KI-Inferenz am Edge - NetApp mit Lenovo ThinkSystem – Lösungsdesign

NetApp Solutions

NetApp
April 18, 2024

This PDF was generated from https://docs.netapp.com/de-de/netapp-solutions/ai/ai-edge-introduction.html on April 18, 2024. Always check docs.netapp.com for the latest.

Inhalt

KI-Inferenz am Edge - NetApp mit Lenovo ThinkSystem – Lösungsdesign . . . 1
TR-4886: KI-Inferenzierung am Edge - NetApp mit Lenovo ThinkSystem - Solution Design . . . 1
Technologieüberblick . . . 6
Testplan . . . 13
Testkonfiguration . . . 14
Testverfahren . . . 16
Testergebnisse . . . 17
Optionen zur Dimensionierung der Architektur . . . 22
Schlussfolgerung . . . 23

KI-Inferenz am Edge - NetApp mit Lenovo ThinkSystem – Lösungsdesign

TR-4886: KI-Inferenzierung am Edge - NetApp mit Lenovo ThinkSystem - Solution Design

Sathish Thyagarajan, NetApp Miroslav Hodak, Lenovo

Zusammenfassung

Verschiedene neue Applikationsszenarien, etwa durch erweiterte Treiberassistenzsysteme (ADAS), Industrie 4.0, Smart Cities und Internet of Things (IoT), erfordern die Verarbeitung von kontinuierlichen Datenströmen bei nahezu null Latenz. In diesem Dokument wird eine Computing- und Storage-Architektur beschrieben, um GPU-basierte Inferenz (KI) auf NetApp Storage-Controllern und Lenovo ThinkSystem Servern in einer Edge-Umgebung zu implementieren, die diese Anforderungen erfüllt. Dieses Dokument bietet auch Performance-Daten für die branchenübliche MLPerf Inferenz-Benchmark. Hierbei werden verschiedene Inferenzaufgaben auf Edge-Servern mit NVIDIA T4-GPUs evaluiert. Wir untersuchen die Performance von Offline-, Single-Stream- und Multistream-Inferenzszenarien und zeigen, dass die Architektur mit einem kostengünstigen Shared-Networked Storage-System eine hohe Performance aufweist und einen zentralen Datenstandort- und Modellmanagement für mehrere Edge-Server ermöglicht.

Einführung

Unternehmen generieren immer mehr riesige Datenmengen am Netzwerkrand. Um Smart-Sensoren und IoT-Daten optimal zu nutzen, suchen Unternehmen eine Echtzeit-Ereignis-Streaming-Lösung, die Edge-Computing ermöglicht. Rechenanspruchsvolle Jobs werden daher zunehmend außerhalb des Datacenters am Edge ausgeführt. KI-Inferenz ist einer der Faktoren für diesen Trend. Edge-Server liefern für diese Workloads genügend Rechenleistung, insbesondere wenn Beschleuniger zum Einsatz kommen. Doch häufig stellt begrenzter Storage ein Problem dar, insbesondere in Umgebungen mit mehreren Servern. In diesem Dokument zeigen wir, wie ein Shared-Storage-System in der Edge-Umgebung implementiert werden kann und wie es von KI-Inferenz-Workloads profitiert, ohne dass sich die Performance beeinträchtigt.

In diesem Dokument wird eine Referenzarchitektur für KI-Inferenz im Edge-Bereich beschrieben. Es kombiniert mehrere Lenovo ThinkSystem Edge Server mit einem NetApp Storage-System zu einer Lösung, die einfach zu implementieren und zu managen ist. Es soll als Basishandbuch für den praktischen Einsatz in verschiedenen Situationen dienen, wie zum Beispiel der Fabrikhalle mit mehreren Kameras und industriellen Sensoren, Point-of-Sale-Systeme (POS) im Einzelhandel oder Full Self-Driving (FSD)-Systeme, die visuelle Anomalien in autonomen Fahrzeugen identifizieren.

Dieses Dokument behandelt Tests und Validierungen einer Computing- und Storage-Konfiguration, die aus Lenovo ThinkSystem SE350 Edge Server sowie einem Storage-System der Einstiegsklasse von NetApp AFF und EF-Series besteht. Die Referenzarchitekturen sind eine effiziente und kostengünstige Lösung für KI-Implementierungen. Gleichzeitig bieten sie umfassende Datenservices, integrierte Datensicherung, nahtlose Skalierbarkeit und Cloud-vernetzten Storage mit NetApp ONTAP und der Datenmanagement-Software NetApp SANtricity.

Zielgruppe

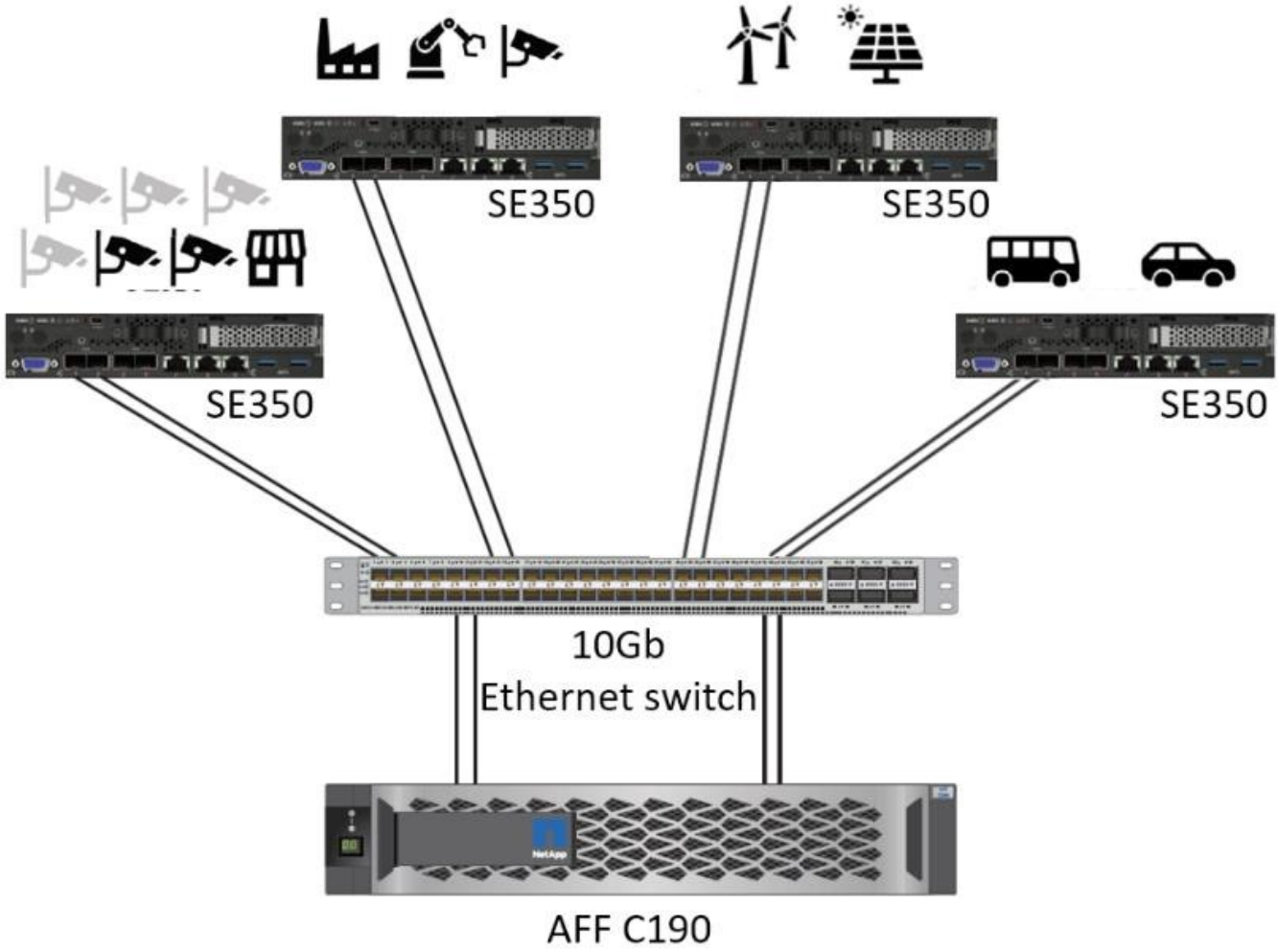
Dieses Dokument richtet sich an folgende Zielgruppen:

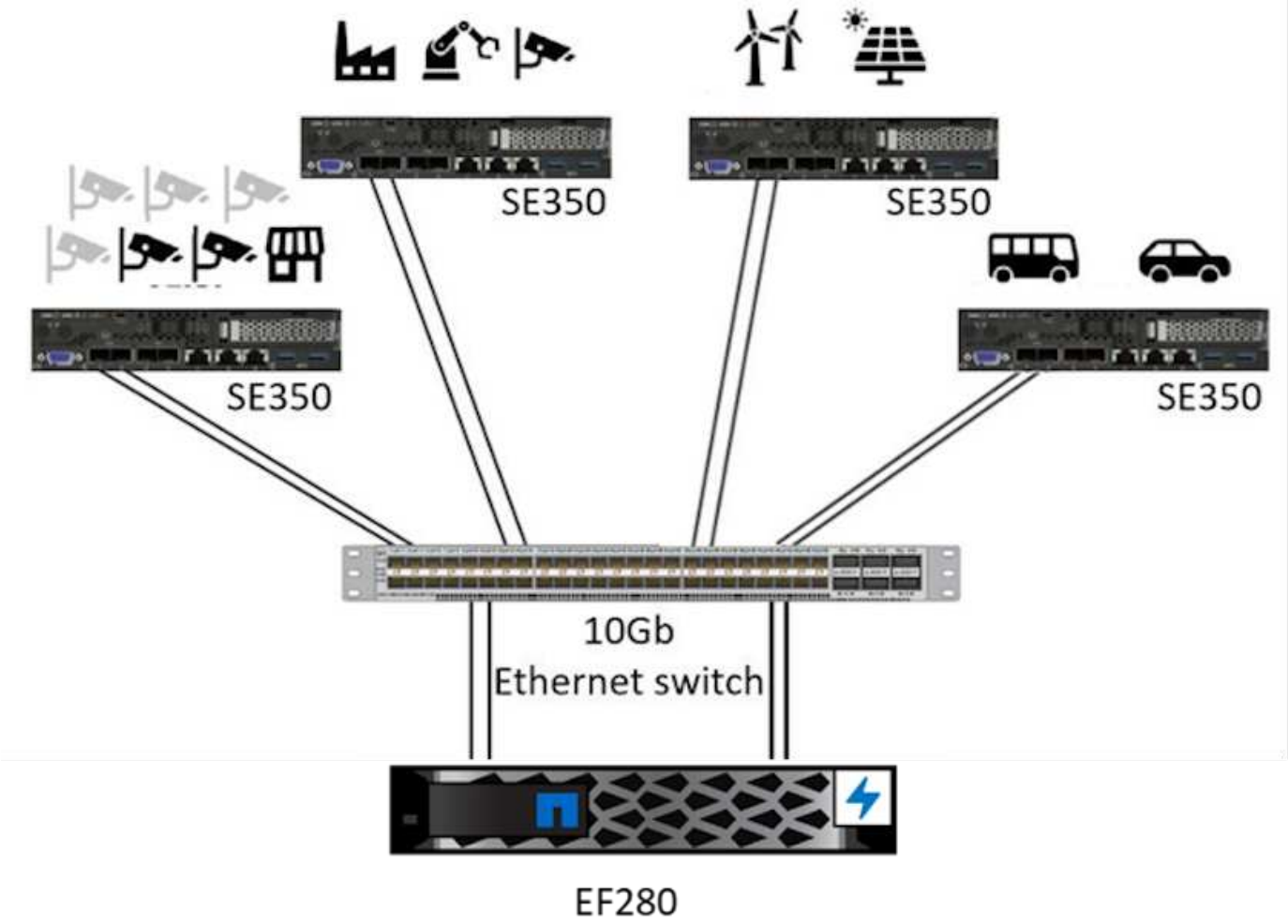
- Führungskräfte und Enterprise-Architekten, die KI im Edge-Bereich productieren möchten

