DOI:10.1145/3552309

An examination of how the technology landscape has changed and possible future directions for HPC operations and innovation.

BY DANIEL REED, DENNIS GANNON, AND JACK DONGARRA

HPC Forecast: Cloudy and Uncertain

COMPUTING PERVADES ALL aspects of society in ways once imagined by only a few. Within science and engineering, computing has often been called the third paradigm, complementing theory and experiment, with big data and artificial intelligence (AI) often called the fourth paradigm.[14] Spanning both data analysis and disciplinary and multidisciplinary modeling, scientific computing systems have grown ever larger and more complex, and today's exascale scientific computing systems rival global scientific facilities in cost and complexity. However, all is not well in the land of scientific computing.

In the initial decades of digital computing, government investments and the insights from designing and deploying supercomputers often shaped the next generation of mainstream and consumer computing products. Today, that economic and technological influence has increasingly shifted to smartphone and cloud service companies. Moreover, the end of Dennard scaling,[3] slowdowns in Moore's Law, and the rising costs for continuing semiconductor advances have made building ever-faster supercomputers more economically challenging and intellectually difficult.

As Figure 1 suggests, we believe current approaches to designing and constructing leading-edge high-performance computing (HPC) systems must change in deep and fundamental ways, embracing end-to-end co-design; custom hardware configurations and packaging; large-scale prototyping; and collaboration between the dominant computing companies, smartphone and cloud computing vendors, and traditional computing vendors.

We distinguish leading-edge HPC—the very highest-performing systems—from the broader mainstream of midrange HPC. For the latter, market forces continue to shape that market's expansion. Let's begin by examining where the technology landscape has changed and then examine possible future directions for HPC innovation and operations.

» key insights

- The future of advanced scientific computing, aka supercomputing or high-performance computing (HPC), is at an important inflection point, being reshaped by a combination of technical challenges and market ecosystem shifts.
- These shifts include semiconductor constraints—the end of Dennard scaling, Moore's Law performance limitations, and rising foundry costs—as well ecosystem shifts due to the rise of cloud hyperscalers and the increasing importance of AI technologies.
- Building the next generation of leading-edge HPC systems will require rethinking many fundamentals and historical approaches by embracing end-to-end co-design; custom hardware configurations and packaging; large-scale prototyping, as was common 30 years ago; and collaborative partnerships with the dominant computing ecosystem companies.

IMAGE BY EVANNOVOSTRO

Ecosystem Shifts

To understand HPC's future potential, one must examine the fundamental shifts in computing technology, which have occurred along two axes: the rise of massive-scale commercial clouds, and the economic and technological challenges associated with the evolution of semiconductor technology.

Cloud innovations. Apple, Samsung, Google, Amazon, Microsoft, and other cloud service companies are now major players in the computing hardware and software ecosystems, both in scale and in technical approach. These companies initially purchased standard servers and networking equipment for deployment in traditional collocation centers (colos). As scale increased, they began designing purpose-built data centers optimized for power usage effectiveness (PUE) and deployed at sites selected via multifactor optimization based on the availability of various factors, such as inexpensive energy, tax incentives and political subsidies, political and geological stability, network access, and customer demand.

As cloud scale, complexity, and operational experience continued to grow, additional optimization and opportunities emerged. These include software defined networking (SDN), protocol offloads, and custom network architectures, greatly reducing dependence on traditional network hardware vendors;[6] quantitative analysis of processor,[17] memory,[36,43] network,[7,10] and disk failure modes,[31,35] with consequent redesign for reliability and lower cost (dictating specifications to vendors via consortia such as Open Compute; custom processor SKUs; custom accelerators (FPGAs and ASICs); and complete processor design—for example, Apple silicon, Google TPUs,[19] and AWS Gravitons). In between, the cloud vendors deployed their own global fiber networks.

This virtuous cycle of insatiable consumer demand for rich services, business outsourcing to the cloud, expanding datacenter capacity, and infrastructure cost optimization has had several effects. Most importantly, it has dramatically lessened, and in

Figure 1. Technical and economic forces reshaping HPC.

