HYPERION RESEARCH

Special Report

Perspectives from SC23

Mark Nossokoff, Bob Sorensen, Tom Sorensen, and Earl Joseph
December 2023

HYPERION RESEARCH OPINION

Surpassing pre-pandemic attendance and achieving new attendance and exhibitor participation records, SC23 in Denver, CO did not disappoint. With over 14,000 on-site attendees and over 435 exhibitors, SC23 provided a vibrant, high-energy 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 the event:

- There were four new entrants in the top 10 of the 62nd edition of the Top500, including one which was only half-configured (Aurora at Argonne National Lab), a major CSP (Eagle at Microsoft), a specific semiconductor provider (Eos at NVIDIA), and the Mare Nostrum 5 at the Barcelona Supercomputer Center.
- LLMs generated a significant amount of interest in its potential to accelerate HPC algorithms and use cases, but questions about the rapidly evolving technology remain, especially for running scientific problems.
- Many HPC end users are actively exploring the challenges and opportunities of integrating quantum computing capabilities into their current and anticipated traditional HPC compute environment.
- Research efforts within the traditional modeling and simulation space to explore the computational advantage of emerging HPC architectures that support mixed precision operation have begun in earnest.
- NVIDIA announced it will be accelerating the cadence of its product release to 12 months, applying increased pressure on the competition to keep up.
- SC23 was not the only major industry event last week. Microsoft's annual Ignite event was happening elsewhere concurrently, at which they announced their internally developed AI chips.
- Liquid cooling in general, and immersion cooling specifically, is emerging as a cooling option but not yet achieving the point of being broadly adopted and deployed beyond a few computing sites.
- An increased number of BoFs, panel discussions, and presentations suggest the global HPC community is moving to address increased diversity within the work force, as well as promoting educational awareness and opportunities for individuals to enter and be successful in the industry.

Top500

The Top 500 list showed some uncharacteristic churn this year: it added four new systems and one upgrade in the Top 10 slots, most notably the long-awaited premier of the Aurora system at DoE's Argonne National Lab at the number two spot, albeit operating on only half of its eventual total system configuration. Noting that the system needed almost 25 MW to run the LINPACK test, questions about the power consumption of the fully configured machine remain. Two new entrants, Microsoft and NVIDIA, representing non-traditional HPC suppliers, also appeared on the Top 10 list at number 3 and 9 respectively. The newly formed Atos spin off, Eviden, entered the Top 10 list for the first time at number 8 with its installation of the Mare Nostrum 5 at the Barcelona Supercomputer Center in Spain.

Overall, the list continued its recent trends toward an overall slowdown in performance gains, with the Top 500 curators projecting that the 10 exaflops barrier likely will not be broken before the end of this decade. Finally, the schism between leadership class HPCs and mainstream HPCs that comprise the bulk of the sector continues to widen. The top seven systems now have the same total compute power as the remaining 493 sites on the list, an historic low and well below the high point for this metric at 90, two decades earlier.

LLMs Generated a Significant Amount of Interest in Its Potential to Accelerate HPC Algorithm and Use Cases, but Questions about This Rapidly Evolving Technology Remain

The potential of Large Language Models (LLMs) and related generative AI techniques was a much talked about topic at the conference, but in reality, the HPC sector has only begun to explore the challenges and opportunities of this emerging capability as a potential accelerator for the wide range of existing HPC applications and use cases. The next 1.5-2.5 years will be important in the development of LLM technology and its role in scientific computing. With a significant number of users on the precipice of mission critical LLM integration and exploring new applications at a rapid pace, the fulfillment of these expectations remains a question.

LLM's have already reached usefulness beyond what could be considered their most logical application area, but how much further could it go? Some questioned the use of LLMs and AI at SC23 for doing actual science, and discussed the issues related to it not following the traditional scientific approach. Organizations that manage to maintain relevant AI literacy, make decisions based realistic expectations, and an openness to cost-friendly innovation stand a good chance of coming out ahead in the rush to adopt these new technologies.

Integration of Quantum Accelerators with Classic HPC Architectures Are Accelerating

HPC end users, eager to catch the next major opportunity to accelerate their most vexing computational problems, are taking their first major steps toward assessing the challenges and opportunities of integrating quantum computing capabilities into their overall classical HPC compute environment. Attendance on the exhibition floor at the "Quantum Village', the collection of quantum computing vendors, was always brisk, and the handful of BoFs that dealt with the technical realities of becoming an early QC adopter were well attended.

