NVIDIA

VISUAL & AI COMPUTING SOLUTION

BayNex

16X Tesla V100 SXM2 32GB

NVIDIA® DGX-2™

16x Tesla V100 SXM2 32GB

81,920 CUDA cores / 10,240 Tensor Cores

FP16 : 1,920 TFLOPS / FP32 : 240 TFLOPS / FP64 : 120 TFLOPS

2x Xeon CPU / 300GB/s NVLink 2.0 / 10U – 10000W

세계 최초 맞춤형 인공지능(AI) 슈퍼컴퓨터 포트폴리오

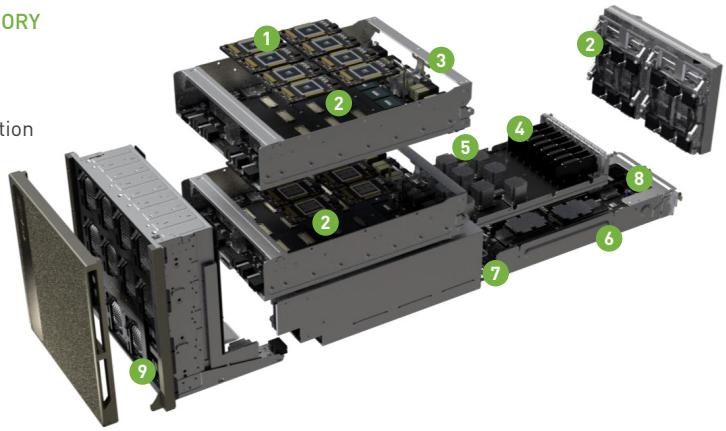
딥 러닝 성능을 5배로 높이기 위해 완전히 상호 연결된 16개의 GPU를 결합하는 최초의 2 페타플롭스 시스템인 NVIDIA DGX-2로 AI 속도 및 확장에 대한 장벽을 뛰어넘으십시오. 이 제품은 NVIDIA® DGX™ 소프트웨어와 NVIDIA NVSwitch 기술을 기반으로 하는 확장 가능한 아키텍처로 구동되므로 세계에서 가장 복잡한 AI 과제를 수행할 수 있습니다.

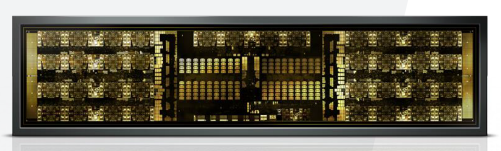
혁신적인 AI 네트워크 패브릭

DGX-2를 이용하면 모델 복잡성과 크기는 더 이상 기존 아키텍처의 한계에 의해 제한되지 않습니다. 이제 여러분은 NVIDIA NVSwitch 네트워크 패브릭을 통해 모델 병렬 트레이닝을 활용하실 수 있습니다. 이는 2.5TB/s의 GPU 간 대역폭을 갖춘 세계 최초의 2 페타플롭스 GPU 가속기를 뒷받침하는 혁신적인 기술로 이전 세대에 비해 12배 향상된 성능과 5배 빠른 솔루션 도출 시간을 제공합니다.

DGX-2의 강력한 요소들을 모두 확인하세요.

1. NVIDIA TESLA V100 32GB, SXM2
2. 16 TOTAL GPUS FOR BOTH BOARDS, 512GB TOTAL HBM2 MEMORY
   각각의 GPU 보드에 탑재된 총 16개의 NVIDIA Tesla V100.
3. 12 TOTAL NVSWITCHES
   HSI(High Speed Interconnect), 2.4 TB/초의 바이섹션 대역폭(bisection bandwidth).
4. 8 EDR INFINIBAND/100 GbE ETHERNET
   1600 Gb/초의 양방향 대역폭(Bi-directional Bandwidth)과 저지연.
5. PCIE SWITCH COMPLEX
6. TWO INTEL XEON PLATINUM CPUS
7. 1.5 TB SYSTEM MEMORY
8. DUAL 10/25 GbE ETHERNET
9. 30 TB NVME SSDS INTERNAL STORAGE





16개의 GPU 연결이 지원되는
첫 on-node 스위치 아키텍처

NVIDIA NVSwitch

멀티 GPU 프로세싱이 DGX-1 대비 개선된
오직 DGX-2에서만 1 NODE 16 GPUs의 NVSwitch가 지원되는 모델입니다.

