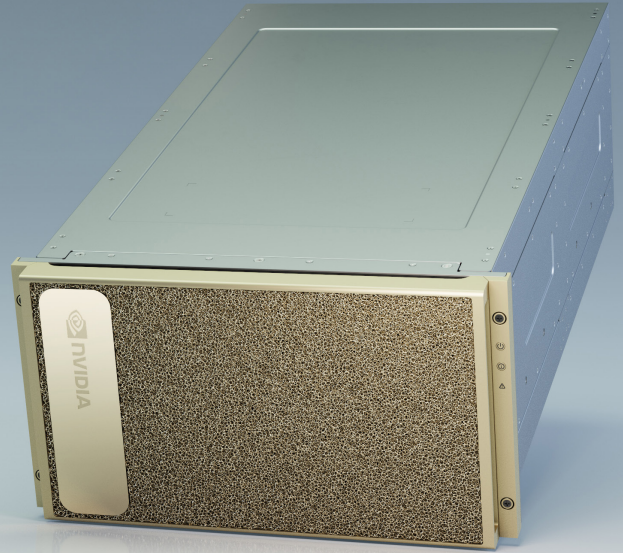




NVIDIA DGX A100 AI 인프라를 위한 범용 시스템



대규모 엔터프라이즈 AI의 도전 과제

모든 비즈니스는 생존을 위해서만이 아니라 도전적인 이 시대에서 번성하려면 인공지능(AI)을 사용해 변화해야 합니다. 기업은 AI 인프라를 위한 플랫폼을 활용하여 전통적인 방식을 개선해야 합니다. 이런 방식은 역사적으로 느린 컴퓨팅 구조를 사용했고 분석, 훈련, 추론 작업으로부터 고립되어 있습니다. 예전 방식은 복잡하고 비용이 높고 확장에 제약이 있어 현대적인 AI에 맞지 않습니다. 기업, 개발자, 데이터 과학자, 연구자들은 모든 AI 작업을 통합하고 인프라를 단순화하고 ROI를 높일 수 있는 새로운 플랫폼이 필요합니다.

모든 AI 작업을 위한 범용 시스템

NVIDIA DGX™ A100은 분석에서부터 훈련과 추론에 이르기까지 모든 AI 작업을 위한 범용 시스템입니다. DGX A100은 5 페타플롭스(petaFLOPS)의 AI 성능을 6U 폼 팩터로 구성하여 컴퓨팅 밀집도에 새로운 기준을 제시합니다. 이 시스템은 구형 컴퓨팅 인프라를 하나의 통합된 시스템으로 바꿀 수 있습니다. 또한 DGX A100은 NVIDIA A100 텐서 코어 GPU에 있는 다중 인스턴스(Multi-Instance) GPU 기능을 사용해 세밀하게 컴퓨팅 자원을 할당하는 종전에 없던 능력을 제공합니다. 따라서 관리자는 작업에 딱 맞추어 자원을 할당할 수 있습니다. 이를 통해 가장 작고 간단한 작업과 가장 크고 복잡한 작업을 모두 처리할 수 있습니다. NGC의 최적화된 소프트웨어와 함께 DGX 소프트웨어 스택을 구성하여 높은 컴퓨팅 성능과 완벽한 작업 유연성을 조합할 수 있습니다. DGX A100은 단일 노드 배포나 NVIDIA DeepOps로 배포한 슬럼(Slurm)과 쿠버네티스(Kubernetes)의 대규모 클러스터를 위한 최상의 선택입니다.

NVIDIA DGXperts와 직접 연결

NVIDIA DGX A100은 하나의 서버 그 이상입니다. 완벽한 하드웨어이자 소프트웨어 플랫폼입니다. 전 세계에서 가장 큰 DGX 시험장인 NVIDIA DGX SATURNV에서 얻은 지식을 기반으로 구축되었고 수천 명의 NVIDIA DGXperts가 지원합니다. DGXperts는 AI에 능숙한 기술자입니다. 이들은 모범적인 가이드와 설계 전문 지식을 제공하여 빠른 AI 전환을 돕습니다. 지난 십여 년간에 걸쳐 구축된 노하우와 경험을 바탕으로 고객의 DGX 투자 가치를 극대화합니다. DGXperts는 중요한 애플리케이션을 빠르게 실행하고 안정적으로 유지하도록 돕습니다. 이런 덕분에 비즈니스에 대한 통찰을 얻기 위한 시간을 극적으로 줄일 수 있습니다.

SYSTEM SPECIFICATIONS

GPUs	8개 NVIDIA A100 텐서 코어 GPUs
GPU 메모리	총 320 GB
성능	5 페타플롭스 AI 10 페타플롭스 INT8
NVIDIA NVSwitches	6
시스템 전력 소모	최대 6.5kW
CPU	듀얼 AMD Rome 7742, 총 128개 코어, 2.25 GHz (기본 클럭), 3.4 GHz (최대 부스트 클럭)
시스템 메모리	1TB
네트워킹	8개 싱글 포트 Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand 1개 듀얼 포트 Mellanox ConnectX-6 VPI 10/25/50/100/200Gb/s 이더넷
스토리지	OS: 2x 1.92TB M.2 NVME 드라이브 내부 스토리지: 15TB (3.84TB 4개) U.2 NVME 드라이브
소프트웨어	우분투 리눅스 OS
시스템 중량	271 lbs (123 kgs)
패키지 시스템 중량	315 lbs (143kgs)
System Dimensions	높이: 10.4 in (264.0 mm) 너비: 최대 19.0 in (482.3 mm) 길이: 최대 35.3 in (897.1 mm)
시스템 제원	5°C to 30°C (41°F to 86°F)

