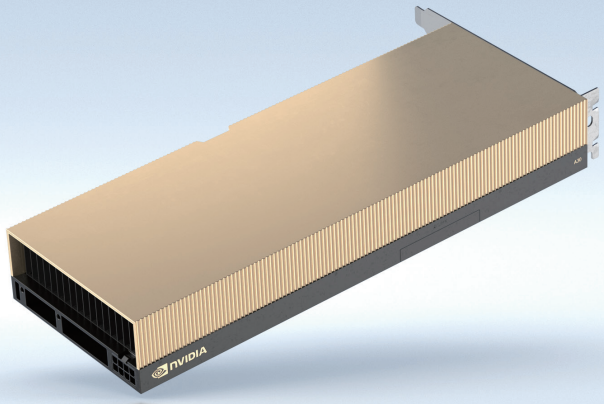
NVIDIA A30
Tenor 코어 GPU

메인스트림 엔터프라이즈 서버를 위한 상황에 맞는 가속 성능 제공



모든 엔터프라이즈를 위한 AI 추론과 메인스트림 컴퓨팅

NVIDIA A30 Tensor 코어 GPU는 AI 추론과 메인스트림 엔터프라이즈 워크로드를 위한 가장 융통성 있는 메인스트림 컴퓨팅 GPU입니다. NVIDIA Ampere 아키텍처 Tensor 코어 기술에 기반, 다양한 수학 정밀도를 지원하고, 모든 워크로드를 가속할 수 있는 단일 가속장치 역할을 수행합니다.

확장 가능한 AI 추론을 위해 개발되어, 동일한 연산 리소스를 이용해 TF32 정밀도로 AI 모델을 신속히 재훈련 시킬 수 있을 뿐 아니라, 고성능 컴퓨팅(HPC) 애플리케이션을 FP64 Tensor 코어를 이용해 가속시킬 수 있습니다. MIG(Multi-Instance GPU)와 FP64 Tensor 코어, 933 GB/s의 빠른 메모리 대역폭이 165W의 낮은 전력 수준에서 제공되며, 모두 메인스트림 서버에 최적화된 PCIe 카드에서 구동됩니다.

3세대 Tensor 코어와 MIG의 결합으로 다양한 워크로드에서 보안성 높은 서비스가 제공되고, 융통성 높은 GPU로 구동되는 만큼 데이터 센터 효율이 보장됩니다. 크고 작은 규모의 워크로드에서 A30의 융통성 높은 컴퓨팅 역량은 메인스트림 기업에 최대의 효용을 제공합니다.

A30는 하드웨어, 네트워킹, 소프트웨어, 라이브러리 및 NVIDIA NGC™ 카탈로그의 최적화된 AI 모델과 애플리케이션으로 이루어진 완전한 NVIDIA 데이터 센터 솔루션의 일부입니다. 데이터 센터를 위한 가장 강력한 엔드-투-엔드 AI 및 HPC 플랫폼을 이용, 리서처들은 실질적인 결과를 도출하고 솔루션을 규모에 맞게 실제 사용을 위해 배포할 수 있습니다.

시스템 사양

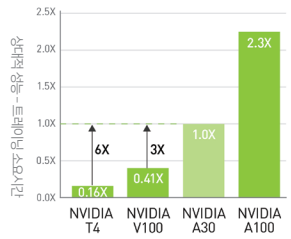
최고 FP64	5.2TF
최고 FP64 Tensor 코어	10.3 TF
최고 FP32	10.3 TF
TF32 Tensor 코어 BFLOAT16 Tensor 코어	82 TF | 165 TF* 165 TF | 330 TF*
최고 FP16 Tensor 코어	165 TF | 330 TF*
최고 INT8 Tensor 코어	330 TOPS | 661 TOPS*
최고 INT4 Tensor 코어	661 TOPS | 1321 TOPS*
미디어 엔진	1개의 광학 플로우 가속기(OFA) 1개의 JPEG 디코더(NVJPEG) 4개의 영상 디코더 (NVDEC)
GPU 메모리	24GB HBM2
GPU 메모리 대역폭	933GB/s
인터커넥트	PCIe Gen4: 64GB/s 3세대 NVIDIA® NVLINK® 200GB/s**
폼 팩터	듀얼 슬롯, 풀 사이즈의 높이·길이(FHFL)
최대 열 설계 전력 (TDP)	165W
MIG (Multi-Instance GPU)	6GB마다 MIG 4개, 12GB마다 MIG 2개, 24GB마다 MIG 1개
가상GPU(vGPU) 소프트웨어 지원	NVIDIA AI Enterprise (VMware NVIDIA 가상컴퓨트서버)

