NetApp Verified Architecture

NetApp ONTAP AI with Mellanox Spectrum Switches

David Arnette and Sung-Han Lin, NetApp
Satinder Nijjar and Jacci Cenci, NVIDIA
Luan Nguyen and John Kim, Mellanox

September 2019 | NVA-1138-DESIGN

Abstract

This document describes a NetApp Verified Architecture for machine learning and artificial workloads using NetApp® AFF A800 storage systems, NVIDIA® DGX-1™ systems, and Mellanox® Spectrum® network switches. It also contains benchmark test results for the architecture as implemented.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Executive Summary

This document contains validation information for the NetApp® ONTAP® AI reference architecture for machine learning and artificial intelligence workloads. This design was implemented using a NetApp AFF A800 all-flash storage system, nine NVIDIA® DGX-1™ systems, and Mellanox® Spectrum® SN2700 100Gb Ethernet (100GbE) switches with Mellanox Onyx™ OS. The operation and performance of this system were validated using industry-standard benchmark tools. Based on the validation testing results, this architecture delivers excellent training and inferencing performance. The results also demonstrate adequate storage headroom for supporting additional DGX-1 systems. Customers can also easily and independently scale compute and storage resources from half-rack to multirack configurations with predictable performance to meet any machine learning workload requirement.

2 Program Summary

The NetApp Verified Architecture program gives customers reference configurations and sizing guidance for specific workloads and use cases. The program 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.

2.1 NetApp ONTAP AI Solution

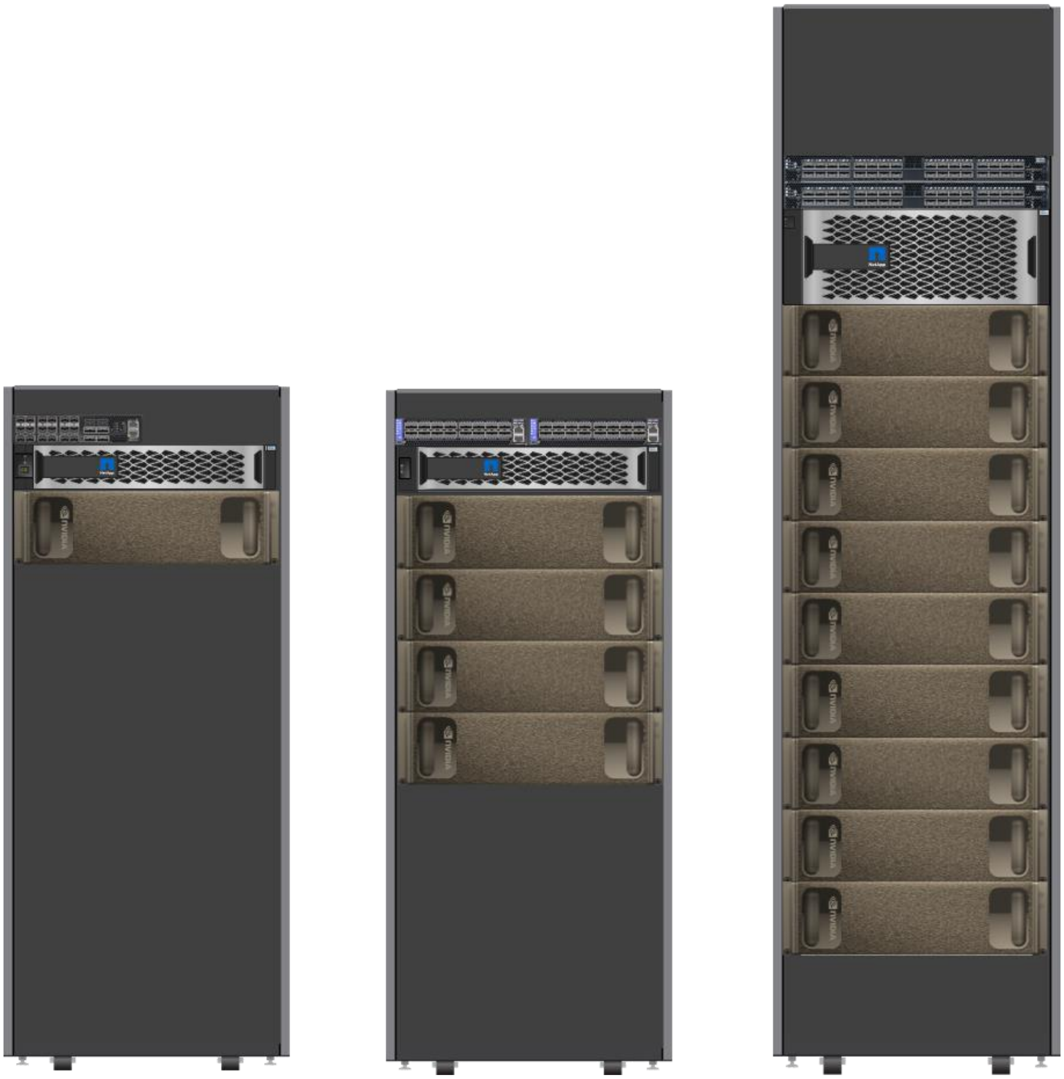
The NetApp ONTAP AI reference architecture, powered by NVIDIA DGX systems and NetApp cloud-connected storage systems, was developed and verified by NetApp and NVIDIA. It gives IT organizations an architecture that:

- 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 NVIDIA DGX systems and NetApp A800 storage systems with state-of-the-art networking. NetApp ONTAP AI with DGX systems 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 three variations in the ONTAP AI family of solutions with Mellanox switches. The AFF A800 system performance has been verified with DGX-1 systems and has demonstrated enough headroom to support nine or more DGX-1 systems without impacting storage throughput or latency. Furthermore, by adding more network switches and storage controller pairs to the ONTAP cluster, the solution can scale to multiple racks to deliver extremely high throughput, accelerating training and inferencing. 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 Mellanox Spectrum.



