Technical Whitepaper

Scalable AI Infrastructure

Designing For Real-World Deep Learning Use Cases

Sundar Ranganathan, NetApp
Santosh Rao, NetApp

June 2018 | WP-7267

In partnership with

NVIDIA

Executive Summary

Deep learning (DL) is enabling rapid advances in some of the grandest challenges in science today: in medicine for better ways to cure cancer, in physics for particle detection and classification, in autonomous vehicles for achieving level-5 autonomy. All have a common element – data. DL is fundamentally driven by data.

Graphics processing units (GPUs) enable new insights that were not previously possible. To meet the rigorous demands of the GPUs in a DL application, the storage systems must be capable of constantly feeding data to the GPUs at low latencies and high throughput, whether that data is text, images, audio, or videos.

As organizations progress from small-scale DL deployments to production, it's crucial to design an infrastructure that can deliver high performance and allow independent and seamless scaling. NVIDIA's leadership in GPUs, along with NetApp's innovation in all flash storage systems, forms a unique solution to accelerate DL applications.

NetApp

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

When IT leaders talk about the value of data in their organizations, the most frequently cited examples are deep learning (DL), and more broadly artificial intelligence (AI). DL is the engine that already powers fraud detection in finance, predictive maintenance in manufacturing, speech recognition in customer support bots, and various levels of autonomy in vehicles.

Moving forward, data and DL applications will be leveraged to improve productivity, identify essential patterns, and design disruptive services, solutions, and products in every imaginable industry. Market research firm IDC forecasts that spending on AI for software, services, and hardware will grow from $12 billion in 2017 to $57.6 billion by 2021[1].

DL systems leverage algorithms that are significantly improved by increasing the size of the neural network, as well as the amount and quality of data used to train the models. Depending on the particular application, DL models work with large amounts of different types of data, such as text, images, audio, video, or time series data. That's why DL applications require a high-performing infrastructure, such as the reference architecture presented here.

Deploying DL workloads on an infrastructure that is purpose-built to handle the unique and growing demands of DL while providing the fastest time to training is essential to the success of an organizational DL strategy.

This infrastructure should meet the following high-level infrastructure requirements:

- Large compute power to train models in less time
- High-performance storage to handle large datasets
- Seamlessly and independently scale compute and storage
- Handle varying types of data traffic
- Optimize costs

An infrastructure that supports flexibility in scaling up and out is vital. As compute and storage scale, performance requirements may evolve, demanding the ability to dynamically alter the ratios of compute to storage systems with zero downtime.

This paper introduces a scalable infrastructure which includes the NVIDIA® DGX-1™ server built on the new NVIDIA® Tesla® V100 graphic processing unit (GPU) platform[2] and the new NetApp® A800™ all flash storage system.

2 Compute – NVIDIA DGX-1

2.1 GPU Accelerated Computing

The computations performed in DL algorithms involve an immense volume of matrix multiplications, run in parallel. The highly parallelized architecture of modern GPUs makes them substantially more efficient than general-purpose central processing units (CPUs) for applications where data processing is done in parallel. Advances in individual and clustered GPU architectures have made them the preferred platform for workloads like high-performance computing, DL, and analytics.

2.2 Purpose-Built for Deep Learning

Assembling and integrating off-the-shelf hardware and software components for DL from multiple vendors can result in increased complexity and deployment times, resulting in valuable data science resources spending considerable effort on systems integration work.

[1] Source: IDC, Worldwide Storage for Cognitive/AI Workloads Forecast, 2018–2022
[2] Powered by the NVIDIA Volta architecture

Once the solution is deployed, many organizations find that they spend excessive cycles on tuning and refining their software stack as their models evolve. Recognizing this, NVIDIA created the DGX-1 server platform, a fully-integrated hardware and software turnkey system that is purpose-built for DL workflows.

Each DGX-1 server is powered by eight Tesla V100 GPUs, configured in a hybrid cube-mesh topology using NVIDIA NVLink™ providing an ultra-high bandwidth, low latency fabric for inter-GPU communications essential to multi-GPU training, eliminating the bottleneck associated with PCIe based interconnect. The DGX-1 server is also equipped with low-latency, high-bandwidth network interconnects for multi-node clustering over RDMA-capable fabrics.

