DDN A³I AI400 Appliance with the EXA5 Filesystem

NVIDIA DGX SuperPOD Reference Architecture

RA-09734-001 | November 2019

Document History

RA-09734-001

Version	Date	Authors	Description of Change
001	2019-11-13	Robert Sohigian and Craig Tierney	Initial release

Abstract

The NVIDIA DGX SuperPOD™ is a first-of-its-kind artificial intelligence (AI) supercomputing infrastructure that delivers groundbreaking performance, deploys in weeks as a fully integrated system, and is designed to solve the world's most challenging AI problems.

The groundbreaking performance delivered by the DGX SuperPOD enables the rapid training of deep learning models at great scale. To create the most accurate image classification, object detection, and natural language models require large amounts of training data. This data must be accessed rapidly across the entire SuperPOD. To maximize the computational capabilities of the DGX SuperPOD, it is essential to pair the DGX SuperPOD with a storage system fitted to the task.

In this paper, the DDN® A³I AI400 appliance was evaluated for suitability for supporting deep learning (DL) workloads when connected to the DGX SuperPOD. The AI400 appliance is a compact and low-power storage solution that provides incredible raw performance with the use of NVMe drives for storage and InfiniBand as its network transport. The AI400 appliance leverages the EXAScaler file system which provides an enterprise version of the Lustre parallel filesystem which features increased hardening and additional data management capabilities.

Parallel filesystems such as Lustre simplify data access and support additional use cases where fast data is required for efficient training and local caching is not adequate. With high single-threaded and multi-threaded read performance, they be used for:

- Training when the datasets cannot be cached locally in DGX-2 system memory or on the DGX-2 NVMe RAID.
- Fast staging of data to local disk.
- Training with large individual data objects (e.g. uncompressed or lossless compressed 1080p images).
- Training with optimized data formats such as TFRecord.
- Workspace for long term storage (LTS) of datasets.

Contents

Storage Overview

Training performance can be limited by the rate at which data can be read and re-read from storage. The key to performance is the ability to read data multiple times. The closer the data are cached to the GPU, the faster they can be read. Storage needs to be designed considering the hierarchy of different storage technologies, either persistent or non-persistent, to balance the needs of performance, capacity, and cost.

Table 1 documents the storage caching hierarchy. Depending on data size and performance needs, each tier of the hierarchy can be leveraged to maximize application performance.

Table 1. DGX SuperPOD storage and caching hierarchy

Storage Hierarchy Level	Technology	Total Capacity	Read Performance
RAM	DDR4	1.5 TB per node	> 100 GB/s
Internal Storage	NVMe RAID	30 TB per node	> 25 GB/s
High-Speed Storage	Generic	Varies depending on specific needs	Required: • Aggregate system read > 32 GB/s • Aggregate system write > 16 GB/s • Single-Node read > 5 GB/s Desired: • Single-Node 1 GB/s read per GPU (16 GB/s)

Caching data in local RAM provides the best performance for reads. This caching is transparent once the data are read from the filesystem. However, the size of RAM is limited to 1.5 TB on a DGX-2 system and that capacity must be shared with the operating system, application, and other system processes. The local storage on the DGX-2 system provides 30 TB of very fast NVMe (and can be upgraded to 60TB if required). While the local storage is fast, it isn't practical to manage a dynamic environment with local disk alone.

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

Datasets today that are 30 TB are still considered large, but we see use cases in automotive and other computer vision tasks where 1080p images are used for training and in some cases are uncompressed. Datasets in these formats can easily exceed 30 TB in size. In these cases, we see a need for 1 GB/s per GPU for read performance.

The metrics above assume variety of workloads, datasets, and needs for training locally and directly from the high-speed storage system. It's best to characterize workload needs before finalizing performance and capacity requirements.

