Edge-NetApp에서 Lenovo ThinkSystem을 사용한 AI 추론 - 솔루션 설계

NetApp Solutions

NetApp
April 20, 2024

This PDF was generated from https://docs.netapp.com/ko-kr/netapp-solutions/ai/ai-edge-introduction.html on April 20, 2024. Always check docs.netapp.com for the latest.

목차

Edge-NetApp에서 Lenovo ThinkSystem을 사용한 AI 추론 - 솔루션 설계 . . . 1
TR-4886: Edge-NetApp에서 Lenovo ThinkSystem - 솔루션 설계를 사용한 AI 추론 . . . 1
기술 개요 . . . 5
테스트 계획 . . . 12
구성을 테스트합니다 . . . 12
테스트 절차 . . . 14
테스트 결과 . . . 15
아키텍처 사이징 옵션 . . . 20
결론 . . . 21

Edge-NetApp에서 Lenovo ThinkSystem을 사용한 AI 추론 - 솔루션 설계

TR-4886: Edge-NetApp에서 Lenovo ThinkSystem - 솔루션 설계를 사용한 AI 추론

Sathish Thyagarajan, NetApp Miroslav Hodak, Lenovo

요약

ADAS(Advanced Driver Assistance Systems), Industry 4.0, 스마트 시티 및 IoT(Internet of Things)와 같은 몇 가지 새로운 애플리케이션 시나리오에서는 지연 시간이 거의 없이 지속적인 데이터 스트림을 처리해야 합니다. 이 문서에서는 이러한 요구사항을 충족하는 에지 환경에서 NetApp 스토리지 컨트롤러 및 Lenovo ThinkSystem 서버에 GPU 기반 인공 지능(AI) 추론을 배포하기 위한 컴퓨팅 및 스토리지 아키텍처에 대해 설명합니다. 또한, NVIDIA T4 GPU가 장착된 에지 서버에서 다양한 추론 작업을 평가하여 업계 표준 MLPerf Inference 벤치마크의 성능 데이터도 제공합니다. 오프라인, 단일 스트림 및 다중 스트림 추론 시나리오의 성능을 조사한 결과, 비용 효율적인 공유 네트워크 스토리지 시스템이 포함된 아키텍처의 성능이 매우 뛰어나며 여러 에지 서버에 대한 데이터 및 모델 관리의 중앙 지점을 제공하는 것으로 나타났습니다.

소개

기업들은 네트워크 에지에 대량의 데이터를 생성하고 있습니다. 스마트 센서 및 IoT 데이터를 활용하여 최대의 가치를 실현하기 위해 조직은 에지 컴퓨팅을 지원하는 실시간 이벤트 스트리밍 솔루션을 찾고 있습니다. 따라서 데이터 센터 외부의 에지에서는 컴퓨팅 작업이 점점 더 많이 수행됩니다. AI 추론을 이러한 트렌드에 동인으로 이끄는 요인 중 하나입니다. 에지 서버는 특히 가속기를 사용할 때 이러한 워크로드에 충분한 연산 능력을 제공하지만 제한된 스토리지는 종종 문제가 됩니다. 특히 다중 서버 환경에서는 더욱 그렇습니다. 이 문서에서는 에지 환경에서 공유 스토리지 시스템을 구축하는 방법과 성능 저하 없이 AI 추론 워크로드의 이점을 활용하는 방법을 설명합니다.

이 문서에서는 에지의 AI 추론을 위한 참조 아키텍처에 대해 설명합니다. 여러 Lenovo ThinkSystem 에지 서버를 NetApp 스토리지 시스템과 결합하여 간편하게 구축 및 관리할 수 있는 솔루션을 구축합니다. 이 가이드는 여러 대의 카메라와 산업용 센서가 장착된 공장 바닥, 소매 거래의 POS(Point-of-Sale) 시스템 또는 자율 차량의 시각적 이상을 식별하는 FSD(Full Self-Driving) 시스템 등 다양한 상황에서 실제 배포를 위한 기본 안내서입니다.

이 문서에서는 Lenovo ThinkSystem SE350 Edge Server와 엔트리 레벨 NetApp AFF 및 EF-Series 스토리지 시스템으로 구성된 컴퓨팅 및 스토리지 구성의 테스트 및 검증을 다룹니다. 참조 아키텍처는 AI 배포를 위한 효율적이고 비용 효율적인 솔루션을 제공하는 동시에 NetApp ONTAP 및 NetApp SANtricity 데이터 관리 소프트웨어를 통해 포괄적인 데이터 서비스, 통합 데이터 보호, 원활한 확장성 및 클라우드 연결 데이터 스토리지를 제공합니다.

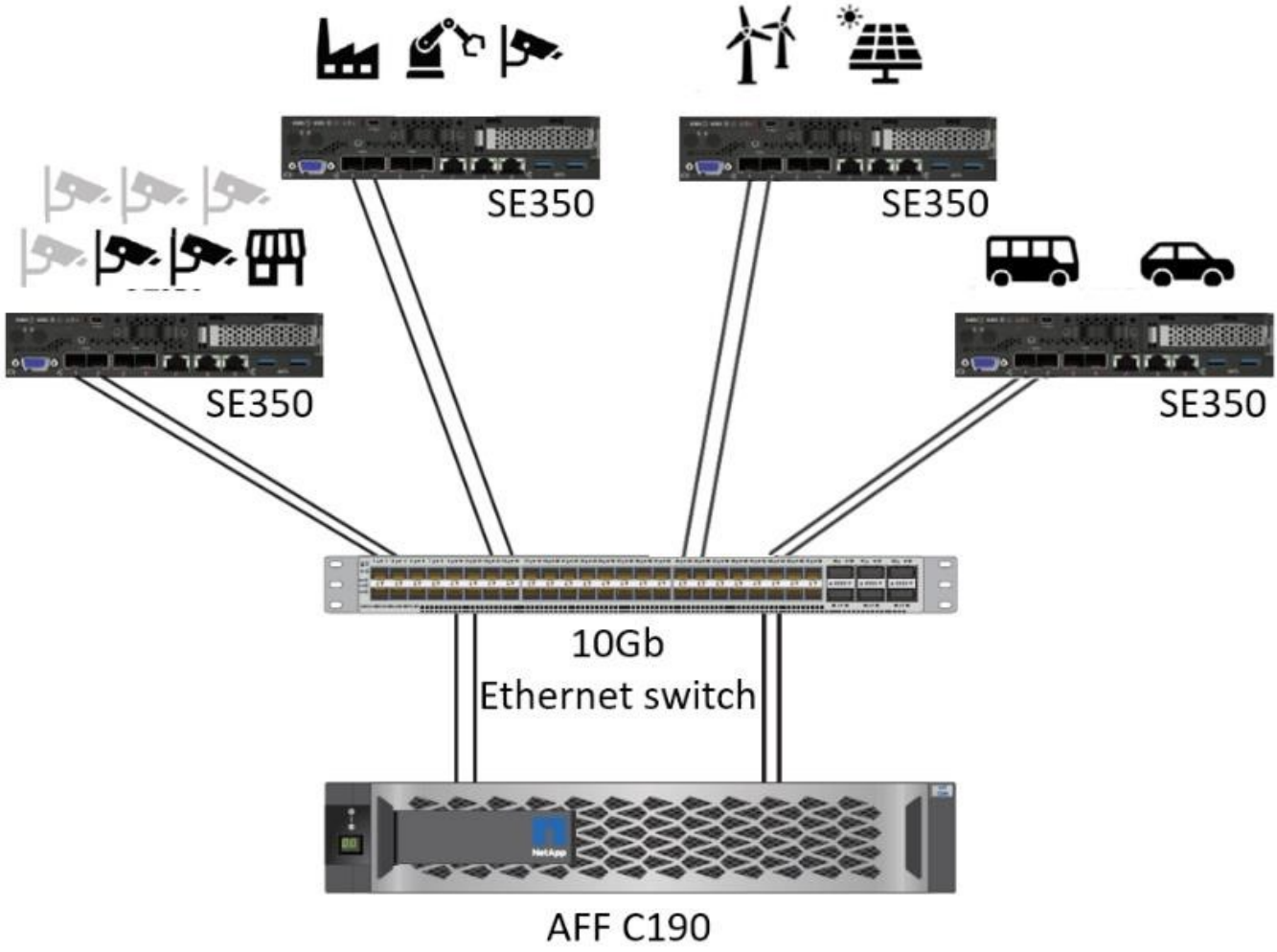
대상

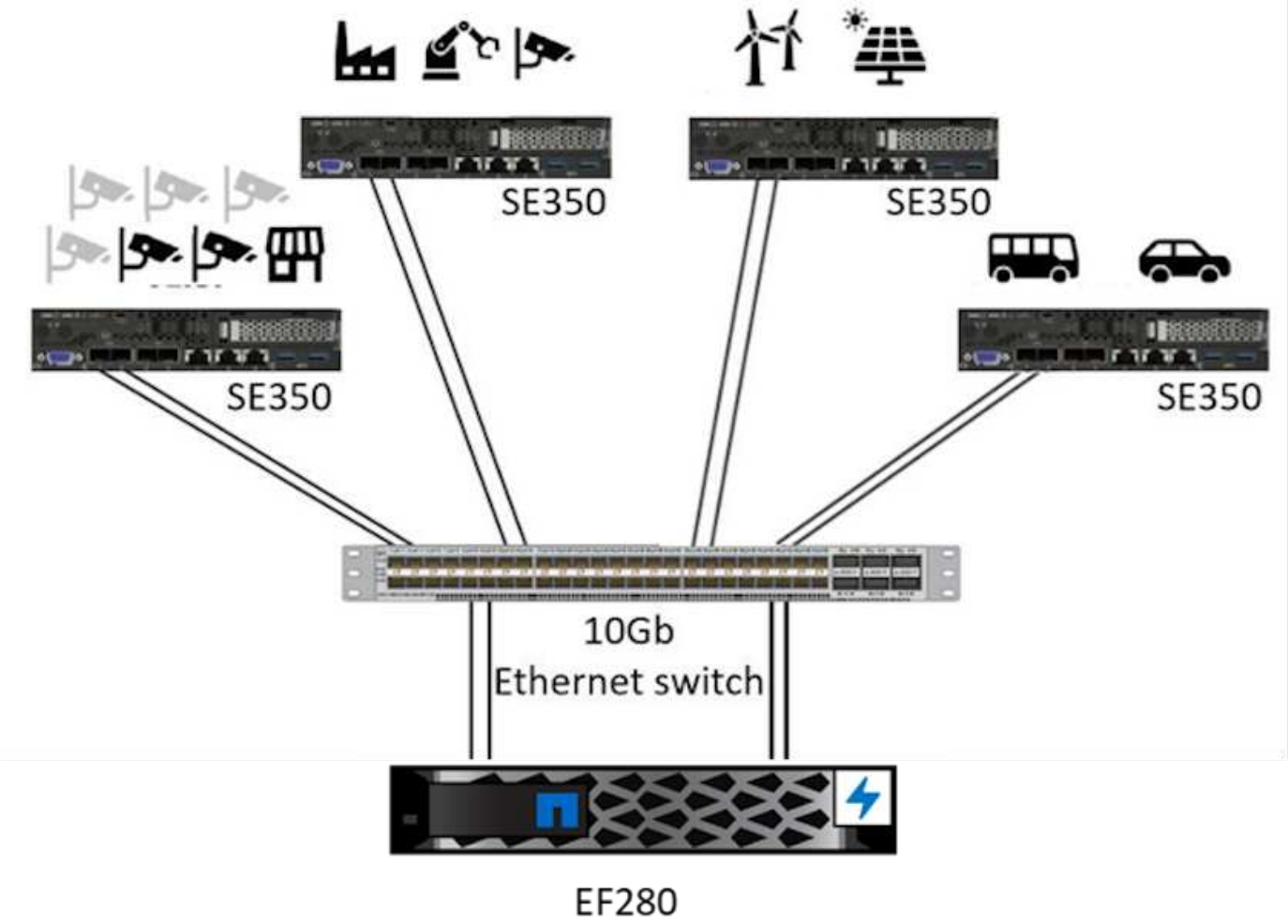
이 문서는 다음 사용자를 대상으로 합니다.

