The Future of High-Performance Computing

Becoming the Driving Force

Data has become the driving force behind business, academic, and social progress, forcing significant advancements in computer processing. By 2025, an estimated 463 exabytes of data will be created each day globally.[1] As institutions embrace a "data everywhere" mentality, high-performance computing (HPC) presents new opportunities to take on emerging challenges in these fields.

HPC arose as a discipline in computer science in which supercomputers are used to solve complex scientific problems. As HPC technologies grow in their computational power, other academic, government, and business institutions have adopted them to meet their own needs for fast computations.[2]

Today, HPC vastly reduces the time, hardware, and costs required to solve mathematical problems critical to core functions. Now an established field for advanced computing, HPC is driving new discoveries in astrophysics, genomics, and medicine, among other academic disciplines; it is driving business value in unlikely industries such as financial services and agriculture as well.

In time, HPC will permeate organizations of all sizes and converge with other advanced computing fields—including cloud computing, AI, machine learning, and even quantum computing—as it becomes a critical element of the industrialized digital world. These connections promise new advancements in learning and productivity, but also new challenges in competitive environments. As HPC evolves and the barriers to adoption disappear, organizations of all sizes will find HPC applications that will shape our collective future.

> "By 2025, an estimated **463 exabytes of data** will be created each day globally."

Understanding HPC: General Requirements, Designs, and Configurations

Unlike traditional computers—whose performance metric is 'million instructions per second' (MIPS)—high-performance computers use 'floating-point calculations,' which are unique operations within each computation instruction. HPC performance is measured in 'floating-point operations per second' (FLOPS), where a leading commercial supercomputer (a type of HPC system) can perform over one-hundred quadrillion FLOPS.[3]

HPC requires the linking of computers to process data in parallel, enabling those systems to achieve higher computing power. To do this, HPC employs clusters of servers, or "nodes," that work in parallel to perform complex calculations. There is no specific number or range of clusters defined to constitute HPC; rather, HPC is classified simply as the aggregation of computing power in this way to solve a given problem. A small group of clusters can be as few as four nodes, and a large cluster can be thousands of nodes.

In each node there is an HPC processor. The processor may use multiple cores in a CPU or a GPU, where GPUs are particularly well-suited to simulations, imaging and modeling; it may also have accelerator cards using GPUs which are more efficient in processing graphics and images. More cores mean more power, and more energy consumed. Multi-core designs were created to split up workloads, requiring that cores communicate with one another. Compute and storage servers have also been networked into clusters to communicate.

Components and servers must keep up with each other. Storage or memory components must feed and ingest data in fast succession, while network components must keep up with the transport of the data between servers, storage, and locations. Without considering these basic requirements and designs, higher levels of HPC performance cannot be realized.

Recent advancements in parallel computing are driving performance improvements in clustered systems, where "high-performance computing with Linux-based clusters are more preferred to build parallel systems that will act as core for next-generation supercomputers," the Academic Journal of Management Sciences reports.[4]

Some enterprises connect large servers in a single cluster, "where the total physical size, energy use or heat output of the computing cluster might become a serious issue," according to Gartner.[5] In fact, "a single data center can take more power than a medium-sized town," The New York Times reports, where an investigation revealed "data centers can waste 90 percent or more of the electricity they pull off the grid."[6] Traditionally, these on-premise HPC comprise hundreds of server racks in data centers and corporate campuses where they are constantly active.

How is HPC Used Today?

HPC is deployed on premise, at the "edge" (i.e. near the data source), or in the cloud. HPC will increasingly be deployed in multi-cloud and hybrid (public and private) environments as well. HPC remains popular in academic and scientific settings and is increasingly used to solve problems in diverse business sectors—such as automotive, finance, economics, and agriculture, among others.
In one modern example, a global energy company uses an AI-driven HPC platform to improve decision making during the early stages of hydrocarbon exploration. The complex task incorporates large amounts of "geological, geochemical, and physical data" to determine likely hydrocarbon deposits. This "cognitive discovery" process aids in the contextualization of information before drilling can begin.[7]

The rapid computational of HPC is helping authorities make life-saving decisions as well. After a careful evaluation process defining its HPC infrastructure, one organization uses HPC to predict the destructive weather patterns of hurricanes before landfall. This allows them to alert local authorities in a timely way, savings lives.[8]

In a very different use case, an ethernet-based vehicle connectivity system featuring a multi-zone architecture delivers a "complete high-performance computing solution" that utilizes machine learning to enhance diagnostic, security, and user-experience features for connected autonomous vehicles. The HPC solution processes data and "provides secure and reliable interconnects" across the entire vehicle network.[9]

In summary, there are applications for HPC in any discipline where computational power and speed are of value. HPC is finding a home in fields as diverse as climate science, digital animation, nuclear physics, and satellite image analysis, as examples.

Now, "The demand for performant and scalable AI solutions has stimulated a convergence of science, algorithm development, and affordable technologies to create a software ecosystem designed to support the data scientist."[10] As we will find, new advancements make it increasingly easier for non-data scientist users to leverage the value of HPC as well.

