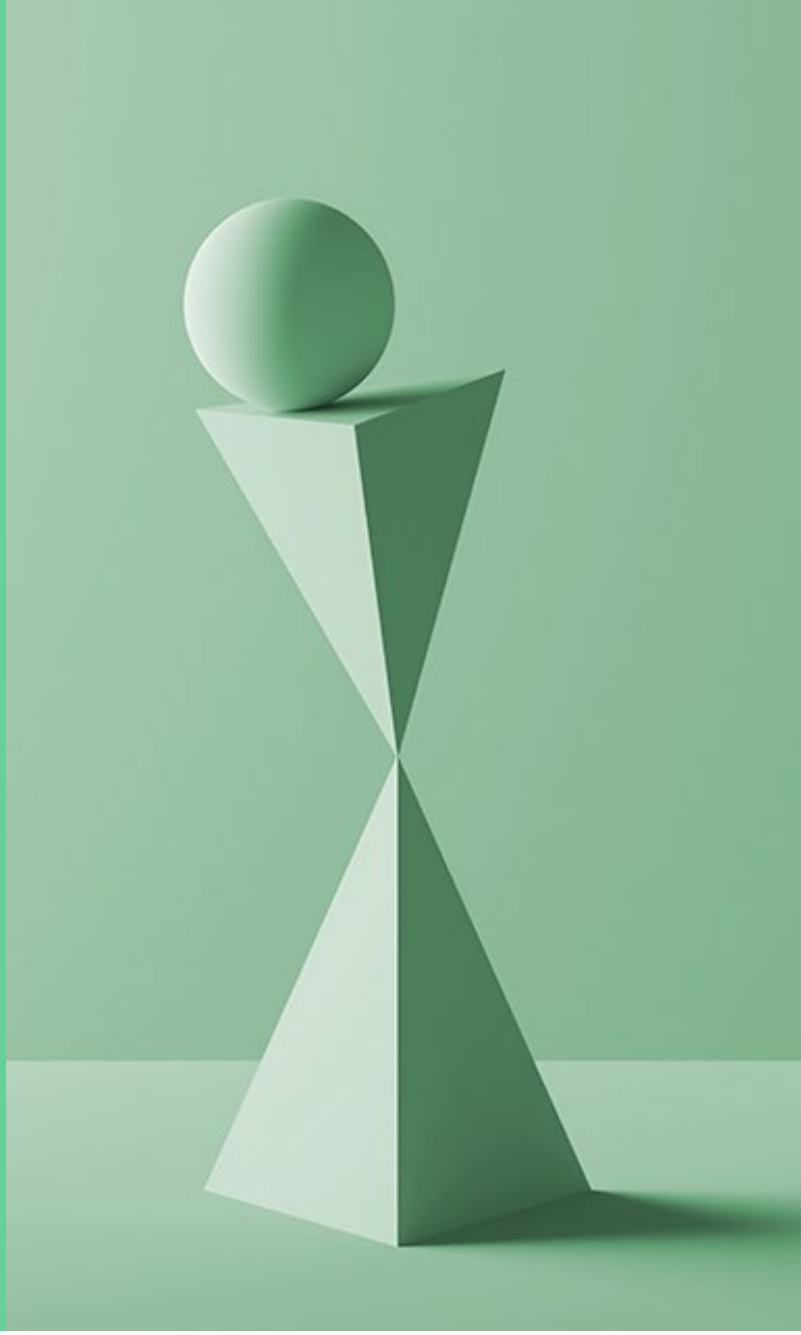
솔루션 개요

ONTAP AI

NetApp 및 NVIDIA를 통해 ML 및 DL을 위한 데이터 파이프라인을 간소화, 가속화, 통합



NetApp | NVIDIA

AI 인프라 과제

점점 더 경쟁이 치열해지는 시장에서 인공 지능 (AI), 머신 러닝 (ML) 및 딥 러닝 (DL) 을 통해 기업은 사기를 탐지하고 고객 관계를 개선하며 공급망을 최적화하고 혁신적인 제품과 서비스를 제공할 수 있습니다 . 귀사도 디지털 혁신을 유도하고 경쟁 우위를 확보하기 위해 새로운 AI 접근법을 활용하는 수많은 조직 중 하나일 수 있습니다 . AI 에서 최대한의 이점을 얻으려면 먼저 몇 가지 주요 과제를 해결해야 합니다 .