The storage hierarchy can be extended further. LTS is often important to preserve data and historical results. LTS could be as large as tens or hundreds of petabytes. Data from this tier is accessed infrequently, and store and recall performance isn't as critical. Solutions for this can be cost-optimized differently than the high-speed storage discussed above. Solutions could be based on slower spinning disk, use S3 compatible object storage technologies, or even use the cloud.

The key metric in filesystem performance for DL training is read, and more specifically re-read, performance. During DL training, data are repeatedly read over and over as the model is iterated to find the most accurate model. When the training dataset is small enough, it can be cached in the local memory or local NVMe disk, to limit access to the remote filesystem. However, as datasets grow, provisioning local systems with enough capacity to store the entire working dataset is not always practical. In these cases, it becomes necessary to train models where data are reread from the network filesystem every epoch. Handling this level of I/O on a large configuration such as the DGX SuperPOD require massive throughput of data of many I/O patterns including large blocks (greater than 1 megabyte), smaller blocks (less than 1 megabyte and even less than 32 kilobytes), and memory-mapped files. For a storage solution to meet the needs of the DGX SuperPOD, it must be able to handle these types of I/O patterns and scale to tens of gigabytes per second of read performance to all nodes simultaneously.

About the DDN AI400 Appliance

The DDN AI400 appliance (Figure 1) meets the requirements above and provides several features that are important for maximizing the performance of the DGX SuperPOD and system management in data centers:

- The DDN AI400 appliance communicates with DGX SuperPOD clients using multiple EDR InfiniBand or 100 GbE network connections for performance, load balancing, and resiliency. The DDN parallel protocol allows storage to be accessed at over 20 GiB/s, exceeding the desired goal of 1 GiB/s per GPU. This performance is necessary for training image-based networks as image sizes grow to 1080p, 4K, and beyond.
- The DDN AI400 appliance is a scalable building block which can be easily aggregated into a single filesystem that can scale seamlessly in capacity, performance, and capability.
- The DDN AI400 appliance can be configured at several different capacities ranging from 30 TiB to 240 TiB.
- The all-NVME architecture of the DDN AI400 appliance provide excellent random read performance, often as fast as sequential read patterns.

Figure 1. DDN AI400 appliance

Native InfiniBand support is a key feature that maximizes performance and minimizes CPU overhead. Since the DGX SuperPOD compute fabric is InfiniBand, designing the storage fabric as InfiniBand means only one high-speed fabric type needs to be managed, simplifying operations.

> **Note:** Since the testing published in this paper was completed, DDN released the next generation DDN AI400X appliance. The AI400X appliance has been updated to provide better IOPS and throughput. In addition, it provides Mellanox HDR100 InfiniBand connections to support next generation HDR fabrics. While the benchmarks provided here are from the current generation of products, the AI400X appliance could provide even better performance for DL storage needs.

Testing Methodology and Results

The filesystem testing tool FIO was used for all single- and multi-node bandwidth performance. MDTest was used for measuring multi-node metadata performance. In addition, some of the MLPerf benchmarks were used to evaluate performance of real applications.

