NetApp Verified Architecture

NetApp ONTAP AI with NVIDIA DGX A100 Systems and Mellanox Spectrum Ethernet Switches

NVA Design

David Arnette and Sung-Han Lin, NetApp
December 2020 | NVA-1153-DESIGN

Abstract

This document describes a NetApp Verified Architecture for machine learning (ML) and artificial intelligence (AI) workloads using NetApp® AFF A800 storage systems, NVIDIA DGX™ A100 systems, and NVIDIA® Mellanox® Spectrum™ SN3700V 200Gb Ethernet switches. This design features RDMA over Converged Ethernet (RoCE) for the compute cluster interconnect fabric to provide customers with a completely ethernet-based architecture for high-performance workloads. This document also includes benchmark test results for the architecture as implemented.

In partnership with

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Executive summary

This document contains validation information for the NetApp ONTAP® AI reference architecture for machine learning (ML) and artificial intelligence (AI) workloads. This design was implemented using a NetApp AFF A800 all-flash storage system, eight DGX A100 systems, and SN3700V switches for both the compute cluster interconnect and storage connectivity. The operation and performance of this system was validated using industry-standard benchmark tools and has proven to deliver excellent training performance. Customers can easily and independently scale compute and storage resources from half-rack to multi-rack configurations with predictable performance to meet any machine learning workload requirement.

Program summary

The NetApp Verified Architecture program provides customers with reference configurations and sizing guidance for specific workloads and use cases. These solutions are:

- Thoroughly tested
- Designed to minimize deployment risks
- Designed to accelerate time to market

This document is for NetApp and partner solutions engineers and customer strategic decision makers. It describes the architecture design considerations that were used to determine the specific equipment, cabling, and configurations required to support the validated workload.

NetApp ONTAP AI solution

The NetApp ONTAP AI reference architecture, powered by DGX A100 systems and NetApp cloud-connected storage systems, was developed and verified by NetApp and NVIDIA. It gives IT organizations an architecture that:

- Eliminates design complexities
- Allows independent scaling of compute and storage
- Enables customers to start small and scale seamlessly
- Offers a range of storage options for various performance and cost points

NetApp ONTAP AI tightly integrates DGX A100 systems and NetApp AFF A800 storage systems with state-of-the-art networking. NetApp ONTAP AI simplifies artificial intelligence deployments by eliminating design complexity and guesswork. Customers can start small and grow nondisruptively while intelligently managing data from the edge to the core to the cloud and back.

Figure 1 shows several variations in the ONTAP AI family of solutions with DGX A100 systems. The AFF A800 system performance has been verified with up to eight DGX A100 systems. By adding storage controller pairs to the ONTAP cluster, the architecture can scale to multiple racks to support many DGX A100 systems and petabytes of storage capacity with linear performance. This approach offers the flexibility to alter compute-to-storage ratios independently based on the size of the data lake, the deep learning (DL) models that are used, and the required performance metrics.

Figure 1) NetApp ONTAP AI family with NVIDIA DGX A100 systems.

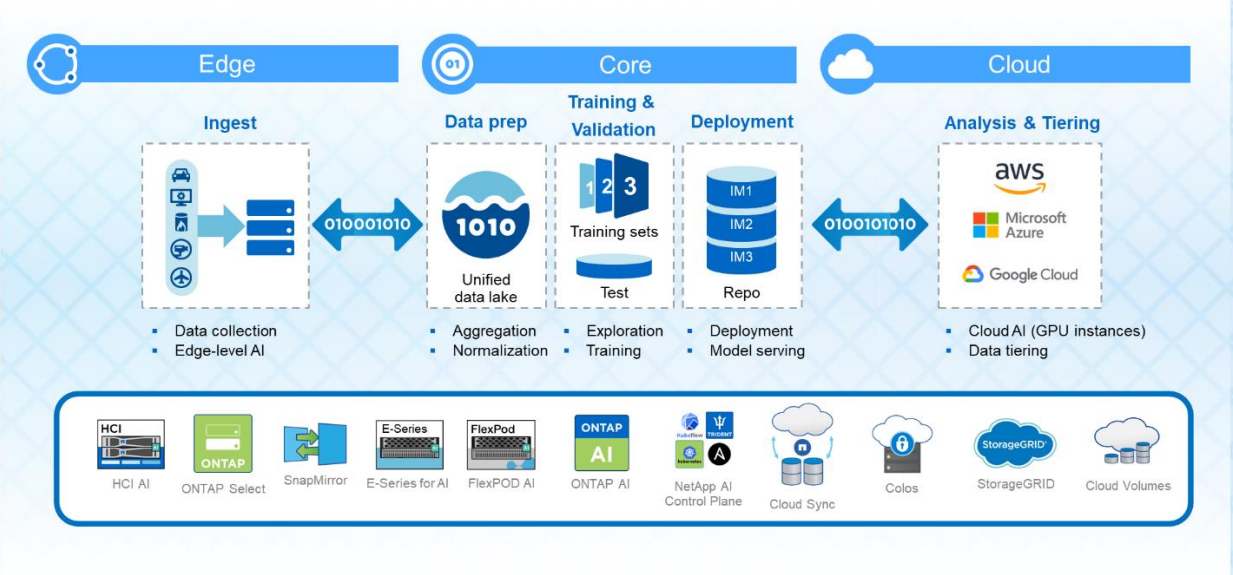
The number of DGX A100 systems and AFF systems per rack depends on the power and cooling specifications of the rack in use. Final placement of the systems is subject to computational fluid dynamics analysis, airflow management, and data center design.

Deep learning data pipeline

Deep learning is the engine that enables businesses to detect fraud, improve customer relationships, optimize supply chains, and deliver innovative products and services in an increasingly competitive marketplace. The performance and accuracy of DL models are significantly improved by increasing the size and complexity of the neural network as well as the amount and quality of data that is used to train the models.

Given the massive datasets required, it is crucial to architect an infrastructure that offers the flexibility to deploy across environments. At a high level, an end-to-end DL deployment consists of three phases through which the data travels: the edge (data ingest and inferencing), the core (training clusters and a data lake), and the cloud (archive, tiering, and dev/test). This is typical of applications such as the Internet of Things (IoT) for which data spans all three realms of the data pipeline. Figure 2 presents an overview of the components in each of the three realms.

Figure 2) Components of the edge-core-cloud data pipeline.



The following list describes some of the activities that occur in one or more of these areas.

- **Ingest.** Data ingestion usually occurs at the edge, for example, by capturing data streaming from autonomous cars or point-of-sale devices. Depending on the use case, an IT infrastructure might be needed at or near the ingestion point. For example, a retailer might need a small footprint in each store that consolidates data from multiple devices.
- **Data prep.** Preprocessing is necessary to normalize and cleanse the data before training. Preprocessing takes place in a data lake, possibly in the cloud, in the form of an Amazon S3 tier or in on-premises storage systems such as a file store or an object store.
- **Training and validation.** For the critical training phase of DL, data is typically copied from the data lake into the training cluster at regular intervals. The servers that are used in this phase use GPUs to parallelize computations, creating a tremendous appetite for data. Meeting the raw I/O bandwidth needs is crucial for maintaining high GPU utilization.
- **Deployment.** The trained models are tested and deployed into production. Alternatively, they could be fed back to the data lake for further adjustments of input weights, or, in IoT applications, the models could be deployed to smart edge devices.
- **Analysis and tiering.** New cloud-based tools become available at a rapid pace, so additional analysis or development work might be conducted in the cloud. Cold data from past iterations might be saved indefinitely. Many AI teams prefer to archive cold data to object storage in either a private or a public cloud. Based on compute requirements, some applications work well with object storage as the primary data tier.