HPC Converges with Artificial Intelligence

The ease of adoption and use of artificial intelligence (AI)—including machine learning (ML) and deep learning (DL) technologies—is driving a convergence between these technologies and HPC. As Intel observes, "AI can solve problems by 'learning' from supervised and unsupervised examples rather than depending on a formulized set of equations or rules. For this reason, the role of a combined HPC and AI solution offers new capabilities and opportunities."[11] Rapid advancements in this area are forcing scientists, academics, and IT decision makers to reevaluate their HPC best practices and identify new use cases that offer performance benefits through technology integrations. As insideHPC reports:

"AI is now becoming practical to deploy at scale... This means HPC scientists can work locally on fast workstations and small clusters using optimized software. If required, HPC scientists can then run at scale on a large supercomputer... to solve leadership-class scientific problems using AI. Thus, it is critical to have both optimized and scalable tools in the AI ecosystem."[12]

In time, "AI will become one of the top workloads that drive infrastructure decisions" through 2023, according to Gartner[13]. AI will permeate the enterprise with touchpoints among both technical and business users. Using AI to augment human decision making or even make decisions autonomously will help in the scaling and delivery of HPC-like performance, even for personnel without backgrounds in data science. As a result, HPC infrastructure that can facilitate the convergence of AI and related data analytics technologies has become paramount:

"AI needs a lot of computational horsepower under the hood, along with an architecture designed for lightning-fast processing, bottleneck-free I/O and the capacity to handle huge datasets. To get there, forward-looking enterprises are rolling out high-performance computing (HPC) infrastructure built and optimized for the challenges of AI workloads, including machine learning, deep learning and high-performance data analytics."[14]

In this way, HPC is critical as we transition from a world where machines simply run calculations to one in which they help with more informed, complex decision making at scale. Decision makers will increasingly rely on these technologies to process enormous datasets and make their own decisions as well—even in three-dimensional environments, as is the case with autonomous vehicles.

But as much as these developments are critical to future computing requirements, experts emphasize the focus remains on data as an opportunity and means for driving value, where software and hardware need to "get out of the way… Major AI software efforts are focusing on assisting the data scientist so they can use familiar software tools that can run anywhere and at scale on both current and future hardware solutions." As we will find, cloud computing offers new solutions for scaling and managing HPC workloads.

How is HPC Used Today?

HPC often calls for a great many cores and clusters, which can be expensive to acquire. Since most organizations can't afford large on-premise supercomputers they choose alternatives, such as HPC cloud computing, less intensive cloud computing services, or bare metal services. Cloud computing is often spoken of in terms of virtualization, but that should not be confused with hyper-converged infrastructure (HCI). HCI is focused on the virtualization of computing, storage, and network resources. Because those resources are virtual instead of physical, they are commonly referred to as being part of a "software defined" architecture—the focus is on allocating resources rather than computational intensity such as in data center applications.

Now, cloud-computing architectures are changing computing as HPC clouds reach equivalency or become superior to on-premise HPC systems in terms of performance. The nearly instant elasticity of the cloud helps organizations scale up for really large projects, then easily scale back down again for smaller tasks.[15]

Proactively integrating cloud-hosted resources with HPC solutions offers unique advantages, including:

- Maximizing OPEX efficiency by curating cloud-friendly HPC licensing strategies
- Optimizing productivity using multi-dimensional HPC cloud selection criteria
- Minimizing challenges associated with cloud adoption by integrating HPC middleware[16]

In fact, organizations that fail to integrate HPC and cloud strategies may face infrastructure delivery challenges and be at a competitive disadvantage as a result. As Newill observes, organizations needn't compromise when moving HPC workloads to the cloud: "You can get the same performance as on-premises, but with all of the benefits of the cloud." He goes on to note that while a 50,000-core on-premise HPC cluster would take two months to build, deliver, and activate, a 50,000-core cloud HPC cluster can be activated with a signature, allowing organizations to schedule millions of jobs that very day.[17]

HPC Architecture and Workloads of the Future

Next-generation HPC is defined by faster interconnections, higher computing densities, scalable storage, higher efficiencies in infrastructure, eco-friendliness, space management, greater security, and increased workload performance. Future HPC architectures will have more powerful compute nodes, use more cores and accelerators, and employ newer memory and cloud-based I/O technologies to handle more intensive workloads and scale. They will feature improved administration and control, response times, environmental hazard protection, and both space and waste management solutions, minimizing organizational friction and environmental impact as well. Next-generation architectures will focus on making applications work better. As such, application porting will have a profound influence for adopting computing architectures beyond x86.