- 에지의 AI를 제품화하려는 비즈니스 리더 및 엔터프라이즈 설계자
- 데이터 과학자, 데이터 엔지니어, AI/기계 학습(ML) 연구원 및 AI 시스템 개발자.
- AI/ML 모델 및 애플리케이션 개발을 위한 솔루션을 설계하는 엔터프라이즈 설계자
- 딥 러닝(DL) 및 ML 모델을 구축하는 효율적인 방법을 찾고 있는 데이터 과학자 및 AI 엔지니어
- 에지 장치 관리자 및 에지 서버 관리자는 에지 추론 모델의 구축과 관리를 담당합니다.

솔루션 아키텍처

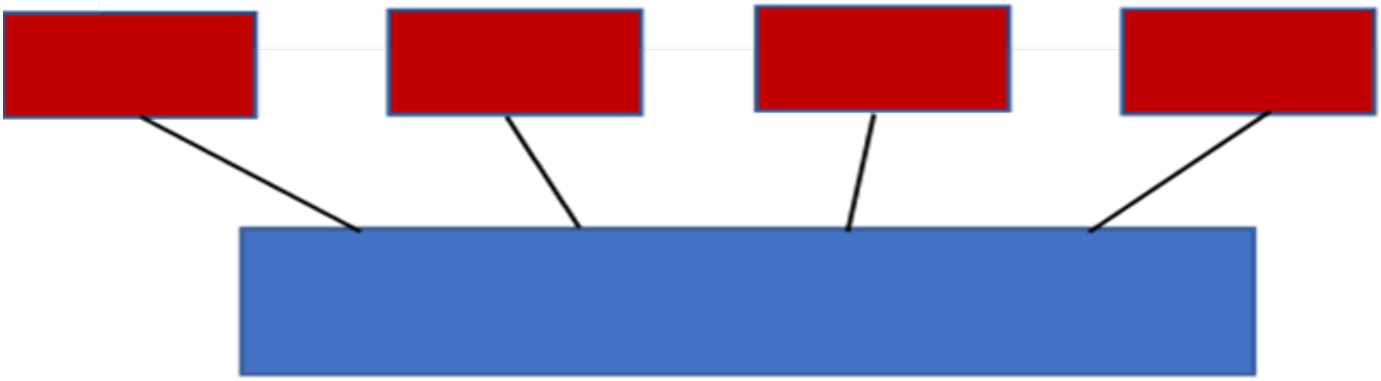
이 Lenovo ThinkSystem 서버 및 NetApp ONTAP 또는 NetApp SANtricity 스토리지 솔루션은 기존 CPU와 함께 GPU의 처리 능력을 사용하여 대규모 데이터 세트에서 AI 추론을 처리하도록 설계되었습니다. 이 검증 방식은 다음 두 그림에 표시된 대로 단일 NetApp AFF 스토리지 시스템과 상호 연결된 단일 또는 다중 Lenovo SR350 에지 서버를 사용하는 아키텍처로 고성능 및 최적의 데이터 관리를 수행하는 것입니다.





다음 그림의 논리적 아키텍처 개요에서는 이 아키텍처의 컴퓨팅 및 스토리지 요소 역할을 보여 줍니다. 특히 다음과 같은 사항이 표시됩니다.

- 에지 컴퓨팅 장치가 카메라, 센서 등의 데이터를 기반으로 추론을 수행합니다.
- 다양한 용도로 사용되는 공유 스토리지 요소:
  - 추론 모델과 추론을 수행하는 데 필요한 다른 데이터를 위한 중심 위치를 제공합니다. 컴퓨팅 서버는 스토리지를 직접 액세스하고 로컬에서 복사할 필요 없이 네트워크 전체에서 추론 모델을 사용합니다.
  - 업데이트된 모델이 여기에 푸시됩니다.
  - 에지 서버가 나중에 분석할 수 있도록 수신하는 입력 데이터를 보관합니다. 예를 들어, 에지 장치가 카메라에 연결된 경우 저장소 요소는 카메라에서 캡처한 비디오를 유지합니다.



빨간색	파란색
Lenovo 컴퓨팅 시스템	NetApp AFF 스토리지 시스템
카메라, 센서 등의 입력에서 추론을 수행하는 에지 장치	추측 모델과 에지 디바이스의 데이터를 저장하는 공유 스토리지로, 추후 분석 지원

이 NetApp 및 Lenovo 솔루션은 다음과 같은 주요 이점을 제공합니다.

- 소규모 지사 또는 부서에서의 GPU 가속 컴퓨팅.
- 공유 스토리지에서 백업 및 관리되는 다중 에지 서버 배포
- 데이터 손실 없이 낮은 RPO(복구 시점 목표) 및 RTO(복구 시간 목표)를 충족하는 강력한 데이터 보호
- NetApp Snapshot 복사본 및 클론을 통해 데이터 관리를 최적화하여 개발 워크플로우를 간소화합니다.

이 아키텍처를 사용하는 방법

이 문서에서는 제안된 아키텍처의 설계 및 성능을 검증합니다. 하지만 NetApp은 특정 소프트웨어 수준의 컨테이너, 워크로드, 모델 관리, 클라우드 또는 온프레미스의 데이터 센터 등과 같은 특정 소프트웨어 레벨 구성 요소를 테스트하지 않았습니다. 이러한 소프트웨어 레벨 구성 요소가 배포 시나리오에 한정되어 있기 때문입니다. 여기에는 여러 개의 선택 사항이 있습니다.

컨테이너 관리 수준에서 Kubernetes 컨테이너 관리는 좋은 선택이며 전체 업스트림 버전(Canonical) 또는 엔터프라이즈 배포에 적합한 수정 버전(Red Hat)에서 지원됩니다. 를 클릭합니다 "NetApp AI Control Plane" NetApp Trident 및 새로 추가된 Trident를 사용합니다 "NetApp DataOps 툴킷" 데이터 과학자 및 데이터 엔지니어가 NetApp 스토리지와 통합할 수 있도록 추적 가능성, 데이터 관리 기능, 인터페이스 및 툴을 기본으로 제공합니다. Kubernetes용 ML 툴킷인 Kubeflow는 추가 AI 기능을 제공하는 동시에 TensorFlow Serving 또는 NVIDIA Triton Inference Server와 같은 여러 플랫폼에서 모델 버전 관리 및 KFServing을 지원합니다. 또 다른 옵션은 NVIDIA EGX 플랫폼으로, GPU 지원 AI 추론 컨테이너 카탈로그에 액세스하여 워크로드 관리를 제공합니다. 그러나 이러한 옵션을 사용하려면 운영 환경에 투입하기 위해 상당한 노력과 전문 지식이 필요할 수 있으며 타사 ISV(독립 소프트웨어 공급업체) 또는 컨설턴트의 도움이 필요할 수 있습니다.

솔루션 영역

AI 추론 및 에지 컴퓨팅의 주요 이점은 지연 시간 없이 높은 수준의 품질로 데이터를 컴퓨팅, 처리 및 분석할 수 있는 장치의 기능입니다. 이 문서에서 설명하는 에지 컴퓨팅 사용 사례는 매우 많지만 다음과 같은 몇 가지 대표적인 사례가 있습니다.

자동차: 자율주행 차량

전형적인 에지 컴퓨팅 일러스트는 자율주행 차량(AV)의 첨단 운전자 지원 시스템(ADAS)에 포함되어 있습니다. 무인 자동차의 AI는 안전하고 성공적인 운전자가 되려면 카메라와 센서의 많은 데이터를 신속하게 처리해야 합니다. 물체와 사람 사이의 해석에 너무 많은 시간이 걸릴경우 생명 또는 사망이 발생할 수 있으므로 데이터를 최대한 차량과 가깝게 처리할 수 있어야 합니다. 이 경우 하나 이상의 에지 컴퓨팅 서버가 카메라, 레이더, LiDAR 및 기타 센서의 입력을 처리하는 동시에 공유 스토리지에는 추론 모델이 저장되고 센서의 입력 데이터가 저장됩니다.

의료: 환자 모니터링

AI 및 에지 컴퓨팅이 미치는 가장 큰 영향 중 하나는 가정 및 중환자실(ICU) 모두에서 만성 질환 환자를 지속적으로 모니터링할 수 있는 기능입니다. 인슐린 수치, 호흡, 신경학적 활동, 심장 리듬 및 위장관 기능을 모니터링하는 에지 장치에서 얻은 데이터는 다른 사람의 생명을 구하기 위한 제한된 시간이 있기 때문에 즉시 실행되어야 하는 데이터에 대한 즉각적인 분석이 필요합니다.

소매**:** 계산원 없는 지불

에지 컴퓨팅은 유통업체가 계산 시간을 단축하고 발트 트래픽을 늘릴 수 있도록 AI 및 ML을 지원합니다. 계산원이 필요 없는 시스템은 다음과 같은 다양한 구성 요소를 지원합니다.

- 인증 및 액세스. 물리적 쇼핑객을 검증된 계정에 연결하고 소매 공간에 대한 액세스를 허용합니다.
- 인벤토리 모니터링. 센서, RFID 태그 및 컴퓨터 비전 시스템을 사용하여 쇼핑객의 아이템 선택 또는 선택 취소를 확인할 수 있습니다.

  여기서 각 에지 서버는 각 계산 카운터를 처리하며 공유 스토리지 시스템은 중앙 동기화 지점으로 사용됩니다.

금융 서비스**:** 키오스크의 인적 안전 및 사기 방지

은행 조직에서는 AI 및 에지 컴퓨팅을 사용하여 혁신을 진행하고 맞춤형 뱅킹 경험을 만들고 있습니다. 실시간 데이터 분석 및 AI 추론을 사용하는 대화형 키오스크는 이제 ATM을 통해 고객이 돈을 인출할 수 있도록 지원할 뿐만 아니라 카메라에서 캡처한 이미지를 통해 키오스크를 사전 예방적으로 모니터링하여 사람의 안전 또는 사기 행위 위험을 식별할 수 있습니다. 이 시나리오에서는 에지 컴퓨팅 서버 및 공유 스토리지 시스템이 대화형 키오스크 및 카메라에 연결되어 은행이 AI 추론 모델로 데이터를 수집하고 처리할 수 있도록 도와줍니다.

제조**: Industry 4.0**

4차 산업혁명(Industry 4.0)은 Smart Factory 및 3D 프린팅과 같은 새로운 트렌드와 함께 시작되었습니다. 데이터 중심의 미래에 대비하기 위해 대규모 M2M(Machine-to-Machine) 통신 및 IoT가 통합되어 사람의 개입 없이 자동화 수준을 높일 수 있습니다. 제조는 이미 고도로 자동화되어 있으며 AI 기능을 추가하는 것은 장기적인 추세를 자연스럽게 이어주는 것입니다. AI를 사용하면 컴퓨터 비전 및 기타 AI 기능을 활용하여 자동화할 수 있는 운영을 자동화할 수 있습니다. 제조 공장이 안전 및 품질 관리에 필요한 ISO 표준을 충족할 수 있도록 제조 공장의 조립 라인에서 자재를 더 빠르게 분석하는 데 있어 인간의 시각이나 의사 결정에 의존하는 품질 관리 또는 작업을 자동화할 수 있습니다. 여기서 각 컴퓨팅 에지 서버는 제조 프로세스를 모니터링하는 센서 배열에 연결되고 필요에 따라 업데이트된 추론 모델이 공유 스토리지로 푸시됩니다.

