NetApp 및 Lenovo를 통한 Edge AI 추론

NetApp artificial intelligence solutions

NetApp
February 12, 2026

목차

NetApp 및 Lenovo를 통한 Edge AI 추론 . . . 1
TR-4886: 엣지에서의 AI 추론 - NetApp 과 Lenovo ThinkSystem - 솔루션 설계 . . . 1
요약 . . . 1
소개 . . . 1
결론 . . . 5
감사의 말 . . . 6
추가 정보를 찾을 수 있는 곳 . . . 6

NetApp 및 Lenovo를 통한 Edge AI 추론

TR-4886: 엣지에서의 AI 추론 - NetApp 과 Lenovo ThinkSystem - 솔루션 설계

Sathish Thyagarajan, NetApp Miroslav Hodak, Lenovo

이 문서에서는 새로운 애플리케이션 시나리오를 충족하는 엣지 환경에서 NetApp 스토리지 컨트롤러와 Lenovo ThinkSystem 서버에 GPU 기반 인공 지능(AI) 추론을 배포하기 위한 컴퓨팅 및 스토리지 아키텍처를 설명합니다.

요약

첨단 운전자 보조 시스템(ADAS), 산업 4.0, 스마트 시티, 사물 인터넷(IoT)과 같은 여러 가지 새로운 애플리케이션 시나리오에서는 거의 0에 가까운 지연 시간으로 연속적인 데이터 스트림을 처리해야 합니다. 이 문서에서는 이러한 요구사항을 충족하는 엣지 환경에서 NetApp 스토리지 컨트롤러와 Lenovo ThinkSystem 서버에 GPU 기반 인공 지능(AI) 추론을 배포하기 위한 컴퓨팅 및 스토리지 아키텍처를 설명합니다. 이 문서에서는 NVIDIA T4 GPU가 장착된 엣지 서버에서 다양한 추론 작업을 평가하는 업계 표준 MLPerf 추론 벤치마크에 대한 성능 데이터도 제공합니다. 오프라인, 단일 스트림, 다중 스트림 추론 시나리오의 성능을 조사하고, 비용 효율적인 공유 네트워크 스토리지 시스템을 갖춘 아키텍처가 매우 성능이 뛰어나고 여러 에지 서버의 데이터 및 모델 관리를 위한 중앙 지점을 제공한다는 것을 보여줍니다.

소개

기업들은 네트워크 엣지에서 점점 더 엄청난 양의 데이터를 생성하고 있습니다. 스마트 센서와 IoT 데이터에서 최대 가치를 얻기 위해 기업들은 엣지 컴퓨팅을 지원하는 실시간 이벤트 스트리밍 솔루션을 찾고 있습니다. 따라서 계산적으로 많은 것을 요구하는 작업은 데이터 센터 외부의 엣지에서 수행되는 경우가 점점 더 많아지고 있습니다. AI 추론은 이러한 추세를 주도하는 요인 중 하나입니다. 엣지 서버는 특히 가속기를 사용할 때 이러한 작업 부하에 충분한 컴퓨팅 성능을 제공하지만, 제한된 저장 용량은 특히 다중 서버 환경에서 종종 문제가 됩니다. 이 문서에서는 엣지 환경에서 공유 스토리지 시스템을 배포하는 방법과 성능 저하 없이 AI 추론 워크로드에 어떤 이점을 제공하는지 보여줍니다.

이 문서에서는 엣지에서의 AI 추론을 위한 참조 아키텍처를 설명합니다. 여러 대의 Lenovo ThinkSystem 엣지 서버와 NetApp 스토리지 시스템을 결합하여 배포와 관리가 쉬운 솔루션을 만듭니다. 이 가이드는 여러 카메라와 산업용 센서가 설치된 공장 현장, 소매 거래에서의 POS(판매 시점 관리) 시스템, 자율주행차에서 시각적 이상을 식별하는 FSD(완전 자율 주행) 시스템 등 다양한 상황에서 실제 배포를 위한 기준 가이드를 제공하고자 작성되었습니다.

