데이터시트

NVIDIA GB200 NVL72

새로운 컴퓨팅 시대의 강자

실시간 조 단위 파라미터 모델 지원

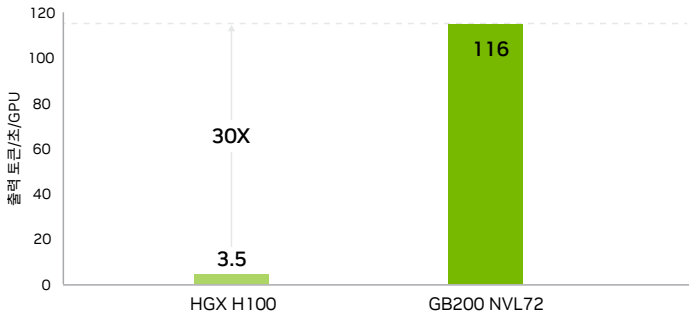
NVIDIA GB200 NVL72는 NVIDIA® NVLink®가 연결된 수냉식 랙 스케일 설계에서 Grace CPU 36개와 Blackwell GPU 72개를 연결합니다. 방대한 용량의 단일 GPU로 사용되기 때문에 조 단위 파라미터 대규모 언어 모델(LLM)에서도 실시간 추론 속도가 30배 더 빠릅니다.

GB200 Grace Blackwell 슈퍼칩은 고성능 NVIDIA Blackwell GPU 2개와 NVIDIA Grace CPU 1개를 NVLink-C2C 인터커넥트와 연결한다는 점에서 **NVIDIA GB200 NVL72**에서 중요한 역할을 하는 구성 요소입니다.

실시간 LLM 추론

GB200 NVL72는 최첨단 기능과 2세대 트랜스포머 엔진을 도입했습니다. 특히 2세대 트랜스포머 엔진으로 FP4 AI를 구현했으며, 여기에 5세대 NVLink까지 결합되면 조 단위 파라미터 언어 모델의 실시간 추론 성능이 30배 더 빨라집니다. 차세대 Tensor 코어로 마이크로스케일링 형식을 새롭게 도입하여 정확도를 높이고 처리량을 늘린 덕분입니다. 또한 GB200 NVL72는 방대한 용량의 단일 72-GPU 랙에서 NVLink와 수냉 방식을 사용하여 통신 병목현상도 해결할 수 있습니다.

GPT-MoE-1.8T 실시간 처리량



LLM 추론 및 에너지 효율: 토큰 간 지연 시간(TTL) = 실시간 50밀리초(ms), 첫 토큰 지연 시간(FTL) = 5초, 32,768 입력/1,024 출력, InfiniBand(IB)를 통해 확장된 NVIDIA HGX™ H100과 GB200 NVL72의 비교

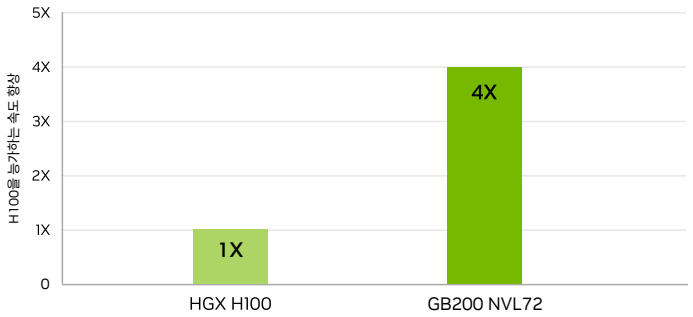
주요 특징

- NVIDIA Grace™ CPU 36개
- NVIDIA Blackwell GPU 72개
- 최대 17테라바이트(TB)의 LPDDR5X 메모리 (오류 보정 코드 포함)
- 최대 13.5TB의 HBM3e 지원
- 최대 30.5TB의 고속 액세스 메모리
- NVLINK 도메인: 초당 130테라바이트(TB/s)의 저지연 GPU 통신

대규모 훈련

GB200 NVL72는 더욱 빨라진 2세대 트랜스포머 엔진이 탑재되어 FP8 정밀도를 구현할 뿐만 아니라 대규모 언어 모델의 훈련 속도가 무려 4배나 더 빠릅니다. 이러한 성능은 5세대 NVLink로 한층 더 향상되어 초당 1.8테라바이트(TB/s)의 GPU-GPU 인터커넥트, InfiniBand 네트워킹, NVIDIA Magnum IO™ 소프트웨어를 제공합니다.

GPT-MoE-1.8T 모델 훈련 속도 향상

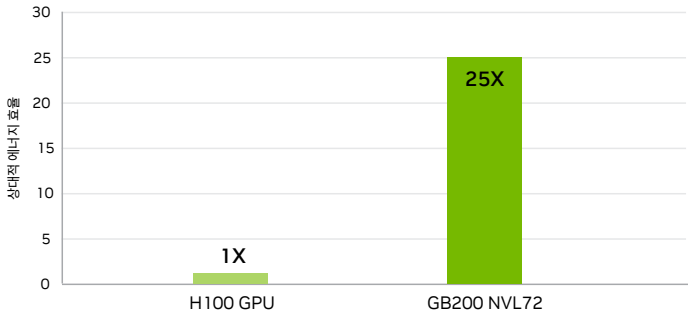


LLM 추론 및 에너지 효율: 토큰 간 지연 시간(TTL) = 실시간 50밀리초(ms), 첫 토큰 지연 시간(FTL) = 5초, 32,768 입력/1,024 출력, InfiniBand(IB)를 통해 확장된 NVIDIA HGX H100과 GB200 NVL72의 비교

