Technology Spotlight

The Importance of HPC Expertise For HPC Cloud Computing

Sponsored by Microsoft

Steve Conway, Alex Norton, Earl Joseph and Bob Sorensen
November 2019

HYPERION RESEARCH OPINION

In recent years, cloud services providers (CSPs) have been paying special attention to the worldwide high performance computing (HPC) market. They have been adding features, functions and partners to make their platforms run a broader spectrum of HPC-related workflows cost- and time-effectively.

There are two main reasons for this sharpened focus.

- First, global spending on HPC servers, storage, software and technical support vaulted from about $2 billion in 1990 to $27.7 billion in 2018, en route to a Hyperion Research forecast $39.2 billion in 2023. When $7.4 billion in revenue from HPC public cloud usage is added to the mix, this number grows to $46.6 billion.
- Second, CSPs all know that HPC is an important factor for success in economically important, emerging markets for artificial intelligence (AI) and high performance data analysis (HPDA). HPC is nearly indispensable at the forefront of R&D for automated driving systems, precision medicine, affinity marketing, business intelligence, cyber security, smart cities and the Internet of Things. Today's HPC activity shows where the mainstream HPDA and AI markets are headed.

CSPs efforts have been paying off. Extrapolated findings from Hyperion Research's most recent worldwide study indicate that HPC sites have rapidly increased the portion of their workloads run on third-party clouds. On average, they now send about 20% of all their HPC workloads to third-party clouds, double the 10% average reported in our studies less than two years ago.

Now that CSPs have overcome many of the *technological barriers* for effectively hosting a broad spectrum of HPC workloads, the next big challenge for CSPs is to ensure that they understand the problems HPC users are trying to solve. These problems are becoming much more diverse. They include not only government and academic research, but design engineering for industrial product development and business operations in enterprise data centers.

Understanding these problems is a personnel issue. It calls on CSPs to hire more people with backgrounds in HPC and associated problem domains. This paper looks at the diversification of HPC use cases and examines Microsoft Azure as an example of a CSP that has been adding more HPC experts to reach the company's next growth stage.

Note: this page is intentionally blank.

RAPID GROWTH OF HPC CLOUD COMPUTING

HPC Cloud Growth

In Hyperion Research's latest global study of HPC sites that are using clouds, the surveyed users on average said they run 33% of all their HPC work in third-party clouds. Extrapolating from this group of cloud users to the whole HPC community drops that average to about 20%, still a significant uptick from the 10% figure in Hyperion Research surveys not long ago. HPC cloud computing is rounding an elbow in the growth curve and is ramping up briskly.

Chief among the reasons driving greater usage is the recognition, at least among major CSPs, that the size and growth rate of the worldwide HPC market make it worth special attention. The market for HPC servers, storage, software and technical support expanded from about $2 billion in 1990 to $27.7 billion in 2018, en route to a Hyperion Research forecast $39.2 billion in 2023. When $7.4 billion in revenue from HPC public cloud usage is added to the mix this number jumps to $46.6 billion. (See Figure 1.)

FIGURE 1

HPC Worldwide Market Forecast ($000)

Revenues by the Broader HPC Market Areas			
	2018	2023	CAGR 18-23
Server	13,706,088	19,979,016	7.8%
Storage	5,547,188	7,771,184	7.0%
Middleware	1,582,892	2,217,801	7.0%
Applications	4,627,492	6,413,592	6.7%
Service	2,229,921	2,858,820	5.1%
Total Revenue	27,693,580	39,240,413	7.2%
Source: Hyperion 2019			

HPC cloud (CSP) usage raises forecast to $46.6B)

Source: Hyperion Research, 2019

CSPs are also aware that HPC is an important factor for success in the emerging markets for artificial intelligence. In one of our recent global surveys of prominent AI experts, 87% said HPC is "extremely important" for advancing AI. HPC is nearly indispensable today at the forefront of R&D for automated driving systems, precision medicine, affinity marketing, business intelligence, cyber security, smart cities and the Internet of Things. Today's HPC activity points to where the mainstream AI and other high performance data analysis (HPDA) markets are headed in the future.