통신**: Rust** 감지**,** 타워 검사 및 네트워크 최적화

통신 업계에서는 컴퓨터 비전과 AI 기술을 사용하여 녹을 자동으로 탐지하고 부식된 셀 타워를 식별하는 이미지를 처리하여 추가적인 검사가 필요합니다. 드론 이미지와 AI 모델을 사용하여 타워의 특정 영역을 식별하고 녹, 표면 균열 및 부식을 분석하는 일이 최근 몇 년 사이에 증가했습니다. 통신 인프라와 셀 타워를 효율적으로 검사하고, 정기적으로 성능 저하를 평가하며, 필요할 때 신속하게 수리할 수 있는 AI 기술에 대한 수요가 지속적으로 증가하고 있습니다.

또한, 데이터 트래픽 패턴을 예측하고 5G 지원 장치를 감지하고 MIMO(다중 입력 및 다중 출력) 에너지 관리를 자동화 및 보강하기 위해 AI 및 ML 알고리즘을 사용하는 것도 통신 업계의 새로운 사용 사례입니다. MIMO 하드웨어는 무선 타워에서 네트워크 용량을 늘리기 위해 사용되지만, 추가 에너지 비용이 필요합니다. 셀 사이트에 배치된 “MIMO 절전 모드”용 ML 모델은 무전기의 효율적인 사용을 예측하고 모바일 네트워크 사업자(MNO)의 에너지 소비 비용을 줄이는 데 도움이 됩니다. AI 추론 및 에지 컴퓨팅 솔루션은 MNO가 데이터 센터로 주고받는 데이터 양을 줄이고, TCO를 낮추고, 네트워크 운영을 최적화하고, 최종 사용자의 전반적인 성능을 개선하는 데 도움이 됩니다.

기술 개요

이 섹션에서는 AI 솔루션의 기술 기반에 대해 설명합니다.

NetApp AFF 시스템

최첨단 NetApp AFF 스토리지 시스템을 사용하면 AI 추론 구축을 통해 에지에서 업계 최고 수준의 성능, 탁월한 유연성,

클라우드 통합, 동급 최고의 데이터 관리로 엔터프라이즈 스토리지 요구사항을 충족할 수 있습니다. 플래시 전용으로 설계된 NetApp AFF 시스템은 비즈니스 크리티컬 데이터를 더 빠르게 처리하고 관리, 보호할 수 있도록 지원합니다.

- 엔트리 레벨 NetApp AFF 스토리지 시스템은 FAS2750 하드웨어 및 SSD 플래시 미디어를 기반으로 합니다
- HA 구성의 컨트롤러 2개

NetApp 엔트리 레벨 AFF C190 스토리지 시스템은 다음 기능을 지원합니다.

- 최대 드라이브 수는 24x 960GB SSD입니다
- 두 가지 가능한 구성:
  - 이더넷(10GbE): 10GBASE-T(RJ-45) 포트 4개
  - 유니파이드(16Gb FC 또는 10GbE): 4x UTA2(Unified Target Adapter 2) 포트
- 최대 50.5TB의 유효 용량

> NAS 워크로드의 경우, 단일 엔트리 레벨 AFF C190 시스템은 연속 읽기의 경우 4.4GBps의 처리량과 작은 랜덤 읽기의 경우 1ms 이하의 지연 시간으로 230K IOPS를 지원합니다.

NetApp AFF A220을 참조하십시오

또한, NetApp은 대규모 구축을 위해 더 뛰어난 성능과 확장성을 제공하는 다른 엔트리급 스토리지 시스템을 제공합니다. NAS 워크로드의 경우 단일 엔트리 레벨 AFF A220 시스템이 다음을 지원합니다.

- 순차적 읽기의 경우 6.2GBps의 처리량
- 375K IOPS, 1ms 미만의 지연 시간으로 소규모 랜덤 읽기 지원
- 최대 드라이브 수는 144x 960GB, 3.8TB 또는 7.6TB SSD입니다
- AFF A220은 1PB 이상의 실제 용량으로 확장됩니다

NetApp AFF A250

- 최대 실제 용량은 35PB이며 최대 스케일아웃 2-24개 노드(HA 쌍 12개)를 지원하는 경우
- AFF A220보다 45% 이상 높은 성능 향상을 제공합니다
- 440k IOPS 랜덤 읽기 @ 1ms
- 최신 NetApp ONTAP 릴리스 ONTAP 9.8을 기반으로 구축
- HA 및 클러스터 인터커넥트에 2개의 25GB 이더넷을 활용합니다

NetApp E-Series EF 시스템

EF-Series는 엔트리 레벨 및 미드레인지 All-Flash SAN 스토리지 어레이 제품군으로, NetApp SANtricity 소프트웨어를 사용하여 데이터에 더 빠르게 액세스하고 가치를 더 빠르게 창출할 수 있습니다. 이러한 시스템은 SAS 및 NVMe 플래시 스토리지를 모두 제공하며 경제적인 가격으로 최고 수준의 IOPS, 100마이크로초 미만의 응답 시간, 최대 44GBps의 대역폭을 제공하므로 AI 추론 및 고성능 컴퓨팅(HPC)과 같은 까다로운 애플리케이션과 혼합 워크로드에 적합합니다.

다음 그림에서는 NetApp EF280 스토리지 시스템을 보여 줍니다.

NetApp EF280

- 32Gb/16Gb FC, 25Gb/10Gb iSCSI 및 12Gb SAS 지원
- 최대 실제 용량은 총 1.5PB의 96개 드라이브입니다
- 10GBps 처리량(순차적 읽기)
- 300K IOPS(랜덤 읽기)
- NetApp EF280은 NetApp 포트폴리오에서 가장 경제적인 All-Flash 어레이(AFA)입니다

NetApp EF300

- 24x NVMe SSD 드라이브로 총 367TB 용량 지원
- 총 240x NL-SAS HDD, 96x SAS SSD 또는 그 조합 확장 옵션
- 100Gb NVMe/IB, NVMe/RoCE, iSER/IB 및 SRP/IB
- 32Gb NVMe/FC, FCP

- 25GB iSCSI
- 20GBps(순차적 읽기)
- 670K IOPS(랜덤 읽기)

> 자세한 내용은 를 참조하십시오 "NetApp EF-Series NetApp EF-Series All-Flash 어레이 EF600, F300, EF570, EF280 데이터시트".

NetApp ONTAP 9

NetApp의 최신 세대 스토리지 관리 소프트웨어인 ONTAP 9.8.1을 통해 기업은 인프라를 현대화하고 클라우드 지원 데이터 센터로 전환할 수 있습니다. ONTAP는 업계 최고 수준의 데이터 관리 기능을 활용하여 데이터가 상주하는 위치와 상관없이 단일 툴셋으로 데이터를 관리하고 보호할 수 있습니다. 필요에 따라 에지, 코어, 클라우드 등 어느 위치로도 데이터를 자유롭게 이동할 수 있습니다. ONTAP 9.8.1에는 데이터 관리를 단순화하고, 중요 데이터를 더 빨리 처리하고 보호하고, 하이브리드 클라우드 아키텍처 전반에서 차세대 인프라 기능을 지원하는 다양한 기능이 포함되어 있습니다.

데이터 관리를 단순화하십시오

애플리케이션 및 데이터 세트에 적절한 리소스가 사용될 수 있도록 데이터 관리는 엔터프라이즈 IT 운영에 매우 중요합니다. ONTAP에는 운영을 간소화 및 단순화하고 총 운영 비용을 절감할 수 있는 다음과 같은 기능이 포함되어 있습니다.

- * 인라인 데이터 컴팩션 및 확대된 중복제거. * 데이터 컴팩션은 스토리지 블록 내부의 낭비되는 공간을 줄이고, 중복제거는 실제 용량을 크게 증가시킵니다. 이는 로컬에 저장된 데이터와 클라우드로 계층화된 데이터에 적용됩니다.
- * 최소, 최대 및 적응형 서비스 품질(AQoS). * 세분화된 서비스 품질(QoS) 제어는 공유 수준이 높은 환경에서 중요 애플리케이션의 성능 수준을 유지하는 데 도움이 됩니다.
- * NetApp FabricPool. * 이 기능은 콜드 데이터를 AWS(Amazon Web Services), Azure, NetApp StorageGRID 스토리지 솔루션을 포함한 퍼블릭 및 프라이빗 클라우드 스토리지 옵션으로 자동 계층화합니다. FabricPool에 대한 자세한 내용은 를 참조하십시오 "TR-4598".

데이터 가속화 및 보호

ONTAP 9은 탁월한 수준의 성능과 데이터 보호를 제공하며 다음과 같은 방법으로 이러한 기능을 확장합니다.

- * 성능 및 낮은 지연 시간 * ONTAP는 가장 짧은 지연 시간으로 가장 높은 처리량을 제공합니다.
- * 데이터 보호. * ONTAP는 모든 플랫폼에서 공통 관리를 지원하는 내장 데이터 보호 기능을 제공합니다.
- * NVE(NetApp 볼륨 암호화). * ONTAP는 온보드 및 외부 키 관리를 모두 지원하여 네이티브 볼륨 레벨 암호화를 제공합니다.
- * 멀티테넌시 및 다단계 인증 * ONTAP를 통해 인프라 리소스를 최고 수준의 보안으로 공유할 수 있습니다.

미래 지향형 인프라

ONTAP 9은 다음과 같은 기능을 통해 지속적으로 변화하는 까다로운 비즈니스 요구사항을 충족할 수 있도록 지원합니다.

- * 원활한 확장 및 무중단 운영 * ONTAP은 기존 컨트롤러 및 스케일아웃 클러스터에 무중단으로 용량을 추가할 수 있도록 지원합니다. 고객은 고비용이 따르는 데이터 마이그레이션이나 운영 중단 없이 NVMe 및 32Gb FC와 같은

최신 기술로 업그레이드할 수 있습니다.

- * 클라우드 연결. * ONTAP은 클라우드에 가장 많이 연결되는 스토리지 관리 소프트웨어로, 모든 퍼블릭 클라우드에서 소프트웨어 정의 스토리지(ONTAP Select) 및 클라우드 네이티브 인스턴스(NetApp Cloud Volumes Service) 옵션이 제공됩니다.
- 새로운 애플리케이션과의 통합 * ONTAP는 기존 엔터프라이즈 앱을 지원하는 인프라와 동일한 인프라를 사용하여 자율주행 차량, 스마트 시티, Industry 4.0과 같은 차세대 플랫폼 및 애플리케이션을 위한 엔터프라이즈급 데이터 서비스를 제공합니다.

NetApp SANtricity를 참조하십시오

NetApp SANtricity는 E-Series 하이브리드 플래시 및 EF-Series All-Flash 어레이에 업계 최고의 성능, 안정성, 단순성을 제공하도록 설계되었습니다. 데이터 분석, 비디오 감시, 백업 및 복구 등 워크로드가 많은 애플리케이션에서 E-Series 하이브리드 플래시 및 EF-Series All-Flash 어레이의 성능과 활용률을 극대화합니다. SANtricity를 사용하면 스토리지를 온라인 상태로 유지하면서 구성 조정, 유지 관리, 용량 확장 및 기타 작업을 완료할 수 있습니다. 또한 SANtricity는 사용하기 쉬운 온박스형 시스템 관리자 인터페이스를 통해 뛰어난 데이터 보호, 사전 예방 모니터링 및 인증 보안을 제공합니다. 자세한 내용은 를 참조하십시오 "NetApp E-Series SANtricity 소프트웨어 데이터시트 를 참조하십시오".

