POLICY FORUM

COMPUTING

High-performance computing at a crossroads

Long-term plans and comprehensive vision are needed

By **Ewa Deelman[1], Jack Dongarra[2,3,4], Bruce Hendrickson[5], Amanda Randles[6], Daniel Reed[7], Edward Seidel[8], Katherine Yelick[9]**

Over the past four decades, high-performance computing (HPC) has enabled considerable advances in scientific discovery and engineering, spurring technological development across the globe. However, with the demand for precision and fidelity of computational models continuing to grow, HPC faces bottlenecks in data handling, algorithm efficiency, and the scalability of new architectures, especially in fields such as chemistry and biology, where molecular simulations increasingly strain hardware and software limits. Governments worldwide are heavily investing in HPC infrastructure to support research, industrial innovation, and national security, each adopting distinct approaches shaped by national interests and regulatory landscapes. Conversely, in the US, there is no long-term plan or comprehensive vision for the next era of HPC advancements, leaving the future trajectory of US HPC and scientific and technological leadership uncertain.

HPC systems are advanced computing ensembles that harness the power of tens of thousands of tightly coupled processors and high-performance storage to deliver massive processing power, parallelism, and scalability. They enable faster computations, high-throughput exploration of ideas, more detailed models, and real-time decision-making in time-critical scenarios. They provide the ability to search massive key spaces for cryptography, conduct biomedical simulations for patient-specific treatments, and analyze petabyte-scale datasets generated by high-energy particle accelerators. Large-scale partial differential equation solvers are being used in a wide spectrum of simulations, from severe weather forecasting and seismic hazard modeling through aircraft and automotive design to managing oil and gas extraction. These solvers and applications require high fidelity and numerical precision because they often involve solving complex, nonlinear systems over millions or even billions of degrees of freedom. The computational intensity and memory demands of these applications also require HPC's massive parallel processing capabilities, high memory bandwidth, and efficient interconnection networks to handle the needed scale and resolution.

The rise of generative artificial intelligence (AI) has intensified the demands on HPC, transforming it from a resource primarily focused on physics-based simulations and large-scale scientific data analyses into a critical foundation for massive neural network training and inference. With AI's ubiquitous applicability to science, commerce, and global competitiveness, HPC's role has expanded, driving unprecedented demand and introducing new computational, economic, and energy requirements.

The 2024 Nobel Prizes awarded in physics, chemistry, and economics all underscored the pivotal role of computing in advancing scientific discovery and economic competitiveness. These achievements are enabled by AI method development and applications that rely on powerful HPC systems to accelerate AI model training, enable advanced AI research through model exploration, and support the large-scale data processing and data generation. However, although these successes capture global attention, they represent only a fraction of the broader ecosystem needed to maintain leadership in computing-driven innovation. Simply put, continued technical advances in HPC are needed for both traditional simulations and to advance the power and reach of AI.

TECHNICAL CHALLENGES

Today, HPC is in a state of transition, shaped by both technology constraints and market forces. As processor floating point operations per second (FLOPS) have grown exponentially, owing to advancements in transistor density, parallelism, and specialized accelerators, the memory bandwidth—which dictates the amount of data that can be moved to the processors per second—has improved much more modestly because of physical constraints, such as latency and power consumption, leading to an increasing FLOPS-to–memory bandwidth ratio. This means that systems can be inefficient, with processors incurring idle time while waiting for data. Along with fast-paced changes in computing hardware and software, the rise of generative AI, the market dominance of large-scale cloud service providers (hyperscalers), and the growth of international competition in innovation and workforce development are all reshaping the computing ecosystem.

Moore's law (*1*) predicted that the number of transistors on a microchip would double approximately every 2 years, leading to exponential increases in processing power and performance at steady-to-declining price points. Over nearly 60 years, this extraordinary, sustained progress has reshaped the modern world, but no exponential lasts forever. With transistor sizes approaching atomic scales, this rate of progress is no longer attainable.

