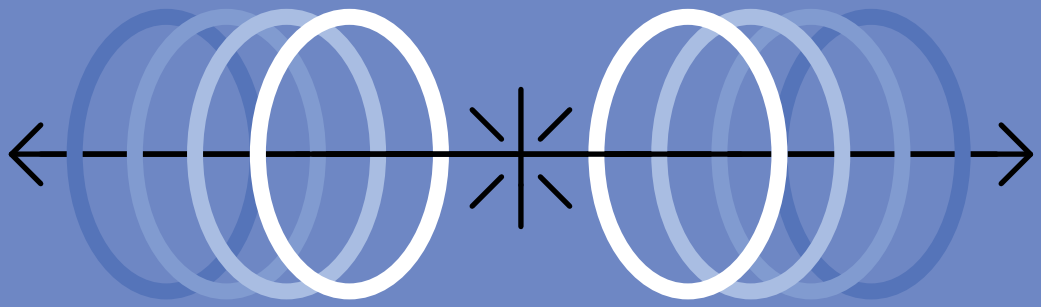
HARVARD BUSINESS REVIEW



Accelerating Innovation with High-Performance Computing in the Intelligent Cloud

HBR Analytic Services **WHITE PAPER**

Sponsored by

Microsoft Azure

Accelerating Innovation with High-Performance Computing in the Intelligent Cloud

High-performance computing (HPC) systems, with enormous computational capabilities, have been the go-to solution for decades for organizations tackling immense scientific, academic, and business challenges. From cosmological simulations of the universe's formation to the development of Covid-19 vaccines and rendering computer graphics for Hollywood movies, HPC systems have addressed some of the world's most notable computational challenges—tasks that would overwhelm traditional computers.

Now HPC and artificial intelligence (AI) are evolving in complementary ways, expanding access beyond traditional scientific and academic institutions into new commercial sectors and applications. "The convergence is real, and it's been real for a while now," says Thuc Hoang, the deputy assistant deputy administrator for Advanced Simulation and Computing (ASC) at the Department of Energy's National Nuclear Security Administration (NNSA). The ASC program has leveraged the strategic advantages of running AI workloads, such as applying inference to make predictions or solve tasks, on HPC systems powered by graphics processing units (GPUs). "In our world, you can't do AI right without the GPUs," she notes, "and you need HPC to do AI training and GPUs for inference, which is where the real promise lies."

HPC offered as a cloud-based service is training some of the largest and best-known AI models. As a platform, it also benefits from AI advancements such as using machine learning to improve simulation outcomes, optimize performance, automate resource allocation, and implement predictive maintenance. This potent combination unlocks the transformational capabilities of cloud-based HPC, enabling it to run workloads more efficiently than ever before on massively parallel processing systems located on-premises (on-prem) and in the cloud. "The learning curves and adjustments to running HPC cloud workloads, as opposed to running on-prem, are much easier than they used to be, say, five to 10

HIGHLIGHTS

While high-performance computing (HPC) has long run simulations, such as those in computational fluid dynamics, aeronautics, and financial risk analysis, it now has capabilities that were previously unavailable to most organizations.

No matter the environment, HPC operations teams face a cascading set of decisions that involve balancing workload requirements against factors such as resource availability, time-to-solution, and projected costs.

The potent yet still emerging combination of HPC and artificial intelligence in the cloud is broadening access to tools that can transform the breadth, speed, and scale of scientific, academic, and commercial research.

Hyperion Research in St. Paul, Minn., reported in April 2025 that the HPC-AI market grew 23.5% in 2024 and may exceed $100 billion by 2028.

years ago," explains Douglas Kothe, chief research officer at Sandia National Laboratories in Albuquerque, N.M.

To be sure, demand for HPC services is robust in the age of AI. Hyperion Research in St. Paul, Minn., reported in April 2025 that the HPC-AI market grew 23.5% in 2024 and may exceed $100 billion by 2028.[1] The cloud will play a pivotal role in scaling this advancement. Most HPC sites operate on-prem and in the cloud, with 44% of sites placing half or more of their HPC workloads in the cloud, according to Hyperion in a September 2024 report called "Cloud-Based AI Activity for HPC." Just 2% of HPC sites eschew the cloud altogether.[2]

The Hyperion study, spanning decision makers in commercial, academic, and government settings, also found that nearly two-thirds (64%) are exploring a range of potential cloud-based AI performance enhancements, with 42% running production-level AI-enabled workloads in the cloud. Hyperion concluded that "across the board, these respondents demonstrated a preference toward the cloud over on-premises options for AI-centric activities."

