NVIDIA DGX SuperPOD: VAST

Reference Architecture

Featuring NVIDIA DGX H100 and DGX A100 Systems

RA-11389-001 v1

June 2023

Abstract

The NVIDIA DGX SuperPOD™ with NVIDIA DGX™ H100 or DGX A100 systems is an artificial intelligence (AI) supercomputing infrastructure, providing the computational power necessary to train today's state-of-the-art deep learning (DL) models and to fuel future innovation. The DGX SuperPOD delivers groundbreaking performance, deploys as a fully integrated system, and is designed to solve the world's most challenging computational problems.

This DGX SuperPOD reference architecture (RA) is the result of collaboration between DL scientists, application performance engineers, and system architects to build a system capable of supporting the widest range of DL workloads. The groundbreaking performance delivered by the DGX SuperPOD with DGX systems enables the rapid training of DL models at great scale. The integrated approach of provisioning, management, compute, networking, and fast storage enables a diverse, multi-tenant system that can span data analytics, model development, and AI inference.

The VAST Data Platform was evaluated for suitability for supporting DL workloads when connected to the DGX SuperPOD. The VAST Data Platform is the first enterprise NAS solution certified for DGX SuperPOD, providing the performance and scalability of parallel file system based architectures, but with the simplicity and ease of use of an enterprise appliance. This fully integrated, turnkey architecture is validated with the DGX SuperPOD, and VAST Data Platform and is fully supported by VAST support services.

Regardless of previous HPC experience, DGX SuperPOD with the VAST Data Platform is designed to make large-scale AI simpler, faster, and easier to manage for every organization and their IT team. The VAST Data Platform as an enterprise NAS solution has key capabilities that benefit organizations as they elevate their AI initiatives to DGX SuperPOD scale:

- Exascale NFS that provides all the performance required for the most demanding HPC and AI workloads without the complexity of parallel file system solutions.
- Seamlessly scale the infrastructure as your data sets and performance requirements grow with zero downtime for upgrades or expansions.
- Exabyte-scale all NVMe namespace with archive tier economics that eliminates data silos and tiers so that every I/O is served in real time.
- Native multi-protocol support enables NFS, SMB, and S3 access to all data without gateways or addons.

Learn more about the NVIDIA / VAST partnership at: https://vastdata.com/nvidia.

Contents

Storage Overview

Training performance may be limited by the rate at which data can be read and reread from storage. The key to performance is the ability to read data multiple times, ideally from local storage. The closer the data is cached to the GPU, the faster it can be read. Both persistent and nonpersistent storage needs to be designed so as to balance the needs of performance, capacity, and cost.

Storage Caching Hierarchy

The storage caching hierarchy for DGX systems is shown in Table 1. Depending on data size and performance needs, each tier of the hierarchy can be leveraged to maximize application performance.

Table 1.DGX system storage and caching hierarchy

Storage Hierarchy Level	Technology	Total Capacity[1]	Performance[1]
RAM	DRAM	2 TB[2]	> 200 GiB/s
Internal storage	Flash storage	30 TB[3]	> 50 GiB/s

1. Total capacity and performance values are per system.
2. Shared between the operating system, application, and other system processes
3. PCIe Gen 4 NVMe SSD storage

Caching data in local RAM provides the best performance for reads. This caching is transparent once the data is read from the filesystem. While local storage is fast, it is not practical to manage a dynamic environment with local disk alone in a multi-node environment. Functionally, centralized storage may be as quick as local storage on many workloads.

Storage Performance Requirements

Performance requirements for high-speed storage greatly depend on the types of AI models and data formats used. The DGX SuperPOD has been designed as a capability-class system that can manage any workload both today and in the future. However, if systems are going to focus on a specific workload, such as natural language processing (NLP), it may be possible to better estimate performance needs of the storage system.

To allow for customers to characterize their own performance requirements, some general guidance on common workloads and datasets is shown Table 2.