- Data Scientists, Data Engineers, KI/Machine Learning-Forscher (ML) und Entwickler von KI-Systemen.
- Enterprise-Architekten, die Lösungen für die Entwicklung von KI/ML-Modellen und -Applikationen entwickeln
- Data Scientists und KI-Ingenieure, die nach effizienten Möglichkeiten suchen, Deep-Learning- (DL) und ML-Modelle zu implementieren.
- Edge-Geräte-Manager und Edge-Serveradministratoren, die für die Implementierung und das Management von Edge-Inferenzmodellen verantwortlich sind

Lösungsarchitektur

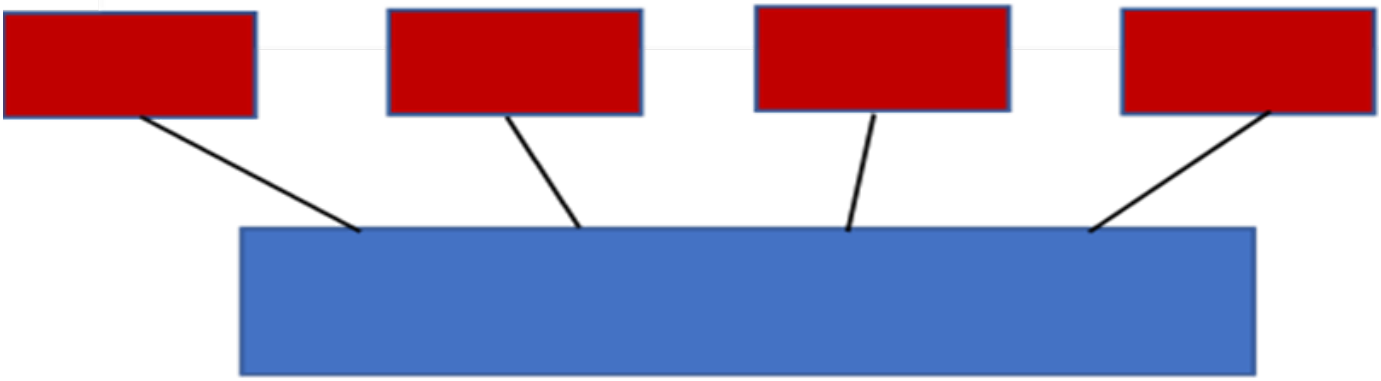
Dieser Lenovo ThinkSystem Server und die NetApp ONTAP oder NetApp SANtricity Storage-Lösung wurden entwickelt, um KI-Inferenz auf großen Datensätzen zu verarbeiten. Dabei wird die Verarbeitungsleistung von GPUs neben herkömmlichen CPUs verwendet. Diese Validierung belegt eine hohe Performance und ein optimales Datenmanagement mit einer Architektur, die entweder einzelne oder mehrere Lenovo SR350 Edge-Server verwendet, die mit einem einzelnen NetApp AFF-Storage-System verbunden sind, wie in den folgenden beiden Abbildungen dargestellt.





Die logische Architekturübersicht in der folgenden Abbildung zeigt die Rollen der Computing- und Storage-Elemente in dieser Architektur. Sie zeigt im Detail Folgendes:

- Edge-Computing-Geräte führen Inferenz für die Daten durch, die von Kameras, Sensoren usw. empfangen werden.
- Ein Shared-Storage-Element, das mehrere Zwecke erfüllt:
  - Ermöglicht den Inferenz-Modelle und andere Daten, die zu Vorgänge erforderlich sind. Compute-Server greifen direkt auf den Storage zu und nutzen Inferenzmodelle im gesamten Netzwerk, ohne dass diese lokal kopiert werden müssen.
  - Aktualisierte Modelle werden hier geschoben.
  - Archivierung von Eingabedaten, die Edge-Server für eine spätere Analyse erhalten Wenn beispielsweise die Edge-Geräte an Kameras angeschlossen sind, speichert das Speicherelement die von den Kameras erfassten Videos.



Rot	Blau
Lenovo Computing-System	NetApp AFF Storage-System
Edge-Geräte, die Inferenz auf Eingänge von Kameras, Sensoren usw. durchführen.	Inferenzmodelle und Daten von Edge-Geräten für Shared Storage werden zur späteren Analyse gespeichert.

Diese Lösung von NetApp und Lenovo bietet folgende wichtige Vorteile:

- GPU-beschleunigtes Computing im Edge-Bereich.
- Implementierung mehrerer Edge-Server, unterstützt und über einen Shared Storage gemanagt
- Robuste Datensicherung zur Einhaltung niedriger Recovery-Zeitpunkte (RPOs) und Recovery-Zeiten (RTOs) ohne Datenverluste
- Optimiertes Datenmanagement mit NetApp Snapshot Kopien und Klonen zur Optimierung von Entwicklungs-Workflows

Wie diese Architektur zu nutzen

Dieses Dokument validiert das Design und die Performance der vorgeschlagenen Architektur. Bestimmte Software-Level-Komponenten wie Container-, Workload- oder Modellmanagement und Datensynchronisation mit On-Premises-Cloud- oder Datacentern haben wir jedoch noch nicht getestet, da sie für ein Implementierungsszenario spezifisch sind. Hier gibt es mehrere Möglichkeiten.

Auf der Ebene der Containerverwaltung ist die Kubernetes-Containerverwaltung eine gute Wahl und wird sowohl in einer vollwertigen Version (Canonical) als auch in einer modifizierten Version unterstützt, die für Enterprise-Bereitstellungen (Red hat) geeignet ist. Der "NetApp AI Control Plane" Welches NetApp Trident und die neu hinzugefügte nutzt "NetApp DataOps Toolkit" Bietet integrierte Rückverfolgbarkeit, Datenmanagementfunktionen, Schnittstellen und Tools für Data Scientists und Data Engineers zur Integration in NetApp Storage. Kubeflow, das ML-Toolkit für Kubernetes, bietet zusätzliche KI-Funktionen sowie Unterstützung für Modellversionierung und KFServing auf mehreren Plattformen wie TensorFlow Serving oder NVIDIA Triton Inferenz Server. Eine weitere Option ist die NVIDIA EGX Plattform, die Workload-Management zusammen mit Zugriff auf einen Katalog von GPU-fähigen KI-Inferenz-Containern ermöglicht. Für diese Optionen ist jedoch unter Umständen ein erheblicher Aufwand und große Fachkenntnisse für die Produktion erforderlich, wobei unter Umständen die Unterstützung eines unabhängigen Softwareanbieters (ISV) oder eines Beraters durch Drittanbieter erforderlich ist.

Lösungsbereiche

Der Hauptvorteil von KI-Inferenz und Edge-Computing ist die Möglichkeit von Geräten, Daten zu berechnen, zu verarbeiten und zu analysieren – ohne Latenz. Es gibt viel zu viele Beispiele für Edge-Computing-

Anwendungsfälle, die in diesem Dokument beschrieben werden. Es gibt jedoch einige prominente Beispiele:

Automobile: Autonome Fahrzeuge

Die klassische Edge Computing-Abbildung zeigt sich in den fortschrittlichen Fahrerassistenzsystemen (ADAS) in autonomen Fahrzeugen (AV). Die KI in selbstfahrenden Autos muss schnell viele Daten von Kameras und Sensoren verarbeiten, um ein erfolgreicher und sicherer Treiber zu sein. Es dauert zu lange, zwischen einem Objekt und einem Menschen zu interpretieren, kann Leben oder Tod bedeuten, daher ist es entscheidend, dass Daten so nah wie möglich am Fahrzeug verarbeitet werden können. In diesem Fall übernimmt ein oder mehrere Edge-Computing-Server die Eingaben von Kameras, RADAR, LiDAR und anderen Sensoren, während Shared Storage Inferenzmodelle enthält und die Eingangsdaten von Sensoren speichert.

Gesundheitswesen: Patientenüberwachung

Einer der größten Auswirkungen von KI und Edge-Computing ist die Fähigkeit, die kontinuierliche Überwachung von Patienten auf chronische Krankheiten zu verbessern, und zwar sowohl in der häuslichen Pflege als auch in Intensivstationen. Daten von Edge-Geräten, die Insulinspiegel, Atmung, neurologische Aktivität, Herzrhythmus und gastrointestinale Funktionen überwachen, erfordern eine sofortige Analyse von Daten, die sofort gehandelt werden müssen, weil es begrenzte Zeit zum Handeln gibt, um das Leben eines Menschen zu retten.

Einzelhandel: Kassierer-lose Zahlung

Edge-Computing unterstützt KI und ML und hilft Einzelhändlern dabei, die Kasse zu verkürzen und den Platzbedarf zu steigern. Kassierer-lose Systeme unterstützen verschiedene Komponenten, wie z. B.:

- Und -Zugriff. Verbinden des physischen Käuferkontos mit einem validierten Konto und Zulassen des Zugangs zu den Einzelhandelsflächen.
- Bestandsüberwachung: Mithilfe von Sensoren, RFID-Tags und Computer-Vision-Systemen können Sie die Auswahl oder das Abwählen von Artikeln durch Kunden bestätigen.

  Hier sind alle Edge-Server für jeden Kassenzähler zuständig, und das gemeinsam genutzte Speichersystem dient als zentraler Synchronisationspunkt.

Finanzdienstleistungen: Menschliche Sicherheit an Kiosken und Betrugsprävention

Bankwesen setzen KI und Edge-Computing ein, um Innovationen voranzutreiben und ein personalisiertes Banking zu ermöglichen. Mithilfe von interaktiven Kiosks, die Datenanalysen in Echtzeit und KI-Inferenz ermöglichen, sind Geldautomaten nun nicht nur in der Lage, Geld zu ziehen, sondern überwachen Kiosksysteme proaktiv über die von Kameras erfassten Bilder, um Risiken für die menschliche Sicherheit oder das betrügerische Verhalten zu erkennen. In diesem Szenario sind Edge-Computing-Server und Shared-Storage-Systeme mit interaktiven Kiosken und Kameras verbunden. So können Banken Daten mit KI-Inferenzmodellen erfassen und verarbeiten.

Fertigung: Industrie 4.0

Die vierte industrielle Revolution (Industrie 4.0) hat begonnen, zusammen mit neuen Trends wie Smart Factory und 3D-Druck. Für die Vorbereitung auf eine datengesteuerte Zukunft werden umfassende Kommunikation zwischen Machine und Machine (M2M) und IoT integriert, um die Automatisierung zu steigern, ohne menschliches Eingreifen erforderlich zu sein. Die Fertigung ist bereits hochautomatisiert und das Hinzufügen von KI-Funktionen stellt eine natürliche Fortsetzung des langfristigen Trends dar. KI ermöglicht automatische Vorgänge, die mit Hilfe der Computervision und anderer KI-Funktionen automatisiert werden können. Sie können die Qualitätskontrolle oder Aufgaben automatisieren, die auf menschliche Vision oder Entscheidungsfindung beruhen, um schnellere Analysen von Materialien auf Montagelinien in Fabrikböden

durchzuführen, um Fertigungsstätten dabei zu unterstützen, die geforderten ISO-Standards des Sicherheits- und Qualitätsmanagements zu erfüllen. Jeder Compute-Edge-Server ist hier mit einer Reihe von Sensoren verbunden, die den Fertigungsprozess überwachen und nach Bedarf aktualisierte Inferenzmodelle in den Shared Storage verschieben.

Telekommunikation: Rosterkennung, Turmüberwachung und Netzwerkoptimierung

Die Telekommunikationsbranche nutzt Computer Vision und KI-Techniken, um Bilder zu verarbeiten, die automatisch Rost erkennen und Zelltürme identifizieren, die Korrosion enthalten, und daher weitere Inspektionen erfordern. Die Verwendung von Drohnenbildern und KI-Modellen zur Identifizierung verschiedener Bereiche eines Turms zur Analyse von Rost, Oberflächenrissen und Korrosion hat in den letzten Jahren zugenommen. Die Nachfrage nach KI-Technologien wächst weiter, mit denen Telekommunikationsinfrastruktur und Zelltürme effizient untersucht, regelmäßig auf Abbaumaßnahmen bewertet und bei Bedarf zeitnah repariert werden können.