최적의 성능

성능에 최적화된 SANtricity 소프트웨어는 모든 데이터 분석, 비디오 감시 및 백업 앱에 높은 IOPS 및 처리량과 짧은 지연 시간으로 데이터를 제공합니다. IOPS가 높고 지연 시간이 짧은 애플리케이션과 대역폭과 처리량이 높은 애플리케이션의 성능을 더욱 높이십시오.

가동 시간 극대화

스토리지가 온라인 상태일 때 모든 관리 작업을 완료하십시오. I/O를 중단하지 않고 구성을 변경하거나, 유지보수를 수행하거나, 용량을 확장할 수 있습니다 자동화된 기능, 온라인 구성, 최첨단 DPP(Dynamic Disk Pool) 기술 등을 통해 동급 최고의 안정성을 실현합니다.

편안한 휴식

SANtricity 소프트웨어는 사용이 간편한 온박스형 시스템 관리자 인터페이스를 통해 뛰어난 데이터 보호, 사전 예방 모니터링 및 인증 보안을 제공합니다. 스토리지 관리 업무를 간소화합니다. 모든 E-Series 스토리지 시스템의 고급 튜닝에 필요한 유연성 확보 언제 어디서나 NetApp E-Series 시스템을 관리할 수 있습니다. NetApp의 온박스 웹 기반 인터페이스는 관리 워크플로우를 간소화합니다.

NetApp 트라이던트

"트라이던트" NetApp은 Docker 및 Kubernetes용 오픈 소스 동적 스토리지 오케스트레이터로서 영구 스토리지의 생성, 관리 및 사용을 단순화합니다. Kubernetes 네이티브 애플리케이션인 Trident는 Kubernetes 클러스터 내에서 직접 실행됩니다. Trident를 사용하면 고객이 DL 컨테이너 이미지를 NetApp 스토리지에 원활하게 배포하고 AI 컨테이너 배포를 위한 엔터프라이즈급 경험을 제공할 수 있습니다. Kubernetes 사용자(예: ML 개발자 및 데이터 과학자)는 오케스트레이션 및 클론 복제를 생성, 관리 및 자동화하여 NetApp 기술이 제공하는 NetApp 고급 데이터 관리 기능을 활용할 수 있습니다.

NetApp BlueXP 복사 및 동기화

"BlueXP 복사 및 동기화" 는 빠르고 안전한 데이터 동기화를 제공하는 NetApp 서비스입니다. 온프레미스 NFS 또는 SMB 파일 공유 간에 파일을 전송해야 하는 경우, 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, 또는 IBM Cloud Object Storage인 BlueXP Copy and Sync는 필요한 파일을 빠르고 안전하게 이동합니다. 데이터가 전송되면 소스와 타겟 모두에서 사용할 수 있습니다. BlueXP 복사 및 동기화는 미리 정의된 일정에 따라 데이터를 지속적으로 동기화하므로 변경된 부분만 이동하므로 데이터 복제에 소비되는 시간과 비용이 최소화됩니다. BlueXP Copy and Sync는 매우 간단하게 설정하고 사용할 수 있는 서비스형 소프트웨어(SaaS) 툴입니다. BlueXP Copy 및 Sync에 의해 트리거되는 데이터 전송은 데이터 브로커에 의해 수행됩니다. AWS, Azure, Google Cloud Platform 또는 사내에 BlueXP Copy 및 Sync 데이터 브로커를 배포할 수 있습니다.

Lenovo ThinkSystem 서버

Lenovo ThinkSystem 서버는 현재 고객의 과제를 해결하고 미래의 과제를 해결할 수 있는 혁신적인 모듈식 설계 접근 방식을 제공하는 혁신적인 하드웨어, 소프트웨어 및 서비스를 갖추고 있습니다. 이러한 서버는 동급 최강의 업계 표준 기술과 차별화된 Lenovo의 혁신적인 기술을 결합하여 x86 서버에서 최대한의 유연성을 제공합니다.

Lenovo ThinkSystem 서버 배포의 주요 이점은 다음과 같습니다.

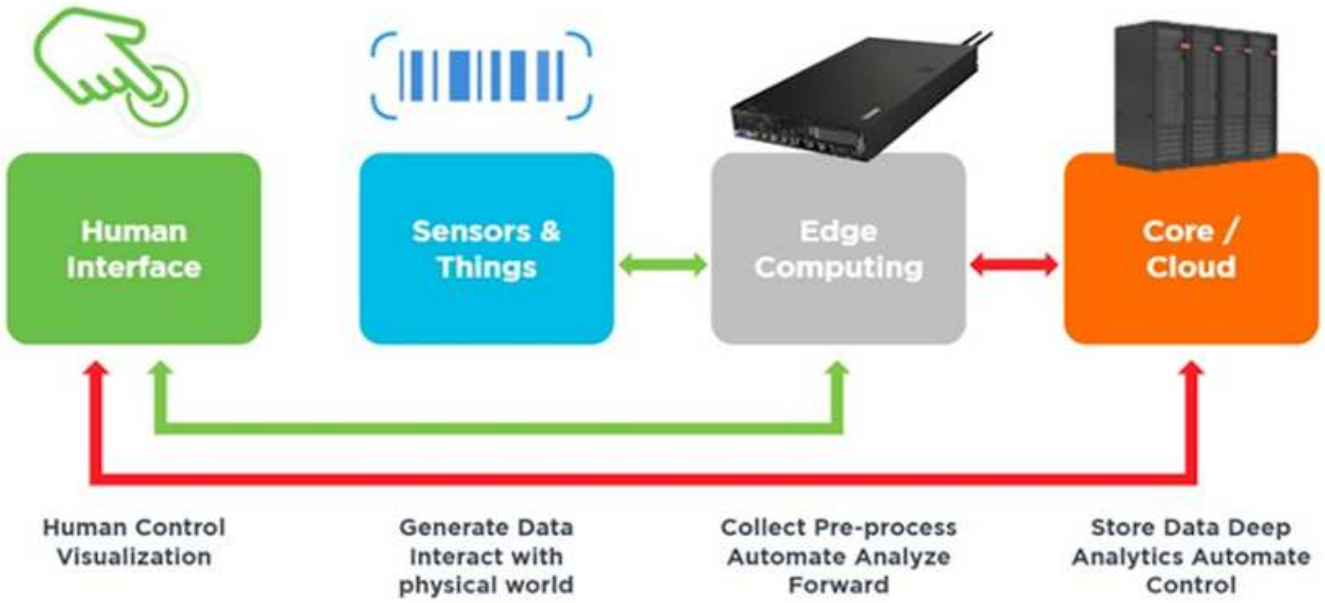
- 비즈니스 성장에 맞춰 확장할 수 있는 모듈식 설계
- 업계 최고 수준의 복원력으로 예기치 못한 가동 중지의 비용이 많이 드는 시간을 절약할 수 있습니다
- 빠른 플래시 기술을 통해 지연 시간을 단축하고, 응답 시간을 단축하며, 데이터 관리를 실시간으로 수행할 수 있습니다

AI 분야에서 Lenovo는 기업들이 워크로드에 대한 ML 및 AI의 이점을 이해하고 적용할 수 있도록 실질적인 접근 방식을 취하고 있습니다. Lenovo 고객은 Lenovo AI Innovation Center의 Lenovo AI 제품을 살펴보고 평가하여 해당 사용 사례의 가치를 완벽하게 파악할 수 있습니다. 가치 창출 시간을 단축하기 위해 이 고객 중심 접근 방식은 AI에 사용하고 최적화할 수 있는 솔루션 개발 플랫폼에 대한 개념 증명을 고객에게 제공합니다.

Lenovo ThinkSystem SE350 Edge 서버

에지 컴퓨팅을 사용하면 데이터 센터 또는 클라우드로 전송되기 전에 네트워크 에지에서 IoT 장치의 데이터를 분석할 수 있습니다. 아래 그림과 같이 Lenovo ThinkSystem SE350은 견고하며 환경 친화적인 소형 폼 팩터에서 유연성, 연결, 보안 및 원격 관리 기능에 중점을 두고 엣지에서의 배포를 위한 고유한 요구 사항을 충족하도록 설계되었습니다.

에지 AI 워크로드에 대한 가속화를 지원할 수 있는 유연성을 갖춘 인텔 제온 D 프로세서를 장착한 SE350은 데이터 센터 외부의 다양한 환경에서 서버 배포의 과제를 해결하기 위해 특별히 제작되었습니다.



MLPerf

MLPerf는 AI 성능 평가를 위한 업계 최고의 벤치마크 제품군입니다. 여기에는 영상 분류, 물체 감지, 의료 영상 및 NLP(자연어 처리)를 비롯한 다양한 적용 AI 영역을 다룹니다. 이 검증에서는 이 검증이 완료될 때 MLPerf 추론의 최신 반복인 Inference v0.7 워크로드를 사용했습니다. 를 클릭합니다 "MLPerf Inference v0.7" 데이터 센터 및 에지 시스템을 위한 새로운 벤치마크 4개가 포함된 제품군:

- * BERT. * Transformers(BERT)의 양방향 Encoder Representation은 Squad 데이터 세트를 사용하여 질문 답변에 맞게 미세 조정되었습니다.
- * DLRM. * DLRM(Deep Learning Recommendation Model)은 CTR(Click-Through Rates)을 최적화하도록 교육받은 개인 설정 및 권장 모델입니다.
- * 3D U-Net. * 3D U-Net 아키텍처는 Brain Tumor Segmentation(뇌종양 분할) 데이터 세트에 대한 교육을 받습니다.
- * RNN-T * Recurrent Neural Network Transducer(RNN-T)는 LibriSpeech의 하위 집합에 대한 교육을 받은 자동 음성 인식(ASR) 모델입니다. MLPerf Inference 결과 및 코드는 공개적으로 사용할 수 있으며 Apache 라이센스에 따라 릴리스됩니다. MLPerf Inference에는 다음과 같은 시나리오를 지원하는 Edge 분산이 있습니다.
- * 단일 스트림. * 이 시나리오는 스마트폰에서 실행되는 오프라인 AI 쿼리와 같이 응답성이 중요한 요소인 시스템을 모방합니다. 개별 쿼리가 시스템으로 전송되고 응답 시간이 기록됩니다. 모든 응답의 90번째 백분위수 지연 시간이 결과로 보고됩니다.
- * 멀티스트림. * 이 벤치마크는 여러 센서의 입력을 처리하는 시스템을 위한 것입니다. 테스트 중에 쿼리는 고정된 시간 간격으로 전송됩니다. QoS 제약(허용되는 최대 지연 시간)이 적용됩니다. QoS 제한을 충족하는 동안 시스템에서 처리할 수 있는 스트림의 수를 보고합니다.
- * Offline. * 배치 처리 응용 프로그램을 다루는 가장 간단한 시나리오이며 메트릭은 초당 샘플 처리량입니다. 모든 데이터를 시스템에서 사용할 수 있으며 벤치마크는 모든 샘플을 처리하는 데 걸리는 시간을 측정합니다.

Lenovo는 이 문서에 사용된 서버인 T4가 포함된 SE350에 대한 MLPerf Inference 점수를 게시했습니다. 의 결과를 참조하십시오 "https://mlperf.org/inference-results-0-7/" 입력 #0.7-145의 "Edge, Closed Division" 섹션에 있습니다.

테스트 계획

이 문서는 MLPerf Inference v0.7을 따릅니다 "코드", MLPerf Inference v1.1 "코드", 및 "규칙". 아래 표에 정의된 대로 에지에서 추론을 위해 설계된 MLPerf 벤치마크를 실행했습니다.