But whether running on-prem or in the cloud, HPC workloads typically require collaboration across multiple teams with different roles, namely domain scientists or engineers who define the research problems, data scientists or mathematicians who design models and algorithms, software engineers who optimize and parallelize code, and HPC system administrators or operators who provision, monitor, and manage the infrastructure. "The real value of HPC and AI as a cloud service, whether the [services] are private on-prem or off-prem, is really for collaborations—big distributed collaborations," says NNSA's Hoang.

This report aims to understand and examine how organizations can accelerate, scale, and drive transformative business and research outcomes by leveraging cloud technologies that unlock HPC's full potential. The report will also explore the increasingly synergistic relationship between HPC and AI, coinciding with partnerships between research institutions and cloud service providers (CSPs), to learn how these dynamic relationships enable organizations to enhance innovation, improve efficiency, and achieve their desired objectives.

High-Performance Computing Cloud Workloads

While HPC has long run simulations, such as those in computational fluid dynamics, aeronautics, and financial risk analysis, it now has capabilities that were previously unavailable to most organizations. The increased demand for cutting-edge performance in the burgeoning AI market has driven rapid changes in processors, including GPUs that excel at performing multiple calculations simultaneously, data management, power requirements, cooling solutions, and low-latency interconnect technologies. Parallelism is HPC's superpower. Training a large language model (LLM) for generative AI requires processing a highly complex calculation distributed across thousands of GPUs. Interconnect technologies such as InfiniBand or NVLink enable GPUs to share this data almost instantaneously.

The advent of AI-augmented HPC workloads is a particularly intriguing development for America's massive exascale supercomputer systems, according to Sandia's Kothe. Sandia is a federally funded research lab operated by NNSA under the auspices of the U.S. Department of Energy (DOE). The DOE employs three of the world's largest exascale systems—each with the capacity to perform at least one quintillion floating-point operations per second. These supercomputers power scientific discovery through simulations and modeling in fields such as national security; energy; discovery

> **“There are not insignificant trade-offs involved in where you elect to compute. Cost is not a minor consideration when you’re talking about scaling up in a cloud-based environment.”**
>
> Bob Sorensen, senior vice president of research, Hyperion Research

science, including astrophysics, cosmology, and high-energy physics; materials science; nuclear physics; engineering; and computer science. Kothe says of AI-augmented HPC, “Overall, the possibilities are endless with regard to how these two approaches can synergistically help each other.”

Meanwhile, HPC systems and cloud clusters share fundamental economic principles, like resource optimization and total cost of ownership considerations, but their distinct architectures lead to different cost optimization strategies for each environment. HPC workloads in public or private clouds often run complex simulations with intense utilization patterns, high data transfer rates, and application licensing models that result in performance-per-dollar considerations different from those of standard public cloud clusters.

Bob Sorensen, senior vice president of research at Hyperion Research, believes that energy costs push operational costs higher in the cloud compared to on-prem. “If a site is using GPUs or other power-intensive computational hardware, it’s much cheaper for the HPC site to operate on-prem because the site doesn’t have to pay for the power; the power bills are paid separately by the overall site or laboratory budget,” he explains. “However, if the computer center goes to the cloud, they could end up paying the power bill because it is wrapped up in the charges for access to computer capabilities. So, there are not insignificant trade-offs involved in where you elect to compute. Cost is not a minor consideration when you’re talking about scaling up in a cloud-based environment.”

Large-scale high-fidelity simulations are typically more cost-effective to run on-prem than on cloud HPC systems. Hoang explains that while price is a “huge factor” in determining where to place workloads, another operational concern is avoiding software and hardware inconsistencies across multiple regions and data centers. Cloud management best practices aim to eliminate unpredictable costs and outcomes.

Although the NNSA runs most of its ASC simulation workloads on-prem, Hoang says the three NNSA labs “constantly talk to CSPs to understand the goodness of their service model,” in part because “we are our own CSP,” providing compute services to sites along with utilizing available HPC resources. “We have end users at some of our NNSA sites who don’t have on-prem HPC-AI resources, but they need those cycles on a bursty level,” she says, referring to the practice of moving a workload from on-prem to the cloud for additional processing capacity on an as-needed basis.

