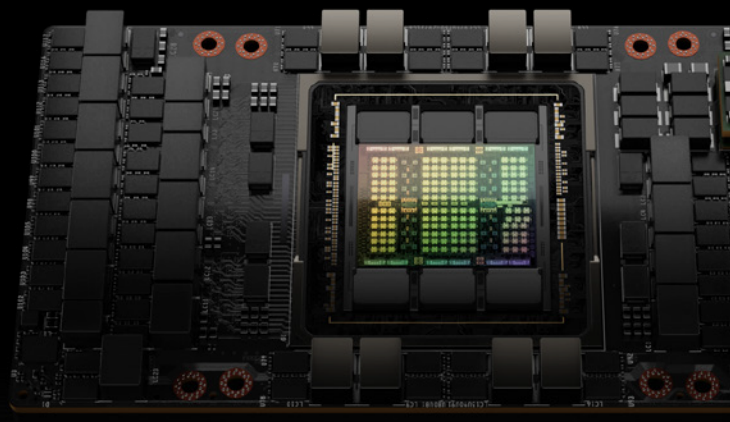
DATASHEET



NVIDIA H100 TENSOR CORE GPU

모든 데이터 센터를 위한 전례 없는 성능,
확장성, 보안

가속 컴퓨팅의 새로운 단계로의 도약

NVIDIA H100 Tensor Core GPU는 모든 워크로드에 대해 전례 없는 성능, 확장성, 보안을 제공합니다. NVIDIA® NVLink® 스위치 시스템과 함께, 최대 256 H100 GPU를 연결해 엑사스케일 워크로드의 가속화가 가능하며, 전용 Transformer Engine은 조 단위의 파라미터 언어 모델을 지원합니다. H100은 NVIDIA Hopper™ 아키텍처의 획기적인 혁신을 통해 업계를 선도하는 대화형 AI를 제공하고, 대규모 언어 모델의 속도를 이전 세대 대비 약 30배 빠르게 해줍니다.

엔터프라이즈에서 엑사스케일로 안전하게 워크로드를 가속화

4세대 Tensor Core와 FP8 precision의 트랜스포머 엔진을 갖춘 NVIDIA H100 GPU는 대규모 언어 모델에서 최대 9배 빠른 훈련 속도와 놀랍게도 약 30배에 달하는 추론 속도 개선으로 NVIDIA의 시장 선도적 AI 리더십을 더욱 확장하고 있습니다. 고성능컴퓨팅(HPC) 어플리케이션의 경우, H100은 FP64의 FLOPS를 3배로 만들고 여기에 동적 프로그래밍(DPX) 명령어를 더해 최대 7배 높은 성능을 제공합니다. NVIDIA 컨피덴셜 컴퓨팅에서 개발된 2세대 Multi-Instance GPU (MIG)와 NVIDIA NVLink 스위치 시스템을 통해, H100은 엔터프라이즈에서 엑사스케일에 이르는 각 데이터 센터의 모든 워크로드를 안전하게 가속화합니다.

H100을 포함하는 완전한 NVIDIA 데이터 센터 솔루션은 하드웨어, 네트워킹, 소프트웨어, 라이브러리 및 NVIDIA NGC™ 카탈로그에서 최적화된 AI 모델과 어플리케이션에 이르는 구성 요소(building blocks)들로 이루어집니다. 데이터 센터를 위한 가장 강력한 엔드 투 엔드 AI와 HPC 플랫폼으로, 연구자들은 대규모의 프로덕션 환경에 현실 세계의 결과를 제공하고 솔루션을 구현할 수 있습니다.

