technologies and beyond-CMOS innovations. Ab initio calculations, molecular dynamics, finite element modelling have become essential for designing, studying and developing novel materials and devices as well as the physical phenomena, especially as devices face miniaturisation challenges which increase quantum effects. These models require using HPC to handle complex calculations and simulations which are the base for designing and enhancing the performance of quantum and nanoscale devices, and related architectures, circuits, and systems.

Complex simulations for global challenges

Global Challenges, like urban air pollution, financial crises, behaviours in social networks, pandemics (like COVID-19), food and water shortages, management of disasters or conflicts, and migration streams require complex simulations. Here often agent-based systems with a huge number of elements are used, but also specific coupling with other simulations (such as weather and climate predictions and other multiphase CFD simulations in 3D models representing digital twins of real locations) increases the complexity. These simulations can be used by policy makers in order to understand how to solve existing global issues or how to mitigate their negative effects. Therefore, it is necessary to advance research on agent-based simulations, coupling technologies and executing ensemble methods, as well as leveraging state-of-the-art HPDA and AI techniques. All of this significantly contributes to the retrieved models, their understanding, and the further extraction of knowledge for scientific-based decision making.

Fundamental sciences

A world class European computational infrastructure will expand the frontiers of fundamental sciences like physics and astronomy, supporting and complementing experiments but also supporting the researchers to go beyond what experiments allow. Researchers will be able to simulate the formation of galaxies, neutron stars and black holes, predict how solar eruptions influence electronics, and model properties of elementary particles. This will explain the source of gamma-ray bursts in the universe, advance our understanding of general relativity, and help us advance understanding of the fundamental structure of matter by means of simulating the theory of strong interactions called quantum chromodynamics. This fundamental research itself leads to advances in the state-of-the-art of scientific computing and helps attract new generations to science, technology, engineering and mathematics.

Digital twins

Digital twins started to emerge in various domains, with Destination Earth being one of the flagship projects. But also in other areas, digital twins are being explored, like in fusion energy and even the operation of HPC systems themselves might be subject to being modelled as a digital twin. Digital twins require huge amounts of computational resources (both CPUs and GPUs) for the multi-physics, multi-scale simulations required as well as huge storage space allowing real-time processing of sensor data.

AI

AI has become an essential tool in many scientific domains and new foundation models are being created virtually every day. The model learning process is time-consuming and requires a huge quantity of GPUs and memory. So is also the fine-tuning process, when generic models are being modified for specific application purposes. For example, the recent Llama 3 training required the use of 16'000 GPUs, and Llama 3-70B needed 6.4 million of GPU hours to complete. HPC infrastructures may be an ideal target for those applications.

Throughout this spectrum of applications, handling of complex data plays an increasingly important role; this is the case for scientists for example with the results of experiments, but also industry increasingly deals with ever more data that is collected continuously, e.g. coming from customers who use their products. Existing research fields are beginning to use deep learning, large language models, and other AI-techniques to generate knowledge directly from data instead of first formulating models of the process, so next-generation infrastructure must be able to handle these applications with dramatically increased data storage and I/O bandwidth capabilities. This will be needed for e.g. autonomous driving, Industry 4.0, Cyber Security and the Internet of Things, but will also enable the application of computing across a whole range of non-traditional areas including the humanities, social sciences, epidemiology, finance, promoting healthy living, determining return-on-investments for infrastructure by considering behavioural patterns and, not least, in helping to develop society and secure democracy.

As the previous topics show, industry can be found almost everywhere when it comes to large scale simulation science. While still, especially when SMEs are using simulation, data analytics, and AI, industry does often not yet require the fastest systems available, it is absolutely crucial that industry has a direct link to those who use these systems and to their technologies in order a) to prepare for their own use as soon as they have a need for this kind of performance and b) to be able to run exceptionally complex tasks, not every day but for particularly large challenges when this compute power is needed to open up new possibilities. If industrial companies are not able to seamlessly connect to the highest-end simulations, data analytics, and AI technologies, they will suffer severe disadvantages compared to those that do. This also requires lowering the entrance bar, through approaches like “simulation as a service” and support from dedicated HPC teams.