In pursuit of HPC business, CSPs have focused heavily on augmenting their technological capabilities to address the demanding requirements of many HPC workloads. They have added HPC system software, applications software (including ISVs applications), InfiniBand and Ethernet interconnects,

beefed-up memory and storage hardware and, more recently, bare metal resources to address remaining performance penalties of virtualization. CSPs have also substantially improved price/performance in recent years.

Now that CSPs have overcome many of the *technological barriers* for effectively hosting a broad spectrum of HPC workloads, the next big challenge for CSPs is to ensure that they understand the problems HPC users are trying to solve. This can be a major challenge for two reasons:

- **The problems are becoming much more diverse.** They include not only government and academic research in nearly every scientific discipline, but mainstream business problems—design engineering for industrial product development (upstream R&D) and, more recently, live business operations in enterprise data centers (downstream jobs), such as business intelligence, affinity marketing, sales analysis, cyber security and Internet of Things workloads. The use of machine learning and other AI-HPDA methods is adding to the challenge of addressing HPC business problems.
- **Understanding these problems is a personnel issue**. The fastest and most effective way for a CSP to gain an effective understanding of HPC business and other problems is by employing people with HPC backgrounds. The idea here is that to handle HPC business, the CSP staff should consist heavily of HPC experts who've gained cloud experience, rather than just cloud experts working to gain HPC experience.

The benefits of employing cloud-proficient HPC experts for HPC customers are self-evident: peer-to-peer interactions help to ensure that the HPC workloads receive the most appropriate, and therefore also the most time- and cost-effective, cloud resources. Among the payoffs for CSPs that employ people with first-hand experience solving business and other HPC problems are closer customer relationships and faster revenue growth.

MICROSOFT AZURE STAFFS UP WITH HPC EXPERTISE AND TECHNOLOGY

Microsoft Azure arguably was the first major CSP to turn its attention seriously to the special requirements of HPC users. The Azure unit has taken major steps not only to meet the technology needs of HPC workloads, but also to meet customers' needs for HPC problem-solving experience.

This strategy has substantially boosted Azure's standing in the HPC cloud computing market. In a recent Hyperion Research worldwide survey in which HPC cloud customers were asked to name all the CSPs they use, 32% of the respondents named Microsoft Azure—not far from market leader AWS' 46% figure. That's a big leap from Azure's single digit following only a few years ago.

Azure Technology Advances for HPC

Especially in the past two years, Azure has substantially expanded technology to support HPC workloads in hybrid and standalone cloud configurations, especially to match on-premise capabilities. The goals here are not to replace on-premise resources but to interoperate with them as seamlessly as possible in hybrid configurations, and to provide leading-edge performance for new HPC users, such as SMBs and others that lack on-premise capabilities. Hyperion Research studies consistently show that most HPC work sent to CSPs taps pent-up demand rather than workloads already being run on premise.

Microsoft recognized that supporting HPC workloads requires more than just computing prowess; it also needs exceptional storage, networking, and software technologies. Microsoft's acquisition of Avere, for

example, targeted the issues of moving excess demand to the cloud via low-latency, high-speed data pipes and avoiding costly, time-consuming application rewrites for the cloud.

Prominent among Microsoft's initiatives to support HPC was the 2017 acquisition of Cycle Computing, a leader in cloud computing orchestration and ease-of-use for experienced and newer HPC users. A decade ago, Cycle was one of the early visionaries in the then-nascent market for HPC computing in the cloud, and many of Cycle's market expectations have materialized.

The company's cloud design point is modeled after on-premise, bare metal capabilities, with the goal of mirroring those abilities with minimal compromises and the added benefits of elastic scaling and sophisticated orchestration. The decision to add a Cray bare metal offering to Azure was consistent with Microsoft's strategy.

Bringing HPC To the Enterprise Market

Microsoft HPC Pack aims to bring the power of HPC to the commercial mainstream. The centralized management and deployment interface is designed to help simplify deployment for compute clusters and enable simple, effective management. HPC Pack includes a scalable job scheduler that supports interactive Service-Oriented Architecture (SOA) applications using High Performance Computing for Windows Communication Foundation (HPC for WCF) and parallel jobs using the Microsoft Message Passing Interface (MS-MPI). Essential applications from key independent software providers (ISVs) can be run on the cluster to help meet business needs.