Depending on the application, DL models work with large amounts of structured and unstructured data. This difference imposes a varied set of requirements on the underlying storage system, both in terms of size of the data that is being stored and the number of files in the dataset.

High-level storage requirements include:

- The ability to store and retrieve millions of files concurrently
- Storage and retrieval of diverse data objects such as images, audio, video, and time-series data
- Delivery of highly parallel performance at low latencies to meet the GPU processing speeds
- Seamless data management and data services that span the edge, the core, and the cloud

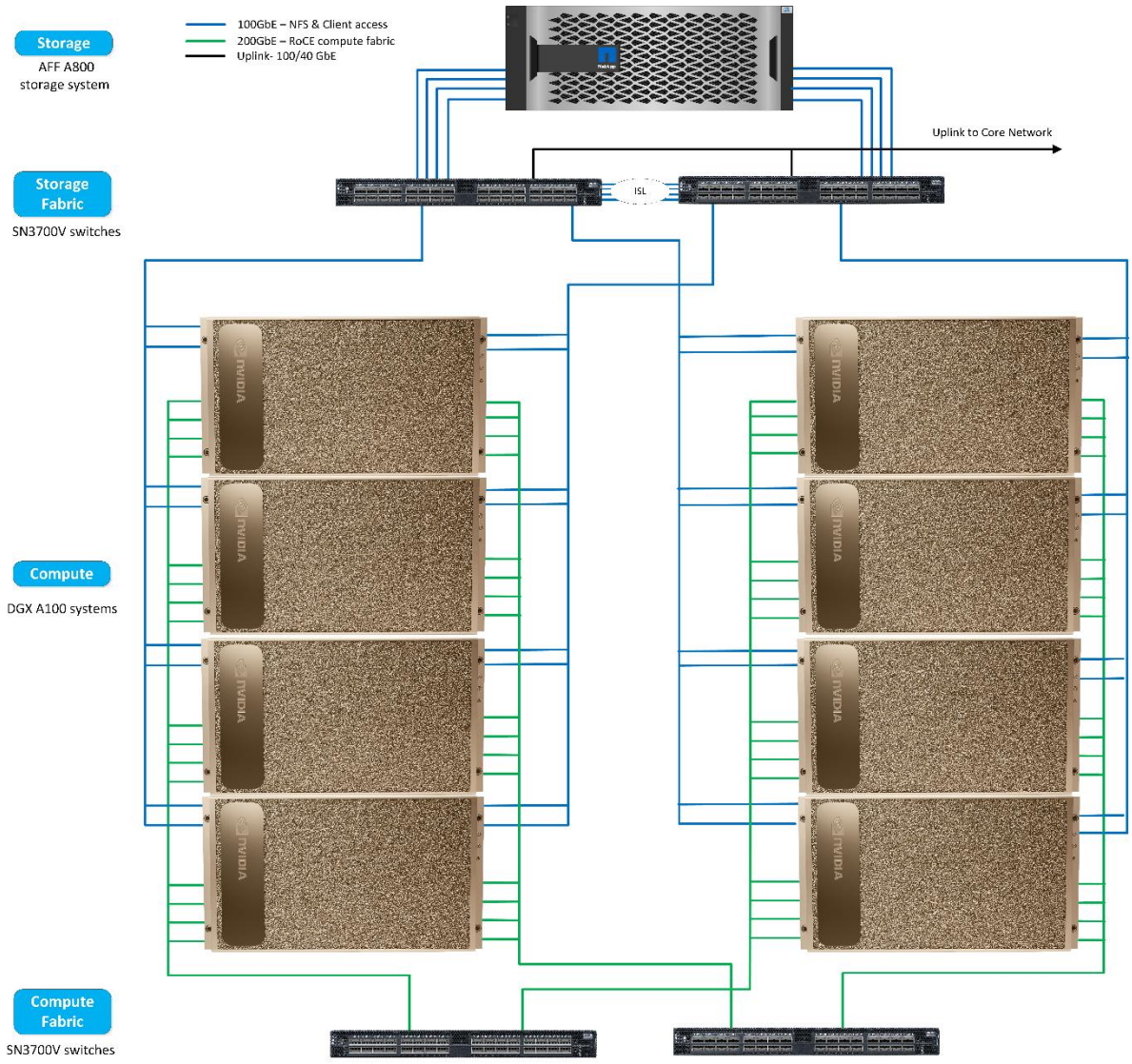
Combined with superior cloud integration and the software-defined capabilities of NetApp ONTAP, NetApp AFF systems support a full range of data pipelines that spans the edge, the core, and the cloud for DL. This document focuses on solutions for the training and inference components of the data pipeline.

Solution overview

DL systems leverage algorithms that are computationally intensive and that are uniquely suited to the architecture of GPUs. Computations that are performed in DL algorithms involve an immense volume of matrix multiplications running in parallel. Advances in individual and clustered GPU computing architectures leveraging DGX systems have made them the preferred platform for workloads such as high-performance computing (HPC), DL, video processing, and analytics. Maximizing performance in these environments requires a supporting infrastructure, including storage and networking, that can keep GPUs fed with data. Dataset access must therefore be provided at ultra-low latencies with high bandwidth.

This reference architecture was validated with one NetApp AFF A800 system, eight DGX A100 systems, two NVIDIA Mellanox Spectrum SN3700V 200Gb Ethernet switches for the compute fabric and two more SN3700V switches for storage and client access. Figure 3 shows the basic solution architecture.

Figure 3) NetApp ONTAP AI verified architecture.



NVIDIA DGX A100 systems

The DGX A100 system is a fully integrated, turnkey hardware and software system that is purpose-built for DL workflows. Each DGX A100 system is powered by eight NVIDIA A100 GPUs that are configured in a hybrid cube-mesh topology that uses NVIDIA NVLink® and NVIDIA NVSwitch® technologies. This configuration provides an ultra-high bandwidth, low-latency fabric for inter-GPU communication within the DGX A100 system. This topology is essential for multi-GPU training, eliminating the bottleneck that is associated with PCIe-based interconnects that cannot deliver linearity of performance as the GPU count increases. The DGX A100 system is also equipped with high-bandwidth, low-latency network interconnects for multinode clustering over RoCE and InfiniBand.

NVIDIA NGC

The DGX A100 system leverages NVIDIA NGC, a cloud-based container registry for GPU-accelerated software. NGC provides containers for today's most popular DL frameworks such as Caffe2, TensorFlow, PyTorch, MXNet, and TensorRT, which are optimized for NVIDIA GPUs. The containers integrate the framework or application, necessary drivers, libraries, and communications primitives, and they are

optimized across the stack by NVIDIA for maximum GPU-accelerated performance. NGC containers incorporate the NVIDIA CUDA Toolkit, which provides the CUDA Basic Linear Algebra Subroutines Library (cuBLAS), the CUDA Deep Neural Network Library (cuDNN), and much more. The NGC containers also include the NVIDIA Collective Communications Library (NCCL) for multi-GPU and multinode collective communication primitives, enabling topology-awareness for DL training. NCCL enables communication between GPUs inside a single DGX A100 system and across multiple DGX A100 systems.

