White Paper

Arnaud Bertrand[a], Jean-Marc Denis[e], Christoph Ebell[d], Hugo Falter[d], Jacques-Charles Lafoucriere[b], Thomas Lippert[c], Jean-Philippe Nominé[b], Estela Suarez[c]

[a] *Atos (ATOS)*
[b] *Commissariat à l'énergie atomique et aux énergies alternatives (CEA)*
[c] *Forschungszentrum Jülich GmbH (FZJ)*
[d] *ParTec AG (PARTEC)*
[e] *Atos until June 2021; now SiPearl*

Executive Summary

As engines for cross-sectoral digital transformations in many scientific, economic and social fields, Exascale and post Exascale supercomputers will be key to advancing automotive, aerospace, chemical, healthcare and energy industries and to strengthening specific European competences in materials science, molecular biology, personalised medicine, neuroscience, AI as well as climate and earth system sciences, in particular. Critical technological building blocks and integration paths are the critical ingredients to realize the objectives of the European High Performance Computing Joint Undertaking (EuroHPC-JU), i.e., the development of a hyper-connected set of Exascale and post Exascale supercomputers for high-end simulation and processing of big data, which can be spearheads and part of a wider vivid HPC ecosystem.

This document describes the high-level strategy -- including concrete key elements and efforts -- needed to realize a vivid ecosystem of European supercomputers, the prospects for a lasting European HPC technology, and the role of pilot platforms in this context. Key assets are the successful Sequana architecture and environment, co-developed by ATOS and CEA, and the proven modular middleware ParaStation Modulo, co-developed by PARTEC and FZJ. A third element is constituted by hardware components developed in Europe, with the EPI processors and the BXI network as major examples. The EPI project was initiated by ATOS, BSC, CEA, and FZJ with many additional partners, including SiPEARL, which will industrialize the EPI general purpose processor. The BXI network initially co-designed by ATOS and CEA is also part of multiple EuroHPC R&I projects. The Modular Supercomputing Architecture (MSA), a truly European approach to heterogeneous computing, combines all these elements putting them at the service of a wide variety of application fields and satisfying their diverse requirements.

1. Introduction

The strategic goals adopted in this design study for extreme and high performance capacity are the development and demonstration of key technologies for the European Exascale and post-Exascale supercomputers. The focus lies on the use and creation of IP in Europe, the construction and operation of powerful computing systems and a scalable software environment, and the stimulation of a vivid, competitive and sustainable European value chain. The vision aims at the highest application efficiency for the benefit of scientific, industrial and societally relevant simulation and data analysis applications. The design promoted integrates important hardware and softsware building blocks of EU-funded FETHPC projects, namely the DEEP[1], and Mont-Blanc[2] project series and many other relevant H2020 measures, then, from 2021 onwards, EuroHPC R&I projects like DEEP-SEA, IO-SEA, RED-SEA projects, EUPEX and HPCQS.

The proposed solution includes ATOS fail-safe and proven Bull Sequana technology (currently BullSequanaX)[3] co-developed within the last 10 years together with CEA[4,5,6], European Processor Technology stemming from EU R&D and design, in particular the plans for a general-purpose European processor that will be produced by SiPearl and a future accelerator, as envisaged in the FPA of the EU-funded EPI project[7], the modular middleware ParaStation Modulo[8], co-developed since 2007 together with the FZJ[9], and the creation of an open and comprehensive Exascale programming and software environment.

On the path to integrating key technologies, prototypes and pilots have been and will be developed in co-design with carefully selected key applications that are among the major challenges of science, industry and society. On the one hand, such applications need highly scalable supercomputing to make progress, and on the other hand they are driving the technical development into the proper direction

The practical use of the Exascale systems is of paramount concern in this design study. HPC must return to the virtues of high application efficiency on the way to Exascale, especially to treat big data-intensive problems by HPC[10], as traditional developments have failed in this respect. Benchmarks relevant to TOP500 and other lists are not in focus, but high application efficiency will certainly lead to first places in such benchmark competitions.[11]

The overall timetable of the EC foresees the first Exascale system in Europe in 2023-25 [12] .