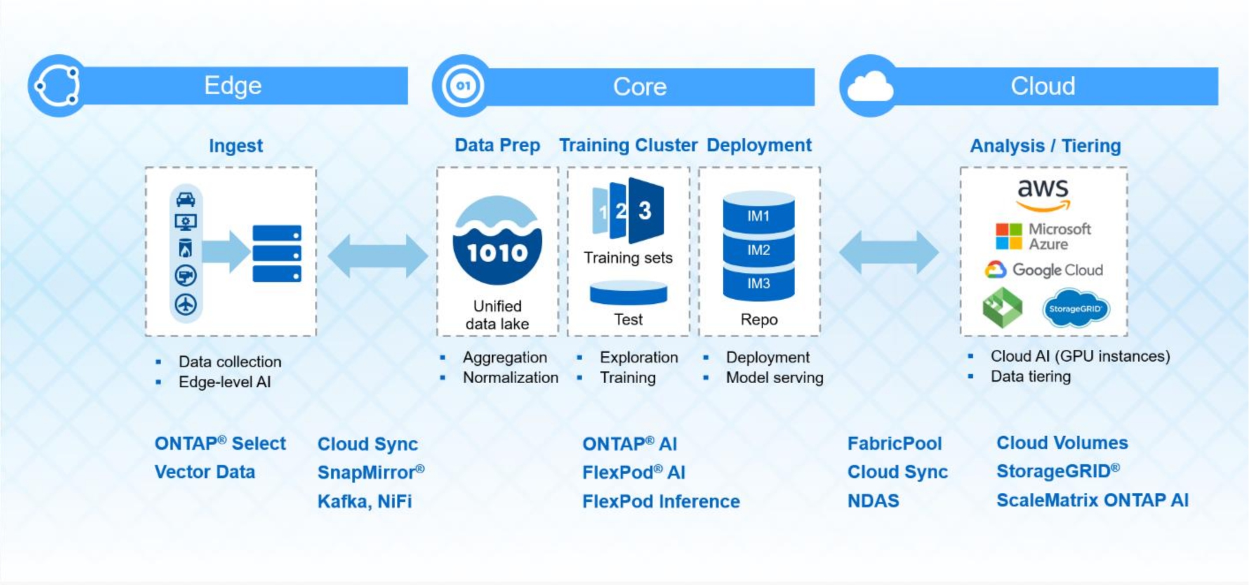
The number of DGX systems and AFF systems per rack depends on the power and cooling specifications of the rack in use. Final placement of the systems is subject to computational fluid dynamics analysis, airflow management, and data center design.

3 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 be deployed 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.



Here are descriptions of 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 instance, 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.** 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 the 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.

Some of the 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 high 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

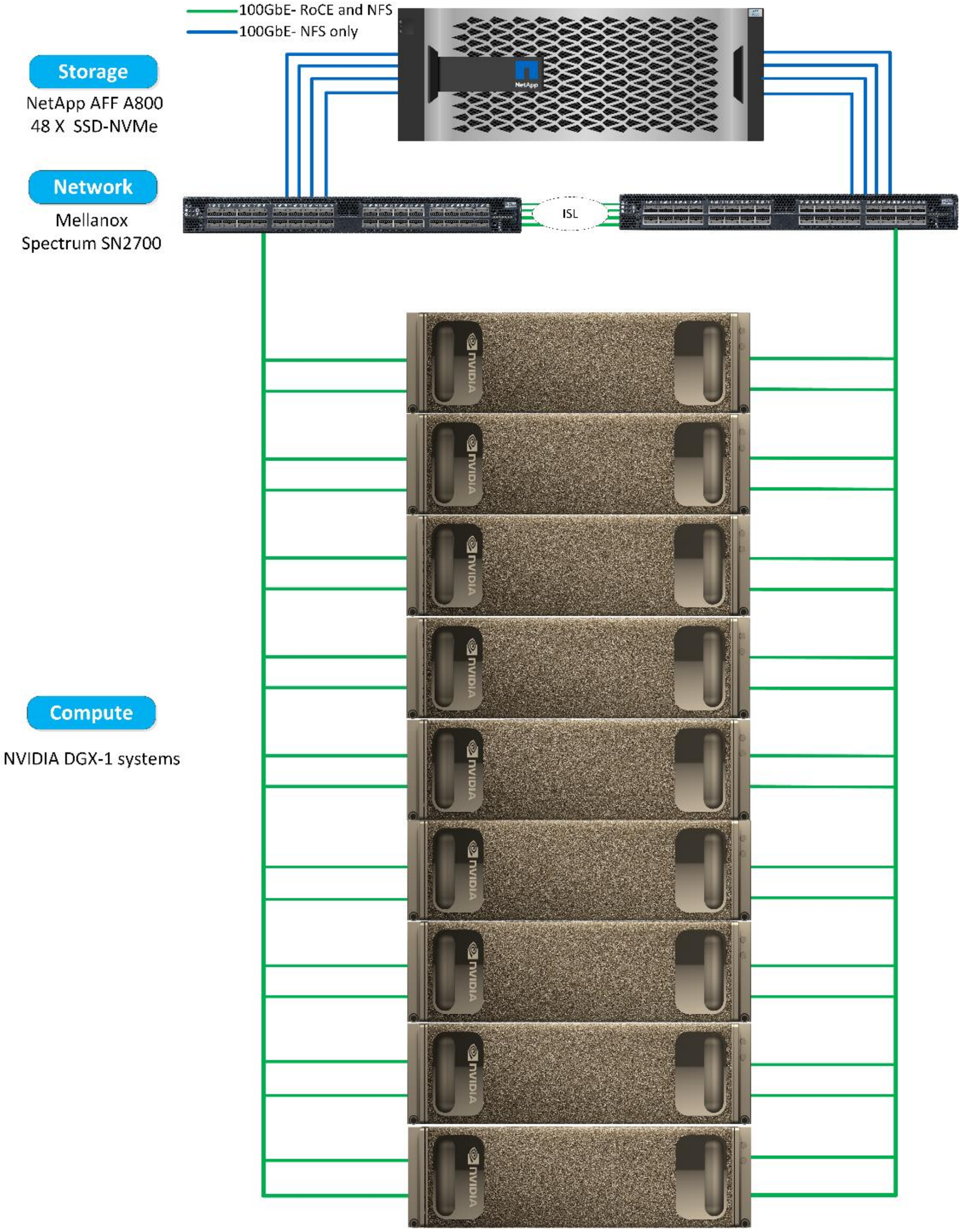
Combined with superior cloud integration and the software-defined capabilities of NetApp ONTAP, NetApp AFF systems support a full range of data pipelines that spans the edge, the core, and the cloud for DL. This document focuses on solutions for the training and inference components of the data pipeline.