NetApp AFF systems

NetApp AFF state-of-the-art storage systems enable IT departments to meet enterprise storage requirements with industry-leading performance, superior flexibility, cloud integration, and best-in-class data management. Designed specifically for flash, AFF systems help accelerate, manage, and protect business-critical data.

The NetApp AFF A800 system is the industry's first end-to-end NVMe solution. For NAS workloads, a single AFF A800 system supports throughput of 25GBps for sequential reads and one million IOPS for small random reads at sub-500µs latencies.

AFF A800 systems support the following features:

- Massive throughput of up to 300GBps and 11.4 million IOPS in a 24-node cluster
- 100GbE and 32Gb FC connectivity
- Up to 30TB solid-state drives (SSDs) with multistream write
- High density with 2PB in a 2U drive shelf
- Scaling from 200TB (2 controllers) to 9.6PB (24 controllers)
- NetApp ONTAP 9.7 includes a complete suite of data protection and replication features for industry-leading data management

Other NetApp storage systems, such as the AFF A700, AFF A400, and AFF A220, offer lower performance and capacity options for smaller deployments at lower cost points.

NetApp ONTAP 9

NetApp ONTAP 9, the latest generation of storage management software from NetApp, enables businesses to modernize infrastructure and transition to a cloud-ready data center. Leveraging industry-leading data management capabilities, ONTAP enables the management and protection of data with a single set of tools, regardless of where that data resides. Data can also be moved freely to wherever it's needed—the edge, the core, or the cloud. ONTAP 9 includes numerous features that simplify data management, accelerate and protect critical data, and enable next-generation infrastructure capabilities across hybrid cloud architectures.

Simplify data management

Data management is crucial to enterprise IT operations so that appropriate resources are used for applications and for datasets. ONTAP includes the following features to streamline and simplify operations and reduce the total cost of operation:

- **Inline data compaction, compression, and deduplication.** Compression delivers the primary benefit for alpha-numeric data often used in ML/DL workloads. Data compaction reduces wasted space inside storage blocks, and deduplication significantly increases effective capacity.
- **Minimum, maximum, and adaptive quality of service (QoS).** Granular QoS controls help maintain performance levels for critical applications in highly shared environments and allows production and development to share infrastructure with guaranteed allocation of resources.

- **ONTAP FabricPool.** This feature provides automatic tiering of cold data to public and private cloud storage options, including Amazon Web Services (AWS), Microsoft Azure, and NetApp StorageGRID®. For more details on FabricPool, please see TR-4598.

Accelerate and protect data

ONTAP delivers superior levels of storage performance and data protection and extends these capabilities with:

- **Performance and lower latency.** ONTAP offers the highest possible throughput at the lowest possible latency.
- **Data protection.** ONTAP provides built-in data protection capabilities with common management across all platforms.
- **NetApp Volume Encryption.** ONTAP offers native volume-level encryption with both onboard and external key management support.
- **Multi-tenancy and multi-factor authentication.** ONTAP enables sharing of infrastructure resources with the highest levels of security.

Future-proof infrastructure

ONTAP 9 helps meet demanding and constantly changing business needs.

- **Seamless scaling and nondisruptive operations.** ONTAP supports the nondisruptive addition of capacity to existing controllers as well as to scale-out clusters. Customers can upgrade to the latest technologies such as NVMe and 32Gb FC without costly data migrations or outages.
- **Cloud connection.** ONTAP is the most cloud-connected storage management software, with options for software-defined storage (ONTAP Select) and cloud-native instances (NetApp Cloud Volumes Service) in all public clouds.
- **Integration with emerging applications.** ONTAP offers enterprise-grade data services for next-generation platforms and applications using the same infrastructure that supports existing enterprise apps.

NetApp FlexGroup volumes

The training dataset is usually a large collection of many and potentially billions of files. Files can include text, audio, video, and other forms of unstructured data that must be stored and processed to be read in parallel. The storage system must store many small files and must read those files in parallel for sequential and random I/O.

A FlexGroup volume is a single namespace that is made up of multiple constituent member volumes and that is managed and acts like a NetApp FlexVol® volume to storage administrators. Files in a FlexGroup volume are allocated to individual member volumes and are not striped across volumes or nodes. They enable the following capabilities:

- FlexGroup volumes enable up to 20PB of capacity and predictable low latency for high-metadata workloads.
- They support up to 400 billion files in the same namespace.
- They support parallelized operations in NAS workloads across CPUs, nodes, aggregates, and constituent FlexVol volumes.

NetApp Trident

Trident from NetApp is an open-source dynamic storage provisioner for Docker and Kubernetes. Combined with NGC and popular orchestrators such as Kubernetes and Docker Swarm, Trident enables customers to seamlessly deploy DL NGC container images onto NetApp storage, which provides an enterprise-grade experience for AI container deployments. These deployments include automated

orchestration, cloning for testing and development, upgraded testing that uses cloning, protection and compliance copies, and many more data management use cases for the NGC AI and DL container images.

NVIDIA Mellanox networking

NVIDIA Mellanox Spectrum switches—The right choice for deep learning workloads

Networking is a critical part of the DL infrastructure that is responsible for moving massive amounts of data between end points efficiently and effectively. Spectrum Ethernet switches with consistent performance, intelligent load balancing, and comprehensive telemetry are an ideal network element for DL workloads.

Consistent performance

Spectrum Ethernet switches provide a high bandwidth and consistently low latency data path for GPU-GPU and GPU-storage communications. Spectrum, along with NVIDIA Mellanox ConnectX® adapters inside the DGX A100 systems, implement a tight and efficient ECN (Explicit Congestion Notification) mechanism that mitigates transient congestion and smooths traffic bursts to maximize network goodput.

Intelligent load balancing

The network is a shared resource, and its bandwidth must be shared in a fair manner across different flows and endpoints. Packet buffer architecture is one of the foundational attributes of the switch that affects performance and traffic fairness. The Spectrum switches feature a flexible and fully shared buffer architecture that ensures fair and balanced performance across all ports even when using a mix of different port speeds. Many high-speed switches in the market use fragmented packet buffers. Switches with fragmented buffers have scheduling issues and can preferentially give more bandwidth to certain ports/flows at the cost of others. This traffic imbalance leads to more performance variation that can hamper distributed DL performance.

Comprehensive telemetry

To reap high return on investment from the DL infrastructure, uptimes must be improved, and the network must be proactively monitored. Traditional methods of centrally processing the telemetry data acquired via SNMP or streaming can quickly become prohibitively expensive at terabit speeds. NVIDIA Mellanox What Just Happened® (WJH) leverages silicon level capabilities to quickly identify and export granular information about issues as soon as they happen. Because this capability is built into the platform, only the data pertinent to the issue is gathered at the central data collector. WJH makes proactive monitoring scalable and practical at terabit speeds. With WJH, customers can dramatically reduce mean time to issue resolution and plan capacity better.

