데이터시트

NVIDIA DGX B200

훈련, 미세 조정 및 추론을 위한
통합 AI 플랫폼

차세대 AI의 역량 강화

인공지능은 업무 자동화, 고객 서비스 강화, 인사이트 생성, 혁신 구현과 같은 다양한 기능을 바탕으로 전반적인 비즈니스 환경을 혁신하고 있습니다. AI는 이제 시대를 앞선 개념이 아닌 비즈니스 운영 방식의 근간을 재구성하고 있는 실제 현실입니다. 하지만 AI 워크로드가 지속적으로 발전하면서 대부분의 엔터프라이즈에서 수용 가능한 수준보다 현저히 높은 컴퓨팅 용량이 요구되기 시작했습니다. 엔터프라이즈가 AI를 활용하기 위해서는 보안, 안정성 및 효율성을 갖춘 고성능 컴퓨팅, 스토리지, 네트워킹 기능이 필요합니다.

NVIDIA DGX 플랫폼에 새롭게 추가된 **NVIDIA DGX™ B200**을 소개합니다. NVIDIA DGX™ B200은 NVIDIA Blackwell GPU 및 고속 상호 연결 기술을 통해 생성형 AI의 새로운 장을 여는 통합 AI 플랫폼입니다. 8개의 Blackwell GPU가 탑재되어 1.4테라바이트(TB)의 GPU 메모리와 초당 64테라바이트(TB/s)의 메모리 대역폭으로 독보적인 생성형 AI 성능을 제공하는 DGX B200은 모든 엔터프라이즈 AI 워크로드를 처리할 수 있도록 특별히 설계되었습니다.

NVIDIA DGX B200을 사용하는 엔터프라이즈의 경우, 데이터 사이언티스트와 개발자는 범용 AI 수퍼 컴퓨터를 통해 인사이트 도출을 가속화하고 AI가 제공하는 비즈니스 이점을 완벽히 누릴 수 있습니다.

개발에서 배포까지의 파이프라인을 위한 단일 플랫폼

AI 워크플로가 계속해서 정교해지면서 엔터프라이즈는 훈련에서 미세 조정, 추론에 이르는 AI 파이프라인 전 단계에서 대규모 데이터세트를 처리해야 합니다. 이는 막대한 양의 컴퓨팅 성능을 요구합니다. NVIDIA DGX B200를 이용하는 엔터프라이즈의 개발자는 워크플로 가속화를 지원하는 단일 통합 플랫폼을 활용할 수 있습니다. DGX B200는 차세대 생성형 AI의 강력한 기능을 통해 엔터프라이즈가 일상 운영과 고객 경험에 AI를 결합할 수 있도록 합니다.

주요 특징

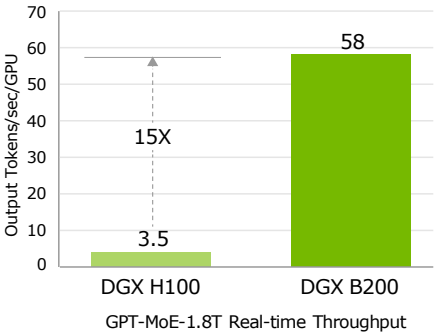
NVIDIA DGX B200

- 8개의 NVIDIA Blackwell GPU 탑재
- 1.4TB의 GPU 메모리 공간
- 72페타플롭스의 훈련 성능
- 144페타플롭스의 추론 성능
- NVIDIA 네트워킹
- 듀얼 5세대 Intel® Xeon® Scalable 프로세서
- NVIDIA DGX BasePOD 및 NVIDIA DGX SuperPOD의 기반
- NVIDIA AI Enterprise 및 NVIDIA Base Command™ 소프트웨어 포함

강력한 AI 성능

NVIDIA는 엔터프라이즈가 마주한 어떠한 복잡한 AI 문제라도 해결할 수 있는 전 세계에서 가장 강력한 차세대 슈퍼컴퓨터를 설계하는 일에 전념하고 있습니다. NVIDIA 가속 컴퓨팅 플랫폼에 새롭게 추가된 DGX B200은 NVIDIA의 이러한 노력을 보여주는 제품입니다. 컴퓨팅 분야에서 혁신적인 NVIDIA Blackwell 아키텍처의 발전을 기반으로 하는 DGX B200은 DGX H100 대비 3배의 훈련 성능과 15배의 추론 성능을 제공합니다. NVIDIA DGX POD™ 참조 아키텍처의 기반인 DGX B200은 **NVIDIA DGX BasePOD™** 및 **NVIDIA DGX SuperPOD™**에 대한 고속 확장성을 제공하여 턴키 AI 인프라 솔루션에서 최고의 성능을 제공합니다.

실시간 거대 언어 모델 추론



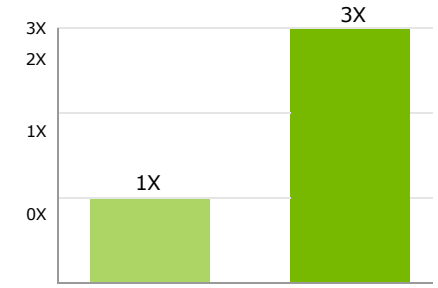
GPT-MoE-1.8T Real-time Throughput

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

입증된 인프라 표준

세계 최초로 NVIDIA Blackwell GPU를 탑재한 시스템인 NVIDIA DGX B200은 거대 언어 모델, 자연어 처리 등 전 세계적으로 어려움을 겪고 있는 복잡한 AI 문제를 위한 혁신적 성능을 제공합니다. DGX B200은 완전히 최적화된 하드웨어 및 소프트웨어 플랫폼으로, 완전한 NVIDIA AI 소프트웨어 스택, 풍부한 타사 지원 에코시스템, NVIDIA 전문 서비스의 전문가 조언이 포함되어 있어 조직이 AI를 이용해 거대하고 복잡한 비즈니스 문제를 해결하는 데 도움을 줍니다.