영역	작업	모델	데이터 세트	**QSL** 크기	품질	멀티스트림 지연 제한
비전	영상 분류	Resnet50v1.5	ImageNet(224 x224)	1024	FP32의 99%	50ms
비전	물체 감지(대형)	SSD-ResNet34	코코(1200x1200)	64	FP32의 99%	66ms
비전	물체 감지(소형)	SSD - MobileNetsv1	코코(300x300)	256	FP32의 99%	50ms
비전	의료 영상 분할	3D UNET	2019 (224x224x160)	16	FP32의 99% 및 99.9%	해당 없음
음성	텍스트 음성 변환	RNNT	리브리스페흐(LiBrispeech) 개발 - 청소	2513	FP32의 99%	해당 없음
언어	언어 처리	베르	스쿼드 v1.1	10833	FP32의 99%	해당 없음

다음 표에는 Edge 벤치마크 시나리오가 나와 있습니다.

영역	작업	시나리오
비전	영상 분류	단일 스트림, 오프라인, 멀티스트림
비전	물체 감지(대형)	단일 스트림, 오프라인, 멀티스트림
비전	물체 감지(소형)	단일 스트림, 오프라인, 멀티스트림
비전	의료 영상 분할	단일 스트림, 오프라인
음성	텍스트 음성 변환	단일 스트림, 오프라인
언어	언어 처리	단일 스트림, 오프라인

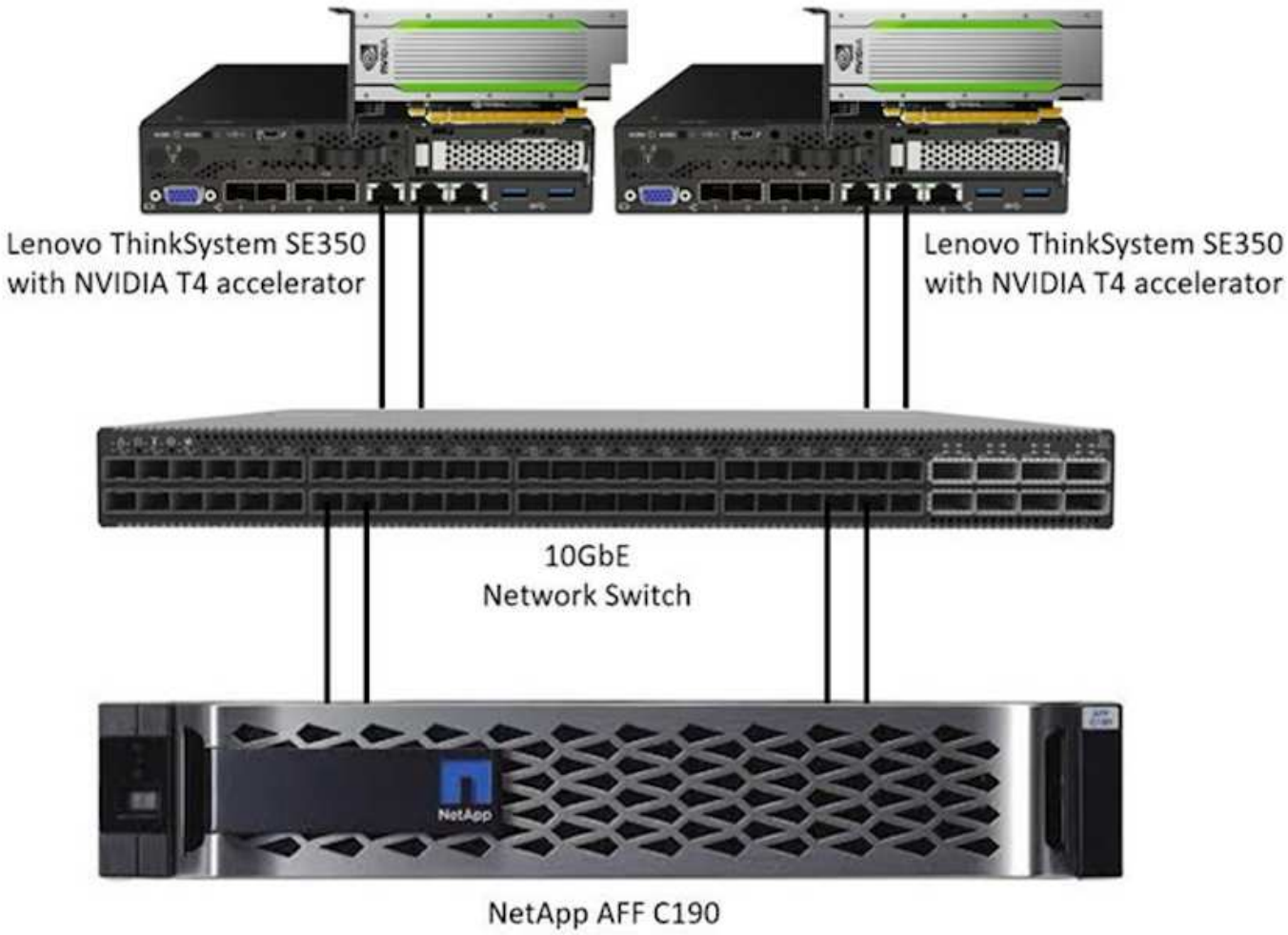
이 검증에서 개발된 네트워크 스토리지 아키텍처를 사용하여 이러한 벤치마크를 수행한 결과 및 이전에 MLPerf에 제출한 에지 서버에서 로컬 실행의 결과를 비교했습니다. 이와 비교하여 공유 스토리지가 추론 성능에 미치는 영향을 확인합니다.

구성을 테스트합니다

다음 그림은 테스트 구성을 보여 줍니다. NetApp AFF C190 스토리지 시스템과 Lenovo ThinkSystem SE350 서버 2대(각각 NVIDIA T4 가속기 1대)를 사용했습니다. 이러한 구성요소는 10GbE 네트워크 스위치를 통해 연결됩니다. 네트워크 스토리지는 검증/테스트 데이터 세트와 사전 교육 모델을 보유하고 있습니다. 서버는 컴퓨팅 기능을 제공하며 스토리지는 NFS 프로토콜을 통해 액세스됩니다.

이 섹션에서는 테스트된 구성, 네트워크 인프라, SE350 서버 및 스토리지 프로비저닝 세부 정보에 대해 설명합니다. 다음 표에서는 솔루션 아키텍처의 기본 구성 요소를 보여 줍니다.

솔루션 구성 요소	세부 정보
Lenovo ThinkSystem 서버	• 각각 NVIDIA T4 GPU 카드 1개가 장착된 SE350 서버 2대
	• 각 서버에는 2.20GHz 및 128GB RAM에서 4개의 물리적 코어가 실행되는 Intel Xeon D-2123IT CPU 1개가 포함되어 있습니다
엔트리 레벨 NetApp AFF 스토리지 시스템(HA 쌍,	• NetApp ONTAP 9 소프트웨어 • 24x 960GB SSD • NFS 프로토콜 • 컨트롤러당 1개의 인터페이스 그룹으로, 마운트 지점에 4개의 논리 IP 주소를 사용합니다



다음 표에는 스토리지 구성이 2RU, 24개 드라이브 슬롯이 포함된 AFF C190에 나와 있습니다.

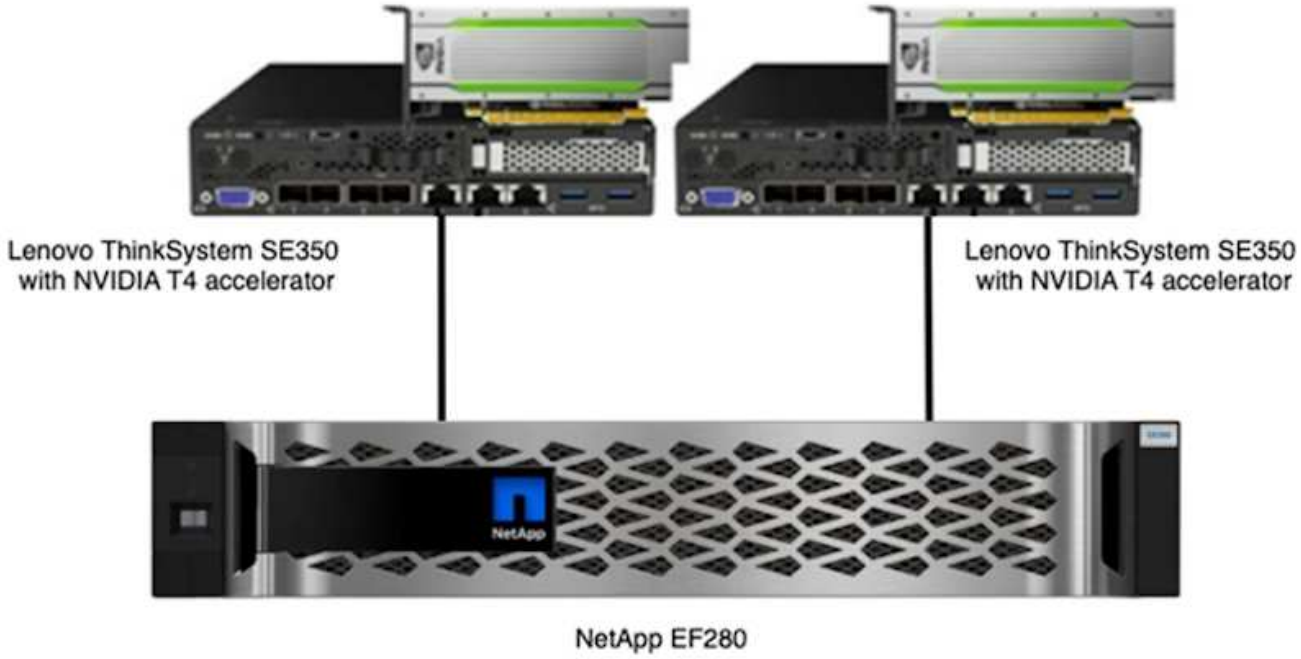
컨트롤러	집계	**FlexGroup** 볼륨	애그리게이트 크기	볼륨 크기	운영 체제 마운트 지점
컨트롤러1	집계1	/netapplenovo_AI_FG	8.42TiB	15TB	/NetApp_Lenovo_FG입니다
컨트롤러 2	집계2		8.42TiB		

/netappLenovo_AI_FG 폴더에는 모델 검증에 사용된 데이터 세트가 포함되어 있습니다.

아래 그림은 테스트 구성을 보여 줍니다. NetApp EF280 스토리지 시스템과 두 개의 Lenovo ThinkSystem SE350 서버(각각 NVIDIA T4 가속기 1개 포함)를 사용했습니다. 이러한 구성요소는 10GbE 네트워크 스위치를 통해 연결됩니다. 네트워크 스토리지는 검증/테스트 데이터 세트와 사전 교육 모델을 보유하고 있습니다. 서버는 컴퓨팅 기능을 제공하며 스토리지는 NFS 프로토콜을 통해 액세스됩니다.

다음 표에는 EF280에 대한 스토리지 구성이 나와 있습니다.

컨트롤러	볼륨 그룹	볼륨	볼륨 크기	**DDPsize**를 참조하십시오	연결 방법
컨트롤러1	DDP1	볼륨 1	8.42TiB	16TB	SE350-1에서 iSCSI LUN 0으로
컨트롤러 2		볼륨 2	8.42TiB		SE350-2를 iSCSI LUN 1로 설정합니다



테스트 절차

이 섹션에서는 이 솔루션을 검증하는 데 사용되는 테스트 절차를 설명합니다.

운영 체제 및 AI 추론 설정

AFF C190의 경우 NVIDIA GPU를 지원하고 MLPerf를 사용하는 NVIDIA 드라이버 및 Docker와 함께 Ubuntu 18.04를 사용했습니다 "코드" MLPerf Inference v0.7에 대한 Lenovo 제출의 일부로 사용할 수 있습니다.

EF280의 경우 NVIDIA GPU 및 MLPerf를 지원하는 Ubuntu 20.04와 NVIDIA 드라이버 및 Docker를 사용했습니다 "코드" MLPerf Inference v1.1에 대한 Lenovo 제출의 일부로 제공됩니다.

