NetApp AFF A400 with Lenovo ThinkSystem SR670 V2 for AI and ML Model Training

NetApp Solutions

NetApp

July 09, 2024

Table of Contents

NetApp AFF A400 with Lenovo ThinkSystem SR670 V2 for AI and ML Model Training

TR-4810: NetApp AFF A400 with Lenovo ThinkSystem SR670 V2 for AI and ML Model Training

Sathish Thyagarajan, David Arnette, NetApp
Mircea Troaca, Lenovo

This solution presents a mid-range cluster architecture using NetApp storage and Lenovo servers optimized for artificial intelligence (AI) workloads. It is meant for small- to medium-sized enterprises for which most compute jobs are single node (single or multi-GPU) or distributed over a few computational nodes. This solution aligns with most day-to-day AI training jobs for many businesses.

This document covers testing and validation of a compute and storage configuration consisting of eight-GPU Lenovo SR670V2 servers, a mid-range NetApp AFF A400 storage system and 100GbE interconnect switch. To measure the performance, we used ResNet50 with the ImageNet dataset, a batch size of 408, half precision, CUDA, and cuDNN. This architecture provides an efficient and cost-effective solution for small and medium-sized organizations just starting out with AI initiatives that require the enterprise-grade capabilities of NetApp ONTAP cloud-connected data storage.

Target audience

This document is intended for the following audiences:

- Data scientists, data engineers, data administrators, and developers of AI systems
- Enterprise architects who design solutions for the development of AI models
- Data scientists and data engineers who are looking for efficient ways to achieve deep learning (DL) and machine learning (ML) development goals
- Business leaders and OT/IT decision makers who want to achieve the fastest possible time to market for AI initiatives

Solution architecture

This solution with Lenovo ThinkSystem servers and NetApp ONTAP with AFF storage is designed to handle AI training on large datasets using the processing power of GPUs alongside traditional CPUs. This validation demonstrates high performance and optimal data management with a scale-out architecture that uses either one, two, or four Lenovo SR670 V2 servers alongside a single NetApp AFF A400 storage system. The following figure provides an architectural overview.

This NetApp and Lenovo solution offers the following key benefits:

- Highly efficient and cost-effective performance when executing multiple training jobs in parallel
- Scalable performance based on different numbers of Lenovo servers and different models of NetApp storage controllers

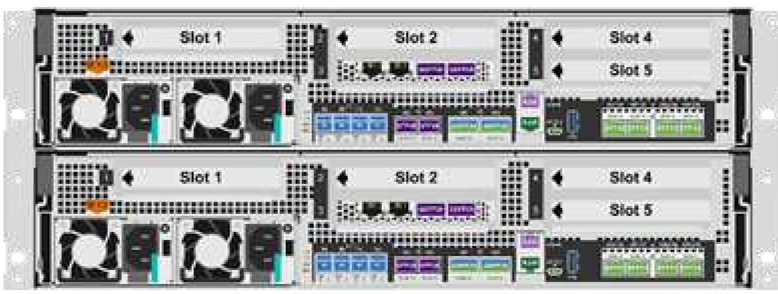
- Robust data protection to meet low recovery point objectives (RPOs) and recovery time objectives (RTOs) with no data loss
- Optimized data management with snapshots and clones to streamline development workflows

Technology overview

This section introduces the major components of this solution in greater detail.

NetApp AFF systems

NetApp AFF storage systems enable businesses 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.



NetApp AFF A400 is a mid-range NVMe flash storage system that includes the following features:

- Maximum effective capacity: ~20PB
- Maximum scale-out: 2-24 nodes (12 HA pairs)
- 25GbE and 16Gb FC host support
- 100GbE RDMA over Converged Ethernet (RoCE) connectivity to NVMe expansion storage shelves
- 100GbE RoCE ports can be used for host network attachment if NVMe shelves aren’t attached
- Full 12Gbps SAS connectivity expansion storage shelves
- Available in two configurations:
    - Ethernet: 4x 25Gb Ethernet (SFP28) ports
    - Fiber Channel: 4x 16Gb FC (SFP+) ports
- 100% 8KB random read @.4 ms 400k IOPS

NetApp AFF A250 features for entry level AI/ML deployments include the following:

- Maximum effective capacity: 35PB

- Maximum scale out: 2-24 nodes (12 HA pairs)
- 440k IOPS random reads @1ms
- Built on the latest NetApp ONTAP release ONTAP 9.8 or later
- Two 25Gb Ethernet ports for HA and cluster interconnect

NetApp also offers other storage systems, such as the AFF A800 and AFF A700 that provide higher performance and scalability for larger-scale AI/ML deployments.