이 문서에서는 Lenovo ThinkSystem SE350 Edge 서버와 엔트리 레벨 NetApp AFF 및 EF 시리즈 스토리지 시스템으로 구성된 컴퓨팅 및 스토리지 구성의 테스트와 검증에 대해 설명합니다. 참조 아키텍처는 NetApp ONTAP 및 NetApp SANtricity 데이터 관리 소프트웨어를 통해 포괄적인 데이터 서비스, 통합 데이터 보호, 원활한 확장성, 클라우드 연결 데이터 스토리지를 제공하는 동시에 AI 배포를 위한 효율적이고 비용 효율적인 솔루션을 제공합니다.

타겟 고객층

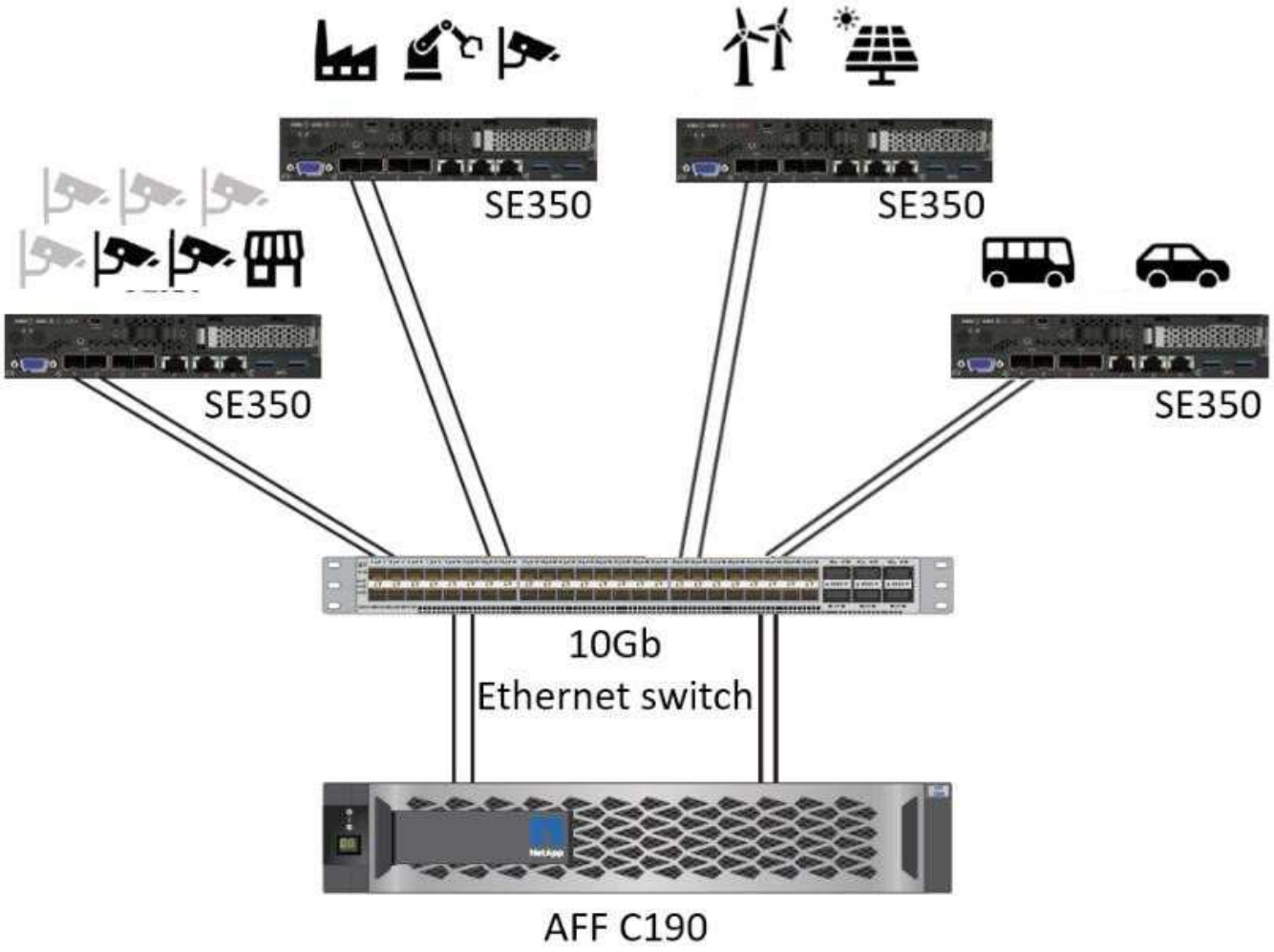
이 문서는 다음과 같은 독자를 대상으로 합니다.