The proposed approach must also pave the way to the future, and it is ready to integrate new disruptive technologies like accelerators or computers based, for example, on quantum or neuromorphic solutions.

[1] DEEP projects: www.deep-projects.eu
[2] Mont-Blanc projects: http://montblanc-project.eu/
[3] Atos. Bull Sequana X. https://atos.net/en/products/high-performance-computing-hpc/bullsequana-x-supercomputer
[4] CEA. Tera. http://www-hpc.cea.fr/en/complexe/tera.htm
[5] Atos. Tera1000. https://atos.net/en/2018/press-release_2018_06_25/atos-cea-place-tera-1000-powerful-european-supercomputer-worlds-top-15
[6] Amiet M, Carribault P, Charon E, Verdière GCd, Deniel P, al. e. Tera100. In Vetter JS. From Petascale toward Exascale.; 2013. p. 46-74
[7] The EPI project: https://www.european-processor-initiative.eu/
[8] ParTec. ParaStation. http://www.par-tec.com
[9] JSC. JUROPA. http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUROPA/JUROPA_node.html
[10] Altintas I. Cetraro, HPC 2018. http://www.hpcc.unical.it/hpc2018/prsnts/altintas.pdf.
[11] This was demonstrated impressively with the JSC JEWELS booster placed #1 in Europe and #7 worldwide (TOP500 2020: https://top500.org/system/179894/)
[12] https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1592
https://ec.europa.eu/digital-single-market/en/news/proposal-council-regulation-establishing-european-high-performance-computing-joint-0

2. Criteria for an Exascale Supercomputer Architecture

Scaling up supercomputer systems by several orders of magnitude poses a multitude of specific technical challenges. On the way from Petascale to Exascale it is understandable that leading, globally established companies build on previous, monolithic concepts that were largely successful up to the Petascale realm. However, one can already observe that the HPC community around the world start to propose computer architectures that combine heterogeneous components in a modular way, in the way that it was proposed originally in Europe with the Modular Supercomputing Architecture (MSA)[13,14,15].

In order to justify this trend, let us summarize here the most important criteria for effective and efficient supercomputing at the Exascale.

Criterion 1 – Modularity. For efficient Exascale Computing, a dynamical assignment of code portions exhibiting different concurrency limits to the appropriate processor types will be essential to operate close to the asymptotic setting of the Generalized Amdahl's Law[16]. It is an mportant requirement for Exascale systems to separately compute low-concurrency code portions and those with high concurrency by means of Modular Supercomputing.

Criterion 2 – **System Reliability, Robustness and Versatility**. Larger overall system sizes, more and more cores, processors, and network components, along with direct water-cooling sub-systems, make a reliable and resilient, fail-safe system architecture the primary important requirement to achieve Exascale. Amongst others, a hierarchical system architecture with the interconnect integrated in the backplane on board- and rack-level is required. Cooling subsystems must be absolutely fail-safe as both footprint and integration density increase. A generic and evolutive platform is also requested to capitalise on best-breed engineering while accommodating the different options of components available at a given moment (processors, interconnects, memories).

Criterion 3 – **Processor Efficiency.** The efficiency of data-intensive applications running on CPUs and accelerators (e.g. GPUs) depends crucially on an appropriate ratio of memory bandwidth to computing capability. Therefore, an important objective is to get processors that are as close as possible to reaching 1 byte of memory performance per floating point operation, while being energy efficient.

Criterion 4 – **Dynamic Evolution**. Post-exascale supercomputers will be used by a very large set of new communities, and supercomputing/data centres will need to adapt quickly to the requirements of new and diverse use-cases. Therefore, future centres need to be able to quickly integrate new technologies, including disruptive approaches like quantum computers and AI specific accelerators. The Modular System Architecture will allow to do exactly this.

