Technology Spotlight

Cloud Computing for HPC Comes of Age

Sponsored by Amazon Web Services

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

HYPERION RESEARCH OPINION

Hyperion Research studies show that the proportion of all HPC sites worldwide that run some workloads in public clouds has shot up more than five-fold in recent years, from 13% of sites in 2011 to 74% in 2018. That's the good news for cloud services providers (CSPs): cloud computing has expanded the HPC market and has further democratized it by making HPC available to new adopters that lack on-premise HPC resources. The less-good news is that the average percentage of the sites' total HPC workloads being sent to CSPs has risen more slowly during this period and stands today at just under 10% (9.8%). So, HPC-in-the-cloud growth has been wide, nicely multiplying the revenue going to CSPs, but has not been nearly as deep.

Our newest studies indicate that this pattern is about to change. In the next two to three years, HPC sites say the average proportion of their workloads headed for CSPs will jump to about 15%, an important upward inflection. The reasons for this increased momentum lead us to believe that cloud adoption among HPC sites is rounding an elbow in the growth curve and the proportion of all HPC workloads run in the cloud will expand at a brisk rate.

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 paying special attention to. The market for HPC servers, storage, software and technical support expanded from about $2 billion in 1990 to $24.3 billion in 2017, en route to a Hyperion Research forecast $38 billion in 2022. Make that $44 billion when revenue from public cloud usage is added to the mix.

But that's not all, CSPs are also aware that HPC is an important factor for success in the emerging markets for artificial intelligence (AI) and high performance data analysis (HPDA) applications. 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 indicates where the mainstream HPDA and AI markets are headed in the future.

With the attractive HPC, HPDA and AI combined market opportunity in mind, leading CSPs have been adding features, functions and partners designed to make their platforms run a broader spectrum of HPC related workloads cost- and time-effectively.

Amazon Webs Services (AWS) enjoys a strong lead among CSPs pursuing business in the worldwide HPC market. This paper discusses the dynamics of this fast-evolving market, using current realities to counter frequent misperceptions, and then explains why Hyperion Research believes AWS is well positioned to benefit strongly from the growth we forecast for the global HPC cloud computing market.

Note: this page is intentionally blank.

SITUATION OVERVIEW

Misperceptions and Realities

In part because HPC cloud computing has come so far so quickly, misperceptions about today's realities abound. In this section of the paper, we try to address the misperceptions we encounter most often in our worldwide HPC market studies.

(All mentions in the paper of "cloud" refer to public and private cloud services offered by CSPs.)

Misperception: It's a zero-sum game, where HPC workloads run in clouds have all been transferred from on-premise HPC data centers. A misperception is that clouds' gains are on-prem data centers' losses.

Reality: It's a bit more complicated than that:

- True, an important subset of HPC workloads run in clouds used to be run in on-premise HPC data centers.
- Other workloads are run in clouds because on-premise data centers are inelastic, and many are oversubscribed. Hyperion Research studies show pent-up demand may exceed on-premise capacity by as much as 2 to 3 times and clouds are helping to handle this demand. Some of this involves periodic spikes in demand ("surge" or "overload" demand) and some is long-standing, pent-up demand.
- In general, when HPC users find better ways to run HPC jobs, they tend to run more jobs and spend more money, in order to do more R&D/science/engineering.
- In still other cases, organizations, especially SMEs, run HPC workloads in clouds because they have never built on-premise HPC data centers, have no experience managing centers of this kind, and want to avoid the capital and operating costs associated with on-prem HPC data centers. In these cases, CSPs play an important role in democratizing the use of HPC. To make adoption easier for new HPC users, leading cloud services have become much more user friendly in recent years.

Misperception: Clouds are only suitable for "embarrassingly parallel" workloads.

Reality: This statement was largely true in 2011, when the DOE-sponsored *Magellan Report on Cloud Computing for Science*[1] found that "scientific applications with minimal communication and I/O are best suited for clouds." But since then, major CSPs have increasingly turned their attention to the global HPC market, both as an attractive growth opportunity in itself and also as a way to exploit HPC's unrivaled capabilities at the forefront of R&D in the high performance data analysis (HPDA) and artificial intelligence (AI) markets.

- Thanks to the addition of high speed networking, enhanced data security, a growing menu of software and hardware options (including bare metal), compliance with government regulations,

[1] https://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Magellan_Final_Report.pdf

and in-house technical support and domain expertise, major CSPs today can efficiently handle a broader range of HPC workloads, including many communications (I/O)-intensive jobs. Through CSPs and their partners, users now routinely run workloads including CFD, structural analysis, fluid-structure interactions, weather and climate modeling, and more in cloud environments.

