Special Analysis

What's In and What's Out for the Global HPC-AI Community in 2026

Mark Nossokoff, Jaclyn Ludema, Tom Sorensen, Earl Joseph, and Bob Sorensen
January 2026

HYPERION RESEARCH OPINION

Breaking tradition somewhat from past predictions, Hyperion Research's 2026 prognostications are embedded within a What's In/What's Out perspective on a diverse range of important areas across the global HPC-AI community. In addition to the high market growth rates of HPC-AI (23% in 2024, close to 20% in 2025, and 2026 looks to be growing in the 15% plus range), these are our predictions on the major changes taking place in the market.

In this special report ten different areas of changes are analyzed in the systems, AI, quantum computing, sustainability, cloud, storage and workforce development areas.

What's In for 2026 includes:

- Deployment of multibillion-$USD, giga-watt scale leadership-class HPC and AI supercomputers driven by broad public/private partnerships.
- Updating system partitions every two to three years to take advantage of new technologies.
- End user focused architectures and data sets for AI training and inference.
- Full integration of AI capabilities into US federal agencies and organizations' mission critical activities, including space exploration, national defense, and intelligence.
- Rapid advancement of QC capabilities across many fronts leading to a potential watershed year of demonstrated use cases.
- Advances in Shor's algorithm design that shrink the time frame before QC systems will be able to break current public-private encryption schemes.
- Data center energy strategies that directly fund new renewable power generation to match actual local electricity consumption.
- Broad adoption of continuum computing into users' overall advanced technical computing infrastructure planning to appropriately leverage on-premises, cloud, and edge computing.
- Data platforms designed to accommodate the heterogeneous workloads of traditional HPC workloads and modern AI workloads, including training, inference, and agentic.
- AI tools partially mitigating the mounting difficulty in attaining qualified and long-lasting advanced computing talent within the industry.

WHAT'S IN AND WHAT'S OUT IN 2026

Table 1 shows What's In and What's Out for ten different areas in 2026.

TABLE 1

What's "In" and "Out" for 2026

Area	"In"	"Out"
Market Dynamics: Extremely Large Systems	Deployment of multibillion-$USD, giga-watt scale leadership-class HPC and AI supercomputers driven by broad public/private partnerships	Government-led and funded leadership-class systems limited by power and budget constraints
More Complex System Upgrades	Updating system partitions every two to three years to take advantage of new technologies	Purchasing large single systems and keeping them for 5 to 6 years
More Focused AI Training	End user focused architectures and data sets for AI training and inference	Wholesale use of "one size fits all" generative AI solutions
AI Becoming Mission Critical	Full integration of AI capabilities into US federal agencies and organizations' mission critical activities, including space exploration, national defense, and intelligence	US federal agencies and organizations' adoption of AI only for experimentation and lower-stakes activities such as office tools, administration, archiving, and IT
Quantum: Real Use Cases	Rapid advancement of QC capabilities across many fronts leading to a potential watershed year of demonstrated use cases	Adoption of QC for primarily experimental use cases by the earliest quantum-curious organizations
Quantum: Breaking Encryption Schemes	Advances in Shor's algorithm design that shrink the time frame before QC systems will be able to break current public-private encryption schemes	Casual, non-urgent approach for transitioning to a quantum safe cybersecurity regime
Sustainability Drives New Renewable Investments	Data center energy strategies that directly fund new renewable power generation to match actual local electricity consumption	Relying on "paper credits" (EACs) to offset carbon emissions on a balance sheet without changing the actual power source
Cloud: Continuum Computing Takes Hold	Broad adoption of continuum computing into users' overall advanced technical computing infrastructure planning to appropriately leverage on-premises, cloud, and edge computing	Biased assessment of cloud capabilities and incomplete accounting of on-premises costs when evaluating workload requirements
Storage: Data Platforms Take Hold	Data platforms designed to accommodate the heterogeneous workloads of traditional HPC workloads and modern AI workloads, including training, inference, and agentic	Storage systems tailored primarily for homogenous traditional HPC workloads focused largely on large-block, sequential checkpoint/restart operations
AI Helps With Workforce Development	AI tools partially mitigating the mounting difficulty in attaining qualified and long-lasting advanced computing talent within the industry	Increasing dearth of early career and junior HPC and AI advanced computing operational talent

Source: Hyperion Research, January 2026

Market Dynamics: Extremely Large Systems

In: Design and deployment of multibillion $USD, giga-watt scale leadership-class HPC and AI supercomputers driven by broad public/private partnerships

Out: Government-led and funded leadership-class systems limited by power and budget constraints

Traditionally the most powerful supercomputers in the world have been chiefly driven and funded by global governments, largely in the US, Japan, and Europe. Many of the most recent machines in this class have been capped at 50 MW and within hundreds of millions $USD budgets.

Recently, with the introduction of AI factories, such as the Stargate program, and the Genesis Mission initiative, investments in the billions $USD could soon become commonplace. Such large investments are possible by the establishment of public/private partnerships between government and industry.