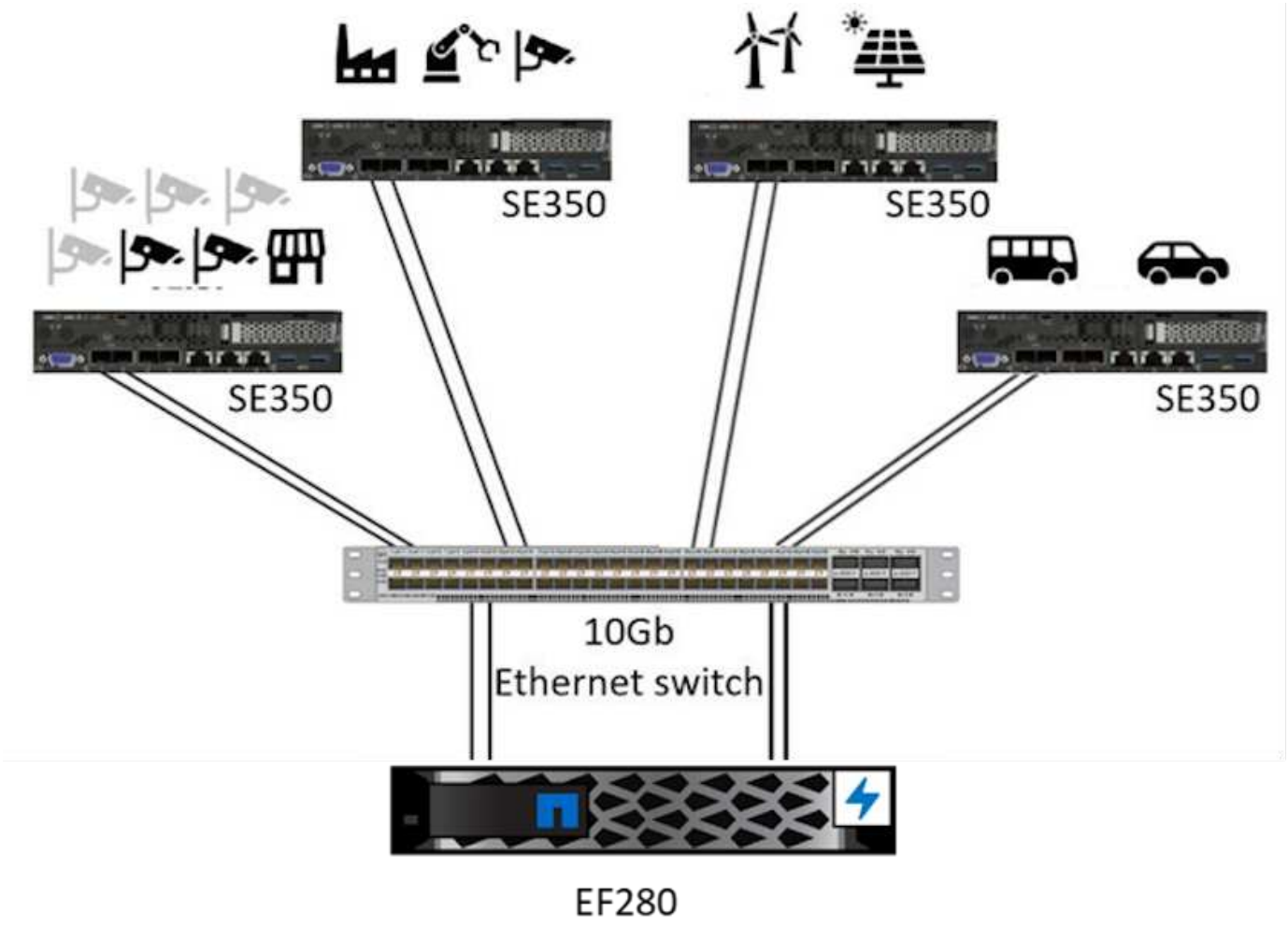
- 엣지에서 AI를 상품화하고자 하는 비즈니스 리더와 엔터프라이즈 아키텍트.
- 데이터 과학자, 데이터 엔지니어, AI/머신러닝(ML) 연구자, AI 시스템 개발자.
- AI/ML 모델과 애플리케이션 개발을 위한 솔루션을 설계하는 엔터프라이즈 아키텍트입니다.