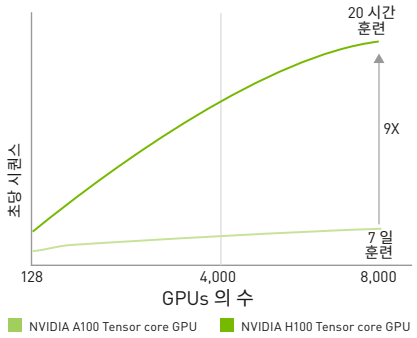
사양

	H100 SXM	H100 PCIe
FP64	**30 TFLOPS**	**24 TFLOPS**
FP64 Tensor Core	**60 TFLOPS**	**48 TFLOPS**
FP32	**60 TFLOPS**	**48 TFLOPS**
TF32 Tensor Core	**1,000 TFLOPS***	**800 TFLOPS***
BFLOAT16 Tensor Core	**2,000 TFLOPS***	**1,600 TFLOPS***
FP16 Tensor Core	**2,000 TFLOPS***	**1,600 TFLOPS***
FP8 Tensor Core	**4,000 TFLOPS***	**3,200 TFLOPS***
INT8 Tensor Core	**4,000 TOPS***	**3,200 TOPS***
GPU 메모리	**80GB**	**80GB**
GPU 메모리 대역폭	**3TB/s**	**2TB/s**
디코더	**7 NVDEC 7 JPEG**	**7 NVDEC 7 JPEG**
최대 열 설계 전력 (TDP)	**700W**	**350W**
Multi-Instance GPUs	**Up to 7 MIGS @ 10GB each**	
폼 팩터	**SXM**	**PCIe dual-slot air-cooled**
인터커넥트	**NVLink: 900GB/s PCIe Gen5: 128GB/s**	**NVLink: 600GB/s PCIe Gen5: 128GB/s**
서버 옵션	**NVIDIA HGX™ H100 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs NVIDIA DGX™ H100 with 8 GPUs**	**Partner and NVIDIA-Certified Systems with 1–8 GPUs**

* 희소성(Sparsity) 적용. 희소성 제외 시 사양은 ½ 낮아짐.

최대 9배나 높은 AI 훈련
(최대 규모 모델)

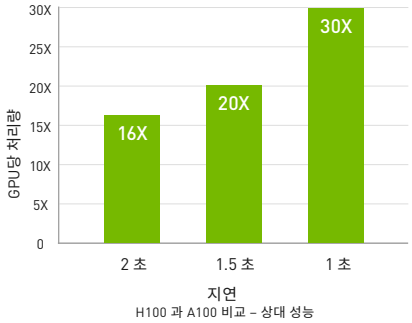
Mixture of Experts
(3950억 퍼라미터)



성능 예상치는 변경될 수 있음. Mixture of Experts (MoE) Transformer Switch-XXL훈련은 1T 토큰 데이터세트 3950억 퍼라미터와 다름 | A100 cluster: HDR IB network | H100 cluster: NVLink Switch System, NDR IB

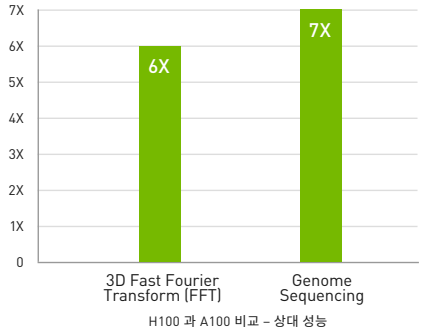
최대 30배 높은 AI 추론 성능
(최대 규모 모델)

Megatron Chatbot Inference
(5300억 퍼라미터)



성능 예상치는 변경될 수 있음. Megatron 5300억 퍼라미터 모델 챗 봇 추론 입력 시퀀스 길이=128, 출력 시퀀스 길이=20 | A100 클러스터: HDR IB 네트워크 | H100 클러스터: 16 H100 구성을 위한 NDR IB 네트워크 | 1과 1.5 초에서 32 A100 대 16 H100 | 2초에서 16 A100 대 8 H100

최대 7배 높은 HPC 어플리케이션 성능



성능 예상치는 변경될 수 있음. 3D FFT (4K^3) 처리량 | A100 클러스터: HDR IB 네트워크 | H100 클러스터: NVLink 스위치 시스템, NDR IB | Genome Sequencing (Smith- Waterman) | 1 A100 | 1 H100

NVIDIA Hopper의 획기적 기술

세계 최고의 첨단 칩

NVIDIA 가속 컴퓨팅 니즈에 맞춤화된 최신 TSMX 4N 공정을 이용해 800억 트랜지스터로 개발된 H100은 지금껏 개발된 칩 중 가장 최첨단의 칩입니다. 데이터 센터 스케일에서 AI, HPC, 메모리 대역폭, 인터커넥트와 통신을 가속화하기위한 주요 개발사항들을 보여줍니다.

트랜스포머 엔진

트랜스포머 엔진은 세계에서 가장 중요한 AI 모델 빌딩 블록인 트랜스포머에서 개발한 모델의 훈련 가속화를 위해 설계된 소프트웨어와 Hopper Tensor Core 기술을 사용합니다. Hopper Tensor Core는 트랜스포머 AI 연산의 극적인 가속화를 위해 FP8와 FP16 precision을 혼합 적용할 수 있습니다.

