Scalable Architecture for Artificial Intelligence and Deep Learning with NVIDIA® DGX-1™

CONTENTS

Introduction 1
NVIDIA DGX-1 Server 2
DDN AI200 Parallel Storage Appliance 3
DDN A^{3}I Reference Architecture 4
Scaling with GPUs and the DDN A^{3}I Reference Architecture 9
DDN A^{3}I Reference Architecture Performance Testing 10
Parallel Architecture Performance for DGX-1 containers 16
Conclusion 19

POWERED BY NVIDIA

Introduction

Artificial Intelligence (AI) is rapidly becoming an essential business and research tool, giving organizations valuable new insights into their data and doing so with high velocity and accuracy. Tremendous resources are being invested by enterprises, universities, and government organizations to further develop and benefit from AI and Deep Learning (DL). With AI technology, autonomous vehicles can circulate unassisted in our cities, real-time fraud detection protects our shopping and internet transactions, natural language translators remove language barriers, augmented reality delivers a far richer entertainment experience, drug discovery gets accelerated, and personalized medicine and remote health diagnostics are fully enabled.

AI and DL are however creating the toughest workloads in modern computing history, quite distinct from those in the traditional enterprise. They place significant strain on the compute, storage and network with ever expanding sets of hot and warm data. An AI-enabled datacenter must be able to service the spectrum of activities involved in the AI and DL process from data ingest and retention through training, verification and inference. The IT infrastructure supporting the AI-enabled datacenter should also accommodate the moving target nature of a developing service; starting small yet allowing real choices of growth strategy as data volumes grow and application workloads become more intensive and diverse.

Breakthrough technologies in processors and storage are acting as catalysts for production AI and DL. Graphical Processing Units (GPUs), such as the NVIDIA® Tesla™ V100, deliver large speedups over CPUs, while Flash Enabled Parallel Storage provides a significant performance boost over traditional hard disk-based storage. The ability to train neural networks with mixed-precision methods permits massively parallel execution of data processing tasks across simpler computational cores. The high-core count of the modern GPU architecture is uniquely suited to provide significant acceleration to AI and DL applications resulting in faster training, and inference capabilities. To ensure maximum application productivity, the data storage and network infrastructure must deliver true end-to-end parallelism from disk to GPU, with high throughput and low latency.

This paper presents the DDN A^3I (Accelerated, Any-Scale AI) scalable architecture which integrates NVIDIA® DGX-1™ servers with DDN AI200 all-flash parallel file storage appliances.

NVIDIA DGX-1 Server

Figure 1. NVIDIA DGX-1 server

The NVIDIA® DGX-1™ is a purpose-built appliance optimized for AI and DL applications. Eight NVIDIA® Tesla® V100 GPUs provide one petaFLOPS of DL training performance and are configured in a hybrid cube mesh topology using NVIDIA® NVLink™ technology. This high-bandwidth, low-latency fabric ensures optimal GPU-to-GPU communication, and maximizes the efficiency of multi-GPU training by removing the bottlenecks of traditional PCIe interconnects. The DGX-1 also includes four high-speed network ports that can be set to either EDR IB or 100 GbE for optimal storage-to-DGX data delivery.

NVIDIA provides containerized versions of popular DL frameworks, specially optimized for the DGX-1, and engineered for maximized GPU-accelerated performance. The NVIDIA GPU Cloud Deep Learning Software Stack includes optimizations from the deep learning framework to the drivers, libraries and communications primitives required for execution and provide a solid foundation that enables data scientists to rapidly develop, deploy and scale applications on the DGX-1 (or multiple DGX-1s). This integrated software stack saves considerable expenditure of software engineering effort that would otherwise be borne by the user.

The DGX-1 provides a fully-integrated hardware and software solution that can be deployed effortlessly across a broad range of environments, providing instant AI and DL application enablement and acceleration, saving time that would otherwise be lost on system design, integration and troubleshooting effort.

DDN AI200 Parallel Storage Appliance

Figure 2. DDN AI200

The DDN AI200 is the scale-out flash solution for the AI datacenter supporting the fastest networking protocols available. The AI200 Active-Active controllers integrate a truly parallel filesystem onto NVMe Flash. The platform offers a simple scale-out model starting at just 2RU and able to expand to hundreds of systems, each adding linearly to a single, unified, shared namespace.

The AI200 parallel storage appliance is highly optimized for the AI Datacenter with several unique characteristics that resolve bottlenecks associated with the AI lifecycle. The AI200 parallel storage appliance performance characteristics include strong ingest performance, capability to deliver full data saturation of GPUs, excellent small data performance and a native accelerated wire protocol for making maximum use of the network. The filesystem has proven robustness at the very largest scales and offers a multi-tenanted security framework allowing the single namespace to be partitioned into private and shared areas.

The DDN A³I Reference Architecture

DDN A³I reference architectures integrate DGX-1 servers and AI200 parallel storage appliances with a set of well-defined topologies and network options for the fastest, most flexible and scalable AI platform. The DDN A³I reference architecture can service the whole of the AI and DL data lifecycle in-place, from initial data ingest through training, verification and inference. By providing shared data access through a unified, scalable namespace, the DDN A³I architecture eliminates the need to copy and manage local storage in the computing nodes or move data in and out of a separate data lake. This central data governance with A³I simplifies management, ensures integrity of collections, and provides robust data protection in case of hardware failure.

The DDN A³I reference architectures integrates high-performance, low-latency, Remote Direct Memory Access (RDMA) capable networks. Both EDR IB and 100 GbE (or 40 GbE) can be used. This choice gives both better whole-environment utilization and also gives the maximum benefit of Flash to AI and DL frameworks. This is particularly important when leveraging distributed processing across multiple multi-GPU servers. RDMA support ensures that data exchanged over the network is delivered directly to the GPUs efficiently, eliminating network transport overhead from the CPU, caches and removing context switch latencies. InfiniBand supports RDMA natively and RDMA over Converged Ethernet (RoCE) brings that capability to Ethernet. This efficient GPU-to-GPU and GPU-to-storage pathing ensures optimal saturation and utilization of GPUs (Figure 3).