[13] Suarez E., Eicker N., Lippert Th., "Modular Supercomputing Architecture: from idea to production", Chapter 9 in Contemporary High Performance Computing: from Petascale toward Exascale, Volume 3, pp 223-251, Ed. Jeffrey S. Vetterm, CRC Press. (2019) [ISBN 9781138487079]
[14] Slide 40 in https://www.r-ccs.riken.jp/R-CCS-Symposium/2019/slides/Wang.pdf
[15] https://www.fz-juelich.de/SharedDocs/Meldungen/IAS/JSC/EN/2020/2020-07-lippert-professur.html?nn=362416
https://aktuelles.uni-frankfurt.de/menschen/physiker-thomas-lippert-als-experte-fuer-supercomputing-an-die-goethe-uni-berufen/
[16] On parallel processing systems: Amdahl's law generalized and some results on optimal design, L. Kleinrock; J.-H. Huang, IEEE Transactions on Software Engineering (Volume: 18, Issue: 5, May 1992)

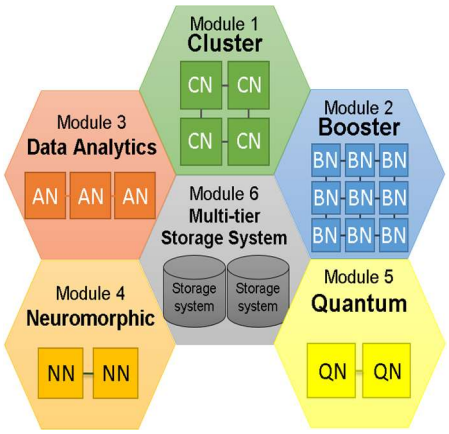
Criterion 5 – Energy Efficiency. Exascale and post-exascale supercomputers must remain within a reasonable energy budget of few tens of MW. This constraint is less technical – in theory we already know how to make a 100 MW supercomputer – than environmental, operational and financial. The goal is to comply with infrastructure capacities, limit the total cost of ownership and overall, contribute the global effort toward a decarbonized world. This efficiency will be achieved by getting more hardware peak performance per Watt, but also by mimimising data movements, improving system management, and optimising the use of resources by applications.

3. Technology solutions

The approach to Exascale in this design study is an integration of a four principal technology pillars[17]. They match the criteria described in the previous section.

3.1 Modular Supercomputing Architecture (MSA)

The need for highest compute performance at the lowest possible energy consumption, leads to computer architectures that combine a large variety of general purpose and acceleration elements. In order to meet the requirements of **Criterion 1 - Modularity**, **Criterion 4 - Dynamic Evolution**, and **Criterion 5 - Energy Efficiency**, system architectures must orchestrate their heterogeneous resources enabling applications to run each of their parts on the best suitable compute elements.



The MSA does it by interconnecting potentially large parallel clusters of CPUs, CPU-accelerator nodes or AI-adapted nodes, whose hardware configuration meets the requirements of a particular type or part of an application, a large storage element, or a future computing system such as a quantum computer (see figure). These “compute modules” are interconnected via a high-speed network and a unified software environment (see section about ParaStation Modulo). This approach gives users full flexibility by allowing them to choose the right mix of computing resources and distribute their code across modules, assigning the resources in a dynamic manner. Such a system is well suited for workflows of different, increasingly complex, and emerging applications for a particular research issue.

3.2 Robust and dynamic generic integration platform

For an Exascale/Post-Exascale-level production system, certain requirements are essential, as required by **Criterion 1 – System Reliability, Robustness and Versatility, Criterion 4 - Dynamic Evolution** and **Criterion 5 - Energy Efficiency.** First and foremost, the system must be fail-safe, adaptable and its system architecture needs to be reusable. In fact, BullSequanaX1000 and its latest evolution XH2000 has proven its perfect availability for high performance production, achieving an availability of >99.98% in production on supercomputers on the Top20 that run business critical Grand Challenge applications.