Table 2: Characterizing different I/O workloads

Storage Performance Level Required	Example Workloads	Dataset Size
Good	NLP	Most to all datasets fit in cache
Better	Image processing with compressed images, ImageNet/ResNet-50	Many to most datasets can fit within the local system's cache.
Best	Training with 1080p, 4K, or uncompressed images, offline inference, ETL	Datasets are too large to fit into cache, massive first epoch I/O requirements, workflows that only read the dataset once

Performance estimates for the storage system necessary to meet the guidelines in Table 2 are in:

- Table 4 of the NVIDIA DGX SuperPOD Reference Architecture—DGX H100 Systems.
- Table 8 of the NVIDIA DGX SuperPOD Reference Architecture—DGX A100 Systems.

Achieving these performance characteristics may require the use of optimized file formats such as TFRecord, RecordIO, or HDF5.

The high-speed storage provides a shared view of an organization's data to all systems. It needs to be optimized for small, random I/O patterns, and provide high peak system performance and high aggregate filesystem performance to meet the variety of training workloads an organization may encounter.

About the Vast Data Platform

The VAST Data Platform meets performance characteristics with an architecture that delivers not just speed and scale but also operational efficiency and ease of use such that any IT team can deploy and support large-scale AI initiatives with DGX SuperPOD.

Figure 1. Highly available, NVIDIA BlueField DPU integrated NVMe enclosure

VAST challenges the long-held assumption that NFS performance is inadequate for AI and HPC workloads. The VAST Disaggregated, Shared-Everything architecture consists of two building blocks that are scaled across a common NVMe fabric. First, the state and storage capacity of the system is built from resilient, high-density NVMe-oF storage enclosures. Second, the logic of the system is implemented by stateless containers that each can connect to and manage all the media in the enclosures. By disaggregating compute from storage, it is possible to spread I/O across the system to achieve levels of parallelism for massive performance.

Other benefits of the VAST Data Platform for DGX SuperPOD include:

- Independent scaling of performance and capacity with support for mixed generations of hardware in a single exabyte-scale namespace.
- Native support for both InfiniBand and Ethernet in the same namespace.
- Archive tier economics via next-generation global data reduction algorithms, support for hyperscale QLC flash with ten years of endurance, and ultra-efficient locally decodable erasure codes.
- Non-disruptive, online system expansions and software upgrades.
- Encryption, authentication, and external key management.
- VAST Catalog, a built-in metadata index allows customers to find and manage data via SQL queries.
- Enterprise-grade data protection with support for n-1 and 1-n replication topologies and up to 1 million ransomware-proof snapshots.
- The VAST Data GUI management interface provides thousands of metrics via an API-first architecture unlocking real-time visibility into performance metrics.

Validation Methodology

Three classes of validation tests are used to evaluate a particular storage technology and its configuration for use with the DGX SuperPOD: microbenchmark performance, real application performance, and functional testing. The microbenchmarks measure key I/O patterns for DL training and are designed so they can be run on CPU-only nodes. This reduces the need for large GPU-based systems to validate storage. Real DL training applications are then run on a DGX SuperPOD to confirm that the applications meet expected performance. Beyond performance, storage solutions are evaluated for robustness and resiliency as part of functional testing.

The NVIDIA DGX SuperPOD storage validation process leverages a “Pass or Fail” methodology. Specific targets are set for the microbenchmark test. Each benchmark result is graded as good, fair, or poor. A passing grade is one where at least 80% of the tests are good, and none are poor. In addition, there must be no catastrophic issues created during testing. For application testing, a passing grade is one where all cases complete within 5% of the roofline performance set by running the same tests with data staged on the DGX RAID. For functional testing, a passing grade is one where all functional tests meet their expected outcomes.

Microbenchmarks

In the Storage Performance Requirements section, there are several high-level performance metrics that storage systems must meet to qualify as a DGX SuperPOD solution. Current testing requires that the solutions meet the “Best” criteria discussed in the table. In addition to these high-level metrics, several groups of tests are run to validate the overall capabilities of the proposed solutions. These include single-node tests where the number of threads is varied and multi-node tests where a single thread count is used and as the number of nodes vary. In addition, each test run in both Buffered and DirectIO modes and when I/O is performed to separate files or when all threads and nodes operate on the same file.

