HYPERION RESEARCH

HYP_Link

AWS FSx for Lustre Bolsters HPC and AI Capabilities with Support for EFA and NVIDIA GPU Direct Storage

Mark Nossokoff and Jaclyn Ludema
December 2024

RECENT DEVELOPMENT

At its recent re:Invent conference in Las Vegas, AWS announced its AWS FSx for Lustre service now supports its Elastic Fabric Adapter (EFA) high performance network interface and NVIDIA's GPUDirect Storage (GDS). These new capabilities aim to reduce overall workload costs and substantially improve workload completion times by reducing inter-node latency for distributed training, more efficiently scaling across multiple GPUs, and eliminating CPU bottlenecks by creating a direct path between GPU and storage and reducing memory copying operations.

ANALYST COMMENTARY

According to recent Hyperion Research studies, respondents are increasing cloud utilization for their HPC-AI workloads (from 27.2% today to 31.1% in 12-18 months) and more intend to run 50% or more of their HPC-AI workloads in the cloud (from 18.8% today to 22.2% in 12-18 months). This announcement shows AWS is addressing the growing demand for cloud-based HPC-AI resources, especially considering other recent AWS announcements such as the Parallel Computing Service (PCS), and Graviton buildout (50% of new compute instance capacity in the last year), among others. AWS is not alone with growing investments to support the rising demand. Google recently released its Parallelstore high performance, managed parallel file service. Hyperscalers have also announced compute instances based on the latest offerings from some or all of NVIDIA (H200, GB200), AMD (EPYC, MI300), and Intel (Xeon, Gaudi 3), not to mention new versions of their own internally developed processors and accelerators.

The strong investment in HPC-AI solutions by the wide range of full service hyperscalers (e.g., AWS, Microsoft Azure, Google Cloud, IBM Cloud HPC, Oracle Cloud Infrastructure) and AIaaS/GPUaaS providers (e.g., Coreweave, Taiga Cloud) further highlights HPC-AI cloud services have reached mainstream adoption and are a well-established and indispensable element for users. Having effectively reframed the question from "…which is better, cloud or on-premises…", to "…how can I leverage continuum computing to balance my cloud and on-premises utilization and investments to most effectively achieve my desired outcomes…?", users may accelerate the time to achieve their desired outcomes in a more efficient manner, fully leveraging HPC-AI resources wherever they may be located.