Challenges for 2025-2029

AI & Co-Design

Hardware development is now mostly driven by AI requirements. Traditional HPC applications, based on FP64 arithmetic, are faced with stagnating, or even declining performance. This is a major challenge as simply investing in new hardware is in most cases not providing a boost in performance anymore. Instead, application developers will have to invest significantly in improving application performance and scaling by improving both, algorithms and implementations. Where possible, applications may also exploit the advances of AI-tuned hardware by using lower precision and specific hardware (e.g. tensor cores), which often also requires changes in the algorithms used.

These changes in algorithms and implementations not only require knowledge about the physical problems to be solved and also knowledge of and sufficient access to these new technologies for application developers. An effective collaboration, including co-design processes, between application and technology developers is crucial for a successful HPC ecosystem to ensure technology can be fully exploited when available. It must also be recognised that not all application maintainers possess the knowledge to independently and effectively conduct the necessary code modifications for new technology. Specialist knowledge and support will also be needed to assess the level of revision needed to move applications which are successful at small scales to large-scale HPC infrastructures. Application co-design will involve providing support to application codes and tools as they move through these processes. In addition, the challenge of transferring knowledge between computational and scientific specialists extends beyond porting and optimisation of applications and should include training existing and upcoming researchers in skills critical to fundamental HPC usage. This is even more essential in research fields that have a relatively immature level of uptake and competence in HPC usage. Incorporating all of these issues into a wider approach will foster an ecosystem of skills and expertise in Europe by retaining scientific talent and enhancing the competitiveness of industries on the global market.

Large Language Models (LLMs) can help with code portability. LLMs are known to help users with different tasks, such as text translation, summarization, and code generation. Nowadays, a diverse range of foundation models are available and generate information in various domains through fine-tuning. But the provided answers of those models are not reliable enough, "hallucinations" often happen. Assuming a large and sufficient dataset of portable codes for newer architectures is available, the use of specialized LLMs will help experts prepare their codes.

Education & Training

Educating new generations of HPC experts and computational scientists is crucial for a healthy HPC environment. In addition, such people need rewarding employment conditions with clear and structured career paths to retain skilled personnel and avoid a brain drain to other parts of the world. Unfortunately, within the traditional academic systems that exist in most European countries this is still difficult for computational experts in certain disciplines and even more so for interdisciplinary experts.

Access to increased computational power and to applications exploiting the features of the system remains crucial to enable more detailed and large-scale (compute intensive) modelling and simulations.

Expanded Scope of HPC Use Cases

New approaches like data-driven computing, High Performance Data Analytics, and AI, and their convergence with classical HPC enable new opportunities and require new capabilities. This includes efficient access to large amounts of data with low latencies and high-bandwidth (HPC system internal) as well as efficient ingest of huge amounts of data into the HPC centres, always following the FAIR principles. Support for new and large comprehensive workflows and ensembles encompassing orders of magnitude more active tasks or computational jobs than today will also be needed. Additionally, some simulations require coupling with others, since their results need to be synchronised, requiring further development in coupling techniques (from data sharing to message passing) and adequate resources access and allocation mechanisms. It should be mentioned that in the US or in Japan, recent large (DoE) Call for Tender are asking for Hybrid architecture integrating closely on-premises HPC with hyperscaler's infrastructure (building a "Cloud" zone in the premises, or expanding logically the infrastructure in the hyperscaler facilities).

New Infrastructure Requirements

Many applications are severely limited by memory bandwidth or communication latency, and as the throughput of floating-point operations has increased faster than the data transport capabilities, even many traditionally floating-point bound applications are now highly memory sensitive. Intra-node and inter-node communication require drastically reduced latencies and high-speed networks, particularly for algorithms based on fast iterations of short tasks so that they can achieve significantly improved performance and strong scaling. In this respect, the process of co-designing the hardware with the applications in mind can already be seen with the addition of High Bandwidth Memory (HBM) interfaces not only to accelerators such as GPUs but also on general purpose processors such as the Fujitsu A64FX processor and in upcoming x86 variants such as the Intel Sapphire Rapids CPU. The impact that this has on application performance has already been demonstrated with the Fugaku system that is based on the A64FX processor and where significant real application speed-ups were obtained on a wide range of scientific codes. Another area that has seen a large degree of co-design between hardware and applications is reduced precision. Modern GPU and CPU processors now offer native FP16 support that mostly targets matrix-matrix operations found in machine learning workloads but there are already a number of applications such as in weather and climate that are adapting parts of their model to exploit reduced precision with promising results both in terms of speed-up and accuracy.

