©SHUTTERSTOCK.COM/MICHAEL TRAITOV

Predicting the Future of Supercomputing

Scott Atchley, Oak Ridge National Laboratory
Rosa M. Badia, Barcelona Supercomputing Center
Bronis R. de Supinski, Lawrence Livermore National Laboratory
Joshua Fryman, Intel
Dieter Kranzlmüller, Leibniz Supercomputing Centre
Srilatha Manne, Advanced Micro Devices, Inc.
Pekka Manninen, IT Center for Science
Satoshi Matsuoka, RIKEN
Dejan Milojicic, Hewlett Packard Labs
Galen Shipman, Los Alamos National Laboratory
Eric Van Hensbergen, Arm
Robert W. Wisniewski, Hewlett Packard Enterprise

The need to solve high-complexity problems using large-scale tightly coupled computing (that is, supercomputing) continues to grow. In this article, we address the needs, challenges, and opportunities for supercomputing over the next decade.

Digital Object Identifier 10.1109/MC.2025.3567232
Date of current version: 27 June 2025

Supercomputing, which involves the use of the highest-performance computing resources available at a given time, has recently seen broader adoption as it is essential for training generative artificial intelligence/machine learning (AI/ML) models. These AI use cases are in addition to the traditional modeling and simulation (modsim) workloads that continue to drive high use at traditional supercomputing centers.

Supercomputing centers are increasingly adopting AI/ML techniques into modsim workloads. This article by leaders from those centers, as well as within the industry, explores the trends and directions that will shape future supercomputers, driven largely by that convergence of modsim and AI/ML techniques. This article extends the predictions of several recent articles that explored the future of supercomputing.[1,2,3,4,5,6]

EDITOR DEJAN MILOJICIC
Hewlett Packard Labs;
dejan.milojicic@hpe.com

INCREASING USES, INCREASING ADOPTION

As we consider the future of supercomputing, we see several factors that will drive changes to the workloads that are run on supercomputers. These changes will continue to broaden the adoption of supercomputing and will affect the technology used to build supercomputers. In this section and the following one, we describe our expectations for future supercomputing workloads and discuss the technologies that will shape their evolution.

While we expect supercomputing workloads to be augmented with new workloads (for example, AI), we expect that traditional supercomputing workloads will remain a significant use case. These traditional workloads serve a wide range of purposes, from advancing science to deepening our understanding of the universe in which we live, addressing humanity's needs in the modern world, to protecting the national interests of governments that deploy such systems. Nonetheless, we expect these traditional workloads to incorporate new algorithmic techniques, starting with the use of AI/ML models, as has already begun.[7,8,9] The adoption of AI/ML techniques includes their use to guide the simulated configurations in ensemble calculations but also their use to accelerate expensive calculations of models of physics and biological phenomena.

With the end of Dennard scaling and the slowing of Moore's law, the automatic increase in performance at constant cost and power is over. Modsim practitioners are faced with modest gains in performance with incremental architecture changes. Future gains are largely coming from the increase in silicon within the package. While providing needed performance boosts, it comes with higher power and higher costs for both the additional silicon and the integration to stitch together several chiplets. When viewed as performance per watt (for example, if a facility has a fixed power budget), then the gains are still modest.

At the same time, the explosive growth in AI, both training and inference, has driven silicon vendors to tailor their products to this lucrative market. It is not clear, however, that modsim can take advantage of lower precision. Some apps will be able to use FP32 for some of their data structures (but not necessarily all) and see benefits compared to lazily promoting everything to FP64. It is not clear if apps will be able to use FP16 for modsim unless it is using AI inferencing in lieu of a component in a multiphysics application, emulation, or iterative refinement. To use ML inferencing, there needs to be an already-trained model. There is a lot of research interest in determining when/if modsim applications can exploit lower precision, which is becoming much more plentiful. There are efforts to see which, if any, apps can use lower precision directly, use lower precision via AI methods, use lower precision via iterative refinement, or use lower precision via emulation. Some apps may be able to do so, while others will not.

The beauty of the General Matrix-Matrix Multiplication (GEMM) emulation methods (that is, Ozaki methods[10]) is that precision is finer-grained than with hardware. Hardware is limited to powers of two (for example, FP64, FP32, and FP16), while Ozaki can provide any multiple of four bits (for example, FP40, FP48, and FP56) to provide just enough precision to converge on a valid solution without providing "too much." While Ozaki's scheme can outperform native cuBLAS in some cases, the downsides to emulation are 1) it can only emulate GEMM (that is, matrix-matrix) instructions but not vector instructions, and 2) it consumes 30–50% of the available memory, thus reducing the solvable problem size. If memory were cheap and plentiful, the latter would not be an issue, but supercomputer users want the fastest memory available. Today, that is high-bandwidth memory, and it is neither cheap nor plentiful. Recently, systems used for AI/ML training have been cast as competitors to supercomputers.

Rather than competitors, the authors view both modsim and AI as having overlapping needs for supercomputer design, except for precision. However, the systems that provide AI/ML capability are best viewed as supercomputers themselves and reflect that AI/ML training has emerged as an important workload for supercomputers. As we look toward the future, not only do we expect that AI/ML training will remain a critical supercomputing workload, but we anticipate that additional new workloads will emerge. We expect that domains that have begun to use supercomputers more extensively due to the success of large-scale AI/ML models, such as finance and retail, will identify new mechanisms to exploit the computational capability available and expand the use of AI/ML in their domain.

