NetApp Verified Architecture

NetApp ONTAP AI, Powered by NVIDIA

Scalable AI Infrastructure: Designing for Real-World Deep Learning Use Cases

David Arnette, Sundar Ranganathan, Amit Borulkar, Sung-Han Lin, and Santosh Rao, NetApp

March 2019 | NVA-1121

In partnership with

NVIDIA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Executive Summary

This document contains validation information for the architecture that is described in the technical white paper WP-7267: Scalable AI Infrastructure. The design from that white paper was implemented by using the NetApp® AFF A800, an all-flash FAS system; NVIDIA® DGX-1™ servers; and Cisco® Nexus® 3232C 100Gb Ethernet switches. We validated the operation and performance of this system by using industry-standard benchmark tools, and, based on the validation testing results, this architecture delivers excellent training and inferencing performance. The results also demonstrate adequate storage headroom for supporting multiple DGX-1 servers. You can also 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.

2 Program Summary

2.1 NetApp Verified Architecture Program

The NetApp Verified Architecture program offers customers a verified architecture for NetApp solutions. With a NetApp Verified Architecture, you get a NetApp solution architecture that:

- Is thoroughly tested
- Is prescriptive in nature
- Minimizes deployment risks
- Accelerates time to market

This document is for NetApp and partner solutions engineers and customer strategic decision makers. The document describes the architecture design considerations that were used to determine the specific equipment, cabling, and configurations that are required for a particular environment.

2.2 NetApp ONTAP AI Solution

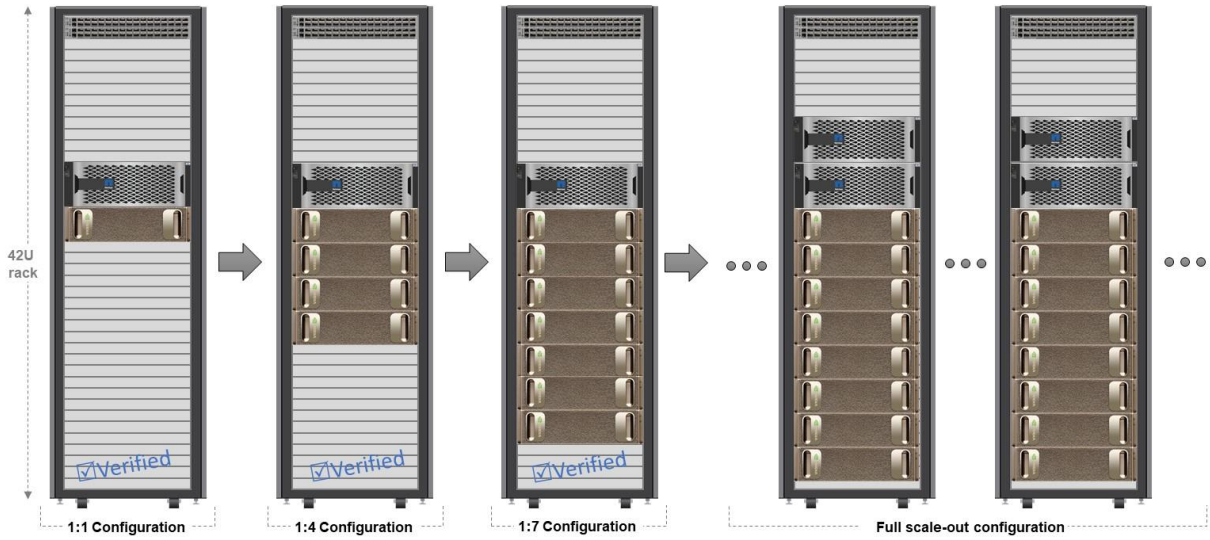
NetApp ONTAP® AI Converged Infrastructure, powered by NVIDIA DGX-1 servers and NetApp cloud-connected storage system, is an architecture that was developed and verified by NetApp and NVIDIA. It provides your organization with a prescriptive architecture that:

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

NetApp ONTAP AI integrates DGX-1 servers, NVIDIA Tesla® V100 GPUs, and a NetApp AFF A800 storage system with state-of-the-art networking. NetApp ONTAP AI simplifies artificial intelligence (AI) deployments by eliminating design complexity and guesswork. Your enterprise can start small and grow non-disruptively while intelligently managing data from the edge to the core to the cloud and back.

Figure 1 shows the scalability of the NetApp ONTAP AI solution. The AFF A800 system has been verified with seven DGX-1 servers and has demonstrated sufficient performance headroom to support more DGX-1 servers 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 the ratios of compute to storage 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 solution rack-scale architecture.



The number of DGX-1 servers 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

DL is the engine that enables you to detect fraud, to improve customer relationships, to optimize your supply chain, and to 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 data sets, it is critical to architect an infrastructure that gives you the flexibility to deploy across environments. At a high level, an end-to-end DL deployment consists of three stages through which the data travels: the edge (data ingest), the core (training clusters and a data lake), and the cloud (archive, tiering, and dev/test). This is very typical in applications such as the Internet of Things (IoT) for which data spans all three realms of the data pipeline.

Figure 2) Edge to core to cloud data pipeline.

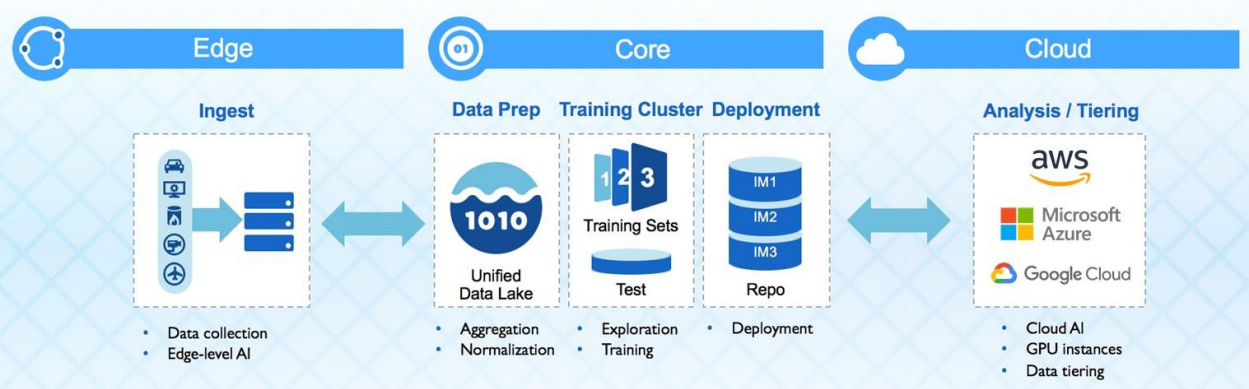


Figure 2 presents an overview of the components in each of the three realms:

- **Ingest.** Data ingestion usually occurs at the edge by, for example, capturing data streaming from autonomous cars or point-of-sale (POS) 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 to 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 utilizations.
- **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, tiering.** New cloud-based tools become available at a rapid pace, so additional analysis or development work may 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.

Depending on the application, DL models work with large amounts of different types of data (both structured and unstructured). 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 to 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, 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. The highly parallelized architecture of modern GPUs makes them substantially more efficient than general-purpose CPUs for applications such as DL, for which data processing is performed in parallel. Advances in individual and in clustered GPU computing architectures that leverage the DGX-1 server have made them the preferred platform for workloads such as high-performance computing (HPC), DL, and analytics. Providing maximized performance in these environments requires a supporting infrastructure that can keep GPUs fed with data. Dataset access must therefore be provided at ultra-low latencies with high bandwidth.

