Special Analysis

Top 10 Predictions for the Global HPC Community in 2023

Mark Nossokoff, Bob Sorensen, Jaclyn Ludema, Melissa Riddle, Tom Sorensen, and Earl Joseph

February 2023

HYPERION RESEARCH OPINION

For 2023 and beyond, Hyperion Research makes these predictions for the worldwide HPC market:

1. Strong growth in the leadership-class segment will support modest growth across the global on-premises HPC market.
2. The advanced computing sector and its associated supply chain will become increasingly driven by national and regional government policies that stress domestic capabilities.
3. Sustainability and energy efficiency considerations will become a dominant factor in many procurements.
4. Cloud utilization will shift from predominantly experimentation to predominantly production workloads as users gain familiarity and confidence with cost, performance, and workflow integration expectations. This will include the initial erosion of on-premises spending in the low end of the HPC market.
5. 2023 will be the year of AI regulation.
6. AI will become more pervasive in production tier deployments due to users' higher confidence in its abilities and it becoming much easier to use.
7. HPC system architectures will bifurcate between systems optimized for one set of applications and those designed to address a myriad of applications.
8. The divergent requirements of traditional HPC modeling/simulation and AI workloads will move HPC architectural focal points from compute to system interconnects and storage systems.
9. Industry sector interest in edge computing for HPC will rise in 2023 sector, but spending will be muted.
10. Growth at many HPC sites in new technology investment areas such as AI, Edge, and alternative HPC architectures will be stunted in 2023 due to the continued difficulty in acquiring and retaining talent and expertise.

TOP 10 PREDICTIONS ANALYSES

1. Strong growth in the leadership-class segment will support modest growth across the global on-premises HPC market.

The on-premises HPC market will be led by strong growth in the leadership-class segment. Multiple pre-exascale and exascale-class machines are projected to be accepted around the world in 2023. The following table details the 5-year on-premises global HPC market forecast of 6.4% overall, and 6.8% for HPC servers.

TABLE 1

On-premises Revenues by the Broader HPC Market Areas

	2021	2022	2023	2024	2025	2026	CAGR 21-26
Server	$14,781	$15,668	$16,901	$18,650	$19,240	$20,585	6.8%
Storage	$5,985	$6,494	$7,098	$7,989	$8,378	$9,072	8.7%
Middleware	$1,733	$1,815	$1,934	$2,110	$2,145	$2,282	5.7%
Applications	$4,960	$5,156	$5,455	$5,898	$6,005	$6,352	5.1%
Service	$2,272	$2,253	$2,276	$2,354	$2,308	$2,309	0.3%
Total Revenue	$29,731	$31,385	$33,664	$37,000	$38,076	$40,599	6.4%

Source: Hyperion Research, January 2023

2. The advanced computing sector and its associated supply chain will become increasingly driven by national and regional government policies that stress domestic capabilities.

National policies stressing sovereign high-tech computing capabilities will come to the fore in 2023. For example, concerns with growing dependence on foreign sources of the most advanced semiconductor components, partially from Taiwan and South Korea, resulted in both EU and US programs to bolster domestic semiconductor production capabilities. Specifically, the US CHIPS and Science Act's $52.7 billion investment in domestic semiconductor manufacturing intended to reduce the likelihood that shocks abroad might disrupt the supply of chips, boost American international economic competitiveness, and protect semiconductors from being sabotaged in the manufacturing process.

- This announcement followed the announcement of a European Chips Act with the same intent and similar implications.

- Observers of these initiatives noted that such efforts were needed primarily to counter existing government semiconductor technology promotion policies already in place in China, South Korea, and Taiwan.

Likewise, the quantum computing sector will continue to be an emerging market typified by strong government support that spans direct R&D support, targeted government procurements, export controls, and workforce development. Many governments already have QC technology and commercial promotion programs in place, several with a stated agenda to build a viable domestic QC supply chain. For example, the European High-Performance Computing Joint Undertaking (EuroHPC JU) recently announced a 2023 €100 million procurement budget for six sites to host what could be the first major round of EU government-sponsored quantum computer (QC) procurements. The QC hardware and software for this effort will draw exclusively on EU technology developed under EU-funded quantum initiatives, related national programs, and private investments.

