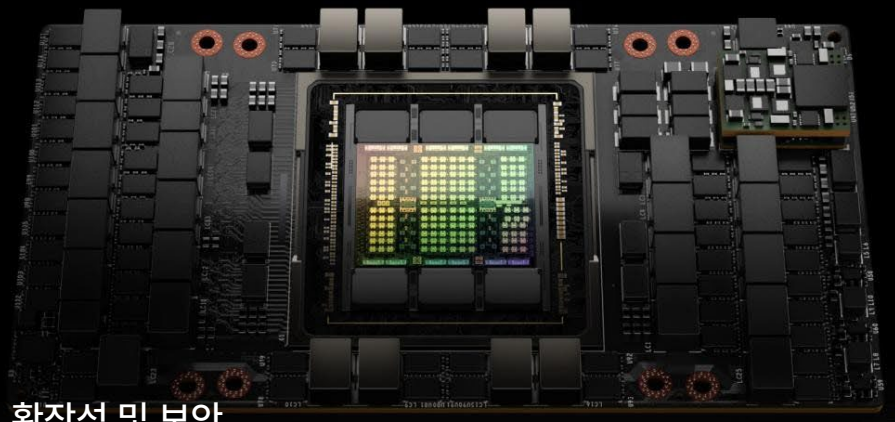
PREFERRED SOLUTION PROVIDER



NVIDIA H100 TENSOR CORE GPU

모든 데이터 센터를 위한 전례 없는 성능, 확장성 및 보안.

가속화된 컴퓨팅에서 한단계 더 도약 합니다

NVIDIA H100 Tensor Core GPU는 모든 워크로드에 전례 없는 성능, 확장성 및 보안을 제공합니다. NVIDIA® NVLink® 스위치 시스템을 사용하면 최대 256개의 H100 GPU를 연결하여 엑사스케일 워크로드를 가속화할 수 있으며 전용 Transformer Engine은 1조 개의 매개변수 언어 모델을 지원합니다.

H100은 NVIDIA Hopper™ 아키텍처의 획기적인 혁신 기술을 사용하여 업계 최고의 대 화형 AI를 제공하고 이전 세대에 비해 대규모 언어 모델의 속도를 30배 높입니다.

엔터프라이즈 AI를 사용할 준비가 되셨습니까?

메인스트림 서버용 NVIDIA H100 Tensor Core GPU 는 NVIDIA AI Enterprise 소프트웨어 제품군에 대한 엔터프라이즈 지원을 포함한 5년 소프트웨어 구독으로 최고의 성능으로 AI 채택을 간소화합니다.

이를 통해 조직은 AI 챗봇, 추천 엔진, 비전 AI 등과 같은 H100 가속 AI 워크플로를 구축하는 데 필요한 AI 프레임워크 및 도구에 액세스할 수 있습니다. 여기 에 서 NVIDIA H100에 대한 NVIDIA AI Enterprise 소프트웨어 구독 및 관련 지원 혜택에 액세스 하십시오.

엔터프라이즈에서 엑사스케일까지 워크로드를 안전하게 가속화합니다.

NVIDIA H100 GPU는 4세대 Tensor 코어와 FP8 정밀도를 갖춘 Transformer Engine을 특징으로 하며, 대규모 언어 모델에서 최대 9배 더 빠른 훈련과 30배 놀라운 추론 속도 향상으로 NVIDIA의 시장 선도 AI 리더십을 더욱 확장합니다.