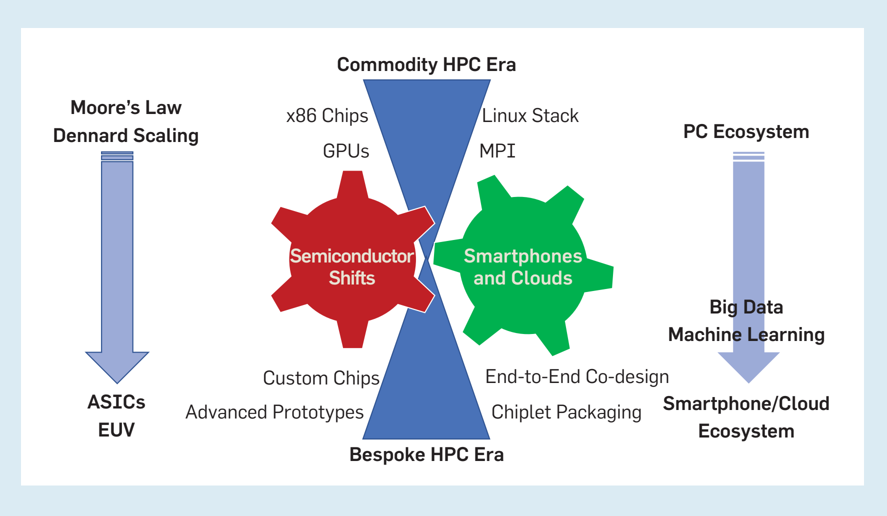
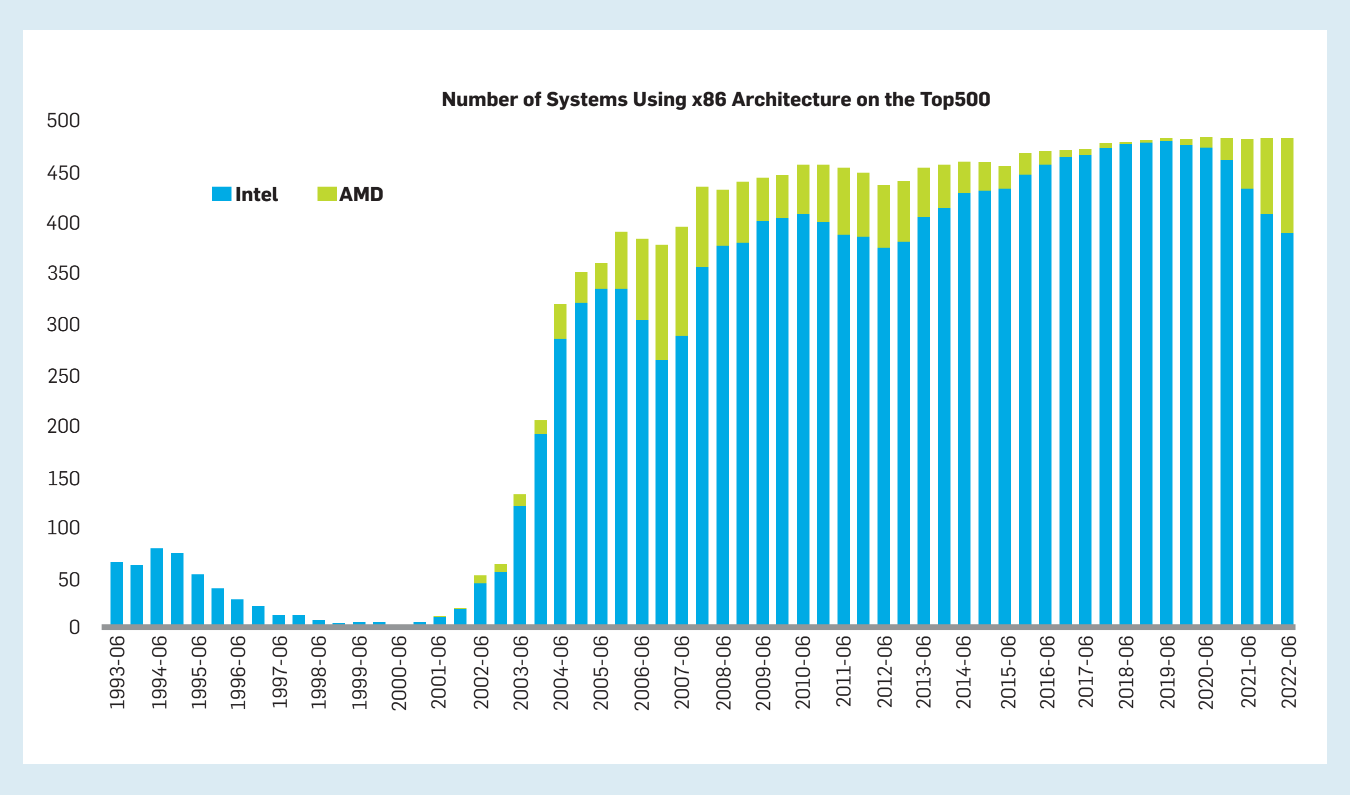


Figure 2. Systems Using the x86-64 architecture on the TOP500.[39]



many cases totally eliminated, their dependence on traditional computing vendors. To see the dramatic shifts in influence and scale, one needs to look no further than cloud service-provider and smartphone-vendor market capitalizations, each near or more than 1 trillion USD. Put another way, the locus of innovation and influence has increasingly shifted from chip vendors and system integrators to cloud service providers.

Semiconductor evolution. Historically, the most reliable engine of performance gains has been the steady rhythm of semiconductor advances: smaller, faster transistors and larger, higher-performance chips. However, as chip feature sizes have approached 5nm and Dennard scaling has ended,[3] the cadence of new technology generations has slowed, even as semiconductor foundry costs have continued to rise. With the shift to extreme ultraviolet (EUV) lithography[4] and gate-all-around FETs,[5] the "minimax problem" of maximizing chip yields, minimizing manufacturing costs, and maximizing chip performance has grown increasingly complex for all computing domains, including HPC.

Chiplets[a,26,29] have emerged to address these issues, while also integrating multiple functions in a single package. Rather than fabricating a monolithic system-on-a-chip (SoC), chiplet technology combines multiple chips, each representing a portion of the desired functionality, possibly fabricated using different processes by different vendors and including IP from multiple sources. Chiplet designs are part of the most recent offerings from Intel and AMD, where the latter's EPYC and Ryzen processors have delivered industry-leading performance via chiplet integration.[29] Similarly, Amazon's Graviton3 uses a chiplet design with seven different chip dies.