Ein weiterer neuer Anwendungsfall in der Telekommunikation ist der Einsatz von KI- und ML-Algorithmen zur Vorhersage von Datenverkehrsmustern, zum Erkennen von 5G-fähigen Geräten sowie zur Automatisierung und Erweiterung des Energiemanagements bei mehreren Eingängen und Mehrfachausgängen (MIMO). MIMO-Hardware wird in Funktürmen verwendet, um die Netzwerkkapazität zu erhöhen; dies ist jedoch mit zusätzlichen Energiekosten verbunden. ML-Modelle für den „MIMO-Schlafmodus“, der an Zellstandorten eingesetzt wird, können die effiziente Nutzung von Funkgeräten vorhersagen und dazu beitragen, die Energiekosten für Mobilfunknetzbetreiber (MNOs) zu senken. Mithilfe von KI-Inferenz- und Edge-Computing-Lösungen können MNOs die Menge der zu Datacentern übertragenen Daten reduzieren, ihre TCO senken, den Netzwerkbetrieb optimieren und die allgemeine Performance für Endbenutzer verbessern.

Technologieüberblick

In diesem Abschnitt wird die technologische Grundlage dieser KI-Lösung beschrieben.

NetApp AFF Systeme

Mit den hochmodernen NetApp AFF-Storage-Systemen können KI-Inferenz-Implementierungen im Edge-Bereich erfolgen. Die Enterprise-Storage-Anforderungen werden mit branchenführender Performance, überlegener Flexibilität, Cloud-Integration und erstklassigem Datenmanagement erfüllt. NetApp AFF Systeme wurden speziell für Flash entwickelt. Sie helfen geschäftskritische Daten zu beschleunigen, zu managen und zu sichern.

- Die NetApp AFF Storage-Systeme der Einstiegsklasse basieren auf der FAS2750 Hardware und SSD-Flash-Medien
- Zwei Controller in HA-Konfiguration

NetApp Storage-Systeme der Einstiegsklasse AFF C190 unterstützen die folgenden Funktionen:

- 1 maximale Laufwerksanzahl: 24 x 960-GB-SSDs
- Zwei mögliche Konfigurationen:
  - Ethernet (10 GbE): 4 x 10GBASE-T-Ports (RJ-45
  - Unified (16 GB FC oder 10 GbE): 4 x Unified Target Adapter 2 (UTA2)-Ports
- Maximal 50,5 TB effektive Kapazität

> Bei NAS-Workloads unterstützt ein einzelnes AFF C190 Einstiegssystem einen Durchsatz von 4,4 GB/s bei sequenziellem Lesen und 230.000 IOPS bei kleinen zufälligen Lesevorgängen mit einer Latenz von 1 ms oder weniger.

NetApp AFF A220

NetApp bietet außerdem andere Storage-Systeme der Einstiegsklasse an, die eine höhere Performance und Skalierbarkeit für größere Implementierungen bieten. Für NAS-Workloads unterstützt ein einzelnes AFF A220 Einstiegssystem:

- Durchsatz von 6,2 GB/s bei sequenziellem Lesen
- 375.000 IOPS für kleine zufällige Lesevorgänge bei einer Latenz von maximal 1 ms
- Maximale Laufwerksanzahl: 144 x 960 GB, 3,8 TB oder 7,6 TB SSDs
- AFF A220 kann auf eine effektive Kapazität von mehr als 1 PB skaliert werden

NetApp AFF A250

- Maximale effektive Kapazität: 35 PB mit maximal horizontal skalierbarer 2-24 Nodes (12 HA-Paare)
- Performance-Steigerung um ≥ 45 % im Vergleich zu AFF A220
- 440.000 IOPS zufälliger Lesezugriff @1 ms
- Basiert auf der neuesten ONTAP-Version von NetApp: ONTAP 9.8
- Zwei 25-GB-Ethernet für HA und Cluster Interconnect

Die NetApp E-Series EF Systeme

Die EF-Series ist eine Produktfamilie aus All-Flash-SAN-Storage-Arrays der Einstiegsklasse und für den Midrange-Bereich, die den Zugriff auf Ihre Daten beschleunigen und mit der NetApp SANtricity Software den maximalen Nutzen aus ihren Daten ziehen können. Diese Systeme bieten sowohl SAS- als auch NVMe-Flash-Storage und bieten kostengünstige bis extreme IOPS, Reaktionszeiten von unter 100 Mikrosekunden und eine Bandbreite von bis zu 44 GB/s. Dadurch eignen sie sich ideal für heterogene Workloads und anspruchsvolle Applikationen wie KI-Inferenz und High-Performance-Computing (HPC).

Die folgende Abbildung zeigt das NetApp EF280 Storage-System.

NetApp EF280

- Unterstützung von 32 GB/16 GB FC, 25 GB/10 GB iSCSI und 12 GB SAS
- Die maximale effektive Kapazität beträgt 96 Laufwerke mit einer Gesamtleistung von 1,5 PB
- Durchsatz von 10 GB/s (sequenzielle Lesevorgänge)
- 300.000 IOPS (zufällige Lesevorgänge)
- Die NetApp EF280 ist das kostengünstigste All-Flash-Array (AFA) im NetApp Portfolio

NetApp EF300

- 24 x NVMe-SSD-Laufwerke für eine Gesamtkapazität von 367 TB
- Erweiterungsoptionen mit insgesamt 240x NL-SAS HDDs, 96 x SAS SSDs oder einer Kombination aus diesen Optionen
- 100-GB-NVMe/IB, NVMe/RoCE, iSER/IB und SRP/IB
- 32-GB-NVME/FC, FCP
- 25 GB iSCSI
- 20 Gbit/s (sequenzielle Lesevorgänge)
- 670.000 IOPS (wahlfreie Lesevorgänge)

Weitere Informationen finden Sie im "NetApp EF-Series All-Flash-Arrays der NetApp EF-Series, F300, EF570 und EF280 Datenblatt".

NetApp ONTAP 9

ONTAP 9.8.1, die neueste Generation der Storage-Managementsoftware von NetApp, ermöglicht Unternehmen, ihre Infrastruktur zu modernisieren und auf ein Cloud-fähiges Datacenter zu wechseln. Dank der erstklassigen Datenmanagementfunktionen lassen sich mit ONTAP sämtliche Daten mit einem einzigen Toolset managen und schützen, ganz gleich, wo sich diese Daten befinden. Zudem können Sie die Daten problemlos dorthin verschieben, wo sie benötigt werden: Zwischen Edge, Core und Cloud. ONTAP 9.8.1 umfasst zahlreiche Funktionen, die das Datenmanagement vereinfachen, geschäftskritische Daten beschleunigen und schützen und Infrastrukturfunktionen der nächsten Generation über Hybrid-Cloud-Architekturen hinweg ermöglichen.

Vereinfachtes Datenmanagement

Für den Enterprise-IT-Betrieb spielt Datenmanagement eine zentrale Rolle, da es Applikationen und Datensätzen die richtigen Ressourcen zuweist. ONTAP umfasst die folgenden Funktionen, um den Betrieb zu optimieren und zu vereinfachen und damit die Gesamtbetriebskosten zu senken:

- **Inline-Data-Compaction und erweiterte Deduplizierung.** Data-Compaction reduziert den ungenutzten Speicherplatz in Storage-Blöcken, während Deduplizierung die effektive Kapazität deutlich steigert. Dies gilt für lokal gespeicherte Daten und für Daten-Tiering in die Cloud.
- **Minimum, Maximum und Adaptive Quality of Service (AQoS).** durch granulare QoS-Kontrollen (Quality of Service) können Performance-Level für kritische Applikationen in Umgebungen mit hohen gemeinsam genutzten Ressourcen aufrechterhalten werden.
- **NetApp FabricPool.** Diese Funktion bietet automatisches Tiering von weniger häufig benötigten Daten in Public- und Private-Cloud-Storage-Optionen, einschließlich Amazon Web Services (AWS), Azure und NetApp StorageGRID Storage-Lösung. Weitere Informationen zu FabricPool finden Sie unter "TR-4598".

Beschleunigung und Sicherung von Daten

ONTAP 9 bietet überdurchschnittliche Performance und Datensicherung, erweitert diese Funktionen auf folgende Weise:

- **Performance und geringere Latenz.** ONTAP bietet den höchstmöglichen Durchsatz bei geringstmöglicher Latenz.
- **Datensicherung.** ONTAP bietet integrierte Datensicherungsfunktionen mit einem einheitlichen Management über alle Plattformen hinweg.
- **NetApp Volume Encryption (NVE).** ONTAP bietet native Verschlüsselung auf Volume-Ebene mit integrierter und externer Unterstützung für das Verschlüsselungsmanagement.
- **Mandantenfähigkeit und Multi-Faktor-Authentifizierung.** ONTAP ermöglicht die gemeinsame Nutzung von Infrastrukturressourcen mit einem Höchstmaß an Sicherheit.

Zukunftssichere Infrastruktur

ONTAP 9 bietet die folgenden Funktionen, um Unternehmen auf ständig wechselnde Geschäftsanforderungen zu reagieren:

- **Nahtlose Skalierung und unterbrechungsfreier Betrieb.** ONTAP unterstützt das unterbrechungsfreie Hinzufügen von Kapazitäten zu bestehenden Controllern sowie das Scale-out von Clustern. Kunden können Upgrades auf die neuesten Technologien wie NVMe und 32 GB FC ohne teure Datenmigrationen oder Ausfälle durchführen.
- **Cloud Connection.** ONTAP ist die Storage-Managementsoftware mit der umfassendsten Cloud-Integration und bietet Optionen für softwaredefinierten Storage (ONTAP Select) und Cloud-native

Instanzen (NetApp Cloud Volumes Service) in allen Public Clouds.

- **Integration in moderne Applikationen.** ONTAP bietet Datenservices der Enterprise-Klasse für Plattformen und Anwendungen der nächsten Generation, wie autonome Fahrzeuge, Smart Cities und Industrie 4.0, mit derselben Infrastruktur, die bereits bestehende Enterprise-Apps unterstützt.

NetApp SANtricity

NetApp SANtricity wurde entwickelt, um für Hybrid-Flash- und EF-Series All-Flash-Arrays eine branchenführende Performance, Zuverlässigkeit und Einfachheit zu bieten. Erzielen Sie maximale Performance und Auslastung Ihrer E-Series Hybrid-Flash- und EF-Series All-Flash-Arrays für anspruchsvolle Workloads, einschließlich Datenanalysen, Videoüberwachung sowie Backup und Recovery. Mit SANtricity können Einstellungen, Wartung, Kapazitätserweiterung und andere Aufgaben abgeschlossen werden, ohne dass der Storage online bleibt. SANtricity bietet zudem erstklassige Datensicherung, proaktives Monitoring und zertifizierte Sicherheit – alles ist über die benutzerfreundliche, integrierte System Manager Schnittstelle zugänglich. Weitere Informationen finden Sie im "NetApp E-Series SANtricity Software – Datenblatt".

Performance-Optimierung

Die Performance-optimierte SANtricity Software stellt Daten für alle Ihre Datenanalyse-, Videoüberwachungs- und Backup-Applikationen bereit – mit hohen IOPS, hohem Durchsatz und niedriger Latenz. Performance-Steigerung für Applikationen mit hohen IOPS und niedriger Latenz sowie für Applikationen mit hoher Bandbreite und hohem Durchsatz

Maximale Verfügbarkeit

Durchführung Ihrer gesamten Managementaufgaben, während der Storage online bleibt. Ganz gleich, ob Sie Konfigurationen anpassen, Wartungsaufgaben durchführen oder die Kapazität erweitern – der I/O-Betrieb wird nicht unterbrochen Hervorragende Zuverlässigkeit dank automatisierter Funktionen, Online-Konfiguration, hochmoderner Dynamic Disk Pools (DPP)-Technologie und mehr

Keine leichte Aufgabe

SANtricity Software bietet erstklassige Datensicherung, proaktives Monitoring und zertifizierte Sicherheit – alles über die benutzerfreundliche, integrierte System Manager Schnittstelle. Vereinfachung von Storage-Managementaufgaben. Sie erhalten die Flexibilität, die Sie zum fortschrittlichen Tuning aller E-Series Storage-Systeme benötigen. Sie können Ihr NetApp E-Series System jederzeit und überall managen. Unsere integrierte, webbasierte Schnittstelle optimiert Ihren Management-Workflow.

