HYPERION RESEARCH

Special Report

Perspectives from SC22

Mark Nossokoff, Bob Sorensen, Alex Norton, Jaclyn Ludema, and Earl Joseph
December 2022

HYPERION RESEARCH OPINION

With over 11,000 on-site attendees approaching pre-pandemic levels, and only 700 virtual participants, SC22 in Dallas, TX far exceeded the high expectations of the broad HPC community. There was no shortage of avenues for participants to obtain the latest knowledge relative to market developments and technology innovations occurring across the industry. The Hyperion Research team of analysts has compiled its primary takeaways and perspectives from the event:

- Little has changed within the Top500, with only one new entrant in the Top 10 and continued absence of any new submissions from China.
- Sustainability and energy consumption has emerged has a key strategic item for user procurement and vendor investments, driven by environmental stewardship initiatives, rising energy costs, and emphasis on TCO, all leading to increasing conversations around the Green500.
- Running HPC applications in the cloud continues to increase for more workloads, as evidenced by virtually all major HPC system vendors offering some hybrid-based control plane and pay-as-you-go service offerings, as well as by several major HPC sites moving largely to the cloud.
- Architectural focus is evolving to address a sector that once relied on custom-built supercomputers running primarily modeling and simulation jobs into one that is a growing assortment of related yet distinct areas of interest.
- Quantum computing is starting to gain mindshare within the classical HPC community as users become more interested in the potential of QC to improve the performance of existing classical HPC applications while opening opportunities in new QC-based ones.
- Some sites in active procurement evaluations are considering delaying their procurements due to perceived and anticipated near-term delays (from 12-18 months) with pending new technology, combined with concerns about the economy.
- Users are expressing confusion in their understanding of and relative positioning of NVIDIA's Omniverse and Meta's Metaverse.
- A dearth of available talent continues to be a major challenge for users and vendors alike.

Top500

Little has changed within the Top500. Leonardo (an Atos BullSequana XH2000 machine at EuroHPC/CINECA in Italy) was introduced at #4. There were only 40 new entrants overall in the November 2022 list. This is after only 40 new entrants were introduced in June at ISC22, reflecting a significant slowdown in new entrants compared to prior years. A primary cause of the slowdown is due to China (which had consisted of approximately 50% of Top500 systems in prior years) continuing to refrain from submitting new systems. The lack of new Chinese participation led to many SC22 observations and commentary regarding the relevance of the Top500 and how it may evolve without continued Chinese presence.

Evidence outside of the Top500, however, confirmed the existence of production Chinese exascale-class machines. Talks given by major Chinese HPC universities at several SC22 conference sessions disclosed system-level node and chip configurations. Details of these machines and the workloads being run were shared in the context of large language and transfer model trainings.

Sustainability and energy consumption

Sustainability could be the next critical HPC performance metric, standing equally along with price and performance. The growing global emphasis on HPC site sustainability could drive HPC suppliers and users to look beyond simple metrics like PUE and instead take a more holistic look towards reducing the carbon footprint of a HPC system as well as its data center. This shift in emphasis could drive a realignment in the design and assembly of large HPCs that may be efficient from a GFlops/watt metric but do not adequately address overall sustainability issues including reduced carbon footprint or inclusion of renewal energy sources.

System vendors and CSPs alike are highlighting ways their respective HPC solutions are addressing higher power demands of key system elements in an environmentally acceptable fashion. With energy costs rising around the world, in some cases 100% over the last year, sustainability is also a strong budget consideration as well. Users are looking at wattage of new components and trying to understand the tradeoff between performance gains and power increases as new technology is deployed (e.g., is a 4X performance increase worthy of a 2X power increase?)

TCO

TCO is becoming a more broadly used performance metric that users are or will be requiring in new RFPs. The cost of running the machine over its intended lifetime is becoming as important, if not more important, than the initial CAPEX of the infrastructure. Large machines may be on an unstainable trajectory of CAPEX, power OPEX, or other OPEX for many sites.

Inflation has exacerbated this and, in some cases, has caused a near doubling in energy costs, making power consideration more prominent in architecture and procurement planning and decisions as the rising energy costs is negatively impacting how much funding is available to be spent on new infrastructure.

Green500

Users and vendors both express concerns relative to the dramatic growth of system power consumption. This is driven by both environmental stewardship and energy consumption perspectives.

As a result, vendor positioning on the Green500 is becoming almost as important as performance on the Top500.

That said, concern has been expressed that the Green500 doesn't adequately reflect the scale of a system. For example, the Frontier “test” system appears above the complete Frontier system on the list. The list may not accurately reflect the value of large-scale systems and may need to be more representative, which could require a new metric alongside the current Gflops per watt.

HPC in the Cloud

Users are continuing to run more HPC workloads in the cloud, largely in response to CSPs providing more appropriate HPC capabilities to address these complex workloads. Inquiries are emerging relative to how the cloud can be leveraged for data access and storage, not just as a place to aggregate data, but also potentially for complete archiving.