AI 추론을 설정하려면 다음 단계를 수행하십시오.

1. 등록이 필요한 데이터 세트, ImageNet 2012 검증 세트, Critio Terabyte 데이터 세트 및 브라츠 2019 교육 세트를 다운로드한 다음 파일의 압축을 풉니다.
2. 최소 1TB의 작업 디렉토리를 생성하고 디렉토리를 참조하는 환경 변수 MLPERF_Scratch_path를 정의합니다.

   네트워크 스토리지 활용 사례나 로컬 데이터로 테스트할 때 로컬 디스크에 대해 공유 스토리지에서 이 디렉토리를 공유해야 합니다.

3. make "prebuild" 명령을 실행하여 필요한 추론 작업을 위해 Docker 컨테이너를 빌드하고 실행합니다.

   > 다음 명령은 실행 중인 Docker 컨테이너 내에서 모두 실행됩니다.

   - MLPerf Inference 태스크에 대한 사전 교육 AI 모델 'MAKE download_model'을 다운로드합니다
   - 무료로 다운로드할 수 있는 추가 데이터셋 'make download_data'를 다운로드하세요
   - 데이터 사전 처리: preprocess_data를 만든다
   - 러닝: 메이크 빌드.
   - 컴퓨팅 서버의 GPU에 최적화된 추론 엔진 'make generate_gservers'를 구축합니다
   - 추론 워크로드를 실행하려면 다음 명령을 실행합니다(하나의 명령).

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

AI 추론 실행

세 가지 유형의 실행이 실행되었습니다.

- 로컬 스토리지를 사용하는 단일 서버 AI 추론
- 네트워크 스토리지를 사용하여 단일 서버 AI 추론
- 네트워크 스토리지를 사용하여 다중 서버 AI 추론

테스트 결과

제안된 아키텍처의 성능을 평가하기 위해 다수의 테스트를 실행했습니다.

6가지 워크로드(영상 분류, 물체 감지[소형], 물체 감지[대형], 의료 영상, 텍스트 음성 변환, 및 NLP(Natural Language Processing))를 사용하여 오프라인, 단일 스트림 및 멀티스트림의 세 가지 시나리오에서 실행할 수 있습니다.

마지막 시나리오는 영상 분류 및 물체 감지에 대해서만 구현됩니다.

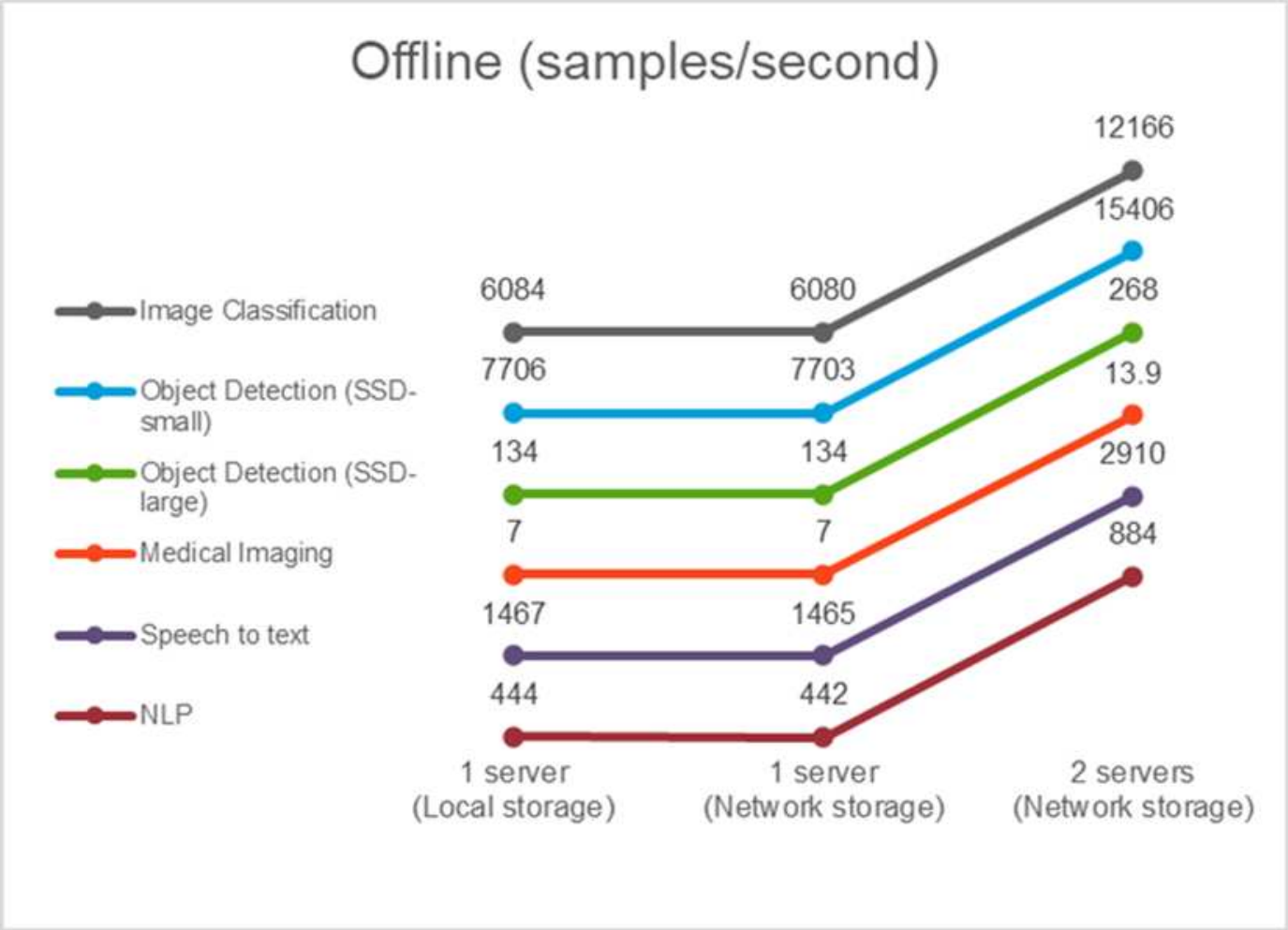
이렇게 하면 다음과 같은 세 가지 다른 설정 하에서 모두 테스트한 15가지 가능한 워크로드가 제공됩니다.

- 단일 서버/로컬 스토리지
- 단일 서버/네트워크 스토리지
- 멀티 서버/네트워크 스토리지

결과는 다음 섹션에 설명되어 있습니다.

AFF의 오프라인 시나리오에서 **AI** 추론을 사용합니다

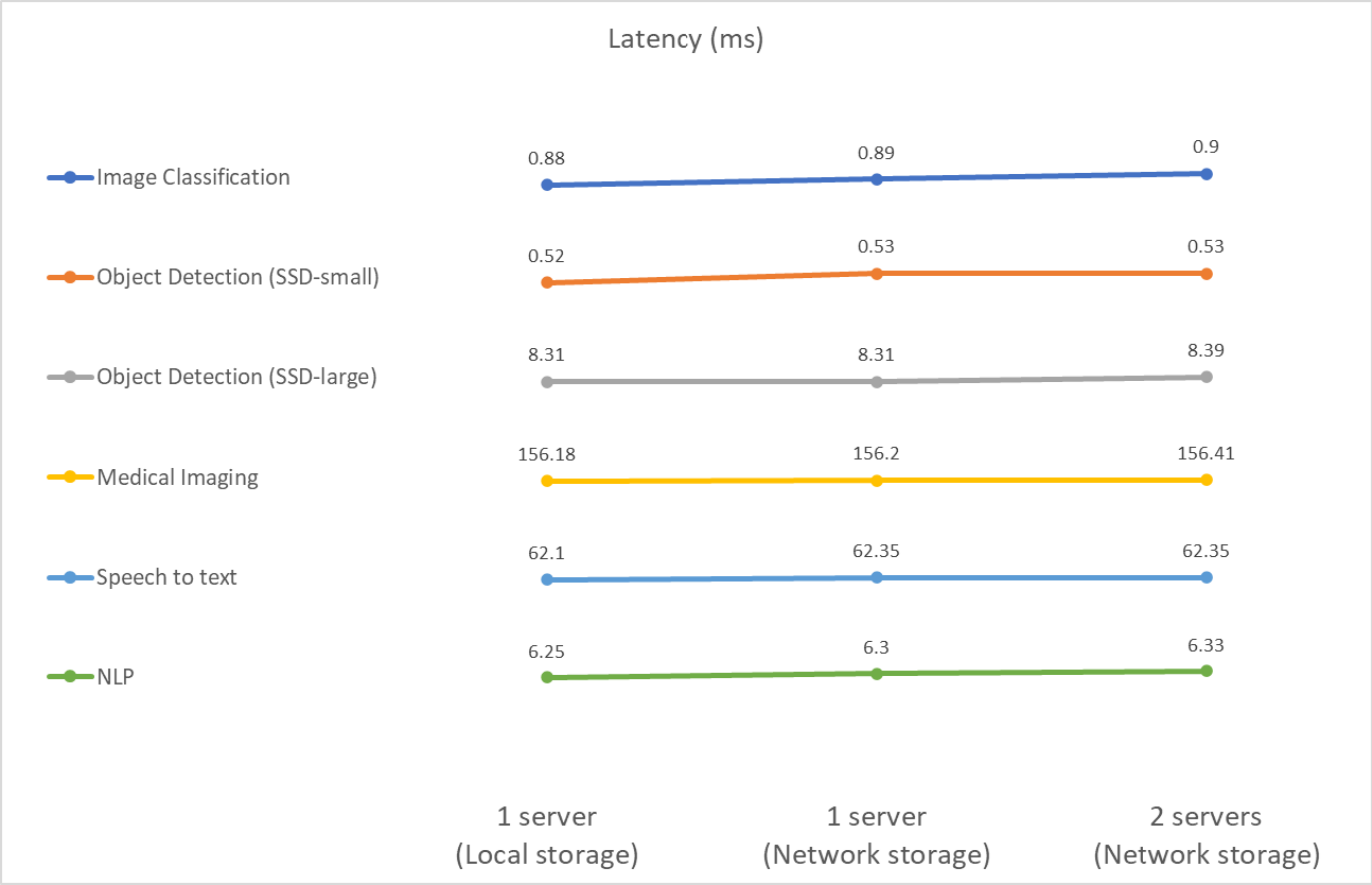
이 시나리오에서는 서버에서 모든 데이터를 사용할 수 있었고 모든 샘플을 처리하는 데 걸린 시간이 측정되었습니다. 테스트 결과로 초당 샘플에 대역폭이 보고됩니다. 두 개 이상의 컴퓨팅 서버를 사용한 경우 모든 서버에 대한 총 대역폭을 합산한 것으로 보고합니다. 아래 그림에서는 세 가지 사용 사례 모두의 결과를 보여 줍니다. 2서버 사례에서는 두 서버의 결합된 대역폭을 보고합니다.



결과에 따르면 네트워크 스토리지는 성능에 부정적인 영향을 주지 않습니다. 변경 사항은 최소이며 일부 작업의 경우 아무것도 발견되지 않습니다. 두 번째 서버를 추가할 때 총 대역폭이 정확히 두 배 또는 최악의 경우 변경률이 1% 미만입니다.

