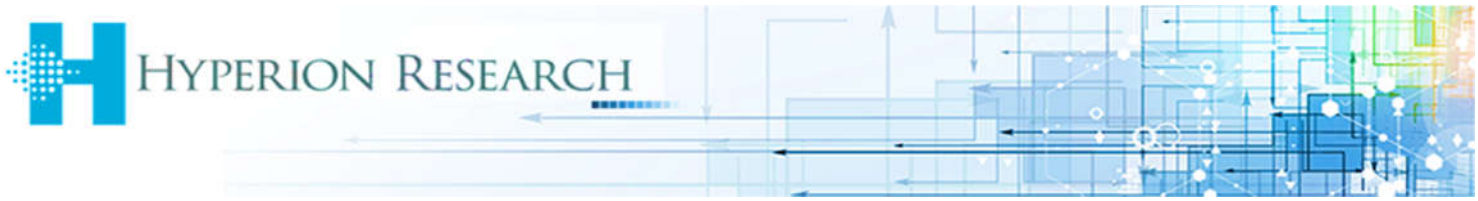


Technology Spotlight

The Future of HPC Cloud Computing

Sponsored by Google

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

HYPERION RESEARCH OPINION

Hyperion Research studies show that the proportion of all HPC sites worldwide that run some workloads in public clouds has risen almost six-fold, from 13% of sites in 2011 to 74% in 2018. It's clear that cloud computing has expanded and democratized the HPC market, first and foremost by accommodating pent-up demand that sometimes exceeds the capacity of HPC sites' on-premise systems by 100% or more, and second by making HPC available to new adopters that lack on-premise HPC resources and expertise.

Although most HPC sites run at least some workloads on cloud services provider (CSP) platforms today, on average they assign only about 10% of all their workloads to external clouds and this figure has risen very slowly during the past decade. Our recent studies indicate that in the next two to three years, the average proportion of HPC sites' workloads will push up more rapidly to more than 15%, and that this accelerating rate of cloud adoption will continue from there. Hyperion Research forecasts that by 2023, revenue for running HPC workloads on CSP platforms will reach approximately $6 billion, or about 14% of the projected $44 billion value of the worldwide HPC ecosystem at that date (which includes servers, storage, software, technical support and cloud usage).

A factor that will increasingly drive cloud usage is the containerization of on-premise HPC. Driven by the need to support extremely diverse workloads (e.g., simulation, analytics, mixed precision), major OEMs are designing systems that can assemble ("compose") the most appropriate resources for each application, partition these with their applications in very lightweight containers using Docker, Singularity or similar tools, and dissolve the containers when the job is finished, releasing the resources to be used again. This architectural direction will render cloud and on-premise environments more and more alike, making it more attractive for existing and new HPC cloud users to exploit CSP platforms in hybrid or standalone deployments.

Another driver of HPC cloud adoption is enterprises installing HPC clusters to support complex, time-sensitive business operations—and bursting out to clouds for additional resources. Some are large corporations that use multiple CSPs ("multi-clouds") in different parts of their businesses today. This group will benefit from a single point from which to manage these diverse, global resources.

This paper summarizes key trends driving the future of HPC cloud computing and zeroes in on Google Cloud Platform (GCP) as a business unit that is making important progress in addressing current and future requirements.

Note: this page is intentionally blank.

(In this paper, unless otherwise specified, the term "cloud" refers to public and private cloud resources offered by CSPs.)

The Growth of HPC Cloud Computing

Hyperion Research studies show that the proportion of all HPC sites worldwide running some workloads in clouds has grown almost six-fold, from 13% in 2011 to 74% in 2018. The percent of these sites' total HPC workloads being run in clouds stands today at just under 10% (9.8%) on average, but a recent Hyperion Research survey indicates that this figure will jump to more than 15% in two to three years, a significant upward swing. HPC cloud computing is approaching an elbow in the growth curve.

Drivers of HPC Cloud Computing Growth

Multiple factors are propelling the growth of HPC workloads on CSP platforms:

Pent-up demand/cloud elasticity. Our studies show that many HPC sites around the world have HPC work that exceeds their on-premise computing capacity by 50% to 200% or more. Cloud elasticity is the number one reason why HPC sites are increasingly turning to CSPs.

Special hardware and software. Users also turn to the cloud frequently to access hardware or software that isn't available on-premise, such as a specific CPU or accelerator, or a version of a software tool that isn't yet available (or is no longer available) on-premise. They may also turn to the cloud to access thousands of accelerators for the duration of a job, avoiding the need to purchase and install this many accelerators on-premise.

Queue avoidance. There is considerable debate surrounding the costs of running HPC workloads in the cloud vs. on-premise—and in truth it depends on the workload—but when using the cloud means avoiding a week's wait in the on-premise queue, the cost equation can look very different.

