HYPERION RESEARCH

Special Report

Perspectives from SC24

Mark Nossokoff, Bob Sorensen, and Earl Joseph
December 2024

HYPERION RESEARCH OPINION

With approximately 18,000 attendees and over 500 exhibitors, including more than 130 new exhibitors, SC24 in Atlanta surpassed organizer expectations and delivered a vibrant and energetic environment for the global HPC community to exchange ideas while establishing or extending collaborative partnerships and relationships. As is our custom, the Hyperion Research team of analysts has compiled its primary takeaways and perspectives from more than 80 customer meetings, workshop and session attendance, and a myriad of other interactions at the event:

- There were two new entrants in the Top10 of the Top500 list, including a new #1, El Capitan at Lawrence Livermore National Lab (LLNL), but the list continues to stagnate.
- Speculation abounded regarding what impact, if any, the new administration may have relative to US government spending on science, technology, and innovation.
- The disproportionate share-of-wallet demanded by GPUs is impacting end users' budgeting processes.
- Conversations continue regarding the impact on industry investment in AI and lower precision workloads will have on mixed and full precision investments within the traditional modeling and simulation space.
- Approaches to multiple dimensions of data management within AI workflows appear to be targeted as differentiators for AI system and storage providers.
- Tier 2 cloud providers, defined nominally as AI clouds, GPU clouds, or GPU as a Service, have emerged as a force in multiple aspects of provisioning resources for AI workloads.
- Investments in and innovations surrounding liquid cooling and related energy and power management appear to be approaching mainstream with a growing presence within the HPC and AI communities.
- Quantum/classical computing integration is seen as the next critical step towards quantum utility.
- Generative AI, particularly large language models (LLM), are receiving much attention, but their considerable compute requirements create challenges for all but the largest and best funded organizations.

PERSPECTIVES AND TAKEAWAYS FROM SC24

There were two new entrants in the Top10 of the Top500 list, including a new #1, El Capitan at Lawrence Livermore National Laboratory (LLNL), but the list continues to stagnate.

El Capitan at the US Department of Energy's (DOE) LLNL was announced as a new #1 on the Top500 list, coming in with an HPL score of 1.742 Eflop/s, displacing DOE's Frontier at Oak Ridge National Laboratory (ORNL). El Capitan's architecture employs a combined 11,039,616 CPU and GPU cores based on AMD 4th generation EPYC processors (24 cores at 1.8GHz and AMD Instinct MI300A accelerators).

- El Capitan also leapfrogs Aurora at DOE's Argonne National Laboratory, questioning whether Aurora can ever achieve #1 on the Top500 when it is fully accepted at its maximum configuration.

Also new to the Top500 is HPC6, installed at Eni, the largest petroleum company in Italy. Employing the same architecture as Frontier at ORNL, HPC6 came in at #5 and is now the fastest system in Europe.

That said, the Top500 list continues to demonstrate that maintaining historical gains in HPC performance is an increasingly costly, complex, and power hungry process. Indeed, Top500 experts indicated that it may take until the end of this decade to see a Top500 system exceed the 10 Eflop/s metric on the Linpack benchmark.

Speculation abounded regarding what impact, if any, the new administration may have relative to US government spending on science, technology, and innovation.

While there was much discussion on the issue, there didn't appear to be a consensus on whether it would be positive, negative, or neutral.

- Attendees postulated both that the scale of the FASST program could be reduced and the proposed Department of Efficiency could positively impact certain areas while re-evaluating government priorities.
- That said, the previous Trump administration was supportive of both cyber and emerging technologies to modernize defense capabilities. This time around, however, Silicon Valley could hold significant sway in overall high-tech policy making.

The disproportionate share-of-wallet demanded by GPUs is impacting end users' budgeting.

End user investments in HPC-AI infrastructure are being driven by organizations adopting and deploying AI. With the current high costs of accelerated computing, particularly GPUs, expected to continue into the foreseeable future, users are challenged to make the most efficient use of their budgets to gain access to as much accelerated computing as they can.

- They are embracing the cloud for access to supply-constrained GPUs and being forced to allocate any extra available funds towards GPUs they would otherwise prefer to spend elsewhere.

- This may open the door for cheaper, less power consumptive, more focused AI-centric accelerators, particularly those offered by hyperscalers including AWS, Google, and Microsoft.