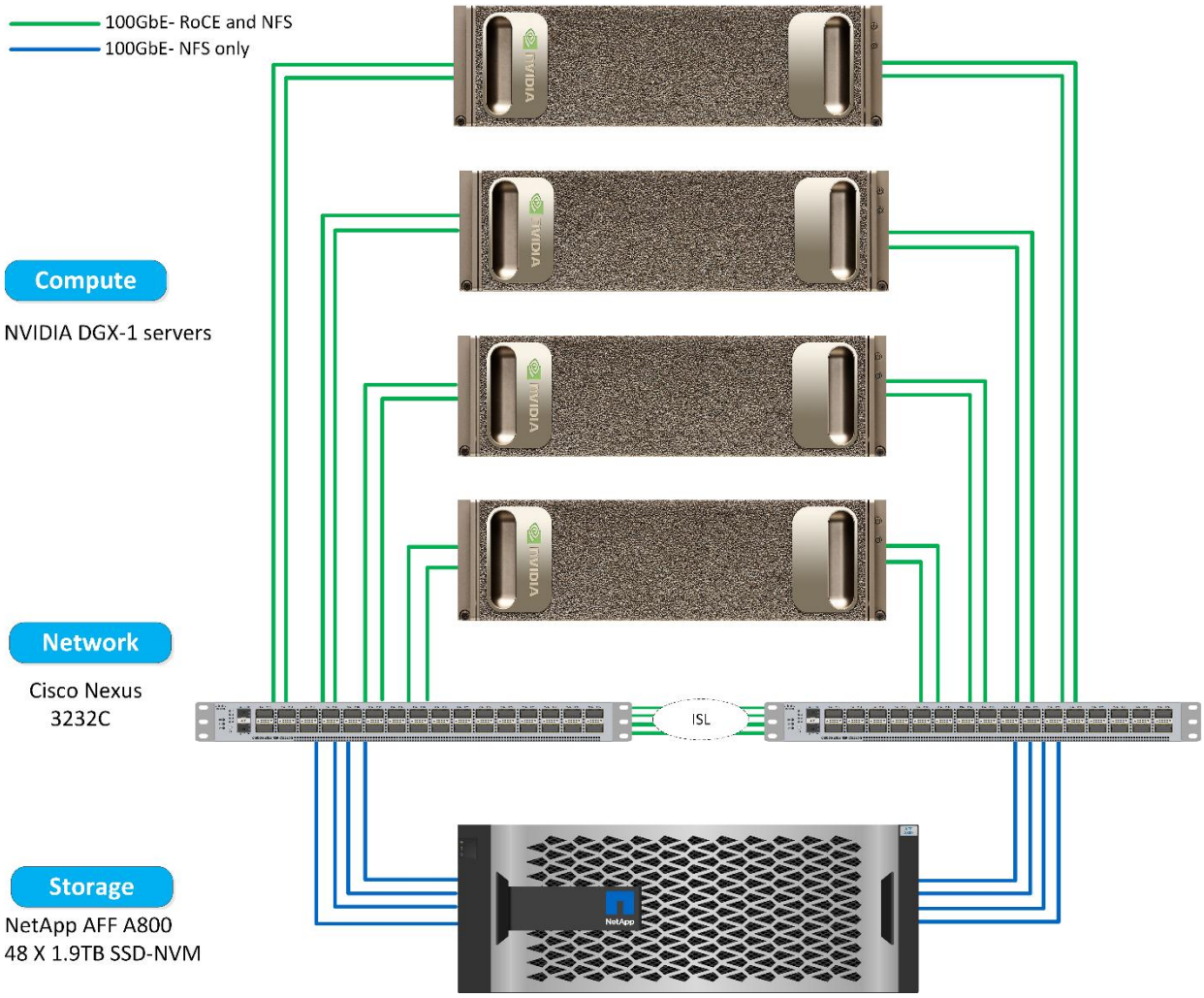
4.1 Solution Technology

This solution was implemented with one NetApp AFF A800 system, four DGX-1 servers, and two Cisco Nexus 3232C 100GbE switches. Each DGX-1 server is connected to the Nexus switches with four 100 GbE connections that are used for inter-GPU communications by using remote direct memory access (RDMA) over Converged Ethernet (RoCE). Traditional IP communications for NFS storage access also

occur on these links. Each storage controller is connected to the network switches by using four 100GbE links.

Traditional HPC infrastructures use RDMA over InfiniBand (IB) for internode connectivity because of its high-bandwidth and low-latency features. As Ethernet technology reaches performance levels that were previously possible only with IB, RoCE enables easier adoption of these capabilities because Ethernet technologies are well understood and are widely deployed in every enterprise data center. Figure 3 shows the basic solution architecture.

Figure 3) NetApp ONTAP AI solution verified architecture.



4.2 NVIDIA DGX-1 Servers

The DGX-1 server is a fully integrated, turnkey hardware and software system that is purpose-built for DL workflows. Each DGX-1 server is powered by eight 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. 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 server is also equipped with high-bandwidth, low-latency network interconnects for multi-node clustering over RDMA-capable fabrics.

The DGX-1 is powered by 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 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 multi-node collective communication primitives, enabling topology-awareness for DL training. NCCL enables communication between GPUs inside a single DGX-1 server and across multiple DGX-1 servers.

4.3 NetApp AFF Systems

NetApp AFF is a state-of-the-art storage system that enable you to meet enterprise storage requirements with the 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 a throughput of 25GB/s for sequential reads and 1 million IOPS for small random reads at sub-500µs latencies. AFF A800 systems support the following features:

- A massive throughput of up to 300GB/s and 11.4 million IOPS in a 24-node cluster
- 100GbE together with 32Gb FC connectivity
- 30TB solid-state drives (SSDs) with multi-stream write (MSW)
- High density with 2PB in a 2U drive shelf
- Scaling from 364TB (2 nodes) to 74PB (24 nodes)
- NetApp ONTAP 9.4, with a complete suite of data protection and replication features for industry-leading data management

The next best storage system in terms of performance is the AFF A700s system, supporting a throughput of 18GB/s for NAS workloads and 40GbE transport. AFF A300 and AFF A220 systems offer sufficient performance at lower cost points.

4.4 NetApp ONTAP 9

ONTAP 9 is the latest generation of storage management software from NetApp that enables you to modernize your infrastructure and transition to a cloud-ready data center. Leveraging industry-leading data management capabilities, ONTAP enables you to manage and to protect data with a single set of tools regardless of where the data resides. Data can also be moved freely to wherever it's needed, either 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 critical to enterprise IT operations so that appropriate resources are used for applications and for data sets. ONTAP includes the following features to streamline and simplify operations and to reduce the TCO:

- **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 you 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), Azure, and the NetApp StorageGRID® solution.

Accelerate and Protect Data

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

- **Performance and lower latency.** ONTAP offers the highest possible throughput at the lowest possible latency.
- **NetApp ONTAP FlexGroup.** A FlexGroup volume is a high-performance data container that can scale linearly up to 20PB and 400 billion files, providing a single name space that simplifies data management.
- **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 you meet demanding and constantly changing business needs:

- **Seamless scaling and non-disruptive operations.** ONTAP supports non-disruptive addition of capacity to existing controllers as well as scale-out clusters. You 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 provides 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.

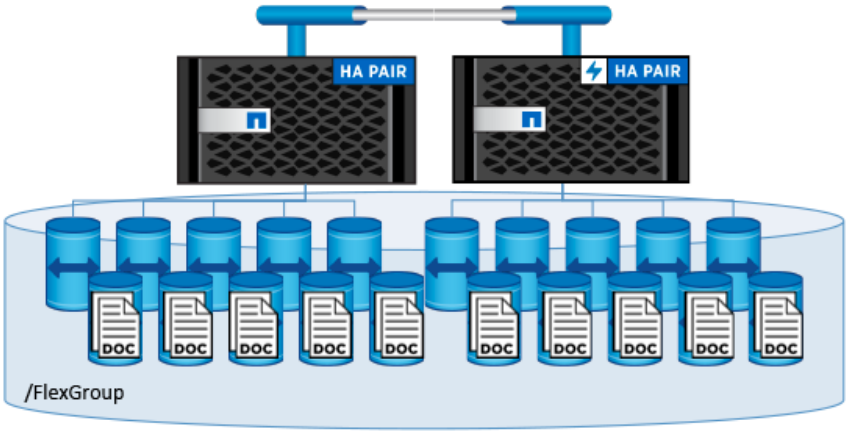
4.5 NetApp FlexGroup Volumes

The training dataset is usually a collection of a large number of files (potentially billions). 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 a large number of small files (potentially billions) and must read those files in parallel for sequential and random I/O.

A FlexGroup volume (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 massive capacity (multiple petabytes) and predictable low latency for high-metadata workloads.
- They support hundreds of billions of 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 NVIDIA GPU Cloud and Trident

NGC provides a catalog of fully integrated and performance engineered Docker images for DL that take full advantage of NVIDIA GPUs. These images include all necessary dependencies such as the CUDA Toolkit and DL libraries. These images are tested, tuned, and certified by NVIDIA for use on DGX-1 servers. Further, to enable portability of images that leverage GPUs, NVIDIA developed NVIDIA Container Runtime for Docker, which enables you to mount the user mode components of NVIDIA drivers and GPUs into the Docker container at launch.

Trident, from NetApp, is an open-source dynamic storage provisioner for Docker and Kubernetes. Combined with NGC and popular orchestrators such as Kubernetes or Docker Swarm, Trident enables you to seamlessly deploy your DL NGC container images onto NetApp storage, which provides an enterprise-grade experience for your 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 Cisco Nexus 3232C Network Switches

The Cisco Nexus 3232C switch (Figure 5) is a low-latency, dense, high-performance, power-efficient 100 Gb/s switch that is designed for the data center. This compact, 1 rack unit (1RU) model offers wire-rate layer 2 and layer 3 switching on all ports with a latency of 450ns. This switch is a member of the Cisco Nexus 3200 platform and runs the industry-leading Cisco NX-OS software operating system, providing you with comprehensive features and functions that are widely deployed. The Cisco Nexus 3232C is a Quad Small Form-Factor Pluggable (QSFP) switch with 32 QSFP28 ports. Each QSFP28 port can operate at 10, 25, 40, 50, and 100Gb/s, up to a maximum of 128 ports of 25Gb/s.

Figure 5) Cisco Nexus switches with NX-OS support for Converged Enhanced Ethernet standards and RoCE v1 and v2.

This solution as tested consumes only half of the available ports on each network switch. Each switch could support up to eight DGX-1 servers with additional storage access ports to provide more GPU power. For even larger implementations, the Cisco Nexus 7000 supports up to 192 ports of 100GbE per

switch. Alternatively, a leaf-spine topology could be implemented with multiple pairs of Nexus 3000 switches that are connected into a central spine switch.

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 by allowing network adapters to do a server-to-server data transfer between application memory by using zero-copy functionality without any OS or device driver involvement. 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 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 the low latency and the 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.

For guaranteed 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.

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

5 Technology Requirements

This section covers the hardware and software that was used in the validation of this solution. All the testing that is documented in section 7, Solution Verification, was performed with the hardware and the software are indicated here.

Note: The configuration that is verified in this reference architecture is based on lab equipment availability and not on the requirements or the limitations of the hardware that was tested.

5.1 Hardware Requirements

Table 1 lists the hardware components that were used to validate this solution. The hardware components that you use in any particular implementation of this solution might vary based on your requirements.

Table 1) Hardware requirements.