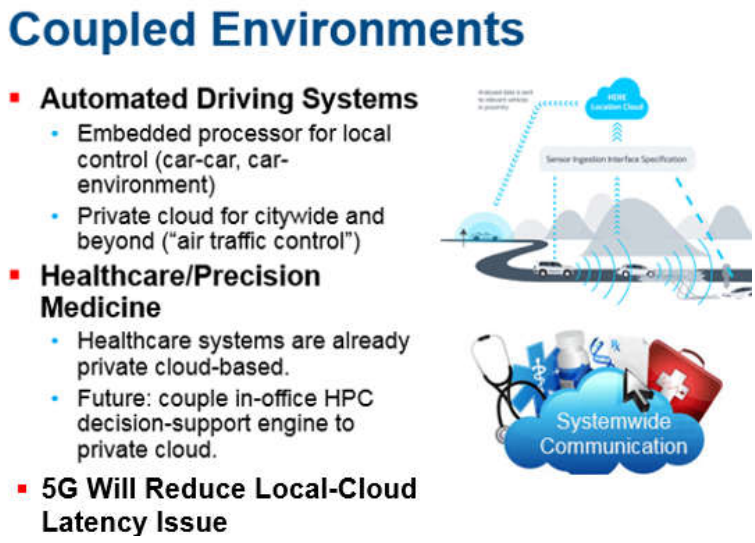
R&D isolation. Another common reason for turning to cloud computing is to run R&D projects, including application development and testing, both to avoid the on-premise queue and to isolate confidential projects from day-to-day production computing workloads.

Networking with the outside world. One other important reason why HPC users turn to CSP platforms is to collaborate and link up with global communities and public research in their scientific, engineering and business domains.

Closely Coupled Hybrid Clouds. Many important emerging AI use cases will need on-premise and cloud computing environments to interoperate closely in real time. Figure 1 illustrates how hybrid clouds will function in two of these use cases, automated driving systems and precision medicine.

FIGURE 1

AI Use Cases Needing Closely Coupled Hybrid Clouds



Source: Hyperion Research, 2019

The Containerization of HPC

HPC on-premise and cloud environments will increasingly look and act alike as on-premise architectures feature container-like partitions to concurrently run workloads with different requirements. CSP platforms for some time have featured containers that assemble the hardware and software resources most appropriate for each workload. One of the big advantages of clouds has been the ability of these containers to offer very heterogeneous resources (see "Special Hardware and Software" in previous section).

Today, on-premise HPC systems are facing a similar challenge—how to support an extremely heterogeneous mix of workloads with varying requirements for precision levels, software stacks, and other needs. This is now common with simulations requiring various accelerators, with machine and deep learning requiring new software and accelerators, and with analytics such as graph analysis. This challenge is nicely described in the January 2018 U.S. Department of Energy paper, *Production Computational Science in the Era of Extreme Heterogeneity*. Leading HPC OEMs are designing next-generation architectures with the goal of supporting this heterogeneity both on-premise and in cloud environments. The growing functional resemblance between on-premise and cloud environments will undoubtedly make it easier for HPC sites to exploit clouds, including the minority of sites that haven't used the cloud yet.

CSP Motivations for Pursuing the HPC Market

The robust growth of the global HPC market is an important reason why CSPs have been paying more attention to this market in recent years—adding features, functions and partners with the aim of addressing a broader spectrum of HPC workloads. 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. Make that $44 billion when revenue from public cloud usage is added to the mix (see Figure 2).

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. Hyperion Research forecasts that by 2023, the HPDA-AI market for HPC servers will reach about $6.4 billion, or about 32% of the $19.9 billion worldwide market for HPC server systems.

FIGURE 2

HPC Ecosystem Market Forecast, with CSP Usage

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 $44 billion

Source: Hyperion Research, 2019

HPC Cloud Trends/Future Directions

Hyperion Research studies indicate that the cloud computing market for HPC will remain in a period of strong innovation and competition, with adequate business to allow multiple players to thrive and grow. Here are important trends now in motion:

HPC containerization. As noted earlier, a major trend that promises to accelerate the use of CSP platforms is the containerization of on-premise HPC applications.

Growing choices. Major CSPs already offer many hardware, software and storage options, but competition among CSPs will escalate the number of choices offered to customers.

Multi-clouds. Large global corporations, especially, already make use of multiple CSP platforms in different parts of the world and for different purposes. This growing trend is creating a need for software products that provide single control points, allowing enterprises to manage their global on-premise and cloud resources as if they were a single, seamless environment.

The long-term goal is for these environments to send user jobs to the most appropriate resource, without the users needing to be concerned about this. Multi-cloud management creates challenges that CSPs will need to overcome:

Managing applications across multiple clouds is difficult.