A Sandbox for New Technologies

Just as many large commercial enterprises apply a service model to running their geographically dispersed data centers and harnessing technology resources, Hoang says the NNSA looks into adopting a similar approach to utilizing HPC, AI, and other shared services. Her program researches how best to manage data transfer between NNSA laboratory sites, examining issues such as “how we authenticate the users and figure out how much of the data generated at one site is needed at another site.”

Yet data transfer isn’t merely a technical issue. Cloud operations can be constrained when data sovereignty or compliance regulations prohibit or severely limit the movement of certain types of data, such as those containing personally identifiable information or classified data related to national security. But the government’s FedRAMP process, which establishes robust cloud security for storing government data, provides a viable cloud approach for federal employees or contractors.

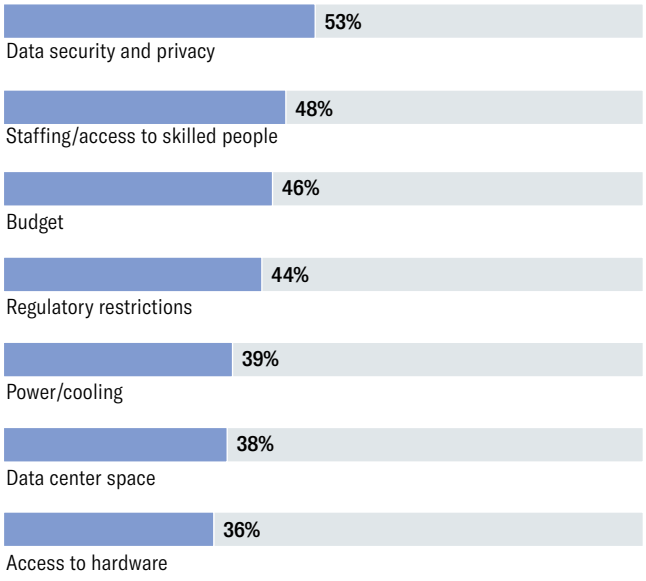
Beyond government work, privacy remains a critical concern for managing HPC workloads that involve data analytics and AI inference, a practice that employs a trained AI model to make predictions or solve research tasks. In a March 2025 survey of 393 business decision makers at midsize-to-large commercial enterprises, known as the “HPC and AI Budget Map,” by Intersect360 Research, a Mountain View, Calif.-based market research firm, security and data privacy concerns were rated as a major or insurmountable barrier by 53% of the respondents, outpolling other impediments

FIGURE 1

The Data Security and Privacy Barrier

Respondents say data security and privacy concerns stand most in the way of implementing artificial intelligence and high-performance computing

What are the barriers to implementing artificial intelligence and high-performance computing?



Source: Intersect360 Research survey, March 2025

such as insufficient talent or budget and regulatory restrictions.[3] FIGURE 1

When organizations, including government, academic, and business enterprises, need to scale their HPC systems, particularly to power large AI workloads, they turn to cloud service providers with available capacity, specifically for access to the latest GPUs. In addition to extra capacity, the public cloud offers test-bed capabilities, most notably a "try-before-you-buy" option.

"Many of us own a car but still take Uber or public transportation sometimes because it fits a particular use case," says Addison Snell, CEO of Intersect360 Research. "I think the better thing to look at with HPC and cloud is what capabilities it can give you beyond what you have on-premises. It can also be a sandbox for trying new technologies before you invest in them. There are new processors, new interconnects, and new storage coming out all the time. It's not practical for me to buy every combination to see what works. I might try them in the cloud and use that as a metric for what makes sense to live in the cloud or to purchase and have on-premises."

High-performance computing encompasses a vast spectrum of architecture, infrastructure, and use cases. An HPC system that runs financial risk analysis, genomics, or product design lacks the scale of an exascale HPC, which may conduct climate simulations, nuclear weapons modeling, and cosmological simulations. For example, a typical HPC cluster may include hundreds of nodes, while an NNSA exascale supercomputer called El Capitan has more than 10,000 nodes—each containing multiple central processing units (CPUs) and GPUs. Despite these apparent differences, even experts who run exascale simulations to test the condition of the U.S. nuclear weapons stockpile say that they're learning from how cloud service providers run HPC workloads.