- Similar high-value government QC procurements in the EU and elsewhere could be a major, if not dominant, source of QC sector revenues in the next few years, playing a key role in determining the overall trajectory of the QC supplier base writ large.
- The expansion of national security driven export controls will continue to drive realignments with the sector. For example, recent US controls metering the international flow of advanced semiconductors as well as related design software and manufacturing equipment has resulted in increased efforts by Chinese chips and systems suppliers to rely more heavily on domestic components that are beyond the control of foreign political efforts. One of the efforts targets the open source RISC-V chip architecture which has the advantage of circumventing US export controls on leading-edge proprietary US chip technology.
  - The Chinese Academy of Sciences is currently developing a RISC-V processor, and a number of leading Chinese IT suppliers, including Alibaba Cloud, Huawei, ZTE, and China's state-backed Institute of Computing Technology Huawei are adopting RISC-V variants for their chip designs.
  - Initial RISC-V technology will likely first appear in low-power devices, such as smart phones, but specialized RISC-V chips could also find their way into next generation Chinese HPCs.

3. Sustainability and energy efficiency considerations will become a dominant factor in many procurements.

Several factors are raising sustainability and energy efficiency as priorities for HPC users as they plan for their future procurements:

- New CPUs/GPUs/xPUs are requiring substantially more power. Each new generation of the most powerful compute elements that are typically architected into HPC systems are requiring more power to deliver their full performance capabilities.
- Energy costs. The cost of energy has increased dramatically around the world, close to doubling in some locations. Large variations in energy costs highly dependent on geographic location, combined with recent inflationary trends, are creating an even wider disparity of energy costs worldwide. At some HPC sites, energy budgets are not currently available to provide the power necessary for their machines to deliver their full capabilities.
- Energy consumption legislation. Recent legislation in several countries is pressing for more energy efficient data centers. Implementation of these new laws is focused on energy-conscious procurement standards, use of renewable energy where possible, and optimizing workloads for efficiency.

- Environmental stewardship. While difficult to quantify, users are placing various metrics and requirements on their vendors to achieve meaningful improvements in responsible HPC environmental stewardship. Goals include lowering overall energy consumption and reducing global carbon footprints.

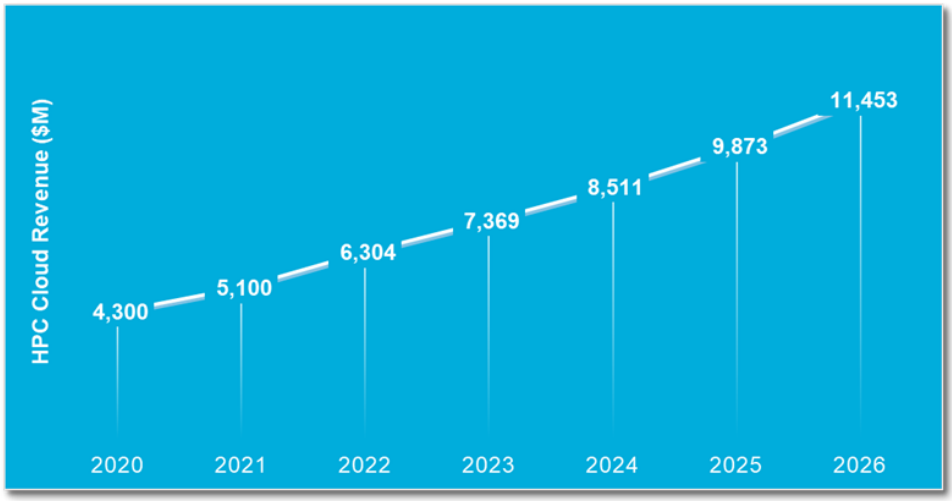
4. Cloud utilization will shift from predominantly experimentation to predominantly production workloads as users gain familiarity and confidence with cost, performance, and workflow integration expectations. This will include the initial erosion of on-premises spending in the low end of the HPC market.

