



Technical Report

AI Inferencing at the Edge

NetApp ONTAP with Lenovo ThinkSystem Solution Design

Sathish Thyagarajan, NetApp
Miroslav Hodak, Lenovo
March 2021 | TR-4886

In partnership with



Abstract

Several emerging application scenarios, such as advanced driver-assistance systems (ADAS), Industry 4.0, smart cities, and Internet of Things (IoT), require the processing of continuous data streams under a near-zero latency. This document describes a compute and storage architecture to deploy GPU-based AI inferencing on NetApp® storage controllers and Lenovo ThinkSystem servers in an edge environment that meets these requirements. This document also provides performance data for the industry standard MLPerf Inference benchmark, evaluating various inference tasks on edge servers equipped with NVIDIA T4 GPUs. We investigate the performance of offline, single stream, and multistream inference scenarios and show that the architecture with a cost-effective shared networked storage system is highly performant and provides a central point for data and model management for multiple edge servers.

TABLE OF CONTENTS

Introduction	4
Target audience.....	4
Solution architecture	4
How to use this architecture	6
Solution areas.....	6
Technology overview	7
NetApp AFF systems.....	7
NetApp AFF A220	8
NetApp AFF A250	8
Lenovo ThinkSystem DM 5100.....	8
NetApp ONTAP 9	8
NetApp Trident	9
NetApp Cloud Sync	10
Test Plan	11
Test configuration.....	12
Test procedure	13
Operating system setup.....	13
AI inference setup.....	13
AI inference runs	14
Test results	14
AI inference in offline scenario.....	14
AI inference in a single stream scenario	15
AI inference in multistream scenario.....	16
Architecture sizing options.....	16
Compute server	16
Shared storage.....	16
Conclusion	17
Acknowledgments	17
Where to find additional information	17
Version history	18

LIST OF TABLES

Table 1) MLPerf Inference: Edge suite benchmarks.	12
Table 2) Edge benchmark scenarios.	12
Table 3) Base components for the solution architecture.	12
Table 4) Storage configuration: AFF C190 with 2RU, 24 drive slots.	13

LIST OF FIGURES

Figure 1) Physical architecture overview.	5
Figure 2) Logical architecture overview.	5
Figure 3) NetApp AFF C190 storage system.	8
Figure 4) Lenovo ThinkSystem SE350 Edge Server.	10
Figure 5) Network topology of tested configuration.	13
Figure 6) Bandwidth in samples per second for various inference tasks.	15
Figure 7) Latency in milliseconds for various inference workloads.	15
Figure 8) Number of streams in the multistream scenario for various inference tasks.	16

Introduction

Companies are increasingly generating massive volumes of data at the network edge. To achieve maximum value from smart sensors and IoT data, organizations are looking for real-time event-streaming solution that enables edge computing. Computationally demanding jobs are therefore increasingly performed at the edge, outside of data centers. Artificial intelligence (AI) inference is one of the drivers of this trend. Edge servers provide sufficient computational power for these workloads, especially when using accelerators, but limited storage is often an issue, especially in multiserver environments. In this document we show how you can deploy a shared storage system in the edge environment and how it benefits AI inference workloads without imposing a performance penalty.

This document describes a reference architecture for AI inference at the edge. It combines multiple Lenovo ThinkSystem edge servers with a NetApp® AFF storage system to create an easy to deploy and manage solution. It is meant as a baseline guide for practical deployments in various situations, such as the factory floor, point of sale (POS), or Full Self-Driving (FSD) systems.

The document covers testing and validation of a compute and storage configuration consisting of Lenovo ThinkSystem SE350 Edge Server and an entry-level NetApp AFF storage system. This architecture provides an efficient and cost-effective solution for AI deployments while also providing comprehensive data services, integrated data protection, seamless scalability, and cloud connected data storage with NetApp ONTAP® data management software.

Target audience

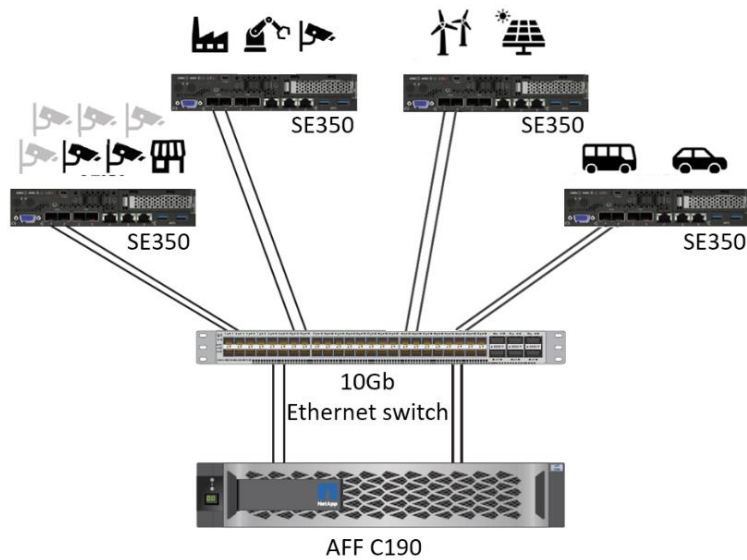
This document is intended for the following audiences:

- Business leaders and enterprise architects who want to productize AI at the edge.
- Data scientists, data engineers, AI/ machine learning (ML) researchers and developers of AI systems.
- Enterprise architects who design solutions for the development of AI/ML models and applications.
- Data scientists and AI engineers looking for efficient ways to deploy deep learning (DL) and ML models.
- Edge device managers and edge server administrators responsible for deployment and management of edge inferencing models.