The BullSequanaX product line was designed from the beginning by ATOS and CEA to meet the requirements of Exascale. The infrastructure is capable of supporting several generations of compute nodes of completely different technologies such as X86, GPU, ARM, FPGA or in the near future the upcoming EPI processors from SiPearl. A wide range of network topology options are possible, including Fat Tree, DragonFly+ and Generalized Hypercube (GHC). With its extreme flexibility, possibly up to dozens of thousands of compute nodes can be managed, which matches current projections for Exascale and Post-Exascale class supercomputers. The BullSequanaX platform is able to integrate any type of HPC network and is optimized for the European BXI network[18] co-designed by ATOS and CEA.

[17] Lippert Th. Cetraro, HPC 2018.http://www.hpcc.unical.it/hpc2018/prsnts/lippert.pdf
[18] See RED-SEA EuroHPC co-funded R&I project.

The BullSequanaX offers highly efficient power and cooling solutions with full direct liquid cooling (DLC) using hot water, resulting in a very high heat dissipation efficiency. It is also one of the few high-end supercomputer technologies with >99% hot water cooling, including all switches and power supply units (PSU) allowing an especially low PUE coming close to 1.01, leading to enormous savings on operating costs.

With the OpenSequana platform, BullSequanaX cabinet specifications will be made available for any hardware developer, so that they can develop they own blades. Their blades will therefore be fully compatible with the BullSequanaX cabinet, the specifications of which are made available and maintained by Atos.

Because quality is mandatory, Atos will also include in OpenSequana the guidelines and processes that hardware developers and manufacturers have to respect when they develop blades for OpenSequana. Their own blades will be qualified and certified by Atos, which will make sure that Supercomputers mixing in OpenSequana cabinets blades from different manufacturers will have the proper level of quality.

On a longer-term perspective, OpenSequana will consider evolving toward the most open standards like Open Compute Project[19] (OCP) and Open Compute Accelerator Module[20] (OAM) which is adopted by more and more accelerator manufacturers. It is an open form factor used for high density accelerators such as GPUs, NPUs, FPGA, etc., and optimized for liquid cooling.

BullSequanaX is therefore the ideal basis for the Modular Supercomputing Architecture (MSA) because it offers a high degree of flexibility, as it allows a large mix of compute elements and technologies, combining them in a seamless hardware and management infrastructure. This is important for modularity as the key to operational Exascale computing. In other words: BullSequanaX is already a modular concept on its own and forms the hardware basis for the European Modularity that we consider in this document.

3.3 European Processor (EPI)

The appropriate ratio of memory bandwidth to computing capability determines the a-priori efficiency of a supercomputing system, as formulated in **Criterion 3 – Processor Efficiency** and **Criterion 5 - Energy Efficiency**.

The European Processor Initiative (EPI), through SiPearl, its industrial hand, strives to create such efficient world-class processors for Exascale machines with architectures concepts that will

- allow end-users to combine openness at the application level thanks to the use of the ARM instruction set,
- offer a reduced operational pressure and enable an enlarged spectrum of applications performing well with no specific programming complexity thanks to a reasonable vector length, typically 256 bits long, and a very high byte per flop ratio
- the possibility to mix with no effort regular HPC models and artificial intelligence technics like machine learning thanks to the very elaborated operations in the vectors, combining classical AVX like operations and BF16 convolutions,

[19] https://www.opencompute.org/
[20] https://www.opencompute.org/blog/new-open-accelerator-infrastructure-oai-sub-project-to-launch-within-the-ocp-server-project

- combine heterogenous memories like HBM and DDR to get the best of both technologies,
- extend the overall openness with an optimal I/O subsystem with more than 6 PCIe Gen5 16X slots allowing to directly connect up to 4 GPUs and two network interfaces, with minimal latency, full bandwidth and advanced remote memory functionalities,
- implement the open Compute Express Link[21][22] (CXL), which will allow the General Purpose Processor and the xPU to share in a coherent way the same memory space. CXL is built on the PCI Express (PCIe) physical and electrical interface.