Another factor driving increased cloud utilization for HPC workloads is executive mandates. Evidence suggests the decision to move to the cloud by several leading industry HPC datacenters has been by executive directive and not from an objective evaluation performed by HPC and general enterprise datacenters. At the other end of the spectrum, some sites are reluctant to move any of their HPC workloads to the cloud, citing security (e.g., NIST certification requirements) as a key reason to stay 100% on-premises.

System vendors continue to provide additional support for hybrid cloud environments. Traditional system vendors each have their own pay-as-you-go service and hybrid cloud access methods (e.g., HPE Greenlake, Dell APEX, Lenovo TruScale, Penguin Sycloud Central Software), and new partnerships and collaborations are also emerging (e.g., NVIDIA's and Microsoft's AI cloud).

Architectural Areas of Focus

HPC has evolved from a sector that relied on custom-built supercomputers running primarily modeling and simulation jobs into one that is a growing assortment of related yet distinct areas of interest, each offering a diverse set of opportunities and challenges to both HPC suppliers and end users. Examples include:

- Cloud vs. on-premises deployments
- Modern AI and big data vs. traditional engineering modeling and simulation workloads
- Classical HPC systems vs. emerging quantum computing
- Proprietary IA and ARM instruction set architectures (ISA) vs open RISC-V ISA
- Traditional Fortran programming vs. modern (e.g., Python) software perspectives
- Bare metal/container applications vs. homogeneous, monolithic vs. heterogenous, composable architectures
- CPU vs. general-purpose GPU vs. custom AI accelerator compute options

The sector is no longer one-size-fits-all as the room under the HPC umbrella continues to embrace new technologies, architectures, applications, and end users.

Maturing conference themes also reflects this evolution. The SC14 theme was HPC Matters, revealing an almost defensive community justifying its existence and relevance in the overall IT environment. Contrast that with the SC22 theme of HPC Accelerates, which showcased HPC's proliferation into new technology and important new use cases.

HPC progress was also reflected in several talks addressing the intersection of machine learning and traditional modeling and simulation. HPC has developed into a much broader scope, enveloping AI and revealing a symbiotic relationship between AI and mod/sim.

Quantum Computing

Quantum computing is starting to gain mindshare within the classical HPC community as users become more interested in the potential of QC to improve the performance of existing classical HPC applications while opening opportunities in new QC-based ones. HPC suppliers and end users are aware of the risks associated with any new technology, especially one that involves mastery of a completely new and somewhat challenging compute paradigm and will move cautiously. However, the HPC community writ large understands the need to be well positioned to embrace the technology once QC performance gains become readily demonstrated for key HPC applications.

Contrasting quantum computing sessions and conversations between November's SC22 and June's ISC22 suggests different regional and governmental approaches and motivations for investment in quantum computing:

- Europe: developing programs designed to create a capable domestic QC user base in key European sectors such as automotive, aerospace, and biosciences
- US: bifurcating investments between defense priorities and creating a vibrant commercial QC sector
- China: focusing primarily on national security agenda, but including some targeted research at CSPs and individual professors at leading academic institutions

Market Impact of Technology Delays

Several site directors running active procurement evaluations suggested they are seriously considering delaying their purchases due to anticipated near-term (~18 months) delays in technology availability. Typical on-premises machine lifetimes are approximately 4 years with recent studies suggesting a trend towards 5 years. Site directors realize their on-premises infrastructure will need to last at least this long, and they are willing to defer significant investment to obtain required infrastructure based on the latest technology.

This trend could bode well for CSPs looking to capture increased mindshare and market share for running HPC workloads. One of the primary benefits users of HPC cloud resources cite is access to the latest technology innovations as they are available in the market without having to commit to the high CAPEX.

End User Confusion Between NVIDIA Omniverse and Meta Metaverse

NVIDIA's Omniverse is being confused with the Metaverse from Meta. The Omniverse offering from NVIDIA is designed to be a digital twin platform based on both AI and traditional physics-based models. The Meta offering is essentially an artificial reality/virtual reality offering to provide users an opportunity to interact in a virtual world. However, there is confusion between the names, adding issues to understanding the specifics of the NVIDIA offering when compared to the Meta/Facebook offering.

Talent Gap

The talent gap identified in recent studies is continuing to worsen. Many in the HPC practitioner workforce are aging and approaching retirement. At the same time, fewer university students and new

college graduates want to do traditional modeling and simulation and prefer other areas. 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.

What is encouraging in this area are the conversations and glimpses of progress being made relative to diversity and inclusion. Attendance and participation in The Women in HPC workshop was standing-room-only, and there was also a full day track on the topic.

FUTURE OUTLOOK

By all accounts, SC22 was a clear success:

- Attendance reaching near pre-pandemic levels
- Quantity and quality of technology on the exhibition floor
- Breadth, depth, and expertise of topics shared at conference sessions

High expectations were set for the coming months relative to availability of new products, solutions, and innovations across all HPC sectors of industry, government, and academia. Scientists, researchers, and engineers are anxious to see their expectations turn into reality.

The entire HPC community is also looking forward to fulfillment of what vendors will deliver, the science and engineering breakthroughs enabled by the hardware and software innovations, and the progress to be made relative to the on-going sustainability, power, and talent challenges.

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
