



NVIDIA DGX B200

Training, Fine-tuning 및
Inference를 위한 통합 AI 플랫폼



차세대 AI의 성장 동력

오늘날 인공지능은 다양한 작업을 자동화하고, 고객 서비스를 개선하고, 인사이트를 창출하고, 혁신을 뒷받침하는 등 거의 모든 비즈니스의 변화를 주도하고 있습니다.

AI는 더 이상 미래 지향적인 개념이 아니고 비즈니스 운영 방식을 근본적으로 재구성하고 있는 현실입니다. 하지만 AI 워크로드가 계속해서 개발됨에 따라, 훨씬 높은 컴퓨팅 성능이 필요해졌습니다. 또한 기업이 AI를 활용하려면 안전하고, 안정적이고, 효율적인 고성능 컴퓨팅, 스토리지 및 네트워킹 기술이 필요합니다.

가장 최근에 **NVIDIA DGX 플랫폼**에 추가된 **NVIDIA DGX™ B200**을 소개합니다. 이 통합 AI 플랫폼은 NVIDIA Blackwell GPU와 고속 인터커넥트를 활용해 생성형 AI의 새로운 길을 제시합니다. Blackwell GPU 8개로 구성되는 DGX B200은 1.4테라바이트(TB)의 대용량 GPU 메모리, 초당 64테라바이트(TB/s)의 HBM3e 메모리 대역폭, 14.4TB/s의 all-to-all GPU 대역폭으로 탁월한 생성형 AI 성능을 자랑하여 어떤 엔터프라이즈 AI 워크로드든 간단히 처리할 수 있습니다.

기업은 NVIDIA DGX B200으로 데이터 사이언티스트와 개발자에게 범용 AI 슈퍼컴퓨터를 제공함으로써 인사이트 추출 시간을 앞당길 뿐만 아니라, AI 기능을 적극적으로 활용하여 비즈니스를 효과적으로 지원할 수 있습니다.

개발-배포 파이프라인을 위한 단일 플랫폼

AI 워크플로우가 더욱 정교해지면, 기업 역시 Training부터 Fine-tuning 및 Inference에 이르기까지 모든 AI 파이프라인 단계에서 대용량 데이터 세트를 처리할 수 있어야 합니다. 이를 위해서는 막대한 용량의 컴퓨팅 파워가 필요합니다.

NVIDIA DGX B200은 기업 개발자에게 워크플로우 가속을 고려해 설계된 단일 통합 플랫폼을 제공합니다. 기업은 DGX B200으로 더욱 강력해진 차세대 생성형 AI 성능을 이용해 AI를 일상적인 업무와 고객 서비스에 활용할 수 있습니다.

주요 특징

NVIDIA DGX B200

- > NVIDIA Blackwell GPU 8개 탑재
- > 1.4TB의 GPU 메모리 공간
- > 72 petaFLOPS의 훈련 성능
- > 144 petaFLOPS의 추론 성능
- > NVIDIA 네트워킹
- > 듀얼 5세대 Intel® Xeon® 확장 가능 프로세서
- > NVIDIA DGX BasePOD 및 NVIDIA DGX SuperPOD 기반
- > NVIDIA AI Enterprise 및 NVIDIA Base Command™ 소프트웨어 포함

AI 성능의 원동력

NVIDIA는 기업이 복잡한 AI 문제도 해결할 수 있도록 세계에서 가장 강력한 차세대 슈퍼컴퓨터를 설계하고 있습니다. DGX B200은 최근에 NVIDIA 가속 컴퓨팅 플랫폼에 추가된 제품입니다. 혁신적인 NVIDIA Blackwell 아키텍처의 컴퓨팅 기술이 적용되어 DGX H100과 비교했을 때 3배 빠른 훈련 성능과 15배 빠른 추론 성능을 제공합니다. 또한 NVIDIA DGX POD™ 참조 아키텍처를 기반으로 **NVIDIA DGX BasePOD™**와 **NVIDIA DGX SuperPOD™**에 고속 확장성을 지원하여 터키 AI 인프라 솔루션 중에서 최고의 성능을 자랑합니다.

검증된 인프라 표준

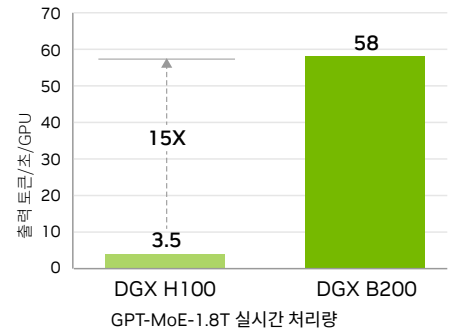
NVIDIA DGX B200은 세계 최초로 NVIDIA Blackwell GPU가 탑재된 시스템으로 대규모 언어 모델, 자연어 처리 등 아무리 복잡한 AI 문제라고 해도 해결할 수 있는 획기적인 성능을 제공합니다. 또한 완전한 NVIDIA AI 소프트웨어 스택, 풍부한 서드파티 지원 에코시스템, NVIDIA 전문 서비스를 통한 전문가 자문 등 최적화된 하드웨어/소프트웨어 플랫폼을 제공하기 때문에 크고, 복잡한 비즈니스 문제도 해결할 수 있습니다.

NVIDIA Base Command 탑재

NVIDIA Base Command는 DGX 플랫폼에 탑재되어 기업들이 혁신적인 NVIDIA 소프트웨어를 최대한 이용할 수 있도록 지원합니다. 덕분에 기업들은 엔터프라이즈급 오케스트레이션 및 클러스터 관리, 컴퓨팅/스토리지/네트워크 인프라를 가속하는 라이브러리, AI 워크로드에 최적화된 운영 체제가 포함되어 이미 성능이 검증된 플랫폼을 이용해 DGX 인프라의 잠재력을 극대화할 수 있습니다. 그 밖에도 AI 개발 및 배포를 간소화할 수 있는 소프트웨어 제품군인 **NVIDIA AI Enterprise**가 DGX 인프라에 포함됩니다.