NetApp Trident

"Trident" NetApp ist ein Open-Source-Orchestrator für den dynamischen Storage von Docker und Kubernetes, das die Erstellung, das Management und die Nutzung von persistentem Storage vereinfacht. Die native Kubernetes-Applikation Trident läuft direkt in einem Kubernetes Cluster. Trident ermöglicht Kunden die nahtlose Implementierung von DL-Container-Images auf NetApp Storage und bietet eine Erfahrung der Enterprise-Klasse für den Einsatz von KI-Containern. Kubernetes-Benutzer (WIE ML-Entwickler und Data Scientists) können die Orchestrierung und das Klonen erstellen, managen und automatisieren, um von den erweiterten Datenmanagement-Funktionen von NetApp Technologie zu profitieren.

NetApp BlueXP Kopie und Synchronisierung

"BlueXP Copy und Sync" Ist ein NetApp Service für schnelle und sichere Datensynchronisierung. Unabhängig davon, ob Sie Dateien zwischen On-Premises-NFS- oder SMB-Dateifreigaben, 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 oder IBM Cloud Object Storage: BlueXP Copy and Sync verschiebt Dateien schnell und sicher an den gewünschten Speicherort. Nach der Übertragung stehen die Daten an der Quelle und am Ziel vollständig zur Verfügung. BlueXP Copy and Sync synchronisiert kontinuierlich die Daten, basierend auf einem vorab definierten Zeitplan. Dabei werden nur die Deltas verschoben, sodass der Zeit- und Kostenaufwand für die Datenreplizierung minimiert wird. BlueXP Copy and Sync ist ein Software-as-a-Service-Tool (SaaS), das sich äußerst einfach einrichten und verwenden lässt. Datentransfers, die durch BlueXP Copy und Sync ausgelöst werden, erfolgen durch Datenmanager. Sie können Datenmanager von BlueXP Copy und Sync in AWS, Azure, Google Cloud Platform oder lokal implementieren.

Lenovo ThinkSystem-Server

Lenovo ThinkSystem Server verfügen über innovative Hardware, Software und Services, die die Herausforderungen der Kunden von heute lösen und einen evolutionären, zweckbezogenen, modularen Designansatz bieten, um den Herausforderungen von morgen gerecht zu werden. Diese Server profitieren von erstklassigen, Industriestandard-Technologien in Verbindung mit differenzierten Lenovo Innovationen, um die größtmögliche Flexibilität bei x86-Servern zu bieten.

Zu den wichtigsten Vorteilen der Bereitstellung von Lenovo ThinkSystem Servern gehören:

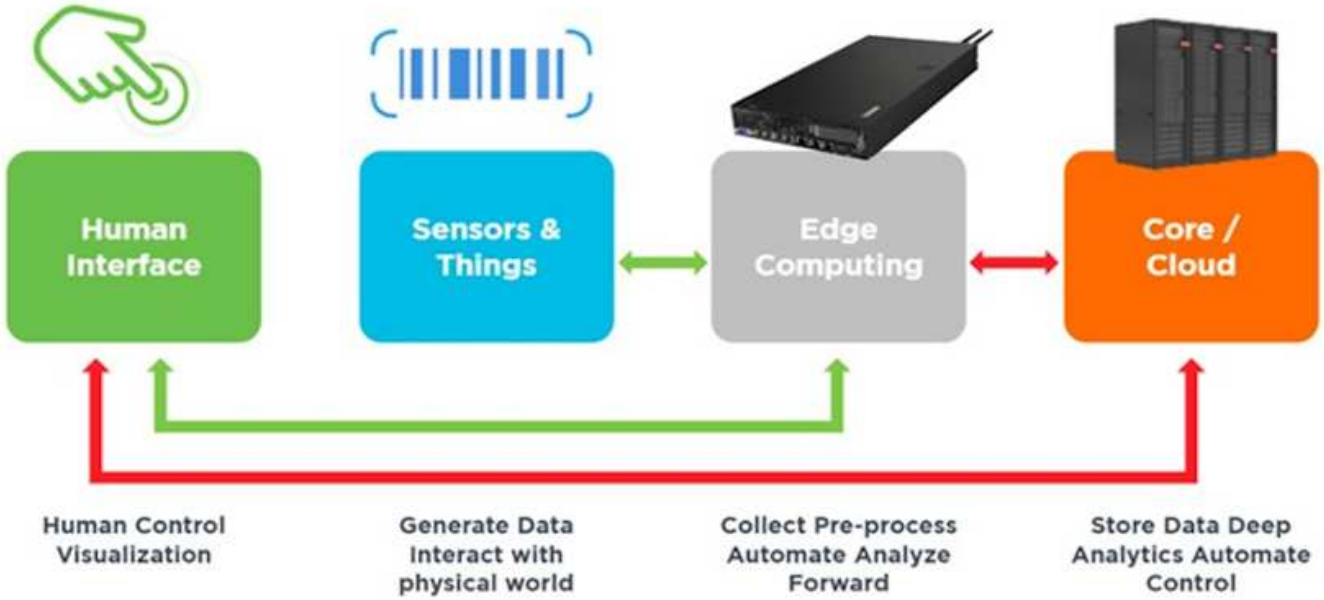
- Hochskalierbare und modulare Designs, die mit dem Unternehmen wachsen können
- Branchenführende Ausfallsicherheit und dadurch Zeitersparnis von Stunden mit teuren, ungeplanten Ausfallzeiten
- Schnelle Flash-Technologien für kürzere Latenzen, schnellere Reaktionszeiten und intelligentes Datenmanagement in Echtzeit

Im KI-Bereich verfolgt Lenovo einen praktischen Ansatz, der Unternehmen dabei hilft, die Vorteile VON ML und KI für ihre Workloads zu verstehen und einzuführen. Lenovo Kunden können die KI-Angebote von Lenovo in Lenovo AI Innovation Centers testen und auswerten, um den Wert für ihren jeweiligen Anwendungsfall zu verstehen. Dieser kundenorientierte Ansatz bietet Kunden ein Proof of Concept für Lösungsplattformen, die sofort einsatzbereit und für KI optimiert sind, zur Verbesserung der Amortisierung.

Lenovo ThinkSystem SE350 Edge Server

Edge Computing ermöglicht die Analyse von Daten von IoT-Geräten am Edge des Netzwerks, bevor sie an das Datacenter oder die Cloud gesendet werden. Das Lenovo ThinkSystem SE350, wie in der Abbildung unten dargestellt, ist für die einzigartigen Anforderungen an den Einsatz am Rand konzipiert, mit dem Schwerpunkt auf Flexibilität, Konnektivität, Sicherheit und Fernverwaltung in einem kompakten robusten und umweltverträglichen Formfaktor.

Mit dem Intel Xeon D Prozessor und der Flexibilität, die Beschleunigung von Edge-KI-Workloads zu unterstützen, wurde der SE350 speziell für die Bewältigung der Herausforderungen von Serverbereitstellungen in verschiedenen Umgebungen außerhalb des Rechenzentrums entwickelt.



MLPerf

MLPerf ist eine branchenführende Benchmark-Suite zur Evaluierung der KI-Performance. Sie deckt zahlreiche Bereiche der angewandten KI ab, darunter Bildklassifizierung, Objekterkennung, medizinische Bildgebung und natürliche Sprachverarbeitung (NLP). In dieser Validierung verwendeten wir Inferenz v0.7 Workloads, was die neueste Version von MLPerf-Inferenz beim Abschluss dieser Validierung ist. Der "MLPerf-Inferenz v0.7" Die Suite enthält vier neue Benchmarks für Datacenter und Edge-Systeme:

- **BERT.** bidirektionale Encoder-Darstellung von Transformatoren (BERT) optimiert für die Beantwortung von Fragen mit Hilfe des Kader-Datensatzes.
- **DLRM.** Deep Learning Recommendation Model (DLRM) ist ein Personalisierungs- und Empfehlungsmodell, das zur Optimierung der Klickraten (CTR) trainiert wird.
- **3D U-Net.** die 3D U-Net-Architektur wird auf dem Datensatz Brain Tumor Segmentation (Briats) trainiert.
- **RNN-T.** Rezidiver Neural Network Transducer (RNN-T) ist ein ASR-Modell (Automatic Speech Rezidicted

Speech Receed), das auf einer Untergruppe von LibriSpeech trainiert wird. MLPerf-Inferenz-Ergebnisse und -Code sind öffentlich verfügbar und unter Apache-Lizenz veröffentlicht. MLPerf-Inferenz verfügt über eine Edge-Abteilung, die die folgenden Szenarien unterstützt:

- **Single Stream.** Dieses Szenario imitiert Systeme, bei denen die Reaktionsfähigkeit ein entscheidender Faktor ist, wie Offline-KI-Abfragen, die auf Smartphones durchgeführt werden. Einzelne Abfragen werden an das System gesendet und Reaktionszeiten werden aufgezeichnet. Die 90. Perzentillatenz aller Antworten wird als Ergebnis gemeldet.
- **Multistream.** dieser Benchmark ist für Systeme, die Input von mehreren Sensoren verarbeiten. Während des Tests werden Abfragen in einem festen Zeitintervall gesendet. Eine QoS-Einschränkung (maximal zulässige Latenz) wurde gestellt. Der Test meldet die Anzahl der Datenströme, die das System verarbeiten kann, während die QoS-Bedingung erfüllt wird.
- **Offline.** Dies ist das einfachste Szenario für Batch-Anwendungen und die Metrik ist der Durchsatz in Proben pro Sekunde. Alle Daten stehen dem System zur Verfügung und der Benchmark misst die Zeit, die für die Verarbeitung aller Proben benötigt wird.

Lenovo hat MLPerf Inference Scores für SE350 mit T4 veröffentlicht, dem Server, der in diesem Dokument verwendet wird. Weitere Informationen finden Sie unter "https://mlperf.org/inference-results-0-7/" Im Abschnitt „Kante, geschlossene Abteilung“ in Eintrag #0.7-145.

Testplan

Dieses Dokument folgt MLPerf Inferenz v0.7 "Codieren", MLPerf Inferenz v1.1 "Codieren", und "Regeln". Wir führten MLPerf-Benchmarks aus, die für die Inferenz am Rand entwickelt wurden, wie in der folgenden Tabelle definiert.

Werden	Aufgabe	Modell	Datensatz	QSL-Größe	Qualität	Einschränkung der Latenz bei mehreren Streams
Vision	Bildklassifizierung	Resnet50v1.5	ImageNet (24 x 224)	1024	99 % des FP32	50 ms
Vision	Objekterkennung (groß)	SSD - Wiedereinsetzen Net34	COCO (1200 x 1200)	64	99 % des FP32	66 ms
Vision	Objekterkennung (klein)	SSD-MobileNetsv1	COCO (300 x 300)	256	99 % des FP32	50 ms
Vision	Segmentierung von medizinischem Bildmaterial	3D-UNET	Briats 2019 (22x24 x 2160)	16	99 % und 99.9 % des FP32	k. A.
Sprache	Sprache-zu-Text	RNNT	Librispeech dev-clean	2513	99 % des FP32	k. A.
Sprache	Sprachverarbeitung	BERT	Kader v1.1	10833	99 % des FP32	k. A.

In der folgenden Tabelle sind Edge Benchmark-Szenarien aufgeführt.