Hardware	Quantity
NVIDIA DGX-1 GPU servers	4
NetApp AFF A800 system	1 high-availability (HA) pair, includes 48x 1.92TB NVMe SSDs
Cisco Nexus 3232C network switches	2

5.2 Software Requirements

Table 2 lists the software components that are required to implement the solution. The software components that you use in any particular implementation of the solution might vary based on your requirements.

Table 2) Software requirements.

Software	Version
NetApp ONTAP	9.4
Cisco NX-OS switch firmware	7.0(3)I6(1)
NVIDIA DGX OS	Ubuntu 16.04 LTS
Docker container platform	18.03.1-ce [9ee9f40]
Container version	netapp_1.7.0.2 based on nvcr.io/nvidia/tensorflow:18.04-py2
Machine learning framework	TensorFlow 1.7.0
Horovod	0.11.3
OpenMPI	3.1.0
Benchmark software	TensorFlow benchmarks [1b1ca8a]

6 Solution Architecture

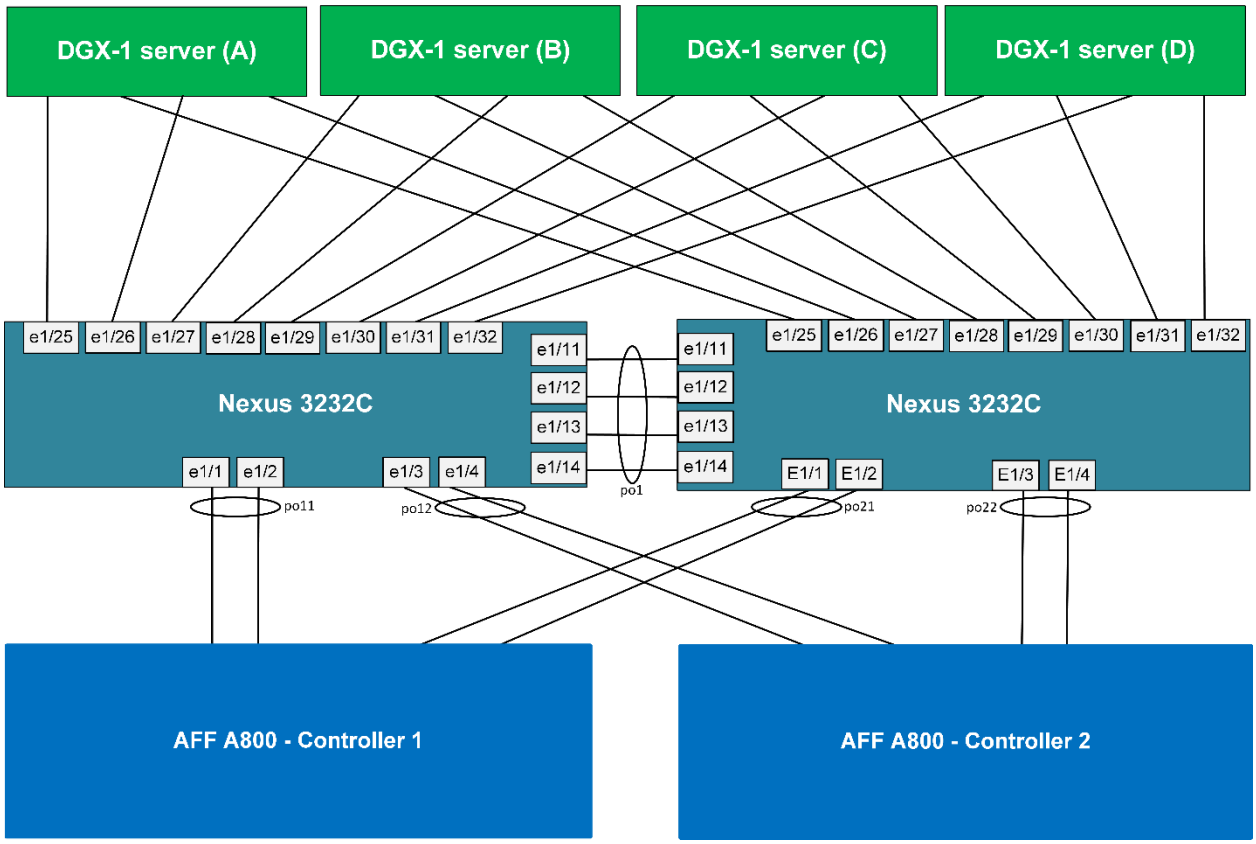
This architecture has been verified to meet the requirements for running DL workloads. This verification enables data scientists to deploy DL frameworks and applications on a pre-validated infrastructure, thereby helping to eliminate risks and allowing businesses to focus on gaining valuable insights 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

For this solution, RoCE is used in place of IB to provide the high-bandwidth, low-latency connectivity that is required for communication between DGX-1 servers. Cisco Nexus switches support RoCE by implementing PFC, which allows users to prioritize RoCE traffic over traditional IP traffic on a shared link and allows the 100GbE links to be used for both RoCE and IP at the same time.

This architecture uses a pair of Cisco Nexus 3232C 100GbE switches for the primary inter-cluster and storage access network. These switches are connected to each other with four 100Gb network 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 Nexus switches with a pair of active-passive bonds, and, to provide link-layer redundancy, each storage controller is connected to each Nexus switch with a two-port LACP port channel. Figure 6 shows the network switch-port configuration.

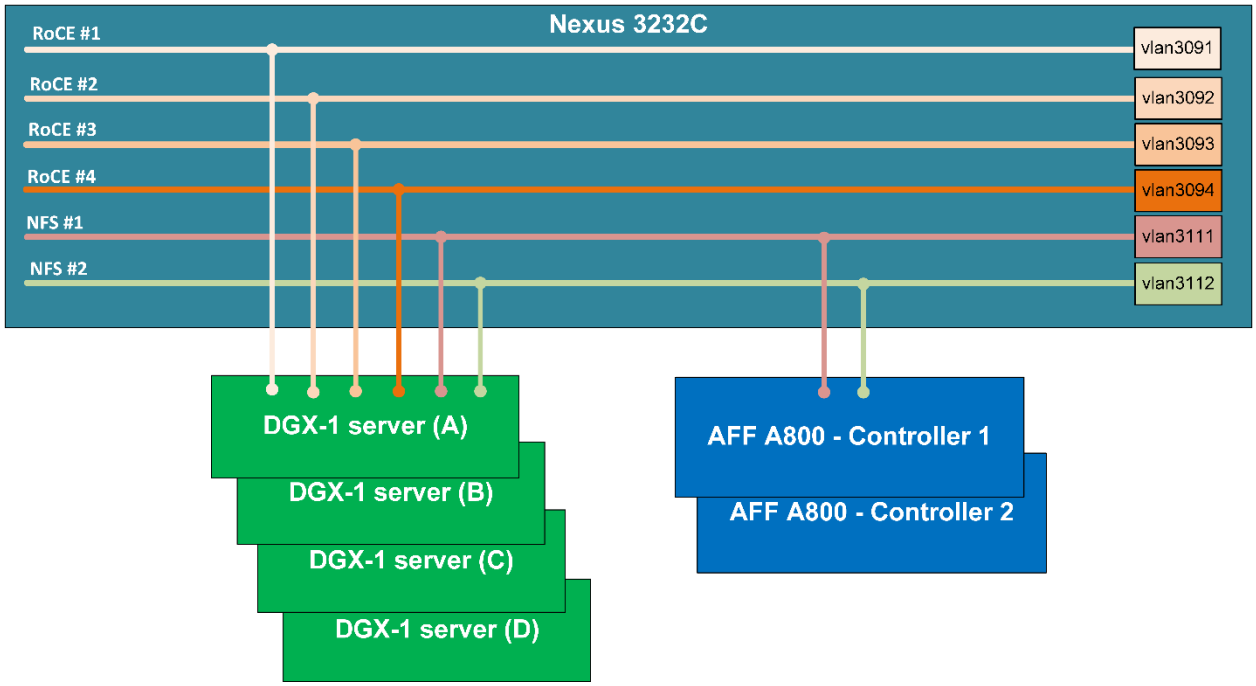
Figure 6) Network switch port configuration.



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 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 class, and the NFS VLANs are assigned to the default best-effort class. All VLANs are configured for jumbo frames with a maximum transmission unit (MTU) size of 9000.

The switch-ports for DGX-1 servers 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 7 shows the VLAN connectivity for the DGX-1 server and storage system ports.

Figure 7) 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.

