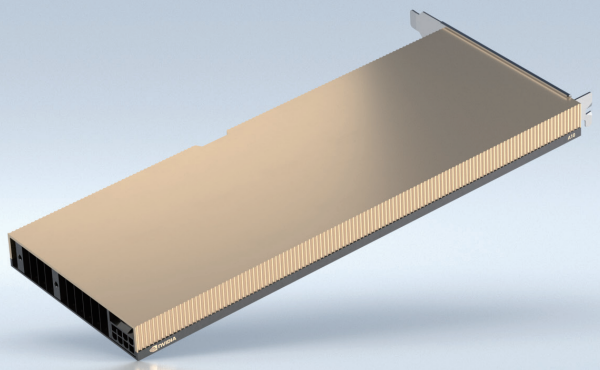
NVIDIA A10

메인스트림 엔터프라이즈 서버용 AI를 통해 가속화된 그래픽 및 영상



강력한 AI로 보강된 그래픽과 영상 애플리케이션

NVIDIA A10 Tensor 코어 GPU는 NVIDIA RTX Virtual Workstation(vWS) 소프트웨어와 결합해 AI 서비스가 적용된 메인스트림 그래픽과 영상을 메인스트림 엔터프라이즈 서버에 제공, 디자이너, 엔지니어, 아티스트, 과학자들의 난제 해결에 필요한 솔루션을 제공합니다. 최신 NVIDIA Ampere 아키텍처에 기반한 A10은 2세대 RT 코어, 3세대 Tensor 코어, 새로운 스트리밍 마이크로프로세서 및 24 기가바이트(GB) 용량의 GDDR6 메모리를 150W 전력 수준에서 제공합니다. 그 결과 다양한 그래픽, 렌더링, AI 및 컴퓨팅 성능이 제공됩니다. 전세계 어디서나 가상 워크스테이션을 통해 액세스, 데이터 센터에 노드를 렌더링하고 다양한 워크로드를 처리할 수 있는 A10은 싱글-와이드, FHFL PCIe 폼 팩터에서 최상의 성능을 제공하도록 구축되었습니다.

NVIDIA A10는 NVIDIA-Certified Systems™의 일부로 온-프레미스 데이터센터, 클라우드, 엣지 모두에서 지원됩니다. NVIDIA NGC™ 카탈로그의 각종 AI 프레임워크, CUDA-X™ 라이브러리, 230만 명 이상의 개발자들 및 1,800개 이상의 GPU 최적화 애플리케이션으로 구성된 다채로운 에코시스템을 기반으로 하는 NVIDIA A10을 활용해 기업의 가장 어려운 난제를 해결할 수 있습니다.

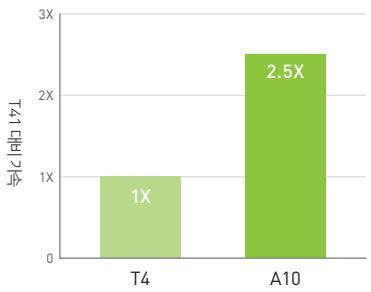
A10 사양

FP32	31.2 TF
TF 32 Tensor 코어	62.5 TF | 125 TF *
BFLOAT16 Tensor 코어	125 TF | 250 TF *
FP16 Tensor 코어	125 TF | 250 TF *
INT8 Tensor 코어	250 TOPS | 500 TOPS*
INT4 Tensor 코어	500 TOPS | 1000 TOPS*
RT 코어	72개
인코딩/디코딩	1개의 인코더, 2개의 디코더, (+AV1 디코딩)
GPU 메모리	24 GB GDDR6
GPU 메모리 대역폭	600 GB/s
인터커넥트	PCIe Gen4 64 GB/s
폼 팩터	단일 슬롯 풀 사이즈의 높이·길이(FHFL)
최대 열 설계 전력(TDP)	150W
vGPU 소프트웨어 지원	NVIDIA vPC/vApps, NVIDIA RTX™ vWS, NVIDIA 가상컴퓨트서버 (vCS)
HW RoT(Root of Trust)를 통한 안전하고 신중한 부팅	해당
NEBS 지원	레벨 3
전원 커넥터	PEX 8 핀

* sparsity 사용시

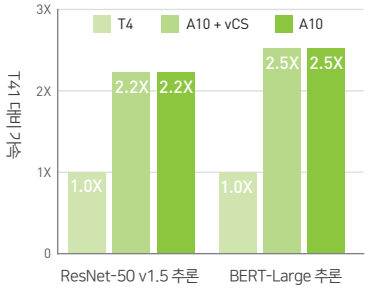
T4[1] 대비 최고 2.5배 빠른 가상 워크스테이션 성능을 제공하는 A10

SPECviewperf 2020 벤치마크



T4[2] 대비 최고 2.5배 많은 추론 성능을 제공하는 A10

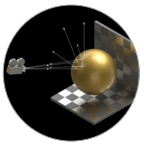
베어 메탈에 맞먹는 성능을 제공하는 NVIDIA vCS



NVIDIA Ampere 아키텍처 소개

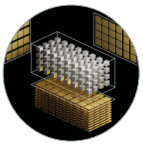
NVIDIA Ampere 아키텍처 CUDA 코어

단정밀도 부동 소수점(FP32) 연산에서 배속 처리 및 개선된 전력 효율 덕분에 복잡한 3D CAD 및 CAE 등 그래픽과 컴퓨팅 워크플로우 성능이 크게 향상됩니다.