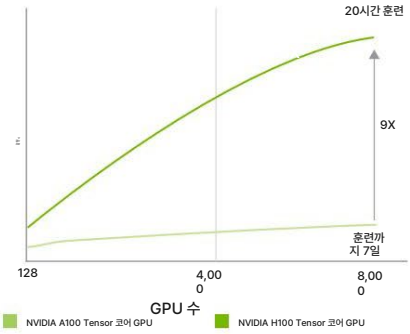
고성능 컴퓨팅 (HPC) 애플리케이션의 경우 H100은 FP64의 FLOPS(초당 부동 소수점 연산)를 3배로 늘리고 동적 프로그래밍(DPX) 명령을 추가하여 최대 7배 더 높은 성능을 제공합니다. 2세 대 멀티 인스턴스 GPU(MIG), 내장형 NVIDIA 기밀 컴퓨팅 및 NVIDIA NVLink 스위치 시스템을 갖춘 H100은 엔터프라이즈에서 엑사스케일에 이르는 모든 데이터 센터의 모든 워크로드를 안전하게 가속화합니다.

	H100 SXM	H100 PCIe
FP64	34 TFLOPS	26 TFLOPS
FP64 Tensor Core	67 TFLOPS	51 TFLOPS
FP32	67 TFLOPS	51 TFLOPS
TF32 Tensor Core	989 TFLOPS *	756 TFLOPS*
비플로트16 Tensor Core	1,979TFLOPS*	1,513 TFLOPS*
FP16 Tensor Core	1,979TFLOPS*	1,513 TFLOPS*
FP8 Tensor Core	3,958TFLOPS*	3,026 TFLOPS*
INT8 텐서 코어	3,958 TOPS*	3,026 TFLOPS *
GPU 메모리	80GB	80GB
GPU 메모리 대역폭	3.35TB/s	2TB/s
디코더	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
최대 열 설계 전력(TDP)	최대 700W (구성 가능)	300-350 W (구성 가능)
다중 인스턴스 GPU	각각 10GB에서 최대 7개의 MIGS	
Form Factor	SXM	PCIe 듀얼 슬롯 에어 쿨링
Interconnect	NV링크: 900GB/s PCIe 5세대: 128GB/s	NV링크: 600GB/s PCIe 5세대: 128GB/s
Server Option	NVIDIA HGX™ H100 partner and NVIDIACertified Systems™ with 4 or 8 GPUs NVIDIA DGX™ H100 with 8 GPUs	Partner and NVIDIA Certified Systems with 1–8 GPUs
NVIDIA AI Enterprise	애드온	포함

* 희소성으로 표시됩니다. 희소성 없이 사양이 1/2 낮습니다.

최대 9배 더 높은 AI 학습 가장 큰 모델

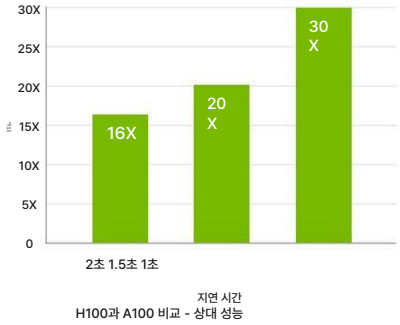
전문가 혼합(3950억 매개변수)



예상 실적은 변경될 수 있습니다. 1T 토큰 데이터 세트에서 395B 매개변수가 포함된 MoE(전문가 혼합) Transformer Switch-XXL 변형 | A100 클러스터: HDR IB 네트 워크 | H100 클러스터: NVLink 스위치 시스템, NDR IB

최대 30배 더 높은 AI 추론 가장 큰 모델의 성능

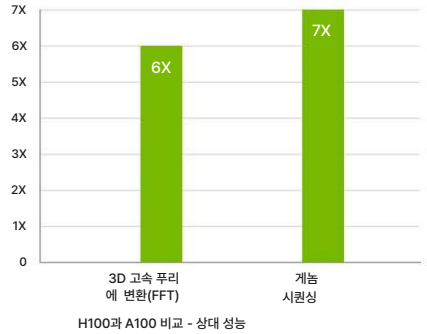
Megatron 챗봇 추론 (5300억 매개변수)



예상 실적은 변경될 수 있습니다. 입력 시퀀스 길이=128, 출력 시퀀스 길이=20에 대한 Megatron 530B 매개변수 모델 챗봇에 대한 추론 | A100 클러스터: HDR IB 네트 워크 | H100 클러스터: 16개의 H100 구성을 위한 NDR IB 네트워크 | 32 A100 vs 16 H100 1초 및 1.5초 | 16 A100