DIY 통합은 복잡합니다 . 기성품인 ML 및 DL 컴퓨팅 , 스토리지 , 네트워킹 및 소프트웨어 구성 요소를 조립 및 통합하면 복잡성이 증가하고 구축 시간이 길어질 수 있습니다 . 결과적으로 시스템 통합 작업에 귀중한 데이터 과학 리소스가 낭비됩니다 .

예측 가능하고 확장 가능한 성능을 달성하는 것은 어렵습니다 . DL 모범 사례에서는 작은 규모로 시작한 후 성장에 따라 확장할 것을 제안합니다 . 기존에는 컴퓨팅과 직접 연결 스토리지가 데이터를 AI 워크플로에 공급하는 데 사용되고 있습니다 . 그러나 기존 스토리지를 확장하면 진행 중인 운영에 장애가 발생할 수 있으며 다운타임이 발생할 수 있습니다 .

작업 중단은 데이터 과학자의 생산성에 영향을 미칩니다 . ML 및 DL 인프라는 수많은 하드웨어 및 소프트웨어 상호 종속성을 수반합니다 . 인프라를 계속 유지하려면 전체 스택 AI 에 대한 깊이 있는 전문 지식이 필요합니다 . 다운타임이나 AI 의 느린 성능은 개발자의 생산성을 저하하고 운영 비용이 통제 불능 상태가 되는 연쇄 반응을 일으킬 수 있습니다 .

솔루션

이제 AI, ML 및 DL 의 약속을 완벽하게 실현할 수 있습니다 . NVIDIA DGX™ 시스템과 NetApp 클라우드 연결형 All-Flash 스토리지 기반의 검증된 NetApp® ONTAP® AI 아키텍처를 사용하여 데이터 파이프라인을 간소화 , 가속 , 통합합니다 . 데이터 흐름을 안정적으로 간소화하고 에지에서 코어 및 클라우드에 이르는 Data Fabric 을 사용하여 분석 , 훈련 , 추론의 속도를 높일 수 있습니다 .

주요 이점

유연하고 검증된 솔루션으로 위험 완화

- 설계의 복잡성과 추측에 의한 작업을 제거하여 신속하게 작업 처리
- 사전 구성된 솔루션으로 구성 및 구축 간소화

정확한 성능 및 확장성 제공

- 작게 시작한 후 중단 없이 확장
- 고성능 솔루션으로 신속하게 결과 확인

통합 데이터 파이프라인 구축

- 에지에서 코어 , 클라우드에 이르기까지 통합 파이프라인으로 데이터를 지능적으로 관리
- 솔루션 구축 시 AI 전문가의 지원 및 간단한 지원 옵션 활용

AI 워크로드 통합

- 인프라 사일로 제거
- 비즈니스 수요에 유연하게 대처

NetApp ONTAP AI 는 세계 최초의 5 페타플롭 AI 시스템인 NVIDIA DGX A100 시스템과 NVIDIA Mellanox 고성능 이더넷 스위치를 통합한 최초의 컨버지드 인프라 스택 중 하나로 , AI 워크로드를 통합하고 구축을 단순화하고 투자 수익률 회수를 앞당겨 줍니다 .

" 딥 러닝은 거의 모든 시장에서 혁명을 일으키고 있습니다 . NetApp 은 다양한 시장에서 딥 러닝을 적용하여 미래의 기술을 주도하고 있습니다 . NVIDIA DGX 시스템 및 NetApp All-Flash 스토리지 기반 NetApp ONTAP AI 는 딥 러닝을 위한 데이터 파이프라인을 단순화 및 가속합니다 ."

Tim Ensor, 인공 지능 책임자
케임브리지 컨설턴트

그림 1) DGX A100, 2, 4, 8 노드 구성의 ONTAP AI 아키텍처

유연하고 검증된 솔루션으로 위험 완화

AI 혁신의 빠른 속도로 인해 AI 인프라를 효과적으로 설계하기 어렵습니다 . ONTAP AI 를 통해 현장에서 검증된 참조 아키텍처를 사용하면 불확실한 추측을 없애고 더 빠르게 시작할 수 있습니다 . 또는 복잡한 설계 및 관리 작업을 수행할 필요 없이 쉽게 조달하고 구축할 수 있는 사전 구성된 통합 솔루션을 선택할 수 있습니다 .