Technology requirements

This section covers the hardware and software that was used for the testing described in the Solution verification section.

Hardware requirements

Table 1 lists the hardware components that were used to verify this solution.

Table 1) Hardware requirements.

Hardware	Quantity
DGX A100 systems	8
AFF A800 storage system	1 high-availability (HA) pair, includes 48x 1.92TB NVMe SSDs
SN3700V ethernet switches	2 for compute cluster interconnect
	2 for storage, client access and out-of-band management

Software requirements

Table 2 lists the software components that were used to validate the solution.

Table 2) Software requirements.

Software	Version
ONTAP storage OS	9.7P6
Network switch OS (all)	Cumulus Linux 4.2.1
DGX OS	4.99.10
Docker container platform	19.03.8
Container version	nvcr.io/nvidia/mxnet:20.06-py3 – MLPerf test tensorflow:20.05-tf2-py3 – other tests
OFED version	5.0-2.1.8
NCCL test version	https://github.com/NVIDIA/nccl-tests/tree/ec1b5e22e618d342698fda659efdd5918da6bd9f
FIO version	3.1

Solution architecture

This reference architecture has been verified to meet the requirements for running deep learning workloads. This enables data scientists to deploy DL frameworks and applications on a prevalidated infrastructure, helping to eliminate risks and allow businesses to focus on gaining valuable insights from their data. This architecture can also deliver exceptional storage performance for other HPC workloads without any modification or tuning of the infrastructure.

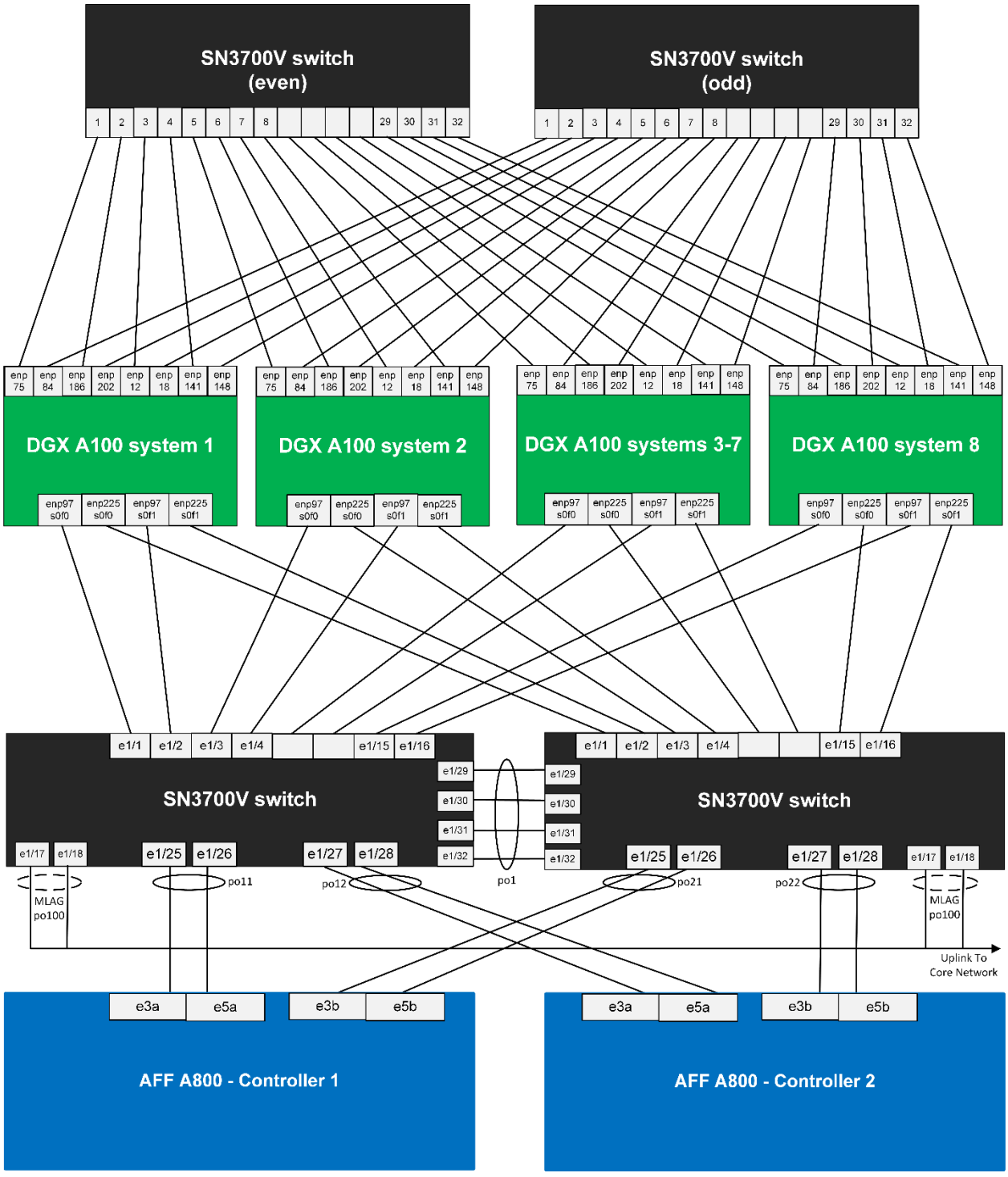
Network topology and switch configuration

This reference architecture leverages separate fabrics for compute-cluster interconnect and storage access. The compute-cluster network uses a pair of SN3700V Ethernet switches operating as independent redundant fabrics. Each DGX A100 system is connected to the switches using eight single-ported ConnectX-6 cards at 200Gbps, with even-numbered ports connected to one switch and odd-numbered ports connected to the other switch. The compute fabric switches are configured for RoCE to enable the lowest possible latency for GPU-to-GPU communications.

Two additional SN3700V switches are used to provide NFS storage connectivity as well as in-band management and client access to the DGX A100 systems. These SN3700V switches are configured for Multi-chassis Link Aggregation (MLAG) to allow aggregation of bandwidth and transparent failover in the event of a switch failure. Two dual-ported ConnectX-6 cards configured for Ethernet are used to provide two ports from each DGX A100 system to each SN3700V switch. One port from each card is configured into a bond dedicated to storage access, and the other port on each card is configured into a bond for in-band management and client access. Each AFF A800 storage system is connected using four 100GbE

ports from each controller, with a two-port LACP bond to each switch to provide balanced workload distribution across the storage controllers. Figure 4 shows the overall network topology.

Figure 4) Network switch port configuration.

