NVIDIA DGX SuperPOD: IBM Spectrum Scale and ESS 3200

Reference Architecture

Featuring NVIDIA DGX A100 Systems

DG-10971-001 | May 2022

Document History

DU-10981-001

Version	Date	Authors	Description of Change
01	2022-05-27	NVIDIA: Craig Tierney and Robert Sohigian IBM: Douglas O'Flaherty and Sanjay Sudam	Initial release

Abstract

The NVIDIA DGX SuperPOD™ with NVIDIA 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 ground-breaking performance, deploys in weeks 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 ground-breaking performance delivered by the DGX SuperPOD with DGX A100 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.

In this paper, the IBM Elastic Storage System (ESS) 3200 was evaluated for suitability in supporting DL workloads when connected to the DGX SuperPOD. IBM ESS 3200 is a modern implementation of software-defined storage with low latency NVMe physical storage, advanced erasure coding, connected using InfiniBand and Ethernet networking. Multiple ESS 3200 systems can be aggregated to create a large global filesystem, or connected to multiple clusters for geographic and cross platform data sharing in a single global namespace. The IBM ESS3200 is a 2U building block that makes it easy to deploy, manage, and grow fast storage for AI with NVIDIA DGX systems.

The DGX SuperPOD is a turnkey solution validated at scale with IBM ESS 3200 systems. Joint testing and integration ensures the NVIDIA DGX SuperPOD is a rapidly deployed and a robust solution for scalable AI development. NVIDIA and IBM jointly test, plan, and install the system, with the storage backed by IBM global deployment and support services.

As configured and tested in the NVIDIA SuperPOD, IBM ESS 3200 systems can be used for all DL workloads including:

- Training models efficiently with directly from Spectrum Scale.
- Automatically leverage local resources as cache to minimizing rereading data across the network.
- Workspace for long-term storage (LTS) of datasets.
- A centralized repository for the acquisition, manipulation and sharing of results via standard protocols like NFS, SMB, and S3.

Contents

Storage Overview

Training performance can 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. Storage needs to be designed considering the hierarchy of different storage technologies, either persistent or nonpersistent, to balance the needs of performance, capacity, and cost.

The storage caching hierarchy of the DGX A100 system 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 A100 system storage and caching hierarchy

Storage Hierarchy Level	Technology	Total Capacity	Performance
RAM	DDR4	2 TB per system	> 200 GB/s per system
Internal Storage	NVMe	30 TB per system	> 50 GB/s per system

Caching data in local RAM provides the best performance for reads. This caching is transparent once the data is read from the filesystem. However, the size of RAM is limited to 2.0 TB on a DGX A100 system and that capacity must be shared with the operating system, application, and other system processes. The local storage on the DGX A100 system provides 30 TB of PCIe Gen 4 NVMe SSD storage. While the local storage is fast, it is not practical to manage a dynamic environment with local disk alone in multi-node environment. Functionally, centralized storage can be as quick as local storage on many workloads.

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

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

Table 2. Characterizing different I/O workloads

Storage Performance Level Required	Example Workloads	Dataset Storage Requirements
Good	Natural language processing (NLP)	Most 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 storage.
Best	Training with 1080p, 4K, or uncompressed images, offline inference, ETL	Datasets are too large to fit into cache.

Table 3 provides performance estimates for the storage system necessary to meet the guidelines in Table 2. To achieve these performance characteristics may require the use of optimized file formats, TFRecord, RecordIO, or HDF5.

Table 3. Guidelines for desired storage performance characteristics

Performance Characteristic	Good, GB/s	Better, GB/s	Best, GB/s
140 system aggregated system read	50	140	450
140 system aggregated system write	20	50	225
Single SU aggregate system read	6	20	65
Single SU aggregate system write	2	6	20
Single system read	2	4	20
Single system write	1	2	5

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 workloads an organization may encounter.