솔루션 출시 시간 단축

NVIDIA DGX A100은 비교할 수 없는 속도를 제공하는 8개의 NVIDIA A100 텐서 코어 GPU를 가지고 있습니다. NVIDIA CUDA-X™ 소프트웨어와 엔드-투-엔드 NVIDIA 데이터 센터 솔루션 스택에 완벽하게 최적화되어 있습니다. NVIDIA A100 GPU는 새로운 TF32 정밀도를 지원합니다. FP32처럼 작동하지만 이전 세대보다 AI 작업에서 20배나 높은 플롭스 FLOPS를 제공합니다. 무엇보다도 어떤 코드 변경도 없이 이런 속도를 얻을 수 있습니다. NVIDIA의 자동 혼합 정밀도를 사용하면 A100은 FP16 정밀도를 사용한 코드에서 단 한 줄만 추가하여 성능을 2배 더 올릴 수 있습니다. 또한 A100 GPU는 지난 세대보다 70% 이상 증가한 초당 1.6 테라바이트(TB/s)의 메모리 대역폭을 제공합니다. 이전에 없던 이런 성능 덕분에 최대한 빠르게 솔루션을 출시할 수 있고 현실적으로 불가능했던 도전 과제를 해결할 수 있습니다.

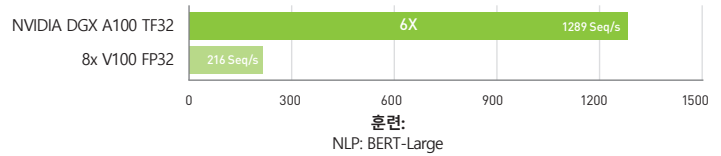
전 세계에서 가장 안전한 엔터프라이즈 AI 시스템

NVIDIA DGX A100은 기업의 AI를 위해 가장 안전한 시스템입니다. 모든 주요 하드웨어와 소프트웨어 구성 요소를 보호하기 위해 다층 구조를 적용했습니다. DGX A100은 베이스보드(baseboard) 관리 컨트롤러(BMC), CPU 보드, GPU 보드, 자기 암호화 드라이브(self-encrypted drive), 보안 부팅으로 확장된 내장 보안 시스템을 가지고 있습니다. 이를 통해 IT가 위험을 평가하고 완화하는데 시간을 뺏기지 않고 AI 운영에 집중할 수 있도록 돕습니다.

비교할 수 없는 Mellanox의 데이터 센터 확장성

DGX 시스템의 빠른 I/O 구조와 더불어 NVIDIA DGX A100은 확장성있는 AI 인프라의 엔터프라이즈급 청사진인 NVIDIA DGX SuperPOD™과 같은 대규모 AI 클러스터를 위한 기본 구성 요소입니다. DGX A100은 클러스터링을 위한 8개의 싱글 포트 Mellanox ConnectX-6 VPI HDR InfiniBand 어댑터, 그리고 스토리지와 네트워킹을 위한 1개의 듀얼 포트 ConnectX-6 VPI Ethernet를 가지고 있습니다. 모두 200Gb/s 전송 능력을 가집니다. GPU 가속을 갖춘 대규모 컴퓨팅을 최고 성능의 네트워킹 하드웨어와 최적화된 소프트웨어와 조합한다는 것은 DGX A100을 수백 또는 수천 개의 노드로 확장하여 대화형 AI나 대규모 이미지 분류 같은 가장 어려운 문제에 도전할 수 있다는 뜻입니다.

DGX A100은 6배 높은 훈련 성능을 제공합니다.



파라미터를 사용한 BERT 사전 훈련 (2/3 단계 1과 1/3 단계 2 포함 | 단계 1 시퀀스 길이 = 128, 단계 2 시퀀스 길이 = 512 | V100: FP32 정밀도를 사용한 V100 8개가 장착된 DGX-1 | DGX A100: TF32 정밀도를 사용한 A100 8개가 장착된 DGX A100

DGX A100은 172배 높은 추론 성능을 제공합니다.



CPU 서버: INT8을 사용한 Intel Platinum 8280 2개 | DGX A100: 구조적 희소성 Structural Sparsity와 INT8을 사용한 A100 8개가 장착된 DGX A100

DGX A100은 13배 높은 데이터 분석 성능을 제공합니다.



3000개의 CPU 서버 vs. 4개의 DGX A100 | 공개된 Common Crawl 데이터셋: 128B 에지, 2.6TB 그래프

믿을 수 있는 데이터 센터 선두 주자와 함께 구축한 검증된 인프라 솔루션

스토리지와 네트워킹 기술을 제공하는 선도 업체와 연합하여 NVIDIA DGX POD™ 구조의 가장 좋은 장점을 포함한 인프라 솔루션의 포트폴리오를 제공합니다. NVIDIA 파트너 네트워크를 통해 완전히 통합되고 즉시 배포 가능하게 제공되는 이 솔루션은 IT를 위해 데이터 센터 AI 배포를 간단하고 빠르게 만들어 줍니다.

NVIDIA DGX A100에 더 자세한 내용을 알고 싶다면 www.nvidia.com/DGXA100을 방문하세요