The convergence of cloud computing and supercomputing has long been expected. However, this convergence has not fully materialized yet, in part due to the requirements of traditional tightly coupled parallel modsim workloads. Nonetheless, cloud providers continue providing more high performance computing (HPC) capability, and cloud computing continues to be a viable economic and technical alternative for embarrassingly parallel workloads and, as of recently, for AI/ML workloads. They are also suitable for offloading or bursting small-scale experiments and development.

Addressing humanity's needs, such as weather forecasting and biomedical research, continues to be an important target of supercomputing. These applications include energy needs and its production using nuclear fission near term and fusion long term—but also for fossil fuels and importantly, carbon and water management. Another use is for new materials, particularly for the continued advancement of technology beyond silicon CMOS device scaling. Yet another use case is mitigating and adapting to climate change, including utilizing digital twins.

Digital twins are virtual representations of physical artifacts, systems, or processes with collected real-time information. They enable monitoring, simulation, and prediction of those physical artifacts. Digital twins often use supercomputers directly in a variety of vertical applications and services (for example, for structural analysis, Earth monitoring, manufacturing, and operations) as well as exploit them peripherally (for example, for monitoring, optimizing operation, anomaly detection, or what-if-analysis). Digital twins are used in areas such as the transportation industry, data centers,[11,12] and even Earth.[13]

Another important use case of traditional supercomputing is helping drive new scientific breakthroughs [that is, helping answer the big questions, for example, performing computation for the follow-on to Laser Interferometer, Gravitational-Wave Observatory (LIGO) or Laser Interferometer Space Antenna (LISA) that will enable sensing of gravitational wavelengths populated by a rich diversity in astrophysical phenomena that are of deep interest to astronomers and astrophysicists]. After a discussion on how the use and adoption of supercomputing evolved, we will next explore how technology evolution impacts workloads.

EVOLVING TECHNOLOGIES AND WORKLOADS

Future supercomputing workloads will reflect recent and anticipated future technological and industry developments. These trends include not only the adoption of AI/ML to serve edge computing and other end-user applications but also productivity enhancements, such as those driving broad consumer adoption of cloud-based computing. Further, architectural and device-level advancements will continue to motivate new supercomputing application enhancements. This section provides a high-level description of these two influences on future supercomputers. We begin by describing the workloads.

- New applications are continuing to demand more computational capability, including bioengineering, climate modeling, national security, fusion energy, and many others.
- HPC and AI will continue to converge and thereby demand more AI-ready infrastructure.
- Large language models and other models have captured the public imagination, and they open new opportunities in supercomputing.
- Physics-informed neural networks and other models, possibly integrated into traditional mod-sim applications, enable the faster exploration of design spaces.
- Some workloads are increasing performance by leveraging mixed precision computation, while others are leveraging multitenancy to increase performance.
- Application demand for scale-up networking, including Ultra Accelerator Link (UALink), will continue to increase per-device bandwidth and the number of directly connected scale-up devices, blurring the boundary between scale-up and scale-out infrastructure.

In the last couple of years, advancements in AI, specifically in generative AI applications, have dramatically influenced private industry toward building large-scale computing infrastructure. Even though these infrastructures are driven by AI requirements, they are becoming increasingly HPC ready. AI and HPC are making significant strides toward convergence, and this development is a major disruptor. We predict the forthcoming technological changes.

- Accelerators, from traditional (for example, compression and crypto) to ones focused on AI (for example, Cerebras, NextSilicon, and SambaNova) to upcoming (for example, neuromorphic and quantum), will address specialized but important demands, and some are already being incorporated into existing supercomputers. 2.5D and 3D memories present obstacles that must be overcome to use, but they provide significant opportunities to help ameliorate memory wall challenges.
- Continued evolution of the scale and latency-sensitive industry-standard or standard-compatible/interoperable interconnects (for example, scale-up merging with scale-out) will occur.
- Increasingly integrated photonics as a means of power reduction, packaging simplicity, and bandwidth enhancement will be seen.
- Improvements in reliability are driven by the need to address resilience (or fault tolerance) at all levels of the system, from hardware to system software to applications.

These technology changes will result in a new macro-political landscape that may influence decisions on next-generation supercomputer procurement. For example

- AI will drive technology directions/priorities, including reduced precision, systolics, and fixed function units.

- Silicon transistor devices are approaching hard limits in scaling, with limited improvements in performance through silicon CMOS scaling, which has implications for specialization, tight integration, and power reduction. These limits introduce a need for deeper co-design alongside other major market forces, such as AI.
- New computing technologies are being explored, including quantum, neuromorphic, and other accelerators that may substantially change the landscape in terms of scaling, reliability, power, and cooling.
- Research in new nonvolatile memories (NVMs) has been occurring for many years. If that work leads to successful productization, it may affect the way we design storage, conduct checkpointing, and in general, manage memory.
- New algorithms (potentially AI inspired and enabled by new accelerators) can also impact performance and scale.

ARCHITECTURE

Two main architectural changes have brought AI and HPC applications closer together. The first is the addition of high-performance GPUs alongside high-performance CPUs for compute, and the second is AI's need for fast and efficient communication within and between compute elements.

One of the biggest shifts in the last decade has been the widespread adoption of GPUs for computation. While accelerated by AI use cases on supercomputers, this trend was occurring independently on HPC systems due to the need for higher compute capabilities while keeping power manageable. Similar motivators (that is, raw performance, performance per watt, performance per area, and performance per dollar) will likely drive the inclusion of accelerator technology (for example, Cerebras, NextSilicon, SambaNova, and potentially quantum or neuromorphic), though the intercept of the latter two's productive use will require additional time.

As GPUs became dominant, the primary architecture of the system remained homogeneous by node. That is, while each node was heterogeneous (microheterogeneity), the overall system was homogeneous. Many of these new accelerators are not as general purpose as GPUs, and therefore, systems are likely to be macro-heterogeneous. What remains open is the tightness of coupling of these macro-heterogeneous partitions.