An HPC Checkpoint

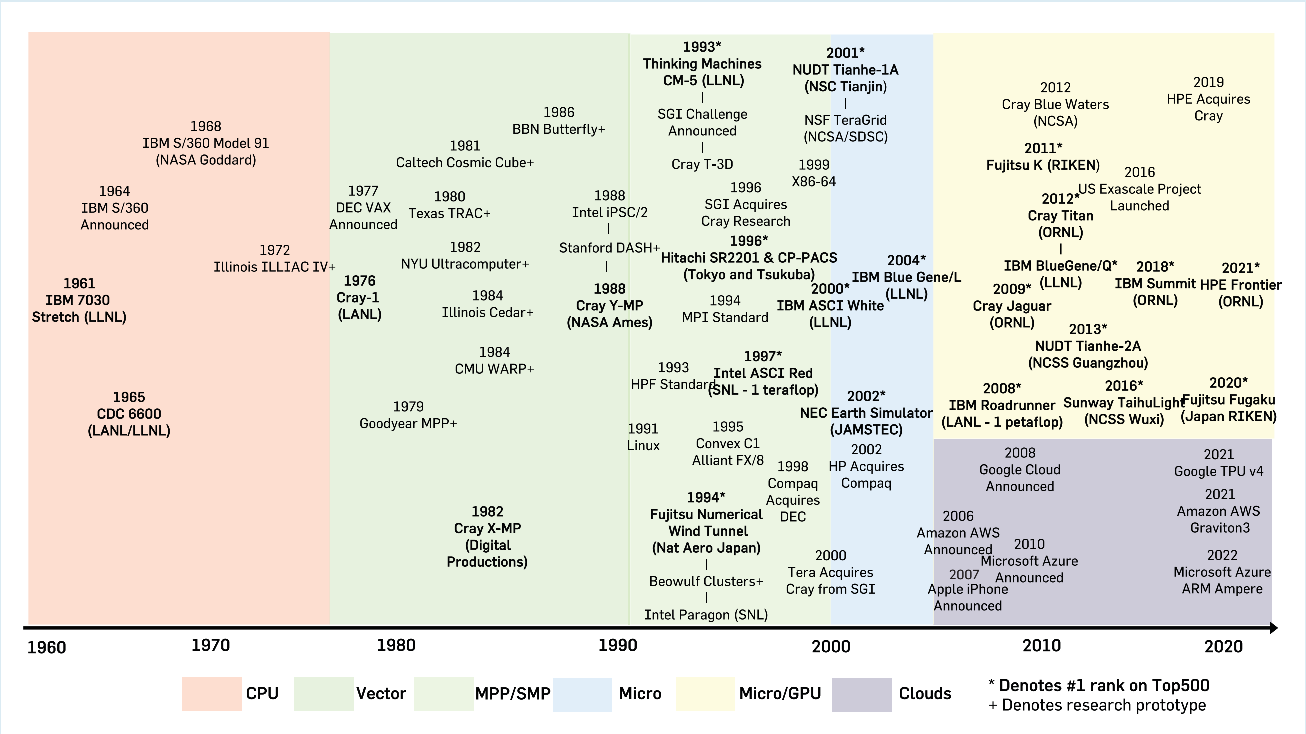
Given the rise of cloud services and increasing constraints on commodity chip performance, it is useful to examine the current state of HPC and how the HPC ecosystem evolved to reach its current structure. From the 1970s to the 1990s, HPC experienced a remarkably active period of architectural creativity and exploration. In the late 1970s, the Cray series of machines[33] introduced vector processing. Companies such as Denelcor and Tera then explored highly multi-threaded parallelism via custom processor design. Universities and companies were also active in exploring new shared memory designs—for example, NYU Ultracomputer,[9] Illinois Cedar,[22] Stanford DASH,[24] and BBN Butterfly.[23]

Finally, distributed-memory, massively parallel computer designs—for example, the Caltech Cosmic Cube,[42] Intel iPSC/2,[40] and Beowulf clusters[38]—established a pattern for hyperscaled performance growth. Riding Moore's Law, the ever-increasing performance of standard microprocessors, together with the cost advantage of volume production, led to the demise of most bespoke HPC systems, a shift often termed the "Attack of the Killer Micros."[27] What followed was academic and industry standardization based on x86-64 processors (see Figure 2) and predominantly gigabit Ethernet and Infiniband networks, the Linux operating system, and message passing via the MPI standard.

By 2000, architectural innovation was limited to node accelerators (for

[a] See Universal Chiplet Interconnect Express (UCIe) standard; https://www.uciexpress.org

Figure 3. Timeline of advanced computing.



example, the addition of GPUs), high-bandwidth memory, and incremental network improvements. The processor, operating system, and network have become the standard interfaces that now define the market boundaries for innovation. At one time, dozens of HPC companies offered competing products. Today, only a few build HPC systems at the largest scales (see Figure 3).

While incremental performance improvements continue with new x86-64 processors and GPU accelerators, basic innovation at the architectural level for supercomputers has largely been lost. However, in the last two years, sparks of architectural creativity are re-emerging, driven by the need to accelerate AI deep learning. Hardware startups, including Graphcore,[18] Groq,[1] and Cerebras,[12] are exploring new architectural avenues. Concurrently, the major cloud service and smartphone providers have also developed custom processor SKUs, custom accelerators (FPGAs and ASICs), and complete processor designs—for example, Apple A15 SoCs, Google TPUs,[19] and AWS Gravitons).