NetApp ONTAP

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

Simplify data management

Data management is crucial to enterprise IT operations so that appropriate resources are used for applications and 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. This applies to data stored locally and data tiered to the cloud.
- **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 automatically tiers cold data to public and private cloud storage options, including Amazon Web Services (AWS), Azure, and NetApp StorageGRID object storage.

Accelerate and protect data

ONTAP delivers superior levels of performance and data protection and extends these capabilities in the following ways:

- **Performance and lower latency.** ONTAP offers the highest possible throughput at the lowest possible latency.
- **Data protection.** ONTAP provides built-in data protection capabilities with common management across all platforms.
- **NetApp Volume Encryption.** ONTAP offers native volume-level encryption with both onboard and external key management support.

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.

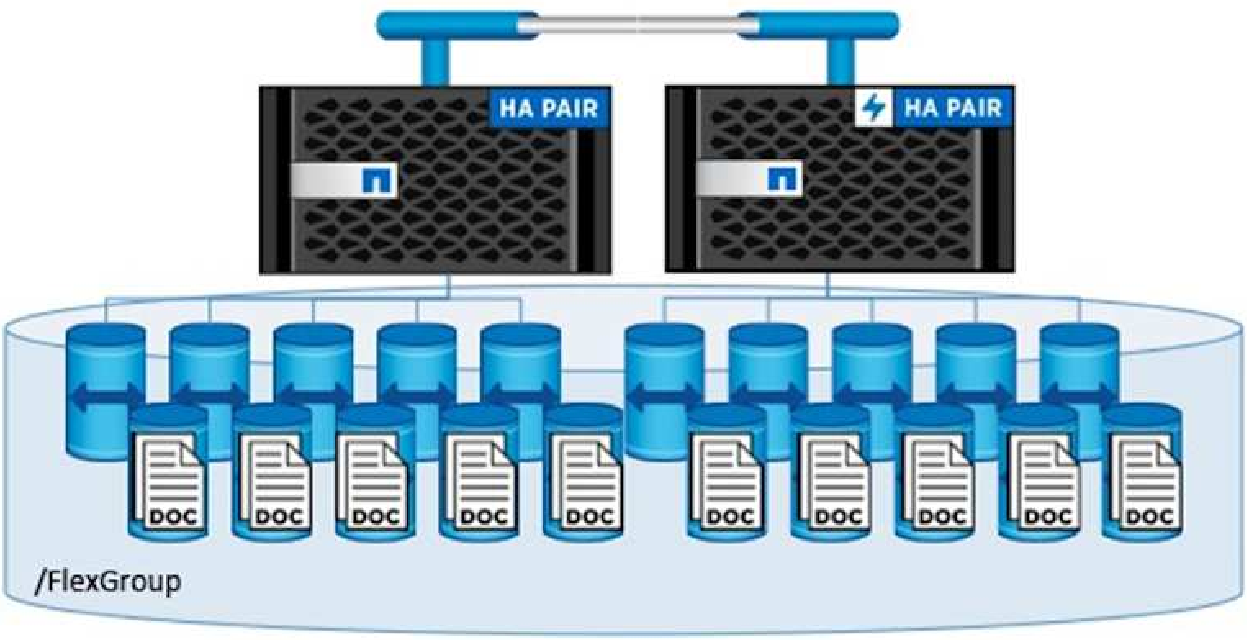
- **Cloud connection.** ONTAP is the most cloud-connected storage management software, with options for software-defined storage (ONTAP Select) and cloud-native instances (NetApp Cloud Volumes Service) in all public clouds.
- **Integration with emerging applications.** ONTAP offers enterprise-grade data services for next-generation platforms and applications such as OpenStack, Hadoop, and MongoDB by using the same infrastructure that supports existing enterprise apps.

