데이터시트

DGX GB200 시스템을 갖춘 NVIDIA DGX SuperPOD

수조 개의 파라미터를 생성하는 AI의 시대

현재 규모를 막론하고 모든 엔터프라이즈는 챗봇 및 코파일럿 개발, 콘텐츠 개인화, 약물 발견 가속화, 시각적 애플리케이션 생성 등에 생성형 AI를 활용하고 있습니다. 오늘날의 최신 기반 모델은 수조 개의 파라미터를 생성하며 페타바이트의 데이터를 학습합니다. 이러한 차세대 고성능 AI 모델은 새로운 아이디어를 효율적으로 반복하고 결과를 빠르게 생성하며 실시간에 가까운 속도로 추론할 수 있는 수천 개의 GPU로 구성된 훈련 및 추론 인프라가 필요합니다.

엔터프라이즈급 생성형 AI 인프라

DGX™ GB200 시스템을 갖춘 NVIDIA DGX SuperPOD™는 엔터프라이즈가 활용도와 생산성의 획기적 증대 및 AI 이니셔티브의 ROI 개선을 달성할 수 있도록 전례 없는 성능과 예측 가능한 가동 시간을 지원합니다. 이는 AI 성능, 안정성 및 확장성의 새로운 표준을 제시합니다.

효율적인 수냉식 랙 스케일 디자인은 최대 수만 개의 GPU로 확장할 수 있으며 NVIDIA GB200 Grace Blackwell 슈퍼칩을 활용해 최신 고급 생성형 AI 애플리케이션에 필요한 1조개 파라미터 AI 모델을 처리합니다.

DGX SuperPOD는 거대 생성형 AI 훈련 및 추론 워크로드를 위한 우수한 성능과 일관된 가동 시간을 제공하도록 특별히 설계된 차세대 솔루션입니다. 엔터프라이즈는 NVIDIA의 내부 클러스터 설계를 기반으로 하며 엔터프라이즈 AI 인프라에서는 최초로 제공되는 풀스택 탄력성을 통해 운영 복잡성이 아닌 혁신에 집중할 수 있습니다.

일정한 가동 시간을 통한 개발자 생산성 극대화

DGX GB200 시스템을 갖춘 DGX SuperPOD는 AI 인프라를 위한 풀스택 탄력성으로 일관된 가동 시간을 제공합니다. 지능형 제어 평면은 하드웨어, 소프트웨어 및 데이터 센터 인프라 전반에서 수천 개의 데이터 포인트를 계속해서 추적하여 지속적인 운영 및 데이터 무결성을 보장합니다. 시스템 관리자 부재 시에도 가동 중단 시간이 방지되도록 예비 하드웨어와 강력한 체크포인트 및 재시작 메커니즘을 사용하는 자동 재구성을 지원합니다.

주요 특징

- NVIDIA GB200 Grace™ Blackwell 슈퍼칩 기반
- 최대 수만 개의 GB200 슈퍼칩으로 확장 가능
- 72개의 NVIDIA Blackwell GPU가 NVIDIA® NVLink®로 하나로 연결
- 효율적인 수냉식 랙 스케일 디자인
- NVIDIA 네트워크
- 가동 시간 극대화를 위한 통합된 예측형 유지 관리
- NVIDIA AI Enterprise 및 NVIDIA Base Command™ 소프트웨어 포함
- 각 DGX SuperPOD는 빠른 현장 배포를 위해 공장에서 완전 조립 및 테스트 완료

