데이터시트

NVIDIA DGX GH200

AI를 위한 대용량 메모리 슈퍼컴퓨팅

AI 사용 방식이 갈수록 정교해지면서 AI 모델의 크기와 복잡성도 날로 커지고 있습니다. 대부분의 조직이 다수의 AI 워크로드를 동시에 처리해야 하는 반면, 사용자들은 단일 워크로드에도 GPU, 또는 대형 멀티 GPU 시스템의 한계를 넘어서는 거대한 메모리 요건을 충족해야 합니다. 이러한 사용자에게는 단순성을 위해 단일 GPU 프로그래밍 모델을 유지하면서 규모 확장에 따라 성능 저하 없이 GPU와 CPU의 메모리와 처리 능력을 확대할 수 있는 새로운 방법이 필요합니다.

클라우드 서비스 제공 업체(CSP), 하이퍼스케일러, 대형 연구 조직, 그 외 AI의 경계를 무너뜨리고 있는 기업 등을 위해 NVIDIA DGX™ GH200은 대형 AI 모델의 개발을 위한 새로운 청사진을 제시합니다.
이 새로운 AI 슈퍼컴퓨터는 GPU 전반에 걸쳐 선형적 확장성과 거대한 공유 메모리 공간을 제공하는 멀티 노드 NVIDIA® NVLink® 기술이 통합된 NVIDIA Grace Hopper™ Superchip을 사용하여 세계 최대 규모의 신경망과 추천 기능, 시뮬레이션 모델 및 생성형 AI 애플리케이션을 개발하는 데 필요한 기능을 제공합니다.

NVIDIA DGX 플랫폼에 속해 있는 DGX GH200은 하드웨어 그 이상의 역할을 합니다. 이는 NVIDIA가 설계하여 제공하는 완전한 소프트웨어 및 하드웨어 솔루션으로, 총체적인 턴키 방식의 경험을 제공합니다. 화이트 글러브 서비스가 복잡성을 해소하고 배포 속도를 단축하며 운영을 간소화하고, 대형 메모리 슈퍼컴퓨터에서 뛰어난 전력 효율을 발휘합니다. DGX GH200은 AI의 새로운 지평을 엽니다. 한층 크고 복잡해진 최첨단 모델로 기술의 수준을 높입니다.

주요 특징

NVIDIA DGX GH200

- NVIDIA Grace Hopper Superchip 32개, NVIDIA NVLink와 상호 연결됨
- 19.5TB 용량의 대형 공유 GPU 메모리
- 900GB/S GPU 간 대역폭
- FP8의 128 petaFLOPS AI 성능
- NVIDIA Base Command™ 및 NVIDIA AI Enterprise 소프트웨어
- 화이트 글러브 실행 경험

대형 모델을 위한 대형 메모리

단일 시스템의 메모리 용량에 맞는 워크로드를 지원하도록 설계된 기존의 AI 슈퍼컴퓨터와 달리 NVIDIA DGX GH200은 Grace Hopper Superchip 32개에 걸쳐 19.5TB의 공유 메모리 공간을 제공하는 유일한 AI 슈퍼컴퓨터로, 대형 모델을 구축할 수 있는 30배 이상 빠른 액세스 속도의 메모리를 제공합니다. DGX GH200은 Grace Hopper Superchip과 NVIDIA NVLink 스위치 시스템을 탑재한 최초의 슈퍼컴퓨터로, GPU 32개를 단일 데이터 센터 크기의 GPU로 통합할 수 있습니다. NVIDIA InfiniBand를 이용해 DGX GH200 시스템을 여러 대 연결하면 연산 능력을 더욱 높일 수 있습니다. 이 아키텍처는 이전 세대 보다 10배 더 큰 대역폭을 제공하여
단일 GPU 프로그래밍의 단순성을 바탕으로 대형 AI 슈퍼컴퓨터의 성능을 제공합니다.

전력 효율이 우수한 컴퓨팅

AI 모델의 복잡성이 증가하면서 이를 개발 및 배포하는 기술의 리소스 사용량이 대폭 증가했습니다. 하지만 DGX GH200는 NVIDIA Grace Hopper Superchip 아키텍처를 이용해 전력 효율을 높였습니다. NVIDIA Grace Hopper Superchip은 CPU와 GPU를 한곳에 탑재한 시스템으로, 매우 빠른 속도의 NVIDIA NVLink-C2C와 연결되어 있습니다. Grace™ CPU는 기존의 DDR5 시스템 메모리 전력의 1/8만 소비하면서 8채널 DDR5 대비 50% 높은 대역폭을 제공하는 LPDDR5X 메모리를 사용합니다. 또한 동일 모듈에서 Grace CPU와 Hopper™ GPU Interconnect가 타 시스템에 사용되는 최신 PCIe 기술 대비 전력을 5배 적게 소모하면서 7배 더 큰 대역폭을 제공합니다.

통합 및 즉시 실행

대형 메모리 애플리케이션 개발에 맞게 조정된 하이퍼스케일 데이터 센터를 설계, 통합 및 운용하는 일은 복잡하고 시간이 많이 소요됩니다. DGX GH200을 사용하면 이 작업을 빠르게 시작할 수 있습니다. 소프트웨어, 컴퓨팅 및 네트워킹이 완전히 테스트된 통합 솔루션입니다. 함께 제공되는 화이트 글러브 서비스에는 설치 및 인프라 관리에서부터 워크로드 최적화를 위한 전문가 자문까지 포함됩니다.

DGX GH200 기술 사양	
CPU 및 GPU	NVIDIA Grace Hopper Superchip 32개
CPU 코어	SVE2 4X 128b의 2,304 Arm® Neoverse V2 코어
공유 메모리	19.5 TB
성능	FP8의 128 petaFLOPS AI 성능
네트워킹	400Gb/s InfiniBand의 OSFP 단일 포트 NVIDIA ConnectX-7 VPI 32개 200Gb/s InfiniBand 및 Ethernet의 듀얼 포트 NVIDIA BlueField®-3 VPI 16개
NVIDIA NVLink 스위치 시스템	L1 NVIDIA NVLink 스위치 9개
관리 네트워크	베이스보드 관리 컨트롤러 호스팅 (BMC), RJ45
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)
지원	3년 약정의 비즈니스 표준 하드웨어 및 소프트웨어 지원

시작할 준비가 되셨나요?

DGX GH200에 대한 자세한 내용은 다음 사이트에서 확인하실 수 있습니다. nvidia.com/dgx-gh200