에너지 효율이 높은 인프라

수냉식 GB200 NVL72 랙은 데이터 센터의 탄소 발자국과 에너지 사용량을 효과적으로 줄여 줍니다. 수냉식은 컴퓨팅 밀도를 높이고, 사용 공간을 줄이는 데 유용할 뿐만 아니라 **NVLink 도메인 아키텍처**를 통해 대역폭이 높고 지연 시간이 낮은 GPU 통신을 지원합니다. 따라서 GB200은 NVIDIA H100 공냉식 인프라와 비교하면 동일한 컴퓨팅 파워로 25배 빠른 성능을 자랑하는 동시에 물 사용량도 줄일 수 있습니다.

에너지 효율

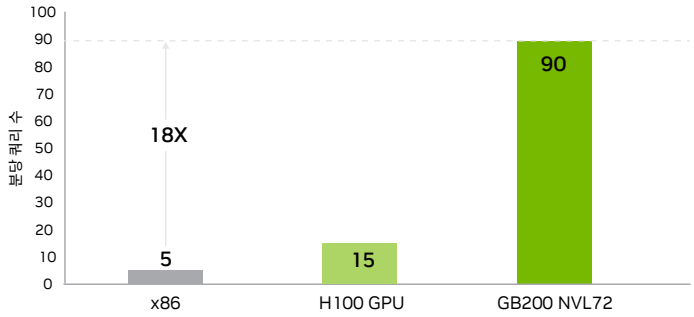


GPT-MoE-1.8T의 실시간 추론 처리량 성능이 동일할 때 65랙 8방향 HGX H100 공냉식과 1랙 GB200 NLV72 수냉식의 에너지 절감량

데이터 처리

데이터베이스는 대용량의 엔터프라이즈 데이터를 처리하고 분석하는 데 매우 중요한 역할을 합니다. GB200은 **NVIDIA Blackwell 아키텍처**에서 고대역폭 메모리 성능, **NVLink-C2C**, 전용 압축해제 엔진을 이용하여 CPU와 비교했을 때 주요 데이터베이스 쿼리의 속도를 18배까지 높여 TCO를 5배 절감합니다.

데이터베이스 조인 쿼리



결과는 변경될 수 있습니다.

완벽한 NVIDIA 플랫폼 지원

NVIDIA GB200 Grace Blackwell 슈퍼칩은 기존에 폭넓고 다양했던 64비트 Arm® 프로세서 에코시스템을 한층 더 확장합니다. 다른 Arm 제품에서 실행되는 컨테이너, 애플리케이션 바이너리 및 운영 체제가 아무런 변경 없이 동일하게 Grace Blackwell에서도 실행되지만 속도는 더욱 빠릅니다. 또한 NVIDIA의 소프트웨어 전문 기술을 이용해 개발하고 싶은 고객을 위해 NVIDIA Grace Blackwell 슈퍼칩은 NVIDIA HPC, NVIDIA AI, NVIDIA Omniverse™ 플랫폼 등 전체 NVIDIA 소프트웨어 스택에서도 지원됩니다.

제품 사양[1]

NVIDIA GB200 Grace Blackwell 슈퍼칩은 다음 두 가지 구성 (GB200 NVL72, GB200 NVL2) 으로 제공됩니다.

Feature	GB200 NVL72	GB200 NVL2	GB200 Grace Blackwell Superchip
Configuration	36 Grace CPUs, 72 Blackwell GPUs	2 Grace CPUs, 2 Blackwell GPUs	1 Grace CPU, 2 Blackwell GPUs
FP4 Tensor Core[2]	1,440 PFLOPS	40 PFLOPS	40 PFLOPS
FP8/FP6 Tensor Core[2]	720 PFLOPS	20 PFLOPS	20 PFLOPS
INT8 Tensor Core[2]	720 POPS	20 POPS	20 POPS
FP16/BF16 Tensor Core[2]	360 PFLOPS	10 PFLOPS	10 PFLOPS
TF32 Tensor Core[2]	180 TFLOPS	5 PFLOPS	5 PFLOPS
FP32	6,480 TFLOPS	180 TFLOPS	180 TFLOPS
FP64	3,240 TFLOPS	90 TFLOPS	90 TFLOPS
FP64 Tensor Core	3,240 TFLOPS	90 TFLOPS	90 TFLOPS
GPU Memory | Bandwidth	Up to 13.5TB HBM3e | 576TB/s	Up to 384GB HBM3e | 16TB/s	Up to 384GB HBM3e | 16TB/s
NVLink Bandwidth	130TB/s	3.6 TB/s	3.6TB/s
CPU Core Count	2,592 Arm Neoverse V2 cores	144 Arm Newoverse V2 cores	72 Arm Neoverse V2 cores
CPU Memory | Bandwidth	Up to 17TB LPDDR5X | Up to 18.4TB/s	Up to 960GB LPDDR5X | Up to 1,024GB/s	Up to 480GB LPDDR5X | Up to 512GB/s
Form Factor	MGX Rack	MGX	Module

1. Preliminary specifications. May be subject to change.
2. With sparsity.

시작할 준비가 되었습니까?

NVIDIA GB200 NVL72에 대해 자세히 알고 싶다면
아래 웹사이트를 방문하십시오.
https://www.nvidia.com/ko-kr/data-center/gb200-nvl72/