State-of-the-art microchip fabrication facilities are technological marvels that now cost in excess of $10 billion. These costs can only be justified by products with very high market demand, and unfortunately, the HPC community is too small to drive markets on its own. Since the mid-1990s, the scientific community has leveraged "commodity" processors designed for other markets. This approach has worked well as central processing units (CPUs) and, more recently, graphics processing units (GPUs) have proved suitable for computational science. However, today's dominant market is AI, which does not require the high-precision arithmetic long common in computational modeling and is leading the design of chips with lower-precision arithmetic (16-bit floating point or 8-bit integer precision, as opposed to the 64-bit floating point precision of traditional processors). This trend raises a very real risk that future commodity hardware will not be appropriate for traditional modeling

[1]Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA. [2]Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, USA. [3]Oak Ridge National Laboratory, Oak Ridge, TN, USA. [4]Department of Mathematics, University of Manchester, Manchester, UK. [5]Computing Directorate, Lawrence Livermore National Laboratory, Livermore, CA, USA. [6]Department of Biomedical Engineering, Duke University, Durham, NC, USA. [7]Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT, USA. [8]Office of the President, University of Wyoming, Laramie, WY, USA. [9]Department of Electical Engineering and Computer Sciences, University of California Berkeley, Berkeley, CA, USA. Email: deelman@isi.edu

and simulation applications that continue to be important for science, engineering, and defense.

To mitigate these risks, algorithm researchers and numerical analysts are exploring ideas for effectively using the low-precision arithmetic that future chips will provide. On the hardware side, new design and fabrication models will allow the different functional units of a chip to be fabricated separately (so-called "chiplets") and then joined together. This should lower the cost of semicustom devices, perhaps allowing specialized HPC chips to be affordable. And, of course, the rapid advances in AI that are driving these trends will create new ways to use computing to advance science. At the same time, AI can also improve chip and HPC system design, reliability, scalability, and performance. Still, it remains to be seen whether these approaches will bear fruit for physics-based simulations.

Another major challenge for HPC is the power consumption of current machines, which is on the order of tens of megawatts. In 1974, Robert Dennard observed that the shrinkage from Moore's law came with a corresponding shrinkage in transistor energy consumption (*2*). For the subsequent three decades, microprocessors grew in performance with minimal increases in energy consumption. Dennard scaling ended in the mid-2000s, and the energy consumption of HPC platforms has been steadily rising ever since. There are ways to reduce power by changing programming models and architectures (for example, GPUs are notably more energy efficient than CPUs). However, sustained progress in energy efficiency will require a dedicated research program involving codesign of hardware and software. The results will potentially not only affect HPC but also improve the sustainability of power-hungry AI.

Lawrence Livermore National Laboratory's El Capitan exascale system is the first supercomputer to use Advanced Micro Device (AMD)'s MI300A accelerated processing units.

PHOTO: GARRY MCLEOD/LLNL

GEOPOLITICAL CONTEXT

In the second half of the 20th century, HPC was a tool for solving problems, often using simulations, that were not solvable using analytical methods. From early scientific and military applications, HPC was also adopted by the industrial sector, and more recently, it became a core driver for AI research and applications. Over time, the global HPC landscape has evolved, placing HPC in a geopolitical arena where nations compete for technological sovereignty and leadership, recognizing the strategic importance of computing leadership in advancing economic, scientific, and military capabilities. Governments worldwide are heavily investing in HPC infrastructure to support research, industrial innovation, and national security, each adopting distinct approaches shaped by national interests and regulatory landscapes. For example, EuroHPC (*3*), a European Union initiative, is Europe's response to concerns over data sovereignty and technological dependency. By building some of the world's fastest supercomputers in locations such as Finland, Italy, and Slovenia, EuroHPC aims to reduce reliance on external technologies, prioritize privacy-centric design, and establish Europe as a leader in fields requiring immense computational power, including climate modeling, personalized medicine, and AI. In Japan, the Fugaku supercomputer (*4*) developed by RIKEN and Fujitsu exemplifies a hybrid approach that balances academic and commercial use cases. This model reflects Japan's commitment to pushing computational boundaries for both fundamental research and industrial applications. Meanwhile, China has rapidly advanced its HPC capabilities, leveraging domestically developed infrastructure and processor technologies, such as those underpinning the Sunway TaihuLight and Tianhe-3 supercomputers (*5*). China's HPC strategy underscores a broader national goal of technological self-sufficiency, aiming to reduce dependence on foreign technology amid trade restrictions. These initiatives reveal deep-seated policy and technical tensions around national security, international collaboration, and market independence, highlighting the essential role of computational power in shaping geopolitical influence and sustaining global competitiveness.