NVLINK 스위치 시스템

NVLink 스위치 시스템은 다수의 서버에서 multi-GPU input/output (IO)가 GPU당 양방향 900 GB/s 로 확장될 수 있도록 해주는데, 이는 PCIe 5세대 대역폭의 7배가 넘습니다. 시스템은 최대 256 H100 클러스터를 지원하며 NVIDIA Ampere 아키텍처의 InfiniBand HDR보다 9배 높은 대역폭을 제공합니다.

NVIDIA 컨피덴셜 컴퓨팅

NVIDIA 컨피덴셜 컴퓨팅은 Hopper에 탑재되어 있는 보안기능으로 NVIDIA H100은 컨피덴셜 컴퓨팅 기능을 갖춘 세계 최초의 가속기입니다. 사용자는 H100 GPU의 가속화를 아무 제약 없이 누리면서 사용중인 데이터와 어플리케이션의 기밀성과 무결성을 보호할 수 있습니다.

2세대 MULTI-INSTANCE GPU (MIG)

Hopper 아키텍처의 2세대 MIG는 가상 환경에서 다중 테넌트, 다중 사용자 구성을 지원합니다. GPU를 격리된 적절한 사이즈의 인스턴스로 안전하게 파티셔닝함으로써 7배나 더 많은 보안이 중요한 테넌트에게 최대한의 QoS를 제공합니다.

DPX 명령어

Hopper의 DPX 명령어는 동적 프로그래밍 알고리즘을 CPU 대비 40배, NVIDIA Ampere 아키텍처 GPU 대비 7배 가속화 합니다. 이를 통해 질병의 진단, 실시간 경로 최적화와 그래프 분석이 놀랄 만큼 빨라집니다.

NVIDIA H100 CNX Converged Accelerator

NVIDIA H100 CNX는 하나의 유니크한 플랫폼에 NVIDIA H100의 파워와 **NVIDIA ConnectX®-7** 스마트 네트워크 인터페이스 카드(SmartNIC)의 첨단 네트워킹 역량을 결합합니다. 이러한 결합을 통해 엔터프라이즈 데이터 센터의 분산형 AI 훈련과 엣지에서의 5G 처리등 GPU 기반의 IO 집약적인 워크로드에서 독보적인 성능을 제공합니다.

NVIDIA H100 CNX에 대해 더 자세히 알아보세요.

Enterprise-Ready

전세계 AI 인프라의 새로운 엔진인 NVIDIA Hopper 아키텍처 기반 NVIDIA H100 Tensor Core GPU는 NVIDIA 데이터 센터 플랫폼의 핵심입니다. 딥 러닝, HPC와 데이터 분석을 위해 개발된 이 플랫폼은 모든 주요 딥 러닝 프레임워크 포함 2,700개 이상의 어플리케이션을 가속화합니다. 추가로, AI와 데이터 분석 소프트웨어의 엔드 투 엔드 클라우드 네이티브 스위트인 NVIDIA AI Enterprise는 하이퍼바이저 기반 가상 인프라에서 Vmware vShpere와 함께 H100에서 실행을 인증 받았습니다. 따라서 하이브리드 클라우드 환경에서 AI 워크로드의 관리와 확장이 가능합니다. 데이터 센터에서 엣지에 이르기까지, 완전한 NVIDIA 플랫폼은 어디서든 사용 가능하며, 놀라운 성능 증가 및 비용 절감의 기회 모두를 제공합니다.

기업에 최적화된 소프트웨어와 서비스

NVIDIA CERTIFIED

모든 딥 러닝 프레임워크

mxnet PYTORCH APACHE Spark TensorFlow

2,000+ GPU 가속 어플리케이션

- Altair nanoFluidX
- Altair ultraFluidX
- AMBER
- ANSYS Fluent
- DS SIMULIA Abaqus
- GAUSSIAN
- GROMACS
- NAMD
- OpenFOAM
- VASP
- WRF
- Simcenter STAR-CCM+

시작할 준비가 되셨나요?

NVIDIA H100 Tensor Core GPU에 대한 자세한 내용은
https://www.nvidia.com/ko-kr/data-center/h100/ 을 방문하세요.

NVIDIA