The severity of the memory bottleneck in generative AI has led to other forms of acceleration reentering consideration, including computation near memory (CNM) as well as processing in memory (PIM). These computational accelerators, coupled with collective acceleration in the network, data processing units (DPUs), and forms of compute near storage, create a more diverse acceleration landscape than that enabled by GPUs. Further, as chiplet-based design points lead to finer-grained customization, the opportunity to intermingle compute acceleration with general purpose compute may become attractive to better balance system performance, power delivery, and thermal dissipation.

A major block to heterogeneity, whether it be at the micro or macro level, is the programming model. Without a productive programming model that enables efficient offload to accelerators, the additional hardware will not provide a good return on area, cost, or power investment. The transition from CPUs to GPUs was made easier via a programming model and tool stack for GPUs, and any accelerators will have to match those capabilities to be viable. For example, circuits for CNM have been known for more than 50 years,[14] but the general programmability problem remains unsolved and generally avoided as "too hard" to solve.

One of the significant challenges in the post-exascale era is communication. This challenge involves moving data from memory to compute and between compute. One way to help address this challenge is to move to more tightly coupled architectures. Memory stacking, 2.5D or 3D, has the potential to reduce power and increase bandwidth between compute and memory.

An important aspect of heterogeneous node architectures is moving data between compute elements, specifically between the main CPU and the accelerator. Coarser parallelism leads to less frequent data movement and more efficient use of the accelerator. Traditional HPC applications need serial cores, and many large AI applications are also increasingly benefiting from the utilization of CPUs. Further, many HPC applications remain bulk synchronous with branchy and data-dependent code between parallelizable kernels; that code runs better on CPUs. The AMD Instinct MI300A accelerated processing unit (APU) brings the CPU and the accelerator computing elements together both physically, via chiplets, and programmatically through a unified memory model; Nvidia's Grace-Hopper provides similar benefits using a full reticle CPU and GPU interconnected through NVLink—a chip-to-chip technology. However, hardware and software challenges, such as software offload launch latencies, remain. Tighter coupling may further help improve performance. For example, 3D stacking would allow more memory bandwidth than 2/2.5D integration.

Moving across compute within the same package or same node offers challenges, but significant performance cliffs occur when moving from high-performance nodes to the network due to lower network byte/flop ratios, high network latencies, and high costs of synchronization across nodes. These inefficiencies require application developers to partition their codes in a coarse-grained manner into serial and parallel compute phases, memory movement phases, and network communicator phases with each one optimized independently. This requirement not only impacts programmer productivity but misses opportunities to optimize power efficiency and memory access across the system. These network inefficiencies also limit strong scaling. The bandwidth and latency cliffs are not the only inhibitors of performance. The model of how memory is accessed can also have a large and potentially greater impact on the performance of applications when they communicate outside the node. The right internode memory model with enhanced capabilities, such as atomics and load/store access to memory within a supernode, pod, or hypernode (collections of tightly coupled nodes with an enhanced memory model), can improve strong-scaled performance by more than an order of magnitude. Nvidia's NVLink and the UALink standard (which AMD is a part of) are specific solutions that can provide tighter coupling between nodes. The general UALink industry-standard effort is moving to create an interoperable fabric for these needs. Competing pressures on interconnects will likely move future interconnects from low radix high diameter to high radix low diameter to improve efficiencies across a wide spectrum of use cases.

Two decades ago, the connection model was flat. A core comprised a node, and each node had a network connection. The topology varied (for example, butterfly, hypercube, or torus), but all compute elements were uniformly separated. With the introduction of multiple cores per chip, multiple chips within a node, and multiple GPUs within a node, two levels of connectivity, inter- and intranode, were introduced. This architecture provided a communication latency and bandwidth advantage between these computing elements that were contained within a node. However, the architecture came at a cost. Applications—and particularly communication runtimes—needed to be aware of the topological structure to exploit it.

The future will determine if ESS leads to a common stack across the community or splinters the community.

Motivated by AI, scale-up networking is creating another layer in the communication hierarchy. Pods, super nodes, wafer scale, or hypernodes, represent an opportunity to connect tens to hundreds (perhaps small thousands) of nodes in a more tightly coupled manner with memory semantics (for example, load/store access and atomic operations). These architectures have better performance for AI and strong-scaled applications but also introduce a programmability cost. Again, the software layers have an opportunity and responsibility to attune the application appropriately for the communication hierarchy.

An open question remains as to the best overall system architecture since this intermediate communication layer (that is, scale-up: between or within a node and across the whole machine) is more expensive from a cost and power perspective than a flat communication architecture. One possibility that shows promise is merging the connectivity emanating from a node into either scale-up or scale-out connectivity. While this approach is a promising notion, no obvious technologies enable it, yet, but the two main standards initiatives in this space, UALink and the Ultra-Ethernet Consortium (UEC), are currently working on it.

Traditionally, the HPC community relied on large-scale hard-drive-based parallel file systems, such as Lustre and GPFS. In recent times, object store file systems optimized for NVM technology, such as DAOS, VAST, and Weka, are gaining popularity and will increase, including the model stores for AI, such as vector databases. Cloud services have innovated object interfaces, such as S3, that AI frameworks use natively.

FACILITIES

Energy has been driving exascale supercomputing as one of the primary constraints. From the beginning of exascale planning, the desire to keep the spending on power to a minimum led to a target of 20 MW.[15,16] This impacted system designs, specifically cooling, space (the number of racks), and the CPU/GPU ratio. Air cooling was not sufficient, and liquid cooling has become the standard solution for capability-class supercomputers and is seeing broader-based adoption.