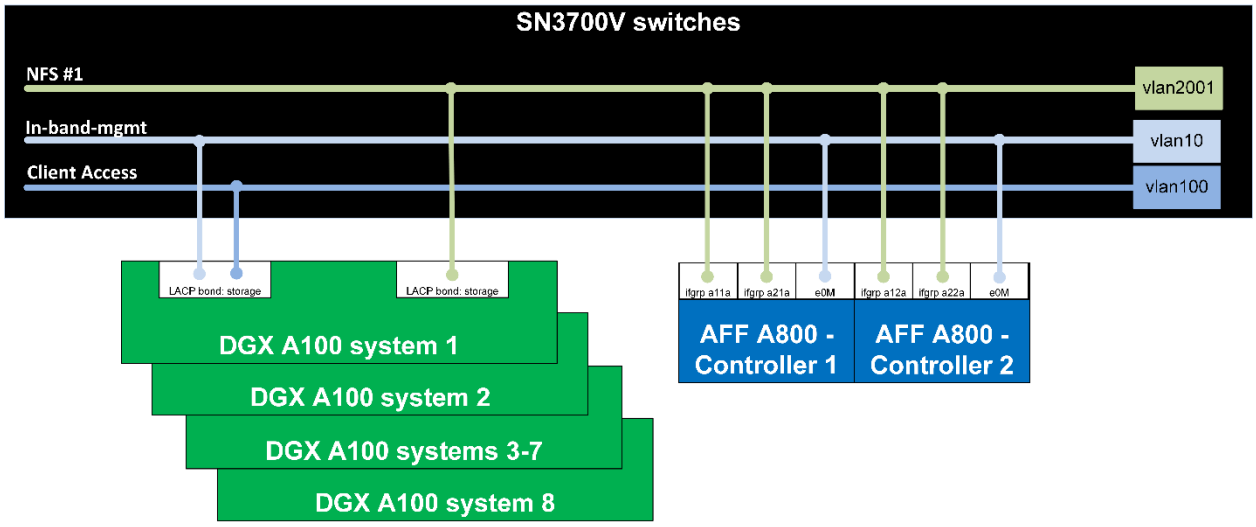


The Ethernet network is configured with multiple VLANs to isolate specific traffic types. NFS storage traffic, in-band management, and client access each have dedicated VLANs to provide the proper

maximum transmission unit (MTU) and other settings for each traffic type. For example, NFS storage traffic requires an MTU of 9000, while other typical Ethernet traffic uses an MTU of 1500.

Figure 5 shows the VLAN connectivity for the hosts and storage system controllers. Note that the AFF A800 storage system controllers have separate 1GbE management interfaces that are plugged into a separate management switch.

Figure 5) VLAN connectivity for DGX A100 and storage system ports.

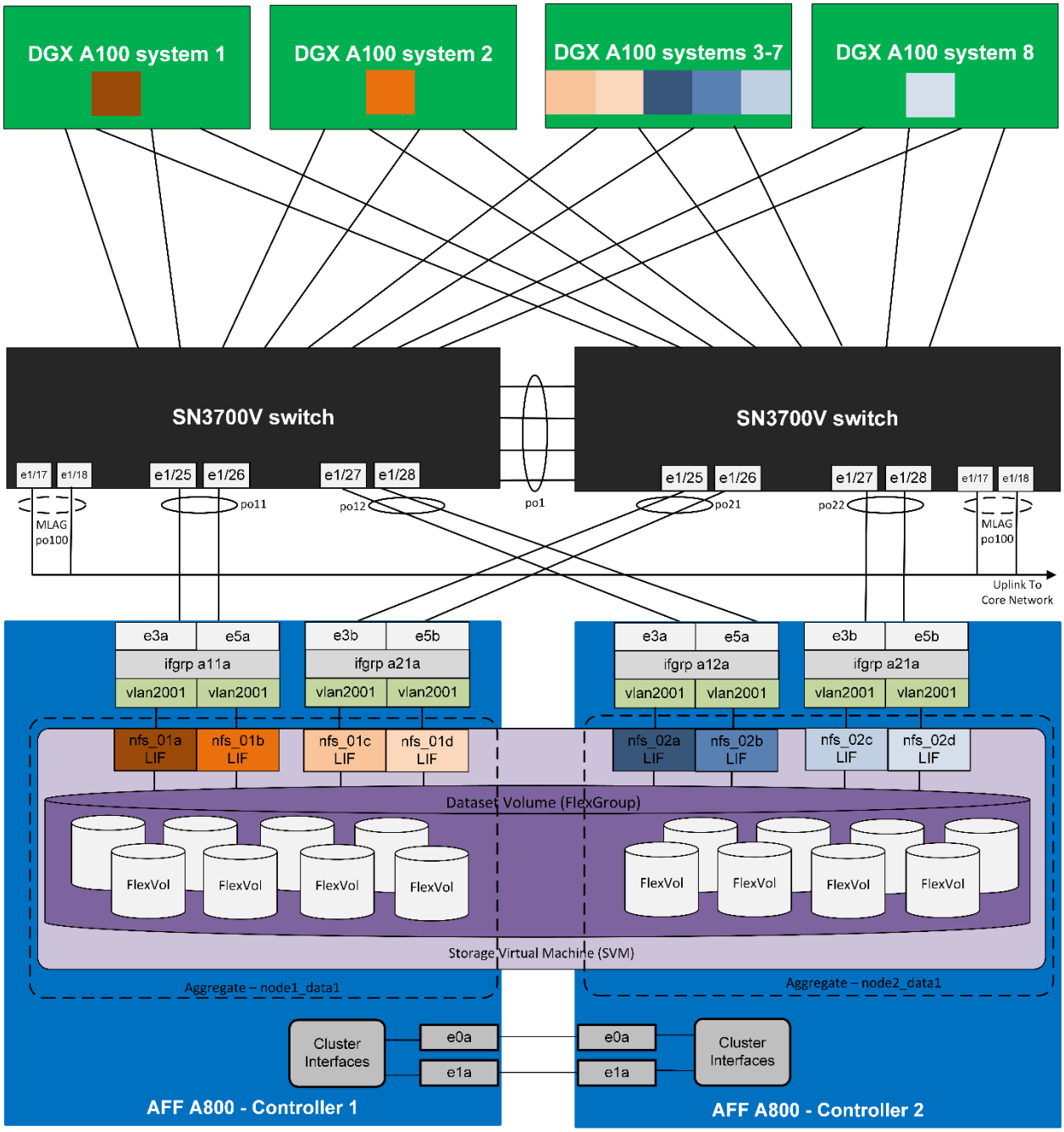


Storage system configuration

To support the storage network requirements of any potential workload on this architecture, each storage controller is provisioned with four 100GbE ports in addition to the onboard ports that are required for storage cluster interconnection. Figure 6 shows the storage system configuration. Each controller is configured with a two-port LACP interface group (ifgrp in Figure 6) to each switch. These interface groups provide up to 200Gbps of resilient connectivity to each switch for data access. Two VLANs are provisioned for NFS storage access, and both storage VLANs are trunked from the switches to each of these interface groups. This configuration allows concurrent access from each host to the data through multiple interfaces, which improves the potential bandwidth that is available to each host.

All data access from the storage system is provided through NFS access from a storage virtual machine (SVM) that is dedicated to this workload. The SVM is configured with a total of four logical interfaces (LIFs), with two LIFs on each storage VLAN. Each interface group hosts a single LIF, resulting in one LIF per VLAN on each controller with a dedicated interface group for each VLAN. However, both VLANs are trunked to both interface groups on each controller. This configuration provides the means for each LIF to fail over to another interface group on the same controller, so that both controllers stay active in the event of a network failure.