About the IBM ESS 3200

The IBM Elastic Storage System 3200, shown in Figure 1, combines the performance of NVMe storage technologies with the reliability and the rich features of IBM Spectrum Scale in a powerful 2U storage system that scales out for performance and capacity.

Figure 1. IBM ESS 3200

IBM Spectrum Scale on NVMe is designed to be the market leader in all-flash performance, and scalability with read performance of 80 GB/s and up-to 55 GB/s write per NVMe all-flash appliance and 100 microseconds latency. Providing data-driven multicloud storage capacity, the NVMe all-flash appliance is deeply integrated with the software-defined capabilities of IBM Spectrum Scale to seamlessly plug it into an analytics, scalable cluster, or AI workload.

Available with multiple drive options and advanced erasure coding, the IBM ESS 3200 provides options to optimize costs for different installation sizes. As with all IBM Spectrum Scale solutions, capacity and performance can be scaled. Combining ESS 3200 systems provides nearly linear performance scalability. IBM ESS 3200 solutions may also be used as an all-flash NVMe performance tier on slower storage, including tape or object storage.

IBM Spectrum Scale

IBM Spectrum Scale is an industry leader in high-performance file systems. The underlying parallel filesystem provides scalable throughput and low-latency data access, as well as superior metadata performance. Unlike other systems that can easily bottleneck, the distributed architecture of a parallel filesystem provides reliable performance for multi-user sequential and random read or write. This is particularly important in HPC and AI clusters where multiple compute nodes may need to read or write to the same file. IBM Spectrum Scale has been proven on the largest AI clusters in the world, including the US National Labs supercomputers Summit and Sierra, as well as the Circe supercomputer built by NVIDIA.

IBM Spectrum Scale is optimized for modern AI workloads with optimizations for common data patterns, NVIDIA GPU Direct® Storage, and run-time controls to align data performance with application needs. It includes optimizations for small files, multiple protocols, and metadata operations. Remarkably fast directory and file management is required for the many workloads that distribute data across multiple directories or many files.

IBM Spectrum Scale enables a common data lakehouse with mixed data types, applications, and protocols. When dealing with multiple applications or clusters, IBM Spectrum Scale creates a single namespace (or data plane) across systems. For users, it is a single repository that can access NFS, SMB, Object, or a high-performance native POSIX filesystem. This single data plane allows the data administrator, analyst, or data scientist to access all the data in place. The entire data pipeline, from ingest to inference can be completed without having to make additional copies or move data between systems. Multiple clusters can be integrated into a single namespace to provide rapid local access to logically or geographically distributed data.

IBM Spectrum Scale provides Container Native Access and Operators to support Kubernetes driven DevOps and Data Ops practices. In addition, IBM Spectrum Scale provides enterprise features such as call-home proactive support, encryption, and audit file logging that works with enterprise security information and event management (SEIM) platforms.

Validation Methodology and Results

Three classes of validation tests are used to evaluate a particular storage technology and 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 crafted so they can be run on nodes with CPU only. This reduces the need for large GPU-based systems only 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 tested for robustness and resiliency as part of a functional test.

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 A100 RAID. For functional testing, a passing grade is one where all functional tests meet their expected outcomes.

Microbenchmarks

Table 3 lists several high-level performance metrics that storage systems must meet to qualify as a DGX SuperPOD solution. Current testing requires that the solution 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 after the first to test 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 test data caching.

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 the quoted performance was close to the measured performance. For the other tests, not every I/O (or any) pattern to be able to achieve this level of performance, but it provides and understanding of the difference between peak and obtainable performance of the I/O patterns of interest.

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 4.

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

Thread Count	Buffered or DirectIO	I/O size (KiB)	Performance				
			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	-	1,024	-

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 5. The optimal number of threads may vary for any particular storage configuration.