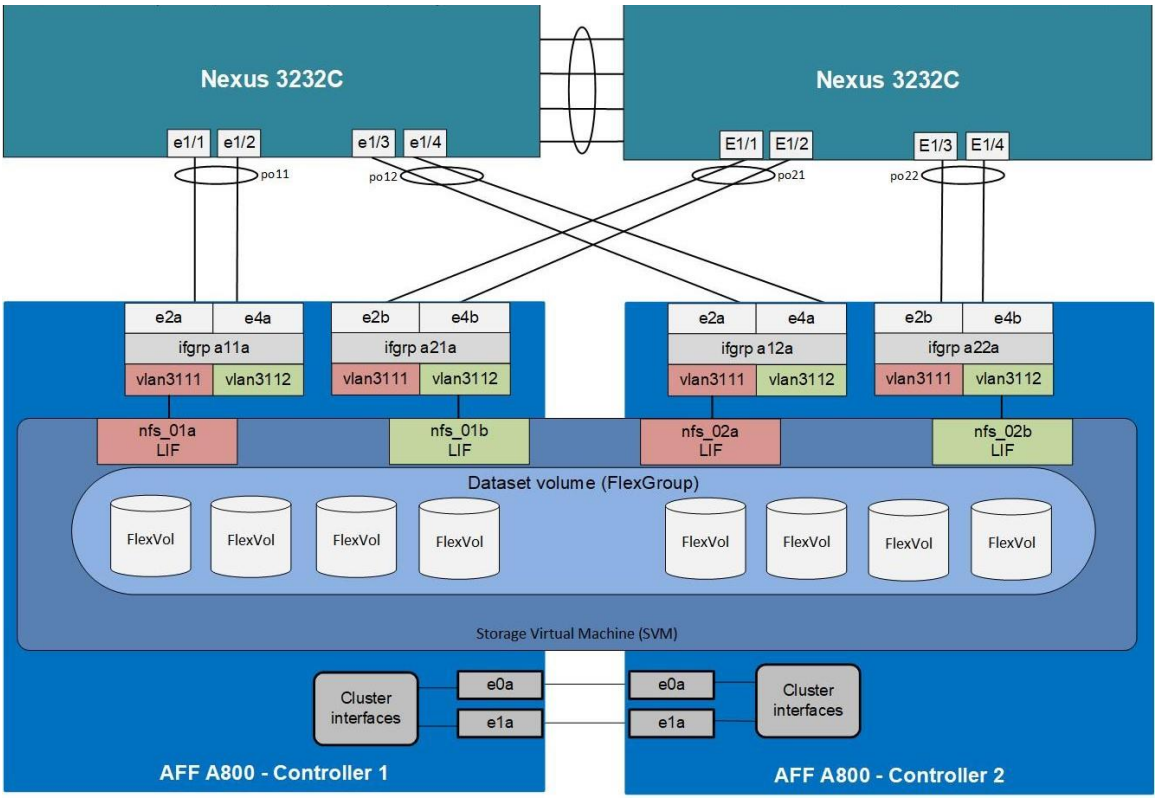
PFC is then enabled on each DGX-1 port, which enables the switch port to send pause frames for specific classes of service to eliminate congestion at the switch. By using ETS to allocate 95% of the bandwidth to RoCE traffic in the event of congestion, this configuration allows dynamic resource allocation between RoCE and NFS traffic while providing priority to node-to-node communication. You can also modify the bandwidth allocation 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 8 shows the storage system configuration. Each controller is configured with a two-port LACP interface group (ifgrp in Figure 8) to each switch. These interface groups provide up to 200Gb/s 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 8) Storage system configuration.

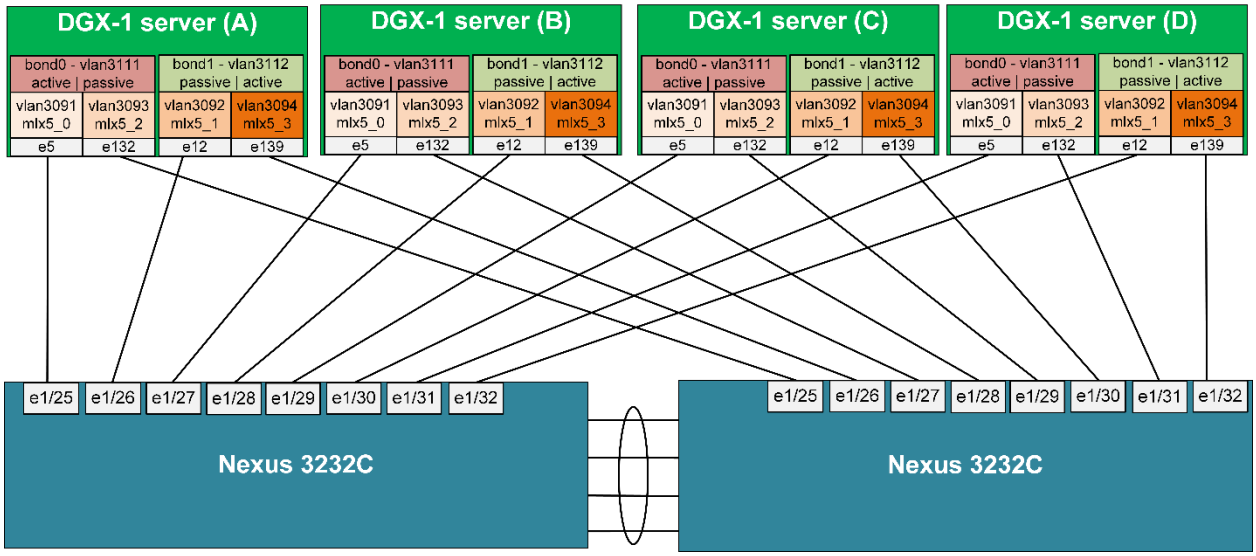


For logical storage provisioning, the solution uses a FlexGroup volume to provide a single pool of storage that is distributed across the nodes in the storage cluster. Each controller hosts an aggregate of 46 disk partitions, with both controllers sharing every disk. When the FlexGroup is deployed on the data SVM, a number of FlexVol volumes are provisioned on each aggregate and then are 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 simply by adding more nodes to the storage cluster.

6.3 Host Configuration

For network connectivity, each DGX-1 is provisioned with four Mellanox ConnectX4 single-port network interface cards. These cards operate at up to 100GbE speeds and support RoCE, providing 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 9 shows the network port and VLAN configuration of the DGX-1 hosts.

Figure 9) 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 provides 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 server to the storage system. Each DGX-1 server 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, providing completely independent connections from the server to the storage system. Although a single NFS mount is capable of delivering 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 that we performed to validate the operation and performance of this solution. We performed all the tests that are described in this section 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 a number of 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 with NFSv3 by up to four DGX-1 servers, 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 with varying degrees of compute and storage complexities were used to demonstrate training rates:

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

Each of these models was tested with various hardware and software configurations to study the effects of each option on performance:

- We tested each model with both synthetic data and the ImageNet reference dataset. Further testing with additional GPUs both internal to the DGX-1 and across multiple DGX-1 servers, assisted in the evaluation of scalability for the compute cluster and the evaluation of storage access performance.
- We used ImageNet data with distortion disabled to reduce the overhead of CPU processing before copying data into GPU memory.
- We tested each model by using Tensor cores and CUDA cores to demonstrate the performance improvements that the Tensor cores provide.
- Increasing the GPU performance also had the effect of increasing storage access requirements and demonstrated the AFF A800 system's ability to easily support those requirements.
- We tested each DL model with various batch sizes. Increasing the batch size has several effects on the system that ultimately result in higher overall training rates, lower inter-GPU communication requirements, and higher storage bandwidth requirements. We tested the following batch sizes with each model:
  - 64, 128, and 256 for ResNet-50
  - 64 and 128 for all other models
- Each model was tested with one, two, and four DGX-1 servers to demonstrate the scalability of each model across multiple GPUs that use RoCE as the interconnect (through Horovod).
- Inference was run by using all the models with the largest batch sizes (256 for ResNet-50 and 128 for all other models), with 32 GPUs (Tensor cores and CUDA cores), and with ImageNet dataset.
- All performance metrics were gathered after at least two epochs. We observed slightly better performance results when we ran training over multiple epochs. Each test was run five times and the mean of the performance metrics that we observed are reported.

7.2 Validation Test Results

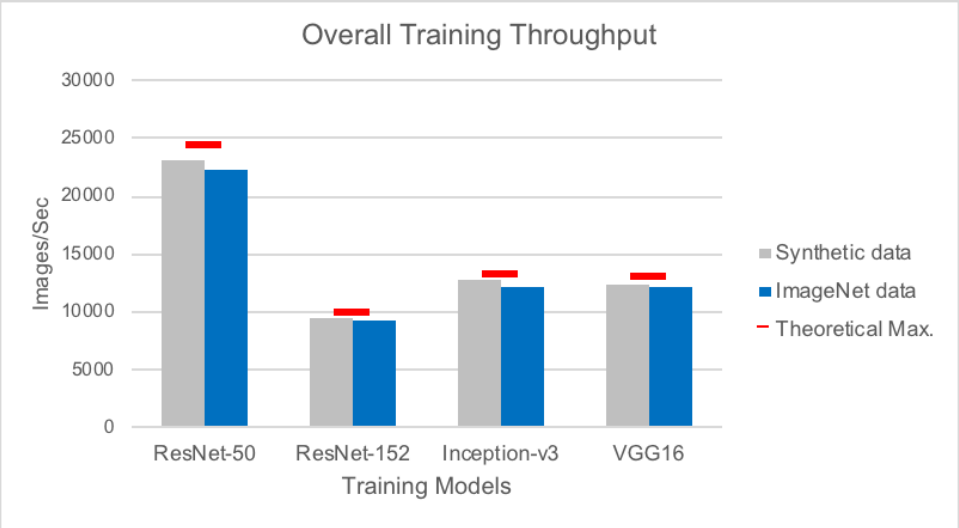
As described previously, we conducted various tests to assess the general operation and performance of this solution. This section contains highlights of the compute and storage performance data that was collected during those tests. Complete detailed test results are in the appendix. Note the following details about the data that is presented in the next subsections of this report:

- Model training performance is measured as images per second.
- Storage performance is measured by using throughput (MB/s) and latency (µs). The storage system CPU was also captured to evaluate the remaining performance capacity on the storage system.
- Each system was tested with multiple batch sizes. Larger batch sizes increase the overall training throughput. Only the largest batch size that was tested for each model is shown here. Data for each batch size that was tested is available in the appendix:
  - ResNet-50 tests used a batch size of 256.
  - ResNet-152, Inception-v3, and VGG16 tests used a batch size of 128.