Solution architecture

This Lenovo ThinkSystem server and NetApp ONTAP storage solution is designed to handle AI inferencing on large datasets using the processing power of GPUs alongside traditional CPUs. This validation demonstrates high performance and optimal data management with an architecture that uses either single or multiple Lenovo SR350 edge servers interconnected with a single NetApp AFF storage system, as shown in Figure 1.

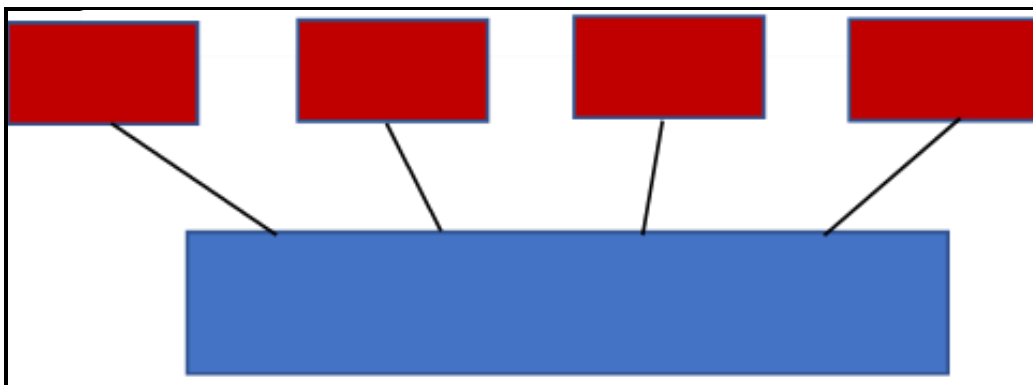
Figure 1) Physical architecture overview.



The logical architecture overview in Figure 2 shows the roles of the compute and storage elements in this architecture. Specifically, it shows the following:

- Edge compute devices performing inference on the data it receives from cameras, sensors, and so on.
- A shared storage element that serves multiple purposes:
 - Provides a central location for inference models and other data needed to perform the inference. Compute servers access the storage directly and use inference models across the network without the need to copy them locally.
 - Updated models are pushed here.
 - Archives input data that edge servers receive for later analysis. For example, if the edge devices are connected to cameras, the storage element keeps the videos captured by the cameras.

Figure 2) Logical architecture overview.



	Lenovo compute system	Edge devices performing Inference on inputs from cameras, sensors, etc.
	NetApp AFF storage system	Shared storage holding inference models and data from edge devices for later analysis.

This NetApp and Lenovo solution offers the following key benefits:

- GPU accelerated computing at the edge.
- Deployment of multiple edge servers backed and managed from a shared storage.
- Robust data protection to meet low recovery point objectives (RPOs) and recovery time objectives (RTOs) with no data loss.
- Optimized data management with NetApp Snapshot™ copies and clones to streamline development workflows.

How to use this architecture

This document validates the design and performance of the proposed architecture. However, we have not tested certain software-level pieces, such as container/workload/model management and data synchronization with cloud or data center on-premises, because they are specific to a deployment scenario. Here, multiple choices exist.

At the container management level, Kubernetes container management is a good choice and is well supported in either a fully upstream version (Canonical) or in a modified version suitable for enterprise deployments (Red Hat). The [NetApp AI Control Plane](#), which leverages NetApp Trident and the newly added [Data Science Toolkit](#), provides built-in traceability, data management functions, interfaces, and tools for data scientists and data engineers to integrate with NetApp storage. Kubeflow, the machine learning toolkit for Kubernetes, provides additional AI capabilities along with a support for model versioning and KFServing on several platforms such as TensorFlow Serving or NVIDIA Triton Inference Server. Another option is NVIDIA EGX platform, which provides workload management along with access to a catalog of GPU-enabled AI inference containers. However, these options might require significant effort and expertise to put them into production and might require the assistance of a third-party independent software vendor (ISV) or consultant.

Solution areas

The key benefit of AI inferencing and edge computing is the ability of devices to compute, process and analyze data with high level of quality without latency. There are far too many examples of edge computing use cases to describe in this document, but here are a few prominent ones:

Automobiles: Autonomous vehicles

The classic edge computing illustration is in the advanced driver-assistance systems (ADAS) in autonomous vehicles (AV). The AI in driverless cars must rapidly process a lot of data from cameras and sensors to be a successful safe driver. Taking too long to interpret between an object and a human can mean life or death, therefore being able to process that data as close to the vehicle as possible is crucial. In this case, one or more edge compute servers handles the input from cameras, RADAR, LiDAR, and other sensors, while shared storage holds inference models and stores input data from sensors.

Healthcare: Patient monitoring

One of the greatest impacts of AI and edge computing is its ability to enhance continuous monitoring of patients for chronic diseases both in at-home care and intensive care units (ICUs). Data from edge devices that monitor insulin levels, respiration, neurological activity, cardiac rhythm, and gastrointestinal functions require instantaneous analysis of data that must be acted on immediately because there is limited time to act to save someone's life.