Against this HPC backdrop, the larger computing ecosystem itself is in flux:

- Dennard scaling[3] has ended and continued performance advances increasingly depend on functional specialization via custom ASICs and chiplet-integrated packages.
- Moore's Law is also at or near an end, and transistor costs are likely to increase as feature sizes continue to decrease.
- Advanced computing of all kinds, including HPC, requires ongoing non-recurring engineering (NRE) investment (that is, endothermic) to develop new technologies and systems.
- The cost to build and deploy leading-edge HPC systems continues to rise, straining traditional acquisition models.
- Smartphone and cloud services companies are cash rich (that is, exothermic); they are designing, building, and deploying their own hardware and software infrastructure at an unprecedented scale.
- AI is fueling a revolution in how both businesses and researchers think about problems and their computational solutions.
- Talent is following the money and intellectual opportunities, which are increasingly at a small number of very large companies or creative startups.

At one time, dozens of HPC companies offered competing products. Today, only a few build HPC systems at the largest scales.

With this backdrop, what is the future of computing? Some of it is obvious, given the current dominance of smartphone vendors and cloud service providers. However, it seems likely that continued innovation in advanced HPC will require rethinking some of our traditional approaches and assumptions, including how, where, and when academia, government laboratories, and companies spend finite resources and how we expand the global talent base.

Leading-Edge HPC Futures

It now seems self-evident that supercomputing, at least at the highest levels, is endothermic, requiring regular infusions of non-revenue capital to fund the NRE costs to develop and deploy new technologies and successive generations of integrated systems. In turn, that capital can come either from other, more profitable divisions of a business or from external sources—for example, government investment. Although most basic computing research is conducted in universities, several large companies (for example, IBM, Microsoft, and Google) still conduct long-term basic research in addition to applied research and development (R&D).

Cloud service companies now offer a variety of HPC clusters of varying size, performance, and price. Given this, one might wonder why cloud service companies are not investing even more deeply in the HPC market. Any business leader must always look at the opportunity cost (that is, the time constant, the talent commitment, and cost of money) for any NRE investments and the expected return on investments. The core business question is always how to make the most money with the money one has, absent some other marketing or cultural reason to spend money on loss leaders, bragging rights, or political positioning. The key phrase here is "the most money;" simply being profitable is not enough, which is why leading-edge HPC is rarely viewed as a primary business opportunity.

The NRE costs for leading-edge

supercomputing are now quite large relative to the revenues and market capitalization of those entities we call "computer companies," and they are increasingly out of reach for most government agencies, at least under current funding envelopes. The days are long past when a few million dollars could buy a Cray-1/X-MP/Y-MP/2 or a commodity cluster, and the resulting system would land in the top 10 of the TOP500 list, a ranking of the world's fastest supercomputers. Today, hundreds of millions of dollars are needed to deploy a machine near the top of the TOP500, and at least similar, if not larger, investments in NRE are needed. In addition, the physical plant and the associated energy and cooling costs for operating such systems are now substantial and continuing to rise. What does this brave new world mean for leading-edge HPC? We believe **five maxims** must guide future HPC government and private sector R&D strategies, for all countries.

Maxim One: Semiconductor constraints dictate new approaches. The "free lunch" of lower-cost, higher-performance transistors via Dennard scaling[3] and faster processors via Moore's Law is at an end. Moreover, the de facto assumption that integrating more devices onto a single chip is always the best way to lower costs and maximize performance no longer holds. Individual transistor costs are now flat to rising as feature sizes approach the 1nm range, due to the interplay of chip yields on 300nm wafers and increasing fabrication facility costs. Today, the investment needed to build state-of-the-art facilities is denominated in billions of dollars per facility.

As recent geopolitical events have shown, there are substantial social, political, economic, and national security risks for any country or region lacking a robust silicon fabrication ecosystem. Fabless semiconductor firms rightly focus on design and innovation, but manufacturing those designs depends on reliable access to state-of-the-art fabrication facilities, as the ongoing global semiconductor shortage has shown. The recently passed U.S. CHIPS and Science Act[b] provides roughly 50 billion USD in subsidies to support construction of semiconductor foundries in the U.S., with similar considerations underway in the EU. To date, Intel, Micron, TSMC, and GlobalFoundries have announced plans to build new chip fabrication facilities in the U.S.

Optimization must balance chip fabrication facility costs, now near 10 billion USD at the leading edge; chip yield per wafer; and chip performance. This optimization process has rekindled interest in packaging multiple chips, often fabricated with distinct processes and feature sizes. Such chiplets[26,29] not only enable the mixing of capabilities from multiple sources; they are an economic and engineering reaction to the interplay of chip defect rates, the cadence of feature size reductions, and semiconductor manufacturing costs. However, this approach requires academic, government, and industry collaborations to establish interoperability standards—for example, the Open Domain-Specific Architecture (OSDA) project[41] within the Open Compute Project[c] and the Universal Chiplet Interconnect Express (UCIe)[d] standard. Open chiplet standards can allow the best ideas from multiple sources to be integrated effectively, in innovative ways, to develop next-generation HPC architectures.