Werden	Aufgabe	Szenarien
Vision	Bildklassifizierung	Single Stream, offline, Multistream
Vision	Objekterkennung (groß)	Single Stream, offline, Multistream
Vision	Objekterkennung (klein)	Single Stream, offline, Multistream
Vision	Segmentierung von medizinischem Bildmaterial	Single Stream offline
Sprache	Sprache-zu-Text	Single Stream offline
Sprache	Sprachverarbeitung	Single Stream offline

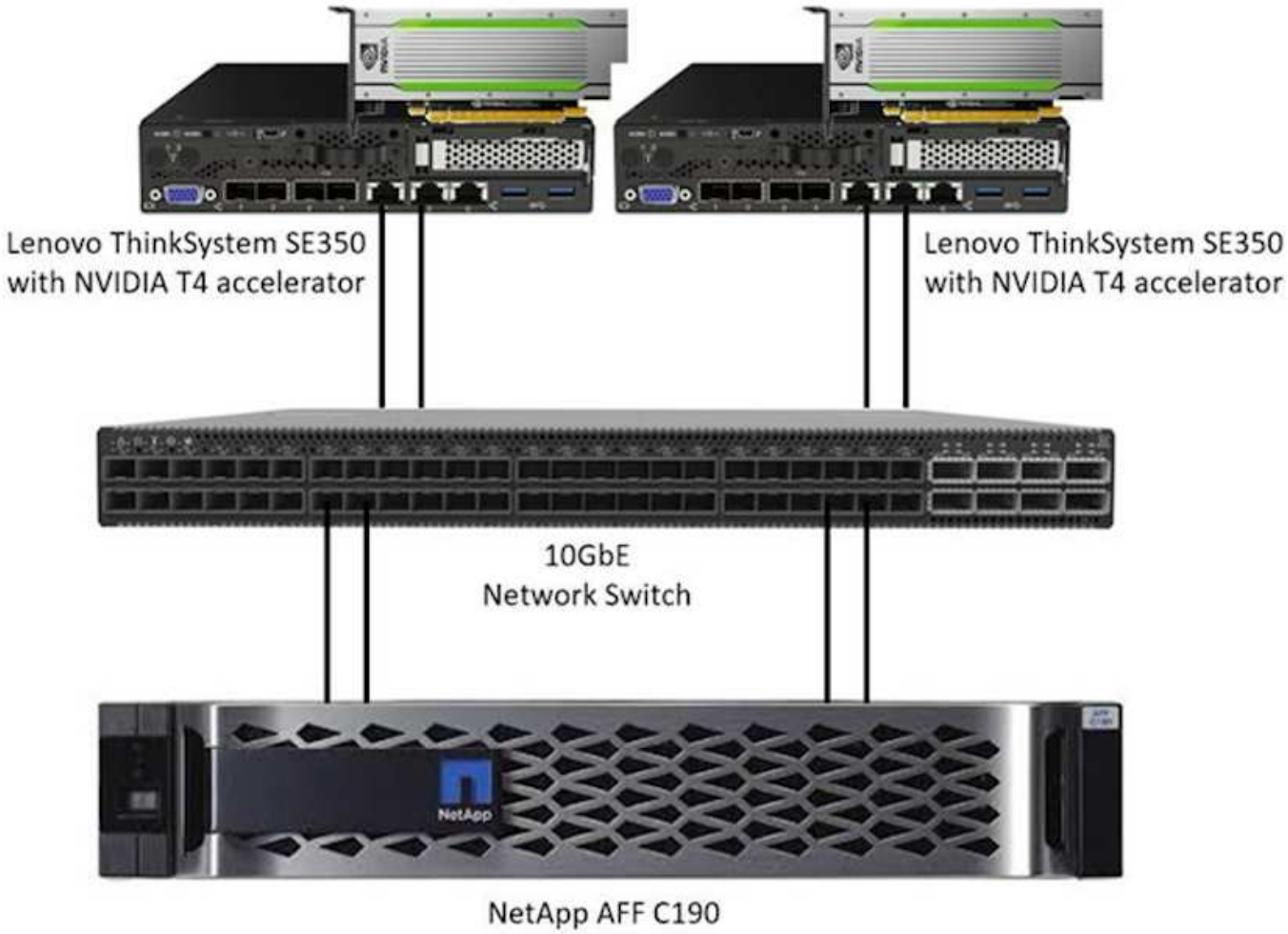
Diese Benchmarks wurden mithilfe der im Rahmen dieser Validierung entwickelten Netzwerk-Storage-Architektur durchgeführt und mit denen von lokalen Ausführung auf den Edge-Servern verglichen, die zuvor MLPerf vorgelegt wurden. Der Vergleich soll ermitteln, welche Auswirkungen der Shared Storage auf die Inferenz-Performance hat.

Testkonfiguration

Die folgende Abbildung zeigt die Testkonfiguration. Wir verwendeten das NetApp AFF C190 Storage-System und zwei Lenovo ThinkSystem SE350 Server (jeweils mit einem NVIDIA T4 Accelerator). Diese Komponenten sind über einen 10-GbE-Netzwerk-Switch verbunden. Der Netzwerk-Storage enthält Validierungs-/Testdatensätze und vortrainierte Modelle. Die Server liefern Berechnungsfunktion, und auf den Storage wird über das NFS-Protokoll zugegriffen.

In diesem Abschnitt werden die getesteten Konfigurationen, die Netzwerkinfrastruktur, der SE350-Server und Details zur Storage-Bereitstellung beschrieben. In der folgenden Tabelle werden die Basiskomponenten für die Lösungsarchitektur aufgeführt.

Lösungskomponenten	Details
Lenovo ThinkSystem-Server	• 2 SE350-Server mit jeweils einer NVIDIA T4-GPU-Karte
	• Jeder Server enthält eine Intel Xeon D-2123IT CPU mit vier physischen Kernen mit 2,20 GHz und 128 GB RAM
NetApp AFF Storage-System der Einstiegsklasse (HA-Paar)	• NetApp ONTAP 9 Software • 24 x 960-GB-SSDs • NFS-Protokoll • Eine Schnittstellengruppe pro Controller, mit vier logischen IP-Adressen für Mount-Punkte



In der folgenden Tabelle ist die Storage-Konfiguration aufgeführt: AFF C190 mit 2RU, 24 Laufwerksschächten.

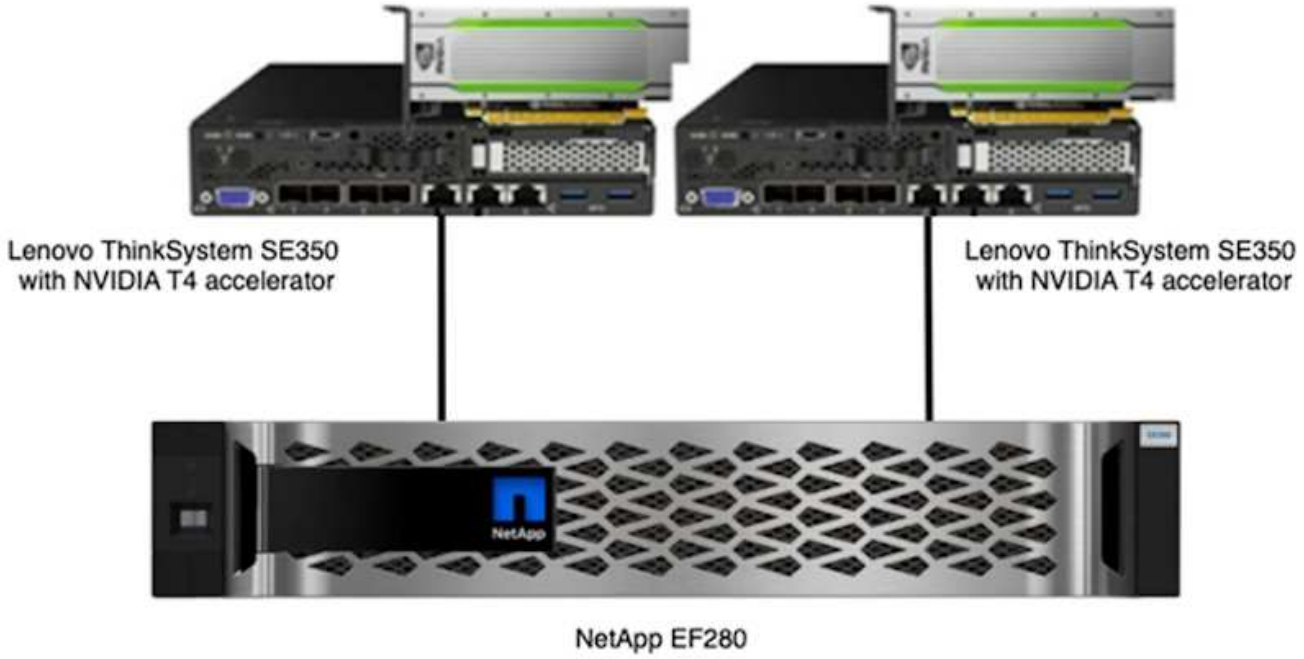
Controller	Aggregat	FlexGroup Volume	Aggregatesize	Volumize	Systemeinhängungspunkt ausführen
Kontroller1	Aggr. 1	/Netapplenovo_AI_fg	8,42 tib	15 TB	/netapp_lenovo_fg
Controller 2	Aggr. 2		8,42 tib		

Der Ordner /netappLenovo_AI_fg enthält die Datensätze, die für die Modellvalidierung verwendet werden.

Die Abbildung unten zeigt die Testkonfiguration. Wir verwendeten das NetApp EF280 Storage-System und zwei Lenovo ThinkSystem SE350 Server (jeweils mit einem NVIDIA T4 Accelerator). Diese Komponenten sind über einen 10-GbE-Netzwerk-Switch verbunden. Der Netzwerk-Storage enthält Validierungs-/Testdatensätze und vortrainierte Modelle. Die Server liefern Berechnungsfunktion, und auf den Storage wird über das NFS-Protokoll zugegriffen.

In der folgenden Tabelle wird die Storage-Konfiguration für EF280 aufgeführt.

Controller	Volume-Gruppe	Datenmenge	Volumize	DDPsize	Verbindungsmethode
Kontroller1	DDP1	Band 1	8,42 tib	16 TB	SE350-1 auf iSCSI-LUN 0
Controller 2		Lautstärke 2	8,42 tib		SE350-2 auf iSCSI-LUN 1



Testverfahren

Dieser Abschnitt beschreibt die Testverfahren zur Validierung dieser Lösung.

Einrichtung von Betriebssystem und KI-Inferenz

Für AFF C190 haben wir Ubuntu 18.04 mit NVIDIA-Treibern und Docker mit Unterstützung für NVIDIA-GPUs verwendet und MLPerf verwendet "Codieren" Erhältlich als Teil der Lenovo Vorlage an MLPerf Inferenz v0.7.

Für EF280 haben wir Ubuntu 20.04 mit NVIDIA-Treibern und Docker mit Unterstützung für NVIDIA-GPUs und MLPerf verwendet "Codieren" Verfügbar als Teil der Lenovo Vorlage für MLPerf Inferenz v1.1.

Zur Einrichtung des KI-Inferenz führen Sie folgende Schritte aus:

1. Laden Sie Datensätze herunter, für die eine Registrierung erforderlich ist, den ImageNet 2012 Validation-Satz, den Criteo Terabyte-Datensatz und den Briats 2019-Trainingssatz, und entpacken Sie die Dateien.
2. Erstellen Sie ein Arbeitsverzeichnis mit mindestens 1 TB und definieren Sie die Umgebungsvariable `MLPERF_SCRATCH_PATH` Bezug auf das Verzeichnis.

   Sie sollten dieses Verzeichnis im gemeinsam genutzten Speicher für den Netzwerkspeicherfall oder die lokale Festplatte beim Testen mit lokalen Daten freigeben.

3. Führen Sie das Make-in-Management `prebuild` Befehl, mit dem der Docker-Container für die erforderlichen Inferenzaufgaben erstellt und gestartet wird.