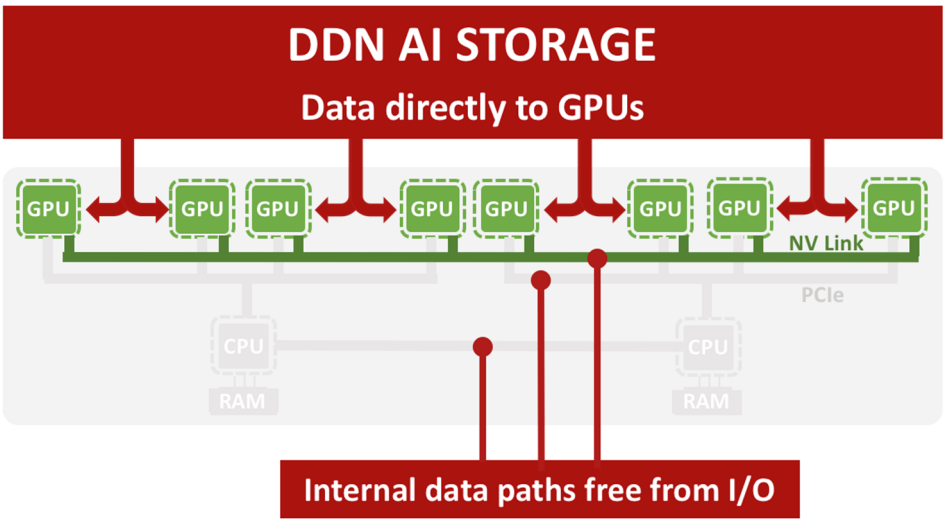

Figure 3. Optimized data delivery for DGX-1 server with DDN A³I

Figures 4 and 5 illustrate the DDN A^{3}I architecture in a 1:1 configuration in which a single DGX-1 server is connected to an AI200 parallel storage appliance through an EDR IB or 100 GbE network. Both the DGX-1 server and AI200 parallel storage appliance connect to a single network switch via four links.

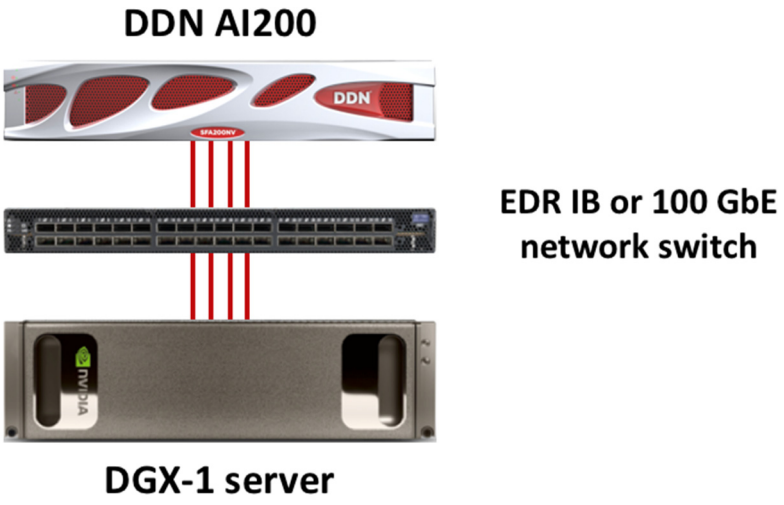

Figure 4. DDN A^{3}I reference architecture in a 1:1 configuration

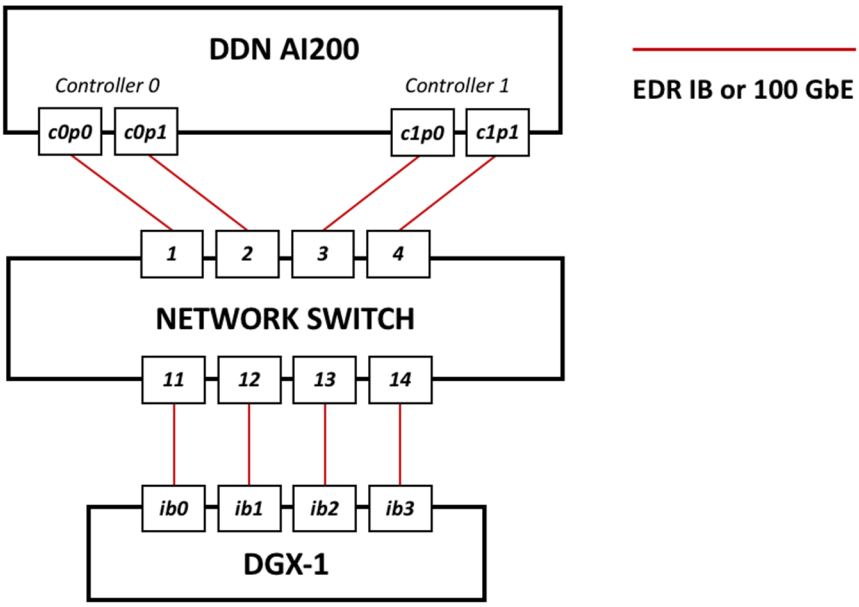

Figure 5. Network diagram of port-level connectivity in a 1:1 configuration

Figures 6 and 7 illustrate the DDN A^{3}I architecture in a 4:1 configuration in which four DGX-1 servers are connected to an AI200 parallel storage appliance through a pair of network switches that are configured for high-availability (HA). Every DGX-1 server connects to each of the network switches via two EDR IB or 100 GbE links. The AI200 parallel storage appliance connects to each of the network switches via two EDR IB or 100 GbE links. The network switches are interconnected with four dedicated links. This ensures non-blocking data exchanges between every device connected to the network. The HA design provides full-redundancy and maximum data availability in case of component failure in one of the devices.