Maxim Two: End-to-end hardware/software co-design is essential. Leveraging the commodity semiconductor ecosystem has led to an HPC monoculture, dominated by x86-64 processors and GPU accelerators. Given current semiconductor constraints, substantial system performance increases will require more intentional end-to-end co-design,[28] from device physics to applications. China and Japan are developing HPC systems outside the conventional path, as seen by the Top500.

The Fugaku supercomputer[34] (Post-K Computer), developed jointly by RIKEN and Fujitsu Limited based on ARM technology with vector instructions, occupied the top spot on the Top500. It also swept the other rankings of supercomputer performance (HPCG, HPL-AI, and Graph500). Fugaku is designed for versatile use based on a co-design approach between application- and system-development teams.

Likewise, the Chinese government, its academic community, and its domestic HPC vendors have made great efforts in the last few years to build a mature, self-designed hardware and software ecosystem and promote the possibility of running large and complex HPC applications on large, domestically produced supercomputers. It has been reported that China has two exaflops systems (OceanLight and Tianhe-3); several Gordon Bell prize submissions ran on OceanLight.[25] Reflecting global tensions surrounding advanced technologies, a new measure by the U.S. Department of Commerce now precludes companies from supplying advanced computing chips, chip-making equipment, and other products to China unless they receive a special license.

Similar application-driven co-designs were evident in the AI hardware startup companies mentioned previously, as well as the cloud vendor accelerators. Such co-design means more than encouraging tweaks of existing products or product plans. Rather, it means looking holistically at the problem space, then envisioning, designing, testing, and fabricating appropriate solutions. In addition to deep partnerships with hardware vendors and cloud ecosystem operators, end-to-end co-design will require substantially expanded government investment in basic R&D, unconstrained by forced deployment timelines. In addition to partnerships with x86-64 vendors, the ARM license model and the open source RISC-V[11] specification offer intriguing possibilities.

Maxim Three: Prototyping at scale is required to test new ideas. Semiconductors, chiplets, AI hardware, cloud innovations—the computing system is now in great flux and not for the first time. As Figure 3 shows, the 1980s and 1990s were filled with innovative computing research projects and companies, many aided by government funding, that built novel hardware, new programming tools,

[b] See U.S. CHIPS and Science Act; https://science.house.gov/chipsandscienceact

[c] See Open Compute Project; https://www.opencompute.org/

[d] See UCIe standard; https://www.uciexpress.org

and system software at large scale. To escape the current HPC monoculture and build systems better suited to current and emerging scientific workloads at the leading edge, we must build real hardware and software prototypes at scale—not just incremental ones, but ones that truly test new ideas using custom silicon and associated software. Implicitly, this means accepting the risk of failure, including at substantial scale, drawing insights from the failure, and building lessons based on those insights. A prototyping project that must succeed is not a research project; it is a product development.

Building such prototypes, whether in industry, national laboratories, or academia, depends on recruiting and sustaining integrated research teams—chip designers, packaging engineers, system software developers, programming environment developers, and application domain experts—in an integrated, end-to-end way. Such opportunities make it intellectually attractive to work on science and engineering problems, particularly given industry partnerships and opportunities to translate research ideas into practice. Implicit in such teams is coordinated funding for workforce development, basic research, and the applied R&D needed to develop and test prototype systems.

Maxim Four: The space of leading-edge HPC applications is far broader now than in the past. Leading-edge HPC originated in domains dominated by complex optimization problems and solutions of time-dependent, partial differential equations on complex meshes. Those domains will always matter, but other areas of advanced computing in science and engineering are of high and growing importance. As an example, the *Science 2021 Breakthrough of the Year*[2] was for AI-enabled protein structure prediction,[20] with transformative implications for biology and biomedicine.

Even in traditional HPC domains, the use of AI for dataset reduction and reconstruction, and for PDE solver acceleration, is transforming computational modeling and simulation. Deep-learning methods developed by the cloud companies are changing the course of computational science, and university collaborations are growing. For example, the University of Washington is working with Microsoft Azure on protein-protein interactions.[16] In other areas, OpenAI is showing that deep learning can solve challenging Math Olympiad problems[32] and can also be used to classify galaxies in astrophysics.[21] Generative adversarial networks (GANs)[8] have been used to understand groundwater flow in superfund sites[42] and deep neural networks have been trained to help design non-photonic structures.[30] More than 20 of the papers written for SC21, supercomputing's flagship conference, were on neural networks. The HPC ecosystem is expanding and engaging new domains and approaches in deep learning, creating new and common ground with cloud service providers.

Maxim Five: Cloud economics have changed the supply-chain ecosystem. The largest HPC systems are now dwarfed by the scale of commercial cloud infrastructure and social media company deployments. A 500 million USD supercomputer acquisition every five years provides limited financial leverage relative to the billions of dollars spent each year by cloud vendors. Driven by market economics, computing hardware and software vendors, themselves increasingly small relative to the large cloud vendors, now respond most directly to cloud vendor needs.