2.3 Pre-Optimized, Enterprise-Grade

The DGX-1 server leverages GPU-optimized software containers from NVIDIA GPU Cloud (NGC), including containers for all of the most popular DL frameworks. The NGC deep learning containers are pre-optimized at every layer, including drivers, libraries and communications primitives, and deliver maximum performance for NVIDA GPUs. These pre-integrated containers insulate users from the constant churn typically associated with today's popular open source DL frameworks, thus providing teams a stable, QA tested stack on which to build enterprise-grade DL applications.

This fully-integrated hardware and software solution, backed by NVIDIA expertise, accelerates application DL deployments, reduces training time from weeks to days or hours, and increases data scientists' productivity, enabling them to spend more time in experimentation rather than systems integration and IT support.

3 Storage – NetApp AFF

As GPUs get faster and datasets increase in size and complexity, using a state-of-the-art storage system is essential to eliminate bottlenecks and maximize system performance. DL applications demand a storage solution that is designed to handle massively parallel DL workloads that require a high degree of concurrency on I/O processing to avoid stalling the GPUs as they wait for data.

Data traffic in many DL applications spans the entire data pipeline, from edge to core to cloud. Designing a storage architecture demands a holistic data management approach, from data ingest and/or edge analytics to data prep and training in the core data center to archiving in the cloud. It's crucial to understand performance requirements, the characteristics of diverse datasets, and the data services needed.

For DL workflows, an ideal storage solution needs to excel in the following high-level requirements.

3.1 High Performance

The bottlenecks in a DL infrastructure most commonly occur during the training phase, when high I/O bandwidth with massive I/O parallelism is required to ensure sustained high GPU utilization. This translates to the ability of the storage architecture to deliver high-throughput performance while maintaining a low-latency profile, which in turn translates to support for high-speed network fabrics.

A single NetApp A800 system supports throughput of 25GB/s for sequential reads and 1 million IOPS for small random reads at sub-500μs latencies[3]. In addition, what sets A800 apart is its 100GbE[4] network support, which accelerates data movement and also fosters balance in the overall training system, because the DGX-1 supports 100GbE RDMA for cluster interconnect. The NetApp A700s system supports multiple 40GbE links to deliver a maximum throughput of 18GB/s.

[3] https://blog.netapp.com/the-future-is-here-ai-ready-cloud-connected-all-flash-storage-with-nvme/
[4] https://www.netapp.com/us/products/storage-systems/all-flash-array/aff-a-series.aspx#technical-specifications

3.2 Scalability

Large data sets are important for increasing model accuracy. DL deployments that start out small (a few terabytes of storage) may soon need to scale out to several petabytes. Furthermore, performance needs can vary based on the training model used and the end application, requiring independent scaling of compute and/or storage. Designing a robust system architecture in a rack-scale environment enables independent scaling.

The NetApp A800 and A700s systems can scale independently, seamlessly, and non-disruptively from 2-nodes (364.8TB) to a 24-node cluster (74.8PB with A800, 39.7PB with A700s). Using ONTAP® FlexGroup™ volumes enables easy data management in a single namespace logical volume of 20PB, supporting more than 400 billion files. For cluster capacities greater than 20PB, one can create multiple FlexGroups to span the required capacity.

3.3 Robust Platform Integration

As organizations accelerate their rate of data collection, the need to introduce automation around that data becomes apparent. Using containers is one way to achieve this; it enables faster deployments by separating applications from the OS and device-driver layer dependencies. Efficient and easy data management is central to reducing the time-to-train.

Trident™ is a NetApp dynamic storage orchestrator for container images that is fully integrated with Docker™ and Kubernetes™. Combined with NVIDIA GPU Cloud (NGC) and popular orchestrators like Kubernetes or Docker Swarm, Trident enables customers to seamlessly deploy their AI/DL NGC Container Images onto NetApp Storage allowing an enterprise grade experience for their AI container deployments. This includes automated orchestration, cloning for test and dev, NGC upgrade testing using cloning, protection and compliance copies and many more data management use cases for the NGC AI container images.

Data traffic in DL can consist of millions of files (images, video/audio, text files). Network File Systems (NFS), are ideally suited for delivering high performance across a diverse range of workloads; they handle both random and sequential I/O well. When used with ONTAP FlexGroup volumes, NetApp AFF can deliver high performance for small file workloads spread across storage systems.

4 Reference Architecture

Architecting an infrastructure that can deliver sustained high I/O throughput at low latencies to exploit the I/O parallelism of the GPUs and also non-disruptively scale compute and storage systems is vital in achieving faster training times. These requirements translate to the need for a storage system that supports a high-speed, high-bandwidth, low-latency network fabric in order to maximize system performance and to keep multiple DGX-1 servers constantly fed with data required for each DL training run.