The HPC cloud market continues to be one of the strongest growth markets in today's HPC ecosystem. Historically, user sites treated cloud resources primarily to dynamically address peaks in workload demand and experimentation. Over the past two years that paradigm has changed dramatically. Hyperion Research has conducted several studies that suggest cloud is becoming a major component in future resource planning for a broad range of HPC user sites. Many sites are now weighing cloud next to on-premises procurement strategies and altering the size and timing of future on-premises deployments to increase their budget for cloud resources.

As a result, cloud providers and on-premises vendors alike are working to provide tools, support, and capabilities to ease access to cloud resources. While the HPC on-premises market is not projected to dramatically change from the steady 6%-8% five-year CAGR seen in years past, the cloud market is showing a projected 17.6% over the forecast period. Within the forecast period, end user spending on cloud resources for HPC workloads is expected to grow to be larger than every segment of the broader HPC market except the server portion of the market. The use of HPC cloud resources is causing a fundamental shift that will alter the future of the HPC market in the coming decades.

FIGURE 1

HPC Cloud Forecast, 2021-2026



Note: Spending represents user spending to run HPC workloads in the cloud, as opposed to CSP spending to provision resources for HPC workloads.

Source: Hyperion Research, 2023

5. 2023 will be the year of AI regulation.

Governments and policy making bodies of all sizes have set their sights on AI technology. Automobile safety regulation groups, the US Copyright Office, state governments, even public schools, have started introducing limitations or stipulations on how AI technology can or should be used. Additionally, there are high-profile international undertakings such as the UK's AI Standards Hub and 'six AI principles', as well as the EU's AI Act, which is reported to be the first law on AI by any major regulator.

- The majority of these endeavors, be they laws, lawsuits, rules, or guidelines, seek to both enable forward-looking innovation while protecting the privacy, agency, and dignity of users and beneficiaries of the technology.
- The mounting efforts to regulate AI technology is a clear indicator of its increasing adoption among commonplace solution spaces as well as an increased exposure to individuals not in the advanced technology sphere.

6. AI will become more pervasive in production tier deployments due to users' higher confidence in its abilities and becoming easier to use.

AI technology is no longer cloistered in a lab or experimental environment. Hospital systems, financial services, military branches, and organizations spanning the gamut of commercial and industrial spaces, are increasingly introducing AI applications into their workflow. This represents the growing confidence not only in the benefits the technology can introduce to an organization, but in its trustworthiness and stability.

- As the pool of high-quality AI expertise, product offerings, and interactive tools grow, the availability and viability of production-ready programs will increase and become more pervasive.

Consumer-facing AI applications have already captured the attention of the public. Creative tools, lifestyle and health management, personal finance, and many more spaces are all currently being transformed by the growing number of functional AI tools available for general consumption.

Small and medium businesses not typically associated with advanced technology will adopt frameworks or tools which leverage AI technology to aid in nearly every vector of business. Retail, IT, business operations, inventory, and manufacturing are all elements of business that can currently be aided by existing AI technology. In accordance with the 'low-tech' status of many of those businesses and individuals who will seek the benefits of AI, these tools will need to feature an increased focus on ease-of-use, trustworthiness, and accountability.

7. HPC system architectures will bifurcate between systems optimized for one set of applications and those designed to address a myriad of applications.

As HPC researchers continue to venture into more diverse and complex applications to solve the problems in their field and push for faster time to solution or higher accuracy of work, new approaches are being folded into their workload mix. One key area of new adoption is in AI methods, both as standalone applications and in conjunction with traditional simulations, pushing users towards a wider variety of hardware options to handle the new requirements. Many AI applications run adequately on traditional CPU-based systems but can experience high performance gains by incorporating accelerators, faster interconnects, different memory schemes, and software improvements.