This attention was amplified by the recent announcement that the first EU-based exascale system, JUPITER, bound for installation at the Jülich Supercomputing Center in Germany in 2024, will include a quantum computing partition alongside its CPU-centric and GPU-centric counterparts.

Opportunities for Performance Gains in Traditional Modeling and Simulation Codes Using Mixed Precision Hardware Is Gaining Momentum

AI algorithms and their specific hardware implementations have long relied on the use of mixed or low precision operations to speed computational performance through more effective use of available hardware while reducing overall data bandwidth and storage requirements. Albeit still in their early stages, SC23 offered up several research efforts that examined the value of using similar techniques to improve the performance of more traditional modeling and simulation codes. Noting that there are cases where high precision, typically 64-bit floating point, numbers may not be required for some portions of an overall computational modeling or simulation run, some researchers are looking at ways to rewrite codes or even revamp their underlying algorithms to take advantage of new hardware support for execute-mixed precision run, seeking the same kinds of performance gains realized in many AI-based codes.

NVIDIA Is Increasing Competitive Pressures with an Accelerated 12 Month Release Cadence for Its Products

NVIDIA announced they will be accelerating the release cadence of its AI-focused GPU product roadmap to 12 months. While this provides potentially increased barriers for its competitors and is intended to support continued acceleration of scientific advancement and discovery, it also introduces some potential challenges and stresses within the broader HPC ecosystem:

- System vendors will need to be very judicious with their system roadmaps regarding how quickly they can absorb and integrate new technologies into their development, test, and validation investments and processes.
- Users who recently have been extending the lifetimes of their on-premises HPC systems from 4 years to 5 or 5.5 years could fall 3 or 4 generations behind leading edge innovations for their installed systems.

Microsoft Announced Its Internally Developed AI Chips for Azure

Microsoft announced the upcoming availability of services based on two internally developed chips: Maia, focused on accelerating AI inferencing, and Cobalt, a consumer Arm-based chip aimed primarily at general-purposed computing.

While Maia is not intended to be direct competition for NVIDIA's H100, it is expected to be competitive with other vendors' solutions, providing Microsoft more opportunity to optimize its performance and infrastructure stack, and allow it to diversify its supply chain.

Immersion Cooling Gaining Traction in Broad HPC Cooling Ecosystem but Challenges Remain

While several vendors (e.g., Lenovo) and cloud service providers (e.g., CGG and DUG) are promoting their innovations and success with providing and implementing immersion cooling solutions, it is not yet broadly adopted beyond specific use cases. This is despite the large presence of immersion cooling providers exhibiting on the SC23 show floor. Several factors inhibiting the broader adoption of immersion cooling include availability of a limited range of immersion-ready server options (necessitating the need to make physical changes on-site to standard servers to convert them to

immersion-ready) and modifications required to existing data centers to take full advantage of immersion cooling benefits. Factors driving the increased utilization of all forms of liquid cooling include the steady increase in the power requirements for leading-edge GPUs and CPUs.

Increased Attention and Activity Towards Diversity in the Global HPC Workforce and Educational Awareness & Opportunities

A dearth of available talent has continually been cited as one of the largest challenges faced today by both HPC users and vendors. Many employees in the HPC workforce are approaching retirement. At the same time, university students and new college graduates appear to be less interested in traditional HPC modeling and simulation and classical HPC technologies. And many of the emerging AI experts don't immediately recognize the relationship between HPC and AI and thus don't consider themselves as HPC end users or express interest in it.

Based on an increased number of BoFs, panel discussions, and presentations on the topic, the global HPC community appears to be moving to address increased diversity within the work force, as well as increasing educational awareness and opportunities for individuals to enter and be successful in the industry. This is further evidenced by graduating cohorts from the EUMaster4HPC program (a pan-European Master of Science (MSc) program in High Performance Computing) and unique and creative initiatives being employed by industry.

FUTURE OUTLOOK

SC23 was a clear success from many perspectives:

- Record attendance, surpassing pre-pandemic levels
- Record 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.

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