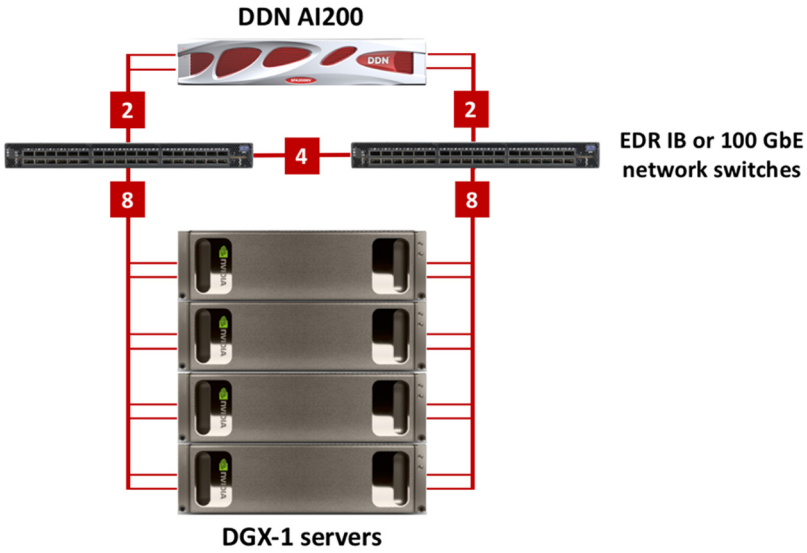

Figure 6. DDN A^{3}I reference architecture in a 4:1 configuration

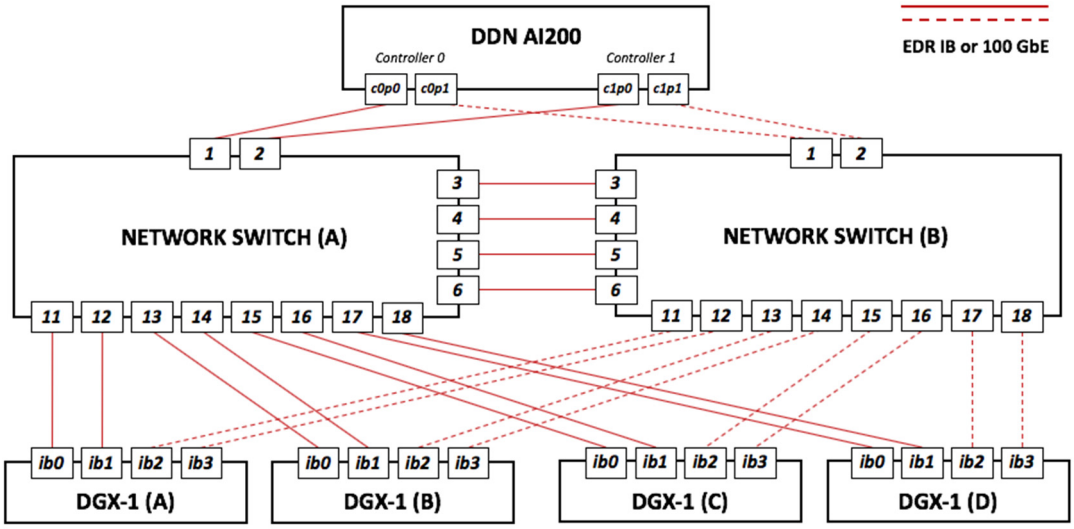

Figure 7. Network diagram of port-level connectivity in a 4:1 configuration

Figures 8 and 9 illustrate the DDN A^{3}I architecture in a 9:1 configuration in which nine DGX-1 servers are connected to an AI200 parallel storage appliances through a pair of network switches that are configured for high-availability (HA). Every DGX-1 server connects to each of the network switches via two EDR IB or 100 GbE links. The AI200 parallel storage appliance connects to each of the network switches via two EDR IB or 100 GbE links. The network switches are interconnected with eight dedicated links. The HA design provides full-redundancy and maximum data availability in case of component failure in one of the devices.