최대 7배 더 높은 성능

HPC 애플리케이션



예상 실적은 변경될 수 있습니다. 3D FFT(4K^3) 처리량 | A100 클러스터: HDR
IB 네트워크 | H100 클러스터: NVLink 스위치 시스템, NDR IB | 게놈 시퀀싱(Smith Waterman) | 1 A100 | 1H100

NVIDIA Hopper의 기술 혁신

엔비디아 H100 Tensor Core GPU

NVIDIA의 가속화된 컴퓨팅 요구 사항에 맞게 맞춤화된 최첨단 TSMC 4N 프로세스를 사용하여 800억 개의 트랜지스터로 구축된 H100은 세계에서 가장 진보된 칩입니다.

데이터 센터 규모에서 AI, HPC, 메모리 대역폭, 상호 연결 및 통신을 가 속화하는 주요 발전 기능을 제공합니다.

트랜스포머 엔진

Transformer Engine은 소프트웨어와 Hopper Tensor Core 기술을 사용하여 세계 에서 가장 중요한 AI 모델 빌딩 블록인 Transformer에서 구축된 모델의 교육을 가속화하도록 설계되었습니다

Hooper Tensor Core는 혼합 FP8 및 FP16 정밀도를 적용하여 변압기에 대한 AI 계산을 획기적으로 가속화할 수 있습니다.

NVLink 스위치 시스템

NVLink 스위치 시스템은 PCIe Gen5 대역폭의 7배 이상인 GPU당 양방향 초당 900기가바이 트(GB/s)로 여러 서버에서 다중 GPU 입/출력(IO)의 확장을 가능하게 합니다. 이 시스템은 최대 256개의 H100 클러스터를 지원하고 NVIDIA Ampere 아키텍처에서 InfiniBand HDR보다 9배 높은 대역폭을 제공합니다.

NVIDIA 컨피덴셜 컴퓨팅

NVIDIA Hooper 아키텍처는 세계 최초로 NVIDIA 컨피덴셜 컴퓨팅 기능을 탑재한 컴퓨팅 플랫폼을 도입했습니다.

사용자는 H100 GPU의 탁월한 가속에 액세스하면서 사용 중인 데이터 및 애플리케이션의 기밀성과 무결성을 보호할 수 있습니다.

2세대 멀티 인스턴스 GPU(MIG)

Hopper 아키텍처의 2세대 MIG는 가상 화된 환경에서 다중 테넌트, 다중 사용자 구성을 지원하여 GPU를 격리된 적절한 크기의 인스턴스로 안전하게 분할하여 7배 더 안전한 서비스 품질 (QoS)을 극대화합니다.

DPX 명령어

Hopper의 DPX 명령어는 동적 프로그래 밍 알고리즘을 CPU에 비해 40배, NVIDIA Ampere 아키텍처 GPU에 비해 7배 가속화 합니다. 이로 인해 질병 진단, 실시간 라우팅 최적화 및 그래프 분석 시간이 크게 단축됩니다.

GPU와 SmartNIC의 융합.

NVIDIA H100 CNX는 NVIDIA H100의 성능과 NVIDIA ConnectX® 의 고급 네트워킹 기능을 결합합 니다. - 하나의 고유한 플랫폼에 있는 7 개의 스마트 네트워크 인터페이스 카드(SmartNIC). 이 컨버전스는 엔터프라이즈 데이터 센터의 분산 AI 교육 및 에지의 5G 처리와 같은 GPU 기반 IO 집약적 워크로드에 탁월 한 성능을 제공합니다.