Managing expenditures across multiple clouds is also challenging.

Managing a separate software environment for each CSP can be inefficient.

New software tools. We also expect strong growth in the development of software tools for model development and training in conjunction with AI methods, as well as for domain-specific science.

GOOGLE CLOUD PLATFORM (GCP)

Providing a detailed technical evaluation of the Google Cloud Platform (GCP) is beyond the scope of this paper. Our aim instead is to present GCP's highlights. Taken together, these should confirm that GCP is an important, rising player in the emerging world of HPC cloud computing. In a recent Hyperion Research global study, for example, 29% of the surveyed HPC sites said they are running HPC workloads on the Google Cloud Platform. That is a substantial increase from a similar study we conducted three years ago, when GCP was just starting to make inroads into the HPC community.

GCP Highlights

Platforms

Anthos. This is Google's hybrid cloud platform for allowing enterprises to run applications in their private data centers, in GCP and to manage multi-cloud resources that extend outside of Google Cloud to other cloud services providers, all from a single interface. What's especially interesting is that Anthos will also support third-party clouds, allowing enterprises to use a single platform, running on Google Cloud, to deploy and manage their applications on any cloud. Enterprises will get a single bill and have a single dashboard to manage their applications. This is an important advance in the multi-cloud arena.

Google Genomics. Google offers a full platform of tools designed specifically for processing and analyzing genomic data at a massive scale.

Google AI Platform. Designed for quick onboarding of organizations to the AI world and for ease-of-use thereafter, the company's AI platform provides a range of software tools to facilitate model development, testing, training and production, along with resources for domain scientists, data scientists and engineers.

Infrastructure

Custom Machine Types. Google Cloud offers flexible configurations with custom machine types, enabling users to add the exact number of cores and amount of memory required for their workload. Users are not limited to predefined instances.

Compute-optimized Virtual Machines (COVM). The new compute-optimized virtual machines are a purpose-built family of instance types for high performance computing workloads. Google reports that COVMs deliver up to a 40% performance improvement compared to current GCP VMs, offer a 3.1 GHz base frequency, 3.8 GHz sustained all-core-turbo clock frequency, and provide full transparency into the architecture of the underlying server platforms, letting users

fine-tune the performance. Users can choose COVMs with up to 60 vCPUs, 240 GBs of memory, and up to 3TB of local storage.

General Purpose Virtual Machines. Google Cloud announced two additions to their general-purpose instance types this August:

AMD EPYC Instances. These instances will support compute workloads which require high memory bandwidth, and offer a 2.25Ghz base frequency, 2.7Ghz all-core-turbo frequency, and 3.3Ghz single-core turbo frequency. These instances are designed to scale to over 200 vCPUs, support custom machine types, and offer RAM-to-vCPU ratios from 1 to 8.

Intel Xeon Scalable Processors. Google Cloud added the 2nd Generation Xeon Scalable Processors to Google Compute Engine's general-purpose machine types. Currently in beta, the new general-purpose machine types (N2), according to Google, offer greater than 20% price-performance improvement for many workloads and support up to 25% more memory per vCPU compared with first generation N1 machines. N2 machine types run at 2.8GHz base frequency, and 3.4GHz sustained all core turbo.

Fair and transparent pricing. GCP offers automatic sustained use discounts (SUDs), which allow customers to save money on long-running workloads (more than 25% of a month) without the need to predict consumption and with no up-front commitment. Resource usage is pooled by time and resources to deliver the best possible discount. HPC users can also take advantage of GCP's per-second billing, commitment savings, and preemptible VMs (up to 80% cheaper than regular instance types) for batch and fault-tolerant workloads, to reduce their overall cost of running HPC workloads in the cloud.

Accelerator Technology. GCP offers a variety of GPUs (including NVIDIA's Tesla V100 and T4) alongside Cloud TPUs, so customers can have the right accelerator to suit their workloads and price-performance requirements.

Networking. Google's private network is designed to provide high performance between VMs and regions, while keeping users secure through encryption in transit and within a global fiber network. Earlier this year, Google Cloud raised the egress bandwidth cap to 32 Gbps for same-zone VM-to-VM traffic for any Skylake or newer VM with at least 16 vCPUs. This comes standard; there is no additional configuration needed to get that 32 Gbps throughput. Meanwhile, 100 Gbps Accelerator VMs are in alpha and soon will be in beta. Any VM with eight NVIDIA V100 or four T4 GPUs attached will have bandwidth caps raised to 100 Gbps.

Storage. GCP has various storage offerings to meet users' workload needs. Google Persistent Disk offers a durable, high-performance block storage for virtual machine instances while Google Cloud Storage is a simple-to-use object storage and offers high availability across all storage classes. Google Cloud Filestore also provides reliability and consistency that latency-sensitive workloads need. Google Cloud's recent acquisition of Elastifile, a provider of scalable, enterprise file storage for the cloud, aims to empower businesses to build industry-specific, high performance applications that need petabyte-scale file storage more quickly and easily.