4 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 solution was implemented with one NetApp AFF A800 system, nine DGX-1 servers, and two Mellanox Spectrum SN2700 100GbE switches. Each DGX-1 system is connected to the Mellanox switches with four 100GbE connections that are used for inter-GPU communications by using remote direct memory access (RDMA) over Converged Ethernet (RoCE). Traditional IP communications for storage access via NFS also occur on these links. Each storage controller is connected to the network switches by using four 100GbE links. Figure 3 shows the basic solution architecture.

Figure 3) NetApp ONTAP AI verified architecture.



4.1 NVIDIA DGX-1 Systems

The DGX-1 system is a fully integrated, turnkey hardware and software system that is purpose-built for DL workflows. Each DGX-1 system is powered by eight NVIDIA Tesla® V100 GPUs that are configured in a hybrid cube-mesh topology that uses NVIDIA NVLink® technology, which provides an ultra-high-bandwidth, low-latency fabric for inter-GPU communication within the DGX 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 GPU count increases. The DGX-1 system is also equipped with high-bandwidth, low-latency network interconnects for multinode clustering over RDMA-capable fabrics such as Ethernet (RoCE) or InfiniBand.

4.2 NVIDIA GPU Cloud

The DGX-1 system leverages the NVIDIA GPU Cloud® (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-1 system and across multiple DGX-1 systems.

4.3 NetApp AFF Systems

NetApp AFF 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-500 µs 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 30 1TB 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 9.5, with a complete suite of data protection and replication features for industry-leading data management

NetApp also offers other storage systems, such as the AFF A700, AFF A320, and AFF A220 that provide lower performance and capacity options for smaller deployments at lower cost points.

4.4 NetApp ONTAP 9

NetApp ONTAP 9 is the latest generation of storage management software from NetApp that 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 future-proof infrastructure across hybrid cloud architectures.

Simplify Data Management

Data management is crucial to enterprise IT operations to make sure 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 and expanded deduplication.** 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.
- **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 the NetApp StorageGRID® solution.

Accelerate and Protect Data

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

- **High performance and low 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.

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 such as OpenStack, Hadoop, and MongoDB by using the same infrastructure that supports existing enterprise apps.

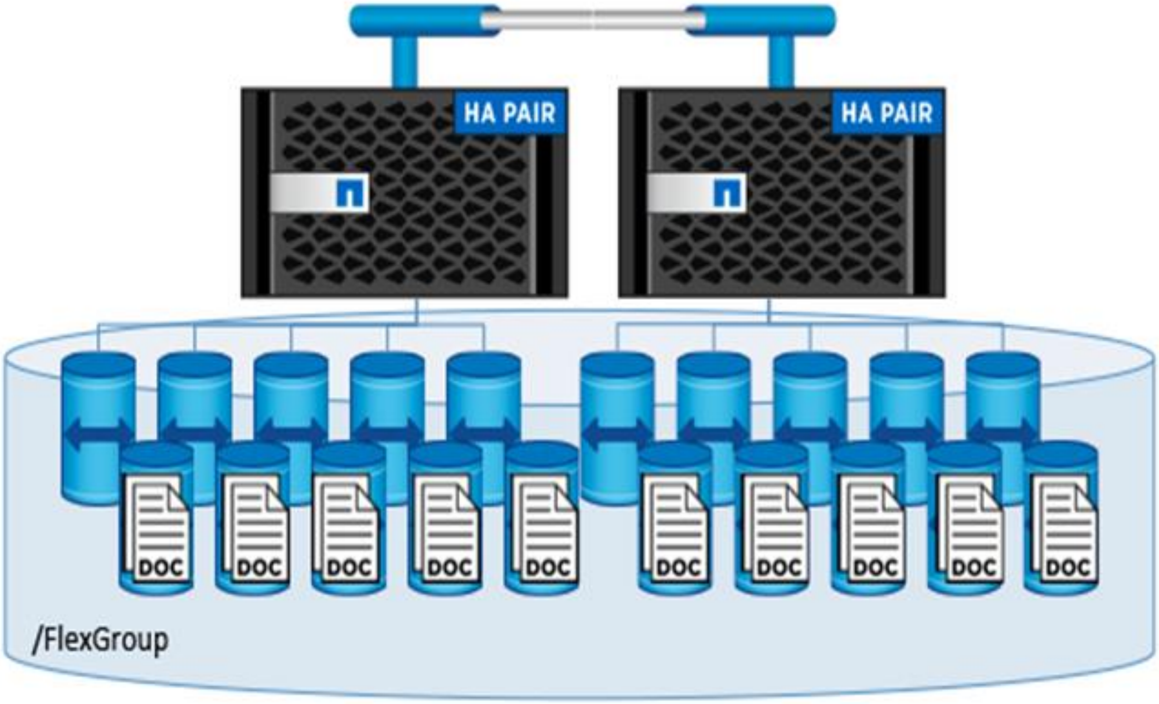
4.5 NetApp FlexGroup Volumes

The training dataset is usually a large collection of many, 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, as shown in Figure 4, 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.

Figure 4) NetApp FlexGroup volumes.



4.6 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, providing 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.

4.7 Mellanox Spectrum Switches—The Right Choice for Deep Learning Workloads

Network is a critical part of the DL infrastructure that is responsible for moving massive amounts of data between the end points efficiently and effectively. Spectrum Ethernet switches with consistent performance, intelligent load balancing, and comprehensive telemetry are the 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. Along with the ConnectX® adapters sitting inside the DGX systems, Spectrum implements a tight and efficient explicit congestion notification mechanism that mitigates transient congestion and smooths traffic burst to maximize network goodput.

Intelligent Load Balancing

The network is a shared resource, and its bandwidth needs to be shared in a fair manner across different flows and different endpoints. Packet buffer architecture is a foundational attribute of the switch that affects performance as well as 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 and flows at a cost to others. This traffic imbalance leads to more performance variation and hampers distributed DL performance.

Comprehensive Telemetry

To reap high return on investment from the DL infrastructure, uptime 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. 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 in to 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 Mellanox WJH, customers can dramatically reduce mean time to issue resolution and plan capacity better.