More Complex System Upgrades

In: Updating system partitions every two to three years to take advantage of new technologies

Out: Purchasing large single systems and keeping them for 5 to 6 years

Research has shown that over the last several years users have been extending the life of their on-premises systems from five to six years, largely driven by budget constraints. One consequence of this trend, as HPC vendors are increasingly accelerating the technology refresh rates of their product roadmaps from 18-24 months down to 12 months, is that users increasingly find themselves multiple generations behind with their technical computing infrastructure and capabilities.

Future procurement intentions appear to signal more frequent, less expensive investments to purchase capabilities targeted at specific workloads while allowing them to keep up with the latest computing advancements and innovations.

More Focused AI Training

In: End-user focused architectures and data sets for AI training and inference

Out: Wholesale use of "one size fits all" generative AI solutions

Current generative AI trends, particularly in the area of large language models (LLMs), draw heavily on proprietary trained models such as OpenAI's GPT, Anthropic's Claude, and Google's Gemini. These training models typically require large training data sets, reaching into the petascale range, consisting of large-scale multilingual datasets of books, articles, web pages, academic papers, and licensed sources. In addition, current training requirements can reach upwards of 1025 floating point operations, requiring training sessions that can last for months.

Increasingly, AI end users will turn towards smaller training models, drawing in a wide availability of open source LLM models to be trained on focused technology, company, or sector specific data that can be orders of magnitude smaller than the proprietary counterparts. These reduced data sets can significantly drive down the training compute requirements while ultimately providing more relevant targeted responses. Such capability also enables more frequent updates to the training process, supporting more current outputs during the inference stage.

AI Becoming Mission Critical

In: US federal agencies and organizations' full integration of AI capabilities into mission critical activities, including space exploration, national defense, and intelligence

Out: US federal agencies and organizations' adoption of AI only for experimentation and lower-stakes activities such as office tools, administration, archiving, and IT

Examples supporting this include:

- DARPA, IARPA, NASA, and defense groups like the USAF are increasingly putting forth research and pilot programs to integrate AI capabilities deeper into their traditional activities.
- The U.S. executive branch has indicated a strong commitment both towards the innovative success of the national AI market as well as its integration into federal workflows.
- Numerous large purchases from mainstream AI tool providers indicate confidence in ongoing future collaboration, for example Google's 'Gemini for Government' plan.

Quantum: Real Use Cases

In: Rapid advancement of QC capabilities across many fronts leading to a potential watershed year of demonstrated use cases as the sector progresses from NISQ-based systems to early fault tolerant error correction systems

Out: Adoption of QC for primarily experimental use cases by the earliest quantum-curious organizations

Examples supporting this include:

- The September 2025 announcement that HSBC, one of the world's largest banking and financial services organizations, and IBM, a leading superconducting QC supplier, applied quantum computing to an algorithmic bond trading scheme that achieved a 34% improvement in predicting trade outcomes compared with traditional classical methods.
- The October 2025 announcement that BlueQubit, a QC software developer, and Quantinuum, a leading trapped ion QC system provider, had collaborated to produce a verifiable quantum advantage demonstration that could be run efficiently on a QC system but extremely time consuming for a classical counterpart. This work represented one of the first attempts to provide a practical verifiable demonstration of quantum advantage, moving beyond the theoretical claims or non-verifiable random circuit sampling experiments common today.
- Multiple announcements from QC vendors including Alice & Bob, Photonic, and QuEra citing significant gains in quantum error correction efficiency using Low-Density Parity-Check (LDPC) versus the currently more used surface codes to lower qubit count requirements. These firms report that although surface codes are mathematically straightforward and locality-friendly, they require a large number of physical qubits to support one logical qubit, and that ratio can rapidly scale with efforts to increase overall logical qubit fidelity. Conversely, LDPC codes can offer greater efficiency (10-20x fewer physical qubits per logical qubit) but demand more complex hardware to implement the required long range qubit interactions, making them a promising option for future, scalable quantum computers despite their complexity.

Quantum: Breaking Encryption Schemes

In: Advances in Shor's algorithm design that will continue to shrink the time frame before QC systems will be able to break current public-private encryption schemes

Out: Casual, non-urgent approach for transitioning to a quantum safe cybersecurity regime

In 2019, leading researchers for Shor's algorithm implementation estimated that it would take 8 hours using 20 million noisy qubits to factor 2048-bit RSA integers using estimates of then current QC hardware capabilities. Advances in algorithmic design have significantly lowered the requirement for the 2048 RSA integer benchmark, and most recently, in 2025 a team from Google postulated an approach that, using the same QC hardware specifics as those from 2019, could theoretically factor a 2048-bit number with less than a million qubits. Despite the increase in the job's runtime from about 8 hours to 1 week, the technique offered a 20X reduction in the number of physical qubits needed and about a 100X improvement in the number of quantum gates. Additional enhanced capabilities at the QC hardware level will continue to only increase the efficiency of the algorithm.