Overall Training Throughput

Figure 10 shows the maximum number of training images per second that was 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 32 GPUs by using ImageNet data and synthetic data for baseline comparison. It also shows the theoretical maximum that is achievable, in which all GPUs train synthetic data independently without updating parameters with each other. As shown in Figure 10 graph, our achieved throughput for ImageNet data is very close to the throughput for synthetic data.

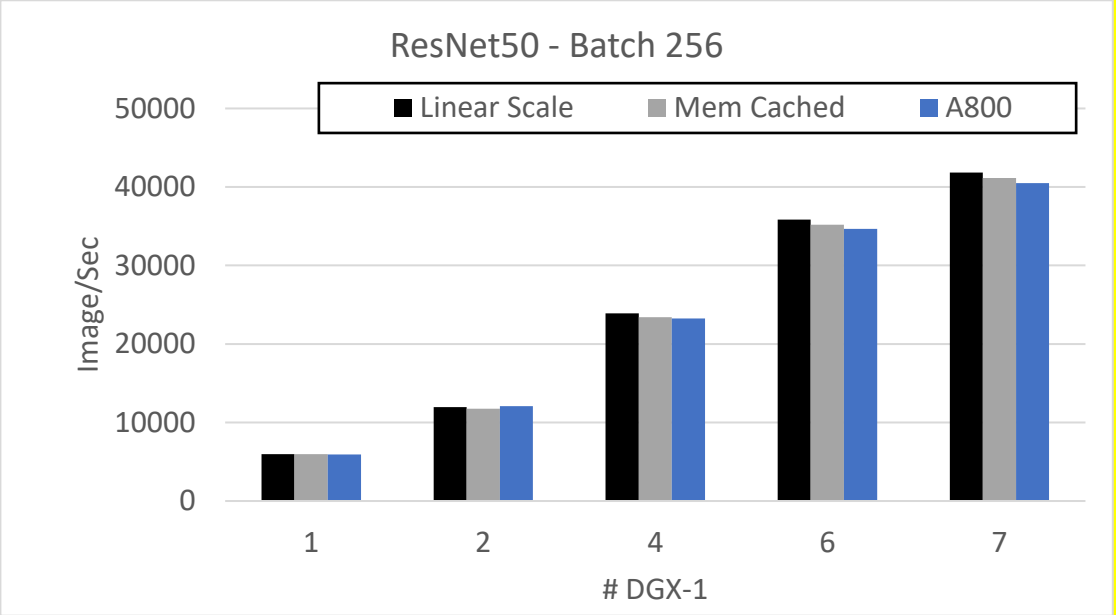
Figure 10) Training throughput for all models.



Large cluster test results

Additional benchmark testing was performed with a larger configuration of 7 DGX-1 servers to demonstrate the performance capabilities of the ONTAP AI infrastructure in general. The graph below shows the training results for the ResNet50 model comparing performance between synthetic data, data cached in DGX-1 server memory, and data read directly from the A800 storage system. The ImageNet data was duplicated 10x times on the A800 to create a dataset larger than DGX-1 memory caching can support.

Figure 11) Training results with up to 7 DGX-1



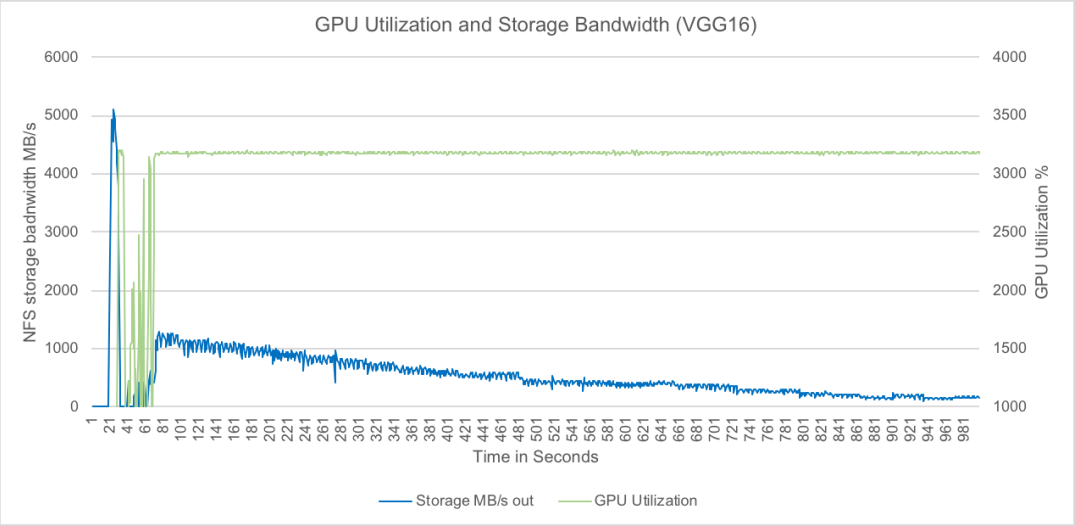
Note that performance from the A800 storage system is comparable to when data is read from memory, indicating the ability to perform training against datasets much larger than DGX-1 system memory would allow with very little performance impact.

GPU Workload Performance

The next set of data demonstrates the ability of the storage system to meet the requirements of the DGX-1 server under a full load. Figure 12 shows the GPU utilization of the DGX-1 servers and the storage bandwidth that was generated when running each model by using 32 GPUs. As seen in the graph, the storage bandwidth starts off very high as the initial data is read from storage into the TensorFlow pipeline cache, and then it drops gradually as a larger portion of the dataset becomes resident in DGX-1 local memory over time.

After all the data is in the local memory, storage access drops to almost nothing. The DGX-1 GPUs begin processing data almost immediately, and GPU utilization remains consistent throughout the test run. This graph shows the results for the VGG16 model with a batch size of 128, which produced the highest level of GPU utilization in our testing. Graphs for the other models are available in the Appendix. Note that the GPU utilization scale is the sum utilization of all GPUs, so, in this case with 32 GPUs tested, the maximum possible utilization is 3200%.

Figure 12) GPU utilization and storage bandwidth (VGG16).



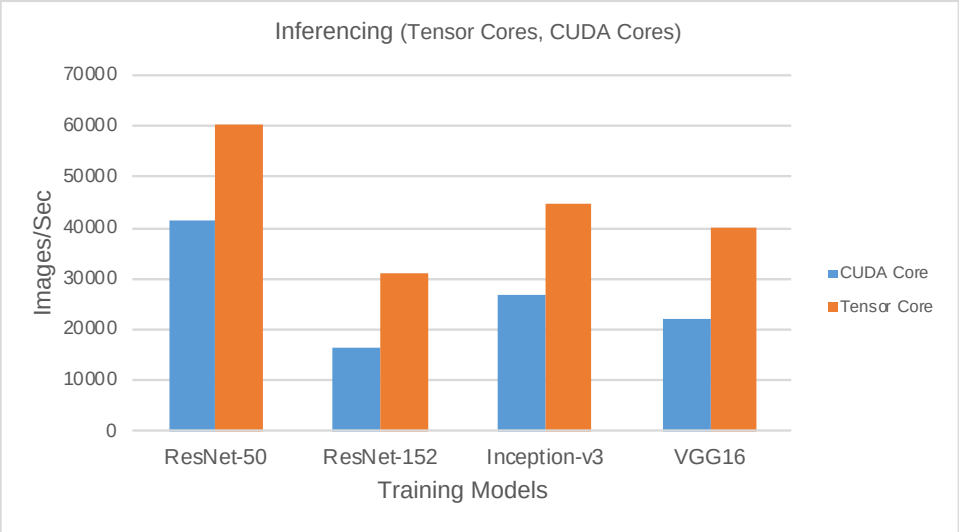
As Figure 12 shows, the GPU utilization remains above 95% for all 32 GPUs and also remains consistent regardless of how much data is coming from the storage system. The storage system delivers an initial 5GB/s of data, then drops from around 2GB/s to almost nothing over the remainder of the training epoch. This result demonstrates that storage access is not a bottleneck to GPU performance with this workload. With a larger data set that exceeds local memory capacity, storage access performance would remain at the steady-state throughput rate until later in the training epoch. In addition, Figure 12 compares the GPU utilization as a function of storage bandwidth. It does not capture the time that is required for the entire training phase because the storage bandwidth gradually drops close to zero as the training phase progresses.

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 data set, 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 13 shows the number of images that can be processed per second during inferencing. This test compares the throughput that was achieved with 32 GPUs that used ImageNet data on each of the models that were tested by using Tensor cores and CUDA cores. With the power of NetApp ONTAP AI, Tensor cores can be used to classify a significant number of images instantaneously.

Figure 13) Inference for all models (Tensor Cores and CUDA Cores).