NetApp FlexGroup volumes

Training datasets are typically a collection of 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 (the following figure) is a single namespace made up of multiple constituent member volumes 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:

- Up to 20 petabytes of capacity and predictable low latency for high-metadata workloads
- Up to 400 billion files in the same namespace
- Parallelized operations in NAS workloads across CPUs, nodes, aggregates, and constituent FlexVol volumes



"

Lenovo ThinkSystem portfolio

Lenovo ThinkSystem servers feature innovative hardware, software, and services that solve customers' challenges today and deliver an evolutionary, fit-for-purpose, modular design approach to address tomorrow's challenges. These servers capitalize on best-in-class, industry-standard technologies coupled with differentiated Lenovo innovations to provide the greatest possible flexibility in x86 servers.

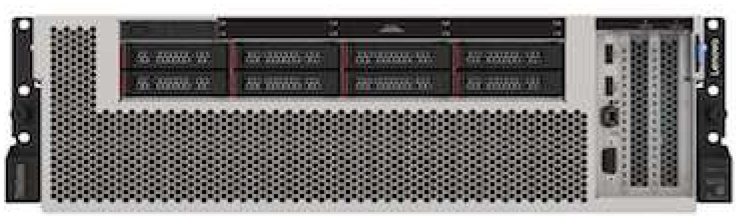
Key advantages of deploying Lenovo ThinkSystem servers include the following:

- Highly scalable, modular designs that grow with your business
- Industry-leading resilience to save hours of costly unscheduled downtime
- Fast flash technologies for lower latencies, quicker response times, and smarter data management in real time

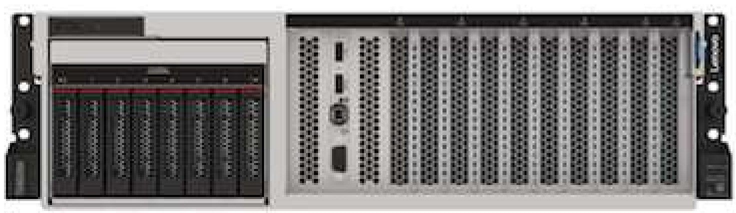
In the AI area, Lenovo is taking a practical approach to helping enterprises understand and adopt the benefits of ML and AI for their workloads. Lenovo customers can explore and evaluate Lenovo AI offerings in Lenovo AI Innovation Centers to fully understand the value for their particular use case. To improve time to value, this customer-centric approach gives customers proofs of concept for solution development platforms that are ready to use and optimized for AI.

Lenovo SR670 V2

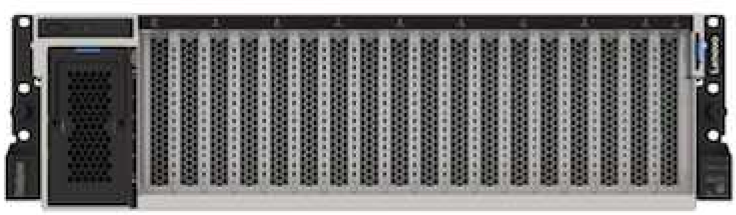
The Lenovo ThinkSystem SR670 V2 rack server delivers optimal performance for accelerated AI and high-performance computing (HPC). Supporting up to eight GPUs, the SR670 V2 is suited for the computationally intensive workload requirements of ML, DL, and inference.



4x SXM GPUs with 8x 2.5-inch HS drives and 2x PCIe I/O slots



4x double-wide or 8x single-wide GPU slots and 2x PCIe I/O slots with 8x 2.5-inch or 4x 3.5-inch HS drives



8x double-wide GPU slots with 6x EDSFF HS drives and 2x PCIe I/O slots

With the latest scalable Intel Xeon CPUs that support high-end GPUs (including the NVIDIA A100 80GB PCIe 8x GPU), the ThinkSystem SR670 V2 delivers optimized, accelerated performance for AI and HPC workloads.