Figure 6) Storage system configuration.



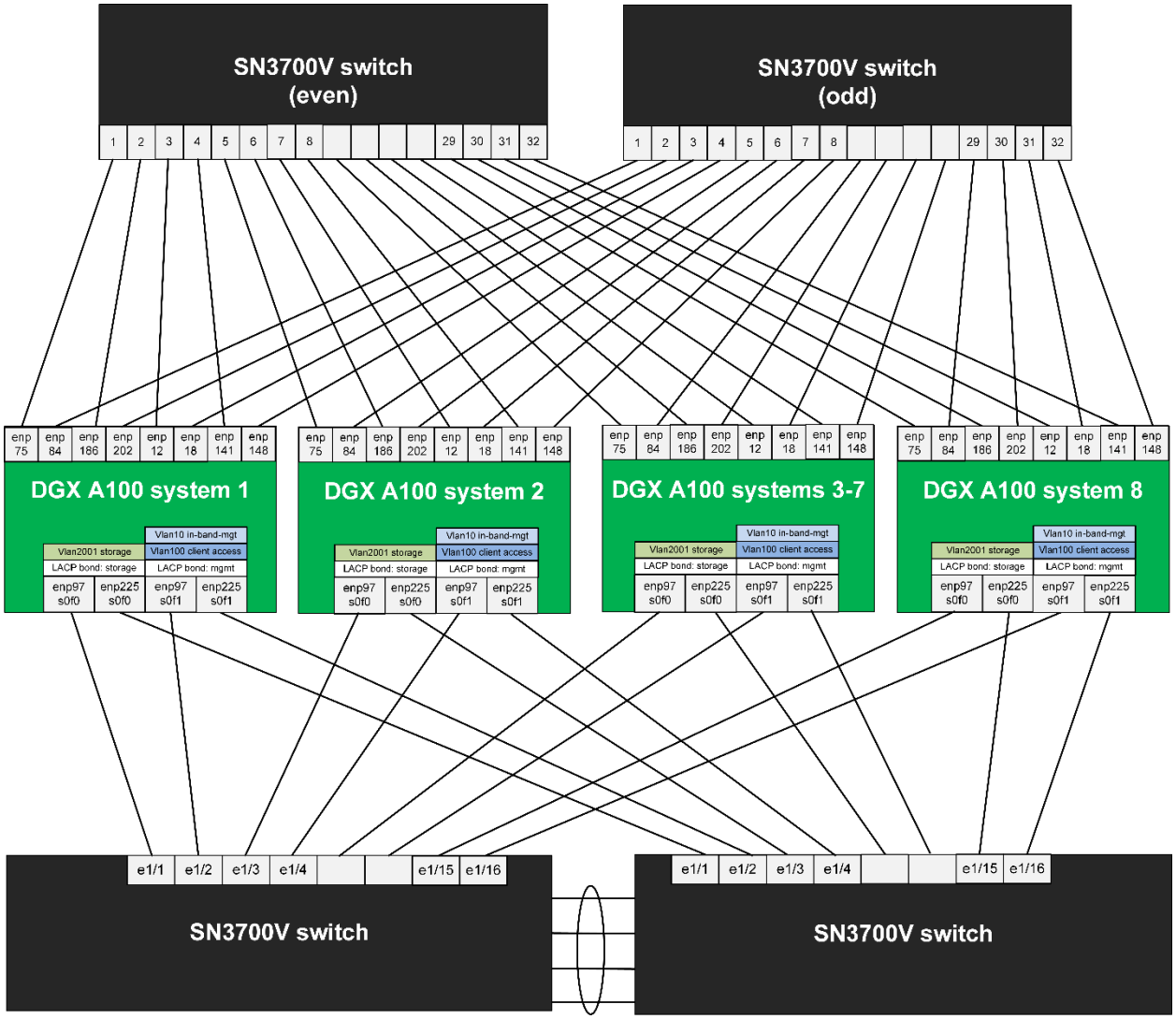
For logical storage provisioning, the solution uses a FlexGroup volume to provide a single pool of storage that is distributed across the nodes in the storage cluster. Each controller hosts an aggregate of 46 disk partitions, with both controllers sharing every disk. When the FlexGroup is deployed on the data SVM, eight FlexVol volumes are provisioned on each aggregate and are then combined into the FlexGroup. This approach allows the storage system to provide a single pool of storage that can scale up to the maximum capacity of the array and provide exceptional performance by leveraging all the SSDs in the array concurrently. NFS clients can access the FlexGroup as a single mount point through any of the LIFs that are provisioned for the SVM. Capacity and client access bandwidth can be increased simply by adding more nodes to the storage cluster. Note that multiple IP addresses are not required to achieve full

performance of either the controllers or FlexGroup volume, but they do allow for better hashing and load distribution in the network.

Host configuration

For network connectivity, each DGX A100 system is provisioned with eight ConnectX-6 single-port network interface cards for compute cluster connectivity and two ConnectX-6 dual-port cards for storage and client access connectivity. These cards support up to 200Gb link speeds for both InfiniBand and Ethernet. In this reference architecture, the eight single-port cards are configured for 200Gb RoCE and connected to a pair of SN3700V switches for compute cluster connectivity. The ports on the dual-ported card are connected to another pair of SN3700V switches for storage and client networking. Figure 7 shows the network port and VLAN configuration of the DGX A100 systems.

Figure 7) Network port and VLAN configuration of the DGX A100 systems.



For Ethernet storage networking, two physical ports are configured as an LACP port-channel on the host side and an MLAG on the switch side. Two additional ports are configured as another LACP bond for in-band management and client access traffic. Due to the high-performance capabilities of the AFF A800 storage system, host-side NFS filesystem caching was disabled for this testing.

DGX OS 4.99 and later use the Linux 5.3 kernel, which includes the NFS nConnect feature that significantly enhances NFSv3 storage performance. nConnect allows a single NFS mount to leverage multiple TCP sessions to increase the available bandwidth potentially up to the wire-speed maximum. This architecture was validated with nConnect to simplify host configuration while delivering performance comparable to previous configurations with multiple mounts. Specific host-side mount parameters used in this testing are listed below:

- `nConnect=8`. Creates eight TCP sessions for each mounted volume to improve overall performance.
- `rsize=262144, wsize=262144`. Sets the maximum read and write transfer size to 256k. ONTAP supports NFS transfer sizes as high as 1MB, but testing has shown that 256k delivers maximum throughput at the lowest latency.

Solution verification

This reference architecture was validated using synthetic benchmark utilities and deep learning benchmark tests to establish baseline performance and operation of the system. Each of the tests described in this section was performed with the specific equipment and software listed in Technology requirements.