Open Source, Partners, and Technologies

Software tools. GCP offers a broad spectrum of useful software frameworks and other tools, including (but not limited to) Tensorflow, Kubernetes, Kubeflow, SinguarityPro, and MapReduce.

Slurm on GCP. Google Cloud and SchedMD have partnered to develop a close integration between the Slurm Workload Manager and Google Cloud Platform, allowing auto-scaling Slurm clusters, and hybrid configurations with on-premise systems.

DDN Lustre. DDN/Whamcloud has partnered with Google Cloud to provide an optimized and supported Lustre parallel file system solution at the click of a button. It's already proven its powerful capabilities by reaching #8 on the IO-500 list upon its release.

GCP Case Histories

This section of the paper presents several case histories to illustrate how customers employ GCP.

The University of South Carolina Advances Research into Climate Change

The University of South Carolina's Research Computing (RC) team serves as a central resource for all research computing on campus. The team wanted a way to redesign their workflow to speed data analysis. In March 2018, the university's Molecular Microbial Biology Lab reached out to the RC team with a problem. The lab was gathering environmental samples at a coastal pond in the Bahamas to help understand how climate change impacts ecosystems. With 10 terabytes of data per sample, their researchers were collecting much more metagenomics data than they could easily process and use, which had created a large processing backlog.

After turning to Google Cloud and Googles' support team, the RC team reported a dramatic improvement. A month's worth of new samples that were expected to have taken three months to process on a local cluster took sixteen hours on GCP. Using 124,352 cores and 3,886 nodes concurrently, GCP ran the job in only 16.5 hours. Google Cloud gives the RC team tools and methods to catch up on a year's backlog of data.

eSilicon Corporation Uses GCP and Elastifile for Chip Design

eSilicon Corporation is an application-specific integrated circuit developer that designs and produces high-end semiconductors for businesses in the 5G infrastructure, artificial intelligence, and networking industries. The company needed a better way to handle peak processing loads during resource-intensive parts of the chip design process.

With GCP, including Compute Engine, and then Google Cloud partner Elastifile, eSilicon built a platform that improves its semiconductor design abilities and provides cost-effective scalability, helping to deliver better solutions to customers sooner.

eSilicon has already seen notable performance gains through its new EDA workflow solution. In internal testing, eSilicon found that Google Compute Engine outperformed its on-premises solution by nearly 15 percent. The company was also able to double the number of chip design cycles that can be concurrently run, which helped facilitate a massive improvement in time to market.

Clear Labs: Disrupting Food Safety Testing with Google Cloud Platform

Clear Labs offers an automated, intelligent next-generation sequencing platform built for food safety testing. Forbes magazine named it one of the 25 most innovative agtech startups of 2018. The company was founded in 2014, has 50 employees, and is headquartered in Menlo Park, California.

Clear Labs uses the Google Genomics Pipeline API to process huge volumes of genomic data; Google Compute Engine as a high-performance virtual machine infrastructure to run its bioinformatics pipeline and Kubernetes Engine to run and manage the company's services and app; Cloud Pub/Sub as a foundation for stream analytics; Cloud SQL for managing databases; and Cloud Storage for unified object storage.

With GCP, Clear Labs is now delivering pathogen test results to enterprise food producers in hours instead of days.

“Google Cloud Platform gives us the infrastructure to scale and quickly process a huge amount of data. No other cloud provider comes close.” —Henrik Gehrmann, Head of Engineering, Clear Labs.

FUTURE OUTLOOK

Hyperion Research forecasts that the global market for running HPC workloads in the cloud will grow quickly to $6 billion in 2023, representing about 14% 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.

The containerization of on-premise HPC environments, driven by the need to support extreme heterogeneity, will start with leadership-class supercomputers at the top of the market and wash over the HPC midrange and entry-level markets over time. This trend, already evident in next-generation HPC system architectures, will cause on-premise HPC and cloud environments to look and act more alike, easing the movement of data and applications between these environments in both hybrid and standalone deployments. The increasing penetration of HPC servers into enterprise data centers and enterprise cloud use will add to this momentum.

Hyperion Research believes that Google Cloud Platform is well positioned to benefit from and help drive the expansion of HPC cloud usage. In one of our recent cloud studies, 29% of the respondents said that they run some of their HPC workloads in GCP—an impressive advance given GCP's relative newness to the HPC cloud scene. In view of the multi-cloud trend, we are especially impressed with GCP's Anthos hybrid cloud platform that's designed to let enterprises run applications in their private data centers and manage multi-cloud resources, including third-party CSP deployments, from a single interface.

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