Because more workloads use the performance of accelerators, the demand for GPU density has increased. Industries such as retail, financial services, energy, and healthcare are using GPUs to extract greater insights and drive innovation with ML, DL, and inference techniques.

The ThinkSystem SR670 V2 is an optimized, enterprise-grade solution for deploying accelerated HPC and AI workloads in production, maximizing system performance while maintaining data center density for supercomputing clusters with next-generation platforms.

Other features include:

- Support for GPU direct RDMA I/O in which high-speed network adapters are directly connected to the GPUs to maximize I/O performance.
- Support for GPU direct storage in which NVMe drives are directly connected to the GPUs to maximize storage performance.

MLPerf

MLPerf is the industry-leading benchmark suite for evaluating AI performance. In this validation, we used its image-classification benchmark with MXNet, one of the most popular AI frameworks. The MXNet_benchmarks training script was used to drive AI training. The script contains implementations of several popular conventional models and is designed to be as fast as possible. It can be run on a single machine or run in distributed mode across multiple hosts.

Test plan

In this validation, we performed image recognition training as specified by MLPerf v2.0. Specifically, we trained the ResNet v2.0 model with the ImageNet dataset until we reached an accuracy of 76.1%. The main metric is the time to reach the desired accuracy. We also report training bandwidth in images per second to better judge scale-out efficiency.

The primary test case evaluated multiple independent training processes (one per node) running concurrently. This simulates the main use case, a shared system used by multiple data scientists. The second test case evaluated scale-out efficiency.

Test results

The following table summarizes the results for all tests performed for this solution.

Test description	Results summary
Image recognition training: multiple concurrent jobs	Highly efficient performance. All jobs ran at full speed even when the cluster was fully used. The NetApp storage systems delivered training performance comparable to local SSD storage while enabling easy sharing of data between servers.
Image recognition training: scale out	Highly efficient for up to four nodes. At that point, scale out was less efficient but still feasible. Using a higher-speed computational network improves scalability. The NetApp storage system delivered training performance comparable to local SSD storage while enabling easy sharing of data between servers.

Test configuration

This section describes the tested configurations, the network infrastructure, the SR670 V2 server, and the NetApp storage provisioning details.

