NVIDIA

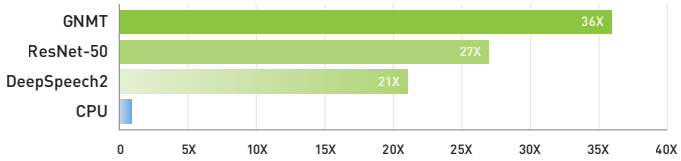
NVIDIA T4
TENSOR CORE GPU

스케일 아웃 AI 훈련 및 추론 능력을 강화하고 싶다면

NVIDIA T4 엔터프라이즈 GPU는 표준 데이터 센터 인프라에 쉽게 적용할 수 있는 세계에서 가장 신뢰할만한 메인스트림 서버를 과급합니다. 본 제품의 로우 프로파일(low profile), 70W 디자인은 NVIDIA Turing™ Tensor Cores로 구동되어 머신 러닝, 딥 러닝 그리고 가상화 데스크탑을 포함한 넓은 범위의 모던 애플리케이션을 가속화하기 위한 혁신적인 다중 정밀도 성능을 제공합니다.

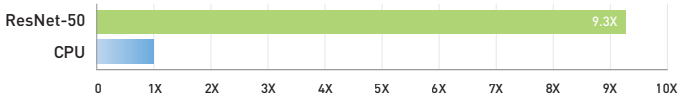
성능이 향상된 본 GPU는 엔터프라이즈 데이터센터와 클라우드에서 유용성을 최대화하기 위해 에너지 효율적인 70W의 소형 PCIe 폼 팩터에 패키징 되어 있습니다.

추론 성능

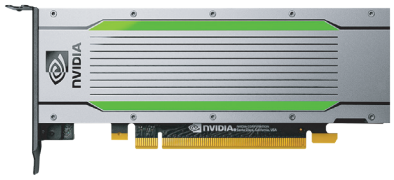


NVIDIA T4 GPU 와 듀얼 소켓 Xeon Gold 6140 CPU 를 사용하는 서버의 비교 분석

훈련 성능



NVIDIA T4 GPU 와 듀얼 소켓 Xeon Gold 6140 CPU 를 사용하는 서버의 비교 분석

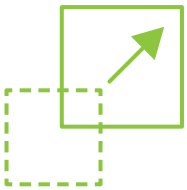


사양

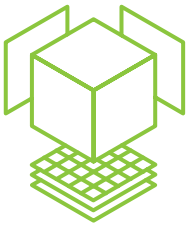
GPU 아키텍처	**NVIDIA Turing**
NVIDIA 튜링 텐서 코어	**320**
NVIDIA CUDA® 코어	**2,560**
단 정밀도 (Single-Precision)	**8.1 TFLOPS**
혼합 정밀도 (Mixed-Precision)	**65 TFLOPS**
INT8	**130 TOPS**
INT4	**260 TOPS**
GPU 메모리	**16GB GDDR6 300 GB/sec**
ECC	**예**
연결 대역폭	**32GB/sec**
시스템 인터페이스	**x16 PCIe**
폼 팩터	**Low-Profile PCIe**
쿨링 솔루션	**수동**
컴퓨팅 APIs	**CUDA, NVIDIA TensorRT™, ONNX**

데이터센터 가속화를 위한 성능

소형 폼 팩터 70와트(W) 설계를 통해 T4는 스케일 아웃 서버구조에 (99%) 최적화되어 CPU 대비 50배 더 높은 에너지 효율을 제공하여 운영 비용을 대폭 절감해줍니다.
지난 2년 동안 NVIDIA의 추론 플랫폼은 효율성을 10배 이상 높였으며, 분산형 AI 훈련 및 추론을 위한 가장 에너지 효율적인 솔루션으로 존재합니다.



NVIDIA T4 데이터 센터 GPU는 분산 컴퓨팅 환경에 이상적인 범용 가속기입니다. 다양한 정밀도 (multi-precision)를 제공하는 혁신적인 성능은 딥러닝, 머신러닝 훈련 및 추론, 비디오 트랜스코딩 및 가상 테스크톱의 가속화를 가능케 합니다. T4는 모든 AI 프레임워크와 네트워크 유형을 지원하며 맞춤 성능 설치에 필요한 제반설비의 성능과 효율성을 극대화합니다.



T4에 탑재된 튜링 텐서 코어 기술은 FP32, FP16 및 INT8, 더 나아가 INT4를 포함한 혼합 정밀도 컴퓨팅을 가능하게 하여 보다 뛰어난 AI 파워를 가능하게 합니다. 훈련에서는 CPU 대비 최대 9.3배, 추론에서는 최대 36배에 달하는 높은 성능을 보입니다.

Turing 의 강력한 RT Cores 는 NVIDIA RTX™ 기술과 결합되어 실시간 레이-트레싱 렌더링이 가능해 물리적으로 정확한 그림자나 반사 및 굴절을 반영한 사실적 객체 및 환경을 제공합니다.

NVIDIA T4에 대한 더 자세한 내용은 **www.nvidia.com/T4** 을 참조하십시오.