4.8 RDMA over Converged Ethernet

Direct memory access (DMA) enables hardware subsystems such as disk drive controllers, sound cards, graphics cards, and network cards to access system memory to perform data read/write without using CPU processing cycles. RDMA extends that capability across different server and storage systems by allowing network adapters to do a server-to-server data transfer between application memory by using zero-copy functionality without involving any OS or device driver. This approach dramatically reduces CPU overhead and latency by bypassing the kernel for read/write and send/receive operations.

RoCE is the most widely deployed and highest-performing implementation of RDMA over Ethernet, and it leverages new Converged Enhanced Ethernet (CEE) standards. It is now available as a standard feature in many high-end network adapters, converged network adapters, and network switches. Traditional Ethernet uses a best-effort delivery mechanism for network traffic and is not suitable for delivering the low latency and high bandwidth that are required for communications between GPU nodes. CEE enables a lossless physical-layer networking medium and the ability to optionally allocate bandwidth to any specific traffic flow on the network.

To ensure lossless, in-order delivery of Ethernet packets, CEE networks use Priority Flow Control (PFC) and Enhanced Transmission Selection (ETS). PFC enables the sending of pause frames for each specific Class of Service (CoS), which allows you to limit specific network traffic while allowing other traffic to flow freely. ETS allows specific bandwidth allocation for each CoS to provide even more granular control over network utilization. For more information, see this article from Mellanox.

The ability to prioritize RoCE over all other traffic allows the 100GbE links to be used for both RoCE and traditional TCP/IP traffic, such as the NFS storage access traffic that is demonstrated in this solution.

4.9 Automation with Ansible

Ansible is a configuration management tool from Red Hat that is quickly becoming the standard for DevOps-style system administration. Ansible accelerates time to value during deployment and improves stability and reduces administrative overhead during daily operations. Ansible’s declarative methodology for hardware and software management allows the administrator to specify the intended state of the configuration in a set of easy-to-read YAML files. The administrator can manage the state of the infrastructure with version controls and change validation processes, just like any other software code.

Ansible was originally designed for Linux administration, but it also includes an extensible framework for management of almost any device. NetApp and Mellanox offer extensive module support, enabling deployment and management of the entire ONTAP AI infrastructure by using Ansible. The infrastructure used in this validation was configured in less than 25 minutes by using Ansible modules that are publicly available in the official distribution of Ansible. For more information, see the blog post and demonstration video at https://blog.netapp.com/how-to-configure-ontap-ai-in-20-minutes-with-ansible-automation/.

5 Technology Requirements

This section covers the hardware and software that were used for all the testing described in Section 7, “Solution Verification.”

5.1 Hardware Requirements

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

Table 1) Hardware requirements.

Hardware	Quantity
NVIDIA DGX-1 systems	9
NetApp AFF A800 system	1 high-availability (HA) pair, includes 48x 1.92TB NVMe SSDs
Mellanox Spectrum SN2700 network switches	2
Management network switch (Mellanox AS4610, optional)	1

5.2 Software Requirements

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

Table 2) Software requirements.

Software	Version
NetApp ONTAP	9.5
Mellanox Onyx switch firmware	3.8.1208
NVIDIA DGX OS	4.0.4 - Ubuntu 18.04 LTS
Docker container platform	18.06.1-ce [e68fc7a]
Container version	netapp_tf_19.02 based on nvcr.io/nvidia/tensorflow:19.02-py2
Machine learning framework	TensorFlow 1.12.2
Horovod	0.15.1
OpenMPI	3.1.3
Benchmark software	TensorFlow benchmarks [7b9e1b4]

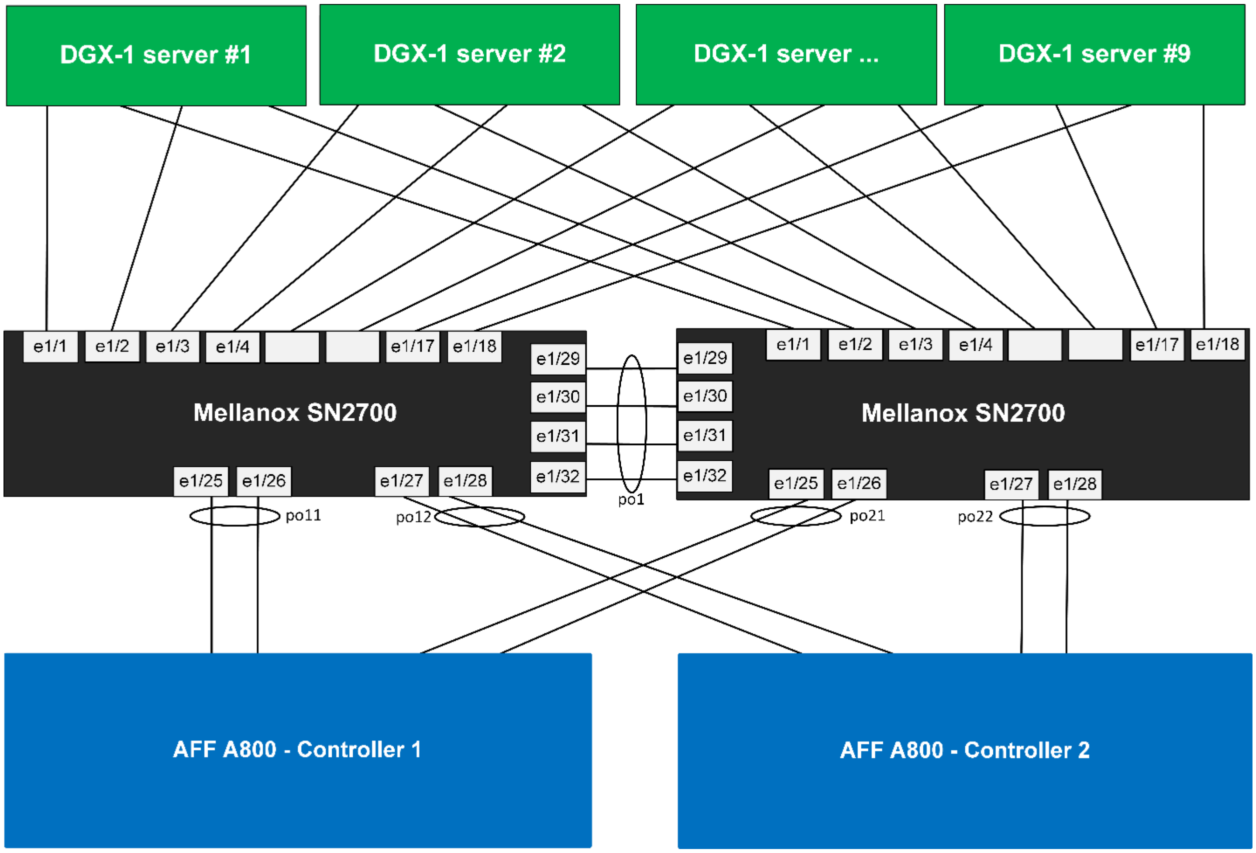
6 Solution Architecture

This solution architecture has been verified to meet the requirements for running DL workloads. This verification enables data scientists to deploy DL frameworks and applications on a prevalidated infrastructure, thereby helping to eliminate risks and allowing 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.