Retail: Cashier-less payment

Edge computing can power AI and ML to help retailers reduce checkout time and increase foot traffic. Cashier-less systems support various components, such as the following:

- **Authentication and access.** Connecting the physical shopper to a validated account and permitting access to the retail space.
- **Inventory monitoring.** Using sensors, RFID tags, and computer vision systems to help confirm the selection or deselection of items by shoppers.

Here, each of the edge servers handle each checkout counter and the shared storage system serves as a central synchronization point.

Financial services: Human safety at kiosks and fraud prevention

Banking organizations are using AI and edge computing to innovate and create personalized banking experiences. Interactive kiosks using real-time data analytics and AI inferencing now enable ATMs to not only help customers withdraw money, but proactively monitor kiosks through the images captured from cameras to identify risk to human safety or fraudulent behavior. In this scenario, edge compute servers and shared storage systems are connected to interactive kiosks and cameras to help banks collect and process data with AI inference models.

Manufacturing: Industry 4.0

The fourth industrial revolution (Industry 4.0) has begun, along with emerging trends such as Smart Factory and 3D printing. To prepare for a data-led future, large-scale machine-to-machine (M2M) communication and IoT are integrated for increased automation without the need for human intervention. Manufacturing is already highly automated and adding AI features is a natural continuation of the long-term trend. AI enables automating operations that can be automated with the help of computer vision and other AI capabilities. You can automate quality control or tasks that rely on human vision or decision making to perform faster analyses of materials on assembly lines in factory floors to help manufacturing plants meet the required ISO standards of safety and quality management. Here, each compute edge server is connected to an array of sensors monitoring the manufacturing process and updated inference models are pushed to the shared storage, as needed.

Telecommunications: Rust detection, tower inspection and network optimization

The telecommunications industry uses computer vision and AI techniques to process images that automatically detect rust and identify cell towers that contain corrosion and, therefore, require further inspection. The use of drone images and AI models to identify distinct regions of a tower to analyze rust, surface cracks, and corrosion has increased in recent years. The demand continues to grow for AI technologies that enable telecommunication infrastructure and cell towers to be inspected efficiently, assessed regularly for degradation, and repaired promptly when required.

Additionally, another emerging use case in telecommunication is the use of AI and ML algorithms to predict data traffic patterns, detect 5G-capable devices, and automate and augment multiple-input and multiple-output (MIMO) energy management. MIMO hardware is used at radio towers to increase network capacity; however, this comes with additional energy costs. ML models for “MIMO sleep mode” deployed at cell sites can predict the efficient use of radios and help reduce energy consumption costs for mobile network operators (MNOs). AI inferencing and edge computing solutions help MNOs reduce the amount of data transmitted back-and-forth to data centers, lower their TCO, optimize network operations, and improve overall performance for end users.

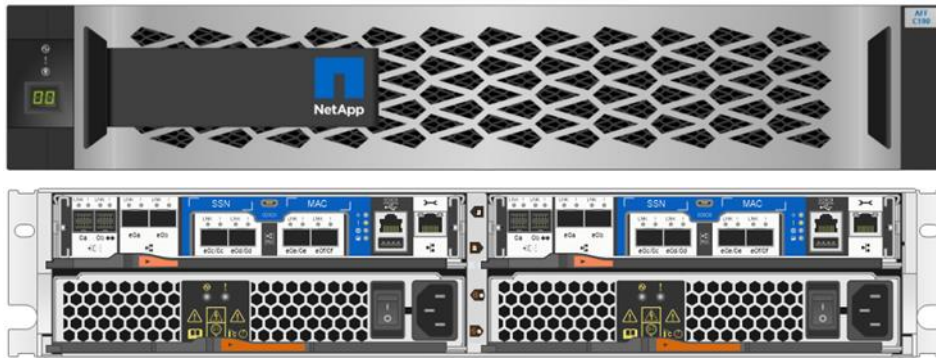
Technology overview

NetApp AFF systems

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

- Entry-level NetApp AFF storage systems are based on FAS2750 hardware and SSD flash media
- Two controllers in HA configuration

Figure 3) NetApp AFF C190 storage system.



NetApp entry-level AFF C190 storage systems support the following features:

- A maximum drive count of 24x 960-GB SSDs
- Two possible configurations:
 - Ethernet (10GbE): 4x 10GBASE-T (RJ-45) ports
 - Unified (16Gb FC or 10GbE): 4x unified target adapter 2 (UTA2) ports
- A maximum of 50.5TB effective capacity

Note: For NAS workloads, a single entry-level AFF C190 system supports throughput of 4.4GBps for sequential reads and 230K IOPS for small random reads at latencies of 1ms or less.