Figure 1 notionally presents the evolution of power efficiency on the left (red curve) versus cooling choices on the right (blue curve) during the past few decades. Power efficiency numbers were taken from Oak Ridge National Laboratory supercomputers (Jaguar, Titan, Summit, Frontier). Due to 3D chips, the power density will continue to increase (more than double from 2021 to 2031) according to the IRDS Roadmap,[17] which will require further innovation in cooling, such as immersive or evaporative spray cooling.

In the longer term, both cooling and power requirements may change substantially. Multiple reasons led to the 20-MW limit in the requirements for exascale supercomputers, including cost and the ability to deliver that much power. The new means of energy production, such as small modular reactors (SMRs), are competitively priced per MW and complemented by onsite renewable energy production (for example, wind and solar). If they succeed, they will address both the cost and power delivery to data centers.[18] The AI compute demand and the boom have further shifted the economics and scale of power generation, altering availability and pricing.

SOFTWARE STACK

The system software stack, as defined by everything below an application and above the hardware, continues to increase in complexity. From a modeling and simulation perspective, as the desired capability has increased, system implementers have increasingly turned toward leveraging open source to provide this capability. This change complicates comprehensive testing. The combinations of open source components exponentially increase the number of possible permutations of the software stack. Insufficient connectivity between these open communities (and interest in being connected) has made comprehensive validation significantly more challenging than when a vendor owned all, or most of, the components in a stack.

OpenHPC created a complete and comprehensive system general software stack. Extreme-scale Scientific Software Stack (E4S) of the Exascale Computing Project (ECP) made strides toward unifying the development environment across many open source components. The High Performance Software Foundation (HPSF), unified by Spack, is making strides toward providing optimized software stacks for well-defined systems. Nonetheless, challenges remain, and a stronger community testing effort, perhaps under HPSF, is still needed.

The inclusion of AI software stacks on supercomputers has significantly increased the number of components of the overall software stack. More importantly, AI infrastructure, including the software stack, is undergoing rapid change. The key contributors are investing significant effort to support this rapidly evolving environment while other organizations are challenged to keep up. Overall, the rapid evolution limits the organizations that can stand up and maintain an AI stack, which further increases the need for community efforts toward testing and maintaining the overall software stack.

While E4S was United States centered, Europe is developing the European Software Stack (ESS). The EuroHPC JU will work with stakeholders to coordinate co-design in the research and investigation of hardware and software activities and ensure that those activities meet user requirements and that developed technologies are deployed. Funding is planned for the different building blocks in HPC, AI, and quantum computing (QC) from innovation to deployment, targeting different technical readiness levels as required by the status of hardware developments. Europe will focus on multiple aspects, such as performance and efficiency, AI-software integration, energy consumption, workflow managers, and support to European processors, among others. The future will determine if ESS leads to a common stack across the community or splinters the community.

As discussed previously, macro-heterogeneity is on the horizon; enhancements of the software will be needed to incorporate the new elements into the system as well as to support macro-heterogeneity generally. To make these accelerators productive, a comprehensive software stack will need to be developed to enable nonexpert application developers. User interfaces, libraries, debuggers, validation tools, high-level programming models, and languages are needed as well as compilers to translate high-level languages to be distributed over coarse-grain reconfigurable architectures or to QC circuits and transpilers that adapt already-compiled circuits to a dedicated technology.

As the software stack becomes more complex and the overall user code moves from a single executable to a complex set of interconnected executables, we will need an overarching workflow infrastructure. Some examples of workflow management exist today, but those capabilities will need to be enhanced to cover the great variety of emerging software stacks. They will also require many new capabilities, such as the control of data movement

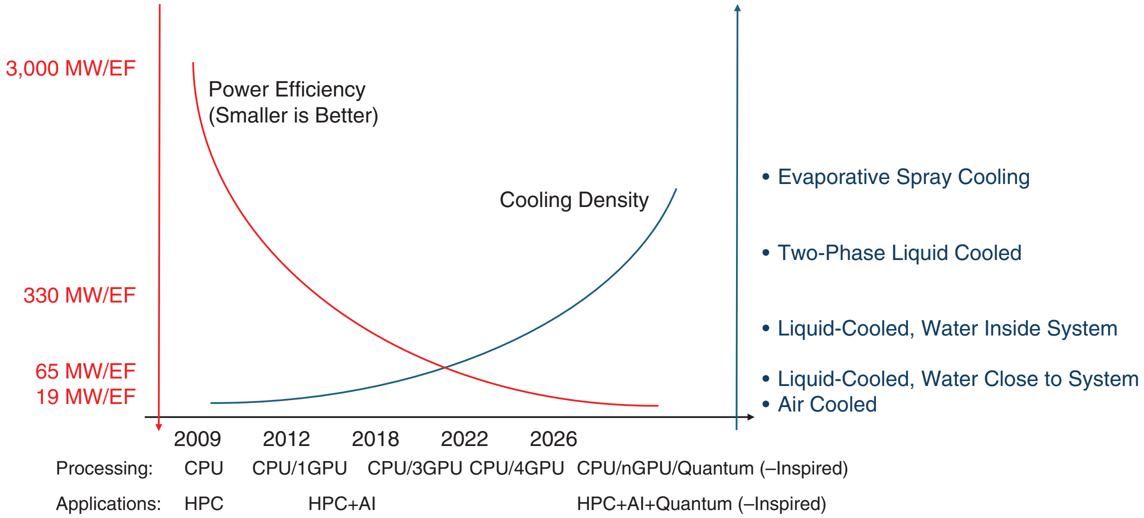

FIGURE 1. Supercomputer power efficiency and cooling over the years.

and enhanced authentication, security, and monitoring.