- It's a safe bet that CSPs will continue to add capabilities designed to extend the range of cloud-friendly HPC and HPDA-AI workloads.

Misperception: Running HPC workloads in clouds is always less expensive (or more expensive) than running them on premise.

Reality: Our HPC end user research confirms that it depends on the workload, the customer and the cloud service.

- A recent Hyperion Research study showed that when cost and other factors are considered, a majority of HPC users (55.8%) find it relatively easy to decide which workloads to run on premise and which to run in a cloud. In 2017, 74% of HPC sites were making use of cloud environments.

Decision Difficulty to Run HPC Workloads on Premise or in a Cloud

Very easy	17.3%
Somewhat easy	38.5%
Somewhat difficult	19.2%
Very difficult	11.5%
Not sure/don't know	13.5%

- Cost comparisons can be misleading; for example, on-premise evaluations need to take CAPEX as well as OPEX into consideration, along with waiting time in on-premise queues (a job that costs 5% more but can be completed days sooner in a cloud environment may be less expensive, all things considered). On the other hand, cloud computing needs to consider data transfer times, performance penalties in cases where virtualization applies, and in-cloud processor and networking capabilities.
- Clouds are evolving to run an ever-broader spectrum of HPC workloads more cost effectively, including many challenging workloads (examples below). For organizations without HPC data centers, including but not limited to many SMBs, clouds are especially attractive compared with the cost and time needed to build and maintain on-prem data centers.
- Finally, leading CSPs and their partners now provide a range of options for monitoring and controlling costs.

Misperception: Data security and data transport in clouds is inadequate.

Reality: Some government agencies and private sector companies have assessed cloud data security and found it still not up to their stringent internal standards, but leading CSPs now make an array of data security and transport measures available:

- Data encryption in transit and at rest
- Role based access control (RBAC)
- Firewalls
- Private networks

- Fast, secure data transport solutions (so called direct connect offerings)
- For some organizations, CSPs offer a higher level of data security than their on-prem data centers.

Hyperion Research studies confirmed that for many HPC user organizations, cloud data security is adequate today, as evidenced by their escalating use of CSPs for computing. Another proof point confirmed in our studies is that in 2018, nearly half of all new HPC private clouds were provided by CSPs, a testament to users' growing trust in cloud data security.

Misperception: Using a cloud service requires a long term commitment.

Reality: Some customers have contracts with CSPs for extended periods, but they and others can also use clouds as a transitory resource that can be turned on to handle an immediate problem, then turned off again. Long term contracts aren't required. Most HPC jobs today are being done via pay-as-you-go arrangements.

Misperception: Clouds may not be compliant with industry regulations.

Reality: CSPs have become compliant with many important industry standards and regulations.

- Example: When AWS became HIPAA compliant, health care organizations quickly began to use AWS more heavily.
- In addition, CSPs today have "quick start" programs designed to automate the process of creating compliant environments even when complex security and privacy regulations apply.

Misperception: CSPs lack domain-specific experience.

Reality: Leading CSPs partner with many organizations that augment the CSPs' own domain-specific capabilities. Most CSPs also have specialized organizations or business units catering to vertical domains. These specialized organizations speak the language of the domain they work with, and often have reference architectures for workloads that are specific to those industries.

AWS: THE GLOBAL MARKET LEADER IN HPC CLOUD COMPUTING

Hyperion Research's most recent study shows that AWS is the primary CSP for 58% of surveyed HPC user organizations that run workloads in the cloud, more than double the percentage for the number two vendor (23%) and more than seven times higher than the third place competitor (8%).

Key AWS Features

The large array of features and functions available on AWS today illustrates how AWS has enhanced its platform in recent years to support a broader spectrum of HPC workloads and user requirements.

Data Security

AWS offers multiple provisions to bolster data security, both in transit and at rest. These include:

- Network firewalls built into Amazon VPC and web application firewall capabilities in AWS WAF. These let users create private networks and control access to instances and applications.
- Encryption in transit with TLS across all services.
- Data encryption capabilities available in AWS storage and database services, including EBS, S3, Glacier, Oracle RDS, SQL Server RDS, and Redshift.
- Connectivity options that enable private or dedicated connections from an office or other on-premise environment.
- AWS Identity and Access Management (IAM) makes it possible to define individual user accounts with permissions across AWS resources.
- Amazon CloudWatch issues alert notifications when specific events occur, or thresholds are exceeded.
- Machine learning-based firewalls and threat detection measures (Amazon Macie and Amazon GuardDuty). A security checklist that walks users through recommended best practices for operating in the cloud.
- Quick Launch environments, such a HIPAA for HCLS workloads, are designed to launch complete, secure and compliant environment in minutes.

