데이터시트

NVIDIA L40S

데이터 센터를 위한 탁월한 AI 및 그래픽 성능

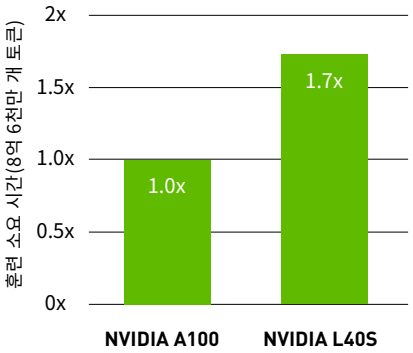
생성형 AI는 모든 산업 분야의 기업들에게 새로운 기회의 지평을 열어주는 등 혁신적인 변화를 촉진하고 있습니다. AI를 통한 혁신을 원하는 기업에게는 계속해서 증가하는 다양하고 복잡한 워크로드의 요구 사항을 충족할 수 있는 더 많은 컴퓨팅 리소스와 더 큰 규모, 다양한 기능이 필요합니다.

NVIDIA L40S GPU는 데이터센터를 위한 가장 강력한 범용 GPU로, **생성형 AI,** 모델 훈련 및 추론부터 3D 그래픽, 렌더링 및 비디오 애플리케이션에 이르는 차세대 AI 지원 애플리케이션을 위한 엔드 투 엔드 가속 솔루션을 제공합니다.

차세대 워크로드 가속화

- 생성형 AI
- LLM(Large Language Model) 훈련 및 추론
- NVIDIA Omniverse™ Enterprise
- 렌더링 및 3D 그래픽
- 스트리밍 및 비디오 콘텐츠

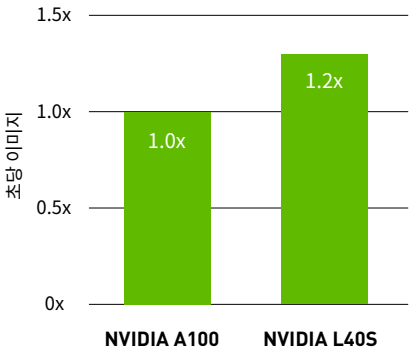
AI 훈련



GPT-408 LoRA (상대적 성능)

LoRA 미세 조정 (GPT-40B): 글로벌 훈련 배치(batch) 크기: 128 (시퀀스), 시퀀스 길이: 256 (토큰). NVIDIA HGX™ A100 8-GPU 대 L40S GPU 4개가 탑재된 2개 시스템. 출시 전 빌드에서의 성능(변경될 수 있음).

생성형 AI



Stable Diffusion, 512x512 (상대적 성능)

Stable Diffusion v2.1. 512 x 512 해상도 이미지 생성 시 상대적 속도 향상. NVIDIA HGX A100 8-GPU 대 L40S GPU 4개가 탑재된 2개 시스템. 출시 전 빌드에서의 성능(변경될 수 있음).

NVIDIA Ada Lovelace 아키텍처 기반

4세대 Tensor 코어

구조적 희소성과 최적화된 TF32 형식이 지원되는 하드웨어가 신속한 AI 및 데이터 과학 모델 훈련을 위해 즉각적인 성능 이점을 제공합니다. 또한, 일부 애플리케이션에서 더 나은 성능으로 해상도를 높일 수 있도록 **DLSS**를 통해 AI 강화 그래픽 기능을 가속화합니다.

3세대 RT 코어

향상된 처리량과 동시 레이 트레이싱 및 셰이딩 기능을 통해 레이 트레이싱 성능을 개선함으로써 제품 디자인 및 아키텍처, 엔지니어링, 건설 워크플로에서 렌더링을 가속화합니다. 또한, 하드웨어 가속 모션 블러와 놀라운 실시간 애니메이션을 통해 디자인이 실제 작동하는 모습을 확인합니다.

트랜스포머 엔진 (Transformer Engine)

트랜스포머 엔진은 AI 성능을 획기적으로 가속화하고 훈련 및 추론 시 메모리 활용도를 높여줍니다. 또한, **Ada Lovelace 4세대 Tensor 코어**의 성능을 활용하여 트랜스포머 아키텍처 신경망의 레이어를 지능적으로 스캔하고, FP8 정밀도와 FP16 정밀도 간을 자동으로 리캐스트하여 더 빠른 AI 성능을 제공하고 훈련 및 추론을 가속화합니다.

데이터 센터에 적합

L40S GPU는 하루 24시간 연중무휴로 운영되는 엔터프라이즈 데이터 센터에 최적화되어 있으며, 최대 성능, 내구성 및 가동 시간을 보장하도록 NVIDIA가 설계, 구축, 테스트 및 지원을 담당하고 있습니다. 또한, 최신 데이터 센터 표준인 NEBS(Network Equipment-Building System) 레벨 3를 충족하는 것은 물론이고 RoT(Root of Trust) 기술을 통한 보안 부팅 기능을 갖추고 있어 데이터 센터에 추가적인 보안 계층을 제공합니다.

Technical Specifications	
GPU Architecture	NVIDIA Ada Lovelace Architecture
GPU Memory	48GB GDDR6 with ECC
Memory Bandwidth	864GB/s
Interconnect Interface	PCIe Gen4 x16: 64GB/s bidirectional
NVIDIA Ada Lovelace Architecture-Based CUDA® Cores	18,176
NVIDIA Third-Generation RT Cores	142
NVIDIA Fourth-Generation Tensor Cores	568
RT Core Performance TFLOPS	209
FP32 TFLOPS	91.6
TF32 Tensor Core TFLOPS	183 I 366*
BFLOAT16 Tensor Core TFLOPS	362.05 I 733*
FP16 Tensor Core	362.05 I 733*
FP8 Tensor Core	733 I 1,466*
Peak INT8 Tensor TOPS	733 I 1,466*
Peak INT4 Tensor TOPS	733 I 1,466*
Form Factor	4.4" (H) x 10.5" (L), dual slot
Display Ports	4x DisplayPort 1.4a
Max Power Consumption	350W
Power Connector	16-pin

Thermal	Passive
Virtual GPU (vGPU) Software Support	Yes
vGPU Profiles Supported	See the **virtual GPU licensing guide**
NVENC I NVDEC	3x I 3x (includes AV1 encode and decode)
Secure Boot With Root of Trust	Yes
NEBS Ready	Level 3
MIG Support	No
NVIDIA® NVLink® Support	No

* With sparsity

시작할 준비가 되셨나요?

NVIDIA L40S에 대한 자세한 내용은 아래 사이트에서 확인하십시오.
www.nvidia.com/l40s