The amount of power consumed by supercomputers is reaching an inflection point where the cost of electricity throughout the life of the system is approaching its capital cost. New software capabilities must be created to enable users to understand and optimize the tradeoff between performance and energy (for example, to allow a user or system administrator to reduce performance by 10% to save 40% on energy). We will also need support to ramp up and down power more smoothly to meet the requirements of electricity providers.

OPERATIONS

The U.S. ECP was a multibillion-dollar effort, with multiple hundred-million-dollar procurements. In addition, the cost to operate an exascale supercomputer is on the order of 100 million U.S. dollars, a significant part of its total cost of ownership.

Producing and procuring a capability-class supercomputer is a complex operation that is not optimal for the participants in the procurement: regulators, users, integrators, and suppliers. Distributed spending with incremental upgrades could be beneficial. Similarly, the operating expense costs are becoming too high to be financially sustainable. New means of producing and delivering supercomputers could prove beneficial for multiple parties.

Current supercomputers are designed to run applications at an extreme scale. While needed for capability-class applications, this model has challenges for maintenance and partial system refreshes. Accelerator road maps are also more frequent and shorter than the lifetime of supercomputers, which makes refreshes more desirable than in the past, from both the performance and power/cost perspective.

NONFUNCTIONAL REQUIREMENTS

Reliability has long been a focus of traditional HPC, extending from high-level software to ensure that it did not have any single points of failure, down to the silicon, including both compute and memory. This focus was needed as the high-level fault tolerance model in applications was that if one node failed, the entire application failed. Thus, as the machine grew in node count, it was imperative that reliability was improved. Nevertheless, the mean time between failure on the largest supercomputers has dropped from around a week on emergent petascale systems to a handful of hours on emergent exascale systems. With each generation, new points of hardware and software reliability failures emerge due to ever increasing hardware complexity and software not planning for significant implications of heterogeneous architecture implementations.

Innovations in checkpointing architecture in conjunction with improved bandwidth for checkpoints have predominantly ameliorated the impact that this decreased reliability has on system availability. However, unless something changes, this trend will be unsustainable for the next three orders of magnitude of system performance improvement. Fewer applications can productively employ a full exaflop of compute than the number that could employ a full petaflop. This potentially implies a different usage model for supercomputers in the next decade. Each facility's workload will determine whether petascale or exascale resources (for example, compute, memory capacity, and memory bandwidth) are needed.

AI has only recently been run at large scales. Thus, GPUs have not focused as much on reliability as CPUs that were designed for supercomputers. The AI software stack has also not had years of focus on reliability and ensuring no single point of failure. Recent data from Meta,[19] Alibaba, Google,[20] and others show the consequences. As AI continues to scale and systems become larger with the desire to run capability-class applications, an increased focus on fault tolerance will be needed, both in designing and implementing more reliable hardware and in changing the application fault tolerance model.

AI applications are inherently more resilient to failures because of the nature of their computation. While academic work has explored application-level fault tolerance for modsim applications, it has not been implemented in practice as most of the work could address only specific computational kernels rather than the resilience of the entire application. In one form or another, reliability will need more focus moving forward.

SUMMARY AND OUTLOOK

In this article, we presented our predictions of the future of supercomputing. We first discussed increased use and adoption, followed by evolving technologies and workloads. We then presented the architecture, facilities, software stack, operation, and nonfunctional requirements. We concluded with some recommendations to critical actors in supercomputing.

Figure 2 and Table 1 summarize our predictions. Figure 2 describes the architecture of future supercomputing, emphasizing the innovations required. Table 1 succinctly presents the evolution of HPC over decades, from traditional to future supercomputing.

Achieving the next level of scale will require innovation, just like it did to get from petascale to exascale. This innovation will likely need to come across the whole system, including new accelerators, interconnects, system software, application and algorithmic innovations, and power and cooling. Some of the scaling may be possible to achieve by leveraging macro-heterogeneity, for example, through the use of AI-specific, quantum or quantum-inspired, or other accelerators in the context of a more traditional GPU-based supercomputer.

Supercomputers will also benefit from the growth in the bandwidth of interconnects. Photonics could help overcome limited processor shoreline performance, power, and packaging. However, additional investments will have to be made to avoid congestion at scale and to address both jitter and tail latency.

In terms of power and cooling, the current limitations will remain and will have to be addressed with onsite power generation, possibly with SMRs and renewable energy sources as complements to grid supplies. Cooling will require new techniques, as discussed in the "Facilities" section. Locating data centers in zones where power is cheap and reliable can also help. Areas with abundant water and favorable climates will assist with cooling challenges.

Sustainability is challenging in supercomputing due to the extreme use of power. Some of the approaches of large-scale enterprise data centers can be applied (for example, following the sun or server consolidation) to a limited extent. Sustainability awareness can help, as can using digital twin techniques to conduct what-if-analyses and understand where opportunities lie.

The use of AI is inherently tied to ethics and is an important topic that will need to be addressed given the widespread use of AI. AI is effective at improving productivity in software development. Productivity in developing supercomputing applications is critical but also hard to automate using AI due to the performance and scale requirements.

RECOMMENDATIONS

We make recommendations to key actors in the supercomputing ecosystem: supercomputer centers, developers, scientists/users, and industry.