AFF의 단일 스트림 시나리오에서 **AI** 추론을 사용합니다

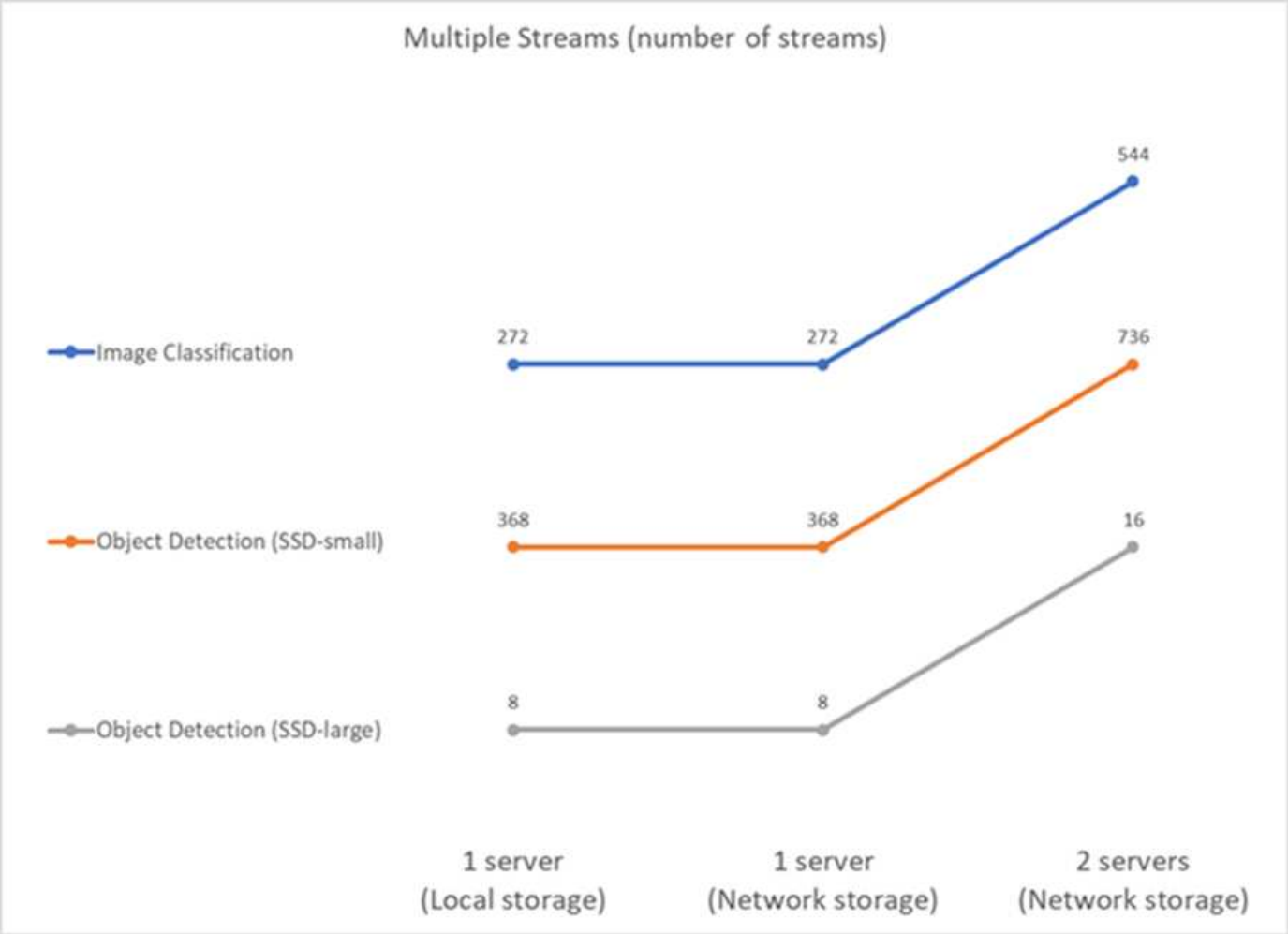
이 벤치마크는 지연 시간을 측정합니다. 여러 계산 서버 사례에서는 평균 지연 시간을 보고합니다. 작업 세트의 결과는 아래 그림에 나와 있습니다. 2서버 사례에서는 두 서버 모두의 평균 지연 시간을 보고합니다.



결과는 네트워크 스토리지가 작업을 처리하기에 충분하다는 것을 다시 한 번 보여 줍니다. 한 서버 케이스에서 로컬 스토리지와 네트워크 스토리지의 차이는 최소 또는 없음입니다. 마찬가지로 두 서버가 동일한 스토리지를 사용하는 경우 두 서버의 지연 시간은 동일하게 유지되거나 매우 적은 양의 변경 사항이 적용됩니다.

AFF의 다중 스트림 시나리오에서 **AI** 추론을 사용합니다

이 경우 결과적으로 QoS 제약 조건을 만족하면서 시스템에서 처리할 수 있는 스트림의 수가 됩니다. 따라서 결과는 항상 정수입니다. 둘 이상의 서버에 대해 모든 서버에 대해 집계된 총 스트림 수를 보고합니다. 모든 워크로드가 이 시나리오를 지원하는 것은 아니지만 이를 실행했습니다. 테스트 결과는 아래 그림에 요약되어 있습니다. 2서버 사례에서는 두 서버 모두에서 스트림 수가 결합된 것으로 보고합니다.



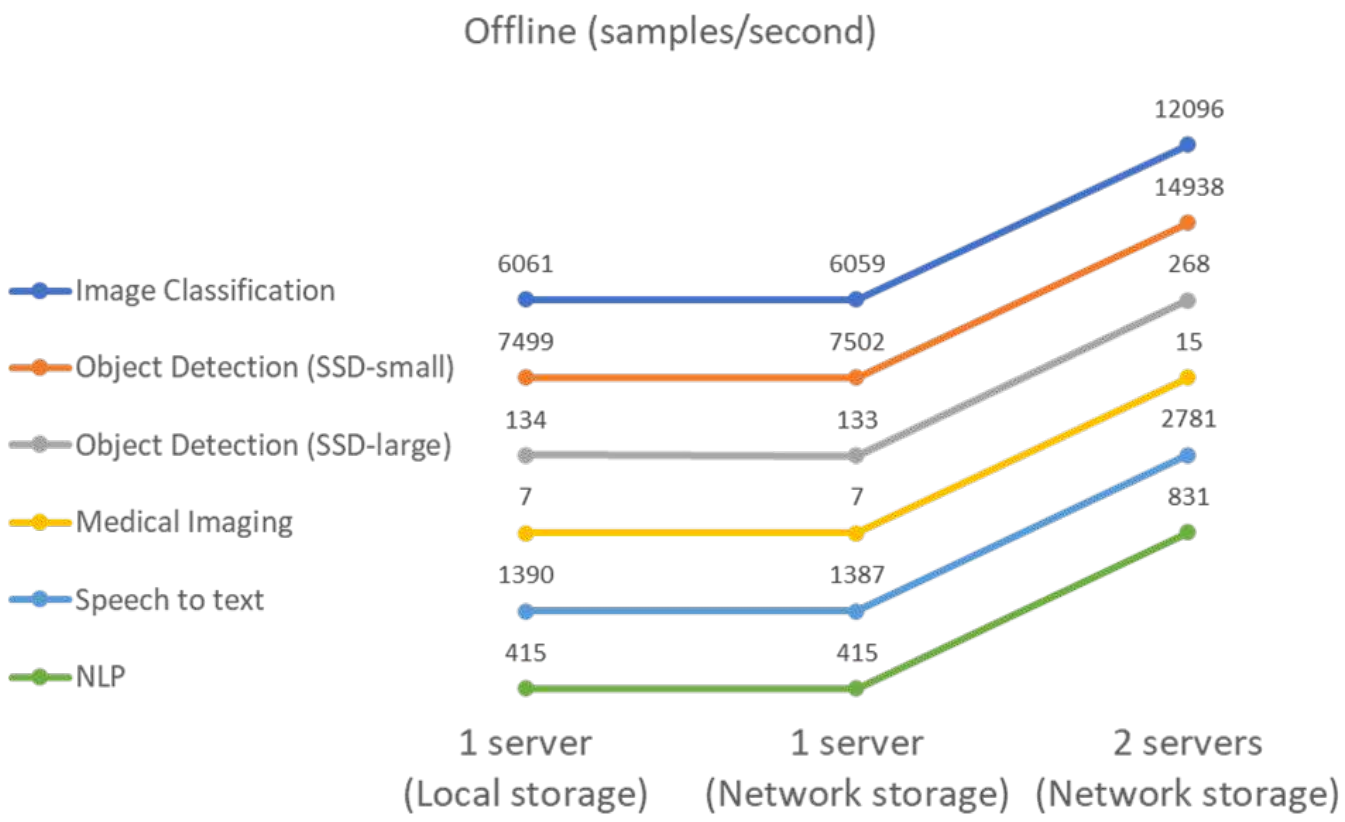
결과는 설정의 완벽한 성능을 보여줍니다. 로컬 및 네트워킹 스토리지는 동일한 결과를 제공하며 두 번째 서버를 추가하면 제안된 설정에서 처리할 수 있는 스트림 수가 두 배가 됩니다.

EF 테스트 결과

제안된 아키텍처의 성능을 평가하기 위해 다수의 테스트를 실행했습니다. 6가지 워크로드(영상 분류, 물체 감지[소형], 물체 감지[대형], 의료 영상, 텍스트 음성 변환, 두 가지 시나리오(오프라인 및 단일 스트림)에서 실행된 자연어 처리[NLP])를 들 수 있습니다. 결과는 다음 섹션에 설명되어 있습니다.

EF의 오프라인 시나리오에서 AI 추론을 사용합니다

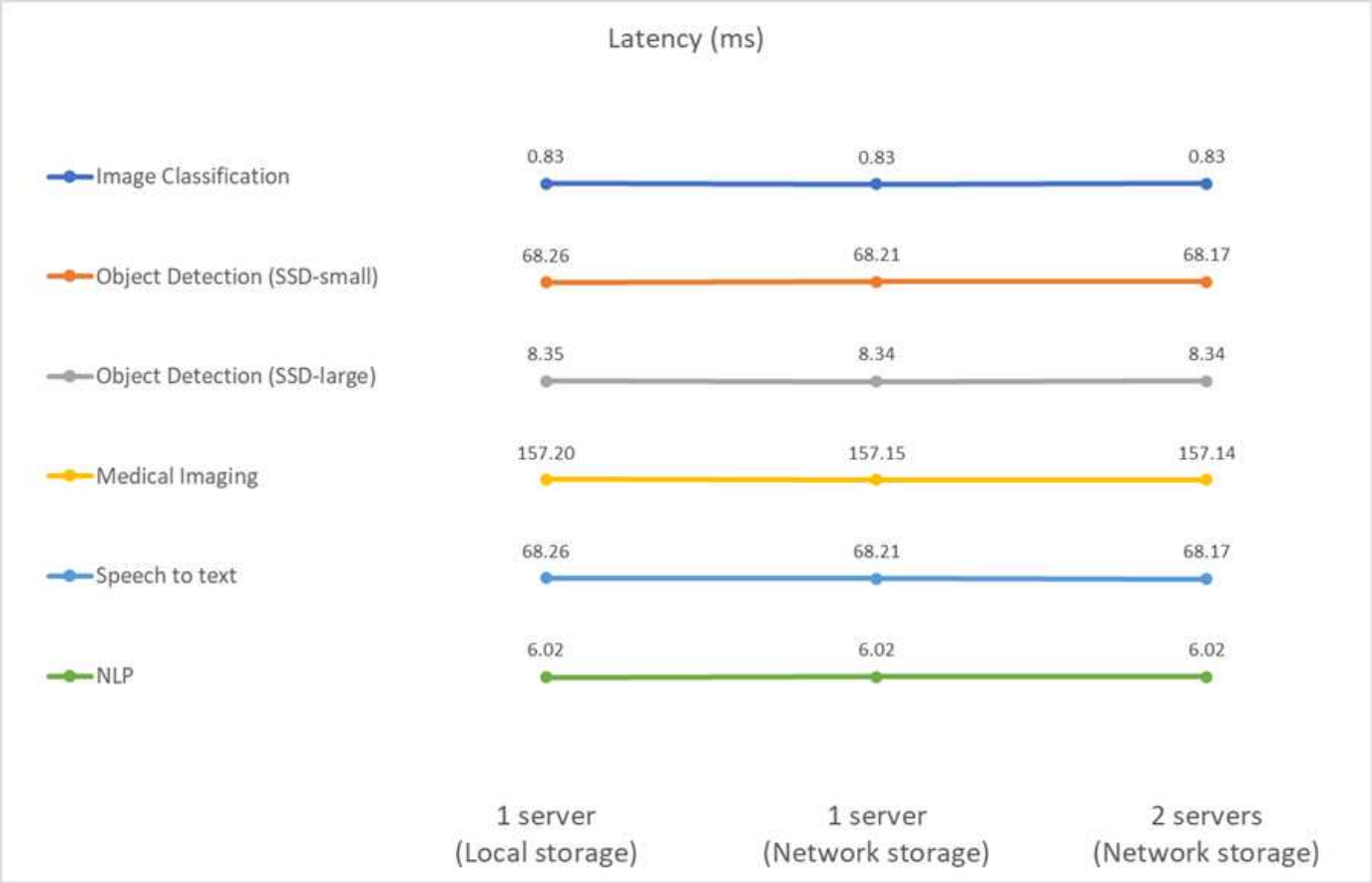
이 시나리오에서는 서버에서 모든 데이터를 사용할 수 있었고 모든 샘플을 처리하는 데 걸린 시간이 측정되었습니다. 테스트 결과로 초당 샘플에 대역폭이 보고됩니다. 단일 노드 실행의 경우 두 서버 모두에서 평균을 보고하며, 두 서버 실행 시 모든 서버에 대해 총 대역폭을 집계합니다. 사용 사례에 대한 결과는 아래 그림에 나와 있습니다.