Perhaps that's why before organizations commit to shifting HPC workloads to the cloud for compute, storage, or both, Kothe recommends that they initiate "a trial run of cloud resources to ensure the programming environment and scientific software stack that our applications rely on can be ported to the cloud and utilized appropriately."

Kothe suggests conducting a detailed cost analysis between on-prem and cloud-based prices for compute, data storage, and data transfer. "Currently, we find that on-prem resources are more cost-effective for many of our heavy-duty workloads, so the cloud provides an appropriate gap filler for some of our smaller and more moderate workloads," he explains.

Emphasis on Time-to-Science

HPC center teams are continually advancing their abilities to measure, manage, and optimize resource utilization both on-prem and in the cloud. Hoang hones her program's overall efficiency by improving the synergy between the NNSA's national laboratories, Sandia and Los Alamos in New Mexico and Lawrence Livermore in California, along with various manufacturing sites. Hoang reports that NNSA's HPC systems operate at a utilization rate of 95% or higher. The NNSA's compute capacity isn't limitless, however, and the lab scientists

> **“Currently, we find that on-prem resources are more cost-effective for many of our heavy-duty workloads, so the cloud provides an appropriate gap filler for some of our smaller and more moderate workloads.”**
>
> Douglas Kothe, chief research officer, Sandia National Laboratories

must apply for compute time to run HPC workloads on the program's systems.

Maximizing the value of HPC cloud workloads requires closely monitoring compute and infrastructure resources to eliminate performance bottlenecks and manage costs. Operations teams use orchestration tools to handle resources such as processors and memory and manage queues of pending jobs. They also track resource utilization across different CPU and GPU types while managing up to thousands of compute nodes and multiple interconnects. Increasingly, HPC teams are applying predictive analytics and machine learning to enhance resource demand forecasts, which enables better scheduling, cost optimization, and capacity planning both on-premises and in the cloud.

“There's always interest in optimizing HPC, and there's never enough budget to do everything that everybody wants to do,” says Intersect360's Snell. “That's the nature of not only HPC but research and development in general.” Snell believes organizations tend to overlook the importance of optimizing data before scaling their AI cloud workloads. He says the process of “holistically managing data over time” should include improvements in data movement, sovereignty, locality, and the overall data management stack.

Hyperion's Sorensen says that on-prem HPC centers with high system utilization rates may also have long average job queues—some can exceed multiple hours—which can be seen as intolerable from a researcher's perspective. “So, if you are looking at, say, a smaller job that may be amenable to going to the cloud and your key performance indicators are turnaround time, queue wait time, [and] the ability to run on different kinds of hardware, then you're going to favor cloud options,” he asserts.

For many workload placement decisions, the “real answer is on-prem working in concert with the cloud, but it depends entirely upon your workload,” adds Sorensen. “If you have a predictable workload with a relatively stable complement of hardware, yes, do it on-prem. But if you're only going to have computer requirements that use, say, 30% or less of a system over time, maybe you don't need to get in line for on-prem capacity. That said, the general consensus is that a progressive HPC center understands that you have to have an integrated capability, a combination of on-prem and in the cloud, if you want to take advantage of the strengths of both and avoid the weaknesses of both.”

A key performance indicator often emphasized in cloud HPC is time-to-science—the total time it takes a researcher to reach meaningful results, says Sorensen. Instant access to HPC cloud resources can cut the time-to-science dramatically, he adds, and it also eliminates unproductive wait periods for highly paid researchers. “If I can go to a cloud instance where my turnaround time is not hours or minutes, it's seconds, I could be much more effective,” he says, adding that HPC teams must prioritize between ensuring that “HPC machines are fully utilized or making sure that my most expensive asset, my scientists and engineers, are being used more effectively to deliver solutions.”

HPC-Artificial Intelligence Convergence

Some widely used LLMs, along with various open-source models, are created on HPC cloud systems because their sheer scale and complexity demand exceptional computational power. For example, an LLM may contain billions of parameters, entail massive calculations, and require vast data sets and storage capacity. Ordinary servers and clusters can't handle this load in a reasonable amount of time. Time is pivotal when the longer it takes to train an LLM, the sooner it becomes out of date. HPC infrastructure requires robust fault tolerance capabilities that can recover from failures without compromising the existing work.