Infrastructure validation

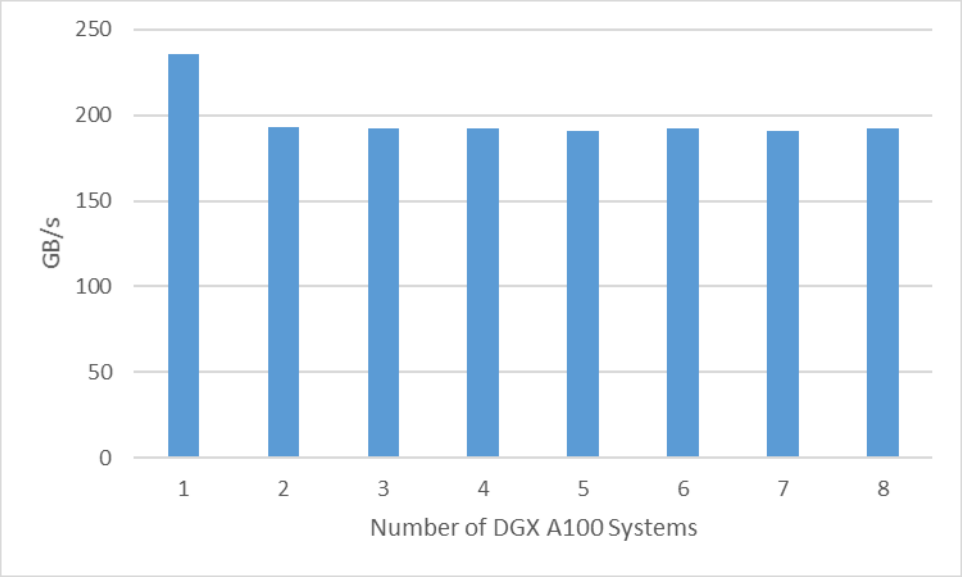
The following tests were performed with one, two, and four DGX A100 systems to validate basic operation and performance of the deployed infrastructure:

- NVIDIA nvsm stress test. This test suite performs a pass/fail verification of many crucial DGX A100 systems. All systems should report a passing status for the tests in this group.
- NVIDIA NCCL all_reduce_perf
- FIO bandwidth test
- FIO I/O operations per second (IOPS) test
- The following sections describe details and results for each of these tests.

NVIDIA NCCL all_reduce_perf test

This test validates the performance of the interconnects between GPUs. For single-node systems, the bottleneck should be the NVIDIA NVLink connection between GPUs. For multi-node systems, the bottleneck should be the Ethernet or InfiniBand connections between DGX A100 systems. This test measures the combined bandwidth between systems using all eight available physical connections. Figure 8 shows the results of the NCCL all_reduce_perf test.

Figure 8) NCCL bandwidth test result.



FIO bandwidth and IOPS tests

These tests are intended to measure the storage system performance using the synthetic I/O generator tool FIO. Two separate configurations were used, one optimized to deliver maximum bandwidth and the other optimized for IOPS. Each configuration was run with both 100% reads and 100% writes, and the files used by FIO were created as a separate step to isolate those activities from the actual test results. Here are the specific FIO configuration parameters for these tests:

- ioengine = posixaio
- direct = 1
- blocksize = 1024k for bandwidth test, 4k for IOPS test
- numjobs = 120 for bandwidth test, 180 for IOPS test
- iodepth = 32
- size = 4194304k

With the workload parameters used in these tests, it is possible to saturate each storage controller using three DGX A100 systems. Figure 9 shows the results of the FIO bandwidth tests with up to eight DGX A100 systems. In this configuration with four hosts mounted to each controller, performance scales in a linear manner until it reaches a single-controller maximum of around 22GBps with three hosts and plateaus at the fourth host. The next four hosts are mounted to the second controller in the HA pair and show similar behavior up to the maximum of roughly 45GBps for both controllers.

Figure 9) FIO bandwidth test results (GBps).

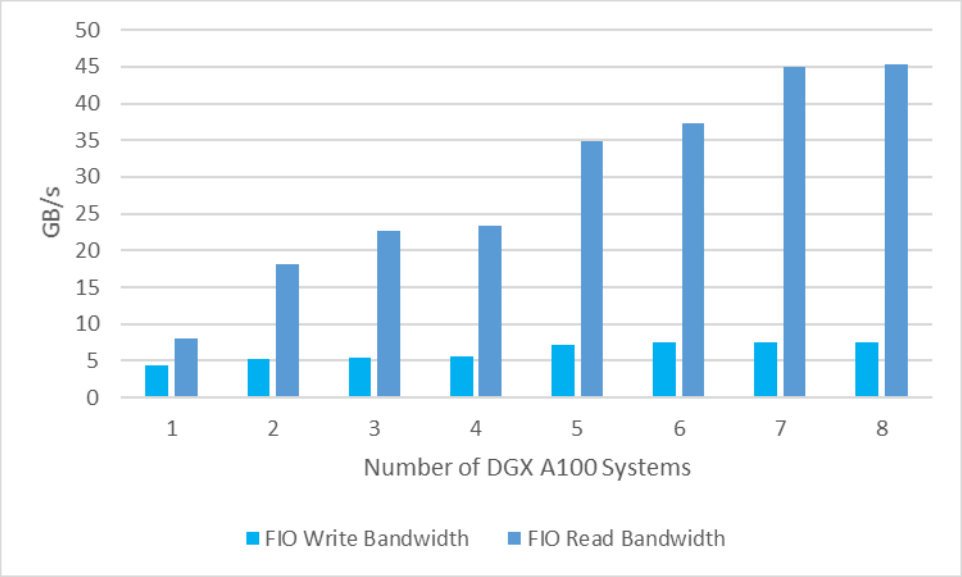
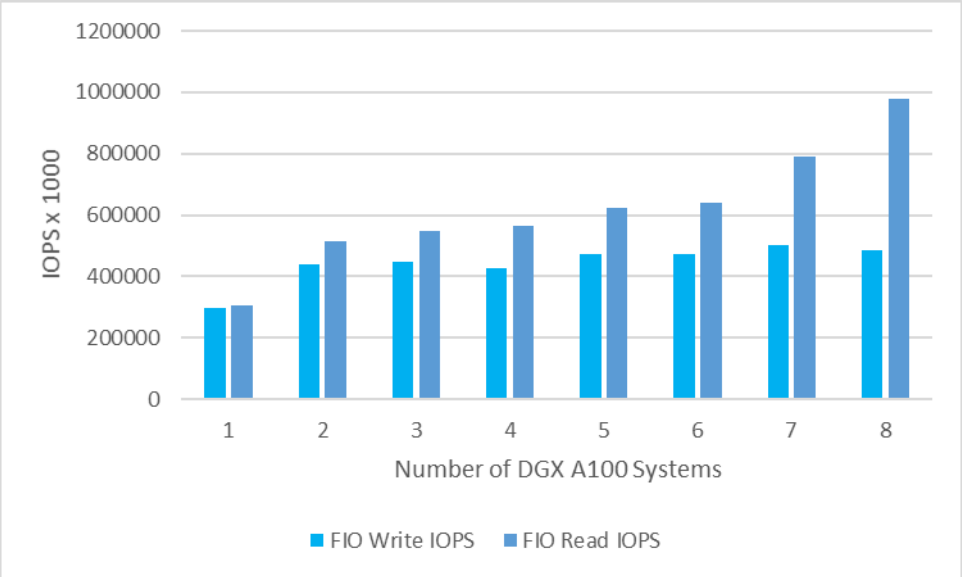


Figure 10 shows the results of the FIO IOPS test.

Figure 10) FIO IOPS test results (operations/sec).