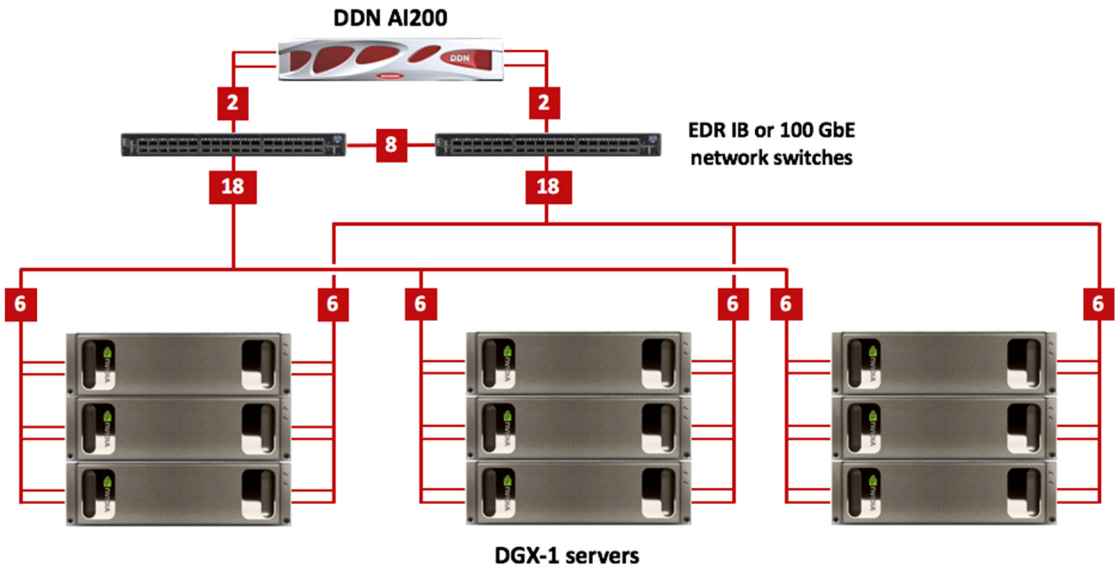

Figure 8. DDN A^{3}I reference architecture in a 9:1 configuration

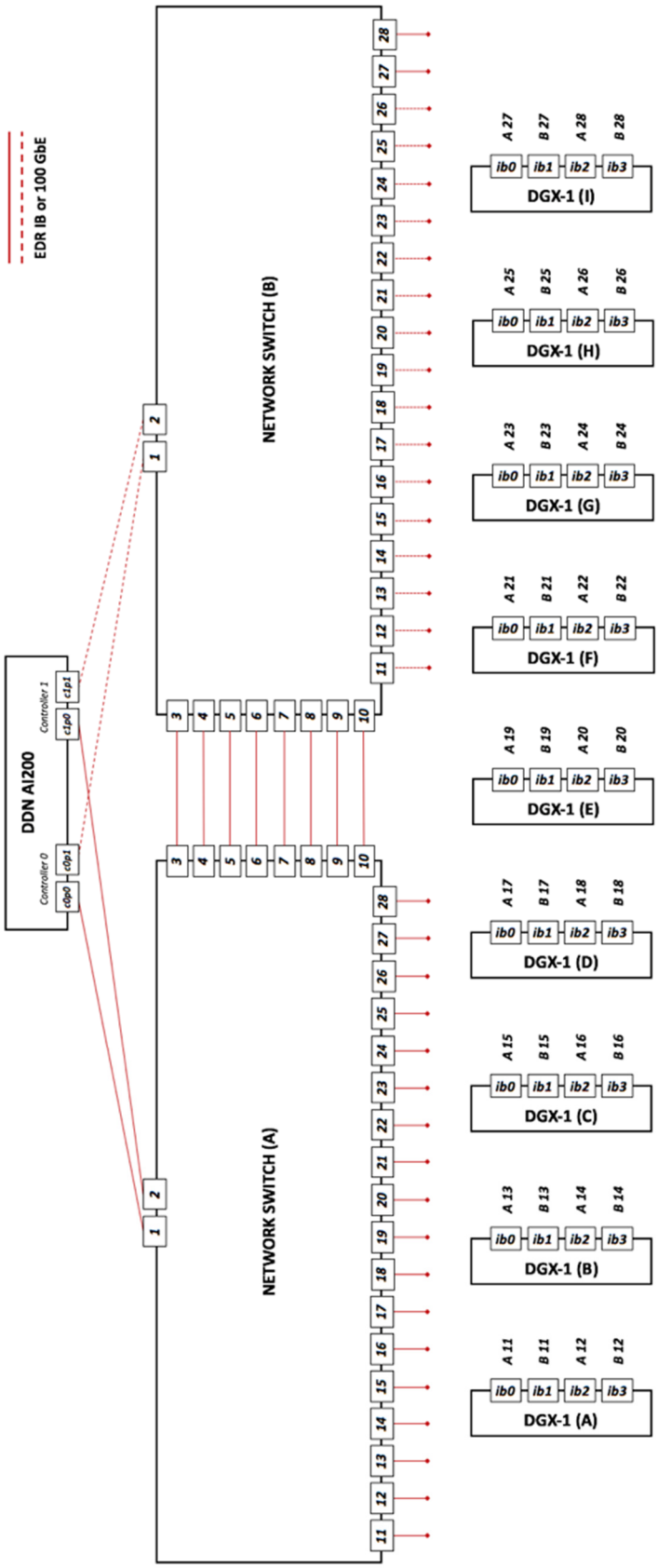

Figure 9. Network diagram for port-level connectivity in a 9:1 configuration

Scaling with GPUs and the DDN A³I Reference Architecture

The DDN A³I architecture is very flexible and can scale seamlessly in capacity, performance and capability. Deployments can start with a single DGX-1 server and a single AI200 parallel storage appliance. Additional DGX-1 servers and AI200 parallel storage appliances can be integrated rapidly to meet evolving workflow and workload requirements. At every deployment scale, the DDN A³I architecture continuously delivers an optimized, extremely cost-effective solution.

With the DDN A³I shared architecture, data can be accessed by multiple compute nodes or GPU appliances with maximum bandwidth simultaneously, enabling fastest data delivery possible in a distributed computing environment. The need to copy data to local cache on compute nodes prior to running the application is eliminated. This reduces the amount of overhead as well as workflow restrictions imposed by the limited capacity of the local cache. The DDN A³I architecture provides simple and effective scaling for AI and DL applications.

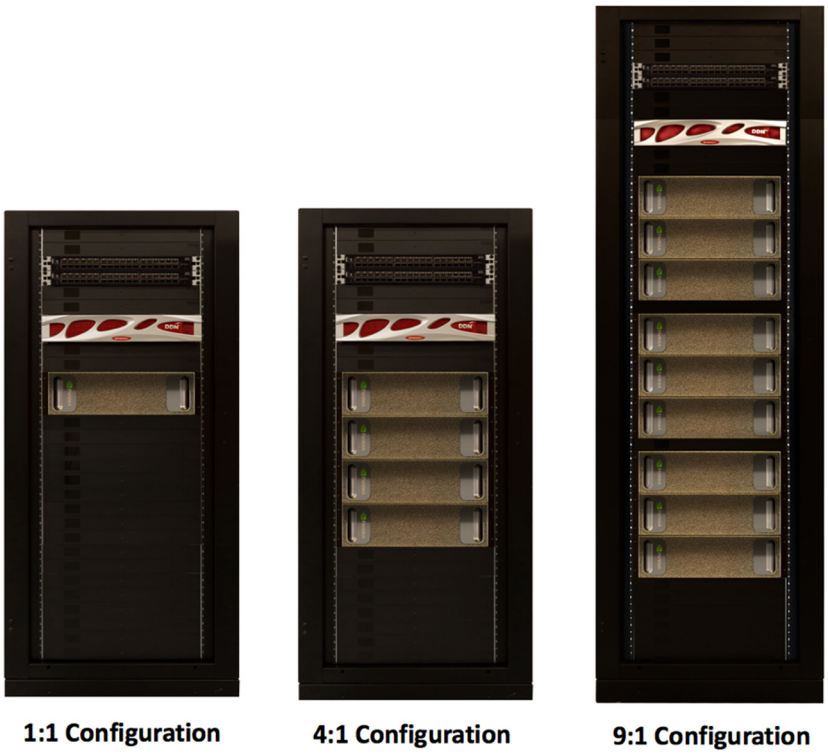

Figure 10. Rack elevations for scaling of DDN A³I 1:1, 4:1 and 9:1 configurations. The rack elevations in this section are for illustrative purposes only. Actual equipment placement will depend upon the type of rack and cooling resources available in the target data center.

DDN A³I Reference Architecture Performance Testing

Performance testing on the DDN A³I architecture has been conducted with synthetic throughput and IOPs testing applications, as well as widely-used DL frameworks. The results demonstrate that using the A³I intelligent client, containerized applications can engage the full capabilities of the data infrastructure, and that the DGX-1 server achieves full GPU saturation consistently for DL workloads.

These tests described below were executed on a DGX-1 server equipped with eight V100 GPUs, running DGX OS Server Software 3.1.6. The AI200 parallel storage appliance is running DDN EXAScaler v4.0.0-r1364. Both the DGX-1 server and the AI200 parallel storage appliance are connected to a Mellanox SB7700 network switch with four EDR IB links each. The switch is running Mellanox OS 3.6.5000.