Solution architecture

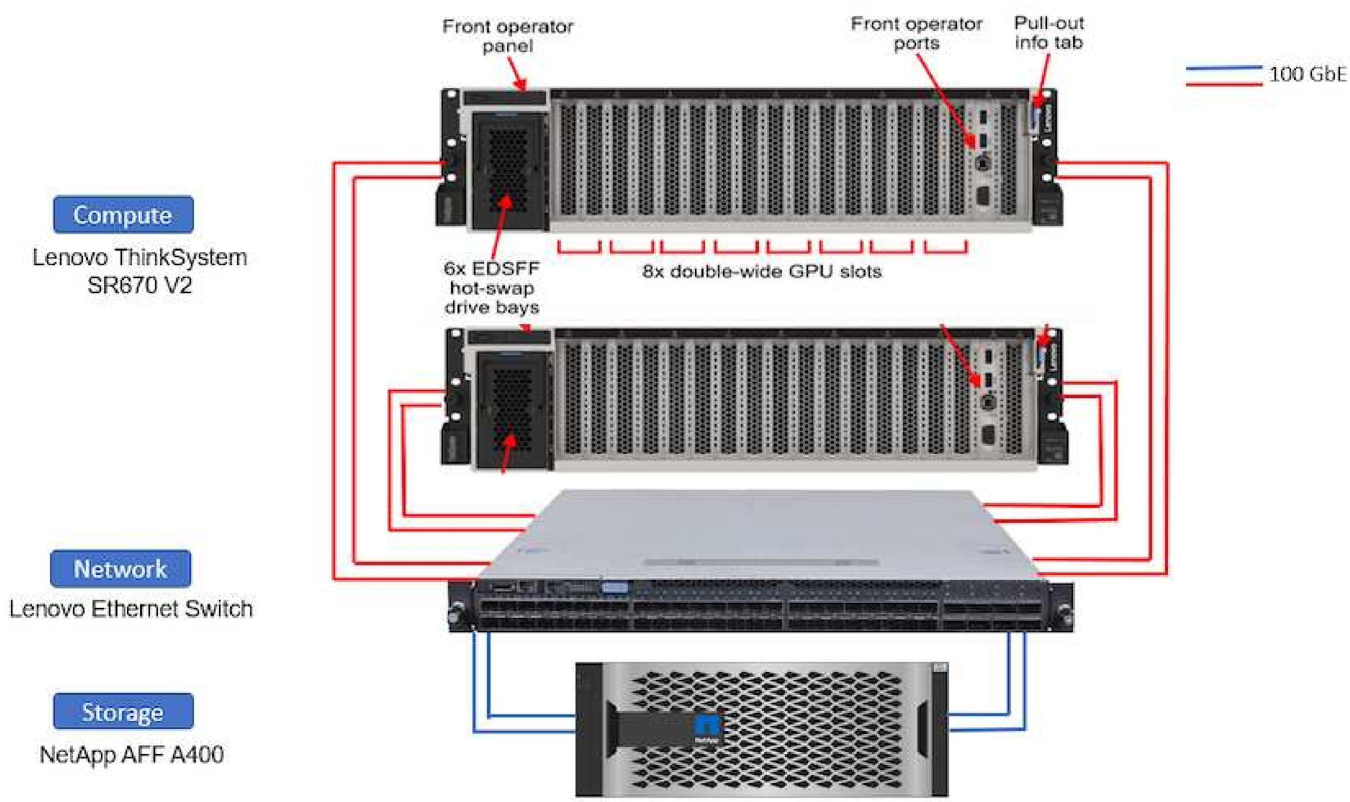
We used the solution components listed in the following table for this validation.

Solution components	Details
Lenovo ThinkSystem servers	• Two SR670 V2 servers each with eight NVIDIA A100 80GB GPU cards • Each server contains 2 Intel Xeon Platinum 8360Y CPUs (28 physical cores) and 1TB RAM
Linux (Ubuntu – 20.04 with CUDA 11.8)	
NetApp AFF storage system (HA pair)	• NetApp ONTAP 9.10.1 software • 24x 960GB SSDs • NFS protocol • 1 interface group (ifgrp) per controller, with four logical IP addresses for mount points

In this validation, we used ResNet v2.0 with the ImageNet basis set as specified by MLPerf v2.0. The dataset is stored in a NetApp AFF storage system with the NFS protocol. The SR670s were connected to the NetApp AFF A400 storage system over a 100GbE switch.

ImageNet is a frequently used image dataset. It contains almost 1.3 million images for a total size of 144GB. The average image size is 108KB.

The following figure depicts the network topology of the tested configuration.



Storage controller

The following table lists the storage configuration.

Controller	Aggregate	FlexGroup volume	Aggregate size	Volume size	Operating system mount point
Controller1	Aggr1	/a400-100g	9.9TB	19TB	/a400-100g
Controller2	Aggr2	/a400-100g	9.9TB		/a400-100g

The /a400-100g folder contains the dataset used for ResNet validation.

Test procedure and detailed results

This section describes the detailed test procedure results.

Image recognition training using ResNet in ONTAP

We ran the ResNet50 benchmark with one and two SR670 V2 servers. This test used the MXNet 22.04-py3 NGC container to run the training.