NetApp AFF A220

NetApp also offers other entry-level storage systems that provide higher performance and scalability for larger-scale deployments. For NAS workloads, a single entry-level AFF A220 system supports:

- Throughput of 6.2GBps for sequential reads
- 375K IOPS for small random reads at latencies of 1ms or less
- Maximum drive count of 144x 960GB, 3.8TB, or 7.6-TB SSDs
- AFF A220 scales to larger than 1PB of effective capacity

NetApp AFF A250

- Maximum effective capacity is 35PB with maximum scale out 2-24 nodes (12 HA pairs)
- Provides $\geq 45\%$ performance increase over AFF A220
- 440k IOPS random reads @1ms
- Built on the latest NetApp ONTAP release: ONTAP 9.8
- Leverages two 25Gb Ethernet for HA and cluster interconnect

Lenovo ThinkSystem DM 5100

The ThinkSystem DM5100F offers the following key features and benefits:

- Scales up to 737TB of raw storage capacity using 48x 15.36-TB drives.
- Up to 440K random read IOPS (8-KB blocks)
- Dual-active controllers (HA pair) with Automatic Load Balancing and failover
- Powered by ONTAP data management software.

NetApp ONTAP 9

ONTAP 9.8, 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. You can also move data freely to wherever it is needed - the edge, the core, or the cloud. ONTAP 9.8 includes numerous features that simplify data management, accelerate, and protect critical data, and enable next generation infrastructure capabilities across hybrid cloud architectures.

Simplify data management

Data management is crucial to enterprise IT operations so that appropriate resources are used for applications and 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 (AQoS).** Granular quality of service (QoS) controls help maintain performance levels for critical applications in highly shared environments.
- **NetApp FabricPool.** This feature provides automatic tiering of cold data to public and private cloud storage options, including Amazon Web Services (AWS), Azure, and NetApp StorageGRID® storage solution. For more information about FabricPool, see [TR-4598](#).

Accelerate and protect data

ONTAP 9 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 (NVE).** ONTAP offers native volume-level encryption with both onboard and External Key Management support.
- **Multitenancy and multifactor authentication.** ONTAP enables sharing of infrastructure resources with the highest levels of security.

Future-proof infrastructure

ONTAP 9 helps meet demanding and constantly changing business needs with the following features:

- **Seamless scaling and nondisruptive operations.** ONTAP supports the nondisruptive addition of capacity to existing controllers and 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 autonomous vehicles, smart cities, and Industry 4.0, by using the same infrastructure that supports existing enterprise apps.

NetApp Trident

[Trident](#) from NetApp is an open-source dynamic storage orchestrator for Docker and Kubernetes that simplifies the creation, management, and consumption of persistent storage. Trident, a Kubernetes-native application, runs directly within a Kubernetes cluster. Trident enables customers to seamlessly deploy DL container images onto NetApp storage and provides an enterprise-grade experience for AI container deployments. Kubernetes users (ML developers, data scientists, and so on.) can create, manage, and automate orchestration and cloning to take advantage of NetApp advanced data management capabilities powered by NetApp technology.

NetApp Cloud Sync

[Cloud Sync](#) is a NetApp service for rapid and secure data synchronization. Whether you need to transfer files between on-premises NFS or SMB file shares, NetApp StorageGRID, NetApp ONTAP S3, NetApp Cloud Volumes Service, Azure NetApp Files, Amazon Simple Storage Service (Amazon S3), Amazon Elastic File System (Amazon EFS), Azure Blob, Google Cloud Storage, or IBM Cloud Object Storage, Cloud Sync moves the files where you need them quickly and securely. After your data is transferred, it is fully available for use on both source and target. Cloud Sync continuously synchronizes the data, based on your predefined schedule, moving only the deltas, so time and money spent on data replication is minimized. Cloud Sync is a software as a service (SaaS) tool that is extremely simple to set up and use. Data transfers that are triggered by Cloud Sync are carried out by data brokers. You can deploy Cloud Sync data brokers in AWS, Azure, Google Cloud Platform, or on-premises.

Lenovo ThinkSystem servers

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:

- Highly scalable, modular designs to 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 proof of concept for solution development platforms that are ready to use and optimized for AI.

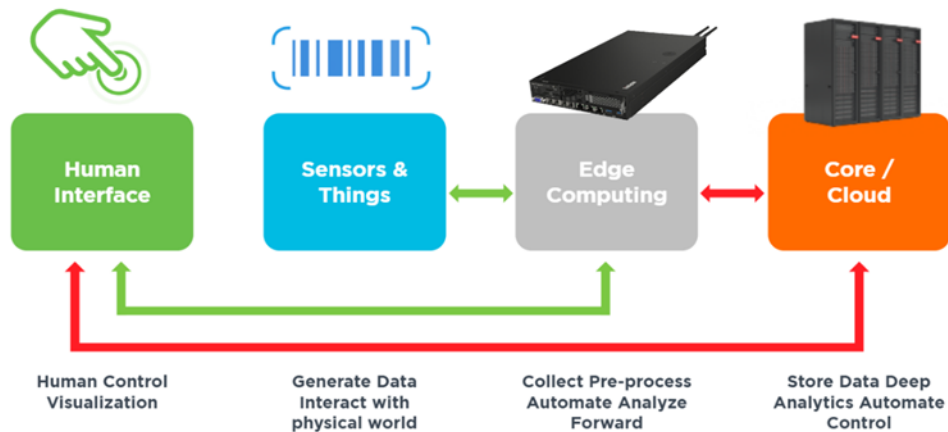
Lenovo ThinkSystem SE350 Edge Server

Edge computing allows data from IoT devices to be analyzed at the edge of the network before being sent to the data center or cloud. The Lenovo ThinkSystem SE350, as shown in Figure 4, is designed for the unique requirements for deployment at the edge, with a focus on flexibility, connectivity, security, and remote manageability in a compact ruggedized and environmentally hardened form factor.