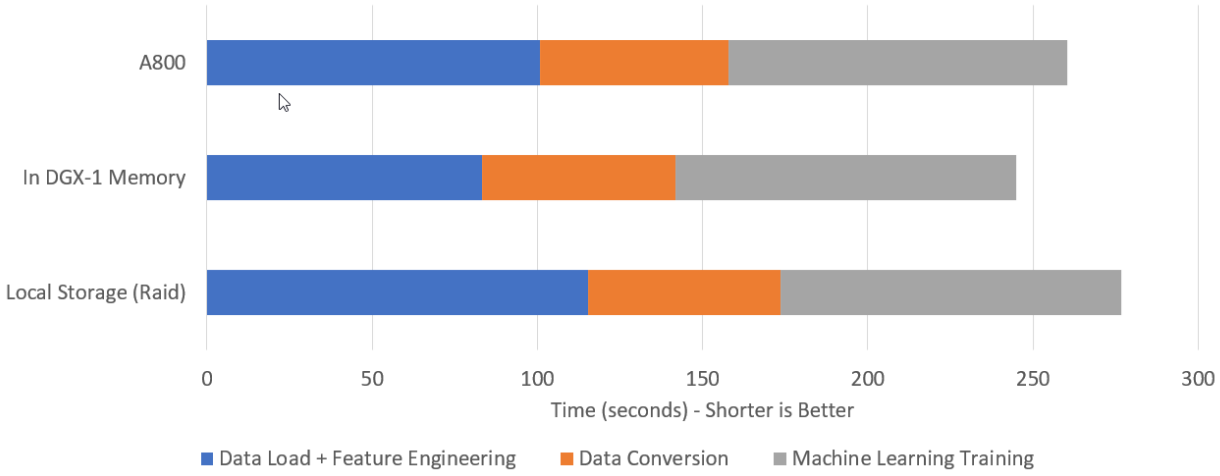


RAPIDS

RAPIDS is a set of libraries designed to integrate data preparation and training into a single GPU-accelerated workflow. RAPIDS uses common data structures and programming interfaces to allow developers to accelerate analytics and deep learning data preparation and model training. To validate the performance of a RAPIDS workflow, mortgage data from (https://rapidsai.github.io/demos/datasets/mortgage-data) was loaded into GPU memory via the RAPIDS CSV reader. The loaded data was then converted to train a gradient boosted decision tree model on the GPU using XGBoost which is one of the RAPIDS libraries. Detailed information on RAPDIS can be found at the RAPIDS website.

The graph below shows the performance of RAPIDS when data is sourced from local RAID storage, system memory, and the A800 storage system. The data load and feature engineering portion of the workflow duration are directly impacted by the performance of the underlying storage. In this case the performance of the workflow from the A800 storage system is faster than the local RAID storage on the DGX-1 server.

Figure 14) End-to-End pipeline performance with RAPIDS on 1 DGX-1

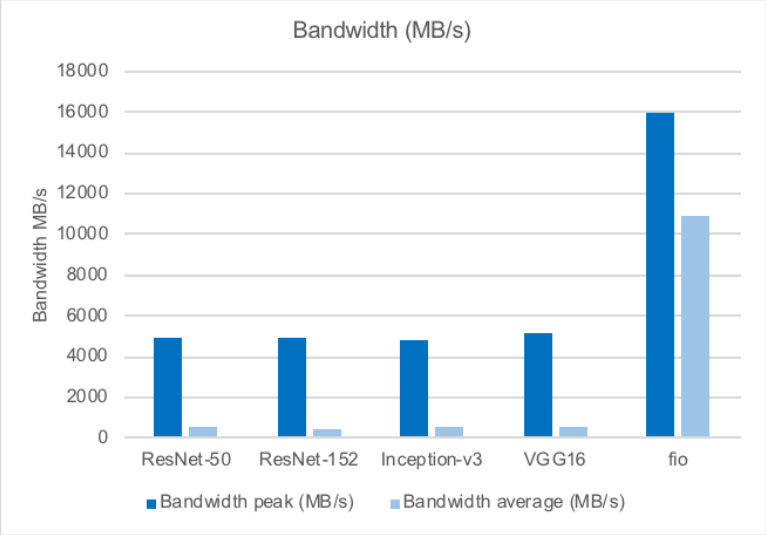


AFF A800 System Performance with AI Training Workloads

Storage bandwidth, latency, and CPU headroom were captured to demonstrate the storage system performance with each of the tested models. Figure 15 through Figure 17 show the storage system metrics for each model when tested with real data. These storage-focused tests were performed with higher batch sizes to increase the storage workload and to demonstrate the worst-case scenario.

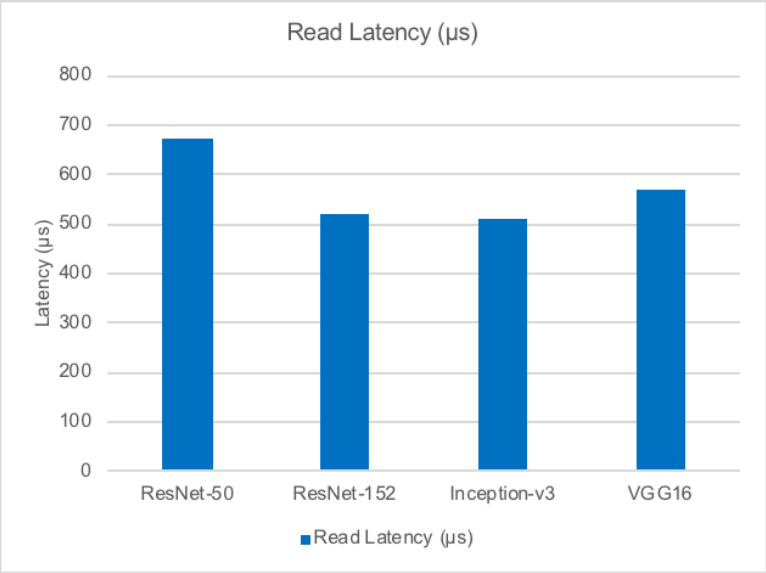
Note that in each metric, the total workload that is generated by each model with 32 GPUs is well within the performance envelope of the AFF A800 system. To provide a frame of reference for the training workload, an artificial workload was generated by using flexible I/O (fio) with a 64K sequential read I/O profile. To achieve the maximum possible throughput with the limited number of DGX-1 servers that were available, additional NFS mounts and multiple fio jobs were used on each server. As seen in Figure 15, throughput for the workload generated with fio peaked at over 15GB/s.

Figure 15) Storage bandwidth for all models.



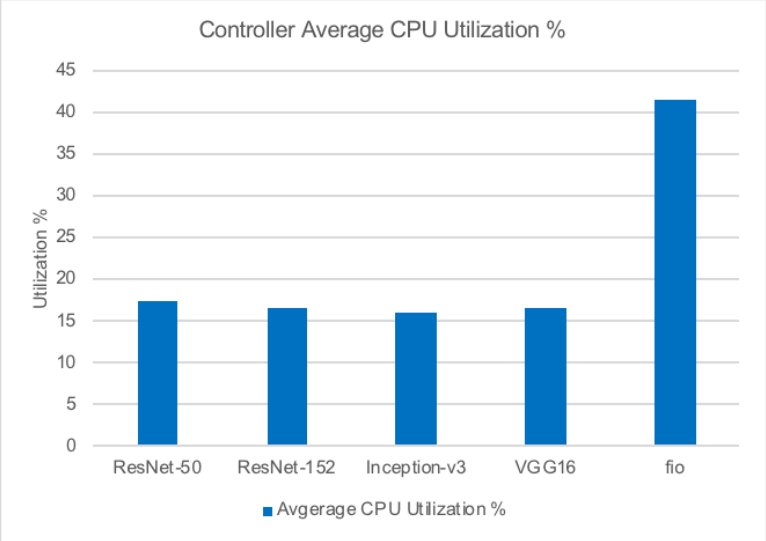
Storage latency for training with all models remained under 1ms as shown in Figure 16.

Figure 16) Storage latency for all models.



As seen in Figure 17, CPU utilization for all models was relatively low, and even with the much higher fio workload controller CPU utilization remained below 50%.

Figure 17) Storage CPU utilization for all models.



Note: A NetApp AFF A800 HA-pair has been proven to support up to 25GB/s under 1ms latency for NAS workloads.

7.3 Solution Sizing Guidance

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

As is demonstrated in this validation, the AFF A800 system easily supports the DL training workload that is generated by four DGX-1 servers, with approximately 70% headroom remaining on the HA pair. Therefore, the AFF A800 system can support additional DGX-1 servers. 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 and, with the FlexGroup technology that is validated in this solution, can provide over 20PB in a single volume. The dataset that we used in this validation was relatively small. However, ONTAP 9 can scale to impressive capacity with linear performance scalability as each HA pair delivers performance comparable to the level that is verified in this document.

For smaller DGX-1 clusters, an AFF A220 or AFF A300 system provides sufficient performance at a lower price point. Because ONTAP 9 supports mixed-model clusters, you can start with a smaller initial footprint and add more or larger storage systems into the cluster as your capacity and performance requirements grow.

From a network perspective, this architecture as verified consumes only 16 of the 32 available ports on each Nexus 3232C switch. Each switch can support up to eight DGX-1 servers with additional storage access ports to significantly increase compute power without additional networking. For larger implementations, the Cisco Nexus 7000 supports up to 192 wire-rate 100GbE ports per switch. Alternatively, you can implement a leaf-spine topology with multiple pairs of Nexus 3000 switches that are connected into a central spine switch.

Based on the validation testing that was performed with this AI training workload, each DGX-1 requires roughly 2GB/s storage throughput. Given that the AFF A800 system has the proven capability of 25GB/s of throughput with a similar workload generated by other means, this architecture could support nine or more DGX-1 servers per AFF A800 HA pair.

8 Conclusion

The DGX-1 server 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 Cisco Nexus switches, you can implement this verified architecture at almost any scale that you need, from a single DGX-1 paired to an AFF A220 system up to potentially 96 DGX-1 servers 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

We gratefully acknowledge the contributions that were made to this NetApp Verified Architecture by our esteemed colleagues from NVIDIA, Darrin Johnson, Tony Paikeday, Robert Sohigian, and James Mauro. We could not have completed this study without the support and guidance of our key NetApp team members, Robert Franz and Kesari Mishra.