In turn, government investment (for example, the U.S. Department of Energy's (DOE) Exascale DesignForward, FastForward, and PathForward programs,[e] and the European Union's HPC-Europa3) is small compared to the scale of commercial cloud investments and their leverage with those same vendors. HPC-Europa3, funded under the EU's Eighth Framework Programme, better known as Horizon 2020, has a budget of only 9.2 million euro.[f] Similarly, the U.S. DOE's multiyear investment of 400 million USD via the FastForward, DesignForward, and PathForward programs as part of the Exascale Computing Project (ECP) targeted reduced power consumption, resilience, and improved network and system integration. The DOE only supplied approximately 100 million USD in NRE for each of the exascale systems under construction. During that same period, the cloud companies invested billions. Market research[17] suggests that China, Japan, the U.S., and the EU may each procure one to two exascale-class systems per year, each estimated at approximately 400 million USD.

The financial implications are clear. Government and academic HPC communities have limited leverage and cannot influence vendors in the same ways they did in the past. New, collaborative models of partnership and funding are needed that recognize and embrace ecosystem changes and their implications, both in use of cloud services and collaborative development of new system architectures. The cloud is evolving as a platform where specialized services, such as attached quantum processors, specialized deep-learning accelerators, and high-performance graph database servers, can be configured and integrated into a variety of scientific workflows. However, that is not the whole HPC story. Massive-scale simulations require irregular sparse data structures, and the best algorithms are extremely inefficient on the current generation of supercomputers. The commercial cloud is only a part of HPC's future. New architecture research and advanced prototyping are also needed.

As we have emphasized, the market capitalizations of the smartphone and cloud services vendors now dominate the computing ecosystem, and the overlap between commercial AI application hardware needs and those of scientific and engineering computing is creating new opportunities. We realize this may be heretical to some, but there are times and places where commercial cloud services can be the best option to support scientific and engineering computing needs.

The performance gaps between cloud services and HPC gaps have lessened substantially over the past

[e] See DoE/Exascale Computing Project's Pathforward; https://www.exascaleproject.org/research-group/pathforward/

[f] See the European Commission's "Transnational Access Programme for a Pan-European Network of HPC Research Infrastructures and Laboratories for Scientific Computing"; https://cordis.europa.eu/project/id/730897

decade, as shown by a recent comparative analysis.[13] Moreover, HPC as a service is now real and effective, both because of its performance and the rich and rapidly expanding set of hardware capabilities and software services. The latter is especially important; cloud services offer some features not readily available in the HPC software ecosystem.

Some in academia and the national laboratory community will immediately say, "But, we can do it cheaper, and our systems are bigger!" Perhaps, if one looks solely at retail prices, but those may not be the appropriate perspectives. Proving such claims means being dispassionate about technological innovation, NRE investments, and opportunity costs. In turn, this requires a mix of economic and cultural realism in making build versus use decisions and taking an expansive view of the application space, unique hardware capabilities, and software tools. Opportunity costs are real, though not often quantified in academia or government. Today, capacity computing (that is, solving an ensemble of smaller problems) can easily be satisfied with a cloud-based solution, and on-demand, episodic computing of both capacity and large-scale scientific computing can benefit from cloud scaling.

Conclusion

The computing ecosystem is in enormous flux, creating both opportunities and challenges for the future of advanced scientific computing. For the past 20 years, the most reliable engine of HPC performance gains has been the steady improvement in commodity CPU technology due to semiconductor advances. But with the slowing of Moore's Law and the end of Dennard scaling, improved performance of supercomputers has increasingly relied on larger scale (that is, building systems with more computing elements) and GPU co-processing. Concurrently, the computing ecosystem has shifted, with the rise of hyperscale cloud vendors that are developing new hardware and software technologies.

Looking forward, it seems increasingly unlikely that future high-end HPC systems will be procured and assembled solely by commercial integrators from only commodity components. Rather, future advances will require embracing end-to-end design, testing, evaluating advanced prototypes, and partnering strategically with not only traditional chip and HPC vendors but also with the new cloud ecosystem vendors. These are likely to involve collaborative partnerships among academia, government laboratories, chip vendors, and cloud providers; increasingly bespoke systems designed and built collaboratively to support key scientific and engineering workload needs; or a combination of these two.

Simply being profitable is not enough, which is why leading-edge HPC is rarely viewed as a primary business opportunity.

Put another way, in contrast to midrange systems, leading-edge HPC systems are increasingly similar to large-scale scientific instruments (for example, the Vera Rubin Observatory, the LIGO gravity wave detector, or the Large Hadron Collider), with limited economic incentives for commercial development. Each contains commercially designed and constructed technology, but each also contains large numbers of custom elements for which there is no sustainable business model. Instead, we build these instruments because we want them to explore open scientific questions, and we recognize that their design and construction requires both government investment and innovative private sector partnerships.

Like many other large-scale scientific instruments, where international collaborations are an increasingly common way to share costs and facilitate research collaborations, leading-edge computing would benefit from increased international partnerships, recognizing that in today's world, national security and economic competitiveness issues will necessarily limit sharing certain "dual-use" technologies. Subject to those very real constraints, if we are to build more performant, leading-edge HPC systems, we believe there is a need for greater government investment in semiconductor futures—both basic research and foundry construction—along with an integrated, long-term R&D program that funds academic, national laboratory, and private-sector partnerships to design, develop, and test advanced computing prototypes.

These investments must be tens, perhaps hundreds of billions of dollars, in scale.

