Quick Take

HPC and Containers — An Intriguing Combination

Mark Nossokoff and Earl Joseph
February 2021

HYPERION RESEARCH OPINION

As more HPC users look to leverage the cloud to augment their on-prem resources or run cloud-natively, containers have emerged as a viable portable method for deploying their workloads more easily, consistently, and repeatedly anywhere (on-prem, in the cloud, or hybrid), decoupled from any one vendor's HPC system or single CSP's framework. Containers also aim to address requirements demanded by both HPC and enterprise IT datacenters (performance, automation, ease of application deployment, security, reliability, and scalability), an important attribute as enterprise IT datacenters increasingly adopt HPC infrastructure to implement their HPDA and AI workflows.

Containers are defined as "a ready-to-run software package bundled with everything needed to run an application quickly and reliably from one computing environment to another." In this case, "everything" may include code, runtime, system libraries, default settings, dependencies and more.

CURRENT SITUATION

Until now, HPC users wanting access to more heterogeneous resources that were unavailable at their own sites often looked for help to public clouds, where they stood a better chance of finding special processors or software tools that would benefit their applications. In cloud environments, these resources, once identified, can be assembled along with the problem data in virtual machines or lightweight containers. But while cloud environments make good sense for some HPC problems, sending others to cloud services providers (CSPs) can be slow and expensive.

Today, on premise HPC systems are facing a new challenge—how to support an extremely heterogeneous mix of workloads with varying requirements: simulation and analytics (machine and deep learning, graph analysis), along with multiple precision levels and other needs. Leading HPC system vendors (OEMs) are designing next-generation architectures with the goal of supporting this heterogeneity seamlessly on premises and in accessed cloud environments.

HPC on-premises and cloud environments will retain some important differences, but these environments will increasingly look and act alike as on-premises architectures feature container-like partitions to concurrently run workloads with different requirements. A single on-premises workflow might traverse a dozen or more lightweight containers that are outfitted on the fly with appropriate hardware, software, and storage resources for the task at hand. The growing functional resemblance between on-premises and cloud environments will undoubtedly make it easier for HPC sites to exploit clouds, including the minority of sites that have not used the cloud yet.

HPC users are becoming more receptive to running their workloads in the cloud. While many HPC applications do not run well in clouds today, many do, as clouds become more capable at running a

larger set of HPC applications effectively. Recent Hyperion Research studies indicate that roughly 20% of HPC users' HPC-enabled AI workloads are currently being run in the cloud, and HPC users are forecast to spend almost $9 billion in the cloud by 2024, reflecting a robust 17.6% CAGR. Table 1 shows HPC users' forecasted spending across the broader market HPC segments, including spending in the cloud.

TABLE 1

Revenues by the Broader HPC Market Areas ($M)

	2018	2019	2020	2021	2022	2023	2024	CAGR 2019-2024
Server	$13,675	$13,710	$11,846	$13,295	$15,817	$17,942	$19,044	6.8%
Storage	$5,381	$5,427	$4,772	$5,410	$6,519	$7,577	$8,099	8.3%
Middleware	$1,590	$1,613	$1,402	$1,576	$1,902	$2,171	$2,317	7.5%
Applications	$4,652	$4,689	$4,062	$4,455	$5,258	$5,862	$6,111	5.4%
Service	$2,248	$2,239	$1,899	$2,040	$2,366	$2,587	$2,643	3.4%
Public Cloud Spend	**$2,466**	**$3,910**	**$4,300**	**$5,300**	**$4,600**	**$7,600**	**$8,800**	**17.6%**
Total On and Off Premises Revenue	$30,012	$31,588	$28,281	$32,076	$36,462	$43,739	$47,014	8.3%

Source: Hyperion Research, 2021

Flexibility and portability, key enablers for HPC users running their workloads in the cloud, are helping drive the growth in HPC cloud spending. Whether utilizing cloud for access to unique resources or as "surge" resources when their on-prem resources are either at 100% utilization or job queue times are too long, HPC users require a mechanism to seamlessly move their workloads securely, reliably, and predictably from one operating environment to another and independently of the underlying infrastructure. At the same time, scale and performance must be maintained to achieve desired runtimes on dynamic and portable workloads.

Container technology was developed to provide such a mechanism. Originally developed for and adopted by users for their stateless applications (temporary, dynamic sets of resource elements that are dispersed when the runtime is complete), containers were inaccessible to many users' workloads that required stateful, persistent elements (such as volumes and storage) to be available and shared between multiple containers and container runtimes. Data intensive AI/ML/DL workloads running on GPUs is one example. Kubernetes has emerged as one of the leading container orchestration services to manage underlying stateless and stateful container systems. The Kubernetes user community quickly recognized the need for stateful elements and created the Container Storage Interface (CSI). CSI was developed as a standard mechanism for exposing arbitrary block and file storage systems to containerized workloads. CSI allows users to have access to multiple third-party storage providers via

an open standard, providing a more secure and reliable system without requiring modifications to core Kubernetes code. Scale and performance continue to improve as container solutions evolve.

Kubernetes solutions and CSI plug-ins are now available from multiple storage vendors, but few are directly addressing the HPC market. One example of an HPC Kubernetes container solution and CSI implementation is from WekaIO. Targeting AI/ML applications and MLOps platforms, Weka's CSI plug-in aims to provide the performance and scale demanded of AI/ML/DL workloads while also providing the simplicity and ease of use required for traditional enterprise IT datacenters adopting HPC-enabled AI infrastructure.

FUTURE OUTLOOK

Users adopting AI/ML/DL and other modern workloads will require high performance shared access to large data sets in a seamless, reliable, secure, consistent, easy-to-use fashion. Both HPC datacenters and traditional enterprise IT datacenters will be well-served to explore leveraging containers and deploying new applications on top of Kubernetes to fully exploit heterogenous on-prem, cloud, and hybrid infrastructures. Extreme care should be taken to ensure workloads are appropriate for the scale and performance afforded by any particular container solution.

A variety of container solutions are available from vendors, ranging from general purpose solutions designed to address a disparate set of heterogenous workloads to focused solutions optimized for more homogenous workloads. Vendors such as WekaIO, who are targeting performance, scale, and ease of use as the primary areas of focus, should be well-positioned to provide HPC and AI container-based solutions.

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

Copyright Notice

Copyright 2021 Hyperion Research LLC. Reproduction is forbidden unless authorized. All rights reserved. Visit www.HyperionResearch.com to learn more. Please contact 612.812.5798 and/or email info@hyperionres.com for information on reprints, additional copies, web rights, or quoting permission.