* sparsity 사용시

다양한 워크로드를 위한 놀라운 성능

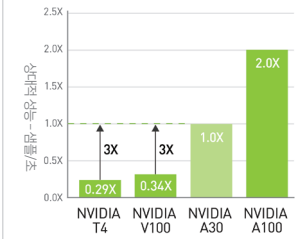
AI 트레이닝 -
V100 대비 최고 3배 높은
처리량과 T41 대비 최고 6배 높은 처리량

BERT Large 사전 트레이닝(일반화)



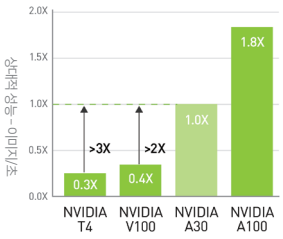
AI 추론 -
실시간 대화형 AI2에서
V100 대비 최고 3배 높은 처리량

BERT Large 추론(일반화),
10ms 미만 지연시간에 대한 처리량



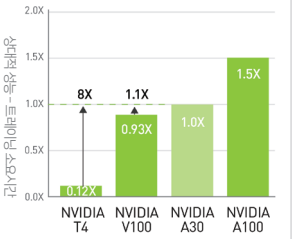
AI 추론 -
실시간 이미지 분류3에서
T4 대비 3배 이상 높은 처리량

RN50 v1.5 추론(일반화),
7ms 미만 지연시간에 대한 처리량

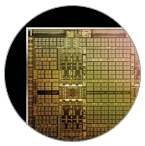


AI 트레이닝 -
V100 대비 최고 3배 높은
처리량과 T41 대비 최고 6배 높은 처리량

BERT Large 사전 트레이닝(일반화)

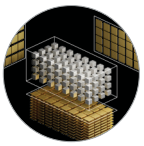


획기적 혁신



NVIDIA Ampere 아키텍처

MIG를 이용해 A30 GPU를 작은 인스턴스로 분할하거나, NVIDIA NVLink로 복수의 GPU를 연결해 보다 큰 규모의 워크로드 속도를 높이는 등, 최소 규모의 작업부터 최대 규모의 멀티-노드 워크로드에 이르기까지 A30는 각 상황에 맞는 가속 성능을 손쉽게 제공할 수 있습니다. 이렇게 융통성 있는 A30를 이용해, IT 관리자는 24시간 내내 메인스트림 서버로 데이터센터 각 GPU의 활용도를 최대치로 끌어올릴 수 있습니다.

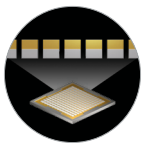


3세대 Tensor 코어

NVIDIA A30는 165 TFLOPS의 TF32 딥러닝 성능을 제공합니다. 이는 NVIDIA T4 Tensor 코어 GPU 대비, 20배 많은 AI 트레이닝 처리량이자 5배 이상의 추론 성능에 해당합니다. HPC 관련 10.3 TFLOPS의 성능이 제공되는데. 이는 NVIDIA V100 Tensor 코어 GPU 대비, 거의 30%나 높은 성능입니다.

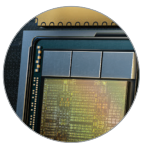
차세대 NVLINK

A30의 NVIDIA NVLink는 이전 세대 대비 처리량이 2배 늘어났습니다. NVLink Bridge를 통해 2개의 A30 PCIe GPU를 연결, 330 TFLOP의 딥러닝 성능을 제공할 수 있습니다.