Our recommendations for supercomputer centers are as follows:

- Workloads of the future will continue to have demands for tightly coupled, highly parallel, and noise-free infrastructure at scale. Therefore, the growth in the needed capabilities of future supercomputers will continue, and centers should continue to plan to procure them.
- Future supercomputers may be supplemented by leveraging offload to a public or private cloud or large AI infrastructures for training or services that enhance productivity. Centers should investigate how to incorporate complex workflow capabilities that allow this interaction as well as intrafacility and interfacility workflows. Infrequent delivery of single large supercomputers puts a strain on providers, users, and maintainers of supercomputers. An alternative incremental delivery should be explored to ensure smooth delivery and secure a more reliable

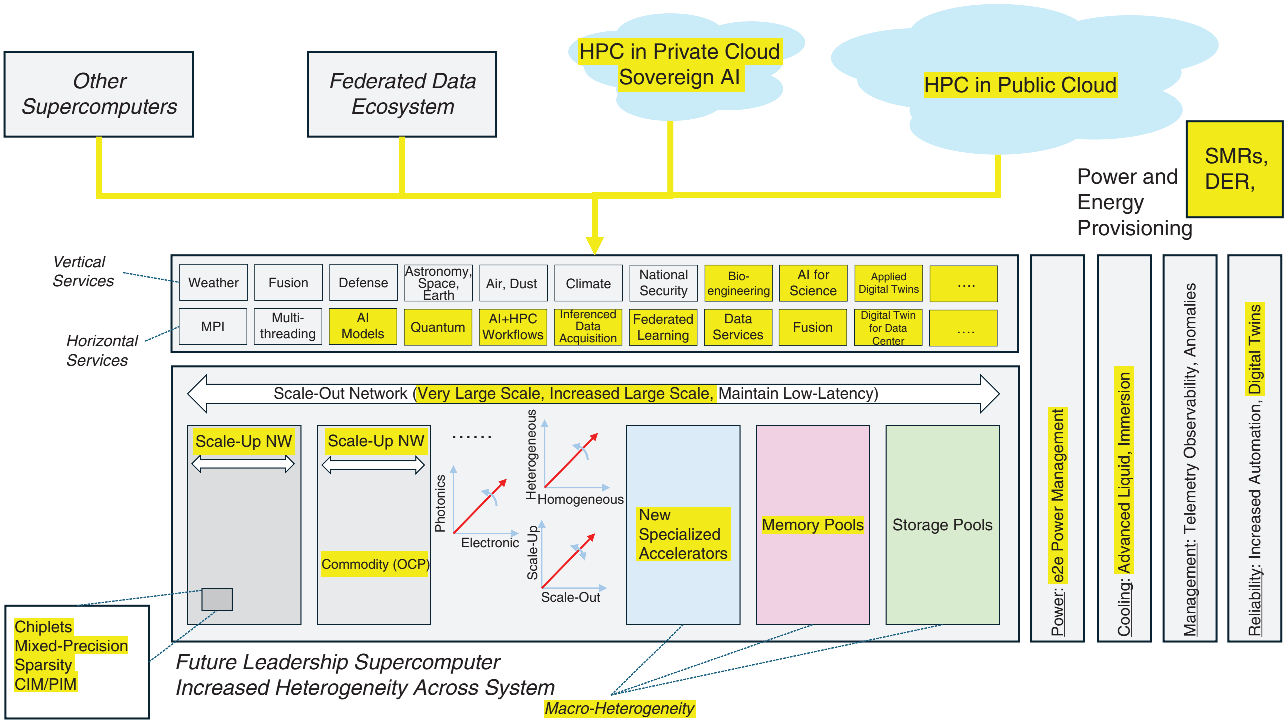

FIGURE 2. High-level supercomputing architecture. Highlighted text in yellow represents new features compared to existing supercomputers. API: application programming interface; SMRs: small modular reactors; DER: distributed energy resources; CIM: computing in memory; OCP: open compute; e2e: end-to-end.

introduction of new features. It also puts HPC at a disadvantage from a performance standpoint. GPU performance is still scaling rapidly, and AI is forcing an acceleration in hardware innovation from compute to networks.

Our recommendations for developers and the open source community are as follows:

- Most of the system software running on supercomputers is becoming open source. The community should become more strategic about planning and delivering new features and secure approaches and infrastructures to be able to develop and test solutions at scale.
- To allow the broadest productive use of software, instilling good software engineering practices into community code will be

TABLE 1. Comparing approaches to building and consuming leadership supercomputing systems.

		Supercomputer eras						
	Comparison criteria	Traditional HPC supercomputer (1990 to present)[*]	Grid (2000–2010)	Cloud (2006 to present)	AI cluster training	AI cluster-inference	HPC cloud	Future HPC supercomputer
How the system is built	Coupling	Very Tight scale-out	Tight scale-out federated	Loose	Tight (scale-up and scale-out)	Loose + scale-up	Loose, scale-up, medium-tight	Configurable
	Scale	<10× exascale	Multisite (federated?)	Multiregions	Collocated	Distributed	Multiregions	>100× Exascale
	Reliability	Job-based restarts	Job-based restarts	Cloud like[†]	Job-based restarts	Cloud like[†]	Cloud like	Job restarts + cloud like[*]
	Elasticity	No	Desired	By design	Moderate	Cloud like	Cloud like	Generally desired, essential for broader workflows
	Storage System	Parallel FS (write intensive)	Grid FS	Block, object (read intensive)	Read-training data write-ckpt (file, object)	Read intensive (mostly objects)	Block, object (read intensive)	Mixture
Consumption model	Business Adoption	Governments	Governments/industry	Consumer/enterprise	Model builders, sovereign AI	Service providers, enterprise	Government/industry/provider	Converged AI + HPC users
	Networking[‡]	No (@ Periphery[§])	Yes	Inherent	Yes	Yes	Yes	Yes
	Multitenancy	Minimal	Yes	Inherent	Moderate (job based)	Yes	Inherent	Yes
	Virtualization	No (well, some containers)	Some	Built-in VMs	Containers + K8s	Containers + K8s	Built-in VMs, containers	Yes
	Optimized for	Mod/sim HPC	(data-intensive) HPC	Content serving, horizontal scale	Training and tuning large AI models	Models at scale, agents, workflows	Loosely coupled HPC, AI	HPC, AI

VMs: virtual machines.