2세대 RT 코어

이전 세대 대비 최고 2배의 처리량, 고급 음영 처리 또는 노이즈 제거와 레이 트레이싱을 동시 실행하는 기능을 갖춘 2세대 RT 코어는 영화 콘텐츠의 사실적 렌더링, 건축 디자인 평가, 제품 디자인의 가상 프로토타입 제작과 같은 작업 속도를 크게 높여줍니다. RT 코어는 또한 레이 트레이싱 모션 블러의 렌더링 속도를 높여줘, 시각적 정확도가 개선된 작업 결과를 더 빠르게 얻을 수 있습니다.



3세대 Tensor 코어

TF32(Tensor Float 32) 정밀도가 이전 세대 대비 최고 5배의 트레이닝 처리량을 제공, 코드 변경 없이도 AI와 데이터 사이언스 모델 트레이닝을 가속해 줍니다. 구조적 희소성을 위한 하드웨어 지원은 추론 처리량을 최고 2배로 높여 줍니다.
Tensor 코어는 또한 딥러닝 슈퍼샘플링(DLSS), AI 노이즈 제거, 특정 애플리케이션에 대한 편집 개선 등의 기능으로 그래픽에서도 AI를 지원합니다.

24 GB GDDR6

초고속 GDDR6 메모리가 렌더링, 데이터 사이언스, 엔지니어링 시뮬레이션 등의 GPU 메모리 집약 작업에 필요한 600 GB/s 대역폭을 제공합니다.

PCI Express Gen 4

PCIe Gen 3의 대역폭을 2배로 증가시켜 AI, 데이터 사이언스, 3D 디자인 등의데이터 집약 작업을 위한 CPU 메모리의 데이터 전송 속도를 개선합니다. 또한 보다 빠른 PCIe 성능으로 GPU DMA(Direct Memory Access) 전송을 가속해, NVIDIA GPUDirect® for Video 지원 디바이스와 GPU 간 영상 데이터 I/O 통신 속도를 높여줍니다. 덕분에 라이브 방송을 위한 강력한 솔루션이 제공됩니다. A10는 이전 버전인 PCI Express Gen 3과 호환되어 배포의 유연성도 보장됩니다.

데이터 센터 효율과 보안

단일 슬롯, FHFL의 전력 효율적 디자인을 갖춘 NVIDIA A10은 전세계 OEM이 생산하는 광범위한 서버에 적합하도록 제작되었습니다. 하드웨어 RoT(Root of Trust) 기술을 이용한 안전하고 신중한 부팅은 펌웨어의 물리적 변경(tempering)과 손상(corruption)을 예방합니다.

NVIDIA A10 Tensor 코어 GPU는 AI가 적용된 메인스트림 그래픽과 영상에 이상적입니다. 2세대 RT 코어와 3세대 Tensor 코어가 메인스트림 서버를 위한 150W TDP 디자인에서 강력한 AI 기능으로 그래픽과 영상 애플리케이션을 보강해 줍니다.
NVIDIA A10는 또한 NVIDIA 가상 GPU(vGPU) 소프트웨어를 이용, 그래픽이 풍부한 VDI에서 고성능 가상 워크스테이션, AI에 이르기까지 다양한 데이터센터 워크로드를 가속시킵니다. 이는 리소스 수요 충족을 위해 확장 가능할 뿐 아니라 관리가 쉽고, 보안성이 높고, 유연한 인프라에서 이루어집니다.

모든 딥러닝 프레임워크

mxnet
PYTORCH
Spark
TensorFlow

전문가용 애플리케이션을 위한 RTX

AUTODESK REVIT
creo
CATIA
Rhinoceros
design, model, present, analyze, realize...
SOLIDWORKS
SIEMENS

NVIDIA A10 Tensor 코어 GPU에 대한 보다 상세한 정보는 다음 사이트를 참조해 주세요.
https://www.nvidia.com/ko-kr/data-center/products/a10-gpu

1. 다음 사양의 서버에서 테스트 실시: 2x 제온 골드 프로세서 6154 3.0GHz (3.7GHz 터보), NVIDIA RTX vWS 소프트웨어, VMware ESXi 7 U2, 호스트/게스트 드라이버 461.33. | SPECviewperf 2020 Subtest 및 HD 3dsmax-07 composite.
2. BERT Large 추론 NVIDIA TensorRT 7.2, 시퀀스 길이=128, 배치 규모=128; NGC 컨테이너: 21.02-py3 | ResNet-50 v1.5: NVIDIA TensorRT 7.2, INT8 정밀도 배치 규모=128 NGC 컨테이너: 20.12-py3 | NVIDIA A10, vCS 소프트웨어, VMware ESXi 7 U2, 호스트/게스트 드라이버 461.33

NVIDIA