Four different read patterns are run. The first read operation is sequential where no data is in the cache. The second read operation is executed immediately thereafter to evaluate the ability for the filesystem to cache data. The cache is purged and then the data are read again, this time randomly. Lastly, the data is reread again randomly, to evaluate data caching.

The IOR benchmark for single-node and multi-node tests was used.

Hero Benchmark Performance

The hero benchmark helps establish the peak performance capability of the entire solution. Storage parameters, such as filesystem settings, I/O size, and controlling CPU affinity, were tuned to achieve the best read and write performance. Storage devices were expected to demonstrate that quoted performance was close to measured performance. Other tests are crafted to demonstrate performance of real workloads.

The delivered solution for a single SU had to demonstrate over 20 GiB/s for writes and 65 GiB/s for reads. Ideally, the write performance should be at least 50% of the read performance. However, some storage architectures have a different balance between read and write performance, so this is only a guideline and read performance is more important than write.

Single-Node, Multi-File Performance

For single-node performance, I/O read and write performance is measured by varying the number of threads in incremental steps. Each thread writes (and reads) to (and from) its own file in the same directory.

For single-node performance tests, the number of threads is varied from 1 to the ideal number of threads to maximize performance (typically more than half the cores 64, but no more than the total physical cores, 128). The I/O size is varied between 128 KiB and 1 MiB and the tests are run with Buffered I/O and Direct I/O.

The target performance for these tests is shown in Table 3.

Table 3. Single-node, multi-file performance targets

Thread Count	Buffered or DirectIO	I/O size (KiB)	Performance (MiB/s)				
			Write	Read	Reread	Random Read	Random Reread
1	Buffered	128	512	1,024	1,536	256	1,536
1	Buffered	1024	800	3,072	4,608	768	1,024
1	Direct	1024	1,024	1,024	N/A	1,024	N/A

When maximizing single-node performance, the thread count may vary, however it is expected that performance does not drop significantly when additional threads are used beyond the optimal thread count.

Target performance for single-node performance with multiple threads is in Table 4. The optimal number of threads may vary for any storage configuration.

Table 4. Single-node, multi-threaded performance targets

Thread Count	Buffered or DirectIO	I/O size (KiB)	Performance (MiB/s)				
			Write	Read	Reread	Random Read	Random Reread
Varies	Buffered	128	8,000	12,000	18,000	12,000	18,000
	Direct	128	8,000	15,000	N/A	15,000	N/A
	Buffered	1024	10,000	20,000	30,000	20,000	30,000
	Direct	1024	10,000	20,000	N/A	20,000	N/A

Reread performance relative to read performance can vary substantially between different storage solutions. The reread performance should be at least 50% of the read performance for both sequential and random reads.

Multi-Node, Multi-File Performance

The next test performed is multi-node I/O read and write test to make sure that the storage appliance can provide the minimum required buffered read and write per system for the DGX SuperPOD. This benchmark determines the capacity of a filesystem to scale performance of different I/O patterns. Performance should scale linearly from one to a few nodes, reach a maximum performance, and not drop off significantly as more nodes are added to the job.

The target performance for a single SU of 20 nodes is 65 GiB/s for reads with I/O size of 128 KiB or 1,024 KiB, and if the I/O is Direct or Buffered. The write performance should be at least 20 GiB/s, but ideally it would be 50% of the read performance. Results from these tests must be interpreted carefully as it is possible to add more hardware to achieve these levels. Overall performance is the goal, but it is desirable that the performance comes from an efficient architecture that is not over-designed for its use.

Single-File I/O Performance

A key I/O pattern is reading data from a single file. Often the fastest way to read data is when all the data is organized into a single file, such as the RecordIO format. This can often be the fastest way to read data because it eliminates any of the open and close operations required when data are organized into multiple large files. Single-file reads are a key I/O pattern on DGX SuperPOD configurations.