Our sincere appreciation and thanks go to all these 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, see the following resources:

- NVIDIA DGX-1 servers
  - NVIDIA DGX-1 servers
    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
- Cisco Nexus networking

  The following links provide more information about Cisco Nexus 3232C series switches:
  - Cisco Nexus 3232C series switches
    https://www.cisco.com/c/en/us/products/switches/nexus-3232c-switch/index.html
  - Cisco Nexus 3232C configuration guide
    https://www.cisco.com/c/en/us/support/switches/nexus-3000-series-switches/products-installation-and-configuration-guides-list.html
  - Cisco Nexus 3232C command line reference
    https://www.cisco.com/c/en/us/support/switches/nexus-3000-series-switches/products-command-reference-list.html
- 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

Appendix

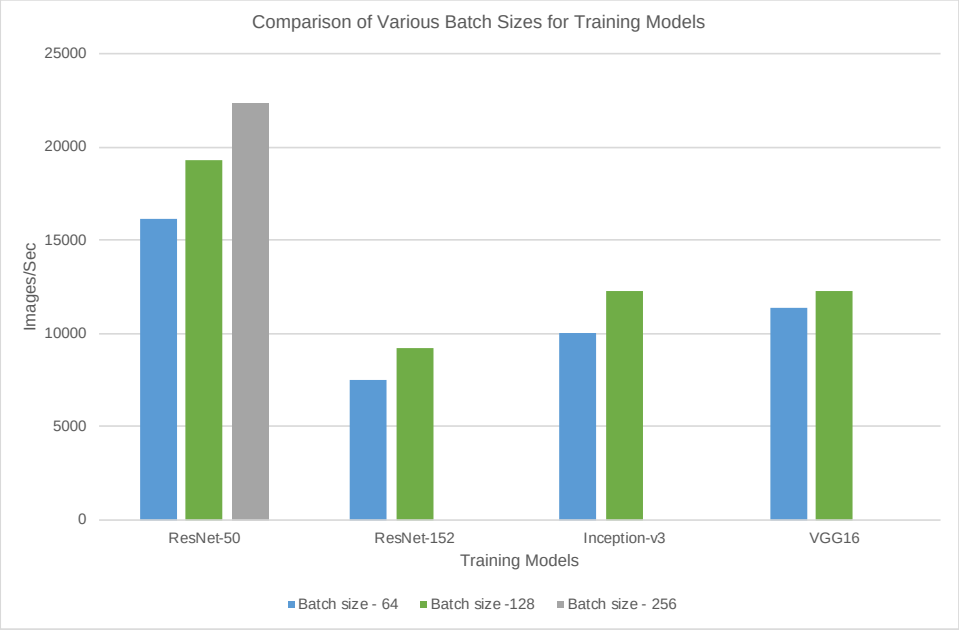
This section contains additional results for the tests that were performed by using this architecture.

Training Rates for Different Batch Sizes for Each Model

Figure 18 shows a comparison of the various batch sizes for the different training models that used the following components:

- Number of GPUs: 32 (4 DGX-1 servers)
- Cores: Tensor cores
- Batch sizes: 64,128, and 256 for ResNet-50; 64 and128 for other models

Figure 18) Comparison of various batch sizes for training models.



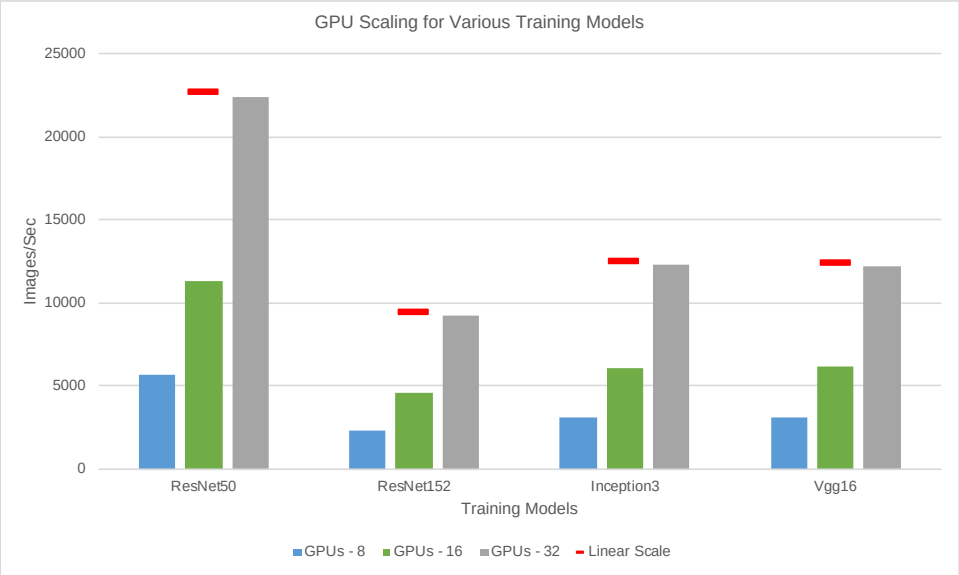
Conclusion: Training throughput performance increases as the batch size increases to 256 or 128.

Comparison of GPU Scaling for Each Model

Figure 19 shows the GPU scaling for the different training models that used the following components:

- Number of GPUs: 8 (1 DGX-1 server), 16 (2 DGX-1 servers), and 32 (4 DGX-1 servers)
- Cores: Tensor cores
- Batch sizes: 256 for ResNet-50 and128 for other models

Figure 19) GPU scaling for various training models.



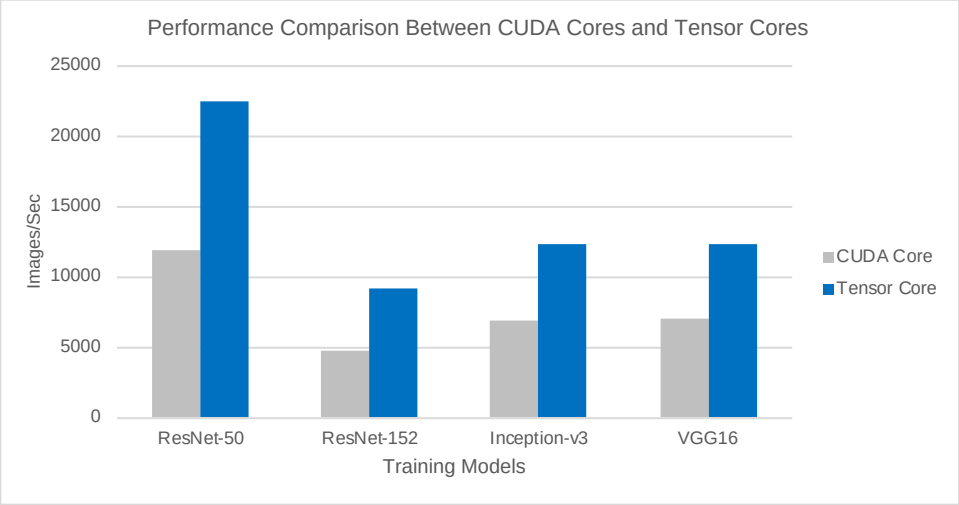
- Conclusion: Linear GPU scaling is observed across all the training models.

Comparison of Tensor Cores and CUDA Cores

Figure 20 shows a performance comparison between CUDA cores and Tensor cores that used the following components:

- Number of GPUs: 32 (4 DGX-1 servers)
- Cores: Tensor cores and CUDA cores
- Batch sizes: 256 for ResNet-50 and128 for other models

Figure 20) Performance comparison between CUDA cores and Tensor cores.



Conclusion: Tensor cores yield better performance than CUDA cores do.

GPU Workload for All Models

Figure 21 through Figure 21 show the GPU utilization and bandwidth for ResNet-50, ResNet-152, and Inception-v3, respectively, that used the following components:

- Number of GPUs: 32 (4 DGX-1 servers)
- Cores: Tensor cores
- Batch sizes: 256 for ResNet-50 and128 for other models

Figure 21) GPU utilization and storage bandwidth for ResNet-50.

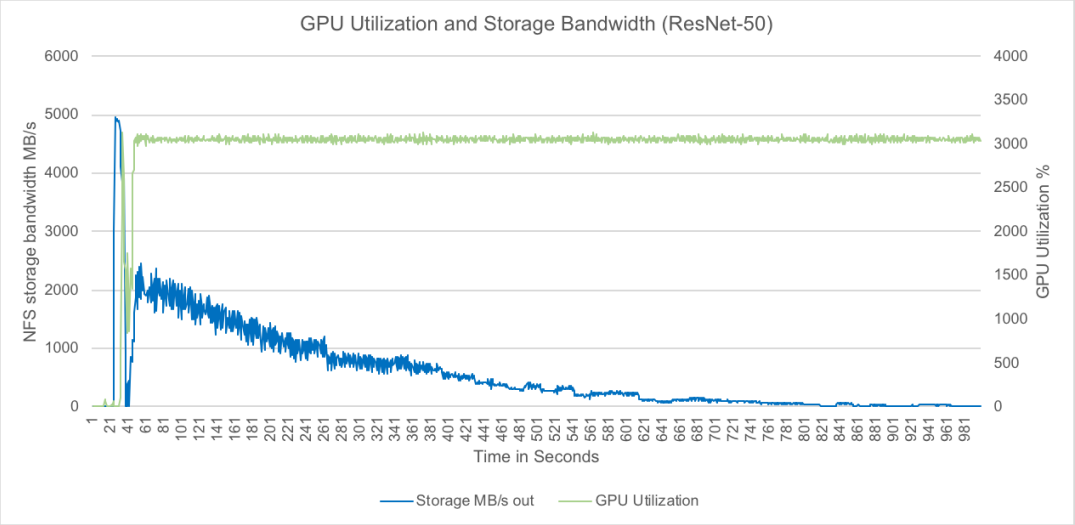


Figure 22 shows the GPU utilization and bandwidth for ResNet-152.