The AI infrastructure market will reach $50.6 billion by 2025, where both HPC accelerators and AI markets (e.g., hardware, natural language processing, context-aware computing) are key players.[18] Fortunately, new HPC products and solutions can help leaders deploy infrastructures ready to support the AI-driven data processing and analytics capabilities their organizations require. The next generation of HPC will include a convergence with quantum computers as well, fundamentally transforming how we perceive value in computing.[19] Noisy Intermediate Scale Quantum (NISQ) computers may be worth hundreds of millions of dollars this decade, Deloitte notes; as with HPC, "NISQ computers are likely to find a place in both government and academia" and lead to "HPC-quantum hybrids" in industrial sectors:

"In the financial sector, for example, it is believed that these intermediate QCs can perform portfolio optimization... trading strategy development, portfolio performance prediction, asset pricing, and risk analysis... One company has described an HPC-quantum hybrid for the simulation and design of a water distribution system; it uses quantum annealing, a restricted version of quantum computation, to narrow down the set of design choices that need to be simulated on the conventional system, with the potential to significantly reduce total computation time."[20]

Technologies' potential aside, identifying the right investments hinges on decision makers' ability today to understand ongoing user and organizational requirements. Optimizing HPC infrastructures for existing and future workloads is now only possible with the right confluence of strategic insight and technological expertise—critical to shaping organizational capabilities across academic and government organizations, and competitive value among enterprises.

AMAX is a leading technology solutions company providing integrated supply chain manufacturing and orchestration services. As a Foxconn Technology Group affiliate, AMAX specializes in the manufacturing of server appliances, high performance computing and storage, AI computing and analytic platforms, and delivers hardware data center solutions at any scale –on any premise.

AMAX HPC and AI Solutions Group (HAS)

AMAX HAS Group provides solutions to organizations that tackle simple-to-complex data workload challenges across hybrid multi-cloud environments. From workload requests and resource allocations to nodes, clusters, and racks, AMAX provides high performance computing hardware solutions for data-centers, server rooms, and GPU AI intensive workloads of all sizes. Partnering across a vast ecosystem of technology and service providers, HAS engineers work with IT and OT organizations to address data, scale and efficiency challenges across computing and storage infrastructure.

Appendix

1. **Desjardins, jeff.** "*How much data is generated each day?*" World Economic Forum. April 17, 2019. https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

2., 4. **Kumar, Dileep et. al**. "*Design and Implementation of High Performance Computing (HPC) Cluster.*" Academic Journal of Management Sciences, January 2018. https://www.researchgate.net/publication/331824476_Design_and_Implementation_of_High_Performance_Computing_HPC_Cluster

3. **Top 500**. "*JUNE 2018.*" June 2018. https://www.top500.org/lists/2018/06/

5. **Gartner Glosssary**. "*High-performance Computing (hpc).*" https://www.gartner.com/en/information-technology/glossary/high-performance-computing-hpc

6. **Glanz, James**. "*Power, Pollution, and the Internet.*" The New York Times, September 22, 2012. https://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html

7., 19. **Moorhead, Patrick**. "*IBM Continues To Invest Big In High-Performance Computing.*" Forbes, September 6, 2019. https://www.forbes.com/sites/patrickmoorhead/2019/09/06/ibm-continues-invest-big-in-high-performance-computing/#4d8d147a6c36

8., 11. **Intel**. "*Bringing AI Into Your Existing HPC Environment, and Scaling It Up.*" https://www.intel.com/content/www/us/en/high-performance-computing/scale-hpc-with-ai-guide.html

9. **Accenture Newsroom**. "*Molex, Accenture and Amazon Web Services Co-Develop Edge Computing Solution Powering NextGen Autonomous Vehicles.*" January 8, 2019. https://newsroom.accenture.com/news/molex-accenture-and-amazon-web-services-co-develop-edge-computing-solution-powering-nextgen-autonomous-vehicles.htm

10., 12. **Farber, Rob**. "*AI-HPC is Happening Now.*" insideHPC, 2017. https://insidehpc.com/white-paper/ai-hpc-happening-now/

18. **Bloomberg Business (Press Release)**. "*AI Infrastructure Market Worth $50.6 Billion by 2025 - Exclusive Report by MarketsandMarkets.*" July 3, 2019. https://www.bloomberg.com/press-releases/2019-07-03/ai-infrastructure-market-worth-50-6-billion-by-2025-exclusive-report-by-marketsandmarkets

13. **Costello, Katie. Gartner Research. February 13, 2019**. *https://www.gartner.com/smarterwithgartner/gartner-predicts-the-future-of-ai-technologies/*

14. **Tracy, Suzanne**. "*Revving Up AI Workloads with HPC under the Hood.*" CIO, April 25, 2019. https://www.cio.com/article/3391212/revving-up-ai-workloads-with-hpc-under-the-hood.html

15., 17. **Zeichick, Alan**. "*Go Big, Really Big, As High-Performance Cloud Computing Scales To Millions Of Jobs.*" Forbes, August 14, 2019. https://www.forbes.com/sites/oracle/2019/08/14/go-big-really-big-as-high-performance-cloud-computing-scales-to-millions-of-jobs/#1c1d54a62efb

16. **Gartner Research**. Top 5 Ways to Successfully Deliver HPC Cloud Strategies, November 29, 2018. https://www.gartner.com/en/documents/3894388/top-5-ways-to-successfully-deliver-hpc-cloud-strategies

20. **Deloitte Insights. Technology, Media, and Telecommunications Predictions 2019. 2018**. https://www2.deloitte.com/content/dam/Deloitte/pe/Documents/technology-media-telecommunications/Quantum%20computers.pdf