Storage and I/O requirements are expected to grow even faster than compute needs, with much larger data sets being used e.g. for data-driven research and machine learning. This is not limited to the amount of storage, but data-heavy applications will also need exceptionally high-bandwidth parallel file systems, and/or advanced data caching solutions on each node. This increase in storage and I/O resources must be coupled with provisioning of a large-scale end-to-end data e-infrastructure to collect, handle, analyse, visualise, and disseminate petabytes to exabytes of data. In addition, some applications like genome analyses or AI applications require ancillary data like reference datasets or underlying models. Making them available on fast storage with pre-configured access mechanisms can significantly speed-up such applications.

Code Maintenance & Portability

Long-term maintenance and portability of codes (both in terms of performance portability as well as compiling/building the codes on new architectures) are other important factors, requiring standardised, open, and supported programming environments and APIs, including container technologies, supporting a wide range of different hardware technologies. Investing in tools automating and tracking deployed software stacks can help to sort out the challenging combination of codes and architectures. In addition, the software engineering practices like unit testing, continuous integration and deployment need to improve. These efforts, typically performed by the development teams, need to be complemented with community benchmarking activities, helping to select the right software for certain environments and ensuring the validity of the results. A particular challenge is the uncertainty about future hardware developments. Porting established codes is a major undertaking and with uncertainties about the long-term future of certain hardware technologies, there is often a reluctance to engage in these expensive endeavours. Performance portability frameworks can shield application developers from hardware changes to some extent, but long term, stable hardware roadmaps are of equal importance.

Many of today's applications are made up of millions of lines of code. Analysing these monolithic codes to identify expensive computational kernels to port to future architectures can be difficult. Creating application dwarfs to simplify the execution and interaction with the algorithmics can be a solution to enhance this task. As an example, in Earth Sciences, dwarfs represent functional units in the forecasting model, such as an advection or a physics parameterisation scheme, which also come with specific computational patterns for processor memory access and data communication. To ensure a smooth user experience, dwarfs should be distributed with an accurate user guide and examples of input files or namelists. In addition, modularisation of codes and the application of modern software engineering best practises are important.

Algorithm Development

While in the longer-term computing leadership will require the development of alternatives to the current technologies, including quantum-, data-flow, unconventional computing (neuromorphic, RNA field-coupled, spintronics…), there is consensus that even today the fundamental mathematical and computer science algorithms that are needed to meet the requirement of leadership science and industrial competitiveness are not in place. The energy-efficient, application-oriented next-generation computing platforms therefore require an ambitious programme of algorithm development integrated with the co-design/co-development of the overall infrastructure and sustained over longer timescales than the usual 3 to 5 years funding cycles. Specifically, we could address the connection with other European programs to enhance the integration between our scientific codes and the emerging RISC-V and ARM architectures and, connected to this, strategies for optimizing our codes for improved vectorization and memory hierarchy, among other aspects. This is particularly true for quantum computing and the new European processor (EPI), where in Europe and globally a vibrant research environment is being built rapidly. World-leading technology and applications innovation in this field requires significant investment in algorithmic and application development, ensuring that the anticipated EPI and quantum computers will be useful for the European Research Area. This must include research funding and support for fundamental new algorithms and quantum information theory, and for robustness, reproducibility, data & I/O, and the convergence with classical computing.