생성형 AI를 위한 슈퍼컴퓨팅

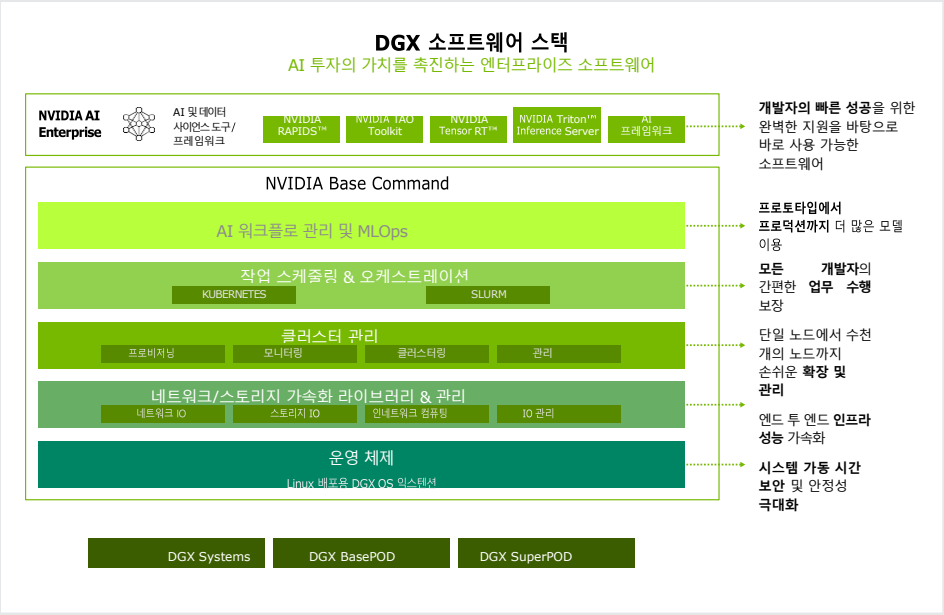
DGX GB200 시스템을 갖춘 DGX SuperPOD는 최대 수만 개의 NVIDIA Grace Blackwell 슈퍼칩으로 확장하여 최첨단 1조 매개변수 모델에 대해 학습 및 지연 없이 빠른 추론을 수행할 수 있습니다. 거대 언어 모델(LLM)에 적합한 DGX SuperPOD 내 각 DGX GB200 시스템은 36개의 NVIDIA Grace CPU와 72개의 NVIDIA Blackwell GPU가 5세대 **NVIDIA NVLink**와 하나로 연결되며, 1.4엑사플롭스의 AI 성능과 30테라바이트(TB)의 고속 메모리, 초당 130 테라바이트(TB/s)의 양방향 GPU 대역폭을 제공합니다. DGX SuperPOD는 NVLink로 연결된 576개의 NVIDIA Blackwell GPU가 탑재된 DGX GB200 시스템을 이용할 수 있는 구성 옵션을 제공합니다. 이를 통해 대규모 공유 메모리 풀을 생성해 추천자, 그래프 신경망(GNN) 및 1조 파라미터 규모의 **전문가 혼합(MoE) LLM**과 같은 메모리 바운드 워크로드의 속도를 크게 높일 수 있습니다. DGX GB200을 갖춘 DGX SuperPOD를 사용하는 엔터프라이즈는 현재 및 미래의 거대 생성형 AI 모델에 대해 학습 및 추론을 손쉽게 수행할 수 있습니다.

NVIDIA Grace Blackwell 기반

DGX SuperPOD를 지원하는 NVIDIA Grace Blackwell 슈퍼칩은 혁신적인 4나노미터 제조 공정, 5세대 NVLink 및 2세대 Transformer Engine이 특징이며, 전 세계에서 가장 효율적인 생성형 AI용 AI 슈퍼컴퓨터를 구축하는 수냉식 랙 스케일 디자인에 통합됩니다. 슈퍼칩에는 각각 두 개의 고성능 NVIDIA Blackwell GPU와 하나의 NVIDIA Grace CPU가 탑재되어 있습니다. GB200 슈퍼칩의 모든 Blackwell GPU는 GPU-to-GPU 연결을 위한 NVLink를 사용해 1.8TB/s의 양방향 처리량을 달성합니다.

통합 AI 소프트웨어

NVIDIA Base Command™는 DGX 플랫폼을 지원하여 조직이 NVIDIA가 제공하는 최상의 소프트웨어 혁신을 활용할 수 있도록 합니다. 엔터프라이즈는 엔터프라이즈급 오케스트레이션 및 클러스터 관리, 컴퓨팅, 스토리지 및 네트워크 인프라를 가속화하는 라이브러리, AI 워크로드에 최적화된 운영 체제를 포함하는 입증된 플랫폼을 이용해 DGX 인프라에 잠재된 가능성을 최대한 구현할 수 있습니다. 또한 DGX 인프라는 AI 개발 및 배포 간소화를 위해 최적화된 소프트웨어 제품군인 **NVIDIA AI Enterprise**를 포함합니다.



기술 사양	
	72-GPU NVLink 도메인(NVL72)
FP4 AI	1,440페타플롭스
FP8 AI	725페타플롭스
FP16 AI	362페타플롭스
GPU	NVIDIA Blackwell GPU 72개가 탑재된 Grace Blackwell Superchip
GPU 메모리 HBM3e	13.3TB
총 고속 메모리	30.2TB
상호 연결	400Gb/s InfiniBand의 72x OSFP 싱글 포트 NVIDIA ConnectX®-7 VPI 200Gb/s InfiniBand/이더넷의 36x 듀얼 포트 NVIDIA BlueField®-3 VPI
NVIDIA NVLink Switch 시스템	L1 NVIDIA NVLink Switch 9개
관리 네트워크	RJ45가 포함된 호스트 베이스보드 관리 컨트롤러(BMC)
소프트웨어	NVIDIA AI Enterprise: 최적화된 AI 소프트웨어 NVIDIA Base Command: 오케스트레이션, 스케줄링 및 클러스터 관리 DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky: 운영 체제
엔터프라이즈 지원	하드웨어 및 소프트웨어에 대한 3년 엔터프라이즈 비즈니스 표준 지원

시작할 준비가 되셨나요?

DGX GB200 시스템을 갖춘 NVIDIA DGX SuperPOD에 대한 추가 정보는 **nvidia.com/dgx-gb200**에서 확인할 수 있습니다.