- 딥 러닝(DL) 및 ML 모델을 배포하는 효율적인 방법을 찾고 있는 데이터 과학자와 AI 엔지니어.
- 에지 추론 모델의 배포와 관리를 담당하는 에지 장치 관리자와 에지 서버 관리자입니다.

솔루션 아키텍처

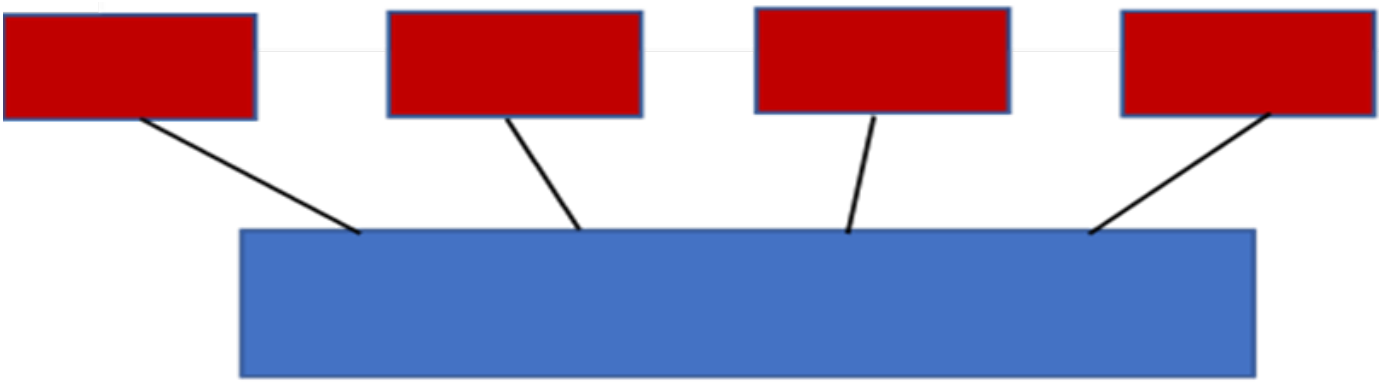
이 Lenovo ThinkSystem 서버와 NetApp ONTAP 또는 NetApp SANtricity 스토리지 솔루션은 기존 CPU와 함께 GPU의 처리 능력을 활용하여 대규모 데이터 세트에 대한 AI 추론을 처리하도록 설계되었습니다. 이 검증은 다음 두 그림에서 볼 수 있듯이 단일 NetApp AFF 스토리지 시스템과 상호 연결된 단일 또는 여러 개의 Lenovo SR350 엣지 서버를 사용하는 아키텍처를 통해 높은 성능과 최적의 데이터 관리를 보여줍니다.





다음 그림의 논리적 아키텍처 개요는 이 아키텍처에서 컴퓨팅 및 스토리지 요소의 역할을 보여줍니다. 구체적으로는 다음 사항을 보여줍니다.

- 카메라, 센서 등에서 수신한 데이터에 대한 추론을 수행하는 엣지 컴퓨팅 장치입니다.
- 여러 가지 목적을 제공하는 공유 저장 요소:
  - 추론을 수행하는 데 필요한 추론 모델 및 기타 데이터의 중앙 위치를 제공합니다. 컴퓨팅 서버는 저장소에 직접 액세스하고 로컬에 복사할 필요 없이 네트워크 전반에서 추론 모델을 사용합니다.
  - 업데이트된 모델이 여기에 게시됩니다.
  - 나중에 분석하기 위해 에지 서버가 수신하는 입력 데이터를 보관합니다. 예를 들어, 에지 장치가 카메라에 연결된 경우, 저장 요소는 카메라가 촬영한 비디오를 보관합니다.