Azure runs on its own OS, a modified version of Windows. But today, most Azure HPC jobs run on Linux. More and more native Azure services also run on Linux. For example, Azure's Software Defined Network (SDN) is based on Linux. Microsoft has close relationships with Red Hat, SUSE, and Canonical. In fact, Microsoft reports that the importance of Linux for Azure extends far beyond the HPC market.

Azure Personnel Advances for HPC

As noted earlier, the final frontier in preparing CSPs for mounting success in the global HPC market is to hire staff with strong HPC backgrounds. The Cycle Computing and Avere acquisitions heavily augmented Microsoft Azure's team members with first-hand HPC experience.

- Azure CycleCloud is designed to enable enterprise IT organizations to provide secure and flexible cloud HPC and Big Compute environments to end users. Dynamic scaling of clusters aims to provide customers with the resources they need at the right time and the right price. Azure CycleCloud’s automated configuration is made to enable IT to focus on providing service to the business users. Azure CycleCloud allows organizations to manage HPC workloads with a wide range of schedulers (e.g., Slurm, Grid Engine, HPC Pack, HTCondor, LSF, PBS Pro and Symphony).

Microsoft has also adopted the practice of making most new hires for Azure from outside the company and placing a strong emphasis on HPC experience. For example, the Azure Compute Program Manager was a center director at NCSA and a former executive with Cycle Computing.

Support For Production Workloads

As part of early acceptance testing, the Azure HPC team benchmarked many widely used HPC applications.

- One common class of applications are those that simulate computational fluid dynamics (CFD). To see how far HB-series virtual machines (VMs) could scale, the team selected the Le Mans 100-million-cell model available to Star-CCM+ customers. Azure reports that the model scaled to 256 VMs across multiple configurations, accounting for up to 11,520 CPU cores. The testing revealed that maximum scaling efficiency was with two MPI ranks per NUMA domain, yielding a top-end scaling efficiency of 71.3 percent. For top-end performance, three MPI ranks per NUMA domain yielded the fastest overall time to solution.
- Another common class of applications are those that simulate the physical and chemical properties of molecules, otherwise known as molecular dynamics (MD). To see how far HC-series VMs could scale, Azure ran a benchmark test using the popular CP2K open source molecular dynamics program. CP2K is widely used in academia and industry. It is one of 13 applications used by PRACE as part of the Unified European Applications Benchmark Suite to drive acceptance testing of supercomputers deployed in Europe.
- The H20-DFT-LS benchmark is a single-point energy calculation using linear-scaling DFT and 2,048 water molecules. Azure reports that HC-series VMs successfully scaled to 392 VMs and 17,248 cores. The Azure team says that 288 VMs offer the optimal balance in price-performance for large scaling on this benchmark.
- The LiHFX benchmark is a single-point energy calculation simulating a 216-atom Lithium Hydride crystal with 432 electrons. According to Azure, HC-series VMs successfully scaled to 512 VMs and 22,528 cores.

Support For AI Workloads

As one aspect of the company's overall focus on AI/ML/DL, Microsoft is investing $1 billion in OpenAI to support artificial general intelligence (AGI) use cases that promise to provide broad economic benefits. OpenAI is the for-profit corporation OpenAI LP, whose parent organization is the non-profit organization OpenAI Inc, which conducts research in the field of artificial intelligence to promote and develop friendly AI in order to benefit humanity as a whole.

Researchers from Azure and other Microsoft business units are partnering to develop a hardware and software platform within Microsoft Azure that will scale to AGI. This team plans to develop new Azure AI supercomputing technologies, and Azure will become the exclusive cloud provider for the OpenAI — extending Microsoft Azure’s capabilities in large-scale AI systems.

Azure's Project Brainwave leverages the massive FPGA infrastructure that Microsoft has been deploying over the past few years. By attaching high-performance FPGAs directly to Microsoft's datacenter network, Azure says, a deep neural network can be mapped to a pool of remote FPGAs and called by a server with no software in the loop. This system architecture is designed to reduce latency, since the CPU does not need to process incoming requests. It aims to allow very high throughput, with the FPGA processing requests as fast as the network can stream them. Project Brainwave is also designed to scale across a wide range of data types.