Conversations continue regarding the impact of industry investment in AI and lower precision workloads on mixed and full precision investments within the traditional modeling and simulation space.

Uncertainly and concern exist on whether there will continue to be advancements relative to features and performance for double precision workflows with vendor investments increasingly being directed towards AI-related areas.

- Vendors suggest investments in FP64 will continue while also supporting performant FP64 emulation with lower precision implementations.
- The trend may, however, may entice some HPC legacy code users to consider converting their FP64 codes to mixed precision support potential performance gains.

Approaches to multiple dimensions of data management within AI workflows appear to be targeted as differentiators for AI system and storage providers and hyperscalers.

Multiple dimensions of data management, such as explainability, governance, traceability, context, flexibility, and tagging, were highlighted by various vendors and service providers as key features in their AI-related infrastructure and service solutions.

- A lack of international consensus on AI regulations likely will complicate these issues significantly.

While some of these aspects are expected to be mandatory due to national compliance initiatives, others are being developed to provide additional perceived business value benefits for users or to avoid legal issues such as copyright and IP protection violations.

Tier 2 cloud providers, defined nominally as AI clouds, GPU clouds, or GPU as a Service, have emerged as a force in multiple aspects of provisioning resources for AI workloads.

A growing class of users do not have the need for the breadth of services or broad selection of compute instances to choose from. This recognition has led to a growing number of service providers focused solely on accelerated computing services consisting of a limited range of bare-metal instances and reduced adjacent service offerings focused solely on AI relative to the full service cloud providers.

Plans for converting to liquid cooling options are being widely considered: everyone is either doing AI or cooling AI.

A growing amount of real estate on the exhibition floor was consumed by cooling, energy, and power management solution vendors. Both an increase in the number of these vendors, as well as larger booth sizes, contributed to the perceived increase in their market presence or at least an increase in end user interest.

- Liquid cooling, both direct-to-chip and immersion implementations, are now recognized as required for leading edge HPC and AI infrastructure solutions and vendors are scrambling to establish technology leadership and market share.
- Conversion can be costly, so change may come more slowly than these vendors anticipate.

Quantum/classical computing integration seen as next critical step towards quantum utility.

Quantum utility, considered the stage where the use of a quantum computer (QC) produces results superior to that when produced by a classical computer, is seen as being in reach within the next three to four years. A growing and increasingly sophisticated QC ecosystem is supporting progress in the field from all directions spanning new qubit designs and manufacturing processes, efforts to support multiple quantum processor architectures, a maturing software stack that supports both software development toolkits as well as third party applications offerings, and perhaps most important, activities to integrate advanced classical computers with quantum counterparts.

Generative AI, particularly large language models (LLMs) are receiving much attention, but their considerable compute requirements create challenges for all but the largest and best funded organizations.

Only a handful of the largest and best funded organizations can commit to train LLMs at their current scale, which can require total flops counts that exceed 10^{25} total operations, a requirement that could consume months on the most powerful HPCs in the world today. In contrast, mainstream generative AI users typically train at 3-4 orders of magnitude lower flops counts and conduct multiple in-house LLM training sessions, trading off model size for model precision and custom understanding. LLM end users are looking to implement efficient, targeted small language models to:

- Reduce computational complexity.
- Ease the requirements for large and sometimes unverified data sets.
- Produce more focused sector, disciplined, or company specific LLMs.

FUTURE OUTLOOK

SC24 was a clear success from many perspectives:

- Record attendance and number of exhibitors providing a broad range of HPC- and AI-related products and solutions.
- Wide breadth, depth, and expertise of topics shared at conference sessions.
- Vibrant and productive conversations, both formally and informally, across a diverse set of venues and networking events.

The global HPC community took full advantage of being able to congregate and collaborate to provide meaningful opportunities for continued advancement in technological innovations. Scientists, engineers, and researchers across government, industry, and academia are directly involved in these conversations and are looking forward to advanced tools and capabilities to accelerate their research and discovery for solutions to the world's most challenging issues.

Attention now turns to SC25 in St. Louis, which is expected to be much more vibrant than when St. Louis hosted the first post-covid SC. Speculation suggests exhibitor space for SC25 may be already sold out.

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 & 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.HyperionResearch.com and www.hpcuserforum.com