빨간색	파란색
레노버 컴퓨팅 시스템	NetApp AFF 스토리지 시스템
카메라, 센서 등에서 입력된 내용에 대한 추론을 수행하는 에지 장치입니다.	추후 분석을 위해 에지 디바이스의 추론 모델과 데이터를 보관하는 공유 스토리지입니다.

NetApp 과 Lenovo 솔루션은 다음과 같은 주요 이점을 제공합니다.

- 엣지에서의 GPU 가속 컴퓨팅.
- 공유 스토리지에서 지원 및 관리되는 여러 개의 엣지 서버를 배포합니다.
- 데이터 손실 없이 낮은 복구 지점 목표(RPO) 및 복구 시간 목표(RTO)를 충족하는 강력한 데이터 보호 기능을 제공합니다.
- NetApp 스냅샷 복사본과 복제본을 사용하여 데이터 관리를 최적화하고 개발 워크플로를 간소화합니다.

이 아키텍처를 사용하는 방법

이 문서는 제안된 아키텍처의 설계와 성능을 검증합니다. 그러나 컨테이너, 워크로드, 모델 관리 및 온프레미스 클라우드나 데이터 센터와의 데이터 동기화 등 특정 소프트웨어 수준의 부분은 배포 시나리오에 따라 달라지기 때문에 테스트하지 않았습니다. 여기에는 여러 가지 선택이 있습니다.

컨테이너 관리 수준에서 Kubernetes 컨테이너 관리가 좋은 선택이며, 완전한 업스트림 버전(Canonical)이나 엔터프라이즈 배포에 적합한 수정된 버전(Red Hat)에서 모두 잘 지원됩니다. 그만큼"NetApp AI 제어 평면" NetApp Trident 와 새로 추가된 기능을 사용하는 "NetApp DataOps 툴킷" 데이터 과학자와 데이터 엔지니어가 NetApp 스토리지와 통합할 수 있는 내장형 추적 기능, 데이터 관리 기능, 인터페이스 및 도구를 제공합니다. Kubernetes용 ML 툴킷인 Kubeflow는 TensorFlow Serving이나 NVIDIA Triton Inference Server 등 여러 플랫폼에서 모델 버전 관리 및 KFServing을 지원하는 것과 함께 추가적인 AI 기능을 제공합니다. 또 다른 옵션은 GPU 지원 AI 추론 컨테이너 카탈로그에 대한 액세스와 함께 워크로드 관리를 제공하는 NVIDIA EGX 플랫폼입니다. 그러나 이러한 옵션을 프로덕션에 적용하려면 상당한 노력과 전문 지식이 필요할 수 있으며, 제3자 독립 소프트웨어 공급업체(ISV)나 컨설턴트의 도움이 필요할 수도 있습니다.

솔루션 영역

AI 추론과 엣지 컴퓨팅의 주요 이점은 장치가 지연 없이 높은 수준의 품질로 데이터를 계산, 처리 및 분석할 수 있는 능력입니다. 이 문서에서 설명하기에는 엣지 컴퓨팅 사용 사례가 너무 많지만, 몇 가지 대표적인 사례를 소개하겠습니다.

자동차: 자율주행차

엣지 컴퓨팅의 대표적인 예는 자율주행차(AV)의 첨단 운전자 지원 시스템(ADAS)입니다. 자율주행차에 사용되는 AI는 카메라와 센서에서 수집한 방대한 데이터를 빠르게 처리해야 안전하게 주행할 수 있습니다. 물체와 사람 간의 정보를 해석하는 데 너무 오랜 시간이 걸리면 생사가 결정될 수 있으므로, 차량에 최대한 가까운 곳에서 해당 데이터를 처리하는 것이 중요합니다. 이 경우, 하나 이상의 엣지 컴퓨팅 서버가 카메라, RADAR, LiDAR 및 기타 센서의 입력을 처리하는 반면, 공유 스토리지는 추론 모델을 보관하고 센서의 입력 데이터를 저장합니다.