결과에 따르면 네트워크 스토리지는 성능에 부정적인 영향을 주지 않습니다. 변경 사항은 최소이며 일부 작업의 경우 아무것도 발견되지 않습니다. 두 번째 서버를 추가할 때 총 대역폭이 정확히 두 배 또는 최악의 경우 변경률이 1% 미만입니다.

EF의 단일 스트림 시나리오에서 **AI** 추론을 사용합니다

이 벤치마크는 지연 시간을 측정합니다. 모든 경우에 대해 실행에 관련된 모든 서버의 평균 지연 시간을 보고합니다. 작업 세트의 결과가 제공됩니다.



결과는 네트워크 스토리지가 작업을 처리하기에 충분하다는 것을 다시 보여줍니다. 한 서버 케이스에서 로컬 스토리지와 네트워크 스토리지의 차이는 Minimal(최소) 또는 None(없음)입니다. 마찬가지로 두 서버가 동일한 스토리지를 사용하는 경우 두 서버의 지연 시간은 동일하게 유지되거나 매우 적은 양의 변경 사항이 적용됩니다.

아키텍처 사이징 옵션

다른 사용 사례에 맞게 검증에 사용된 설정을 조정할 수 있습니다.

컴퓨팅 서버

우리는 SE350에서 지원되는 최저 수준의 CPU인 Intel Xeon D-2123IT CPU를 4개의 물리적 코어와 60W TDP로 사용했습니다. 서버는 CPU 교체를 지원하지 않지만 보다 강력한 CPU로 주문할 수 있습니다. 지원되는 최상위 CPU는 16개의 코어가 있는 Intel Xeon D-2183IT, 2.20GHz에서 실행되는 100W입니다. 이렇게 하면 CPU 계산 기능이 크게 향상됩니다. CPU는 추론 워크로드 자체를 실행하는 데 병목 지점이 되지 않지만, 데이터 처리와 추론과 관련된 다른 작업에 도움이 됩니다. 현재 NVIDIA T4는 에지 사용 사례에 사용할 수 있는 유일한 GPU이므로, 현재는 GPU를 업그레이드하거나 다운그레이드할 수 없습니다.

공유 스토리지

테스트 및 검증을 위해 최대 스토리지 용량이 50.5TB, 순차적 읽기의 경우 처리량 4.4GBps, 소규모 랜덤 읽기의 경우 230K IOPS를 지원하는 NetApp AFF C190 시스템이 이 문서의 목적에 사용되었으며 에지 추론 워크로드에 적합한 것으로 입증되었습니다.

그러나 스토리지 용량 또는 더 빠른 네트워킹 속도가 필요한 경우 NetApp AFF A220 또는 을 사용해야 합니다 "NetApp AFF A250" 기술을 자세히 소개합니다. 또한 최대 1.5PB의 용량을 가진 NetApp EF280 시스템도 이 솔루션 검증을

위해 10Gbps 대역폭 사용이 사용되었습니다. 더 높은 대역폭으로 더 많은 스토리지 용량을 원하는 경우, "NetApp EF300" 사용할 수 있습니다.

결론

AI 기반 자동화 및 에지 컴퓨팅은 비즈니스 조직이 디지털 혁신을 달성하고 운영 효율성과 안전을 극대화할 수 있도록 지원하는 선도적인 접근 방식입니다. 에지 컴퓨팅은 데이터 센터와 데이터를 전송할 필요가 없기 때문에 훨씬 더 빠르게 처리됩니다. 따라서 데이터를 데이터 센터 또는 클라우드로 전송하는 데 따른 비용이 절감됩니다. 에지에 구축된 AI 추론 모델을 사용하여 거의 실시간으로 의사 결정을 내려야 하는 경우 지연 시간이 단축되고 속도가 빨라질 수 있습니다.

NetApp 스토리지 시스템은 로컬 SSD 스토리지와 동일하거나 더 우수한 성능을 제공하여 데이터 과학자, 데이터 엔지니어, AI/ML 개발자 및 비즈니스 또는 IT 의사 결정자에게 다음과 같은 이점을 제공합니다.

- AI 시스템, 분석 및 기타 중요한 비즈니스 시스템 간에 데이터를 손쉽게 공유 이러한 데이터 공유는 인프라 오버헤드를 줄이고 성능을 향상하며 기업 전체에서 데이터 관리를 간소화합니다.
- 컴퓨팅과 스토리지를 독립적으로 확장하므로 비용을 최소화하고 리소스 사용량을 높일 수 있습니다.
- 즉각적이고 공간 효율적인 사용자 작업 공간, 통합 버전 제어 및 자동화된 구축을 위해 통합 Snapshot 복사본과 클론을 사용하여 개발 및 구축 워크플로우를 간소화했습니다.
- 재해 복구 및 비즈니스 연속성을 위한 엔터프라이즈급 데이터 보호 기능 이 문서에 제공된 NetApp 및 Lenovo 솔루션은 에지에서 엔터프라이즈급 AI 추론 배포에 이상적인 유연한 스케일아웃 아키텍처입니다.

감사의 말

- J.J. Falkanger, 선임 Lenovo, HPC 및 AI 솔루션 매니저
- Dave Arnette, NetApp 기술 마케팅 엔지니어
- Joey Parnell, 기술 팀장 E-Series AI 솔루션, NetApp
- Cody Harryman, NetApp QA 엔지니어

추가 정보를 찾을 수 있는 위치

이 문서에 설명된 정보에 대한 자세한 내용은 다음 문서 및/또는 웹 사이트를 참조하십시오.

- NetApp AFF A-Series 어레이 제품 페이지 를 참조하십시오

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

- NetApp ONTAP 데이터 관리 소프트웨어 - ONTAP 9 정보 라이브러리

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

- TR-4727: NetApp EF-Series 소개

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

- NetApp E-Series SANtricity 소프트웨어 데이터시트 를 참조하십시오

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

- 컨테이너용 NetApp 영구 스토리지 - 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 복사 및 동기화

  "https://docs.netapp.com/us-en/occm/concept_cloud_sync.html#how-cloud-sync-works"

- TensorFlow 벤치마크

  "https://github.com/tensorflow/benchmarks"

- Lenovo ThinkSystem SE350 Edge 서버

  "https://lenovopress.com/lp1168"

- Lenovo ThinkSystem DM5100F 유니파이드 플래시 스토리지 어레이

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

저작권 정보

Copyright © 2024 NetApp, Inc. All Rights Reserved. 미국에서 인쇄된 본 문서의 어떠한 부분도 저작권 소유자의 사전 서면 승인 없이는 어떠한 형식이나 수단(복사, 녹음, 녹화 또는 전자 검색 시스템에 저장하는 것을 비롯한 그래픽, 전자적 또는 기계적 방법)으로도 복제될 수 없습니다.

NetApp이 저작권을 가진 자료에 있는 소프트웨어에는 아래의 라이센스와 고지사항이 적용됩니다.

본 소프트웨어는 NetApp에 의해 '있는 그대로' 제공되며 상품성 및 특정 목적에의 적합성에 대한 명시적 또는 묵시적 보증을 포함하여(이에 제한되지 않음) 어떠한 보증도 하지 않습니다. NetApp은 대체품 또는 대체 서비스의 조달, 사용 불능, 데이터 손실, 이익 손실, 영업 중단을 포함하여(이에 국한되지 않음), 이 소프트웨어의 사용으로 인해 발생하는 모든 직접 및 간접 손해, 우발적 손해, 특별 손해, 징벌적 손해, 결과적 손해의 발생에 대하여 그 발생 이유, 책임론, 계약 여부, 엄격한 책임, 불법 행위(과실 또는 그렇지 않은 경우)와 관계없이 어떠한 책임도 지지 않으며, 이와 같은 손실의 발생 가능성이 통지되었다 하더라도 마찬가지입니다.

NetApp은 본 문서에 설명된 제품을 언제든지 예고 없이 변경할 권리를 보유합니다. NetApp은 NetApp의 명시적인 서면 동의를 받은 경우를 제외하고 본 문서에 설명된 제품을 사용하여 발생하는 어떠한 문제에도 책임을 지지 않습니다. 본 제품의 사용 또는 구매의 경우 NetApp에서는 어떠한 특허권, 상표권 또는 기타 지적 재산권이 적용되는 라이센스도 제공하지 않습니다.

본 설명서에 설명된 제품은 하나 이상의 미국 특허, 해외 특허 또는 출원 중인 특허로 보호됩니다.

제한적 권리 표시: 정부에 의한 사용, 복제 또는 공개에는 DFARS 252.227-7013(2014년 2월) 및 FAR 52.227-19(2007년 12월)의 기술 데이터-비상업적 품목에 대한 권리(Rights in Technical Data -Noncommercial Items) 조항의 하위 조항 (b)(3)에 설명된 제한사항이 적용됩니다.

여기에 포함된 데이터는 상업용 제품 및/또는 상업용 서비스(FAR 2.101에 정의)에 해당하며 NetApp, Inc.의 독점 자산입니다. 본 계약에 따라 제공되는 모든 NetApp 기술 데이터 및 컴퓨터 소프트웨어는 본질적으로 상업용이며 개인 비용만으로 개발되었습니다. 미국 정부는 데이터가 제공된 미국 계약과 관련하여 해당 계약을 지원하는 데에만 데이터에 대한 전 세계적으로 비독점적이고 양도할 수 없으며 재사용이 불가능하며 취소 불가능한 라이센스를 제한적으로 가집니다. 여기에 제공된 경우를 제외하고 NetApp, Inc.의 사전 서면 승인 없이는 이 데이터를 사용, 공개, 재생산, 수정, 수행 또는 표시할 수 없습니다. 미국 국방부에 대한 정부 라이센스는 DFARS 조항 252.227-7015(b)(2014년 2월)에 명시된 권한으로 제한됩니다.

상표 정보

NETAPP, NETAPP 로고 및 http://www.netapp.com/TM에 나열된 마크는 NetApp, Inc.의 상표입니다. 기타 회사 및 제품 이름은 해당 소유자의 상표일 수 있습니다.