We used the following test procedure in this validation:

1. We cleared the host cache before running the script to make sure that data was not already cached:

```
sync ; sudo /sbin/sysctl vm.drop_caches=3
```

2. We ran the benchmark script with the ImageNet dataset in server storage (local SSD storage) as well as on the NetApp AFF storage system.
3. We validated network and local storage performance using the `dd` command.
4. For the single-node run, we used the following command:

```
python train_imagenet.py --gpus 0,1,2,3,4,5,6,7 --batch-size 408 --kv
-store horovod --lr 10.5 --mom 0.9 --lr-step-epochs pow2 --lars-eta
0.001 --label-smoothing 0.1 --wd 5.0e-05 --warmup-epochs 2 --eval-period
4 --eval-offset 2 --optimizer sgdwfastlars --network resnet-v1b-stats-fl
--num-layers 50 --num-epochs 37 --accuracy-threshold 0.759 --seed 27081
--dtype float16 --disp-batches 20 --image-shape 4,224,224 --fuse-bn-relu
1 --fuse-bn-add-relu 1 --bn-group 1 --min-random-area 0.05 --max-random
-area 1.0 --conv-algo 1 --force-tensor-core 1 --input-layout NHWC --conv
-layout NHWC --batchnorm-layout NHWC --pooling-layout NHWC --batchnorm
-mom 0.9 --batchnorm-eps 1e-5 --data-train /data/train.rec --data-train
-idx /data/train.idx --data-val /data/val.rec --data-val-idx
/data/val.idx --dali-dont-use-mmap 0 --dali-hw-decoder-load 0 --dali
-prefetch-queue 5 --dali-nvjpeg-memory-padding 256 --input-batch
-multiplier 1 --dali- threads 6 --dali-cache-size 0 --dali-roi-decode 1
--dali-preallocate-width 5980 --dali-preallocate-height 6430 --dali-tmp
-buffer-hint 355568328 --dali-decoder-buffer-hint 1315942 --dali-crop
-buffer-hint 165581 --dali-normalize-buffer-hint 441549 --profile 0
--e2e-cuda-graphs 0 --use-dali
```

5. For the distributed runs, we used the parameter server's parallelization model. We used two parameter servers per node, and we set the number of epochs to be the same as for the single-node run. We did this because distributed training often takes more epochs due to imperfect synchronization between processes. The different number of epochs can skew comparisons between single-node and distributed cases.

Data read speed: Local versus network storage

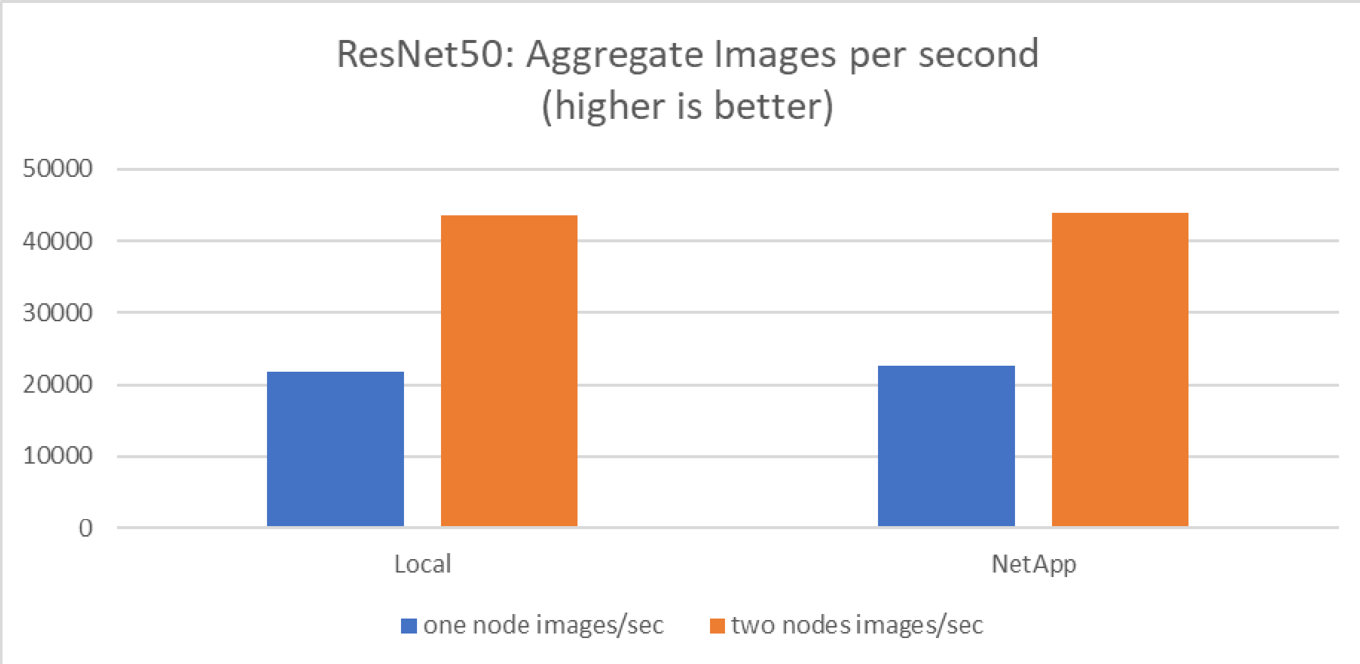
The read speed was tested by using the `dd` command on one of the files for the ImageNet dataset. Specifically, we ran the following commands for both local and network data:

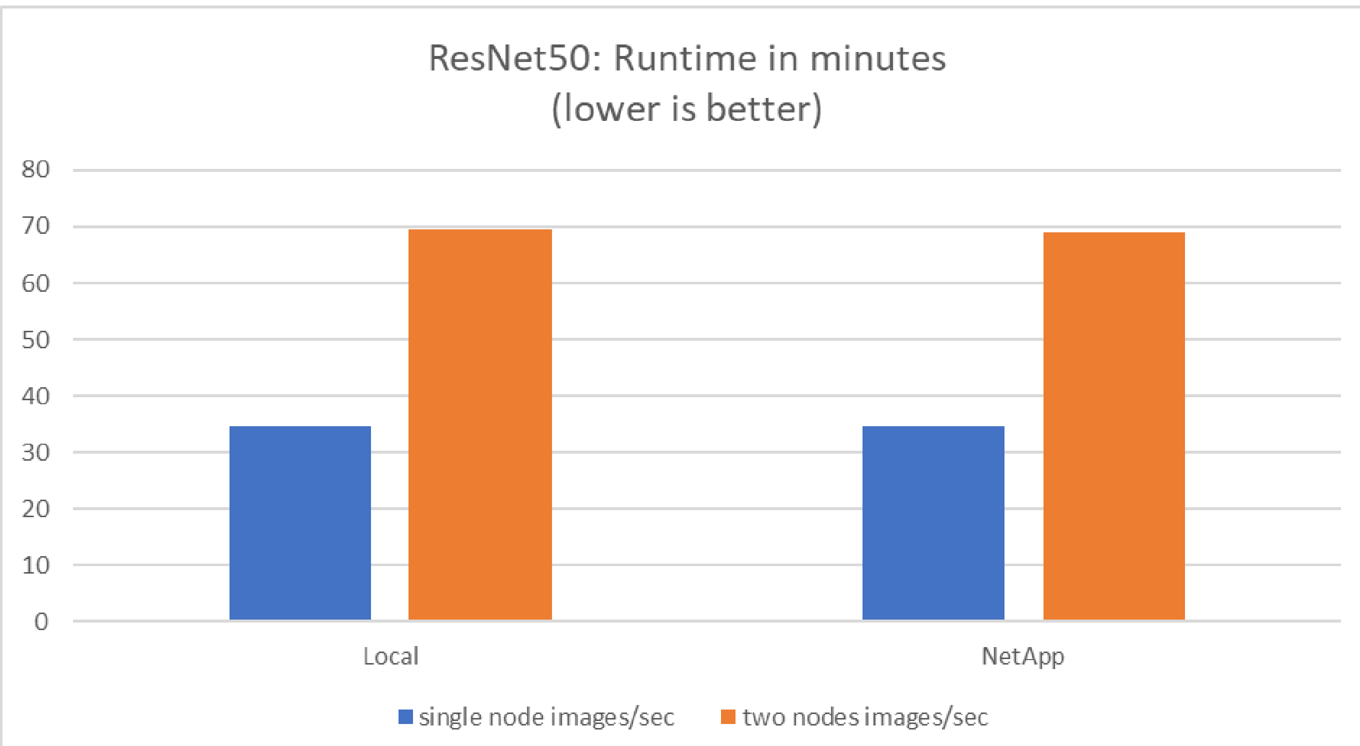
```
sync ; sudo /sbin/sysctl vm.drop_caches=3dd if=/a400-100g/netapp-
ra/resnet/data/preprocessed_data/train.rec of=/dev/null bs=512k
count=2048Results (average of 5 runs):
Local storage: 1.7 GB/s Network storage: 1.5 GB/s.
```

Both values are similar, demonstrating that the network storage can deliver data at a rate similar to local storage.