   Die folgenden Befehle werden alle im laufenden Docker-Container ausgeführt:

   - Laden Sie vortrainierte KI-Modelle für MLPerf-Inferenz-Aufgaben herunter: `make download_model`
   - Laden Sie zusätzliche Datensätze herunter, die kostenlos heruntergeladen werden können: `make download_data`
   - Daten vorverarbeiten: Make `preprocess_data`
   - Ausführen: `make build`.
   - Aufbau der für die GPU optimierten Inferenz-Engines in Computing-Servern: `make generate_engines`
   - Führen Sie zum Ausführen von Inferenz-Workloads den folgenden Befehl aus (ein Befehl):

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

KI-Inferenz wird ausgeführt

Es wurden drei Läufe ausgeführt:

- KI-Inferenz eines einzelnen Servers mit lokalem Storage
- KI-Inferenz eines einzelnen Servers mittels Netzwerk-Storage
- Multi-Server-KI-Inferenz mit Netzwerk-Storage

Testergebnisse

Zur Bewertung der Performance der vorgeschlagenen Architektur wurden zahlreiche Tests durchgeführt.

Es gibt sechs verschiedene Workloads (Bildklassifizierung, Objekterkennung [klein], Objekterkennung [groß], medizinische Bildgebung, Speech-to-Text, Und natürliche Sprachverarbeitung [NLP]), die Sie in drei verschiedenen Szenarien ausführen können: Offline, Single Stream und Multistream.

Das letzte Szenario wird nur für die Bildklassifizierung und Objekterkennung implementiert.

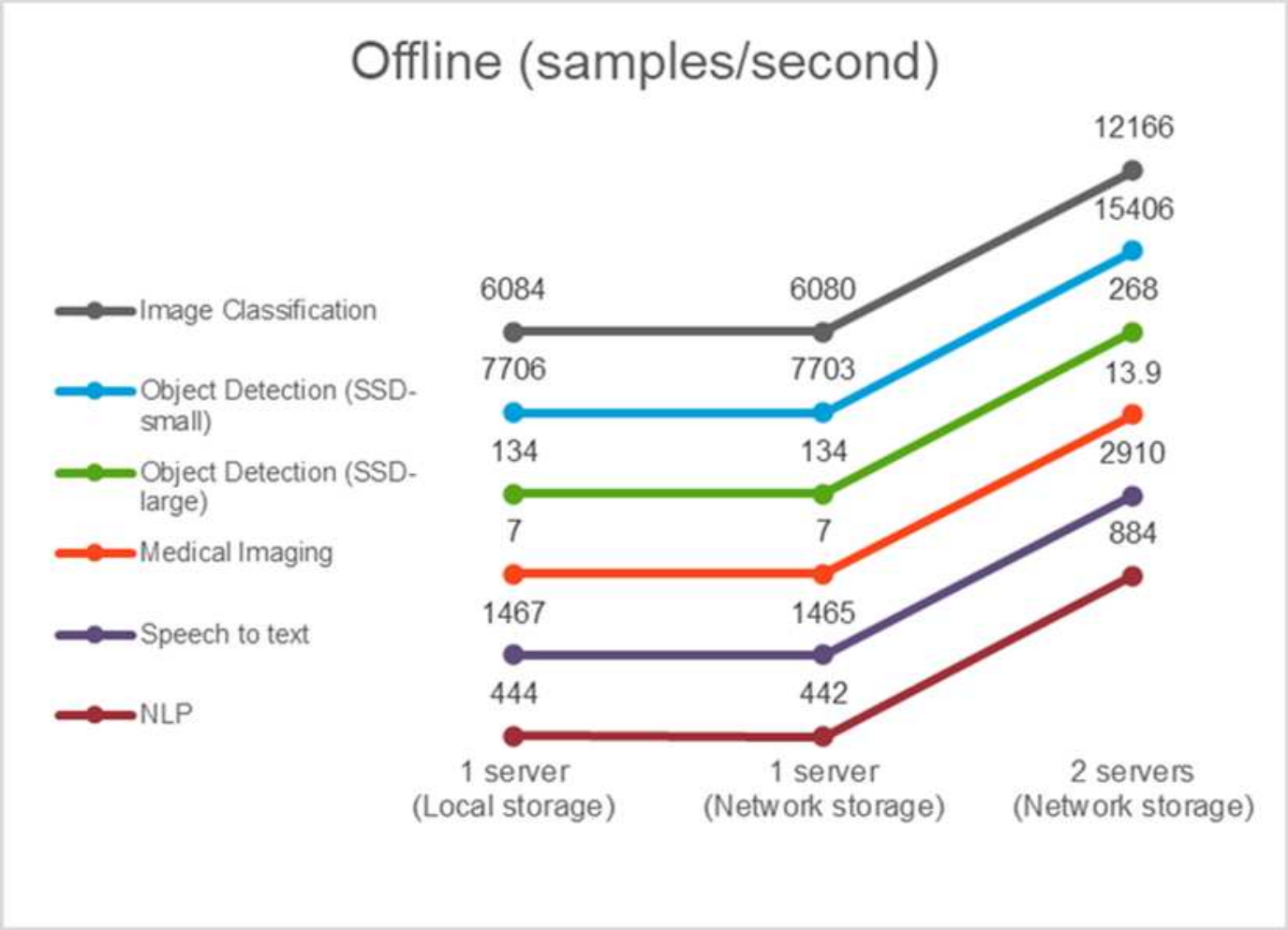
Dies ergibt 15 mögliche Workloads, die alle unter drei verschiedenen Setups getestet wurden:

- Einzelner Server/lokaler Storage
- Einzelner Server/Netzwerk-Storage
- Multi-Server-/Netzwerk-Storage

Die Ergebnisse werden in den folgenden Abschnitten beschrieben.

KI-Inferenz für AFF im Offline-Szenario

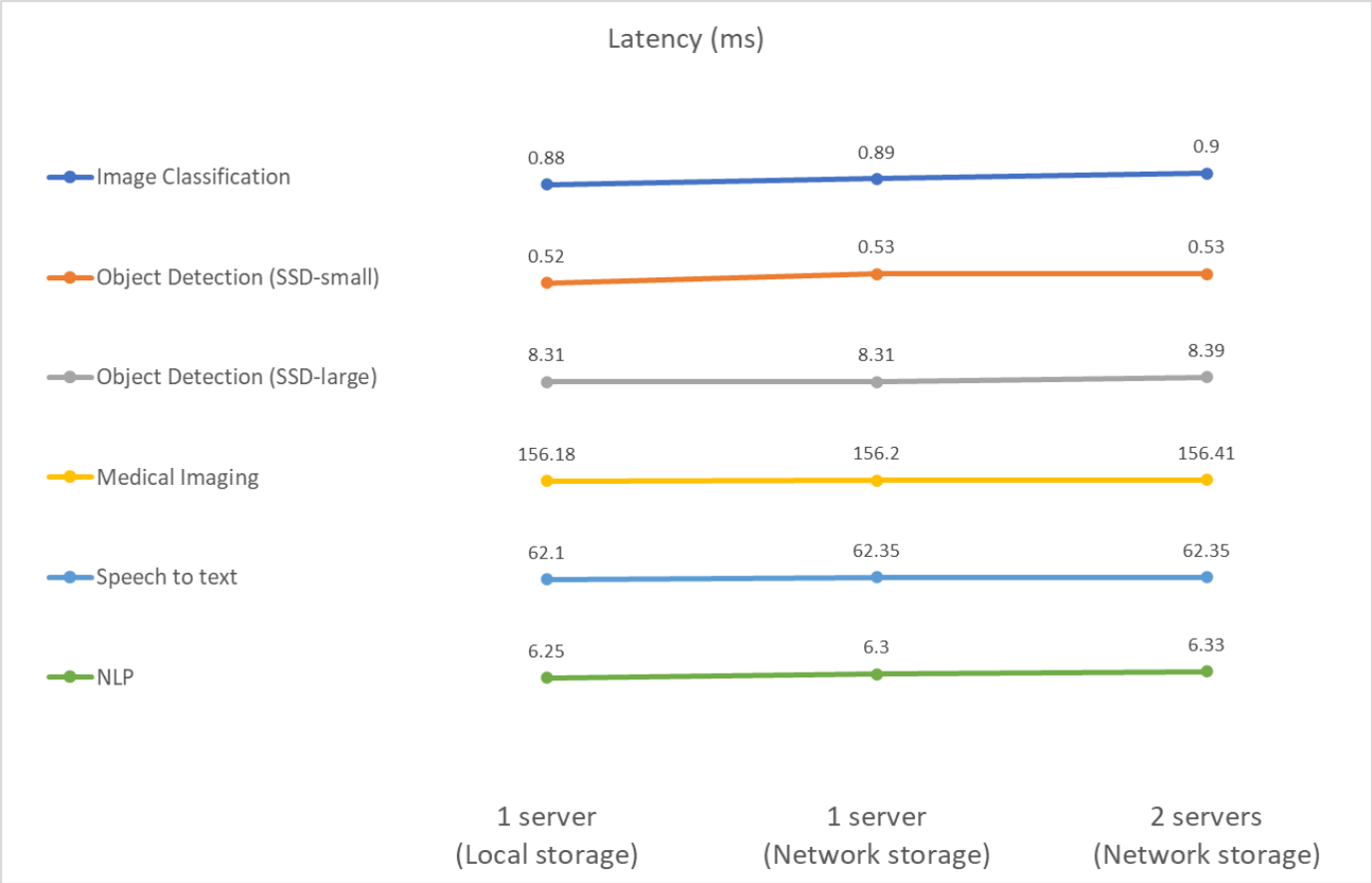
In diesem Szenario waren alle Daten dem Server zur Verfügung und die Zeit, die für die Verarbeitung aller Proben erforderlich war, wurde gemessen. Als Ergebnisse der Tests berichten wir Bandbreiten in Proben pro Sekunde. Wenn mehr als ein Server verwendet wurde, berichten wir über die gesamte Bandbreite, die über alle Server zusammengefasst wurde. Die Ergebnisse für alle drei Anwendungsfälle sind in der folgenden Abbildung dargestellt. Für den Fall mit zwei Servern wird die kombinierte Bandbreite von beiden Servern angegeben.



Die Ergebnisse zeigen, dass Netzwerkspeicher sich nicht negativ auf die Performance auswirken – die Änderung ist minimal und bei einigen Aufgaben wird keine gefunden. Beim Hinzufügen des zweiten Servers wird die Gesamtbandbreite entweder genau verdoppelt oder im schlimmsten Fall ist die Änderung weniger als 1 %.