ONTAP AI 통합 솔루션은 사전 구성된 4 가지 옵션으로 제공되며 선택에 따라 용량을 확장하거나 고급 소프트웨어를 사용할 수 있습니다 . 이 통합 솔루션은 한 번의 통화로 현장 설치 및 사고 보고에서 해결에 이르기까지 포괄적인 지원을 제공하므로 복잡성이 더욱더 감소합니다 .

정확한 성능 및 확장성 제공

DL 훈련 루틴에는 엄청난 양의 컴퓨팅 파워가 필요합니다 . 더 빠른 이미지 훈련을 통해 전체적인 컴퓨팅 비용을 줄이는 동시에 AI 혁신과 생산성을 높일 수 있습니다 .

새로운 NVIDIA Ampere 아키텍처를 사용하여 구축된 DGX A100 시스템은 이전 세대보다 최대 6 배의 훈련 성능을 제공합니다 . 이제 데이터 센터에 상응하는 분석 , 훈련 및 추론용 컴퓨팅 인프라를 단일 시스템에 통합할 수 있습니다 . DGX A100 시스템은 CPU 시스템과 비교하여 1/25 의 공간과 1/20 의 전력을 사용하며 비용은 단 1/10 에 불과합니다 .

첨단 컴퓨팅에 투자하려면 초당 수천 개의 훈련 이미지를 처리할 수 있는 최첨단 스토리지가 필요합니다 . 가장 까다로운 DL 훈련 워크로드를 지원할 수 있는 고성능 데이터 서비스 솔루션이 필요합니다 .

NetApp All-Flash 스토리지를 활용하면 2GBps 이상의 일관된 처리량 (최대 5GBps) 을 확보할 수 있습니다 . 또한 , 1 밀리초 미만의 지연 시간과 95% 이상의 GPU 활용률을 실현할 수 있습니다 . NAS 워크로드의 경우 단일 NetApp AFF A800 시스템은 순차적 읽기에서 25GBps 의 처리량을 지원하고 , 크기가 작은 랜덤 읽기에서 500µs 미만의 지연 시간으로 100 만 IOPS 를 지원합니다 .

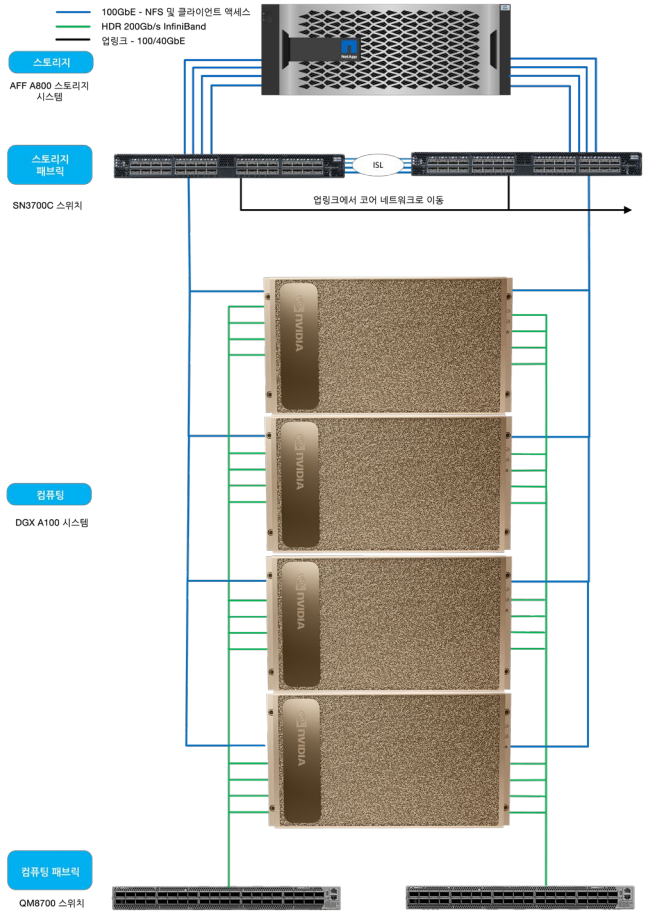


그림 2) Mellanox Spectrum 100GbE 스위치를 포함한 ONTAP AI 4 노드 구성

NetApp 랙 확장 아키텍처에서 All-Flash 스토리지를 통해 수십 테라바이트에서 수십 페타바이트로 확장할 수 있습니다 . NetApp ONTAP FlexGroup 을 사용하면 최대 20PB 의 단일 네임스페이스로 4 천억 개 이상의 파일을 처리할 수 있습니다 .