EPI and SiPearl concentrates on the following lines of development:

A. SiPearl: Produce a high-end CPU based on the ZEUS core technology from ARM. This GP-EPI processor (code name RHEA) aims at a memory- to computing ratio as close as possible to 1 byte/flop. In addition, it addresses energy efficiency by optimizing the operation in low voltage. RHEA will serve as a processor for the universal, general purpose module of the EUPEX Pilot and as a processor for managing additional accelerators (GPUs and beyond) on the nodes of a highly scalable module.
B. EPI: Develop an accelerator platform and technologies based on the RISC-V instruction set. Its approach enables integrating different application specific acceleration innovations. The first version will include a long vector engine and serves as general-purpose vector processor add-on for Exascale machines. When production-ready devices are available in the development roadmap, these can be used in a highly scalable HPC module.
C. EPI and SiPearl: Provide a common platform to merge general-purpose processor and accelerator elements; this seamless integration allows energy efficient computation offloads and data exchange between different hardware units.

The *EUPEX* EuroHPC pilot project plans to integrate EPI GPP from SiPearl in HPC clusters in 2023, while *The European PILOT* EuroHPC project plans to leverage EPI accelerators developments.

3.4 Middleware and software stack

The need for dynamic adaptation to the limits of Amdahl's law as required by **Criterion 1 - Modularity** and **Criterion 4 - Dynamic Evolution** is the main driver for the world's first modular middleware, ParaStation Modulo, designed by PARTEC to operate MSA systems.

The efficient and robust integration, deployment and use of the modules and their components as required by Criteria 2, 3 and 4 more generally call for a consistent assembly of middleware and tools within a solid software stack. In particular, operating system (OS), cluster management software, resource management and scheduling tools, compilers, programming environment, libraries, and performance analysis and optimisation tools, need to be all installed to create a near-production software stack, prepared and validated for its installation in the later European Exascale systems.

A rich directory of best-breed packages and components, together with good practices regarding large-scale processes to build and deploy software stack configurations, should be constituted. The DEEP-SEA and IO-SEA projects, co-funded by EuroHPC, which have started early-2021, are good examples of the kind of integration of European software components and tools that is considered here. Such an

[21] https://en.wikipedia.org/wiki/Compute_Express_Link
[22] https://www.computeexpresslink.org

approach should be open in the sense of offering flexible ways of including and make mutually compliant developments from different sources (academia, industry, open source or not).

Of particular importance, the software concept offers ways to combine general purpose cluster nodes with advanced, autonomous accelerators (so-called Booster nodes). In ParaStation Modulo, applications can be split into less scalable and highly scalable code parts that run separately and simultaneously on nodes of either the Cluster Module or the Highly Scalable Booster Module. The Global MPI bridges seamlessly between different modules and, at the same time, offers high-speed connectivity by using different communication technologies while abstracting the details of communication between any two peers, whether they are pure Cluster, pure Booster or mixed node pairs. In addition, the parallel process environment is highly integrated into the batch system and support for Slurm's job steps and job packs within ParaStation Modulo enables planning and orchestration of heterogeneous workloads transparently across different modules.

To scale up to Exascale, communication libraries must not only be extremely scalable but also quite robust in view of the vast amount of cores and processes cooperating. ParaStation Modulo features both, and in particular, its robustness is to be highlighted here: So-called Connection Guards constantly monitor all activated communication channels between the processes during runtime. In addition, a proper cleanup is guaranteed in the event of a connection failure. In situations with connected MPI sessions, ParaStation Modulo can handle parent and child groups independently, so that higher-level checkpoint and restart features can kick in and provide resiliency features to the applications.