Deep learning workload validation

The operation of deep learning workloads on the deployed infrastructure was validated using the MLPerf Training v0.7 ResNet-50 benchmark test. This test uses the MLPerf v0.7 testing criteria for validating performance of systems using the ResNet-50 model with parameters and dataset specified in the MLPerf v0.7 testing specification.

The following section contains specific details and results for this test.

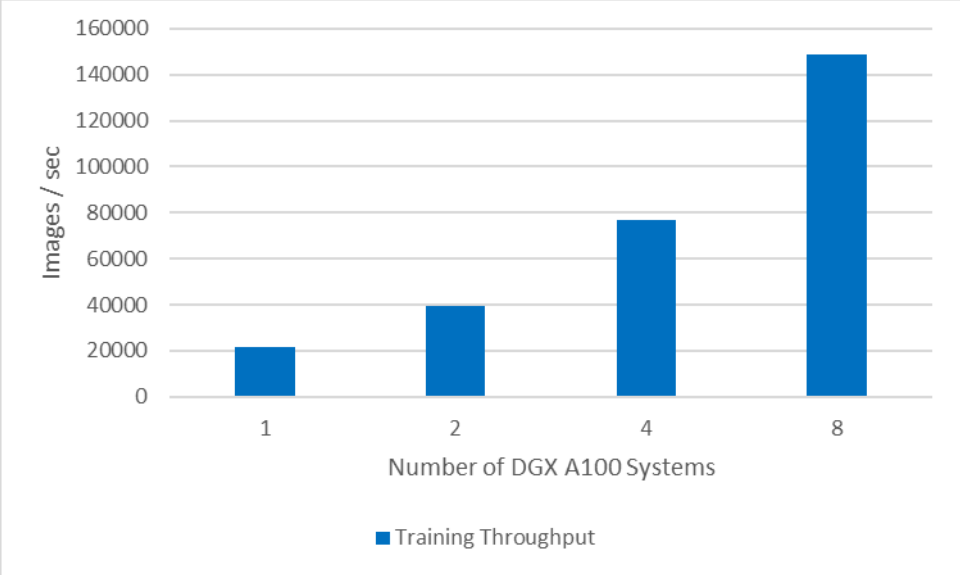
MLPerf Training v0.7 ResNet-50

This reference architecture was tested using a MLPerf Training v0.7 benchmark to validate the operation of deep learning workloads on the deployed infrastructure. MLPerf is an industry-standard benchmark implementation of various neural networks for validating the performance of deep learning infrastructure. This test used the MXNet implementation of ResNet-50 along with the ImageNet dataset in IORecord format to validate model training performance. DALI was used to accelerate ingest and preprocessing of data, and Horovod was used to distribute the training across multiple DGX A100 systems. The results presented maintain a consistent batch size per system of 408 images as the workload is scaled (weak scaling).

The base container image used for these tests is the 20.06 MXNet image from NGC. MLPerf benchmark tests are deliberately not optimized for any specific hardware implementation so overall system performance in these tests can be increased by tuning parameters such as concurrency.

Figure 11 shows the average images per second for the training run duration of 45 epochs.

Figure 11) MLPerf Training v0.7 average images per second.



Solution sizing guidance

This architecture is intended as a reference for customers and partners who would like to implement a DL infrastructure with NVIDIA DGX A100 systems and a NetApp AFF system.

As is demonstrated in this validation, the AFF A800 system easily supports the DL training workload generated by eight DGX A100 systems. For larger deployments with higher storage performance requirements, additional AFF A800 systems can be added to the NetApp ONTAP cluster. ONTAP 9 supports up to 12 HA pairs (24 nodes) in a single cluster. With the FlexGroup technology validated in this solution, a 24-node cluster can provide over 20 PB and up to 300 GBps throughput in a single volume. Although the dataset used in this validation was relatively small, ONTAP 9 can scale to impressive capacity with linear performance scalability, because each HA pair delivers performance comparable to the level verified in this document.

Other NetApp storage systems such as the AFF A400 offer lower performance and capacity options for smaller deployments at lower cost points. Based on the results of this testing, an AFF A400 storage system can support one or two DGX A100 systems with the workloads that were tested. Because ONTAP 9 supports mixed-model clusters, customers can start with a smaller initial footprint and add more or larger storage systems to the cluster as capacity and performance requirements grow.

Conclusion

The DGX A100 system is a next-generation deep learning platform that requires equally advanced storage and data management capabilities. By combining DGX A100 with NetApp AFF systems, this verified architecture can be implemented at almost any scale, from a single DGX A100 paired with an AFF A400 storage system up to potentially 48 DGX A100 systems on a 12-node AFF A800 cluster. Combined with the superior cloud integration and software-defined capabilities of NetApp ONTAP, AFF enables a full range of data pipelines that spans the edge, the core, and the cloud for successful DL projects.

Acknowledgments

The authors gratefully acknowledge the contributions that were made to this technical report by our esteemed colleagues from NVIDIA and NetApp. Our sincere appreciation and thanks go to all the individuals who provided insight and expertise that greatly assisted in the research for this paper.

Where to find additional information

To learn more about the information that is described in this document, review the following resources:

- NVA-1153-DEPLOY: NetApp ONTAP AI with NVIDIA DGX A100 Systems and Mellanox Spectrum Ethernet Switches:
  www.netapp.com/pdf.html?item=/media/21789-nva-1153-deploy.pdf

NetApp AFF systems:

- AFF datasheet
  https://www.netapp.com/us/media/ds-3582.pdf
- NetApp Flash Advantage for AFF
  https://www.netapp.com/us/media/ds-3733.pdf
- ONTAP 9.x documentation
  http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286
- NetApp FlexGroup technical report
  https://www.netapp.com/us/media/tr-4557.pdf

NetApp Interoperability Matrix:

- NetApp Interoperability Matrix Tool
  http://support.netapp.com/matrix

NetApp Trident:

- https://netapp.io/persistent-storage-provisioner-for-kubernetes/
- https://netapp-trident.readthedocs.io/en/stable-v19.04/kubernetes/index.html
- https://github.com/NetApp/trident

NVIDIA DGX A100 systems:

- NVIDIA DGX A100 systems
  https://www.nvidia.com/en-us/data-center/dgx-a100/
- NVIDIA A100 Tensor core GPU
  https://www.nvidia.com/en-us/data-center/a100/
- NVIDIA GPU Cloud
  https://www.nvidia.com/en-us/gpu-cloud/

NVIDIA Mellanox networking:

- NVIDIA Mellanox Spectrum SN3000 series switches
  https://www.mellanox.com/products/ethernet-switches/sn3000

Machine learning frameworks:

- TensorFlow: An Open-Source Machine Learning Framework for Everyone
  https://www.tensorflow.org/
- Horovod: Uber's Open-Source Distributed Deep Learning Framework for TensorFlow
  https://eng.uber.com/horovod/
- Enabling GPUs in the Container Runtime Ecosystem
  https://devblogs.nvidia.com/gpu-containers-runtime/

Dataset and benchmarks:

- ImageNet
  http://www.image-net.org/
- MLPerf training and inference benchmarks
  https://mlperf.org/

Version history

Version	Date	Document Version History
Version 1.0	November 2020	Initial release

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

NetApp