Figures 1 shows a NetApp architecture in a 1:5 configuration that consists of five DGX-1 servers fed by one A800 high availability (HA) pair via two switches. Each DGX-1 server connects to each of the two switches via two 100GbE links. The A800 connects via four 100GbE links to each switch. The switches can have two to four 100Gb inter-switch links, designed for failover scenarios. The HA design is active-active, so maximum throughput can be sustained across all network connections in the absence of a failure.

Figure 1) Reference architecture in a 1:5 configuration.

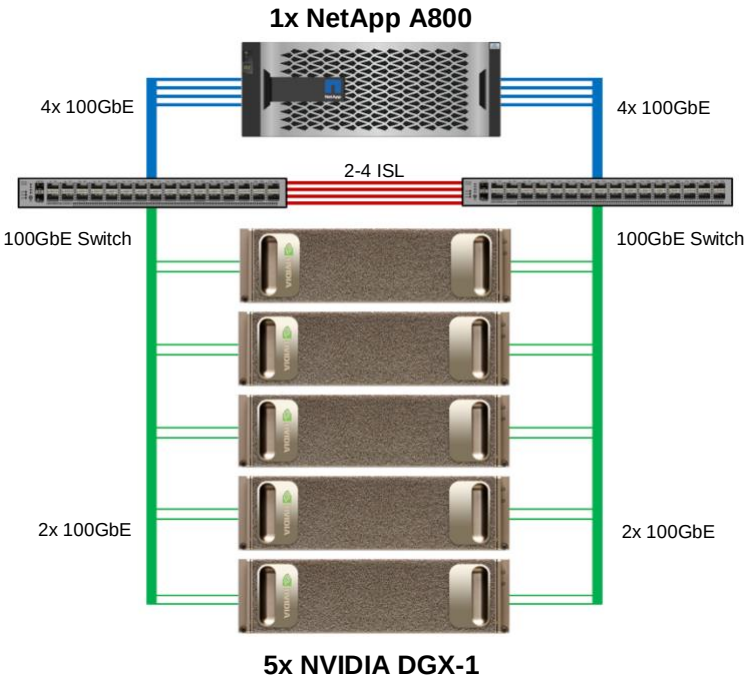
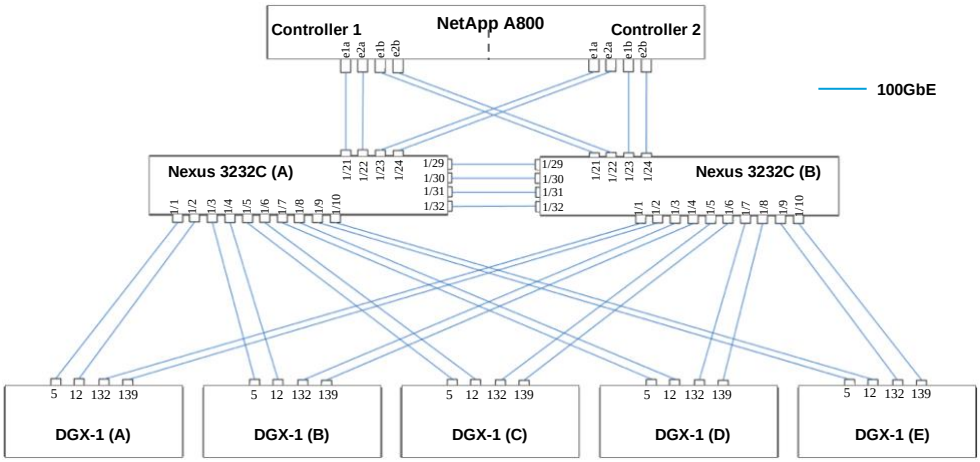


Figure 2 shows the port level connectivity of the architecture. This architecture consists of two Cisco Nexus 3232C 100GbE switches as well as a 1GbE management switch (not shown). There are five DGX-1 servers connected to the network with four 100GbE links each, with two links from each DGX-1 connected to each Nexus switch. Storage is provided by a NetApp A800 HA pair, with each of the two storage controllers connected to each Nexus switch with two 100GbE links.

Figure 2) Network diagram showing port-level connectivity for a 1:5 configuration.