AI-augmented HPC workloads run more efficiently on systems tuned for inference or training rather than on general-purpose HPC systems, explains Sorensen. “There are really two different worlds out there,” he posits. “Most high-speed general-purpose HPCs aren't as effective on training. Likewise, most of your training systems aren't as effective on

> **“AI might narrow 100,000 [prospects] down to 1,000 or a few hundred, and then simulation might narrow it down further to the dozen that are [the] most likely [candidates] before you get to the physical experimentation with the most likely prospects.”**
>
> Addison Snell, CEO, Intersect360 Research

traditional HPC workloads such as modeling and simulation applications that are kind of the stock-in-trade of the classic HPC environment.”

“We see AI for what you could term preprocessing, where you do the AI prior to modeling and simulation,” explains Snell. He cites the example of “target reduction,” in which a chemical engineering company might consider how best to design a new molecule but has “100,000 targets it might explore. It’s not feasible from a chemistry standpoint to wet lab all of those or even to simulate them all. But AI might narrow 100,000 [prospects] down to 1,000 or a few hundred, and then simulation might narrow it down further to the dozen that are [the] most likely [candidates] before you get to the physical experimentation with the most likely prospects.”

Running an AI model inside an HPC simulation can “dramatically increase the efficiency of the overall simulation,” says Andrew Chien, the William Eckhardt Professor of Computer Science at the University of Chicago. “Not in the traditional way one would think about it, which is to make that simulation go faster or make that loop go faster, but in the sense that AI can jump from some initial conditions to very close to [approximating] the solution. That’s the very kind of estimation or approximate answer construction you see AI being good at. Then they use the traditional methods to [generate a] very nice high-precision solution.”

AI can help optimize HPC scientific workloads in real time using a technique known as mesh refinement, Snell reports. AI can monitor a running simulation, such as a large-scale physics job, and “redirect computational resources” to where they are needed most. “With HPC jobs, generally you’re talking about running jobs in parallel, across many computational elements all at once,” he continues. “Inevitably, there are parts of a model that are busier than other parts, and sometimes you want to redistribute that work while it’s in flight if possible.”

AI enables other innovative methods on HPC platforms, such as reverse engineering a path to produce novel material, according to Sorensen. He says this example is akin to “backward planning. I want a material that has a certain weight, strength, and cost. AI can help you examine large databases to work backward to determine the material composition and production process that would get me there. That’s a nontrivial task in the HPC world.”

Though it is becoming a difference maker, Sorensen notes that AI generally isn’t directly written into existing HPC code, such as “nesting itself into a computational fluid dynamics program.” AI is functioning as “more of a recommender system,” he explains, that manages the flow of incoming and outgoing data and presents reasonable options for review by subject-matter experts.

“AI presents options to the subject-matter experts and says, ‘Here are the 10 most promising simulations,’” Sorensen says. “So, it’s more of a productivity enhancer for a researcher than it is something that says, ‘OK, just tell me what to do and I’m going to design you a new aircraft.’” Or, to put it another way, he adds, the HPC-AI convergence is “more about improving productivity between the traditional HPC computations than it is [about] diving down in and altering the simulations.”

Sorensen sees some limits to AI’s role in HPC value generation. At this point, trust is a gating factor. “The biggest reason for that is the inability, right now, to validate and verify why an AI program made a decision that it made,” he notes. “The inability to peel back the hood of AI, especially in generative AI—to ask, ‘Why did you make this decision?’—makes it very difficult for the classical HPC subject-matter expert to trust it implicitly. It is going to be a long road, and I think AI is always going to be used in conjunction with and as an advisor to subject-matter experts, as opposed to a replacement for them.”

Snell sees the potential but also would set boundaries for HPC and AI convergence. “I hope that that’s really, truly the path that we stay on for the majority of HPC workloads,” he says. “AI is going to be around—the genie is out of the bottle—but just like I would not recommend replacing all human decision making with analytics, I would not recommend displacing all physics-based simulation with AI predictions. There will be some areas where maybe that truly is good enough. But I think for a lot of scientific endeavors, what you want to see is a balance of AI and HPC, not AI replacing HPC.”