Figure 22) GPU utilization and storage bandwidth for ResNet-152.

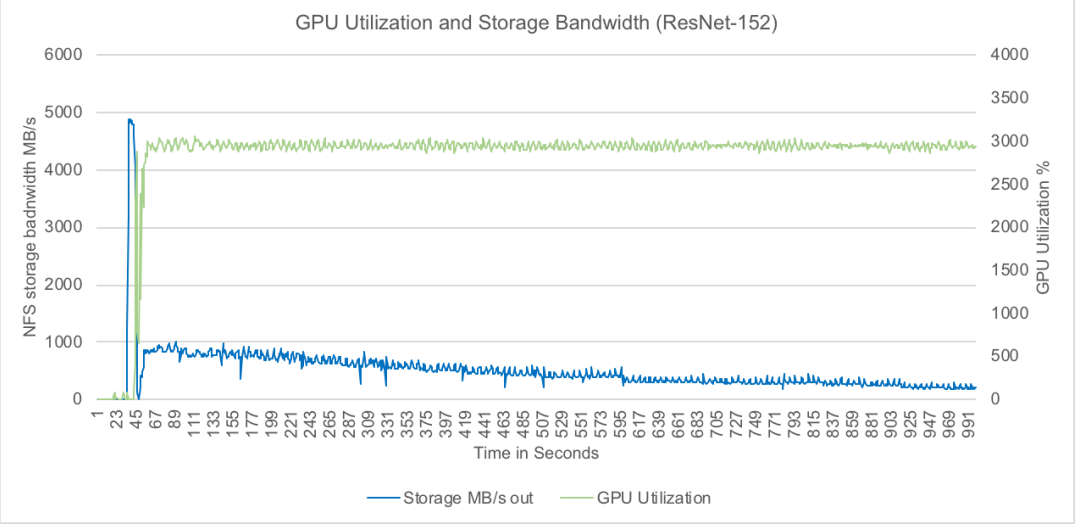
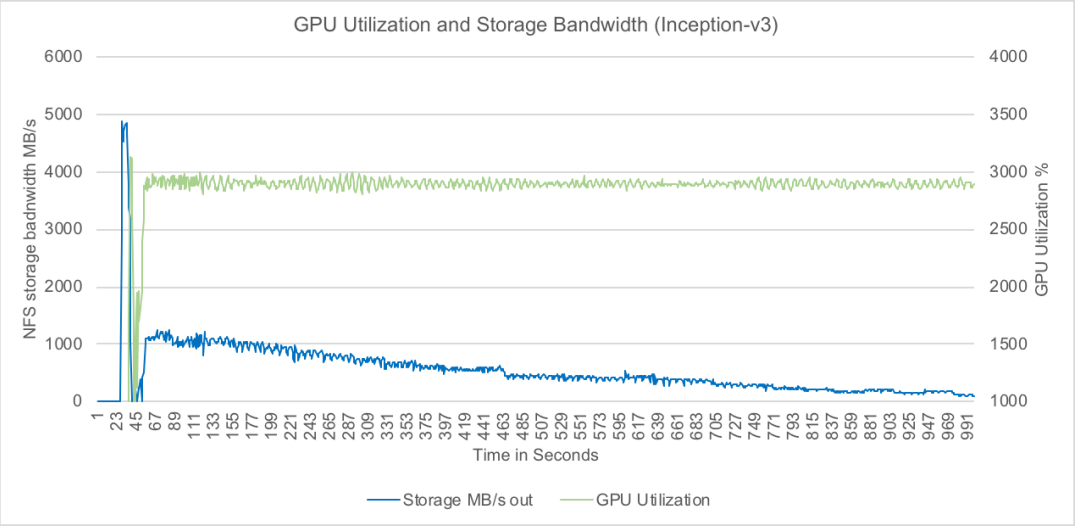


Figure 23 shows the GPU utilization and bandwidth for Inception-v3.

Figure 23) GPU utilization and storage bandwidth for Inception-v3.



Large cluster results for All Models

Figure 24 through Figure 26 show the overall training results for ResNet-152, Inception-v3, and VGG16 respectively, that used the following components:

- Number of GPUs: up to 56 (7 DGX-1 servers)
- Cores: Tensor cores
- Batch sizes: 256 for ResNet-50 and128 for all models
- Synthetic data vs. real data in memory vs. 10x data on A800

Figure 24) Large cluster results- ResNet152

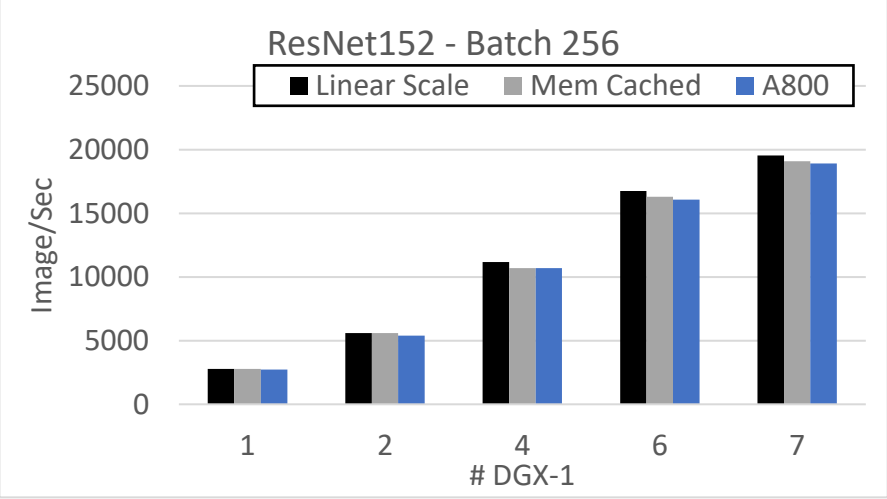


Figure 25) Large cluster results- Inception3

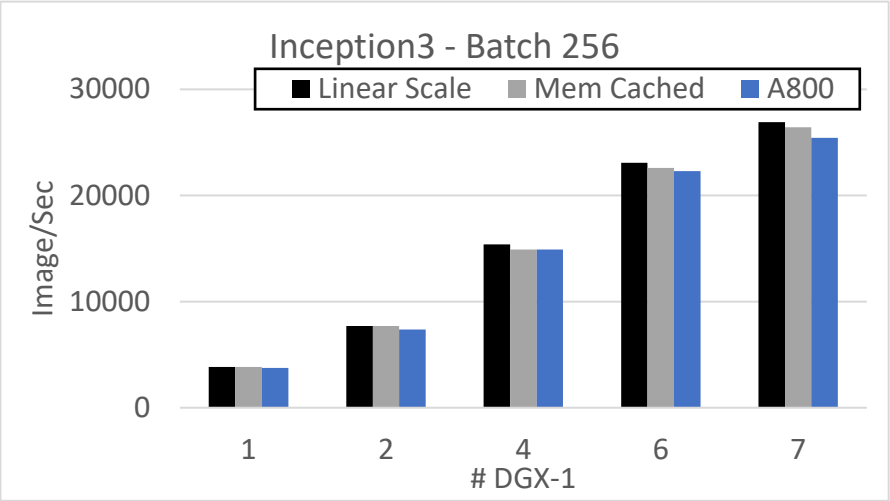
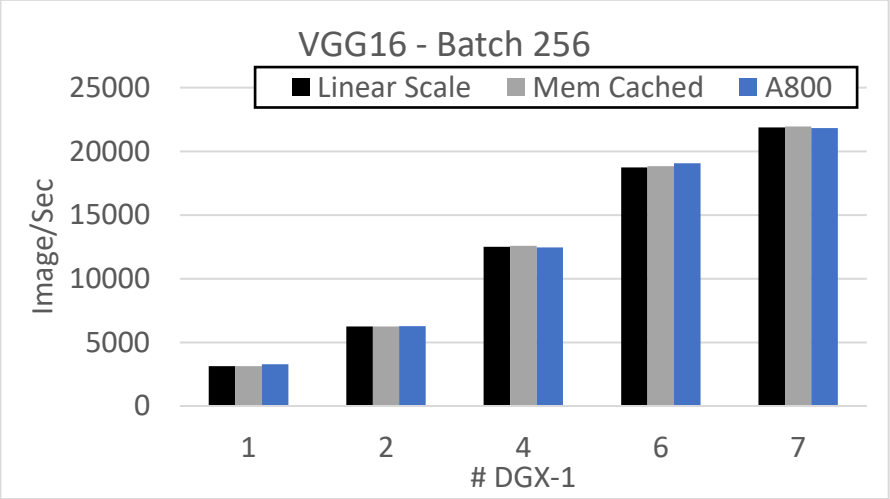


Figure 26) Large cluster results- Vgg16



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

BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

Data contained herein pertains to a commercial item (as defined in FAR 2.101) and is proprietary to NetApp, Inc. The U.S. Government has a non-exclusive, non-transferrable, non-sublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b).

Trademark Information

NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.