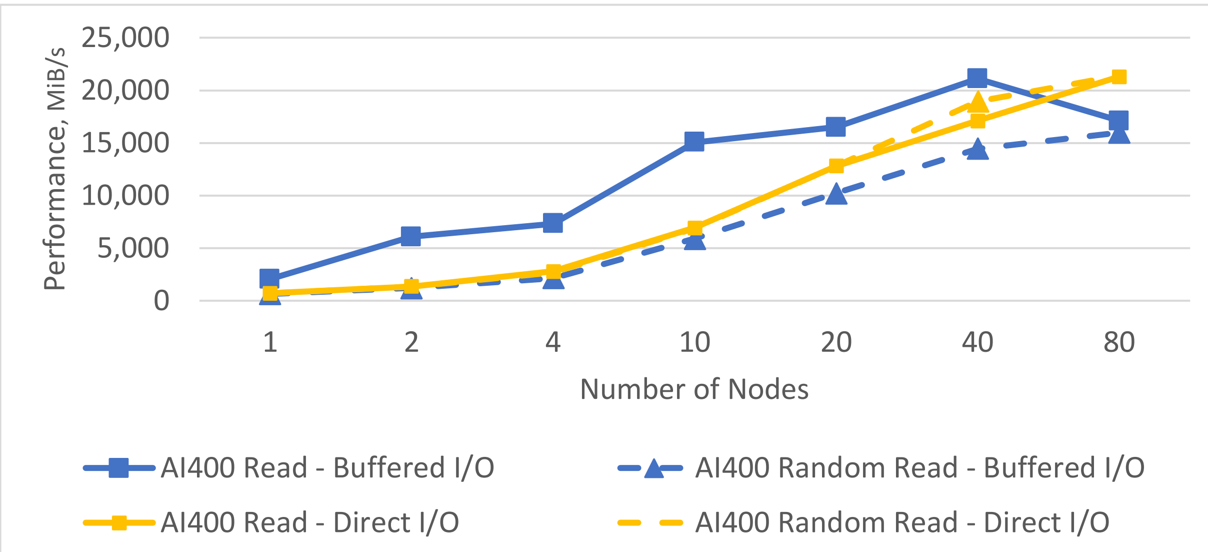
The DDN AI400 appliance was configured with EXA5 filesystem, which is based on an enterprise release of Lustre 2.12. The storage was attached to a separate InfiniBand fabric on the DGX SuperPOD. Each system of the DGX SuperPOD was connected to the storage fabric with two Mellanox InfiniBand adapters, allowing for over 20 GiB/s per system.

The following sections highlight a few of the tests used to gauge overall capability of the DDN AI400 appliance for DL workloads. The three tests are read performance by thread, read performance by node, and DL training performance. The read benchmarks best correlate to the requirements needed for DL applications. The DL training benchmark provides an example of how the different performance benefits of the NVMe based DDN AI400 appliance translates into real application performance.

Read Performance by Thread

Read performance, both sequential and random, are important for DL training. Figure 2 shows the sequential read and random read performance of the DDN AI400 appliance as thread count is increased. For buffered I/O, single thread performance of sequential reads is above 1.5 GiB/s and above 640 MiB/s for random reads. Read performance scales to over 20 GiB/s when 80 threads are used for Direct I/O. This exceeds are target performance of 16 GiB/s per node (or 1 GiB/s per GPU). Performance for sequential buffered I/O is better than direct I/O at all but the highest thread count.