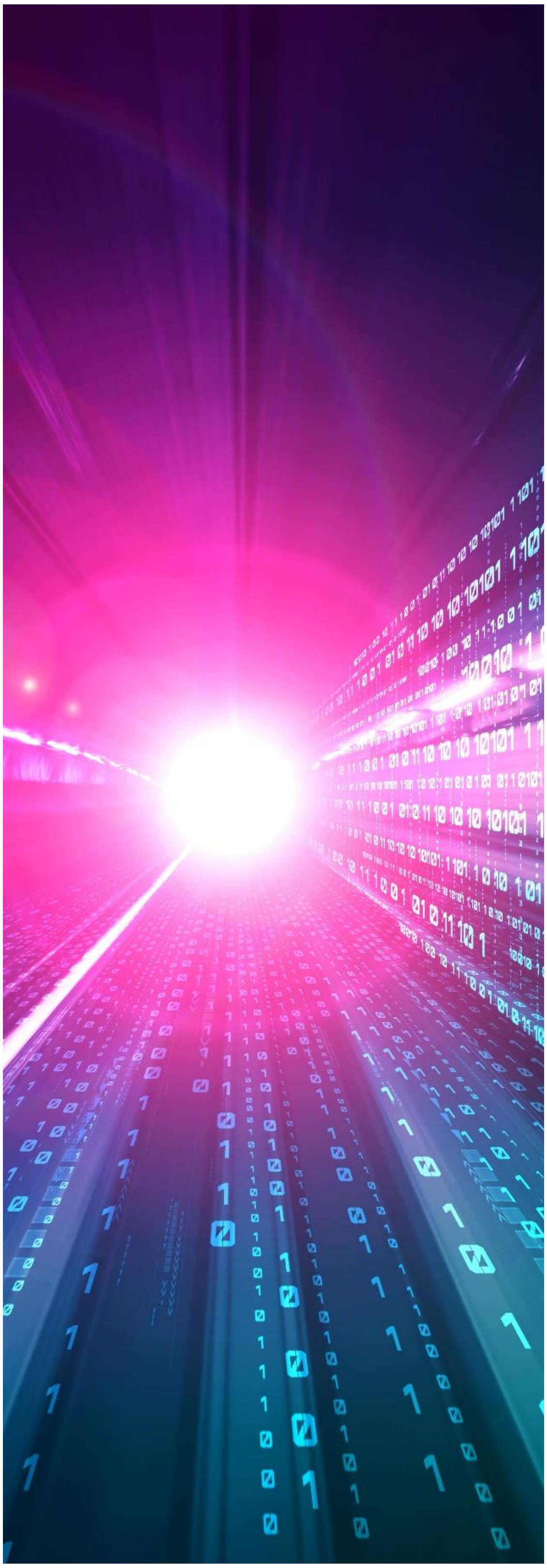
All these approaches require co-design activities involving architectures, OS, communication libraries, workload management and end-user applications to achieve the intended results.

R&I Priorities

- Strengthen the role AI plays in scientific workflows (“AI for Science”)
- Sustained long-term efforts for maintenance, adaptation, and development of scientific applications; adoption of software engineering best practices; adoption of performance portability frameworks, high productivity programming approaches, etc.
- Development of novel methods and algorithms that can work with lower precision which are able to exploit modern hardware better
- Further exploration which applications could profit from disruptive technologies (quantum, neuromorphic)