US-CENTRIC CONCERNS

In 2024, the US celebrated the success of its Exascale Computing Project (ECP) (*6*), a $1.8 billion project launched by the Department of Energy (DOE) in 2016. This effort culminated in deploying the first US exascale supercomputers (capable of 10^{18} operations per second) at Oak Ridge, Argonne, and Lawrence Livermore National Laboratories. The ECP was a collaborative, multiyear effort that brought together national laboratories, academia, and industry to develop more than 20 new applications running on these exascale systems along with the underlying software stack and advanced hardware features. Today, the US lacks a strategic roadmap and a broad and coordinated federal HPC investment strategy, which puts the US at a crossroads, especially as other global players—notably China, Japan, and the European Union—aggressively pursue ambitious plans to develop their own advanced computing ecosystems.

In the past 2 years, we have coauthored multiple papers (*7*–*10*) that reviewed the state of the art in HPC, examined US leadership in this area, and explored potential future research and development (R&D). These reports have also called for a coordinated national R&D and funding strategy to advance HPC hardware designs, algorithms, software, and their applications. Such a strategy should be sustained over at least the next decade and across multiple federal agencies and should involve companies, including hyperscalers, universities, national laboratories, and strategic international partners. Because the technology and application landscapes are changing extremely rapidly, we advocated for developing prototype systems that would allow exploration of new hardware and software solutions. We also advocated for a holistic codesign to integrate hardware and software systems to optimize performance and efficiency across the computing ecosystem to support a new era of applications that will use AI, simulation, and their combination.

Given the importance of HPC to future economic competitiveness and national security, we are dismayed by the lack of coordinated action to address the recommendations in these reports, and we foresee long-term adverse outcomes for the US. With this Policy Forum, we aim to bring attention to the totality of challenges and opportuni-

ties in HPC and advocate for a multiagency, "whole-nation," and internationally collaborative effort to reenergize HPC R&D.

A key area of focus should be high-end computational science and engineering, where the US has a deep foundation in applied mathematics, particularly in scientific machine learning, optimization, and numerical algorithm development. These fields are essential in building applications and software for future national priorities and for harnessing the potential of emerging computing architectures. Moreover, sustained investment in core computer science disciplines—such as programming models, algorithmic complexity, AI, data management, system architectures, and network research—will be critical to drive future HPC innovations. There is also exciting research into new computing models, in particular quantum computing, which will require deep interdisciplinary efforts to realize. This should be seen as a promising future technology with the potential to transform the feasibility of computational solutions for important applications, such as cryptography, drug discovery, and molecular modeling, but not as a replacement for the breadth and ubiquity of traditional computing in the near term.

The 2022 CHIPS and Science Act (*11*) was a step in the right direction, bolstering semiconductor manufacturing (for example providing almost $8 billion to expand semiconductor facilities across Arizona, New Mexico, Ohio, and Oregon) and creating a new Directorate for Technology, Innovation and Partnerships within the US National Science Foundation geared toward transitioning research into practice and tighter engagement with industry (for example, funding regional microelectronics hubs). Another key advance is a renewed focus on data life cycles and data ecosystems. More recently, the creation of the new Vision for American Science & Technology (VAST) task force to advise decision-makers in the federal government is also another positive development (*12*) because charting the course of US science and technology will undoubtably require investments in HPC computing to solve complex problems. The US Congress has also made recommendations that would help maintain US leadership in AI research, industry adoption, and private sector innovation (*13*), noting that responsible AI innovation requires HPC support to ensure ethical guardrails and sustainable development. Meanwhile, the private sector is investing hundreds of billions of dollars in AI data center infrastructure.

Although laudable, these steps have yet to bear fruit, and more importantly, they do not address the central challenge of federal support for HPC-specific R&D that would result in innovative HPC solutions to the outlined technical challenges. Lack of progress in HPC puts US competitiveness at risk in the race against countries with integrated public-private strategies. The US must urgently pursue meaningful collaborations with industry, including deeper research in computing hardware, software systems, algorithm development, and applications, to capitalize fully on progress in microelectronics.