Data Upload and Data Transfer

AWS Direct Connect is a cloud service solution designed to make it easy to establish a dedicated network connection from customers' premises to AWS. AWS Direct Connect lets customers establish private connectivity between AWS and their data center, office, or colocation environment. AWS customers report that in many cases, this reduces network costs, increases bandwidth throughput, and provides a more consistent network experience than Internet-based connections. AWS Snowball and AWS Snowmobile are other data transport solutions that use devices designed for securely transferring large amounts of data into and out of the AWS Cloud. AWS customers today use Snowball to migrate analytics data, genomics data, video libraries, image repositories, and backups. Transferring data with these utilities can cost as little as one-fifth the cost of transferring data via high-speed Internet.

Support for Communications-Intensive Workloads

AWS' Elastic Fabric Adapter (EFA) is a network interface aimed at customers needing high levels of inter-instance communication. At its core, EFA relies on a custom-built OS-bypass technique to speed up communication between instances. EFA supports applications using the Message Passing Interface (MPI) by employing the industry standard libfabric APIs. As a result, codes using a supported MPI library can get by with little or no source modifications. According to AWS, the EFA technology will enable customers to scale their tightly coupled applications to tens of thousands of CPU cores. It's aimed at typical scalable HPC codes, including communications-intensive codes such as computational fluid dynamics, weather modeling, and reservoir simulation.

Support for High-Performance File Systems

Amazon FSx for Lustre provides high-performance file systems that are optimized for HPC. Amazon FSx for Lustre is designed to provide sub-millisecond access to data stored in durable, long-term data stores (e.g., Amazon S3). It can read and write data at speeds of up to hundreds of GBs per second. FSx for Lustre is natively integrated with Amazon S3 for processing HPC data sets stored in Amazon S3. It can also be used as a standalone high-performance file system for workloads that have to burst to the cloud. By copying on-premises data to an FSx for Lustre file system, that data can be available for fast processing by compute instances running on AWS.

Domain-Specific Workloads

AWS provides special services to support requirements for established and emerging domain-specific workloads, such as the following:

- Automated Driving Systems:
  - AWS provides a full suite of services to support Advanced Driver Assistance Systems (ADAS) and Autonomous Vehicle development and deployment.
  - AWS provides support for deep learning frameworks such as Apache MXNet, TensorFlow and PyTorch to accelerate algorithm training and testing.
  - AWS Greengrass provides edge computing with machine learning inference capabilities for real-time processing of local rules and events in the vehicle while minimizing the cost of transmitting data to the cloud.
- Precision Medicine:
  - PrecisionFDA. This initiative is led by the U.S. Food and Drug Administration. The goal is to define the next-generation standard of care for genomics in precision medicine.
  - The Precision Medicine Platform - powered by Amazon Web Services - is a strategic initiative of the American Heart Association's Institute for Precision Cardiovascular Medicine. This platform gives researchers the ability to collaborate and analyze datasets, to better predict and intervene in cardiovascular disease and stroke.
  - Deloitte ConvergeHEALTH gives healthcare and life sciences organizations the ability to analyze their disparate datasets on a singular real world evidence platform.

Transitory Workloads

Many small ("$20") workloads can be handled on the fly. When peta-class public data sets are made available for free in the cloud, for example, many small workloads go to the cloud that otherwise might never be run. This can sometimes prevent an entire organization from grinding to a halt waiting for a result from an on-prem cluster. The cloud cluster bursts into existence to deal with a problem and then just as quickly disappears when that part of the pipeline is done.

Cost Controls and SLAs

Many AWS partners have built cost control models on the Amazon platform. In addition:

- The AWS Billing Dashboard lets customers view the status of their month-to-date AWS expenditure, pinpoint the services that account for the majority of the overall expenditure, and understand at a high level how their costs are trending.
- Amazon CloudWatch is a monitoring and management service built for developers, system operators, site reliability engineers (SRE), and IT managers. CloudWatch provides customers with data and actionable insights to monitor applications, understand and respond to system-wide performance changes, optimize resource utilization, and get a unified view of operational health.
- AWS Budgets gives customers the ability to set custom budgets that alert them when costs or usage exceed (or are forecasted to exceed) the budgeted amount.