> **“I think one of our biggest learnings is that there are a lot of different cloud servers, and the key is to choose the ones that provide you the best return on investment, even if it’s not the fastest.”**
>
> Shay Rootman, vice president of business development, Cognata

Graphics Processing Unit Difference Maker

Long before GPUs containing multiple “cores”—or small processors—optimized for massive parallelism made a transformative impact on AI innovation, they were best known as graphics accelerators for gaming and computer-aided graphics. It was in this context that Cognata Ltd., an Israeli provider of digital twin simulation platforms for the autonomous vehicle (AV) market, first began using these GPUs. But the company’s GPU usage is evolving. “We aspire to take an artificial intelligence engine and train it to be a safer and more reliable driver across a wide range of situations and conditions,” explains Shay Rootman, vice president of business development at Cognata. Since its inception in 2016, Cognata has tapped HPC in the cloud to build and deploy city and off-road simulators, software that enables manufacturers of AVs and driver assistance systems to run complex driving scenarios in various virtual geographic locations.

Cognata’s simulations, which rely on GPUs for real-time rendering, sensor physics, and AI workloads, enable companies to bypass extensive road testing in the physical world. Simulations are preferable, Rootman explains, “mainly because driving on the road takes a very, very long time and it’s very, very expensive.” The simulated driving tests validate that the manufacturers’ algorithms are “robust enough to handle all types of driving, whether it’s in complicated driving weather or light conditions,” he notes. “What we’re offering, basically, is the ability to run millions of different scenarios.”

Cognata is in the process of rolling out a new AI-based HPC cloud service called DriveMatriX, which relies on customer data but changes “important elements of the weather and light conditions,” Rootman explains. “What we’re offering is to take your existing [road test] data and repurpose it for different conditions, such as cloudy and foggy, night and rain, and so on. These simulations are very heavy AI-related workloads.”

As a startup, Rootman says, Cognata never seriously considered investing “millions of dollars” in on-prem HPC capacity in its early years. Today the company primarily runs in the cloud and also supports on-premises deployments for customers that require it. “I think one of our biggest learnings is that there are a lot of different cloud servers,” he says, “and the key is to choose the ones that provide you the best return on investment, even if it’s not the fastest.” However, he adds, they’ve learned that “if you have a better GPU, you will get better results and better throughput.”

No matter the environment, HPC operations teams face a cascading set of decisions that involve balancing workload requirements against factors such as resource availability, time-to-solution, and projected costs. There are use cases that call primarily for CPUs or GPUs. If the workload and application code support dividing the research task into discrete yet similar calculations, they may opt for GPUs. On the other hand, CPUs provide advantages such as better management of large memory requirements or sequential processing—akin to steps that must occur in a prescribed order.

In an HPC system, a GPU becomes a “computational accelerator,” says Snell. “The types of neural networks that would feed into machine learning, AI training, and AI inference are also very well suited to this type of acceleration.” Yet, he adds, there are “possible trade-offs of managing HPC and AI in a shared environment, and it’s a reason why you might go to the cloud for some workloads.”

Snell explains that an HPC system optimized for pure AI processing may prioritize GPU speed over precision—the exacting 64-bit computational accuracy required for scientific workloads. On the other hand, some companies may “buy something on-premises that [they] think will serve all [their] needs,” he adds. “That becomes the cost-performance trade-off that anyone’s going to be looking at.”

There are numerous price and performance variables to balance, especially in the cloud, but Sorensen adds that “architecture shouldn’t be dictating workload. When you go to the cloud, you have a near-infinite possibility in terms of mixes of what kind of instance you want, how much GPU, how much memory, how much CPU,” he asserts. “Maybe you’re concerned more with low cost than you are with finding the

> **“The AI folks are benefiting from data center, interconnect, and cooling designs that were scaled up in the HPC world.”**
>
> Andrew Chien, William Eckhardt Professor of Computer Science, University of Chicago

right instance. Finding that right instance may be a nontrivial task because there are so many options.”

Trying different resources before you buy them is a well-trodden path to the HPC cloud. “One of the more attractive aspects of cloud resources is that they undergo tech refreshes, for example, providing new generations of GPUs, in a timely fashion,” Kothe explains. “These effectively enable us to exploit these resources as test beds in certain situations.”