Another important point is the integration of the supercomputer in the global network and the service interface provided to the user. Previously the users/developers had to comply with each of the computing centres environment. As many new communities do not have the HPC history and knowledge of numerical simulation communities (like physicists or climate scientists), this time is over. Computing centres now have to adapt themselves to wider communities' practices by using virtualization technologies, and by providing cloud-like interfaces, offering more advanced and flexible services to the users. More global compute/storage resource management is also required: supercomputing centres have to be hyper-connected to allow fast and reliable data and job transfers across all Europe. The Fenix federation of top European computing centres is the germ of this new organisation of high performance resources. HPCQS, another EuroHPC pilot project, will also be a pioneering demonstration of such principles, with the first cross-European association of Quantum Computing with HPC in an hybrid platform.

4. Integration paths

Until 2020, the technologies introduced in Section 3 had been developed quite independently from each other through various initiatives, with the support of institutional, national, and EU-funding. This leads to different levels of maturity for the various elements, with Sequana, ParaStation Modulo, and the Modular Supercomputing Architecture already deployed in various production systems, while the SiPearl's processors based on EPI processor technologies are still in development phase.

In order to fulfil the Criteria presented in Section 2 all these technologies need to be jointly deployed in the upcoming European Exascale platforms. Therefore, it is mandatory to previously bring them together and demonstrate their interoperability at pre-Exascale level. Pilot platforms are crucial vehicles to achieve these goals.

Any intermediate integration or pilot approach should aim at:

- Integrating and development of processor technologies into medium-size platforms, leveraging modular approaches (with general purpose, accelerated, quantum-hybridized parts) and demonstrating their scalability potential;
- Enabling Software Stack, leveraging such pilot hardware platforms for software development and integration. In particular, operating system (OS), cluster management software, resource management and scheduling, compilers, programming environment, libraries, and performance analysis tools, need to be all installed to create a near-production software stack, prepared and validated for its installation in the later European Exascale systems. The pilot projects also have to support the longer-term development of compiler technologies and low-level software elements for future processors beyond EPI first generations;
- Co-designing, Adapting and Optimizing Applications: the final validation of the hardware and software elements of the Pilot projects has to be certified with the performance and energy-consumption metrics of a set of key applications – a small number selected based on their importance and impact on the European Scientific community - while more widely collaborating with scientific and industrial user communities via CoEs and other European application-focused projects for them to prepare their codes for the upcoming Exascale platforms.

Starting from 2021, EUPEX, a EuroHPC project, is an implementation of this pilot concept.

The modularity concept also offers a nice approach to progressive integration and evolution of high end computing systems:

- New technologies can be added as "modules" to a production system as soon as they reach sufficient readiness
- This is also a globally more cost effective approach, since the newest and most efficient technologies can be deployed at any relevant moment without requesting global system updates

5. Strategy and Roadmap 2022-2027

Given the demonstrable success of the MSA approach, the strategy of the EA-MHPC consists of a two-pronged approach:

1. Expand and enable the European ecosystem around the MSA. This includes a rich software and technology stack built and co-designed with key application partners and communities. EA-MHPC is committed to working with a wide range of European partners, on both the push and the pull ends of the innovation system.
2. Provide continued technology leadership based on the MSA. This includes prototypes and pilots, and enabling hard- and software for future seamless Quantum Accelerator integration, Data Center integration, hybridization and a focus on security and resilience.

Post-Exascale requirements are already on the radar while the current focus is firmly on delivering reliable, energy-efficient, and accessible Exascale performance. Our members' strategic research agenda integrates and contributes to the European HPC roadmaps, set by EuroHPC, industry organizations, and individual states. EA-MHPC publishes its strategy and technology roadmap and adds regular updates on its website and other channels.

6. Copyright/Licence