Featuring the Intel Xeon D processor with the flexibility to support acceleration for edge AI workloads, the SE350 is purpose-built for addressing the challenge of server deployments in a variety of environments outside the data center.

Figure 4) Lenovo ThinkSystem SE350 Edge Server.





MLPerf

MLPerf is the industry-leading benchmark suite for evaluating AI performance. It covers many areas of applied AI including image classification, object detection, medical imaging, and natural language processing (NLP). In this validation, we used Inference v0.7 workloads, which is the latest iteration of the MLPerf Inference at the completion of this validation. The [MLPerf Inference v0.7](#) suite includes four new benchmarks for data center and edge systems:

- **BERT.** Bi-directional Encoder Representation from Transformers (BERT) fine-tuned for question answering by using the SQuAD dataset.
- **DLRM.** Deep Learning Recommendation Model (DLRM) is a personalization and recommendation model that is trained to optimize click-through rates (CTR).
- **3D U-Net.** 3D U-Net architecture is trained on the Brain Tumor Segmentation (BraTS) dataset.
- **RNN-T.** Recurrent Neural Network Transducer (RNN-T) is an automatic speech recognition (ASR) model that is trained on a subset of LibriSpeech.

MLPerf Inference results and code are publicly available and released under Apache license. MLPerf Inference has an Edge division, which supports the following scenarios:

- **Single stream.** This scenario mimics systems where responsiveness is a critical factor, such as offline AI queries performed on smartphones. Individual queries are sent to the system and response times are recorded. 90th percentile latency of all the responses is reported as the result.
- **Multistream.** This benchmark is for systems that process input from multiple sensors. During the test, queries are sent at a fixed time interval. A QoS constraint (maximum allowed latency) is imposed. The test reports the number of streams that the system can process while meeting the QoS constraint.
- **Offline.** This is the simplest scenario covering batch processing applications and the metric is throughput in samples per second. All data is available to the system and the benchmark measures the time it takes to process all the samples.

Lenovo has published MLPerf Inference scores for SE350 with T4, the server used in this document. See the results at <https://mlperf.org/inference-results-0-7/> in the “Edge, Closed Division” section in entry #0.7-145.

Test Plan

This document follows MLPerf Inference v0.7 [code](#) and [rules](#). We ran benchmarks designed for inference at the edge as defined in Table 1 and Table 2.

Table 1) [MLPerf Inference](#): Edge suite benchmarks.

Area	Task	Model	Dataset	QSL size	Quality	Multi-stream latency constraint
Vision	Image classification	Resnet50-v1.5	ImageNet (224x224)	1024	99% of FP32	50ms
Vision	Object detection (large)	SSD-ResNet34	COCO (1200x1200)	64	99% of FP32	66ms
Vision	Object detection (small)	SSD-MobileNets-v1	COCO (300x300)	256	99% of FP32	50ms
Vision	Medical image segmentation	3D UNET	BraTS 2019 (224x224x160)	16	99% and 99.9% of FP32	n/a
Speech	Speech-to-text	RNNT	Librispeech dev-clean	2513	99% of FP32	n/a
Language	Language processing	BERT	SQuAD v1.1	10833	99% of FP32	n/a

Table 2) Edge benchmark scenarios.

Area	Task	Scenarios
Vision	Image classification	Single stream, offline, multi-stream
Vision	Object detection (large)	Single stream, offline, multi-stream
Vision	Object detection (small)	Single stream, offline, multi-stream
Vision	Medical image segmentation	Single stream, offline
Speech	Speech-to-text	Single stream, offline
Language	Language processing	Single stream, offline

We performed these benchmarks using the networked storage architecture developed in this validation and compared results to those from local runs on the edge servers previously submitted to MLPerf. The comparison is to determine how much impact the shared storage has on inference performance.

Test configuration

Figure 5 shows the test configuration. We used the NetApp AFF C190 storage system and two Lenovo ThinkSystem SE350 servers (each with one NVIDIA T4 accelerator). These components are connected through a 10GbE network switch. The network storage holds validation/test datasets and pretrained models. The servers provide computational capability, and the storage is accessed over NFS protocol.

This section describes the tested configurations, the network infrastructure, the SE350 server, and the storage provisioning details.

Table 3) Base components for the solution architecture.

Solution components	Details
Lenovo ThinkSystem servers	<ul style="list-style-type: none"> 2x SE350 servers each with 1 NVIDIA T4 GPU card Each server contains 1 Intel Xeon D-2123IT CPU with 4 physical cores running at 2.20GHz and 128 GB RAM

Solution components	Details
Entry-level NetApp AFF storage system (HA pair)	<ul style="list-style-type: none"> • NetApp ONTAP 9 software • 24 x 960-GB SSDs • NFS protocol • 1 interface group per controller, with 4 logical IP addresses for mount points

Figure 5) Network topology of tested configuration.