6.1 Network Topology and Switch Configuration

This network architecture uses a pair of Mellanox SN2700 100GbE switches for the primary intercluster network using RoCE and for the storage access network using NFS. These switches are connected to each other with four 100GbE ports that are configured as a standard port channel. This Inter-Switch Link (ISL) port channel allows traffic to flow between the switches during host or storage system link failures. Each host is connected to the Mellanox switches with a pair of active-passive bonds. Also, to provide link-layer redundancy, each storage controller is connected to each SN2700 switch with a two-port LACP port channel. Figure 5 shows the network switch-port configuration.

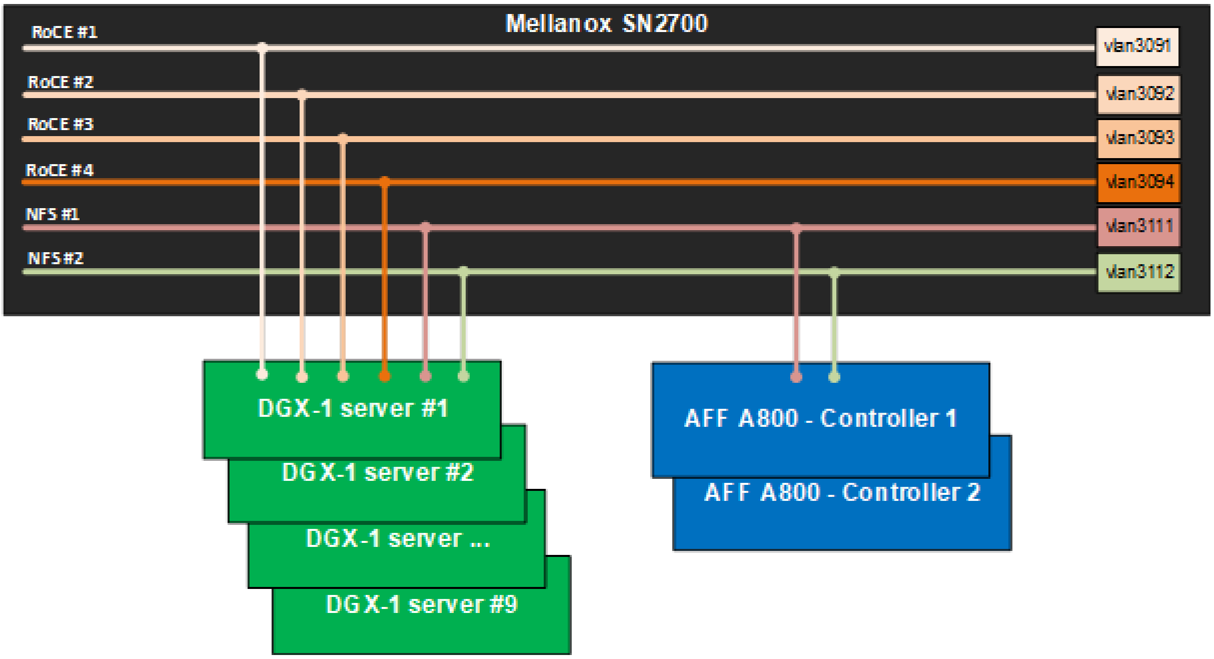
Figure 5) Network switch port configuration.



Multiple virtual LANs (VLANs) were provisioned to support both RoCE and NFS storage traffic. Four VLANs are dedicated to RoCE traffic, and two VLANs are dedicated to NFS storage traffic. Four discrete VLANs and IP ranges are used to provide symmetrical routing for each RoCE connection, and the software stack manages these connections for bandwidth aggregation and fault tolerance. For storage access, this solution uses NFSv3, which does not support multipath access, so two VLANs are used to enable multiple dedicated NFS mounts. This approach does not provide any additional fault tolerance, but it does enable multiple links to be used to increase available bandwidth. PFC is configured on each switch to assign all four RoCE VLANs to the priority traffic class, and the NFS VLANs are assigned to the default best-effort traffic class. All VLANs are configured for jumbo frames with a maximum transmission unit (MTU) size of 9000.

The switch ports for DGX-1 systems are configured as trunk ports, and all RoCE and NFS VLANs are permitted. The port channels that are configured for the storage system controllers are also trunk ports, but only the NFS VLANs are permitted. Figure 6 shows the VLAN connectivity for the DGX-1 system and storage system ports.

Figure 6) VLAN connectivity for DGX-1 and storage system ports.



To provide priority service for RoCE traffic, the host network adapter assigns a CoS value of 4 to traffic on each RoCE VLAN. The switch is configured with a QoS policy that provides no-drop service to traffic with this CoS value. NFS traffic is assigned the default CoS value of 0, which falls into the default QoS policy on the switch and provides best-effort service.