We have long relied on the commercial market for the building blocks of leading-edge HPC systems. Although this has leveraged commodity economics, it has also resulted in systems ill-matched to the algorithmic needs of scientific and engineering applications. With the end of Moore's Law, we now have both the opportunity and the pressing need to invest in first principles design.

Investing in the future is never easy, but it is critical if we are to continue to develop and deploy new generations of HPC systems, ones that leverage economic shifts, commercial practices, and emerging technologies to advance scientific discovery. Intel's Andrew Grove was right when he said, "Only the paranoid survive," but paranoia alone is not enough. Successful competitors also need substantial financial resources and a commitment to technological opportunities and scientific innovation.

Acknowledgments

Jack Dongarra was supported in part by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. The authors are grateful for thoughtful feedback on earlier drafts of this article from the anonymous reviewers, as well as Doug Burger (Microsoft), Tony Hey (U.K. Science and Technology Facilities Council), and John Shalf (Lawrence Berkeley National Laboratory).

References

1. Abts, D. et al. Think fast: A Tensor streaming processor (TSP) for accelerating deep learning workloads. In *Proceedings of the ACM/IEEE 47th Annual Intern. Symp. on Computer Architecture*, IEEE Press (2020), 145–158; http://bit.ly/3PA2a87.
2. Beckwith, W. Science's 2021 breakthrough: AI-powered protein prediction. *AAAS* (December 2021); http://bit.ly/3W4j6Gd.
3. Bohr, M. A 30 year retrospective on Dennard's MOSFET scaling paper. *IEEE Solid-State Circuits Society Newsletter 12*, 1 (2007), 11–13; https://doi.org/10.1109/N-SSC.2007.4785534.
4. Chang, Y-W., Liu, R-G., and Fang, S-Y. EUV and e-beam manufacturability: Challenges and solutions. In *Proceedings of the 52nd Annual Design Automation Conf,* Association for Computing Machinery (June 2015), 1-6; https://doi.org/10.1145/2744769.2747925.
5. Dey, S. et al. Performance and opportunities of gate-all-around vertically stacked nanowire transistors at 3nm technology nodes. In *2019 Devices for Integrated Circuit*, 94–98; https://doi.org/10.1109/DEVIC.2019.8783385.
6. Firestone, D. et al. Azure accelerated networking: SmartNICs in the public cloud. *15th USENIX Symp. on Networked Systems Design and Implementation 15*. (2015); https://bit.ly/3HFyNzd.
7. Gill, P., Jain, N., and Nagappan, N. Understanding network failures in data centers: Measurement, analysis, and implications. In *Proceedings of the ACM SIGCOMM 2011 Conf*, 350–361; https://doi.org/10.1145/2018436.2018477.
8. Goodfellow, I. et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27* (2014), Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Eds.), Curran Associates, Inc.; https://bit.ly/3BI6IDy.
9. Gottlieb, A. et al. The NYU Ultracomputer—Designing a MIMD, shared-memory parallel machine. In *Proceedings of the 9th Annual Symp. on Computer Architecture, IEEE Computer Society Press* (1982), 27–42.
10. Govindan, R. et al. Evolve or die: High-availability design principles drawn from Google's network infrastructure. In *Proceedings of the 2016 ACM SIGCOMM Conf.*, 58–72; https://doi.org/10.1145/2934872.2934891.
11. Greengard, S. Will RISC-V revolutionize computing? *Communications of the ACM 63*, 5 (April 2020), 30–32; https://doi.org/10.1145/3386377.
12. Groeneveld, P. Wafer scale interconnect and pathfinding for machine learning hardware. In *Proceedings of the Workshop on System-Level Interconnect: Problems and Pathfinding Workshop* (2020); https://doi.org/10.1145/ 3414622.3432992.
13. Guidi, G. et al. 10 years later: Cloud computing is closing the performance gap. *Companion of the ACM/SPEC Intern. Conf. on Performance Engineering* (2021), 41–48; https://bit.ly/3FBFa3I.
14. Hey, T., Tansley, S., and Tolle, K. (Eds.). *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research (2009); https://bit.ly/3v3qXrn.
15. Hochschild, P.H. et al. Cores that don't count. In *Proceedings of the Workshop on Hot Topics in Operating Systems.* Association for Computing Machinery (2021), 9–16; https://bit.ly/3hxVSZT.
16. Horvitz, E. A leap forward in bioscience (2022); https://erichorvitz.com/Leap_forward_bioscience.htm.
17. HPC market update briefing during SC21. *Supercomputing 2021*; https://bit.ly/3HR6Kg2.
18. Jia, Z., Tillman, B., Maggioni, M., and Scarpazza, D.P. Dissecting the Graphcore IPU architecture via microbenchmarking. *CoRR*, abs/1912.03413 (2019). arXiv:1912.03413; http://arxiv.org/abs/1912.03413.
19. Jouppi, N.P. et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th Annual Intern. Symp. on Computer Architecture*, Association for Computing Machinery (2017), 1–12; https://doi.org/10.1145/3079856.3080246.
20. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* (August 2021), 1476–4687; https://doi.org/10.1038/s41586-021-03819-2.