Table 5. Single-node, multi-threaded performance targets

Thread Count	Buffered or DirectIO	I/O size (KiB)	Performance				
			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		15,000	
	Buffered	1024	10,000	20,000	30,000	20,000	30,000
	Direct	1024	10,000	20,000		20,000	

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. The purpose of this benchmark is to test a given filesystems capability to scale performance of different I/O patterns. Performance should scale linearly from 1 to a few nodes, reach a maximum performance, then 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 have to be interpreted carefully as it is possible to just 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 when all the data are 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 tested. In addition, it is expected that buffered reread performance be similar to the multi-file reread performance.

Target performance for single file I/O is in Table 6.

Table 6. Single file read performance targets.

Node Count	Buffered or DirectIO	I/O size (KiB)	Performance			
			Read	Reread	Random Read	Random Reread
1	Buffered	128	2,500	1[1]	2,500	1[1]
1	Direct	128	15,000	-	15,000	-
1	Buffered	1024	3,000	1[1]	3.000	1[1]
1	Direct	1024	20,000	-	20,000	-
20	Buffered	128	65,000	1[1]	65,000	1[1]
20	Direct	128	65,000	-	65,000	-
20	Buffered	1024	65,000	1[1]	65,000	1[1]
20	Direct	1024	65,000	-	65,000	-
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 are used to validate storage performance and function. Here, both single-node and multi-node configurations are tested to ensure that the filesystem can support different I/O patterns and workloads. Training performance when data are staged on the DGX A100 RAID was used as the baseline for performance. The performance goal is for the total to train when data are staged on the shared filesystem are within 5% of those measured when data are 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. It is dataset size is over 100 GiB and has a requirement for fast data ingestion. On a DGX A100 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, which is one large file for the entire dataset since this provides the best performance for MLPerf. Since it is a single file, this can stress share filesystem architectures that do not distribute the data across multiple targets or controllers.

Natural Language Processing–BERT

BERT is the reference standard natural language processing 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 test to 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 Table 3. 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.

Functional Testing

In addition to performance, it is necessary to ensure that the storage solution is designed to meet the requirements of DGX SuperPOD integration and that it has been designed for the highest levels of uptime. To validate the solution's resiliency, it is put through several tests to ensure that failure of one component will not interrupt the DGX SuperPOD operation. The tests are shown in Table 7.

Table 7. Functional tests

Test	Expected Outcome
Client build and installation under DGX OS	This should work as documented, with no additional manual steps
Client rebuild and upgrade under DGX OS	This should work as documented, with no additional manual steps
Client InfiniBand failure, disconnect a single InfiniBand cable used for storage access while under stress testing.	Test continues to operate successfully, with minimal pausing, possibly at lower performance.
Server InfiniBand failure, disconnect InfiniBand links from storage servers while under stress testing	Test continues to operate successfully, with minimal pausing, possibly at lower performance
Power loss to server, disconnect single power source to one or more storage servers	Test continues to operate successfully, at full performance
Server hang, pull all power from single server, while running stress tests.	Test continues to operate successfully, with minimal pausing, possibly at lower performance
Simulate server drive failure, induce drive failure (through SW or hot pull) while running stress tests.	Test continues to operate successfully, with minimal pausing, possibly at lower performance
InfiniBand switch failure, catastrophically power down single InfiniBand switch connecting storage to clients, while running stress tests.	Test continues to operate successfully, with minimal pausing, possibly at lower performance

Summary

NVIDIA evaluations show that the IBM ESS 3200 storage systems with Spectrum Scale file system meets the DGX SuperPOD performance and functionality requirements. It is a great choice to pair with a DGX SuperPOD to meet current and future storage needs.

As storage requirements grow, ESS 3200 building blocks can be added to seamlessly scale capacity, performance, and capability. The combination of NVME hardware and IBM Spectrum Scale parallel file system architecture provides excellent random read performance, often just as fast as sequential read patterns. Testing has validated that each ESS3200 can deliver the highest levels of per node performance and meet all of our application performance requirements. The IBM Spectrum Scale parallel file system provides a platform that is fully supported with the NVIDIA DGX SuperPOD.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA DGX , NVIDIA DGX POD, and NVIDIA DGX SuperPOD are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

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