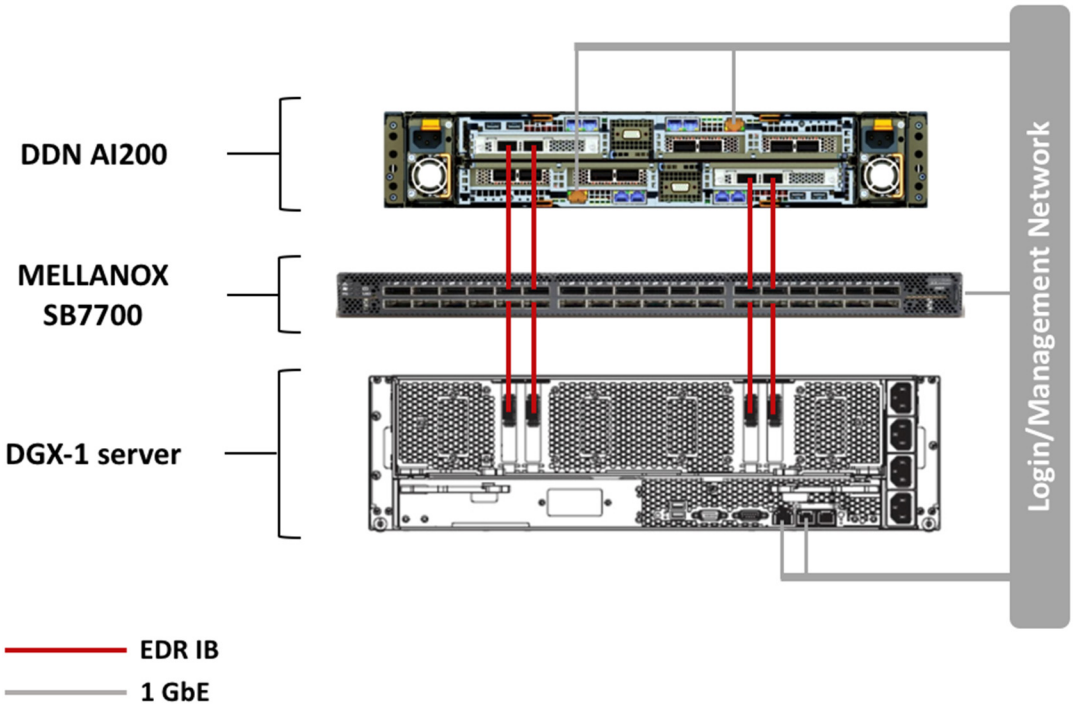

Figure 11. Network diagram of DDN A³I testing environment

Infrastructure Performance – Single Container

For single containers, DDN A^{3}I architecture delivers 10GB/s throughput, and random read performance of more than 100k IOPS. Data is consistently delivered extremely low latencies of under 1.5ms for small random IOs.

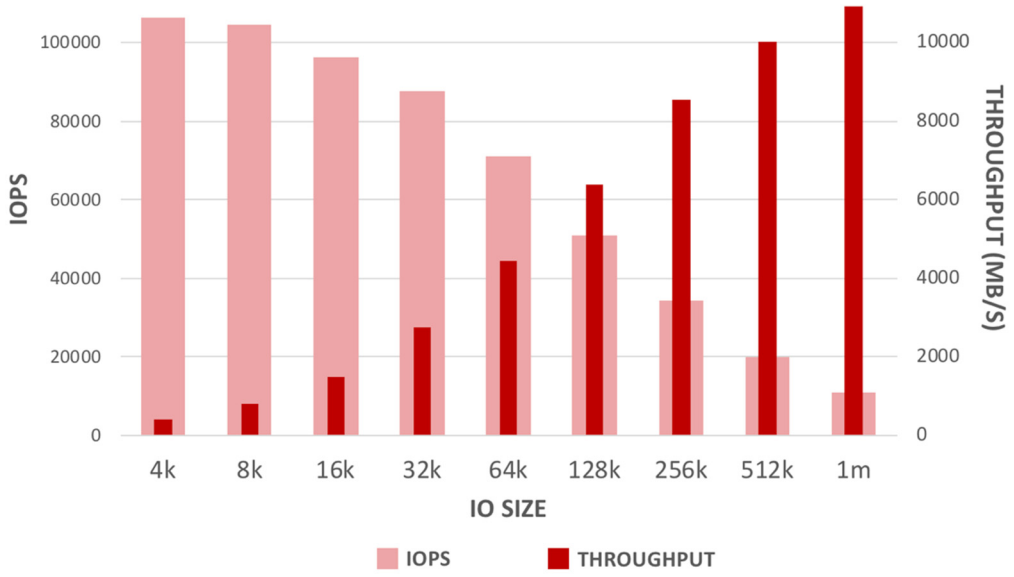

Figure 12. DGX-1 single container IOPS and throughput performance

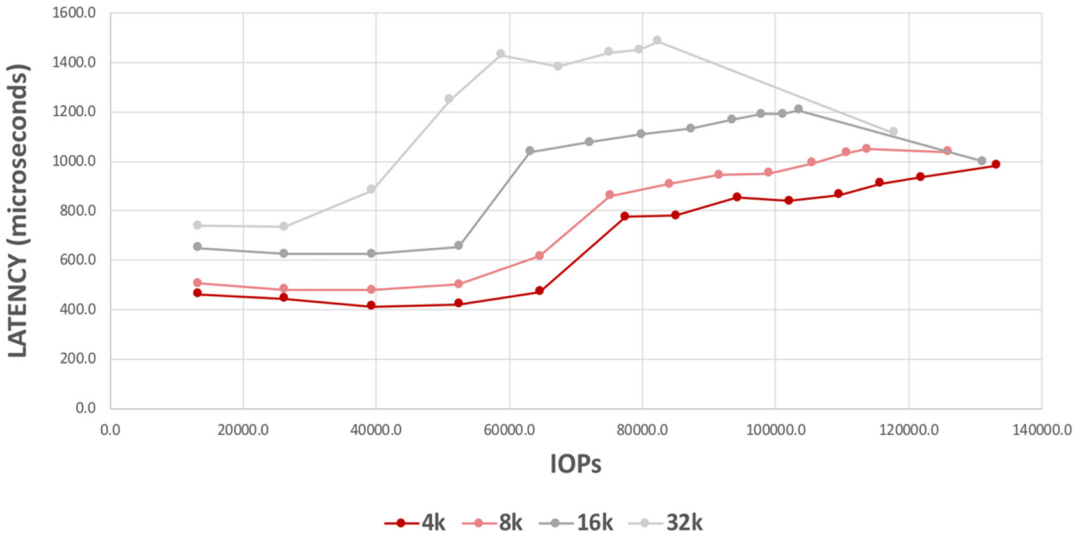

Figure 13. DGX-1 single container IO latency

Infrastructure Performance – Multiple Containers

For multiple containers, the DDN A^{3}I architecture demonstrates linear scale up to full saturation of the DGX-1 platform. The graphs below show concurrent IO activity from 2 and 4 containers simultaneously running on an 8 GPU DGX-1 server with aggregate delivered performance of 20GB/s.