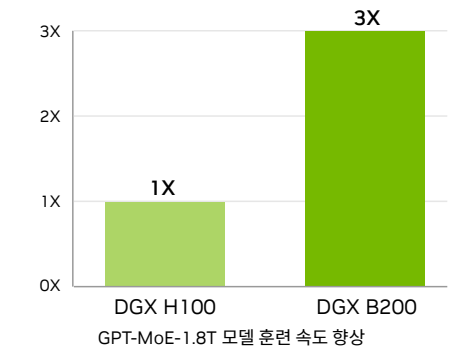
실시간 대규모 언어

모델 추론

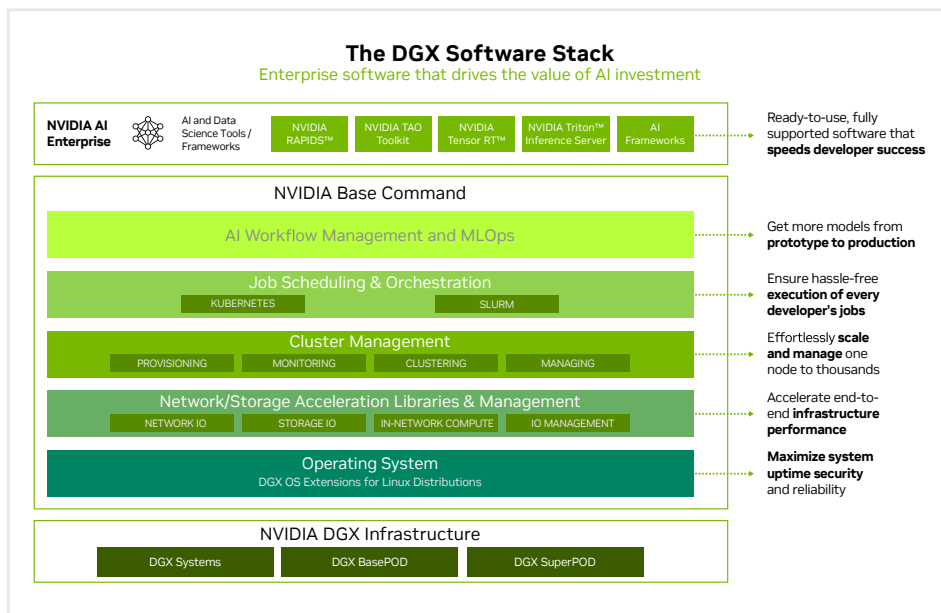


예상 성능은 변경될 수 있습니다. 토큰 간 지연 시간(TTL) = 실시간 50ms, 첫 토큰 지연 시간(FTL) = 5,000ms, 입력 시퀀스 길이 = 32,768, 출력 시퀀스 길이 = 1,028, 8x 8방향 DGX H100 GPU 공냉식 vs 1x 8방향 DGX B200 공냉식, GPU당 성능 비교입니다.

더욱 강력해진 AI 훈련 성능



예상 성능은 변경될 수 있습니다. 32,768 GPU 규모, 4,096x 8방향 DGX H100 공냉식 클러스터; 400G IB 네트워크, 4,096x 8방향 DGX B200 공냉식 클러스터; 400G IB 네트워크



DGX B200 Technical Specifications	
GPU	8x NVIDIA Blackwell GPUs
GPU Memory	1,440GB total, 64TB/s HBM3e bandwidth
Performance	72 petaFLOPS FP8 training and 144 petaFLOPS FP4 inference
NVIDIA® NVSwitch™	2x
NVIDIA NVLink Bandwidth	14.4 TB/s aggregate bandwidth
System Power Usage	~14.3kW max
CPU	2 Intel® Xeon® Platinum 8570 Processors 112 Cores total, 2.1 GHz (Base), 4 GHz (Max Boost)
System Memory	2TB, configurable to 4TB
Networking	4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA BlueField-3 DPU > Up to 400Gb/s InfiniBand/Ethernet
Management Network	10Gb/s onboard NIC with RJ45 100Gb/s dual-port ethernet NIC Host baseboard management controller (BMC) with RJ45
Storage	OS: 2x 1.9TB NVMe M.2 Internal storage: 8x 3.84TB NVMe U.2
Software	NVIDIA AI Enterprise – Optimized AI Software NVIDIA Base Command – Orchestration, Scheduling, and Cluster Management DGX OS / Ubuntu – Operating System
Rack Units (RU)	10 RU
System Dimensions	Height: 17.5in (444mm) Width: 19.0in (482.2mm) Length: 35.3in (897.1mm)
Operating Temperature	5–30°C (41–86°F)
Enterprise Support	Three-year Enterprise Business-Standard Support for hardware and software 24/7 Enterprise Support portal access Live agent support during local business hours

시작할 준비가 되었습니까?

NVIDIA DGX B200에 대한 자세한 내용은 nvidia.com/dgx-b200에서 확인할 수 있습니다.

© 2024 NVIDIA Corporation and affiliates. All rights reserved.
NVIDIA, NVIDIA 로고, Base Command, BlueField, ConnectX, DGX, DGX BasePOD, DGX POD,
DGX SuperPOD, NVSwitch는 미국을 비롯한 기타 국가에 위치한 NVIDIA 기업과 계열사의 상표 및/또는 등록 상표입니다.
기타 회사 및 제품명은 관련된 각 소유권자의 상표일 수 있습니다. 3237805. MAY24