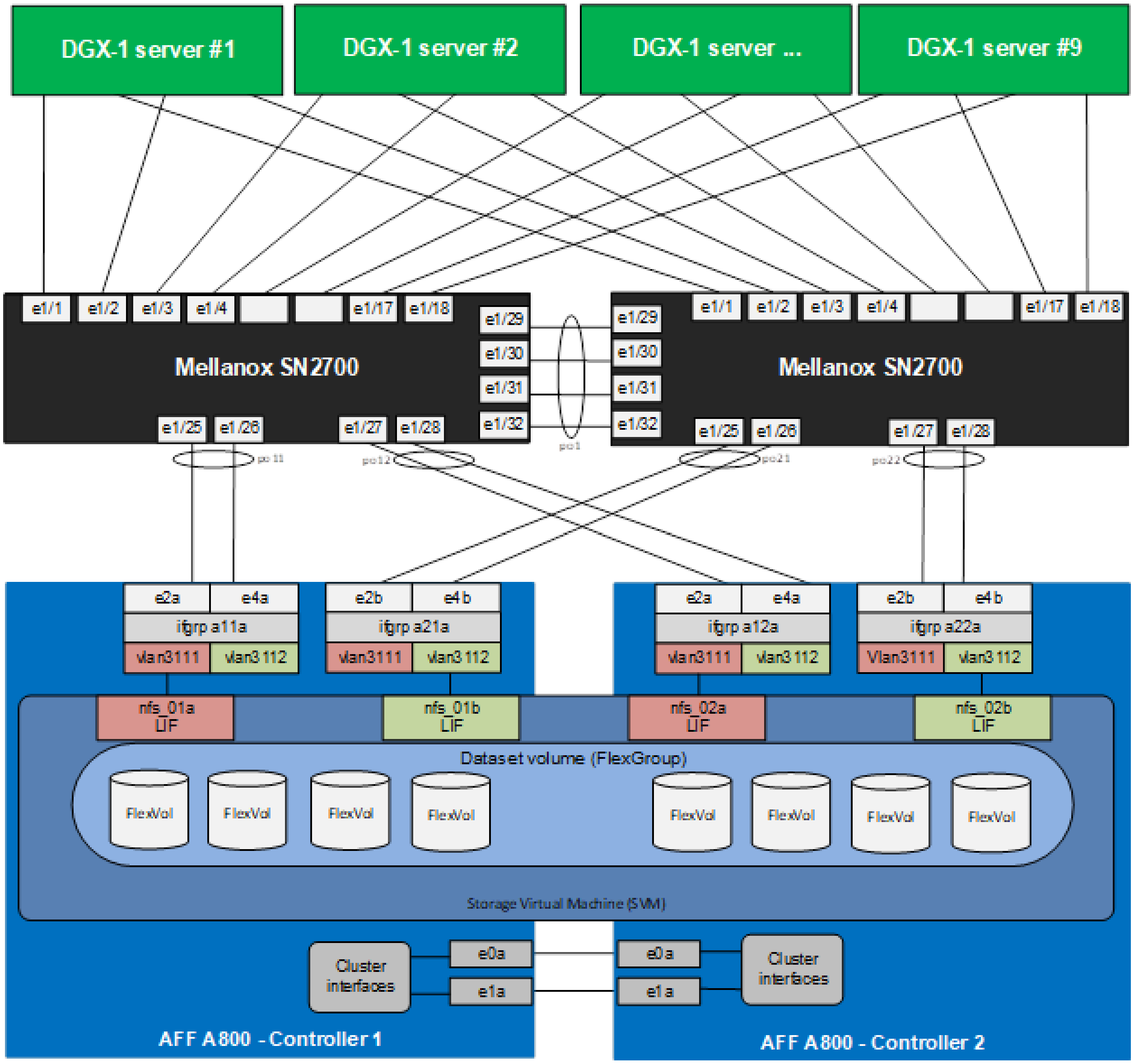
PFC is then enabled on each DGX-1 port, which enables the switch port to send pause frames for specific classes of service to eliminate congestion at the switch. By using ETS to allocate 95% of the bandwidth to RoCE traffic in the event of congestion, this configuration allows dynamic resource allocation between RoCE and NFS traffic while providing priority to node-to-node communication. Most of the time the network provides more than enough bandwidth to support both the GPU-to-GPU traffic on RoCE and the NFS storage traffic on TCP/IP. However, if a link is temporarily saturated, the QoS policy allows the network to prioritize the DGX GPU traffic while ensuring that NFS and other TCP/IP-based traffic will still get through. Bandwidth allocation can be modified dynamically to optimize for workloads that require higher storage performance and less internode communication.

6.2 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 7 shows the storage system configuration. Each controller is configured with a two-port LACP interface group (ifgrp in Figure 7) 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 7) Storage system configuration.



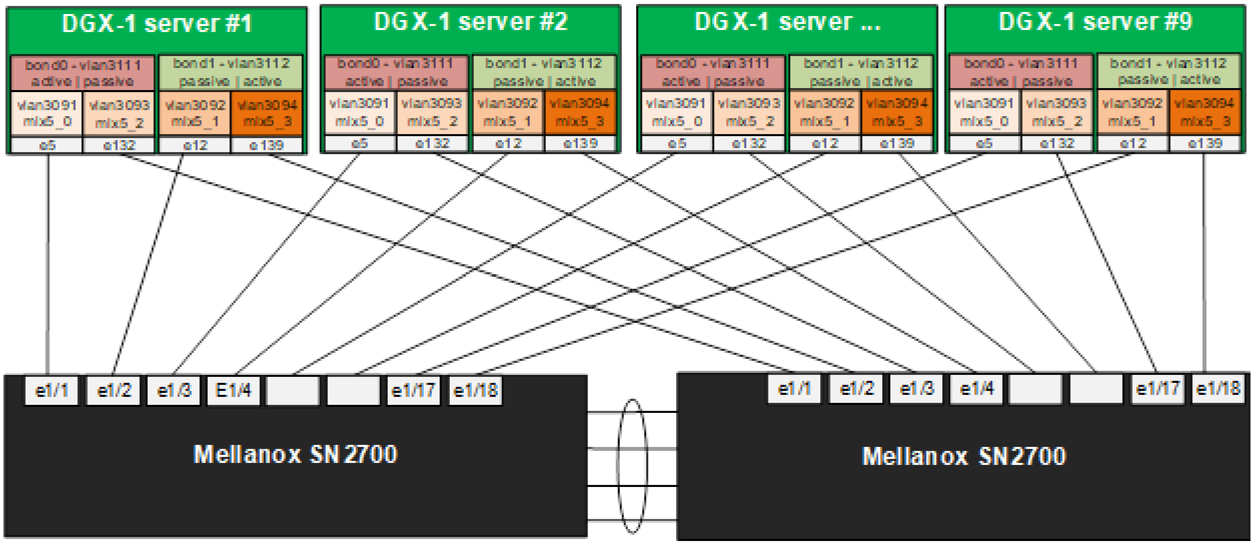
For logical storage provisioning, the solution uses a FlexGroup volume to provide a single pool of storage that is distributed across the nodes in the storage cluster. Each controller hosts an aggregate of 46 disk partitions, with both controllers sharing every disk. When the FlexGroup is deployed on the data SVM, several 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. You can increase capacity and client access bandwidth by simply adding more nodes to the storage cluster.