에지에서 코어 , 클라우드에 이르는 통합 데이터 파이프라인 구축

ONTAP AI 는 Data Fabric 을 사용하여 단일 플랫폼으로 데이터 파이프라인에서 데이터 관리를 통합합니다 . 동일한 도구를 사용하여 이동 중이거나 사용 중인 데이터 또는 유휴 데이터를 안전하게 제어 및 보호하고 자신 있게 규정 준수 요구사항을 충족할 수 있습니다 . DL 환경에서 문제가 발생하면 검증된 지원 모델을 통해 문제를 해결하고 지침을 제공할 수 있습니다 .

AI 워크로드 통합

이제 조직은 사용량이 적거나 AI 워크로드를 처리하지 못하는 인프라 사일로를 제거할 수 있습니다 . ONTAP AI 를 사용하면 DGX A100 시스템을 기반으로 하는 범용 AI 인프라 솔루션을 구축할 수 있습니다 . 이 솔루션은 분석 , 훈련 , 추론을 단일 플랫폼으로 통합하여 비즈니스 수요에 유연하게 대처할 수 있습니다 . 또한 , 기존 아키텍처보다 TCO 가 더 우수합니다 .

NetApp 과 NVIDIA: 함께 혁신 추진

ONTAP AI 의 핵심에는 훈련 , 추론 , 데이터 과학 , 기타 고성능 워크로드를 지원하는 데이터 센터 AI 용 범용 구성 요소인 DGX A100 시스템이 있습니다 . 각 DGX A100 시스템은 8 개 NVIDIA A100 Tensor Core GPU 와 듀얼 2 세대 AMD EPYC ™ 프로세서로 구동되며 최신 고속 NVIDIA Mellanox 100/200Gb 이더넷과 InfiniBand 지원 ConnectX-6 어댑터 인터커넥트가 통합되어 있습니다 .

새로운 NVIDIA 멀티 인스턴스 GPU(MIG) 기술로 DGX A100 을 시스템당 최대 56 개의 인스턴스로 파티셔닝하여 여러 작은 워크로드의 성능을 높일 수 있습니다 . 빠른 속도를 통해 조직은 ONTAP AI 에 GPU 성능을 효율적으로 할당할 수 있습니다 . 기업의 데이터 과학팀은 더 빠르게 추론하고 재현 기능을 자동화하여 , 우수한 품질의 AI 프로젝트를 최대 3 개월 더 빠르게 완수할 수 있습니다 .

NetApp AFF 시스템은 업계에서 세계 최초의 엔드 투 엔드 NVMe 기술을 제공하는 가장 빠르고 유연한 All-Flash 스토리지를 통해 ML 및 DL 프로세스의 데이터 흐름을 유지합니다 . AFF A800 시스템은 DGX 시스템에 경쟁 솔루션보다 최대 4 배 더 빠르게 데이터를 공급할 수 있습니다 .[1]

ONTAP AI 솔루션은 Mellanox Spectrum 이더넷 스위치와 통합되어 제공됩니다 . 이 스위치는 AI 환경에서 요구하는 짧은 지연 시간 , 높은 밀도 , 고성능 , 전력 효율성을 제공합니다 .

1. All-Flash 클러스터당 최대 300GBps 의 읽기 처리량 (주요 경쟁업체의 75GBps 대비)

NetApp 에서 지원하는 Data Fabric 은 동급 최강의 데이터 관리 및 클라우드 통합을 제공하여 중요한 데이터를 관리 및 보호하면서 DL 의 속도를 높여줍니다 . ONTAP 은 직접 연결 스토리지와 비교하여 탁월한 22:1 의 전체 데이터 감소율과 최대 54% 의 TCO 를 제공합니다 .