KI-Inferenz in einem einzelnen Stream-Szenario für AFF

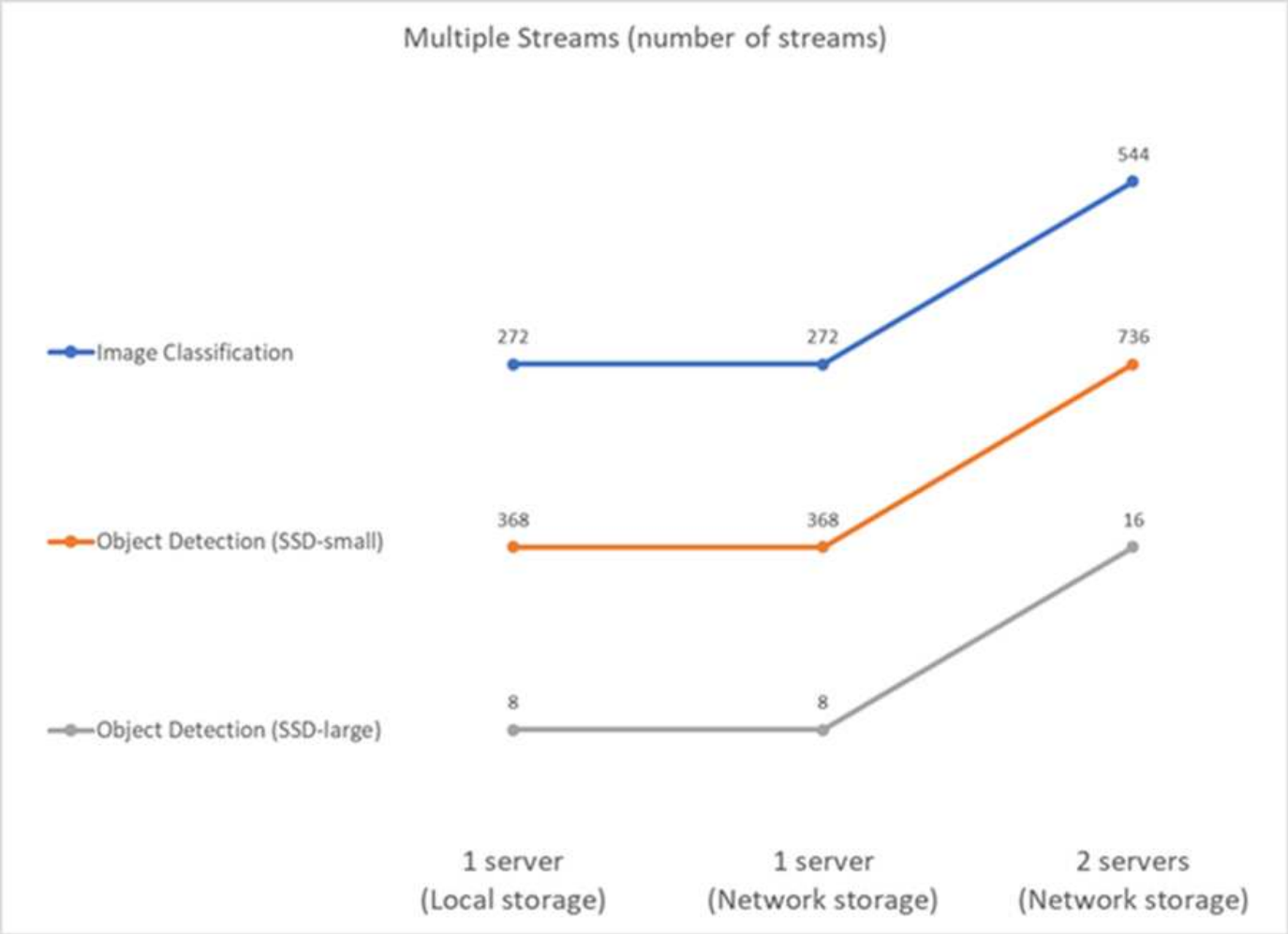
Dieser Benchmark misst die Latenz. Für den Fall mehrerer Computing-Server wird die durchschnittliche Latenz angegeben. Die Ergebnisse für die Aufgabensuite finden Sie in der folgenden Abbildung. Bei zwei Servern wird die durchschnittliche Latenz von beiden Servern angegeben.



Die Ergebnisse zeigen auch hier, dass der Netzwerkspeicher ausreichend ist, um die Aufgaben zu bewältigen. Der Unterschied zwischen lokalem und Netzwerk-Speicher in einem Server-Fall ist minimal oder nicht. Gleiches gilt, wenn zwei Server denselben Storage nutzen – die Latenz auf beiden Servern bleibt bei sehr geringen Änderungen gleich.

KI-Inferenz in Multistream-Szenario für AFF

In diesem Fall führt das System zu der Anzahl der Streams, die das System verarbeiten kann, und erfüllt gleichzeitig die QoS-Einschränkung. Das Ergebnis ist also immer eine ganze Zahl. Für mehrere Server wird die Gesamtanzahl der Streams, die über alle Server zusammengefasst wurden, angegeben. Nicht alle Workloads unterstützen dieses Szenario, aber wir haben sie auch ausgeführt. Die Ergebnisse unserer Tests sind in der folgenden Abbildung zusammengefasst. Im Fall von zwei Servern wird die kombinierte Anzahl von Streams beider Server angegeben.



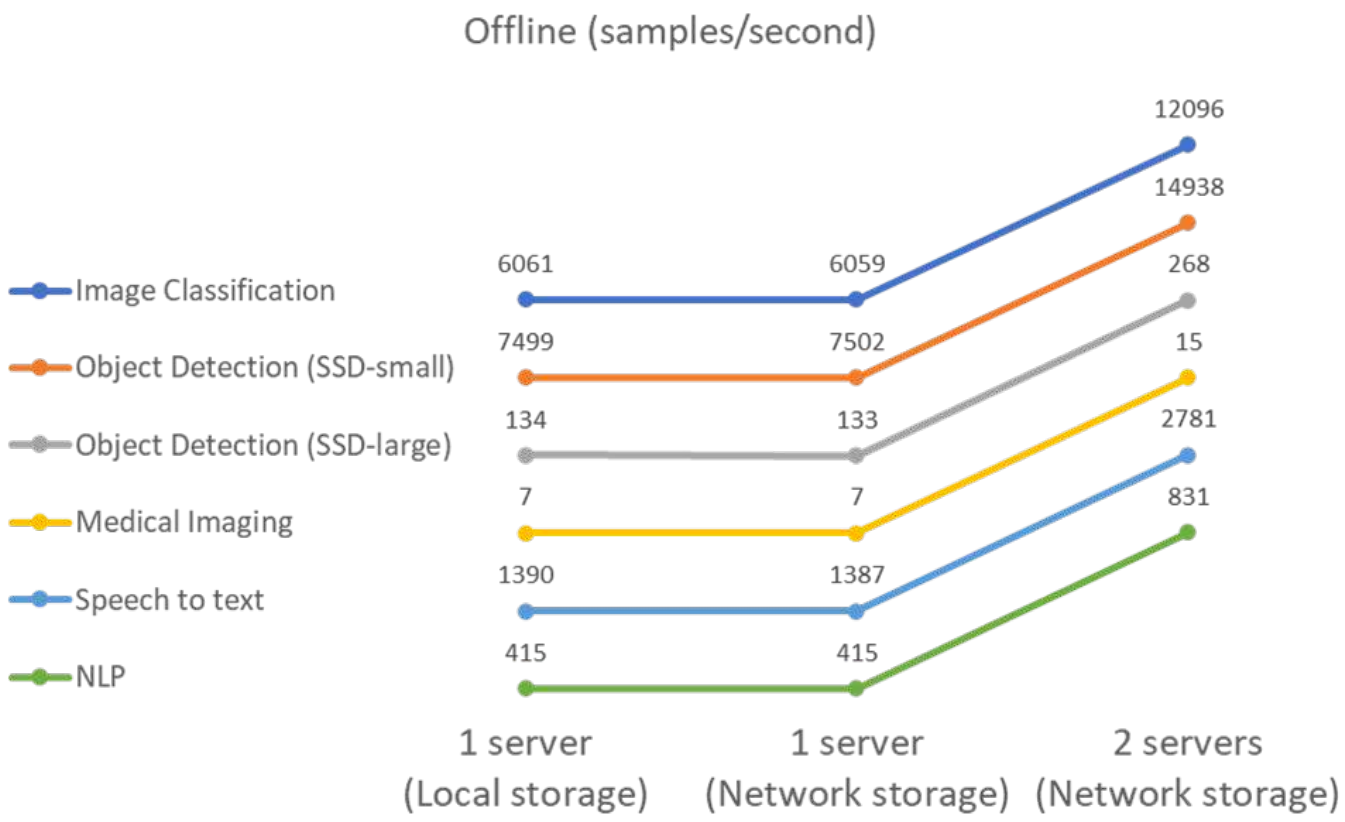
Die Ergebnisse zeigen die perfekte Performance der Einrichtung: Lokaler Speicher und Netzwerk-Storage liefern die gleichen Ergebnisse, und mit dem zweiten Server verdoppelt sich die Anzahl der Streams, die durch das vorgeschlagene Setup bewältigt werden können.

Testergebnisse für EF

Zur Bewertung der Performance der vorgeschlagenen Architektur wurden zahlreiche Tests durchgeführt. Es gibt sechs verschiedene Workloads (Bildklassifizierung, Objekterkennung [klein], Objekterkennung [groß], medizinische Bildgebung, Speech-to-Text, Und natürliche Sprachverarbeitung [NLP]), die in zwei verschiedenen Szenarien ausgeführt wurden: Offline und Single Stream. Die Ergebnisse werden in den folgenden Abschnitten beschrieben.

KI-Inferenz im Offline-Szenario für EF

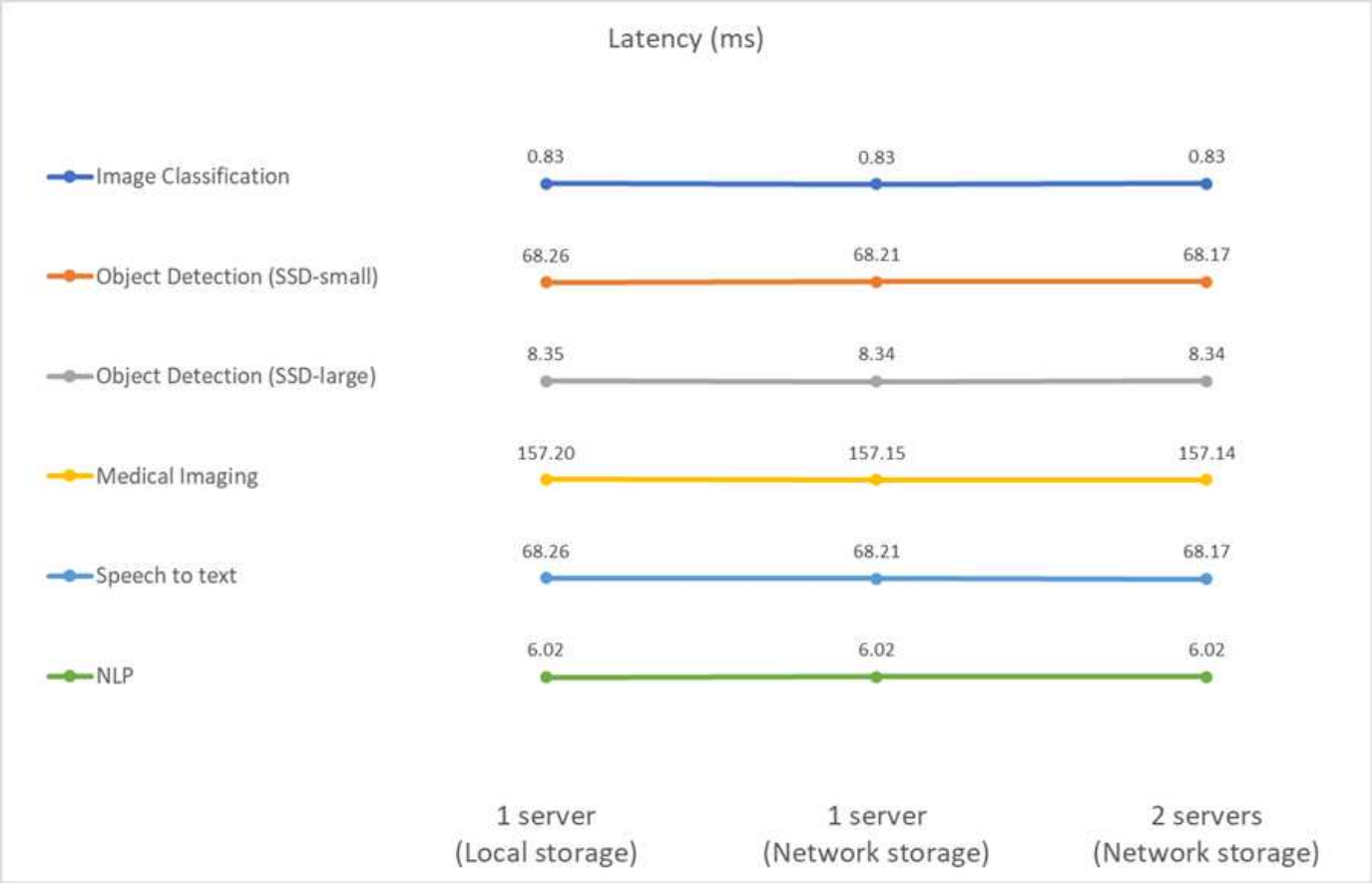
In diesem Szenario waren alle Daten dem Server zur Verfügung und die Zeit, die für die Verarbeitung aller Proben erforderlich war, wurde gemessen. Als Ergebnisse der Tests berichten wir Bandbreiten in Proben pro Sekunde. Für Einzelknoten-Ausführung berichten wir über den Durchschnitt beider Server, während für zwei Server-Ausführung die gesamte Bandbreite über alle Server zusammengefasst angegeben wird. Die Ergebnisse für Anwendungsfälle sind in der Abbildung unten dargestellt.