With A700s, the architecture shown in Figure 1 changes to a 1:4 configuration (1 A700s : 4 DGX-1), with four 40GbE links from the A700s to each switch on the storage side and two 100GbE links from each DGX-1 to each of two switches. In addition to 100GbE, the A800 system also supports 40GbE. Both architectures can scale up and out with zero down time as the datasets grow in size.

5 Rack Scale – Start Small, Scale Big

Scale-out means that as the storage environment grows, additional storage capacity and/or compute nodes are added seamlessly to the resource pool residing on a shared storage infrastructure. Host and client connections as well as datastores can move seamlessly anywhere in the resource pool. Therefore, existing workloads can be easily balanced over the available resources and new workloads can be easily deployed. Technology refreshes (adding or replacing drive shelves and/or storage controllers) are accomplished while the environment remains online and continues serving data.

NetApp has combined the compute power of DGX-1 servers with the high-performance architecture of A800 and A700s systems to offer a compelling solution that enables organizations to deploy DL workflows in a few hours and seamlessly scale out as needed.

Organizations starting out with DL might start with a 1:1 configuration and scale out as the data grows to a 1:5 configuration and beyond in a scale-out mode. Table 1 highlights the capacity and performance scaling that can be achieved with a spectrum of configurations with DGX-1 and A800 using ONTAP 9.4.

Table 1) Capacity and performance metrics in scale-out scenarios with A800.

# of **A800** Storage Systems	# of DGX-1 Servers	Throughput	Typical Raw Capacity[5]	Raw Capacity w/ Expansion[5]
1 HA pair	5	25GB/s	364.8TB	6.2PB

The information in Table 1 is based on A800 and ONTAP 9.4 performance metrics. Delivering 25GB/s of throughput, each A800 is able to handle traffic from 5 DGX-1 systems while providing the option of scaling to 6.2PB of storage.

Table 2) Capacity and performance metrics in scale-out scenarios with A700s.

# of **A700s** Storage Systems	# of DGX-1 Servers	Throughput	Typical Raw Capacity[5]	Raw Capacity w/ Expansion[5]
1 HA pair	4	18GB/s	367.2TB	3.3PB

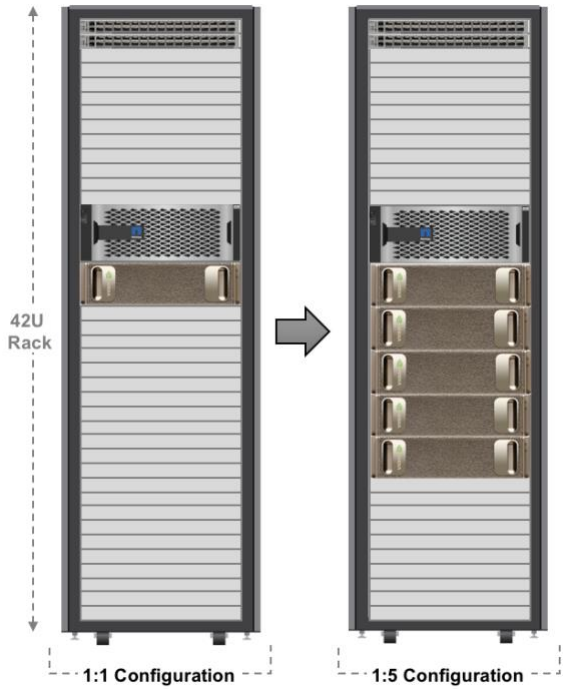
The information in Table 2 is based on A700s and ONTAP 9.4 performance metrics. The A700s system can support a throughput of 18GB/s and is ideal for a 1:4 configuration as a starting point.

Organizations with a need to start at a lower storage footprint and cost can use the NetApp A300 or A220 storage systems which also support seamless scalability.

Based on the scaling information in Table 1, Figure 3 illustrates how a 1:1 configuration can scale to a 1:5 configuration deployment in a data center. This approach gives the flexibility of altering the ratios of compute to storage based on the size of the data lake, the DL models used, and the required performance metrics.

[5] https://www.netapp.com/us/media/ds-3582.pdf

Figure 3) Rack-level scaling from a 1:1 to a 1:5 configuration with A800.



The number of DGX-1 servers and AFF storage 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.

6 Performance Testing

TensorFlow benchmarks were run on a 1:1 configuration setup (one DGX-1 server and one A700 storage system) with ImageNet dataset (143GB) stored on a FlexGroup volume on the A700 system. NFSv3 was the file system of choice for these tests.