Future system designs for HPC users will have to factor in these new requirements, highlighted by the rise in accelerator, and specifically GPU, adoption in HPC. In the near future, there will be a large

segment of the HPC ecosystem that will be deploying more complex systems, incorporating various technologies intended to accelerate specific subsets of the workloads being run at the site.

Many sites will choose to deploy multiple different systems, each architected for a specific type of application, and each distinctly different from the others, whether through the choice in processors and accelerators, the interconnect scheme, or various other architectural options.

As systems become more diverse, users will have to be keenly aware of resource management and allocation to ensure applications are running on the most efficient parts of future systems. Understanding software complexities and, in many cases, refactoring code to take advantage of accelerators or different architectural choices in a system will be beneficial despite the effort needed at the front end.

It will be crucial to understand the right distribution of technologies to handle the anticipated composition of workloads, both established workloads and emergent applications. Working closely with vendors on system design and proof of concept tests will be critical in ensuring future systems optimize application performance.

8. The divergent requirements of traditional HPC modeling/simulation and AI workloads will move HPC architectural focal points from compute to system interconnects and storage systems.

Computing elements (e.g., CPUs, GPUs, FPGAs, ASICs) have evolved to address the demanding and divergent requirements of HPC and AI workloads. System interconnects and storage systems are now becoming the differentiation flashpoint of HPC system architectures.

Connectivity of elements within a node will demand greater attention as composability of system resources becomes more widespread. Composability will be adopted to provide more flexibility of HPC systems to be re-configured as different workloads and workflows are run on the system over time.

System interconnects are becoming more critical to address how to connect composable elements of the system (compute, memory, storage), both internal and external to the compute nodes. External system networks that provide server-server and server-storage connectivity are largely dominated by InfiniBand and ethernet. Both will continue to evolve to overcome network congestion caused by heavy system utilization supporting diverse compute-intensive and data-intensive workloads, as well as improving predictability and duration of workload completion times.

Storage systems once optimized for bandwidth performance will evolve to support the performance needs of both compute-intensive and data-intensive workloads. Different approaches will emerge to address the challenges presented by disparate demands such as IO profiles (large block sequential vs. small block random), access methods (file vs. block vs. object), access frequency (hot vs. archive vs. cold) and locality (centralized on-premises vs. cloud service provider vs. edge).

9.Industry sector interest in edge computing for HPC will rise in 2023, but spending will be muted.

Edge computing is continuing to draw attention as a potential addition to HPC ecosystems. According to recent survey data, a quarter of users (28%) either currently employ edge computing or expect to within 1-2 years. This figure was highest in the industry sector, where more than a third (37%) of surveyed HPC users reported plans for edge computing. The main perceived benefit among these users was improvement to real-time data collection and processing capabilities. Other prominent

attractors included accelerating HPC applications, access to IoT devices for data collection, and supporting a wider range of sensor data.

The raw data being generated from a growing array of globally distributed data sources (e.g., vehicles and traffic sensors, large science experiments, medical devices, product manufacturing lines, and military sensors) is becoming more complex and at such a scale that moving it for processing and analysis is often not broadly economically feasible. Computing at the edge provides the advantage of addressing data scalability and offers the opportunity to perform a first pass computational look at the data and reduce it to a manageable amount to be shared with a primary on-premises HPC datacenter for simulation, training, and inference.

However, edge computing continues to have challenges that are preventing it from becoming ubiquitous in the near future. HPC users are concerned about the complexity of incorporating the edge and acquiring the talent required to integrate the edge (see Prediction #10 below). Many users are hoping that edge vendors will increase their support to lower barriers for new entrants. Cost is also a consistent concern, as is uncertainty about privacy or safety. Throughout 2023, Hyperion Research expects that several HPC sites will experiment with and invest funding in edge, but production-level usage will remain limited.

10. Growth at many HPC sites in new technology investment areas such as AI, edge, and alternative HPC architectures will be stunted in 2023 due to the continued difficulty in acquiring and retaining talent and expertise.