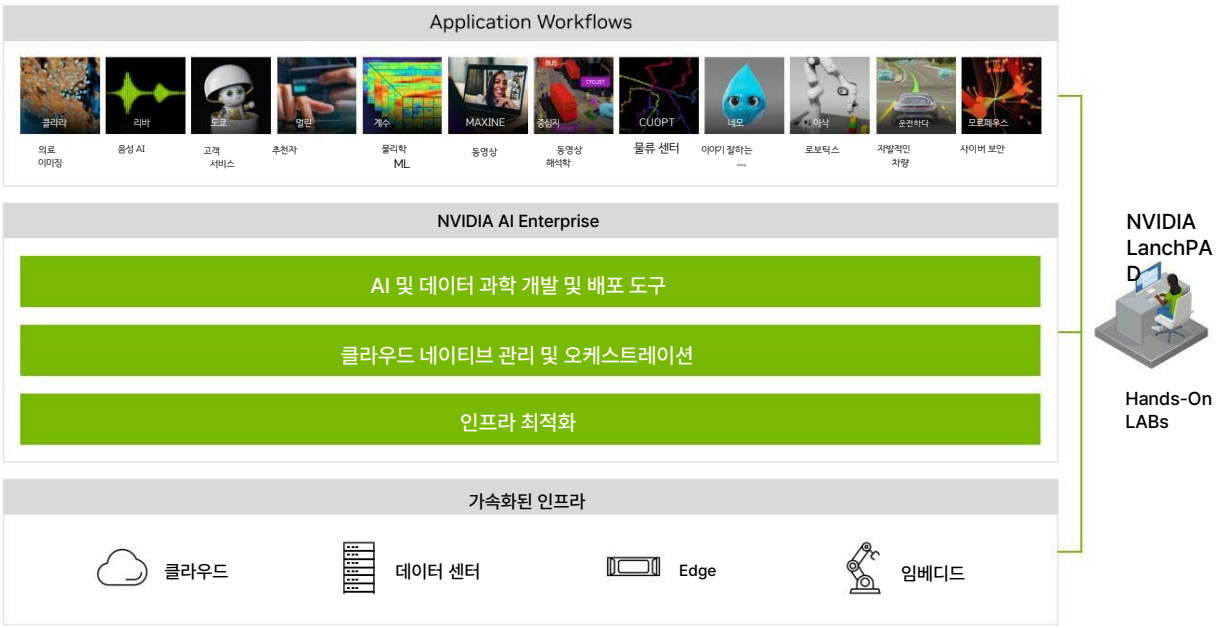
NVIDIA H100 CNX에 대해 자세히 알아보십시오.

어디서나 모든 워크로드를 가속화합니다.

NVIDIA H100은 NVIDIA 데이터 센터 플랫폼의 필수적인 부분입니다. AI, HPC 및 데이터 분석을 위해 구축된 이 플랫폼은 3,000개 이상의 애플리케이션을 가속화하고 데이터 센터에서 에지까지 어디에서나 사용할 수 있 어 극적인 성능 향상과 비용 절감 기회를 제공합니다.

NVIDIA AI 플랫폼으로 H100을 배포하십시오.

NVIDIA AI는 NVIDIA H100 GPU에 구축된 프로덕션 AI를 위한 종단 간 개방형 플랫폼입니다. 여기에는 NVIDIA 가속 컴퓨팅 인프라, 인프라 최적화 및 AI 개발 및 배포를 위한 소프트웨어 스택, 시장 출시 시간을 단 축하기 위한 애플리케이션 워크플로가 포함됩니다. NVIDIA LaunchPad 에서 NVIDIA AI 및 NVIDIA H100 을 경험할 수 있으며 무료 실습 자료가 준비 되어 있습니다.



제품 문의

02-6206-5024

jysuh@krinfra.co.kr

(주) 한국인프라
Infra Information Technology Co., Ltd.

06288) 서울특별시 강남구 삼성로 150(대치동,극동교회빌딩 3층) | Tel. 02-6204-5000 | Fax. 02-6204-5099
51139) 경남 창원시 의창구 용동로 85 한마음빌딩 805호 | Tel. 055-262-8086 | Fax. 055-722-7218