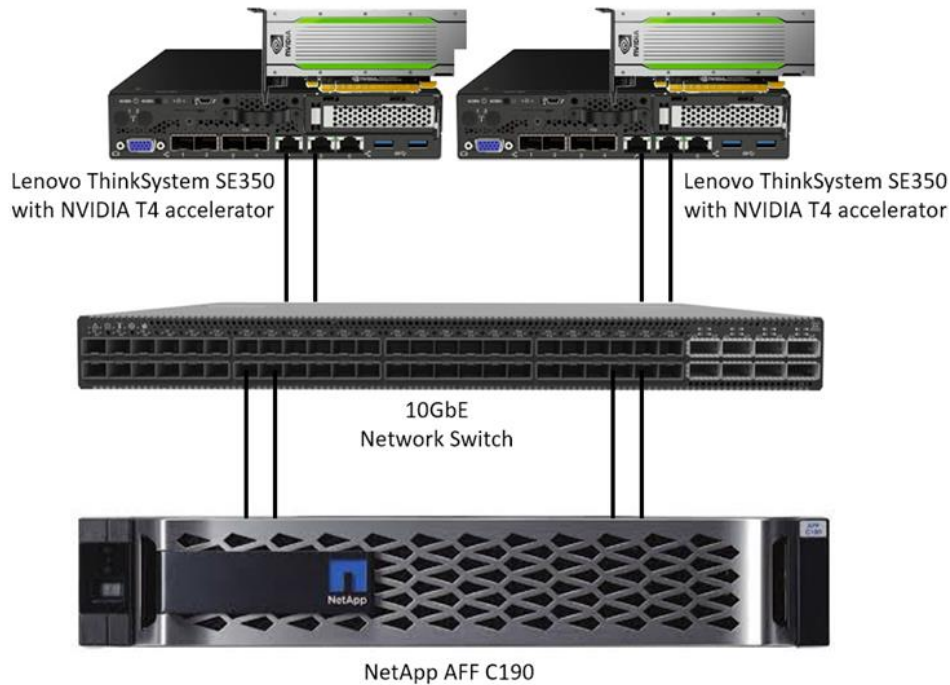


Table 4) Storage configuration: AFF C190 with 2RU, 24 drive slots.

Controller	Aggregate	FlexGroup volume	Aggregate size	Volume size	Operating system mount point
Controller1	Aggr1	/netapplenovo_AI_fg	8.42TiB	15TB	/netapp_lenovo_fg
Controller2	Aggr2		8.42TiB		

The /netappLenovo_AI_fg folder contains the datasets used for model validation.

Test procedure

We used the following test procedure in this validation.

Operating system setup

We used Ubuntu 18.04 with NVIDIA drivers and docker with support for NVIDIA GPUs.

AI inference setup

We used MLPerf [code](#) available as a part of the Lenovo submission to MLPerf Inference v0.7.

To set up the AI inference, follow these steps:

1. Download datasets that require registration: ImageNet 2012 Validation set, Criteo Terabyte dataset, and BraTS 2019 Training set, and then unzip the files.

2. Create a working directory with at least 1TB and define environmental variable `MLPERF_SCRATCH_PATH` referring to the directory.
You should share this directory on the shared storage for the network storage use case, or the local disk when testing with local data.
3. Run the `make prebuild` command, which builds and launches the docker container for the required inference tasks.
Note: The following commands are all executed from within the running docker container.
4. Download pretrained AI models for MLPerf Inference tasks: `make download_model`
5. Download additional datasets that are freely downloadable: `make download_data`
6. Pre-process the data: `make preprocess_data`
7. Run: `make build`.
8. Build inference engines optimized for the GPU in compute servers: `make generate_engines`
9. To run Inference workloads, use the following (one command):

```
make run_harness RUN_ARGS="--benchmarks=<BENCHMARKS> --scenarios=<SCENARIOS>"
```

AI inference runs

Three types of runs were executed:

- Single server AI Inference using local storage
- Single server AI Inference using network storage
- Multiserver AI Inference using network storage

Test results

A multitude of tests were run to evaluate the performance of the proposed architecture. There are six different workloads (image classification, object detection (small), object detection (large), medical imaging, speech-to-text, and natural language processing (NLP)), which you can run in three different scenarios: offline, single stream, and multistream.

Note: The last scenario is implemented only for image classification and object detection.

This gives 15 possible workloads, which were all tested under three different setups:

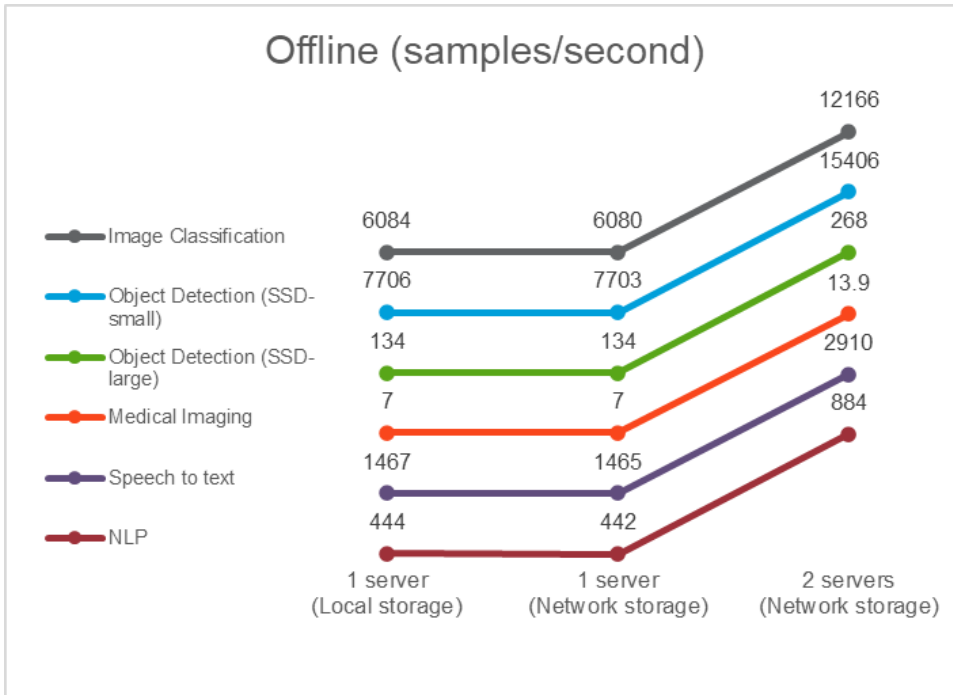
- Single server/local storage
- Single server/network storage
- Multiserver/ network storage

The results are described in the following sections.

AI inference in offline scenario

In this scenario, all the data was available to the server and the time it took to process all the samples was measured. We report bandwidths in samples per second as the results of the tests. When more than one compute server was used, we report total bandwidth summed over all the servers. The results for all three use cases are shown in Figure 6. For the two-server case, we report combined bandwidth from both servers.

Figure 6) Bandwidth in samples per second for various inference tasks.

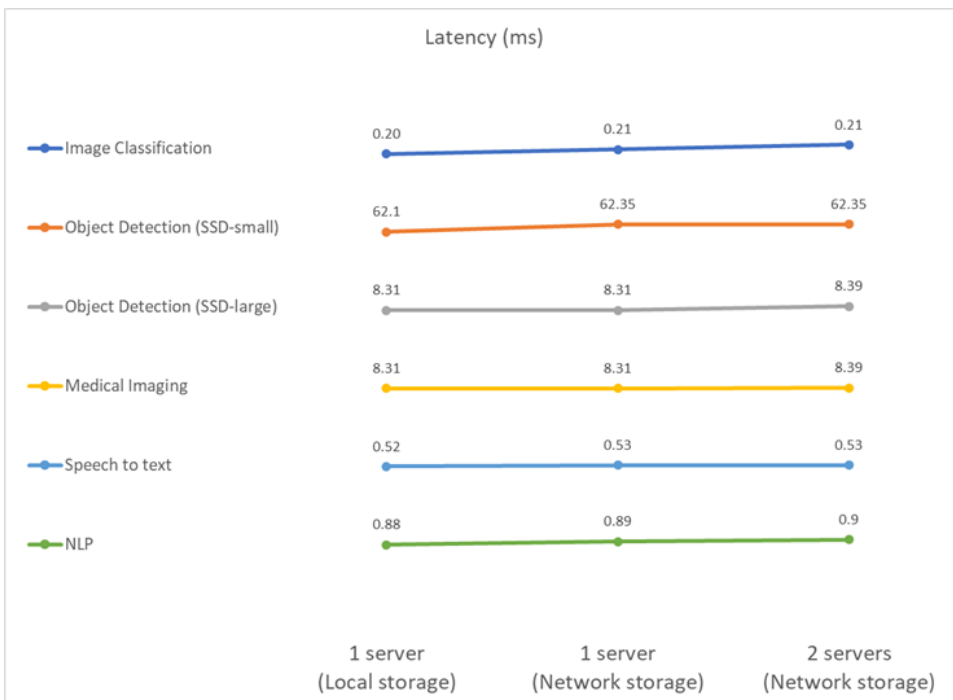


The results show that network storage does not negatively affect the performance—the change is minimal and for some tasks, none is found. When adding the second server, the total bandwidth either exactly doubles, or at worst, the change is less than 1%.

AI inference in a single stream scenario

This benchmark measures latency. For the multiple computational server case, we report the average latency. The results for the suite of tasks are given in Figure 7. For the two-server case, we report the average latency from both servers.

Figure 7) Latency in milliseconds for various inference workloads.

