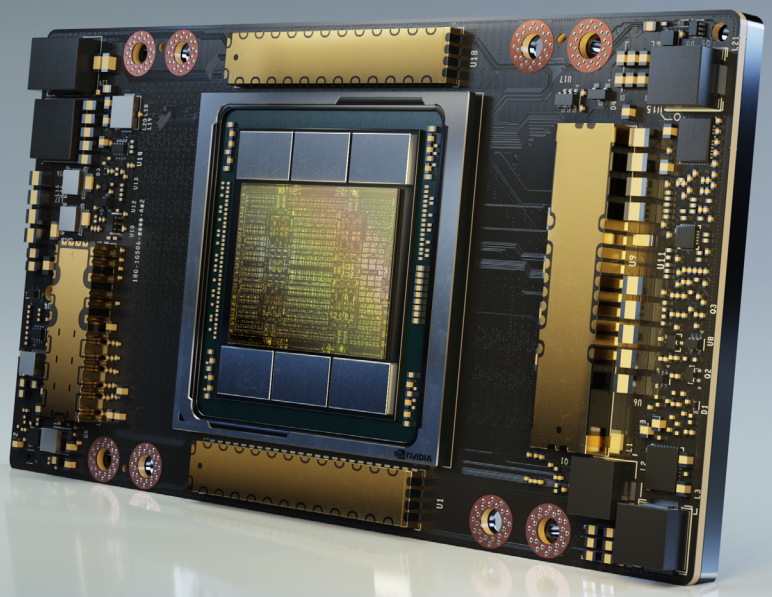
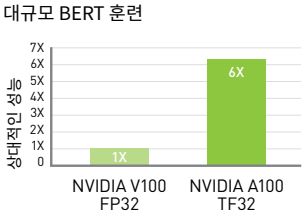
NVIDIA A100
TENSOR 코어 GPU



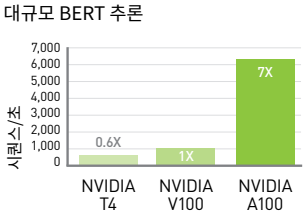
모든 스케일에 걸친 전례없는 가속

NVIDIA A100 Tensor 코어 GPU는 모든 스케일의 AI, 데이터 분석,HPC 분야에서 전례없는 가속을 제공하여 세상에서 가장 어려운 컴퓨팅 문제를 해결합니다. A100은 NVIDIA 데이터 센터 플랫폼의 엔진으로서 수천 개의 GPU로 수직 확장할 수 있습니다. 또는 새로운 멀티-인스턴스 GPU(MIG) 기술을 사용해 7개의 독립된 GPU 인스턴스로 나뉘어 다양한 크기의 작업을 가속할 수 있습니다. A100의 3세대 Tensor 코어 기술은 다양한 작업을 위해 여러 종류의 정밀도를 지원하므로 빠르게 인사이트를 얻고 시장에 출시할 수 있게 합니다.

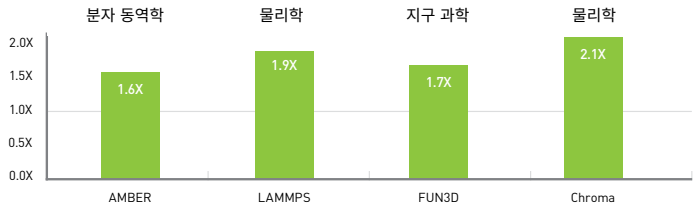
TF32를 사용한 기본 AI 훈련[1] 성능 6배 향상



멀티-인스턴스 GPU(MIG)를 사용하여 AI 추론[2] 성능 7배 향상



HPC 성능[3] 2배 이상 향상



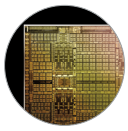
시스템 사양 (최고 성능)

	NVIDIA HGX™ 용 NVIDIA A100 SXM4	NVIDIA A100 PCIe GPU
GPU 아키텍쳐	NVIDIA Ampere	
배정밀도 성능	FP64: 9.7 TFLOPS FP64 Tensor 코어: 19.5 TFLOPS	
단정밀도 성능	FP32: 19.5 TFLOPS Tensor Float 32 (TF32): 156 TFLOPS | 312 TFLOPS*	
반정밀도 성능	312 TFLOPS | 624 TFLOPS*	
Bfloat16	312 TFLOPS | 624 TFLOPS*	
정수 성능	INT8: 624 TOPS | 1,248 TOPS* INT4: 1,248 TOPS | 2,496 TOPS*	
GPU 메모리	40 GB HBM2	
메모리 대역폭	1.6 TB/sec	
오류 정정 부호	Yes	
인터커넥트 인터페이스	PCIe Gen4: 64 GB/sec 3세대 NVIDIA® NVLink®: 600 GB/sec**	PCIe Gen4: 64 GB/sec 3세대 NVIDIA® NVLink®: 600 GB/sec**
폼 팩터	NVIDIA hGX™ A100에 있는 4/8 SXM GPU	PCIe
멀티-인스턴스 GPU (MIG)	최대7 GPU 인스턴스	
최대 전력 소비	400 W	250 W
Delivered Performance for Top Apps	100%	90%
쿨링 솔루션	패시브 (Passive)	
컴퓨팅 API	CUDA®, DirectCompute, OpenCL™, OpenACC®	

* 구조적 희소성 사용

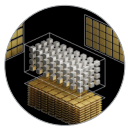
** HGX A100 서버 보드를 사용한 SXM GPU; 최대 2 GPU까지 NVLink Bridge를 사용한 PCIe GPU

놀라운 혁신



NVIDIA AMPERE 아키텍처

A100은 크고 작은 워크로드를 가속합니다. MIG를 사용해 A100 GPU를 작은 인스턴스로 나누거나 NVLink를 사용해 여러 GPU를 연결해 대규모 워크로드를 가속할 수 있습니다. A100은 작은 업무에서 멀티-노드 워크로드까지 다양한 사이즈의 가속 요구사항을 손쉽게 처리합니다. A100의 다재다능함 덕분에 IT 관리자는 데이터 센터에 있는 모든 GPU의 사용률을 항상 극대화할 수 있습니다.