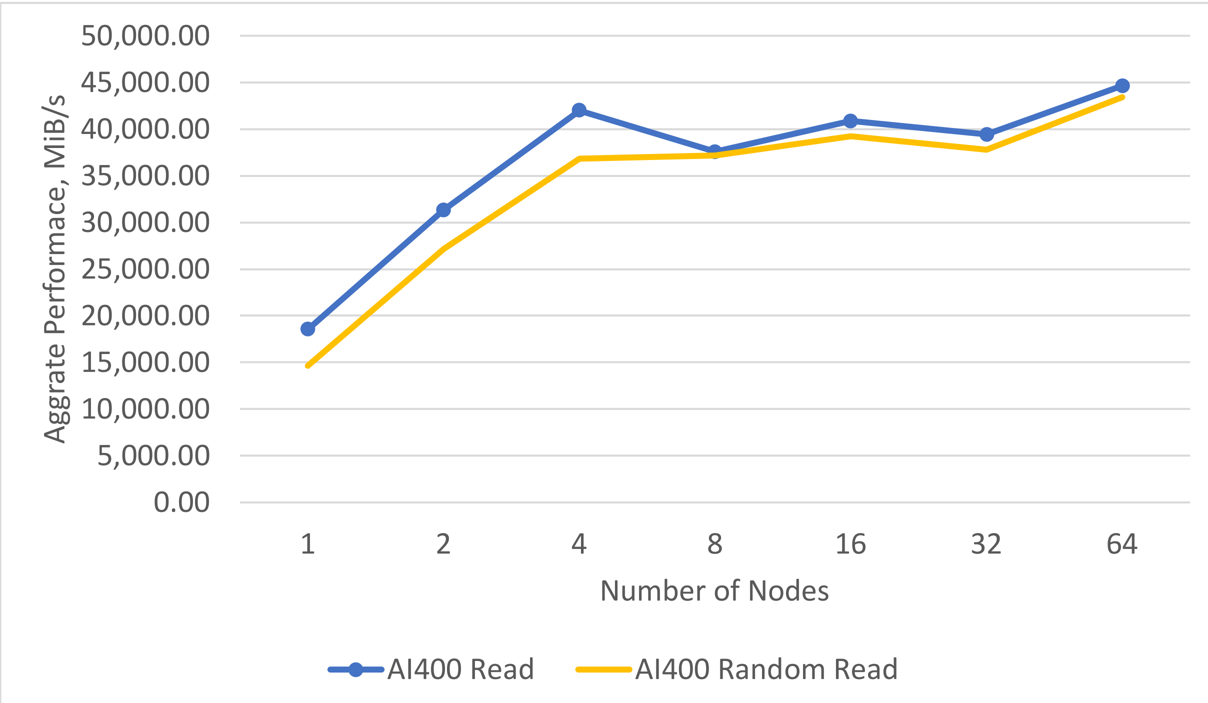
Figure 2. Sequential read and random read performance of the DDN AI400 appliance



Read Performance by Node

Figure 3 shows the system verify the peak aggregate performance of a single DDN AI400 appliance. As illustrated in the figure, large block sequential and random reads have similar performance. Single node performance is in excess of 16 GiB/s. Maximum read performance is over 44 GiB/s and is similar for both sequential and random reads. Aggregate performance should scale nearly linearly as more DDN AI400 appliances are used.