6.3 Host Configuration

For network connectivity, each DGX-1 system is provisioned with four Mellanox ConnectX-4 single-port network interface cards. These cards operate at up to 100GbE speeds and support both RoCE and InfiniBand, with RoCE offering a lower-cost alternative to IB for cluster interconnect applications. Each 100Gb port is configured as a trunk port on the appropriate switch, with four RoCE and two NFS VLANs allowed on each. Figure 8 shows the network port and VLAN configuration of the DGX-1 systems.

Figure 8) Network port and VLAN configuration of the DGX-1 hosts.



For RoCE connectivity, each physical port hosts a VLAN interface and IP address on one of the four RoCE VLANs. The Mellanox drivers are configured to apply a network CoS value of 4 to each of the RoCE VLANs, and PFC is configured on the switches to guarantee priority lossless service to the RoCE class. RoCE does not support aggregating multiple links into a single logical connection, but the NCCL communication software can use multiple links for bandwidth aggregation and fault tolerance.

For NFS storage access, two active-passive bonds are created by using a link to each switch. Each bond hosts a VLAN interface and IP address on one of the two NFS VLANs, and each bond's active port is connected to a different switch. This configuration offers up to 100Gb of bandwidth in each NFS VLAN and provides redundancy in the event of any host link or switch failure scenario. To provide optimal performance for the RoCE connections, all NFS traffic is assigned to the default best-effort QoS class. All physical interfaces and the bond interfaces are configured with an MTU of 9000.

To increase data access performance, multiple NFSv3 mounts are made from the DGX-1 system to the storage system. Each DGX-1 system is configured with two NFS VLANs, with an IP interface on each VLAN. The FlexGroup volume on the AFF A800 system is mounted on each of these VLANs on each DGX-1 system, providing completely independent connections from the server to the storage system. Although a single NFS mount can deliver the performance that is required for this workload, multiple mount points are defined to enable the use of additional storage access bandwidth for other workloads that are more storage intensive.

7 Solution Verification

This section describes the testing performed to validate the operation and performance of this solution. All of the tests described in this section were performed with the specific equipment and software listed in section 5, “Technology Requirements.”

7.1 Validation Test Plan

This solution was verified by using standard benchmarks with several compute configurations to demonstrate the scalability of the architecture. The ImageNet dataset was hosted on the AFF A800 system by using a single FlexGroup volume that was accessed by up to nine DGX-1 systems using NFSv3, as recommended by NVIDIA for external storage access. TensorFlow was used as the machine learning framework for all the models that were tested, and compute and storage performance metrics were captured for each test case. Highlights of that data are presented in section 7.2, “Validation Test Results.”

The following convolutional neural network (CNN) models were used to demonstrate training rates. These models are used for image recognition and classification and represent varying degrees of compute and storage complexity:

- **ResNet-152** is generally considered to be the most accurate training model.
- **ResNet-50** delivers faster processing time than ResNet-152 but with potentially lower accuracy.
- **VGG16** produces the highest inter-GPU communication.
- **Inception-v3** is another common image classification model.

The following configuration details were used across all models tested:

- ImageNet data was used with distortion disabled to reduce the overhead of CPU processing before copying data into GPU memory.
- Each model was tested with batch size 256.
- Each model was tested with one, three, five, seven, and nine DGX-1 systems to demonstrate the scalability of each model across multiple GPUs using RoCE as the interconnect with Horovod for distributed training.
- Inferencing was tested by using all the models with the ImageNet dataset, batch size 256, and all GPUs in nine DGX-1 systems.

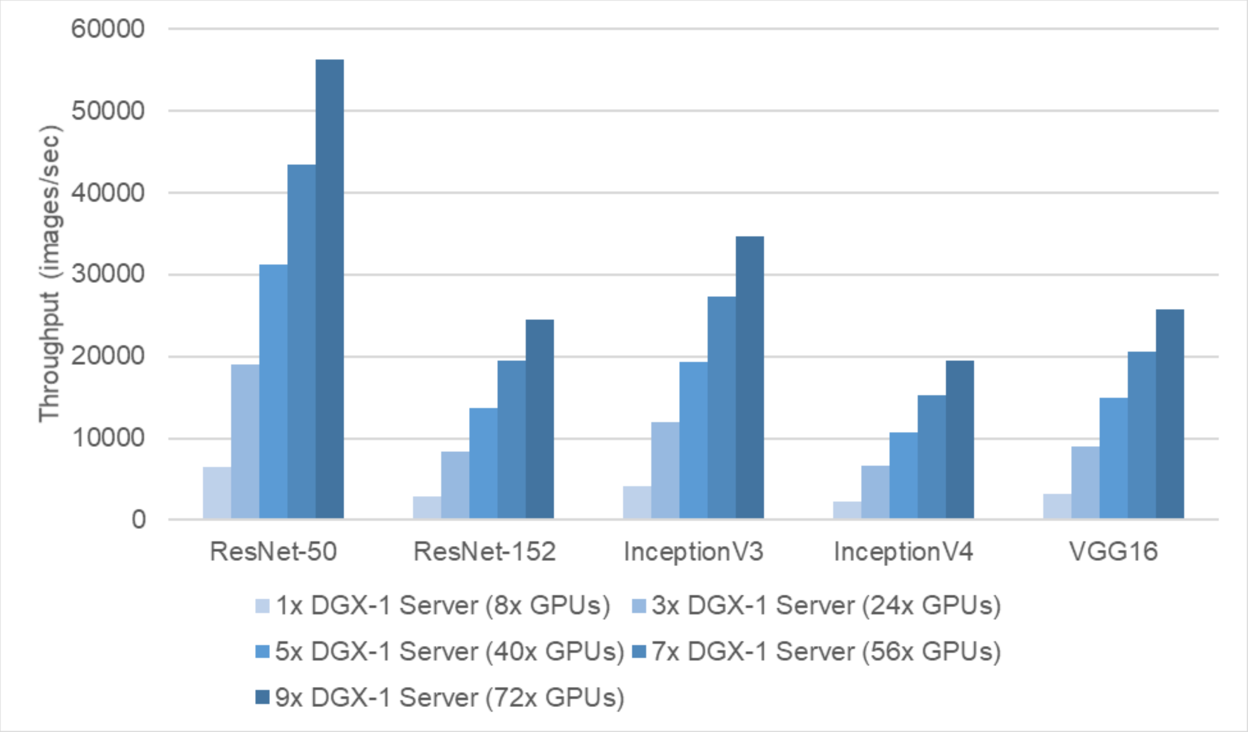
7.2 Validation Test Results

As described in the previous section, various tests were conducted to assess the general operation and performance of this solution. This section contains overall training and inferencing performance data that was collected during those tests.