Targeted performance and expected I/O behavior is that the single-node, multi-threaded, writes can successfully create the file, that sequential read and random read performance is good, and that read performance scales as more nodes are used. Multi-node, multi-threaded, single file writes are not evaluated. In addition, it is expected that buffered reread performance is like the multi-file reread performance.

Target performance for single file I/O is in Table 5.

Table 5. Single-file read performance targets

Node Count	Buffered or DirectIO	I/O size (KiB)	Performance (MiB/s)			
			Read	Reread	Random Read	Random Reread
1	Buffered	128	2,500	1[1]	2,500	1[1]
1	Direct	128	15,000	N/A	15,000	N/A
1	Buffered	1024	3,000	1[1]	3.000	1[1]
1	Direct	1024	20,000	N/A	20,000	N/A
20	Buffered	128	65,000	1[1]	65,000	1[1]
20	Direct	128	65,000	N/A	65,000	N/A
20	Buffered	1024	65,000	1[1]	65,000	1[1]
20	Direct	1024	65,000	N/A	65,000	N/A

1. Reread performance of cached data should be near in performance to the results from the multi-file reread test

Application Testing

Microbenchmarks provide indications of the peak performance of key metrics. However, it is application performance that is most important. A subset of the MLPerf Training benchmarks is used to validate storage performance and function. Here, both single-node and multi-node configurations are evaluated to ensure that the filesystem can support different I/O patterns and workloads. Training performance when data is staged on the DGX RAID was used as the baseline for performance. The performance goal is for the total time to train when data is staged on the shared filesystem to be within 5% of those measured when data is staged on the local RAID. This is not just for individual runs, but also when multiple cases are run across the DGX SuperPOD at the same time.

ResNet-50

ResNet-50 is the canonical image classification benchmark. Its dataset size is over 100 GiB and it has a requirement for fast data ingestion. On a DGX system, a single node training requires approximately 3 GiB per second and the dataset is small enough that it can fit into cache. Preprocessing can vary, but the typical image size is approximately 128 KiB. One challenge of this benchmark is that at NVIDIA the processed images are stored in the RecordIO format (i.e. one large file for the entire dataset) since this provides the best performance for MLPerf. Since it is a single file, this can stress shared filesystem architectures that do not distribute the data across multiple targets or controllers.

NLP—BERT

BERT is the reference standard NLP model. In this test, the system is filled with two eight node jobs and four single node jobs (or less if not all 20 nodes of the SU are available). It is expected that the total time to train is within 5% of that measured when training from the local raid. This test does not stress the filesystem but does ensure that local caching is operating as needed.

Recommender—DLRM

The recommender model has different training characteristics than ResNet-50 and BERT in that the model trains in less than a single epoch. This means that the data set is read no more than once, and local caching of data cannot be used. To achieve full training performance, DLRM must be able to read data at over 6 GiB/s. In addition, the file reader uses DirectIO that stresses the filesystem differently than the other two files. The data are formatted into a single file.

This test is only run as a single node test; however, several tests are run where the number of simultaneous jobs vary from one to the total number of nodes available. It is expected that the shared filesystem only sustains performance up to the peak performance measured from the hero test. For 20 simultaneously cases, the storage system would have to provide of over 120 GiB/s of sustained read performance, more than what is prescribed in the Storage Performance Requirements section. Even the best performance outlined in this table is not meant to support every possible workload. It is meant to provide a balance of high throughput while not over-architecting the system.

Summary

The VAST Data Platform meets the DGX SuperPOD performance and functionality requirements. As an enterprise NAS solution, the VAST Data Platform paired with DGX SuperPOD enables customers to take on demanding AI workloads without the complexity and specialized skills typically associated with HPC storage solutions.

As requirements grow the VAST Data Platform may be seamlessly scaled by adding compute and or storage resources tailored to meet performance and capacity targets. The VAST Data Platform supports multiple generations of infrastructure in a cluster, allowing customers to mitigate supply chain issues and select from a range of VAST certified hardware solutions.

DGX SuperPOD customers can be confident that the VAST Data Platform will meet their most challenging AI workloads at any scale.

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com