Figure 14. DGX-1 container aggregate performance (2 containers)

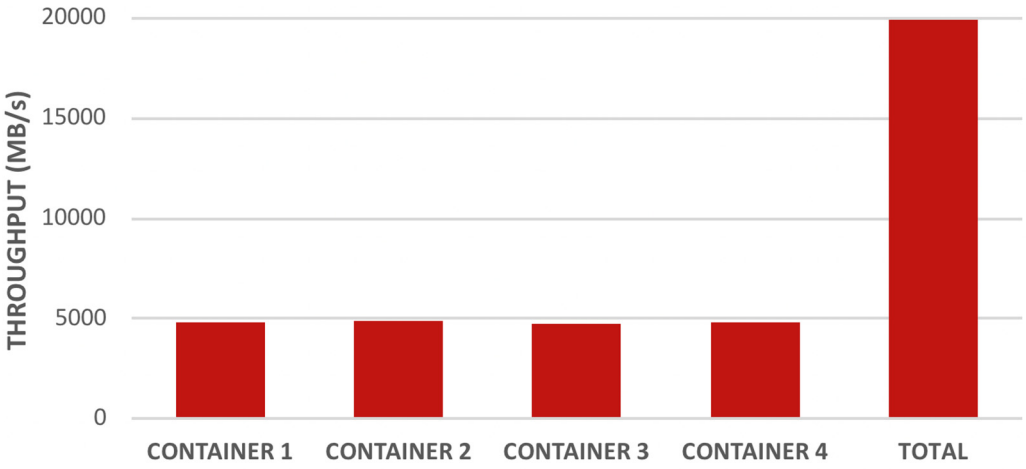

Figure 15. DGX-1 container aggregate performance (4 containers)

The infrastructure performance tests were performed using fio version 3.2, an industry standard synthetic IO benchmarking application. It was compiled for DGX OS version 3.1.6.

AI and DL Application Training Performance

The DDN A^{3}I architecture provides high-throughput and low-latency data delivery for AI and DL frameworks on the DGX-1 server. Extensive interoperability and performance testing has been completed using popular DL frameworks, notably TensorFlow, Horovod, Torch, PyTorch, NVIDIA® TensorRT™, Caffe, Caffe2, CNTK, MXNET, and Theano. This effort is being led by DDN in close collaboration with customers and NVIDIA. Full details and results of this effort are published in the DDN A^{3}I Solutions Guide.

The test below demonstrates training application performance with resnet-50, resnet-152 and inceptionV3 models using different numbers of GPUs on a single DGX-1 server. The resnet-152 and inceptionV3 tests were executed with the NVIDIA TensorFlow 18.03-py2[1] dockerfile and a data set from the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). The resnet-50 test was executed with the NVIDIA TensorFlow 18.09-py3[2] dockerfile and same data set. The CNN benchmark from the alsrgv[3] package was used for the test (commit 3b90c14 05/31/2018).

The application, data set and execution protocols used reflect a real-world use case. The system cache is cleared prior to every iteration of the test to ensure that the results are not biased by cached data.

The results show linear training application performance scaling. The results also demonstrate that the DGX-1 server achieves and maintains full saturation of all eight GPUs consistently for the duration of the training application.

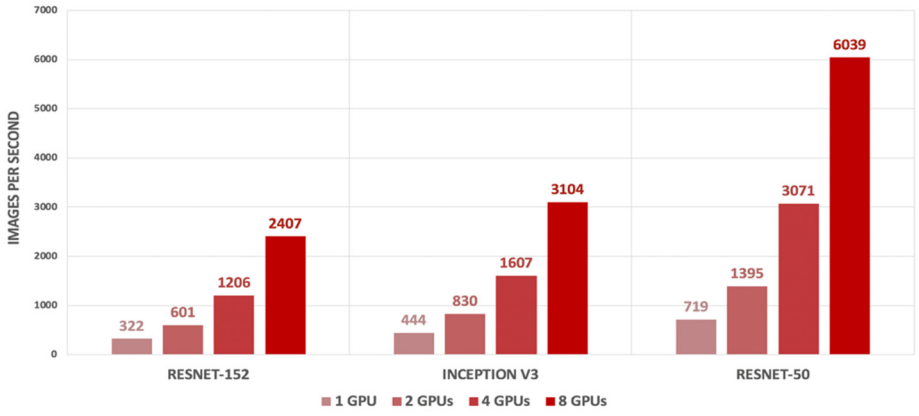

Figure 16. Training rate using TensorFlow with different models and GPU configurations

[1] Available from https://ngc.nvidia.com/registry/nvidia-tensorflow
[2] Ibidem
[3] Available from https://github.com/alsrgv/benchmarks

Figure 17 illustrates the GPU utilization and read activity from the AI200 parallel storage appliance. The GPUs achieve maximum utilization, and the AI200 parallel storage appliance delivers a steady stream of data to the application during the training process. The training application takes 933 seconds to complete. At approximately 660 seconds, the data set is fully loaded into the DGX-1 server memory and the application no longer needs to read the data from the AI200 parallel storage appliance.