Overall, Kothe says, “the CPU has become less and less important during a given application run. CPUs are still useful and needed for tasks such as off-processor communication, [input/output], and other bookkeeping tasks. That said, we’re seeing more and more intimate integration of CPUs and GPUs. In other words, an AI application running on a few GPUs in concert with an HPC application running on a few CPUs—all on the same node. That way, they can share data and interact with one another. The HPC application is sharing data with the AI model, and the AI model is sending inference answers back to the HPC application.”

Conclusion

The potent yet still emerging combination of HPC and AI in the cloud is broadening access to tools that can transform the breadth, speed, and scale of scientific, academic, and commercial research. AI isn’t merely optimizing HPC workload cost or performance management; it’s enabling scientists to make accurate real-time judgments that can speed time-to-science in research or, in business terms, time-to-market.

It’s no minor effort to muster the scientific talent and resources necessary to implement HPC on-prem or in the cloud efficiently. Not every project justifies the expense. A compelling business case for the resources might include climate modeling, Monte Carlo simulations, or an enormous data set that might take weeks or months of processing on standard hardware. Still, with HPC, it’s solvable in hours or days.

When researchers and market analysts speak of an HPC-AI convergence, it’s also a virtuous cycle that is upleveling both disciplines. “For the last 30 years, it turns out the HPC people have been building super-high-density computational systems, and so they pioneered at-scale water cooling, direct-to-chip cooling that is now the mainstream technology,” adds the University of Chicago’s Chien. “The AI folks are benefiting from data center, interconnect, and cooling designs that were scaled up in the HPC world.”

True to form, HPC advocates constantly seek the next breakthrough. At Congress’ request, the NNSA, which manages the U.S. nuclear weapons stockpile, issued a 120-page report in 2023 examining what will follow the current generation of exascale systems.[4] The report established a plan for a next-generation system emphasizing the importance of collaboration and partnerships, noting, “The roadmap should be explicit about traditional and nontraditional partnerships, including commercial computing and cloud providers, and academic, government laboratory, and broader cross-government coordination.”

But how do you build a next-generation system when innovation is happening at unprecedented speed and scale and when data center power consumption is placing enormous stress on the U.S. energy grid?

“I think we’re going to be surprised, possibly shocked, as to what’s possible in three to five years,” predicts Kothe. “With Big Tech in the U.S. driving for more and larger data centers, we at Sandia need to foster and grow public-private partnerships with the major cloud vendors to make sure our specialized workloads and associated requirements—such as using and creating sensitive data—are met moving forward.”

Endnotes

1 Hyperion Research, “Hyperion: HPC AI Market Grew 23.5% in 2024, to Exceed $100B by 2028,” April 2025. https://insidehpc.com/2025/04/hyperion-hpc-ai-market-grew-23-5-in-2024-to-exceed-100b-by-2028/.

2 Bob Sorensen and Tom Sorensen, “Cloud-Based AI Activity for HPC: Widespread but Primarily Exploratory,” Hyperion Research, September 2024. https://hyperionresearch.com/wp-content/uploads/2024/09/Hyperion-Research-Special-Report-AI-in-the-Cloud-September-2024.pdf.

3 Addison Snell, Antonia Maar, Kevin Jackson, et al., “Intersect360 Research HPC and AI and Budget Map,” March 2025. https://www.intersect360.com/.

4 National Nuclear Security Administration, “Charting a Path in a Shifting Technical and Geopolitical Landscape: Post-Exascale Computing for the National Nuclear Security Administration 2023,” April 2023. https://nap.nationalacademies.org/catalog/26916/charting-a-path-in-a-shifting-technical-and-geopolitical-landscape.

HARVARD BUSINESS REVIEW

VISIT US ONLINE

hbr.org/hbr-analytic-services

Harvard Business Review Analytic Services is an independent commercial research unit within Harvard Business Review Group, conducting research and comparative analysis on important management challenges and emerging business opportunities. Seeking to provide business intelligence and peer-group insight, each report is published based on the findings of original quantitative and/or qualitative research and analysis. Quantitative surveys are conducted with the HBR Advisory Council, HBR's global research panel, and qualitative research is conducted with senior business executives and subject-matter experts from within and beyond the *Harvard Business Review* author community. Email us at hbranalyticservices@hbr.org.

CRE73800126