Overall Training Throughput

Figure 9 shows the maximum number of training images per second achieved with each of the models that were tested by using Tensor cores for maximum performance. The graph compares the training throughput that was achieved with up to nine DGX-1 systems and demonstrates the linear scalability of this solution using NFS and RoCE.

Figure 9) Training throughput for all models.

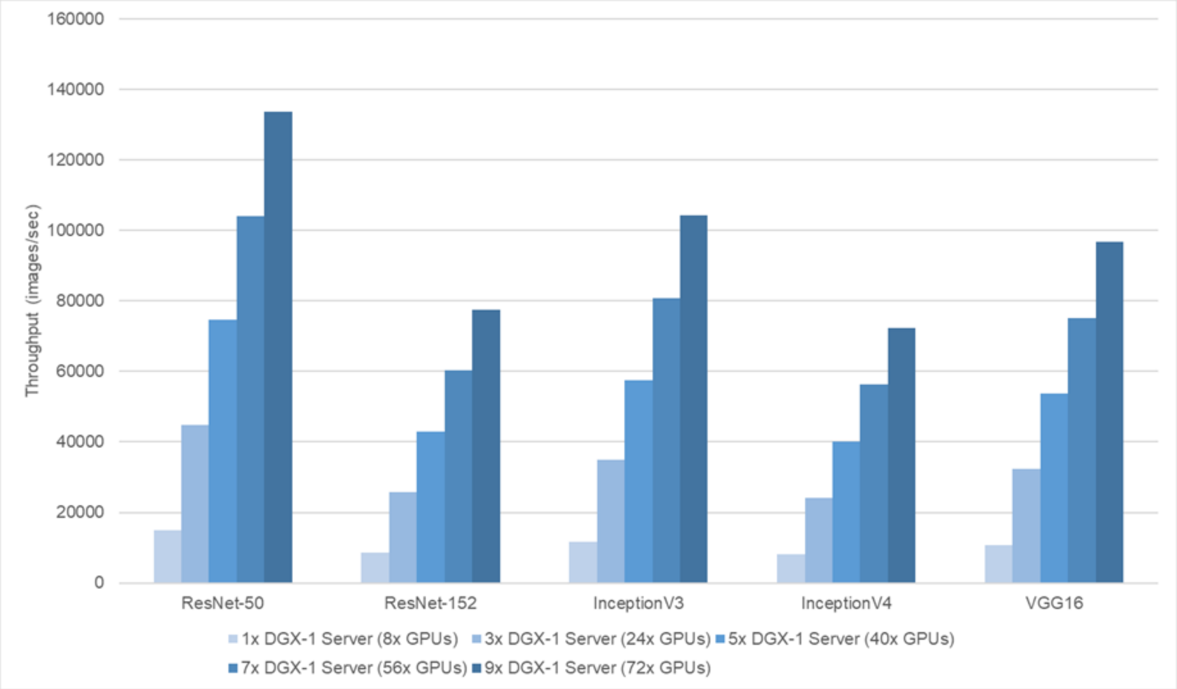


Inference with GPUs

Inferencing is the process of deploying the DL model to assess a new set of objects and making predictions with similar predictive accuracies as observed during the training phases. In an application with an image dataset, the goal of inferencing is to classify the input images and to respond to the requesters as quickly as possible. In addition to achieving high throughput, minimizing latency becomes important.

NetApp ONTAP AI was used to demonstrate inferencing and to measure throughput metrics in this phase. Figure 10 shows the number of images that can be processed per second during inferencing.

Figure 10) Inference for all models.



7.3 Solution Sizing Guidance

This architecture is intended as a reference for customers and partners who want to implement a high-performance computing (HPC) infrastructure with NVIDIA DGX-1 systems and a NetApp AFF system.

As demonstrated in this validation, the AFF A800 system easily supports the DL training workload generated by nine DGX-1 systems. For even larger deployments with even 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 20PB and up to 300GBps throughput in a single volume. Although the dataset that was 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.

NetApp also offers other storage systems, such as the AFF A700, AFF A320, and AFF A220 that provide lower performance and capacity options for smaller deployments at lower cost points. 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 their capacity and performance requirements grow.

This solution as tested consumed almost all the ports on each Mellanox SN2700 switch. Remaining ports could be used for uplinks to a core or spine network as the infrastructure grows.

8 Conclusion

The DGX-1 system is an extremely powerful DL platform that benefits from equally powerful storage and network infrastructure to deliver maximum value. By combining NetApp AFF systems with Mellanox Spectrum switches, this verified architecture can be implemented at almost any scale, from a single DGX-1 paired to an AFF A220 system up to potentially 96 DGX-1 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 Mellanox 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:

- NVIDIA DGX-1 systems:
  - NVIDIA DGX-1 systems
    https://www.nvidia.com/en-us/data-center/dgx-1/
  - NVIDIA Tesla V100 Tensor core GPU
    https://www.nvidia.com/en-us/data-center/tesla-v100/
  - NVIDIA GPU Cloud
    https://www.nvidia.com/en-us/gpu-cloud/
- 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
- Mellanox Spectrum SN2000 series switches
  - https://www.mellanox.com/page/products_dyn?product_family=251&mtag=sn2000
- Machine learning framework:
  - 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/
  - TensorFlow benchmarks
    https://www.tensorflow.org/performance/benchmarks

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.

Copyright Information

Trademark Information

NVA-1138-DESIGN-0919

NetApp®