VAST + Lenovo

UNLEASH AI TRAINING AND INFERENCING

Artificial intelligence (AI) workloads, including Machine Learning, Deep Learning (DL), and Inference workloads present a new set of storage considerations requiring architects to rethink data center infrastructure. Modern DL training workloads such as Computer Vision and Natural Language Processing (NLP) predominantly rely on unstructured data, including video, images, and sound, that can only be optimized by training them on the largest data sets possible.

These data-intensive neural networks require performant flash infrastructure to randomly read across massive amounts of files. Unfortunately, legacy NAS solutions and parallel file systems built on spinning-disks fail to deliver the performance necessary for random data access required across the lifecycle of an AI application. And although all-flash configurations for these same storage systems do exist, they are cost-prohibitive at petabyte-plus scale, forcing organizations to implement multiple storage tiers, introducing additional data management complexity.

ENTER VAST DATA & LENOVO AI SOLUTIONS

VAST Data's Universal Storage and Lenovo SR670 V2 systems equipped with NVIDIA A100 Tensor Core GPUs provides a turnkey, high-performance all-flash solution for training and inferencing workloads across 100s of petabytes. Jointly built and tested by Lenovo and VAST Data, this solution delivers predictable performance to meet the needs of your applications in real time, across the entire AI lifecycle – from data ingestion to model development and validation before launching to production. Now, you can finally consolidate ALL your data on a single tier of affordable flash, thereby accelerating inference and training and unlocking previously impossible insights.

WHY VAST+LENOVO

UNPARALLELED PERFORMANCE AT SCALE

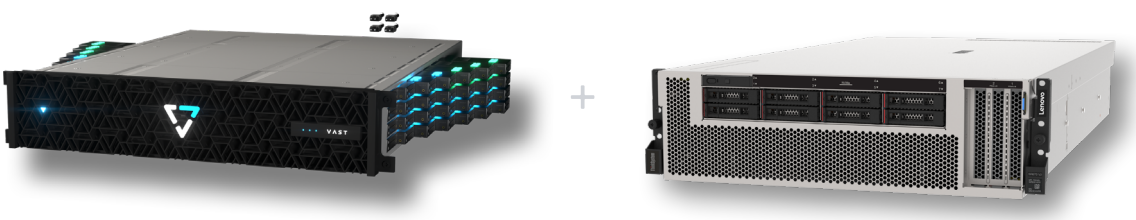
Scale to TB/s, millions of IOPS; 100GB/s+ per AI client over Ethernet or InfiniBand

OPERATIONAL SIMPLICITY

Turnkey Lenovo Scalable Infrastructure (LeSI) with VAST's all-flash Universal Storage NAS appliance delivers integrated environment with no client software dependency or complex management

ARCHIVE ECONOMICS

Radical flash storage economics to make flash affordable for all AI datasets at petabyte scale



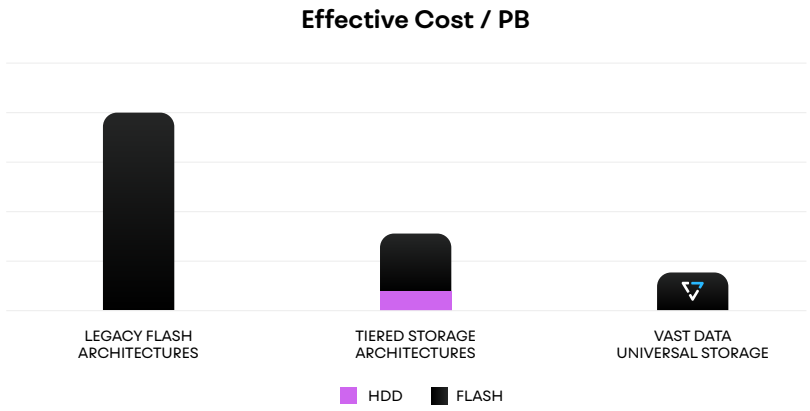
LENOVO SR670 SYSTEM

The Lenovo ThinkSystem SR670 V2 is a versatile GPU-rich 3U rack server that supports eight double-wide NVIDIA A100 Tensor Core GPUs, or the NVIDIA HGX A100 4-GPU offering with NVLink switch all-to-all GPU connection and Lenovo Neptune hybrid liquid-to-air cooling. The server is based on the new third-generation Intel Xeon Scalable processor family (formerly codenamed "Ice Lake") and the new Intel Optane Persistent Memory 200 Series. By using the processing power of the NVIDIA GPUs the server delivers optimal performance for machine learning (ML) and deep learning (DL) to extract greater insights and drive innovation across an array of industries. Also, the SR670 V2 employs Lenovo Neptune liquid-to-air (L2A) hybrid cooling technology to remove the heat of the HGX NVIDIA A100 GPUs through a unique closed loop liquid-to-air heat exchanger that delivers the benefits of liquid cooling such as higher density, lower power consumption, quiet operation, and higher performance without adding plumbing.

Dense & Flexible	3U Form Factor, up 8 double wide GPUs
Cooling with no datacenter changes	Hybrid Liquid-to-Air
NVLink	Supported
GPU Direct Storage (GDS)	Supported
NFS Multipath & RDMA	Supported

VAST DATA UNIVERSAL STORAGE

VAST Data, the pioneer of the Universal Storage concept, delivers all-flash storage systems for a wide range of workloads and use-cases. While built for high-performance, VAST eliminates the tradeoff between storage capacity and performance by delivering high-capacity and high-performance, at a price-point significantly lower than traditional all-flash storage systems.



Universal Storage brings flash TCO in line with tiered storage approaches to de-risk customer storage decisions all while freeing up HW capital for AI processors.

Our formula for compounded flash savings: QLC + 2.5% Erasure Codes + Similarity-Based Data Reduction

VAST Data's LightSpeed, both a product and storage philosophy for AI-based workloads, is optimized for both simplicity and speed. New VAST Data Universal Storage clusters built using LightSpeed enclosures deliver 2X the I/O throughput compared to previous enclosures. In addition, VAST supports several NFS client-side enhancements to achieve best-in-class throughput for AI applications and GPU servers without requiring complex parallel file systems.

NFS-over-RDMA

The Network File System (NFS) is a popular protocol used for accessing files remotely over networks and is widely used in enterprise environments for numerous applications because of its simplicity. In specialized environments, however, NFS has historically been deemed too slow for high-performance requirements. But by adding support for Remote Direct Memory Access (RDMA), NFS can now deliver greater efficiency of data transfers by bypassing client-side CPU and memory when leveraging InfiniBand or RDMA-over-Converged Ethernet (RoCE) networks. This results in higher throughput and lower latency while preserving the operational simplicity and familiarity with this industry standard storage protocol.