Environment Setup:

- OS: Ubuntu 16.04 LTS
- Docker: 18.03.1-ce [9ee9f40]
- Dockerfile: nvcr.io/nvidia/tensorflow:18.04-py2
- Framework: Tensorflow 1.7.0
- Benchmarks: Tensorflow Benchmarks [26a8b0a]

As part of our initial tests, we ran benchmarks based on synthetic data to study the performance of GPUs without potential TensorFlow pipeline or storage related bottlenecks. Training was run with synthetic data on both CUDA cores and Tensor cores for all the models in the TensorFlow Benchmark.

As a next step, real data with distortion was used for all tests.

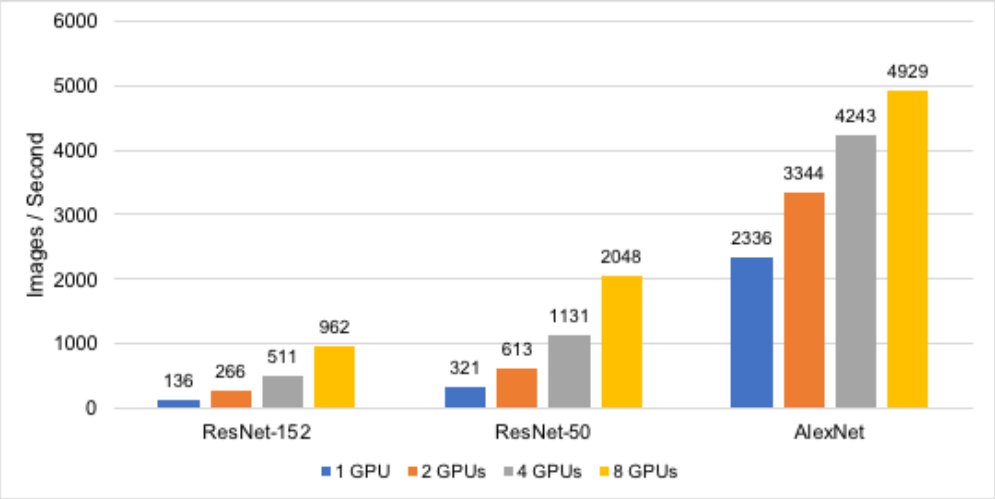
The following points highlight the salient aspects of the tests:

- Model training performance is measured as number of images processed per second.
- To demonstrate the training rates that can be achieved, three popular models representing varying degrees of computational complexity and predictive accuracy were chosen – ResNet-152, ResNet-50, and AlexNet.
- To reflect real life model training scenarios, distortion (image pre-processing steps) was enabled to stress the system from a storage and GPU processing standpoint.

- Performance metrics were measured with varying number of GPUs enabled on the DGX-1 server.
- GPU utilizations were maintained close to 100% during the entire training epoch indicating that the A700 system is able to feed data fast enough to the GPUs while maintaining high training rates.
- AlexNet, being the most storage I/O intensive model was chosen to stress the pipeline and illustrate extreme use cases. This model may not be the most accurate and is known to be restrictive in scaling scenarios.
- Batch size used – 64 for ResNet-152 and ResNet-50 and 512 for AlexNet.

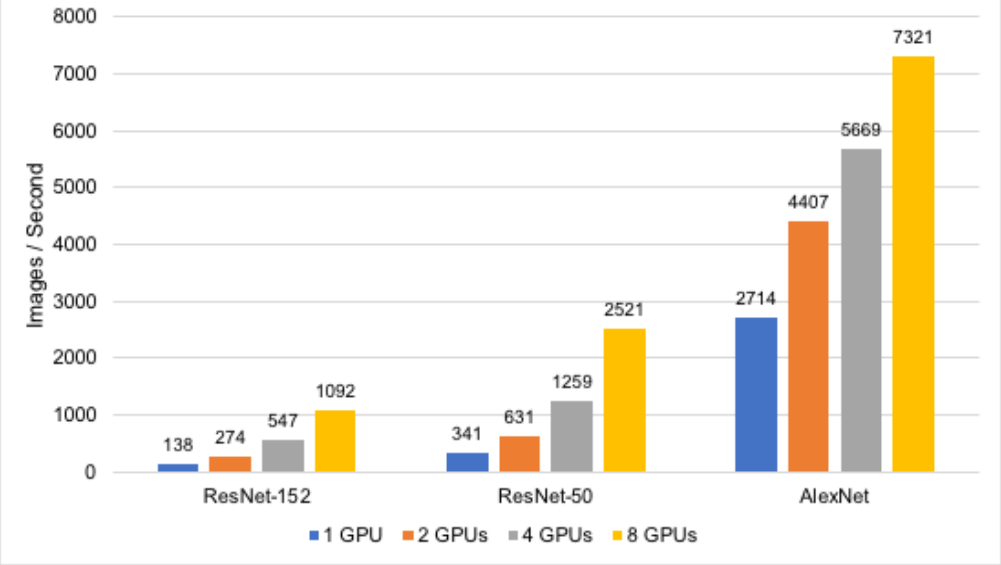
Figures 4 and 5 summarize the training performance measured with each of the three DL models with one, two, four, and eight GPUs.