Die Ergebnisse zeigen, dass Netzwerkspeicher sich nicht negativ auf die Performance auswirken – die Änderung ist minimal und bei einigen Aufgaben wird keine gefunden. Beim Hinzufügen des zweiten Servers wird die Gesamtbandbreite entweder genau verdoppelt oder im schlimmsten Fall ist die Änderung weniger als 1 %.

KI-Inferenz in einem einzelnen Stream-Szenario für die EF-Series

Dieser Benchmark misst die Latenz. In allen Fällen wird die durchschnittliche Latenz über alle Server angegeben, die in den Ausführung involviert sind. Die Ergebnisse für die Aufgabensuite sind gegeben.



Die Ergebnisse zeigen wieder, dass der Netzwerkspeicher zur Bewältigung der Aufgaben ausreicht. Der Unterschied zwischen dem lokalen Speicher und dem Netzwerk-Speicher in einem Server-Fall ist minimal oder gar nicht. Gleiches gilt, wenn zwei Server denselben Storage nutzen – die Latenz auf beiden Servern bleibt bei sehr geringen Änderungen gleich.

Optionen zur Dimensionierung der Architektur

Sie können das für die Validierung verwendete Setup an andere Anwendungsfälle anpassen.

Computing Server

Wir haben eine Intel Xeon D-2123IT CPU mit vier physischen Kernen und 60 W TDP verwendet, die in SE350 die niedrigste CPU-Stufe unterstützt. Der Server unterstützt zwar keinen Austausch von CPUs, kann aber mit einer leistungsstärkeren CPU bestellt werden. Die wichtigste unterstützte CPU ist Intel Xeon D-2183IT mit 16 Kernen, 100 W mit 2,20 GHz. Dadurch erhöht sich die CPU-Rechenleistung erheblich. Während die CPU keinen Engpass mehr für die Ausführung der Inferenz-Workloads darstellt, hilft sie bei der Datenverarbeitung und anderen Vorgänge im Zusammenhang mit Inferenz. Derzeit ist NVIDIA T4 die einzige GPU für Edge-Anwendungsfälle und ist daher derzeit keine Möglichkeit, die GPU zu aktualisieren oder herunterzustufen.

Shared Storage

Das NetApp AFF C190 System mit einer maximalen Storage-Kapazität von 50,5 TB, einem Durchsatz von 4,4 GB/s bei sequenziellem Lesen und 230.000 IOPS für kleine zufällige Lesevorgänge wurde in diesem Dokument eingesetzt. Es hat sich als gut für Edge-Inferenz-Workloads erwiesen.

Wenn Sie jedoch mehr Storage-Kapazität oder schnellere Netzwerkgeschwindigkeiten benötigen, sollten Sie das NetApp AFF A220 oder "NetApp AFF A250" Storage-Systeme. Außerdem wurde das NetApp EF280 System mit einer maximalen Kapazität von 1,5 PB und einer Bandbreite von 10 GB/s im Rahmen dieser Lösungsvalidierung ebenfalls verwendet. Wenn Sie mehr Storage-Kapazität mit höherer Bandbreite bevorzugen, "NetApp EF300" Kann verwendet werden.

Schlussfolgerung

KI-gestützte Automatisierung und Edge-Computing sind eine führende Vorgehensweise, um Unternehmen bei der digitalen Transformation zu unterstützen und ihre betriebliche Effizienz und Sicherheit zu maximieren. Mithilfe von Edge Computing werden Daten wesentlich schneller verarbeitet, da sie nicht mehr zum und von einem Datacenter entfernt werden müssen. Somit sinken die Kosten für den Datentransfer in Datacenter oder in die Cloud. Niedrigere Latenz und eine höhere Geschwindigkeit können von Vorteil sein, wenn Unternehmen Entscheidungen mithilfe von im Edge-Bereich bereitgestellten KI-Inferenzmodellen in nahezu Echtzeit treffen müssen.

NetApp Storage-Systeme bieten die gleiche oder eine bessere Performance wie lokaler SSD-Storage und bieten Data Scientists, Data Engineers, KI/ML-Entwickler und Business- oder IT-Entscheidungsträger die folgenden Vorteile:

- Problemlose gemeinsame Nutzung von Daten zwischen KI-Systemen, Big-Data-Analysen und anderen geschäftskritischen Systemen Diese gemeinsame Nutzung von Daten verringert den Infrastruktur-Overhead, verbessert die Performance und optimiert das Datenmanagement im gesamten Unternehmen.
- Unabhängig skalierbare Computing- und Storage-Ressourcen minimieren Kosten und verbessern die Ressourcenauslastung.
- Optimierte Entwicklungs- und Implementierungs-Workflows mithilfe integrierter Snapshot Kopien und Klone für sofortige und platzsparende Benutzer-Workspaces, integrierte Versionskontrolle und automatisierte Implementierung
- Datensicherung der Enterprise-Klasse für Disaster Recovery und Business Continuity Die in diesem Dokument vorgestellte NetApp und Lenovo Lösung ist eine flexible Scale-out-Architektur, die sich ideal für Edge-Implementierungen von KI-Inferenz der Enterprise-Klasse eignet.

Danksagungen

- J.J. Falkanger, Sr Manager, HPC & AI Solutions, Lenovo
- Dave Arnette, Technical Marketing Engineer, NetApp
- Joey Parnell, Tech Lead E-Series KI-Lösungen, NetApp
- Cody Harryman, QA Engineer, NetApp

Wo Sie weitere Informationen finden

Weitere Informationen zu den in diesem Dokument beschriebenen Daten finden Sie in den folgenden Dokumenten bzw. auf den folgenden Websites:

- Produktseite zu NetApp AFF A-Series Arrays

  "https://www.netapp.com/data-storage/aff-a-series/"

- NetApp ONTAP Datenmanagement-Software: Informationsbibliothek unter ONTAP 9

  http://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62286

- TR-4727: NetApp EF-Series Introduction

  https://www.netapp.com/pdf.html?item=/media/17179-tr4727pdf.pdf

- NetApp E-Series SANtricity Software Datenblatt

  https://www.netapp.com/pdf.html?item=/media/19775-ds-3171-66862.pdf

- NetApp persistenter Storage für Container – NetApp Trident

  "https://netapp.io/persistent-storage-provisioner-for-kubernetes/"

- MLPerf
  - "https://mlcommons.org/en/"
  - "http://www.image-net.org/"
  - "https://mlcommons.org/en/news/mlperf-inference-v11/"
- NetApp BlueXP Kopie und Synchronisierung

  "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"

- Lenovo ThinkSystem DM5100F Unified Flash Storage Array

  "https://lenovopress.com/lp1365-thinksystem-dm5100f-unified-flash-storage-array"

Copyright-Informationen

Copyright © 2024 NetApp. Alle Rechte vorbehalten. Gedruckt in den USA. Dieses urheberrechtlich geschützte Dokument darf ohne die vorherige schriftliche Genehmigung des Urheberrechtsinhabers in keiner Form und durch keine Mittel – weder grafische noch elektronische oder mechanische, einschließlich Fotokopieren, Aufnehmen oder Speichern in einem elektronischen Abrufsystem – auch nicht in Teilen, vervielfältigt werden.

Software, die von urheberrechtlich geschütztem NetApp Material abgeleitet wird, unterliegt der folgenden Lizenz und dem folgenden Haftungsausschluss:

DIE VORLIEGENDE SOFTWARE WIRD IN DER VORLIEGENDEN FORM VON NETAPP ZUR VERFÜGUNG GESTELLT, D. H. OHNE JEGLICHE EXPLIZITE ODER IMPLIZITE GEWÄHRLEISTUNG, EINSCHLIESSLICH, JEDOCH NICHT BESCHRÄNKT AUF DIE STILLSCHWEIGENDE GEWÄHRLEISTUNG DER MARKTGÄNGIGKEIT UND EIGNUNG FÜR EINEN BESTIMMTEN ZWECK, DIE HIERMIT AUSGESCHLOSSEN WERDEN. NETAPP ÜBERNIMMT KEINERLEI HAFTUNG FÜR DIREKTE, INDIREKTE, ZUFÄLLIGE, BESONDERE, BEISPIELHAFTE SCHÄDEN ODER FOLGESCHÄDEN (EINSCHLIESSLICH, JEDOCH NICHT BESCHRÄNKT AUF DIE BESCHAFFUNG VON ERSATZWAREN ODER -DIENSTLEISTUNGEN, NUTZUNGS-, DATEN- ODER GEWINNVERLUSTE ODER UNTERBRECHUNG DES GESCHÄFTSBETRIEBS), UNABHÄNGIG DAVON, WIE SIE VERURSACHT WURDEN UND AUF WELCHER HAFTUNGSTHEORIE SIE BERUHEN, OB AUS VERTRAGLICH FESTGELEGTER HAFTUNG, VERSCHULDENSUNABHÄNGIGER HAFTUNG ODER DELIKTSHAFTUNG (EINSCHLIESSLICH FAHRLÄSSIGKEIT ODER AUF ANDEREM WEGE), DIE IN IRGENDEINER WEISE AUS DER NUTZUNG DIESER SOFTWARE RESULTIEREN, SELBST WENN AUF DIE MÖGLICHKEIT DERARTIGER SCHÄDEN HINGEWIESEN WURDE.

NetApp behält sich das Recht vor, die hierin beschriebenen Produkte jederzeit und ohne Vorankündigung zu ändern. NetApp übernimmt keine Verantwortung oder Haftung, die sich aus der Verwendung der hier beschriebenen Produkte ergibt, es sei denn, NetApp hat dem ausdrücklich in schriftlicher Form zugestimmt. Die Verwendung oder der Erwerb dieses Produkts stellt keine Lizenzierung im Rahmen eines Patentrechts, Markenrechts oder eines anderen Rechts an geistigem Eigentum von NetApp dar.

Das in diesem Dokument beschriebene Produkt kann durch ein oder mehrere US-amerikanische Patente, ausländische Patente oder anhängige Patentanmeldungen geschützt sein.

ERLÄUTERUNG ZU „RESTRICTED RIGHTS“: Nutzung, Vervielfältigung oder Offenlegung durch die US-Regierung unterliegt den Einschränkungen gemäß Unterabschnitt (b)(3) der Klausel „Rights in Technical Data – Noncommercial Items“ in DFARS 252.227-7013 (Februar 2014) und FAR 52.227-19 (Dezember 2007).

Die hierin enthaltenen Daten beziehen sich auf ein kommerzielles Produkt und/oder einen kommerziellen Service (wie in FAR 2.101 definiert) und sind Eigentum von NetApp, Inc. Alle technischen Daten und die Computersoftware von NetApp, die unter diesem Vertrag bereitgestellt werden, sind gewerblicher Natur und wurden ausschließlich unter Verwendung privater Mittel entwickelt. Die US-Regierung besitzt eine nicht ausschließliche, nicht übertragbare, nicht unterlizenzierbare, weltweite, limitierte unwiderrufliche Lizenz zur Nutzung der Daten nur in Verbindung mit und zur Unterstützung des Vertrags der US-Regierung, unter dem die Daten bereitgestellt wurden. Sofern in den vorliegenden Bedingungen nicht anders angegeben, dürfen die Daten ohne vorherige schriftliche Genehmigung von NetApp, Inc. nicht verwendet, offengelegt, vervielfältigt, geändert, aufgeführt oder angezeigt werden. Die Lizenzrechte der US-Regierung für das US-Verteidigungsministerium sind auf die in DFARS-Klausel 252.227-7015(b) (Februar 2014) genannten Rechte beschränkt.

Markeninformationen

NETAPP, das NETAPP Logo und die unter http://www.netapp.com/TM aufgeführten Marken sind Marken von NetApp, Inc. Andere Firmen und Produktnamen können Marken der jeweiligen Eigentümer sein.