초고속 AI 학습 성능

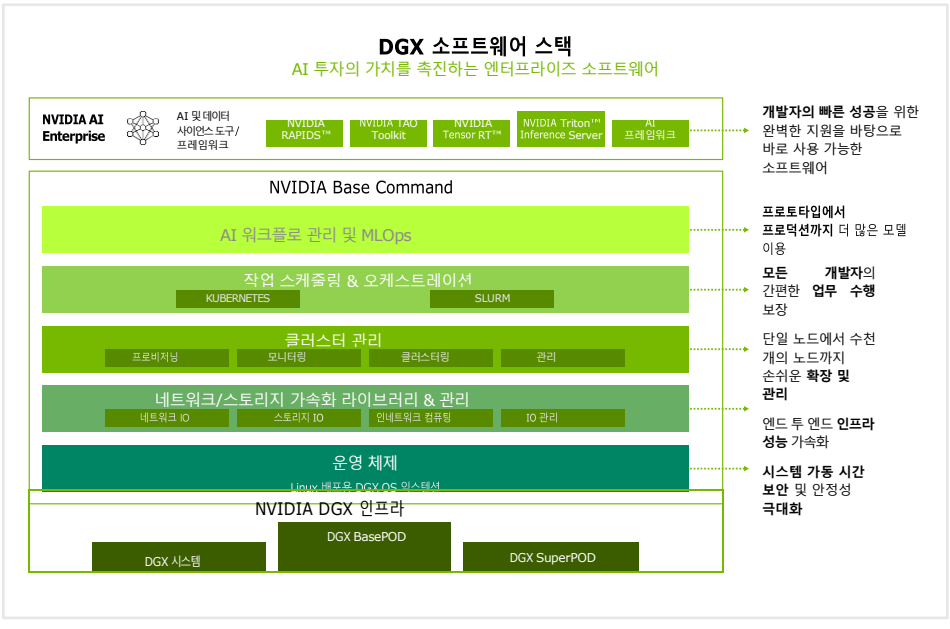


GPT-MoE-1.8T Model Training Speed-Up

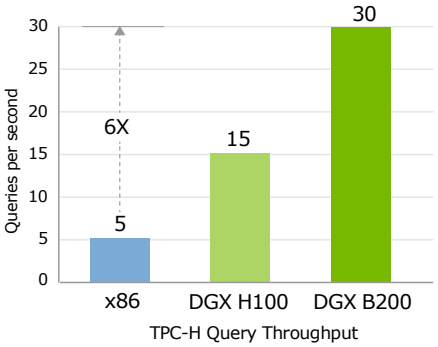
예상 성능은 변경될 수 있습니다. 32,768개의 GPU 규모, 4,096배 8방향 DGX H100 공랭식 클러스터: 400G IB 네트워크, 4,096x 8방향 DGX B200 공랭식 클러스터: 400G IB 네트워크.

NVIDIA Base Command의 지원

NVIDIA Base Command는 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.



데이터 처리 가속화



TPC-H Query Throughput

예상 성능은 변경될 수 있습니다. 데이터베이스 조인 쿼리(Snappy 지원) / TPC-H Q4 쿼리에서 파생된 Deflate 압축. 1x x86, 1x H100 GPU 및 1x Blackwell 단일 GPU.

DGX B200 기술 사양	
GPU	8x NVIDIA Blackwell GPU
GPU 메모리	총 1,440GB
성능	72페타플롭스 훈련 및 144페타플롭스 추론
NVIDIA® NVSwitch™	2x
시스템 소비 전력	최대 약 14.3kW
CPU	Intel® Xeon® Platinum 8570 프로세서 2개 총 112개 코어, 2.1GHz(기본), 4 GHz (최대 부스트)
시스템 메모리	최대 4TB
네트워킹	싱글 포트 NVIDIA ConnectX-7 VPI 8개를 제공하는 OSFP 포트 4개 > 최대 400Gb/s InfiniBand/이더넷 2x 듀얼 포트 QSFP112 NVIDIA BlueField-3 DPU > 최대 400Gb/s InfiniBand/이더넷
관리 네트워크	10Gb/s 온보드 NIC(RJ45 포함) 100Gb/s 듀얼 포트 이더넷 NIC RJ45가 포함된 호스트 베이스보드 관리 컨트롤러(BMC)
스토리지	OS: 2x 1.9TB NVMe M.2 내부 스토리지: 8x 3.84TB NVMe U.2
소프트웨어	NVIDIA AI Enterprise – 최적화된 AI 소프트웨어 NVIDIA Base Command – 오케스트레이션, 스케줄링 및 클러스터 관리 DGX OS / Ubuntu – 운영 체제
랙 장치(RU)	10 RU
시스템 크기	높이: 444mm(17.5인치) 너비: 482.2mm(19.0인치) 길이: 897.1mm(35.3인치)
작동 온도	5~30°C(41~86°F)
엔터프라이즈 지원	하드웨어 및 소프트웨어에 대한 3년 엔터프라이즈 비즈니스 표준 지원 연중무휴 엔터프라이즈 지원 포털 액세스 현지 업무 시간 동안의 실시간 에이전트 지원

시작할 준비가 되셨나요?

NVIDIA DGX B200에 대한 추가 정보는 nvidia.com/dgx-b200에서 확인할 수 있습니다.