MICROSOFT AZURE SUCCESS STORIES

City of Hope Medical Research and Treatment Center

Computer scientists at the City of Hope medical research and treatment center use high-performance computing to understand diseases like cancer and diabetes at the molecular modeling level. A team under Dr. Nagarajan Vaidehi, Director of the Computational Therapeutics Core (CTC), uses Linux-based Microsoft Azure Virtual Machines to slash protein-modeling simulations from weeks to days. The

computational methods developed by her team play a critical role in speeding up the drug discovery process.

However, there are never enough compute resources to analyze the huge volumes of genomic, proteomic, and structural data being generated by researchers. The CTC has its own HPC clusters that run around the clock. These clusters are primarily made up of high-performance CPUs, but, increasingly, Vaidehi's team relies on graphics-processing units (GPUs) to accelerate simulations.

With the addition of Azure Virtual Machines N-Series, Vaidehi and her team were able to rapidly scale their GPU cluster footprint. Before the availability of GPUs in the cloud, Vaidehi's team had to put together proposals for additional HPC resources and wait months for budget approval—if it ever came. Talented scientists were diverted from research and forced to focus on procurement issues.

Dr. Vaidehi and her team plan to keep their "extension" datacenter in Azure running constantly, churning through simulations. Other City of Hope teams are also using Azure HPC resources, both in continuous-use and burst models.

MetLife Insurance Company

Voted the most admired life and health insurance company by Fortune magazine in 2015, MetLife is a global provider of life insurance, annuities, employee benefits and asset management. Critical to its success is the ability to perform complex actuarial modeling of business data. To this end, the company shifted some of the high-performance computing and data processing this modeling requires to the Microsoft Azure cloud platform.

MetLife created a processing environment called the MetLife Integrated Actuarial Modeling Environment (MIAME). This end-to-end solution, based on a high performance computing grid, takes advantage of Microsoft technologies including HPC Pack, Windows Server, Microsoft Analytics Platform System, big data, and Microsoft SQL Server. One of the technology goals of the MIAME program is to continuously improve capability and control costs by evolving with new technologies as they become available. This flexible framework allows for these technical paradigm changes without the need to rebuild processing from scratch.

By using Microsoft cloud offerings rather than continuing to build out its own calculation infrastructure, MetLife expects to save between 45 percent and 55 percent — as well as year-over-year savings — in infrastructure costs. The increased speed at which MetLife actuaries can run models has given them the ability to deliver results faster in order to better serve its customers and other stakeholders. MetLife isn't stopping there. The company is exploring Microsoft Power BI for data visualization and graphical processing for actuarial calculations.

MetLife says it benefited from the flexibility and scalability of Azure data processing capabilities to achieve faster, more accurate actuarial calculations and save significant infrastructure costs, which resulted in more value for the company, its customers, and other stakeholders around the world.

FUTURE OUTLOOK

Hyperion Research forecasts that the global market for running HPC workloads in third-party clouds will reach about $7.4 billion in 2023, representing about 16% of our projected $46.6 billion in overall spending on HPC in that year. CSPs have been turning more attention to the global HPC market, because this market has become a sizable opportunity and because HPC is at the forefront of R&D for

economically important, emerging HPDA-AI use cases. Their efforts to date have been rewarded with a sharp upturn in the average percentage of HPC sites' workloads being assigned to third-party clouds, from about 10% two years ago to about 20% today.

Hyperion Research believes that HPC cloud computing has rounded an elbow in the growth curve. We forecast continued robust growth for this part of the HPC market through 2023, although not at the same stellar rate CSPs benefited from in the past two years by adapting to capture HPC "low-hanging fruit," i.e., workloads that could be captured primarily via improvements in CSP technology, security and business processes.

Realizing the next growth stage for HPC cloud computing will require a deeper understanding of HPC problems themselves, even as the problem set grows more heterogeneous, complex and time-critical with the rise of HPC-enabled AI and enterprise business operations. CSPs wanting to make the most of this opportunity should add more staff with direct experience in HPC problem domains, to complement cloud computing experts who are learning about HPC.

Hyperion Research believes that Microsoft Azure is already far along on this important path and is therefore positioned well to benefit from, and help drive, the next growth stage in HPC cloud computing.

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 multiuser 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