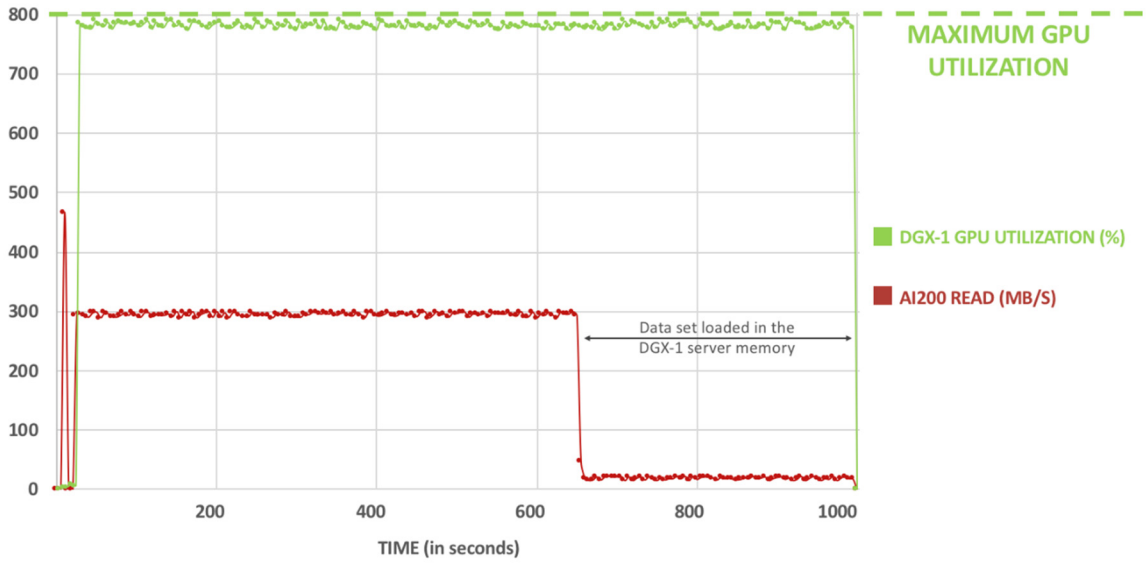

Figure 17. GPU utilization and AI200 parallel storage appliance read throughput during Inception V3 training

AI and DL Data Preparation Performance

TFRecord is a highly optimized TensorFlow file format that enables the conversion of discrete data and metadata asset collections into series of streamlined binary files. This process significantly reduces the amount of dataset preparation time required before executing the TensorFlow application. To be utilized, discrete assets must be split into training, testing, and validation sets that are stored in a specific folder structure and shuffled to avoid biased data distribution. This requires tedious data handling and attention to maintain proper shuffling. TFRecords provide a consolidated dataset that is easy to maintain and distribute and that eliminates the need for file manipulation.

TFRecords also streamline TensorFlow at runtime. Discrete assets must be opened individually, generating tremendous overhead for the data delivery and storage systems. A consolidated TFRecord binary file is more efficient as it only requires a single file open operation and allows the entire dataset to be held into a block of memory. This also enables applications to shuffle data at random places throughout the workflow and dynamically split training, testing and validation sets. This provides tremendous agility, efficiency and acceleration to TensorFlow applications.

The DDN A^{3}I parallel architecture furthers these benefits by allowing concurrent delivery of discrete data and metadata assets from source datasets to the conversion application, and rapid write of the binary file to persistent storage. In the example below, a dataset with 1.9 million data and metadata files spread across thousands of folders is being condensed to 1150 TFRecords binary files in a single directory.

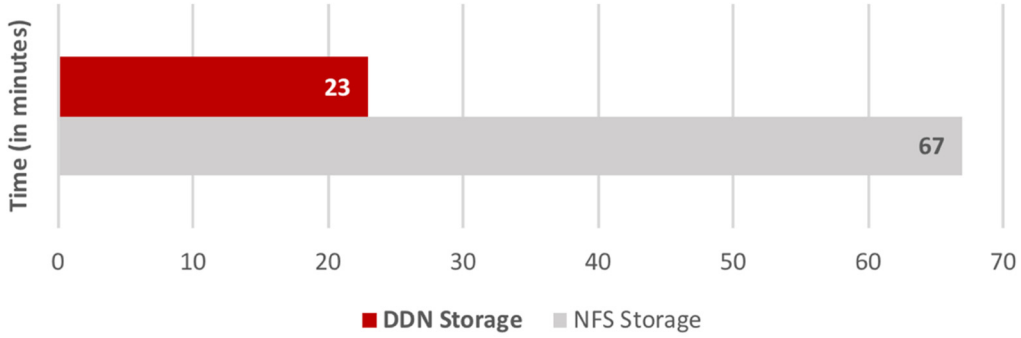

Figure 18. TFRecords conversion operation duration

This test was executed with the NVIDIA TensorFlow 18.04-py2 dockerfile using a data set from the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). The NFS storage setup used for the test is detailed at the bottom of page 20.

Parallel Architecture Performance for DGX-1 containers

NFS-based storage systems, commonly found in traditional IT infrastructures, are woefully inadequate in handling the demanding needs of AI and DL. Designed to handle more modest workloads, lower scalability, limited performance needs and small data volumes, these platforms are highly bottlenecked and lack the fundamental capabilities needed for AI-enabled deployments.

An NFS architecture associates client IP-addresses with the container and, using a hashing algorithm, NFS calls from these IP addresses are redirected to a particular storage blade in the cluster, restricting maximum performance achievable by each client or container to a fraction of the overall storage system capability. To try to circumvent this such systems create several container instances, assigning each one a unique IP-Address, and attempt to load balance them across server instances.

As the number of containers scales up the problem gets worse. For example, if 21 containers are created and seven storage blades are available, this would require the load balancing of three client virtual IP-addresses per storage blade creating significant inefficiency, complexity, wasted GPU cycles and degraded performance.

NFS has high overhead and a poor latency software layer which brokers traffic between network clients and storage. NFS is prone to severe traffic contention when engaged from multiple clients, lacks redundancy features, and will lose access to data with a single point failure. The DDN A^3I architecture on the other hand uses high speed parallelized data paths with direct multi-port connections to the filesystem, allowing for massive concurrency in transactions. Redundancy and automatic failover capability ensures continuous data availability, even in case of network or server connection unavailability.

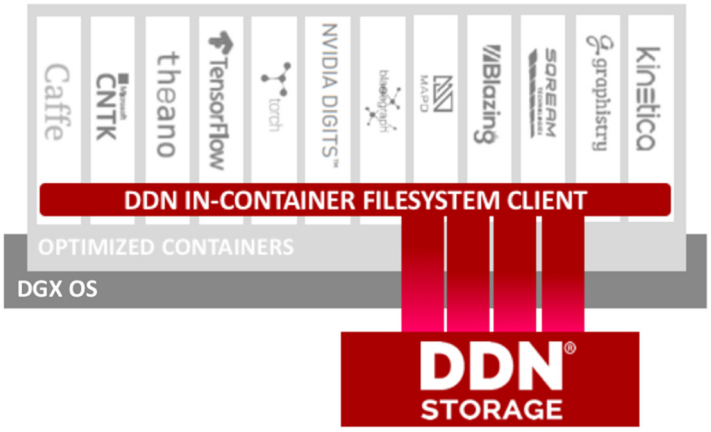

Figure 19. DDN A^{3}I in-container parallel filesystem client for widely used DL frameworks

As part of its integration with NVIDIA platforms, DDN has developed an intelligent client software for DGX-1 containers that enables file-level access to the shared storage system directly from containerized applications at runtime. This high-throughput, low-latency, parallel connection between DGX-1 server application containers and the storage system seamlessly provides fastest data access possible. Additionally, the limitations of sharing a single host-level connection to storage between multiple containers disappear. The in-container filesystem mounting capability is added at runtime through a universal wrapper that does not require any modification to the application or container.