[*] While the first supercomputer was delivered in 1964, we started counting from 1990 when the first modern scale-out computer was delivered.

[†] Cloud-like reliability: 1) stateless/fungible VMs; 2) reliable persistence layer (S3, etc.); 3) restartable service requests; and 4) eventual consistency for distributed tasks.

[‡] Most HPC and AI training is dominantly East-West, while cloud and AI serving are dominantly North-South (N-S). The difference with AI is that it is N-S + scale-up (multi-GPU networks), while the traditional cloud is largely N-S.

[§] Supercomputers are connected to the outside—but only at the periphery of the system, with a different network.

- beneficial (for example, the work E4S did made its components more accessible to a wider community). HPSF is a good step in this direction.
- As AI is becoming more prevalent in almost every aspect of programming, the models should be treated the same way as open software. The data that were used for training should be made available and documented. While enhancement based on private data will be necessary for some use cases, the data on which open models are based must also be open.
- In general, but especially for science applications, focus on the explainability of AI methods.
- Open hardware is becoming an alternative that needs to be carefully evaluated and considered in supercomputing solutions. Open firmware is also an interesting direction to enhance security and maintainability.
- Work on leveraging low-precision hardware to emulate or perform high-precision calculations is essential. Ultimately, scientific applications need a more rigorous error-based approach to numerical precision.

Our recommendations for scientists and users of supercomputers are as follows:

- Adjust to using cloud infrastructure and AI programming models combined with the existing traditional HPC algorithms.
- Continue to be innovative in terms of continuously increased scale and alternative programming models offered by new hardware (for example, AI accelerators and quantum).
- Invent new algorithms and applications to leverage the new AI and future computing and memory technology.

Our recommendations for industry, integrators, and system vendors are as follows:

- Ensure sufficient interoperability across the components and interconnects to enable reusability across supercomputers.
- Provide sufficient documentation and interfaces for using hardware and core system software.
- Support interfaces and software for the maintenance and management of supercomputers at scale.
- Provide the capability to combine AI capability productively into existing applications.

The need for supercomputing continues to grow. In addition to the needs of traditional scientific computing, AI's needs are driving the evolution of computing hardware and software. The authors lay out several challenges and opportunities for the next decade for computing facilities; developers, scientists and users; and industry.

ACKNOWLEDGMENT

We would like to thank Cullen Bash, Paolo Faraboschi, and Samantika Sury for their advice and insights that substantially helped improve this article. This manuscript has been authored in part by Lawrence Livermore National Security, LLC under Contract DE-AC52-07NA27344 with the U.S. Department of Energy (DOE) (Grant LLNL-JRNL-2005193); it is based in part upon work supported by the DOE, National Nuclear Security Administration under Contract 89233218CNA000001; and it used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. LA-UR-25-23959 (approved for public release). Rosa M. Badia acknowledges funding from Projects CEX2021-001148-S, and PID2023-147979NB-C21 from the MCIN/AEI and MICIU/AEI/10.13039/501100011033 and by FEDER, UE, by the Departament de Recerca i Universitats de la Generalitat de Catalunya, research group MPiEDist (2021 SGR 00412). AMD, the AMD Arrow logo, AMD CDNA, AMD Instinct, AMD ROCm, AMD Infinity Cache, AMD Infinity Fabric, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

REFERENCES

1. R. M. Badia, I. Foster, and D. Milojicic, “Future of HPC,” *IEEE Internet Comput.*, vol. 27, no. 1, pp. 5–6, Jan./Feb. 2023, doi: 10.1109/MIC.2022.3228323.
2. D. Milojicic, P. Faraboschi, N. Dube, and D. Roweth, “Future of HPC: Diversifying heterogeneity,” in *Proc. Design, Automat. Test Europe Conf. Exhib. (DATE)*, 2021, pp. 276–281, doi: 10.23919/DATE51398.2021.9474063.
3. N. Dube, D. Roweth, P. Faraboschi, and D. Milojicic, “Future of HPC: The internet of workflows,” *IEEE Internet Comput.*, vol. 25, no. 5, pp. 26–34, Sep./Oct. 2021, doi: 10.1109/MIC.2021.3103236.
4. G. M. Shipman et al., “The future of HPC in nuclear security,” *IEEE Internet Comput.*, vol. 27, no. 1, pp. 16–23, Jan./Feb. 2023, doi: 10.1109/MIC.2022.3229037.
5. E. Deelman et al., “High-performance computing at a crossroads,” *Science*, vol. 387, no. 6736, pp. 829–831, 2025, doi: 10.1126/science.adu0801.
6. R. Stevens, V. Taylor, J. Nichols, A. B. Maccabe, K. Yellick, and D. Brown, “AI for science: Report on the Department of Energy (DOE) Town Halls on artificial intelligence (AI) for science,” Argonne National Lab. (ANL), Argonne, IL, USA, Tech. Rep. ANL-20/17; 158802; TRN: US2103893, Feb. 2020.
7. W. Jia et al., “Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning,” in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Atlanta, GA, USA, 2020, pp. 1–14.