21. Kim, E.J. and Brunner, R.J. Star-galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society 464*, 4 (2016).
22. Kuck, D.J., Davidson, E.S., Lawrie, D.H., and Sameh, A.H. Parallel supercomputing today and the cedar approach. *Science 231*, 4741 (1986), 967–974; https://doi.org/10.1126/science.231.4741.967; arXiv:https://www.science.org/doi/pdf/10.1126/science.231.4741.967.
23. LeBlanc, T.J., Scott, M.L., and Brown, C.M. Large-scale parallel programming: Experience with BBN Butterfly parallel processor. *SIGPLAN Not. 23*, 9 (January 1988), 161–172; https://doi.org/10.1145/62116.62131.
24. Lenoski, D. et al. The Stanford DASH Multiprocessor. *Computer 25*, 3 (March 1992), 63–79; https://doi.org/10.1109/2.121510.
25. Liu, Y. et al. Closing the "quantum supremacy" gap: Achieving real-time simulation of a random quantum circuit using a new Sunway supercomputer. In *Proceedings of the Intern. Conf. for High Performance Computing, Networking, Storage and Analysis*, Association for Computing Machinery (2021); https://doi.org/10.1145/3458817.3487399.
26. Loh, G.H., Naffziger, S., and Lepak, K. Understanding chiplets today to anticipate future integration opportunities and limits. In *2021 Design, Automation Test in Europe Conf. Exhibition*, 142–145; https://doi.org/10.23919/DATE51398.2021.9474021.
27. Markoff, J. The attack of the killer micros. *The New York Times* (May 6, 1991); https://bit.ly/3FDzhDr.
28. Murray, C. et al. Basic research needs for microelectronics. U.S. DoE Office of Science, (October 2018); https://doi.org/10.2172/1616249.
29. Naffziger, S. et al. Pioneering chiplet technology and design for the AMD EPYC™ and Ryzen™ processor families: Industrial product. *2021 ACM/IEEE 48th Annual Intern. Symp. on Computer Architecture*, 57–70; https://doi.org/10.1109/ISCA52012.2021.00014.
30. Peurifoy, J. et al. Nanophotonic particle simulation and inverse design using artificial neural networks. *Science Advances 8*, 5 (2022).
31. Pinheiro, E., Weber, W-D., and Barroso, L.A. Failure trends in a large disk drive population. In *Proceedings of the 5th USENIX Conf. on File and Storage Technologies*, USENIX Association, (2007), 2.
32. Polu, S., Han, J.M., and Sutskever, I. Solving (some) formal math Olympiad problems. Open AI (2022); https://openai.com/blog/formal-math.
33. Russell, R.M. The CRAY-1 computer system. *Communications of the ACM 21*, 1 (January 1978), 63–72; https://doi.org/10.1145/359327.359336.
34. Sato, M. et al. Co-design for A64FX Manycore processor and "Fugaku". In *Proceedings of the Intern. Conf. for High Performance Computing, Networking, Storage and Analysis*, IEEE Press (2020).
35. Schroeder, B. and Gibson, G.A. Understanding disk failure rates: What does an MTTF of 1,000,000 hours mean to you? *ACM Trans. Storage 3*, 3 (October 2007), 8–es; https://doi.org/10.1145/1288783.1288785.
36. Schroeder, B., Pinheiro, E., and Weber, W-D. DRAM errors in the wild: A large-scale field study. *Communications of the ACM 54*, 2 (February 2011), 100–107; https://doi.org/10.1145/1897816.1897844.
37. Seitz, C.L. The cosmic cube. *Communications of the ACM 28*, 1 (1985), 22–33; https://bit.ly/3FWoLIy.
38. Sterling, T. et al. BEOWULF: A parallel workstation for scientific computation. In *Proceedings of the 24th Intern. Conf. on Parallel Processing 1*, CRC Press (1995), I:11–14.
39. Strohmaier, E., Meuer, H.W., Dongarra, J., and Simon, H.D. The TOP500 list and progress in high-performance computing. *Computer 48*, 11 (2015), 42–49; https://doi.org/10.1109/MC.2015.338.
40. The Intel iPSC/2 system: The concurrent supercomputer for production applications. In *Proceedings of the 3rd Conf. on Hypercube Concurrent Computers and Applications: Architecture, Software, Computer Systems, and General Issues 1*, Association for Computing Machinery (January 1988), 843–846; https://doi.org/10.1145/62297.62412.
41. Vinnakota, B. The open domain-specific architecture: Next steps to production. In *Proceedings of the 8th Annual ACM Intern. Conf. on Nanoscale Computing and Communication*. Association for Computing Machinery (2021); https://bit.ly/3PBPtK0.
42. Yang, L. et al. Highly scalable, physics-informed GANs for learning solutions of stochastic PDEs. *ArXiv* abs/1910.13444v1 (2021).
43. Zivanovic, D. et al. DRAM errors in the field: A statistical approach. In *Proceedings of the Intern. Symp. on Memory Systems,* Association for Computing Machinery (2019), 69–84; https://bit.ly/3HI3EuR.

Daniel Reed is a Presidential Professor at the University of Utah, Computer Science and Electrical & Computer Engineering, Salt Lake City, Utah, USA.

Dennis Gannon is Professor Emeritus at the Indiana University, Luddy School of Informatics, Computing and Engineering, Bloomington, Indiana, USA.

Jack Dongarra is Professor Emeritus at the University of Tennessee, Innovative Computing Laboratory, EECS Department, Knoxville, Tennessee, USA.