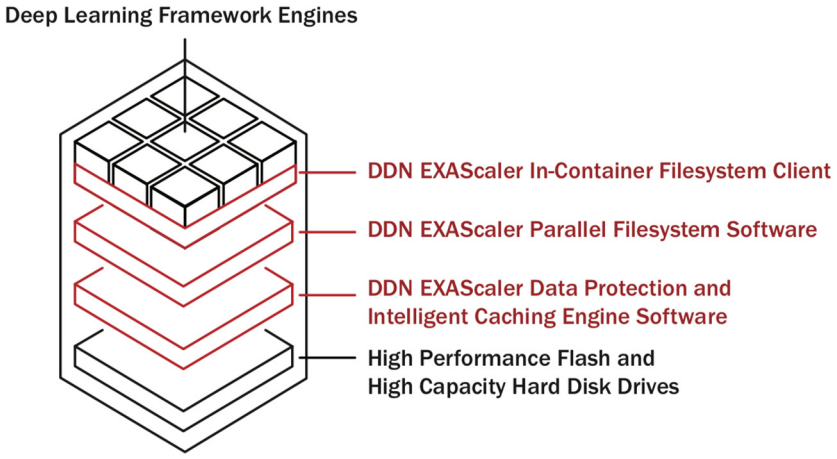

Figure 20. Multi-layer integration of the DDN A^{3}I architecture

At the application level, data is accessed through a standard highly interoperable POSIX file interface, for a familiar and intuitive user experience. In addition, security features and controls allow for compartmentalized access to data at the system or container level, enabling multi-tenancy options for customers that require trusted levels of segregation.

The container-level integration of the DDN A^{3}I parallel architecture provides significant training rate performance and faster completion times for a majority of commonly used DL frameworks. In this example Caffe GoogleNet training rate is increased 2.4X and training time is twice as fast with the DDN parallel architecture.

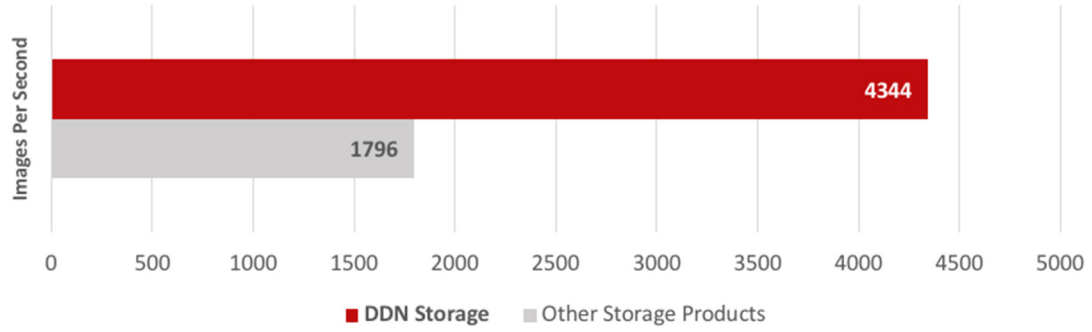

Figure 21. Comparing Caffe GoogleNet training rate between NFS and DDN Storage

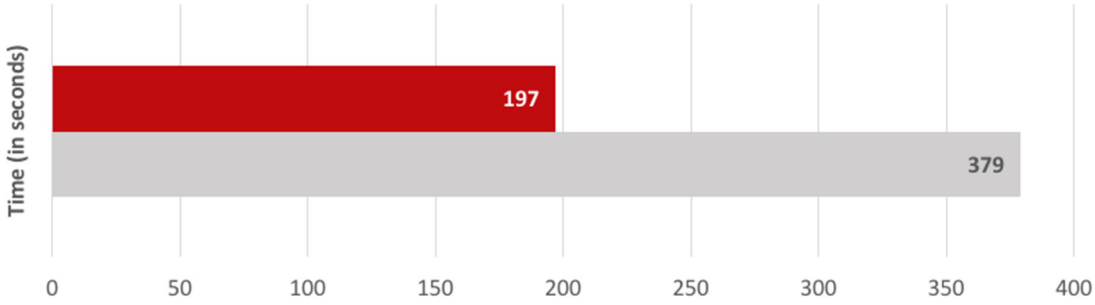

Figure 22. Comparing Caffe GoogleNet training duration between NFS and DDN Storage

This test was executed with NVIDIA Caffe 18.03-py2 dockerfile using a data set from the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). For NFS comparison tests, the DGX-1 server is connected to a Mellanox MSN2700-CS2F switch with four 100 GbE links. An NFS server is connected to switch with four 100 GbE links.

Conclusion

The DDN A^{3}I architecture is fully-integrated and optimized for the DGX-1 server. It provides a high-performance parallel architecture that delivers data to applications with high bandwidth and low latency. This ensures full GPU resource utilization even with distributed applications running on multiple DGX-1 servers.

The DDN A^{3}I shared architecture provides concurrent and efficient service for the entire spectrum of activities involved in the AI and DL process, including data ingest, data manipulation, training and inference. The DDN A^{3}I architecture is very flexible and can scale seamlessly in capacity, performance and capacity to match ever evolving workflow requirements.

DDN has long been a partner of choice for organizations pursuing data-intensive projects at any scale. DDN has successfully deployed data-at-scale systems across all areas of AI and DL, from autonomous vehicles to data security and fraud detection, augmented reality and healthcare, personalized marketing and nature language processing. In addition, DDN solutions are continuously tested and optimized with AI applications, network topologies and latest GPU technology, thereby ensuring the performance of data fulfillment from storage-to-GPU and GPU-to-GPU.

Developed in close collaboration with NVIDIA, the DDN A^{3}I architecture enables organizations everywhere to generate value and accelerate time to insight from their data using AI and DL with maximum velocity and efficiency.

ABOUT DDN®

DataDirect Networks (DDN) is the world's leading big data storage supplier to data-intensive, global organizations. DDN has designed, developed, deployed, and optimized systems, software, and solutions that enable enterprises, service providers, research facilities, and government agencies to generate more value and to accelerate time to insight from their data and information, on premise and in the cloud. For more information, visit our website www.ddn.com or call 1-800-837-2298.

SALES@DDN.COM
+1.800.837.2298

V4 09/18