8. S. Das et al., “Large-scale materials modeling at quantum accuracy: Ab initio simulations of quasicrystals and interacting extended defects in metallic alloys,” in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Denver, CO, USA, 2023, pp. 1–12.
9. G. Dharuman et al., “MProt-DPO: Breaking the ExaFLOPS barrier for multimodal protein design workflows with direct preference optimization,” in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Atlanta, GA, USA, 2024, pp. 1–13.
10. H. Ootomo, K. Ozaki, and R. Yokota, “DGEMM on integer matrix multiplication unit,” 2024, *arXiv:2306.11975*.
11. J. Athavale et al., “Digital twins for data centers,” *Computer*, vol. 57, no. 10, pp. 151–158, Oct. 2024, doi: 10.1109/MC.2024.3436945.
12. W. Brewer et al., “A digital twin framework for liquid-cooled supercomputers as demonstrated at exascale,” in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Atlanta, GA, USA, 2024, pp. 1–18.
13. L. R. Leung, “Earth system modeling or actionable science,” Pacific Northwest Nat. Lab., Richland, WA, USA, Jul. 2024. [Online]. Available: https://www.nersc.gov/assets/Uploads/NERSC_Leung_2024_final.pdf
14. R.R. Seeber, “Associative self-sorting memory,” *presented at the Eastern Joint IRE-AIEE-ACM Comput. Conf. (IRE-AIEE-ACM)*, New York, NY, USA: ACM, Dec. 13–15, 1960, pp. 179–187.
15. P. Kogge et al., “ExaScale computing study: Technology challenges in achieving exascale systems,” DARPA, Arlington, VI, USA, Sep. 2008. [Online]. Available: https://ftp.eecs.berkeley.edu/~yelick/papers/Exascale_final_report.pdf
16. S. Atchley et al., “Frontier: Exploring exascale the system architecture of the first exascale supercomputer,” in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Denver, CO, USA, 2023, pp. 1–16, doi: 10.1145/3581784.3607089.
17. “International Roadmap for Devices and Systems™ 2023 update: Systems and architectures,” IEEE, Piscataway, NJ, USA, 2023. [Online]. Available: https://irds.ieee.org/images/files/pdf/2023/2023IRDS_Perspectives.pdf
18. C. Bash, J. Bian, D. Milojicic, C. D. Patel, L. Strezoski, and V. Terzija, “Energy supplies for future data centers,” *Computer*, vol. 57, no. 7, pp. 126–134, Jul. 2024, doi: 10.1109/MC.2024.3393248.
19. A. Grattafiori et al., “The Llama 3 herd of models,” 2024, *arXiv:2407.21783*.
20. N. Jouppi et al., “TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings,” in *Proc. 50th Annu. Int. Symp. Comput. Archit. (ISCA)*, New York, NY, USA: ACM, 2023, pp. 1–14.
21. C. Kesselman and I. Foster, *The Grid: Blueprint for a New Computing Infrastructure*. Burlington, MA, USA: Morgan Kaufmann, 1999.
22. A. Gupta et al., “Evaluating and improving the performance and scheduling of HPC applications in cloud,” *IEEE Trans. Cloud Comput.*, vol. 4, no. 3, pp. 307–321, Jul./Sep. 2016, doi: 10.1109/TCC.2014.2339858.
23. T. Gamblin et al., “HPC center of the future: R&D acquisition intent,” Lawrence Livermore Nat. Lab., Livermore, CA, USA, Tech. Rep. LLNL-TR-871269, Nov. 14, 2014.

SCOTT ATCHLEY is the CTO at Oak Ridge National Laboratory’s National Center for Computational Science, Oak Ridge, TN 37830 USA. Contact him at scott@ornl.gov.

ROSA M. BADIA is a workflow and distributed computing manager at the Barcelona Supercomputing Center, 08034 Barcellona, Spain. Contact her at rosa.m.badia@bsc.es.

BRONIS R. DE SUPINSKI is the CTO for Livermore Computing at the Lawrence Livermore National Laboratory, Livermore, CA 94550 USA. Contact him at bronis@llnl.gov.

JOSHUA FRYMAN is a fellow at Intel, Hillsboro, OR 97124 USA. Contact him at joshua.b.fryman@intel.com.

DIETER KRANZLMÜLLER is the chair of the board of directors at the Leibniz Supercomputing Centre (LRZ), 85748 Garching bei München, Germany. Contact him at dieter.kranzlmueller@lrz.de.

SRILATHA MANNE is a senior fellow at Advanced Micro Devices, Inc., Seattle, WA 98103 USA. Contact her at srilatha.manne@amd.com.

PEKKA MANNINEN is the director of science and technology at CSC, the Finnish IT Center for Science, 02101 Espoo, Finland. Contact him at pekka.manninen@csc.fi.

SATOSHI MATSUOKA is the director of THE RIKEN Center for Computational Science, Saitama 351-01, Japan. Contact him at matsu@acm.org.

DEJAN MILOJICIC is an HPE fellow and vice president at Hewlett Packard Labs, Milpitas, CA 95035 USA. Contact him at dejan.milojicic@hpe.com.

GALEN SHIPMAN is a computer scientist at the Los Alamos National Laboratory, Los Alamos, NM 87545 USA. Contact him at gshipman@lanl.gov.

ERIC VAN HENSBERGEN is a fellow at ARM, Austin, TX 78735 USA. Contact him at eric.vanhensbergen@arm.com.

ROBERT W. WISNIEWSKI is an HPE fellow, chief architect, and vice president of AI and HPC Solutions at Hewlett Packard Enterprise, Spring TX 77389 USA. Contact him at robert.wisniewski@hpe.com.