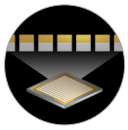


3세대 TENSOR 코어

A100은 312 TFLOPS의 딥러닝 성능을 제공합니다. 이는 NVIDIA Volta™ GPU보다 딥러닝 훈련에서 20배 높은 Tensor FLOPS이고 딥러닝 추론에서 20배 높은 Tensor TOPS입니다.

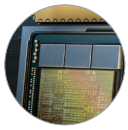
차세대 NVLINK

A100의 NVIDIA NVLink는 이전 세대보다 2배 높은 처리량을 제공합니다. NVIDIA NVSwitch™를 사용하면 최대 초당 600 기가바이트로 A100 GPU를 16개까지 연결하여 단일 서버에서 가능한 가장 높은 성능에 도달할 수 있습니다.
HGX A100 서버 보드를 통한 A100 SXM GPU와 최대 2개 GPU까지 NVLink Bridge로 연결한 PCIe GPU에서 사용할 수 있습니다.



멀티-인스턴스 GPU (MIG)

하나의 A100 GPU는 7개의 GPU 인스턴스로 나뉠 수 있습니다.
하드웨어 수준에서 완전히 독립적이며 각자의 고대역 메모리, 캐시, 컴퓨팅 코어를 가집니다.
MIG 덕분에 개발자들은 모든 애플리케이션을 가속할 수 있고 IT 관리자는 모든 작업에 적절한 크기의 GPU 가속을 제공할 수 있습니다.
동시에 사용률을 최적화하고 모든 사용자와 애플리케이션에 가속 서비스를 제공할 수 있습니다.



HBM2

40 기가바이트 고대역 메모리 (HBM2) 덕분에 A100은 1.6TB / sec의 향상된 대역폭을 제공합니다. 또한 95%의 높은 DRAM 사용 효율을 제공합니다. A100은 이전 세대보다 1.7배 높은 메모리 대역폭을 제공합니다.

구조적 희소성

AI는 수백 만에서 수십억 개에 이르는 파라미터를 가진 대규모 네트워크입니다. 정확한 예측에 모든 파라미터가 필요하지 않기 때문에 정확도를 희생하지 않으면서 일부 파라미터를 0으로 바꾸어 희소(sparse) 모델을 만들 수 있습니다.
A100의 Tensor 코어는 희소 모델의 성능을 2배로 높입니다. 이런 희소 기능은 AI 추론에 당연히 도움이 되고 모델 훈련의 성능도 향상시킬 수 있습니다.

NVIDIA A100 Tensor Core GPU는 딥러닝, HPC, 데이터 분석을 위한 NVIDIA 데이터 센터 플랫폼의 대표 제품입니다. 이 플랫폼은 700개 이상의 HPC 애플리케이션과 모든 주요 딥러닝 프레임워크를 가속화합니다. 데스크톱에서 서버, 클라우드 서비스에 이르기까지 어디에서나 사용할 수 있으며 성능을 극적으로 향상시키고 비용을 절감시켜 줍니다.

모든 딥러닝 프레임워크

Caffe2 · Microsoft Cognitive Toolkit · mxnet · PYTORCH · APACHE Spark · TensorFlow · theano

700개 이상의 GPU 가속 애플리케이션

AMBER · ANSYS Fluent · GAUSSIAN · GROMACS · LS-DYNA · NAMD · OpenFOAM · Simulia Abaqus · VASP · WRF

NVIDIA A100 Tensor 코어 **GPU**에 대한 더 자세한 내용은 **www.nvidia.com/a100** 을 참고하세요.

1 파이토치를 사용한 BERT 사전 훈련 처리량, 단계 1(2/3)과 단계 2(1/3) | 단계 1의 시퀀스 길이 = 128, 단계 2의 시퀀스 길이 = 512 | V100: NVIDIA DGX-1™ 서버, NVIDIA V100 텐서 Core GPU 8개, FP32 정밀도 | A100: NVIDIA DGX™ A100 서버, A100 8개, TF32 정밀도.

2 BERT 대규모 추론 | NVIDIA T4 Tensor 코어 GPU: NVIDIA TensorRT™ (TRT) 7.1, INT8 정밀도, 배치 크기 256 | V100: TRT 7.1, FP16 정밀도, 배치 크기 256 | A100, 1g.5gb 7 MIG 인스턴스; TRT 시제품, 배치 크기 94, 희소한 INT8 정밀도.

3 사용된 V100은 단일 V100 SXM2. 사용된 A100은 단일 A100 SXM4. PME-Cellulose 기반의 AMBER, Atomic Fluid LJ-2.5를 사용한 LAMMPS, dpw를 사용한 FUN3D, szscl21_24_128을 사용한 Chroma.

NVIDIA