Compliance with Major Industry Regulations

When AWS became HIPAA compliant, health care organizations quickly began to use AWS more heavily. AWS' compliance with this and other important regulations also makes it easier and less costly for organizations to become compliant internally if they aren't already.

AWS Worldwide Partnership Program

AWS partner programs are designed to support the unique business models of members by providing them with increased prominence and additional support from AWS partner teams. The AWS partner community already includes modelling & simulation ISVs and partners who specialize in orchestration, resource management, system integration and consultancy. Licenses for many common HPC codes can be directly purchased from the AWS Marketplace.

AMAZON SUCCESS STORIES

BP

BP, which employs 75,000 people in 72 countries, extracts about 3.3 million barrels of oil equivalent each day. It falls to the company's downstream segment to decide which of these feedstocks—which vary widely in terms of quality, yield, and cost of extraction—should be used to manufacture which of the many different fuels, lubricants, and petrochemicals that a petroleum refinery can produce.

To support this crucial decision-making, BP downstream was using Schneider Electric Spiral Suite software to run linear programming models involving complex calculations based on thousands of inputs. But the company couldn't take full advantage of the software's potential power because of the long processing times resulting from its deployment in the company's on-premise data centers.

Today, BP is running Spiral Suite on AWS. By using AWS Auto Scaling, BP can increase or decrease the number of Amazon EC2 instances Spiral Suite is using as calculation nodes on demand, ensuring Spiral Suite has access to as much processing power as necessary during complex calculations while avoiding paying for resources when they aren't needed. For BP, the biggest benefit of the move to AWS is how much faster Spiral Suite can now execute complex calculations. A problem that once would have required about seven hours of calculation time completes in less than four minutes, which helps BP adapt to market changes in almost real time.

Celgene

Celgene is a global biopharmaceutical company that develops drug therapies for cancer and inflammatory disorders. Headquartered in New Jersey and employing more than 8,000 globally, the company is committed to improving the lives of patients through the delivery of innovative treatments.

In the pharmaceutical industry a failed clinical trial may result in product failure that can exceed a billion dollars in total expense to pharmaceutical companies. To avoid this scenario, Celgene's scientific researchers take extreme measures to analyze compounds and focus only on those with a high probability of success. In order to "fail fast," Celgene needed agility and speed in the HPC and analytics workflows.

Using AWS, Celgene scientists have dramatically reduced the time it takes to complete HPC jobs needed for cancer drug research. For their informatics researchers, computational jobs on AWS can be reduced to hours, compared to weeks or months when they were managing their HPC cluster on premise. As a result, researchers can run many more queries. By spinning up a few hundred nodes on AWS and getting results in less than a day, their scientific researchers have a lot more freedom to ask questions that weren't even possible before.

TLG Aerospace

Seattle-based TLG Aerospace is an aerospace engineering services company that provides CFD analysis and other services to customers worldwide. The company provides engineering for small and large aircraft. It focuses on vehicle analysis and design, including static and dynamic loads, flutter, stability and control, aerodynamics, and airframe stress analysis and design.

The company wanted to reduce the costs associated with running simulations. TLG also wanted to take on larger simulations, but its internal HPC cluster was limited to a small number of nodes and couldn't allocate enough memory to run large scale problems.

TLG decided to use Amazon Web Services (AWS) to run CFD simulations, taking advantage of Amazon EC2 Spot instances to exploit unused computing capacity at a discounted price. The company saw a 75 percent reduction in the cost per CFD simulation and has been able to pass those savings along to customers—allowing TLG and its customers to become more competitive.

FUTURE OUTLOOK

Hyperion Research forecasts that the global market for running HPC workloads in the cloud will grow quickly to $6 billion in 2022, representing about 16% of our projected $44 billion in overall spending on HPC in that year for servers, storage, software, support and cloud usage. 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 including automated driving, precision medicine, affinity marketing, business intelligence, cyber security, smart cities and the Internet of Things. Leading CSPs will continue to enhance their capabilities in order to address a growing portion of existing and emerging HPC workloads time- and cost-effectively.

CSPs will benefit from being able to tap into pent-up demand from on-premise HPC data centers as well as new demand from traditional HPC buyers and first-time adopters in commercial markets, along with SMEs and others who want to avoid building and operating their own HPC data centers. Amazon Web Services has a strong lead today in the worldwide market for HPC cloud computing. Hyperion Research believes that Amazon Web Services is well positioned to maintain its global leadership in this fast growing market and has the potential to widen its lead.

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