의료: 환자 모니터링

AI와 엣지 컴퓨팅의 가장 큰 영향 중 하나는 재택 치료와 중환자실(ICU) 모두에서 만성 질환 환자의 지속적인 모니터링을 강화할 수 있는 능력입니다. 인슐린 수치, 호흡, 신경 활동, 심장 리듬, 위장 기능을 모니터링하는 에지 디바이스에서 수집된 데이터는 즉각적인 분석이 필요하며, 누군가의 생명을 구하기 위해 조치를 취할 시간이 제한되어 있기 때문에 즉각적인 조치가 필요합니다.

소매: 계산원 없는 결제

엣지 컴퓨팅은 AI와 ML을 구동하여 소매업체가 결제 시간을 줄이고 고객 수를 늘리는 데 도움이 될 수 있습니다. 무계산 시스템은 다음과 같은 다양한 구성 요소를 지원합니다.

- 인증 및 접근. 실제 쇼핑객을 검증된 계정에 연결하고 소매 공간에 대한 접근을 허용합니다.
- 재고 모니터링. 센서, RFID 태그, 컴퓨터 비전 시스템을 사용하여 쇼핑객이 품목을 선택하거나 선택 취소한 것을 확인하는 데 도움을 줍니다.

  여기에서는 각 엣지 서버가 각 체크아웃 카운터를 처리하고 공유 저장 시스템이 중앙 동기화 지점 역할을 합니다.

금융 서비스: 키오스크에서의 인적 안전 및 사기 방지

은행 기관들은 AI와 엣지 컴퓨팅을 활용해 혁신을 이루고 개인화된 은행 경험을 창출하고 있습니다. 실시간 데이터 분석과 AI 추론을 활용하는 대화형 키오스크를 통해 ATM은 고객이 돈을 인출하는 것을 도울 뿐만 아니라 카메라에서 촬영한 이미지를 통해 키오스크를 사전에 모니터링하여 인간의 안전에 대한 위험이나 사기 행위를 파악할 수 있습니다. 이 시나리오에서는 엣지 컴퓨팅 서버와 공유 스토리지 시스템이 대화형 키오스크와 카메라에 연결되어 은행이 AI 추론 모델을 사용하여 데이터를 수집하고 처리하는 데 도움이 됩니다.

제조업: 산업 **4.0**

스마트 팩토리, 3D 프린팅 등의 새로운 트렌드와 함께 제4차 산업혁명(Industry 4.0)이 시작되었습니다. 데이터 중심의 미래에 대비하기 위해 대규모 M2M(기계 간 통신)과 IoT가 통합되어 인간의 개입 없이도 자동화가 더욱 강화됩니다. 제조업은 이미 높은 수준으로 자동화되어 있으며, AI 기능을 추가하는 것은 이러한 장기적 추세의 자연스러운 지속입니다. AI는 컴퓨터 비전과 기타 AI 기능의 도움으로 자동화할 수 있는 작업을 자동화할 수 있습니다. 공장 현장의 조립 라인에서 재료를 더 빠르게 분석하여 품질 관리나 인간의 시각 또는 의사 결정에 의존하는 작업을 자동화하여 제조 공장이 안전 및 품질 관리에 대한 필수 ISO 표준을 충족하도록 도울 수 있습니다. 여기에서 각 컴퓨팅 엣지 서버는 제조 공정을 모니터링하는 센서 어레이에 연결되고, 필요에 따라 업데이트된 추론 모델이 공유 스토리지에 푸시됩니다.

통신: 녹 탐지, 타워 검사 및 네트워크 최적화

통신 산업에서는 컴퓨터 비전과 AI 기술을 사용하여 이미지를 처리하여 녹을 자동으로 감지하고 부식이 있는 셀 타워를 식별하여 추가 검사가 필요합니다. 최근 몇 년 동안 드론 이미지와 AI 모델을 사용하여 타워의 특정 영역을 식별하고 녹, 표면 균열 및 부식을 분석하는 방식이 증가했습니다. 통신 인프라와 셀 타워를 효율적으로 검사하고, 정기적으로 성능 저하 여부를 평가하고, 필요할 경우 신속하게 수리할 수 있는 AI 기술에 대한 수요는 계속해서 증가하고 있습니다.