Organizations across the entire IT ecosystem, but especially those with national security workloads or those that have critical proprietary or personal information, should more actively monitor developments in both Shor's algorithm as well as new post quantum encryption schemes currently under development. Although there are a number of algorithms that have already been standardized by the US government's NIST for general use, the daunting process of seamlessly deploying such new-end-to-end encryption schemes could realistically take years, and any organization's planning process to do so should be put in place sooner rather than later.

Sustainability Drives New Renewable Investments

In: Data center energy strategies that directly fund new renewable power generation to match actual local electricity consumption

Out: Relying on "paper credits" (EACs) to offset carbon emissions on a balance sheet without changing the actual power source

Historically, Energy Attribute Certificates (EACs) have served as a primary mechanism for data centers to claim, "carbon neutrality". This approach allowed facilities to procure renewable attributes equivalent to their consumption, regardless of the physical power mix of their local grid. This model reached a peak in 2025, which saw a massive proliferation of EAC utilization across multiple industries as organizations moved to meet interim Environmental, Social, and Governance (ESG) targets.

However, the sheer scale of AI-driven power demand is rendering this "accounting-first" approach obsolete. The disconnect between a clean balance sheet and a dirty local grid is becoming too obvious for regulators and stakeholders to ignore. The focus is shifting toward "24/7 carbon-free energy" (CFE) matching. This methodology prioritizes time-aligned and location-based procurement, ensuring that energy consumption is matched by renewable generation on the same grid at the same time. While EACs remain a component of the market, the 2026 outlook suggests a pivot where HPC/AI centers will be increasingly evaluated on their ability to stimulate "additionality" (bringing new renewable capacity online locally), rather than solely offsetting consumption through global certificate markets.

Cloud: Continuum Computing Takes Hold

In: Broad adoption of continuum computing into users' overall computing infrastructure planning to appropriately leverage on-premises, cloud, and edge computing

Out: Biased assessment of cloud capabilities and incomplete accounting of on-premises costs when evaluating workload requirements

Users will increasingly embrace the concept of "continuum computing," seamlessly integrating cloud resources with on-premises infrastructure when planning for future procurements and fulfilling RFPs and grant proposals. This hybrid approach allows organizations to optimize their computing resources based on workload requirements and desired outcomes.

The ability to dynamically allocate resources between cloud and on-premises solutions will likely enhance operational efficiency and agility. This trend may also lead to advancements in orchestration tools that facilitate smooth transitions between different computing environments, ensuring that organizations can adapt quickly to changing demands.

Storage: Data Platforms Take Hold

In: Data platforms designed to accommodate the heterogeneous workloads of traditional HPC workloads and modern AI workloads, including training, inference, and agentic

Out: Storage systems tailored primarily for homogenous traditional HPC workloads focused largely on large-block, sequential checkpoint/restart operations

The importance of storage solutions within the overall architecture of modern advanced technical computing systems has continued to grow in importance with the evolution and dominance of AI

workloads. Storage solutions have morphed into data platforms to shift from supporting primarily homogenous I/O profiles to being capable of high performance heterogenous workloads. Beyond performance and the "-abilities" (reliability, availability, serviceability, usability, installability, durability), data platforms have and will become much more content-aware, particularly relative to LLMs and agentic AI.

Several vendors have brought their implementation of data platforms to market in recent years (e.g., DDN, Hammerspace, NetApp, Weka, Pure, VAST). While there is ample market opportunity to support multiple data platform vendors, given the business attractiveness and importance of the market, acquisitions, finance rounds, and IPOs in this area may increase.

AI Helps with Workforce Development

In: AI tools mitigate the mounting difficulty in attaining qualified and long-lasting advanced computing talent

Out: Increasing dearth of early career and junior HPC and AI advanced computing operational talent

The challenge of building, acquiring, and retaining talent for the operation of advanced computing systems such as HPC or advanced AI centers has been a mounting issue across many industries for some time. While the fundamental contributing factors of this challenge remain and, in some cases,

are continuing to grow, advancements in AI capabilities such as code and text generation or data management allow for junior positions to be shrunk, eliminated or turned into higher performing employees.

While these solutions offer continued relief for those who can effectively implement and operate them, they present a new challenge in the form of the disruption of the traditional junior position to senior position pipeline on which industries rely to ultimately populate the most important roles in their ranks.

Decision makers implementing AI tools to replace junior positions would do well to maintain some level of expertise training infrastructure to ensure a future supply of those with advanced knowledge and experience of their operations to avoid facing challenges in the future brought on by shifts in workforce (e.g., retirement) and overdependency on AI. In addition, the new AI tools can make existing employees more productive, which helps address the employee shortage.

Future Outlook

2026 is expected to be a year of change in the advanced computing, HPC and AI for science world, with many changes in what's in and what's out in systems, AI, quantum computing, sustainability, cloud, storage and workforce development.

Many large new AI systems will likely come online in 2026 and a number of new quantum systems will be available to a broad set of researchers, providing the tools to explore innovative ways of approaching science, engineering and advanced computing.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology, and related trend analysis, and both user and vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors, and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.hpcuserforum.com and www.HyperionResearch.com