Figure 3. Read and random-read performance across the DGX SuperPOD



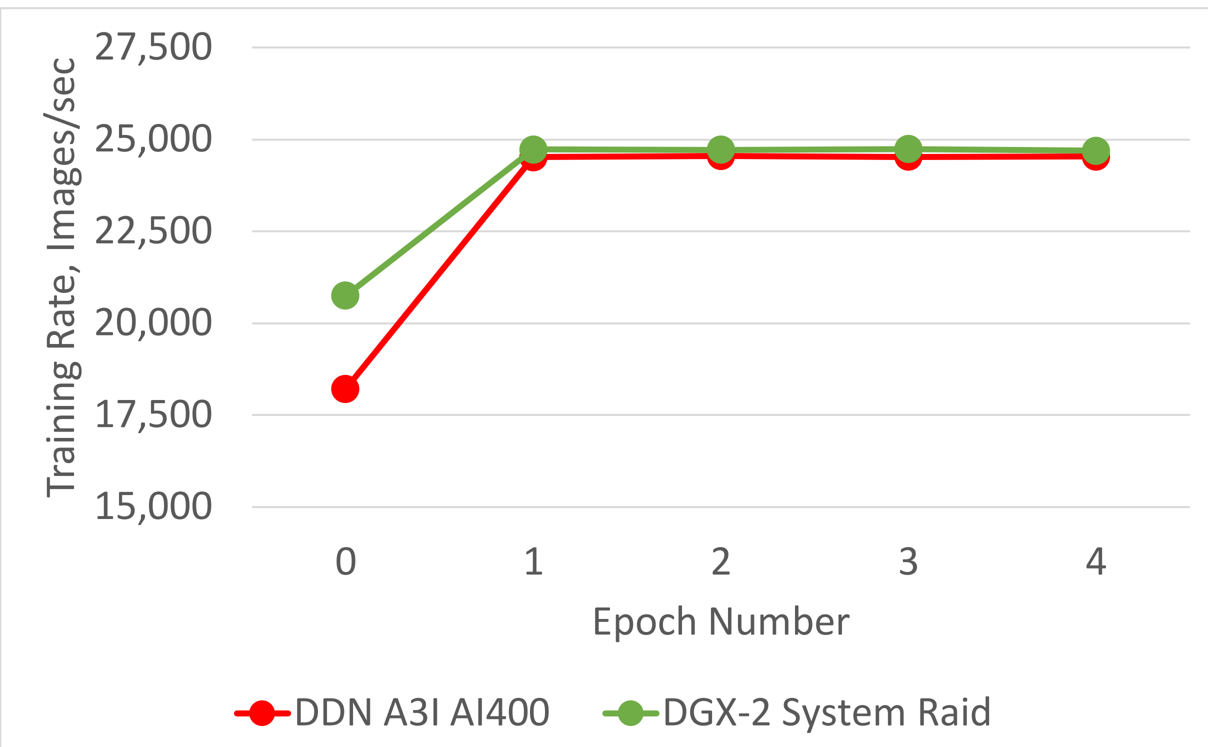
MLPerf and ResNet-50 Performance

While microbenchmarks such as FIO and MDTEST are important tools for characterizing filesystem performance, real application performance is the best metric. The MLPerf benchmarks provide a variety of different DL models with different I/O requirements that are good representations of DL workloads today.

The most I/O intense of these workloads is ResNet-50. On a DGX-2 system, the training rate exceeds 24,700 images per second. The average size of an image in the ImageNet database is 122 KiB. To achieve robust models, it is important that the data are processed randomly for each epoch. This translates into an I/O requirement of over 3 GiB/s of random reads of relatively small files across 16 GPUs. This is further complicated by the fact that the data are read as memory mapped files. Reading files as memory mapped files is a good optimization for local filesystems, but is more difficult on a network filesystem because of the performance of reads on page boundaries (in this case 4 KiB) and the overhead required to ensure that no process is trying to write to a page while another tries to read from multiple nodes.

Figure 4 shows the training performance measured in images per second for ResNet-50 from MLPerf v0.6. In this benchmark, the data are formatted as RecordIO files. RecordIO files are used by the MXNet framework. All the files are written into a single large file. RecordIO files are read using MMAP. In the NVIDIA MLPerf reference code, the NVIDIA Data Loading Library (DALI) framework is used for reading the data.

Figure 4. Training performance of MLPerf v0.6 and ResNet-50



Results are shown where data are read from the local RAID disks on the DGX-2 system and read from the DDN AI400 appliance. The first item to note is the uncached performance when training on the RAID. Even when reading from the local RAID , the uncached epoch (epoch 0) is slower than subsequent epochs when the data are cached. The RAID disks on the DGX-2 system are capable of sustaining over 25 GiB/s reads. Achieving good performance is more than just read performance.

The DDN AI400 appliance can achieve 87% of the RAID performance for the first epoch. Since the files are small and mmap is used, this performance is considered excellent. In addition, when the data are cached, training performance is within 1%. The DDN A^3I solution is providing excellent performance for the most challenging MLPerf benchmark in terms of I/O requirements.

Summary

The DDN A^{3}I AI400 appliance with EXA5 is a great building block for scalable shared high-performance storage for the DGX SuperPOD. The DDN AI400 appliance is a scalable building block which can be easily aggregated into a single filesystem that can scale seamlessly in capacity, performance and capability. The all-NVME architecture of the DDN AI400 appliance provide excellent random read performance, often as fast as sequential read patterns. It provides all this performance in a 2U building block that meets all the performance requirements of the DGX SuperPOD. In addition, it meets larger desired goals of 16 GiB/s per node (or 1 GiB/s per GPU).

The DDN AI400 appliance is a great choice to pair with a DGX SuperPOD to meet current and future storage needs.

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051

http://www.nvidia.com