Post-Exascale Vision

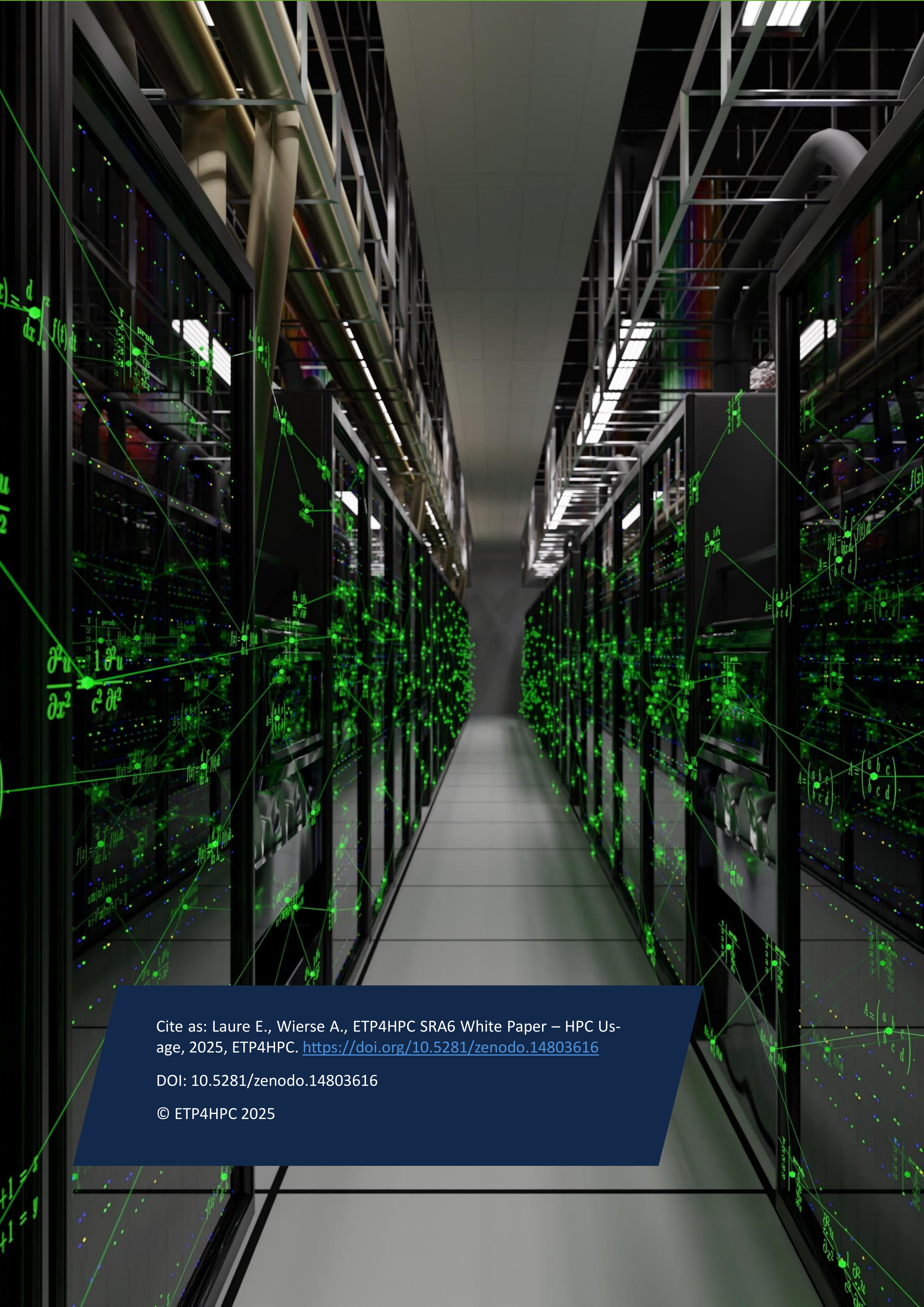
The availability of exascale-class computing has unlocked novel scientific capabilities, paving the way to novel scientific discoveries, which were not possible before. This, often in combination with AI technologies, is changing the scientific workflows in many fields. To fully capitalize on these capabilities, scientific methods need to be adapted and novel methods developed; applications require innovation in productivity, performance, and sustainability; and highly efficient software building blocks (libraries, frameworks, tools) are needed. Close coordination among application, algorithm and software development to address key application development challenges is a must - as well as exploiting synergies across application domains as addressing these challenges independently for each area will not be possible.



Contributing Authors

Erwin Laure is Director of the Max Planck Computing and Data Facility and Professor at the Technical University Munich in Garching, Germany. Prior to this he was Director of the PDC Centre for High Performance Computing at Stockholm, Sweden. He is a long-term member of the ETP4HPC, was a member of the EuroHPC Infrastructure Advisory Group and has over 25 years of practical experiences in HPC.

Since 2011, **Dr. Andreas Wierse** has been Managing Director of SICOS BW GmbH, located in Stuttgart. The company was founded by the University of Stuttgart and the Karlsruhe Institute of Technology (KIT) to support small and medium sized enterprises in the uptake of smart data technology and is financially supported significantly by the ministry for science, research and art of Baden-Württemberg and its shareholders. He is one of the founders of the Smart Data Solution Centre BW (SDSC-BW); the experience he and his project partner Dr. Till Riedel gained there formed the basis for the „Smart Data Analytics" Compendium for entrepreneurs, published by DeGruyter in 2017. Since 2014 he has also been Managing Director of HWW GmbH, a public/private partnership for industrial use of HPC including shareholders such as Porsche, T-Systems and the state of Baden-Württemberg.



Cite as: Laure E., Wierse A., ETP4HPC SRA6 White Paper – HPC Usage, 2025, ETP4HPC. https://doi.org/10.5281/zenodo.14803616

DOI: 10.5281/zenodo.14803616