DGX A100 은 AI 및 데이터 과학 워크로드에 최적화된 NVIDIA DGX 소프트웨어 스택을 기반으로 하므로 , 성능을 극대화하여 기업에서 AI 인프라에 대한 투자 수익률을 더 빠르게 회수할 수 있습니다 .

NetApp AI Control Plane 을 사용하면 Kubernetes 및 Kubeflow 를 NetApp Data Fabric 과 통합하여 AI 데이터 관리를 간소화할 수 있습니다 . 이 통합 솔루션은 에지 , 코어 , 클라우드에서 최적의 데이터 가용성과 이동성을 제공합니다 . AI Control Plane 은 NetApp DataOps Toolkit 을 통해 개선합니다 . 이는 데이터 과학자와 데이터 엔지니어가 다양한 데이터 관리 작업을 쉽게 수행할 수 있도록 하는 Python 라이브러리입니다 . 예를 들어 , 새로운 데이터 볼륨을 프로비저닝하고 , 데이터 볼륨을 즉시 복제하고 , 추적 기능과 기준선 설정을 위한 NetApp Snapshot™ 데이터 볼륨 복사본을 생성할 수 있습니다 .

성공을 위해서는 적합한 툴을 사용해야 합니다 . 이것이 바로 ONTAP AI 가 Domino Data Lab, Iguazio 등을 포함하여 선도적인 MLOps(머신 러닝 운영) 소프트웨어로 검증된 이유입니다 . 익숙한 툴을 사용하여 AI 환경의 가치를 극대화하고 더 빠르게 통찰력을 얻을 수 있습니다 .

솔루션 구성 요소

- NVIDIA DGX A100 시스템
- NetApp AFF A-Series 스토리지 시스템 (ONTAP 9 포함)
- NVIDIA Mellanox Spectrum SN3700C, NVIDIA Mellanox Quantum QM8700 및 / 또는 NVIDIA Mellanox Spectrum SN3700-V
- NVIDIA DGX 소프트웨어 스택
- NetApp AI Control Plane
- NetApp DataOps 툴킷

참조 아키텍처

NetApp 은 특정 산업의 사용 사례를 대상으로 다음과 같은 ONTAP AI 기반 참조 아키텍처를 출시했습니다 .

- ONTAP AI Reference Architecture for Healthcare: 진단 이미징
- ONTAP AI Reference Architecture for Autonomous Driving Workloads: 솔루션 설계
- ONTAP AI Reference Architecture for Financial Services Workloads: 솔루션 설계

NetApp 정보

평범함으로 가득한 세상에서 NetApp 은 특별함을 선사합니다 . NetApp 은 귀사가 데이터를 최대한 활용할 수 있도록 돕는다는 한 가지 목표에 주력하고 있습니다 . NetApp 은 귀사에서 사용 중인 엔터프라이즈급 데이터 서비스를 클라우드로 전환하고 , 클라우드의 유연성을 데이터 센터에 제공합니다 . 업계 최고 수준의 NetApp 솔루션은 다양한 고객 환경과 세계 최대의 퍼블릭 클라우드에서 작동합니다 .

클라우드 주도형 데이터 중심 소프트웨어 회사인 NetApp 만이 고유한 Data Fabric 을 구축하고 , 클라우드를 단순화하고 연결하며 , 언제 어디서나 원하는 사람에게 원하는 데이터와 서비스 , 애플리케이션을 안전하게 제공하도록 지원할 수 있습니다 . www.netapp.com/kr

NVIDIA 정보

NVIDIA 는 1999 년 GPU 를 발명하여 PC 게임 시장 성장에 핵심적인 역할을 했으며 , 현대적 컴퓨터 그래픽을 재정의하고 병렬 컴퓨팅의 변혁을 일으켰습니다 . 최근 GPU 딥 러닝은 컴퓨팅의 다음 시대라 할 수 있는 현대적 인공지능 (AI) 의 포문을 열었습니다 . GPU 는 컴퓨터와 로봇 , 자율주행차 등에 탑재되어 세계를 인지하고 이해할 수 있도록 돕는 뇌 역할을 하고 있습니다 . 오늘날 , NVIDIA 는 'AI 컴퓨팅 기업 ' 으로서의 인지도를 높여가고 있습니다 . 자세한 내용은 www.nvidia.com 을 참조하십시오 .

NetApp