또한, 통신 분야에서 떠오르는 또 다른 활용 사례는 AI와 ML 알고리즘을 사용하여 데이터 트래픽 패턴을 예측하고, 5G 지원 장치를 감지하고, 다중 입력 및 다중 출력(MIMO) 에너지 관리를 자동화하고 증강하는 것입니다. MIMO 하드웨어는 네트워크 용량을 늘리기 위해 무선 타워에 사용됩니다. 그러나 여기에는 추가 에너지 비용이 발생합니다. 셀 사이트에 배치된 "MIMO 슬립 모드"를 위한 ML 모델은 무선 장치의 효율적인 사용을 예측하고 이동통신 사업자(MNO)의 에너지 소비 비용을 줄이는 데 도움이 될 수 있습니다. AI 추론 및 엣지 컴퓨팅 솔루션은 MNO가 데이터 센터로 전송되는 데이터 양을 줄이고, TCO를 낮추고, 네트워크 운영을 최적화하고, 최종 사용자를 위한 전반적인 성능을 개선하는 데 도움이 됩니다.

결론

AI 기반 자동화 및 엣지 컴퓨팅은 기업 조직이 디지털 전환을 달성하고 운영 효율성과 안전성을 극대화하는 데 도움이 되는 선도적인 접근 방식입니다. 엣지 컴퓨팅을 사용하면 데이터가 데이터 센터로 이동하거나 데이터 센터에서 전송되지 않으므로 훨씬 빠르게 처리됩니다. 따라서 데이터 센터나 클라우드로 데이터를 주고받는 데 드는 비용이 절감됩니다. 기업이 엣지에 구축된 AI 추론 모델을 사용하여 거의 실시간으로 의사 결정을 내려야 하는 경우, 지연 시간을 줄이고 속도를

높이는 것이 유익할 수 있습니다.

NetApp 스토리지 시스템은 로컬 SSD 스토리지와 동일하거나 더 나은 성능을 제공하며 데이터 과학자, 데이터 엔지니어, AI/ML 개발자, 비즈니스 또는 IT 의사 결정권자에게 다음과 같은 이점을 제공합니다.

- AI 시스템, 분석 및 기타 중요 비즈니스 시스템 간에 데이터를 손쉽게 공유할 수 있습니다. 이러한 데이터 공유를 통해 인프라 오버헤드가 줄어들고, 성능이 향상되며, 기업 전체의 데이터 관리가 간소화됩니다.
- 비용을 최소화하고 리소스 사용을 개선하기 위해 독립적으로 확장 가능한 컴퓨팅 및 스토리지.
- 즉각적이고 공간 효율적인 사용자 작업 공간, 통합 버전 제어, 자동화된 배포를 위한 통합 스냅샷 복사본과 복제를 사용하여 개발 및 배포 워크플로를 간소화합니다.
- 재해 복구 및 비즈니스 연속성을 위한 엔터프라이즈급 데이터 보호. 이 문서에 소개된 NetApp 과 Lenovo 솔루션은 엣지에서 엔터프라이즈급 AI 추론을 배포하는 데 이상적인 유연하고 확장 가능한 아키텍처입니다.

감사의 말

- 제이제이 Falkanger, Lenovo HPC 및 AI 솔루션 수석 관리자
- Dave Arnette, NetApp 기술 마케팅 엔지니어
- Joey Parnell, NetApp E-Series AI 솔루션 기술 책임자
- 코디 해리먼, NetApp QA 엔지니어

추가 정보를 찾을 수 있는 곳

이 문서에 설명된 정보에 대해 자세히 알아보려면 다음 문서 및/또는 웹사이트를 참조하세요.

- NetApp AFF A-시리즈 어레이 제품 페이지

  "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 시리즈 소개

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

- NetApp E-시리즈 SANtricity 소프트웨어 데이터시트

  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"

- 레노버 ThinkSystem SE350 엣지 서버

  "https://lenovopress.com/lp1168"

- 레노버 ThinkSystem DM5100F 통합 플래시 스토리지 어레이

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

저작권 정보

상표 정보