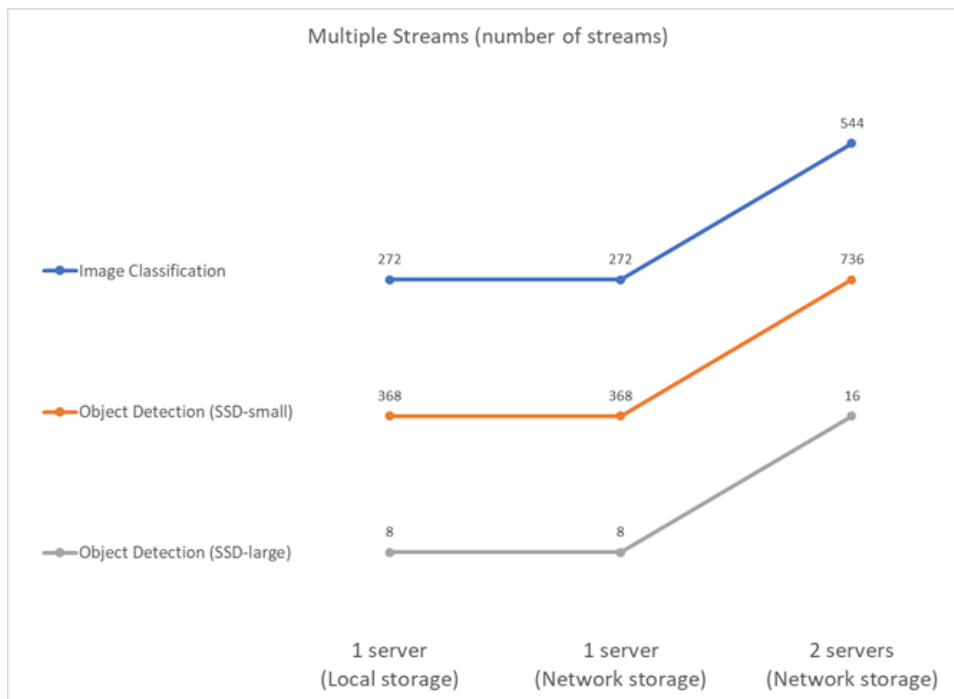


The results, again, show that the network storage is sufficient to handle the tasks. The difference between local and network storage in the one server case is minimal or none. Similarly, when two servers use the same storage, the latency on both servers stays the same or changes by a very small amount.

AI inference in multistream scenario

In this case, the result is the number of streams that the system can handle while satisfying the QoS constraint. Thus, the result is always an integer. For more than one server, we report total number of streams summed over all the servers. Not all workloads support this scenario, but we have executed those that do. The results of our tests are summarized in Figure 8. For the two-server case, we report the combined number of streams from both servers.

Figure 8) Number of streams in the multistream scenario for various inference tasks.



The results show perfect performance of the setup—local and networking storage give the same results and adding the second server doubles the number of streams the proposed setup can handle.

Architecture sizing options

You can adjust the setup used for the validation to fit other use cases.

Compute server

We used an Intel Xeon D-2123IT CPU, which is the lowest level of CPU supported in SE350, with four physical cores and 60W TDP. While the server does not support replacing CPUs, it can be ordered with a more powerful CPU. The top CPU supported is Intel Xeon D-2183IT with 16 cores, 100W running at 2.20GHz. This will increase CPU computational capability considerably. While CPU was not a bottleneck for running the inference workloads themselves, it will help with data processing and other tasks related to inference. At present, NVIDIA T4 is the only GPU available for edge use cases, so, currently, there is no ability to upgrade/downgrade the GPU.

Shared storage

For testing and validation, the NetApp AFF C190 system, which has maximum storage capacity of 50.5TB, throughput of 4.4GBps for sequential reads, and 230K IOPS for small random reads, was used for the purpose of this document and is proven to be well suited for edge inference workloads.

However, if you require more storage capacity or faster networking speeds, you should use the NetApp AFF A220, Lenovo DM5100F, or [NetApp AFF A250](#) storage systems.

Conclusion

AI driven automation and edge computing is a leading approach to help business organizations achieve digital transformation and maximize operational efficiency and safety. With edge computing, data is processed much faster because it does not have to travel to and from a data center. Therefore, the cost associated with sending data back and forth to data centers or the cloud is diminished. Lower latency and increased speed can be beneficial when businesses must make decisions in near-real time using AI inferencing models deployed at the edge.

NetApp storage systems deliver the same or better performance as local SSD storage and offer the following benefits to data scientists, data engineers, AI/ML developers, and business or 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 usage.
- Streamlined development and deployment workflows using integrated Snapshot copies and clones for instantaneous and space-efficient user workspaces, integrated version control, and automated deployment.
- Enterprise-grade data protection for disaster recovery and business continuity.

The NetApp and Lenovo solution presented in this document is a flexible, scale-out architecture that is ideal for enterprise-grade AI inference deployments at the edge.

Acknowledgments

- J.J. Falkanger, Sr. Manager, HPC&AI Solutions, Lenovo
- Dave Arnett, Technical Marketing Engineer, NetApp

Where to find additional information

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

- NetApp AFF A-Series arrays product page
<https://www.netapp.com/data-storage/aff-a-series/>
- NetApp ONTAP data management software—ONTAP 9 information library
<http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286>
- NetApp Persistent Storage for Containers—NetApp Trident
<https://netapp.io/persistent-storage-provisioner-for-kubernetes/>
- MLPerf
<https://mlperf.org>
<http://www.image-net.org/>
<https://mlperf.org/inference-overview>
- NetApp Cloud Sync
https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works
- TensorFlow benchmark
<https://github.com/tensorflow/benchmarks>
- Lenovo ThinkSystem SE350 Edge Server
<https://lenovopress.com/lp1168-thinksystem-se350-edge-server>

- [Lenovo ThinkSystem DM5100F Unified Flash Storage Array](https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array)
<https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array>

Version history

Version	Date	Document Version History
Version 1.0	March 2021	Initial release

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

Copyright Information

Copyright © 2021 NetApp, Inc. All rights reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR 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.

Lenovo, the Lenovo logo, and ThinkServer are trademarks or registered trademarks of Lenovo. Other company, product, and service names may be trademarks or service marks of others.

TR-4886-0321