To navigate HPC's future, we urge the US federal government to organize a task force charged with creating a national, 10-year roadmap for HPC in the post-exascale, post-ECP era. The roadmap should encompass the entire HPC ecosystem, which in addition to hardware acquisition, includes application and system software as well as a well-trained workforce. The task force should include participation from academia, national laboratories, industry, and government. The needed roadmap should include investment at the federal level in computational science and engineering, integrated with AI advancements, and exploration of customized HPC systems tailored to address the distinct demands of multidisciplinary scientific and engineering simulations as well as their commercial and national security applications. Such efforts include creating real hardware and software prototypes at scale, incorporating custom silicon designs to test emerging ideas, and education and training of future researchers and engineers that can contribute to the HPC-AI ecosystem. To realize the roadmap, it is essential to move beyond planning to deliberate implementation. The enactment should promote broad participation, building on ideas developed in programs such as the National Artificial Intelligence Research Resource (NAIRR) (*14*), which aims to broaden access to AI research resources (computational systems, datasets, and educational materials) and to address key barriers that limit participation. Because technologies are evolving at a rapid pace, such a roadmap would need to be a living document. The roadmap and associated actions would need to be periodically revisited and adapted both to the national needs and priorities and to the changing technosocial landscape.

CONCLUSIONS

Recently, the Council on Competitiveness and its National Commission on Innovation and Competitiveness Frontiers have called on the new administration and the new Congress to "act strategically and boldly toward a transformative goal for US competitiveness: boosting U.S. innovation tenfold" (*15*) and to specifically pursue a whole-nation approach to drive technological innovation. There are many lessons to learn from previous efforts and many existing programmatic elements to build upon. With international competition for leadership in computing intensifying, without a renewed commitment, we fear that the US will soon lose scientific computing leadership and technological independence, which will have deeply worrying implications for the US economy, national security, and the international science community. ■

REFERENCES AND NOTES

1. G. E. Moore, *Electronics* **38**, 114 (1965).
2. R. H. Dennard *et al*., *IEEE J. Solid-State Circuits* **9**, 256 (1974).
3. The European High Performance Computing Joint Undertaking (EuroHPC JU); https://eurohpc-ju.europa.eu/about/discover-eurohpc-ju_en.
4. S. Matsuoka, "Fugaku and A64FX: The first exascale supercomputer and its innovative arm CPU" in *2021 Symposium on VLSI Circuits* (IEEE, 2021), pp. 1–3.
5. Z. Chen, "The progress of high performance computing in China and the development trend of international high performance computing" in *China's E-Science Blue Book 2020*, Chinese Academy of Sciences *et al*., Eds. (Springer, 2021), pp. 43–60.
6. P. Messina, *Comput. Sci. Eng.* **19**, 63 (2017).
7. J. Dongarra *et al*., "Can the United States Maintain Its Leadership in High-Performance Computing? - A report from the ASCAC Subcommittee on American Competitiveness and Innovation to the ASCR Office" (US DOE Office of Science, 2023); https://www.osti.gov/biblio/1989107.
8. E. Seidel *et al*., "2024 Advanced Scientific Computing Advisory Committee (ASCR): Facilities Subcommittee Recommendations" (US DOE, 2024); https://www.osti.gov/biblio/2370379.
9. E. Alhajjar, T. Islam, "SIAM Task Force Anticipates Future Directions of Computational Science" (Society for Industrial and Applied Mathematics, 2024); https://www.siam.org/publications/siam-news/articles/siam-task-force-anticipates-future-directions-of-computational-science/.
10. National Academies of Sciences, Engineering, and Medicine, *Charting a Path in a Shifting Technical and Geopolitical Landscape: Post-Exascale Computing for the National Nuclear Security Administration* (National Academies Press, 2023).
11. T. Ryan [D-OH-13], CHIPS and Science Act (117th Congress, 2022); https://www.congress.gov/bill/117th-congress/house-bill/4346.
12. S. S. Parikh, M. K. McNutt, D. Gil, *Science* **386**, 947 (2024).
13. 118th Congress, "Bipartisan House Task Force Report on Artificial Intelligence" (2024); https://www.speaker.gov/wp-content/uploads/2024/12/AI-Task-Force-Report-FINAL.pdf.
14. National Artificial Intelligence Research Resource Task Force, "Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource" (2023); https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf.
15. National Commission on Innovation & Competitiveness Frontiers, "Competing in the Next Economy: Innovating in the Age of Disruption and Discontinuity" (2024); https://compete.org/wp-content/uploads/coc-disruption_discontinuity-call-to-action-final_12.13.24.pdf.

ACKNOWLEDGMENTS

B.H. leads the Computing Directorate at Lawrence Livermore National Security, LLC, under contract no. DE-AC52-07NA2 7344 with the US DOE. D.R. is a Microsoft stockholder. E.S. is a member of the DOE Advanced Scientific Computing Advisory Committee and a member of the Scientific Advisory Board to the president of the Helmholtz Association.

10.1126/science.adu0801