MIG(Multi-Instance GPU)

각 A30 GPU는 자체 고대역폭 메모리, 캐쉬 및 컴퓨팅 코어를 갖추고 하드웨어 수준에서 완전히 분리된, 최고 4개의 GPU 인스턴스로 분할될 수 있습니다. MIG를 이용해 개발자는 모든 애플리케이션을 획기적으로 가속할 수 있습니다. 그리고 IT 관리자는 각 작업마다 알맞은 수준의 GPU 가속 성능을 제공, 활용을 최적화하고 모든 사용자와 애플리케이션으로 액세스를 확장할 수 있습니다.



HBM2

최고 24GB의 고대역폭 메모리(HBM2)로 A30는 메인스트림 서버에서 다양한 AI와 HPC 작업 처리에 최적인 933GB/s의 GPU 메모리 대역폭을 제공합니다.

구조적 희소성

AI 네트워크의 매개변수는 그 수가 수백만 개에서 수십억 개에 달하지만 이 매개변수가 정확한 예측에 모두 필요한 것은 아니므로, 일부는 정확성을 훼손시키지 않고 모델을 "희소"하게 만들기 위해 0으로 변환할 수 있습니다. A30의 Tensor 코어는 희소한 모델에 대해 최고 2배 높은 성능을 제공할 수 있습니다. 희소성 기능이 AI 추론에 보다 유용하지만, 모델 트레이닝의 성능 개선에도 사용할 있습니다.

엔터프라이즈용 엔드-투-엔드 솔루션

최신 데이터 센터의 핵심에 해당하는 NVIDIA Ampere 아키텍처에 기반한 NVIDIA A30 Tensor 코어 GPU는 NVIDIA 데이터 센터 플랫폼에서 빼놓을 수 없는 요소입니다. 딥러닝, HPC, 데이터분석을 위해 구축된 플랫폼은 모든 주요 딥러닝 프레임워크를 포함. 2,000개 이상의 애플리케이션을 가속화시키고 있습니다. 또한 AI와 데이터 분석 소프트웨어로 구성된 엔드-투-엔드, 클라우드-기반 스위트인 NVIDIA AI Enterprise는 VMware vSphere를 통해 하이퍼바이저-기반 가상 인프라에서 A30를 이용, 구동됨이 인증되었습니다. 덕분에 하이브리드 클라우드 환경에서 AI 워크로드를 관리 및 확장할 수 있습니다. 완전한 NVIDIA 플랫폼이 데이터센터에서 엣지에 이르기까지 어디서나 제공되며, 극적인 성능 개선과 비용 절감을 가능하게 합니다.

엔터프라이즈를 위해
최적화된 소프트웨어 및 서비스

NVIDIA CERTIFIED

모든 딥러닝 프레임워크

mxnet PYTORCH APACHE Spark TensorFlow

2,000개 이상의 GPU-가속 애플리케이션

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

NVIDIA A30 Tensor 코어 GPU에 대한 보다 상세한 정보는 다음 사이트를 참조해 주세요.
https://www.nvidia.com/ko-kr/data-center/products/a30-gpu

1. BERT-Large 사전 트레이닝 (9/10 에포크) 1 단계 및 (1/10 에포크) 2 단계, 1단계 시퀀스 길이=128, 2단계 시퀀스 길이=512, 데이터셋=real, NGC™ 컨테이너 =21.03,8x GPU: T4 (FP32, BS=8, 2) | V100 PCIE 16GB (FP32, BS=8, 2) | A30 (TF32, BS=8, 2) | A100 PCIE 40GB (TF32, BS=54, 8) | 표시된 배치 크기는 각각 1단계와 2단계에 해당
2. NVIDIA® TensorRT®, 정밀도=INT8, 시퀀스 길이=384, NGC 컨테이너 20.12, 지연시간<10ms, 데이터셋=Synthetic; 1x GPU: A100 PCIe 40GB (BS=8) | A30 (BS=4) | V100 SXM2 16GB (BS=1) | T4 (BS=1)
3. TensorRT, NGC 컨테이너 20.12, 지연시간<7ms, 데이터셋=Synthetic,; 1x GPU: T4 (BS=31, INT8) | V100 (BS=43, 혼합 정밀도) | A30 (BS=96, INT8) | A100 (BS=174, INT8)
4. 데이터셋: ReaxFF/C, FP64 | 4x GPU: T4, V100 PCIE 16GB, A30

NVIDIA