Shared use case: Multiple, independent, simultaneous jobs

This test simulated the expected use case for this solution: multi-job, multi-user AI training. Each node ran its own training while using the shared network storage. The results are displayed in the following figure, which shows that the solution case provided excellent performance with all jobs running at essentially the same speed as individual jobs. The total throughput scaled linearly with the number of nodes.





These graphs present the runtime in minutes and the aggregate images per second for compute nodes that used eight GPUs from each server on 100 GbE client networking, combining both the concurrent training model and the single training model. The average runtime for the training model was 35 minutes and 9 seconds. The individual runtimes were 34 minutes and 32 seconds, 36 minutes and 21 seconds, 34 minutes and 37 seconds, 35 minutes and 25 seconds, and 34 minutes and 31 seconds. The average images per second for the training model were 22,573, and the individual images per second were 21,764; 23,438; 22,556; 22,564; and 22,547.

Based on our validation, one independent training model with a NetApp data runtime was 34 minutes and 54 seconds with 22,231 images/sec. One independent training model with a local data (DAS) runtime was 34 minutes and 21 seconds with 22,102 images/sec. During those runs the average GPU utilization was 96%, as observed on nvidia-smi. Note that this average includes the testing phase, during which GPUs were not used, while CPU utilization was 40% as measured by mpstat. This demonstrates that the data delivery rate is sufficient in each case.

Architecture adjustments

The setup used for this validation can be adjusted to fit other use cases.

CPU Adjustments

We used a Skylake Intel Xeon Platinum 8360Y processor for this validation, as recommended by Lenovo. We expect that the equivalent Cascade Lake CPU, an Intel Xeon Gold 6330 processor, would deliver similar performance because this workload is not CPU bound.

Storage Capacity Increase

Based on your storage capacity needs, you can increase the share storage (NFS volume) on demand, provided that you have the additional disk shelves and controller models. You can do this from the CLI or from the NetApp web interface of the storage controller as the admin user.

Conclusion

The NetApp and Lenovo solution validated here is a flexible scale-out architecture that is ideal for entry into mid-level enterprise AI.

NetApp storage delivers the same or better performance as local SSD storage and offers the following benefits to data scientists, data engineers, and IT decision makers:

- Effortless sharing of data between AI systems, analytics, and other critical business systems. This data sharing reduces infrastructure overhead, improves performance, and streamlines data management across the enterprise.
- Independently scalable compute and storage to minimize costs and improve resource utilization.
- Streamlined development and deployment workflows using integrated snapshots and clones for instantaneous and space-efficient user workspaces, integrated version control, and automated deployment.
- Enterprise-grade data protection for disaster recovery and business continuance.

Acknowledgments

- Karthikeyan Nagalingam, Technical Marketing Engineer, NetApp
- Jarrett Upton, Admin, AI Lab Systems, Lenovo

Where to find additional information

To learn more about the information described in this document, refer to the following documents and/or websites:

- NetApp All Flash Arrays product page

  https://www.netapp.com/us/products/storage-systems/all-flash-array/aff-a-series.aspx

- NetApp AFF A400 page

  https://docs.netapp.com/us-en/ontap-systems/a400/index.html

- NetApp ONTAP data management software product page

  http://www.netapp.com/us/products/data-management-software/ontap.aspx

- MLPerf

  https://mlperf.org

- TensorFlow benchmark

  https://github.com/tensorflow/benchmarks

- NVIDIA SMI (nvidia-smi)

  https://developer.nvidia.com/nvidia-system-management-interface