Figure 4) Training rates with real data and image distortions enabled.



* Data rounded to the nearest decimal value

Figure 5) Training rates with real data and image distortions disabled.



* Data rounded to the nearest decimal value

Figure 6 illustrates the GPU utilizations achieved while training the ResNet-50 model with 8 GPUs. The green curve shows the sum of utilizations of all 8 GPUs and the orange curve indicates the read throughput from A700. To maintain high GPU utilizations and a training rate of ~2500 images/sec, the read throughput from A700 reaches ~300MB/s. It took ~500 seconds to load the 143GB dataset into the DGX-1 memory. Before and after the 500 second mark, the GPU utilization and training rates are identical. This indicates there were no storage I/O or other bottlenecks in the pipeline feeding the GPUs at this training rate.

Figure 6) GPU utilization and A700 read throughput with ResNet-50 at ~2500 images/sec rate.

7 Conclusion

AI requires massive compute power and an infrastructure architecture that can keep pace. Disciplines like DL demand extreme performance in order to feed the GPUs that power data-hungry algorithms. As AI becomes a core business capability, virtually all organizations will rely on insights generated from the large datasets they're generating. To meet their objectives, they will need an infrastructure that can be stood up quickly, deliver the required parallel performance, scale effortlessly, and be managed easily.

The DGX-1 server has made it possible to glean insights from massive data lakes by accelerating model training speeds. Combining the state-of-the-art GPU hardware and GPU-optimized containers from NGC makes deploying DL applications fast and efficient.

As the industry leader in NFS, NetApp has a suite of products with the AFF product family, ONTAP, FlexGroup, Trident and leading storage efficiency features to meet the massively parallel performance requirements of DL applications coupled with the real-world expertise in deploying AI solutions.

NetApp has partnered with NVIDIA to introduce a rack-scale architecture enabling organizations to start small and grow the infrastructure seamlessly as the number of projects and dataset sizes increase. This architecture has been designed to relieve organizations from dealing with increasing infrastructure complexity and helps them focus on developing better DL applications. Adopting these AI solutions empower businesses to meet even the most demanding performance requirements, leading to a new era of intelligent applications.

8 Appendix: Component List

Table 3 lists the components that were used for the architectural designs described in this report.

Table 3) Component list.

Component	Qty	Description
Base Server	1	Dual Intel Xeon CPU motherboard with x2 9.6 GT/s QPI, 8 Channel with 2 DPC DDR4, Intel X99 Chipset, AST2400 BMC
	1	GPU baseboard supporting 8 SXM2 modules (hybrid cube mesh) and 4 PCIE x16 slots for InfiniBand NICs
Connectivity Ports	1	10/100BASE-T IPMI port
	1	RS232 serial port
	2	USB 3.0 ports
CPU	2	Intel Xeon E5-2698 v4, 20-core, 2.2GHz, 135W
GPU	8	Tesla V100: 1 petaFLOPS, mixed precision 32GB memory per GPU 40,960 NVIDIA CUDA® cores 5120 NVIDIA Tensor cores
System Memory	16	32GB DDR4 LRDIMM (512GB total)
SAS Raid Controller	1	8-port LSI SAS 3108 RAID Mezzanine
Storage (RAID 0) (Data)	4	1.92TB, 6Gb/s, SATA 3.0 SSD
Storage (OS)	1	480GB, 6Gb/s, SATA 3.0 SSD
10GbE NIC	1	Dual port, 10GBASE-T, network adapter Mezzanine
Ethernet / InfiniBand EDR NIC	4	Single port, x16 PCIe, Mellanox ConnectX-4 VPI MCX455A-ECAT
NetApp A800/A700/A300	1	All-flash array, 1 HA-pair
Cisco Nexus 3232C	2	100Gb Ethernet switches
Cisco Nexus 3048-TP	1	1Gb Ethernet management switch

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