HPC experts are an aging workforce. The attrition of HPC staffs due to retirement or semi-retirement is increasingly becoming an issue for many HPC sites. Hiring and onboarding replacements while maintaining continuity in the center's functions is a top priority, but this is becoming more difficult as the pipeline of new HPC staff entering the workforce does not adequately match the outflow of retirees. Meanwhile, HPC expertise demands are also increasing at a fast rate.

The number of HPC user types is increasing as traditional enterprise IT users are entering the HPC space, often lured in by AI and big data. The staffing and expertise required at existing sites is also expanding. Many HPC sites plan to purchase more HPC systems and diversify them by incorporating different technologies such as multiple processor types, co-processors, accelerators, and other specialized hardware. In addition, most HPC sites are using the cloud part of the time. As HPC sites incorporate more variety and increase the number of systems running, it becomes less likely that the same number of HPC staff can manage everything.

Together, these forces mean that demand for expertise to efficiently run such varied setups is far outpacing the supply. This will lead to intensifying competition when it comes to hiring HPC staff. Areas with relatively fewer HPC experts may find it harder to attract foreign employees, especially amid generally ongoing uncertainty about future global travel restrictions. Sites with limited budgets (such as in academia) may be especially hard-hit if they cannot offer competitive salaries.

2022 SCORECARD

How did the Hyperion Research 2022 predictions fare? With as much humility and objectiveness as could be mustered, here is an introspective self-assessment:

TABLE 2

2022 Predictions Scorecard

#	2022 Prediction	Grade	Commentary
1	The strong market rebound in 2021 will carry into 2022 as more buyers look to HPC solutions to address new opportunities and compete more aggressively.	A	HPC growth in 2021 was 9.3% while 2022 technical computing growth is currently projected to be 6.4%, in large part due to Frontier being accepted at ORNL.
2	The global exascale rollout will officially begin with the acceptance of the first US exascale system.	A	Frontier was formally accepted at ORNL in December 2022.
3	2022 will be a high growth year for HPC cloud computing, building upon the fundamental changes in buying behavior seen in 2021.	A	2022 actual revenues are not, yet, available, but 2022 is projected to be a high growth year.
4	Artificial intelligence (AI), along with big data, will become far more effective and pervasive, creating new applications and solutions that greatly impact many sectors.	A	A majority of HPC sites are leveraging AI platforms for some application. Beyond that, AI use cases have expanded and all user types from the advanced tech to consumer grade are using it.
5	The squeeze on advanced semiconductor production for the HPC sector will worsen in both the short and long terms.	A	New semiconductor supply chain investments will require years to have any effect. Ups and downs in demand continued to fuel supply chain issues.
6	HPC system architectures will bifurcate between systems optimized for one set of applications and those designed to address a myriad of applications.	C	In planning for their next procurements, bifurcation is being explored at many sites, however new deployments in 2022 did not bear this out.
7	The divergent requirements of traditional HPC modeling/simulation and AI workloads will move HPC architectural focal points from compute to system interconnects and storage systems.	C	This prediction was premature. While architectural planning increased around interconnects, compute and acceleration requirements remained at the center of procurement decisions.
8	HPC solutions will be used in many new areas, including supporting the high growth in edge computing.	B	HPC solutions are being increasingly adopted in many newer areas, while edge deployments, are occurring at a less modest rate.
9	Global competition and tensions in leadership-class HPCs will intensify.	A	Growing government policies stressing indigenous supply of HPCs, targeted government procurements, and secure domestic HPC supply chains.
10	The growing scarcity of HPC expertise to implement new technologies and integrate them	A	Scarcity of talent is noted almost universally as a top challenge globally. There is evidence that difficulty in

TABLE 2

2022 Predictions Scorecard

#	2022 Prediction	Grade	Commentary
	into the higher complexity of systems will become the number one roadblock for many HPC sites.		identifying and attracting talent is stunting the pace of adoption of new technologies.
11	Interest in quantum computing will grow but demonstrated use cases will still be needed.	A	The QC market grew by >20% in 2022 as more QC early adopters found ways to use QC systems.

Source: Hyperion Research, 2023

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