세계 최초 딥 러닝 용 슈퍼컴퓨터

NVIDIA® DGX-1™

8x Tesla V100 SXM2 16GB

40,960 CUDA cores / 5,120 Tensor Cores

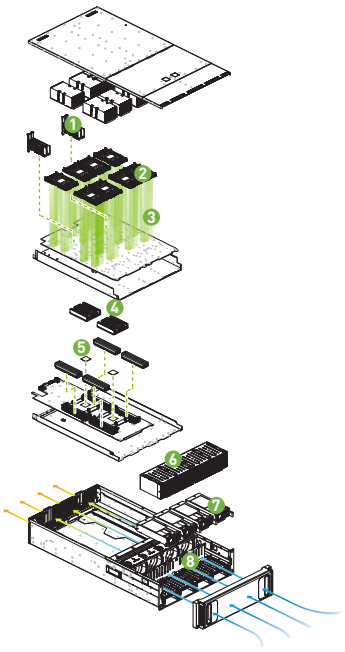
FP16 : 960 TFLOPS / FP32 : 120 TFLOPS / FP64 : 60 TFLOPS

2x Xeon CPU / 300GB/s NVLink 2.0 / 3U – 3200W

세계 최초 맞춤형 인공지능(AI) 슈퍼컴퓨터 포트폴리오

딥 러닝 및 분석 요구 사항에서 영감을 받은 NVIDIA® DGX™ 시스템은 새롭고 혁신적인 NVIDIA Volta GPU 플랫폼을 기반으로 구축되었습니다. NVIDIA® DGX™ 시스템은 데이터 과학자들에게 책상과 데이터센터, 클라우드를 넘나들며 AI를 탐구할 수 있는 가장 강력한 도구를 제공하도록 설계되었습니다.

Hardware



1. NETWORK INTERCONNECT
4X InfiniBand 100 Gbps EDR
2X 10 GbE

2. GPUs
8X NVIDIA Tesla® V100 32 GB/GPU
40,960 Total NVIDIA CUDA® Cores
5,120 Tensor Cores

3. GPU INTERCONNECT
NVIDIA NVLink™
Hybrid Cube Mesh

4. SYSTEM MEMORY
512 GB DDR4 LRDIMM

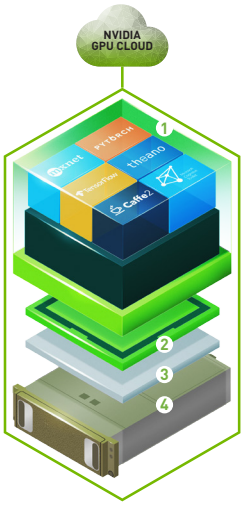
5. CPUs
2X 20-Core Intel® Xeon®
E5-2698 v4 2.2 GHz

6. STREAMING CACHE
4X 1.92 TB SSDs RAID 0

7. POWER
4X 1600 W PSUs
(3200 W TDP)

8. COOLING
Efficient Front-to-Back
Airflow

Software



1. DGX DEEP LEARNING STACK
DEEP LEARNING FRAMEWORKS
Caffe Caffe2 mxnet
PYTORCH TensorFlow theano torch
DEEP LEARNING USER SOFTWARE
NVIDIA DIGITS™
NVIDIA DEEP LEARNING SDK
CONTAINERIZATION TOOL
NVIDIA Docker
Docker

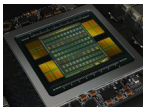
2. GPU DRIVER
NVIDIA Driver

3. SYSTEM
Host OS

4. NVIDIA DGX-1

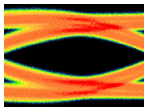
NVIDIA® Tesla® V100 GPU 8개와 함께하세요.

Built on the latest NVIDIA Volta GPU Architecture

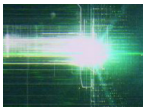


The power of 800 x86 CPUs

3X deep learning training speed-up



10X speed-up with NVLink™ vs PCIe



30% faster performance with DGX software stack

생산성을 극대화 할 수 있는 플랫폼

Get started in as little as 2 hours with NVIDIA® DGX™

Deploy Quickly and Simply

Plug-and-play setup that takes you from Power-on to deep learning in minutes

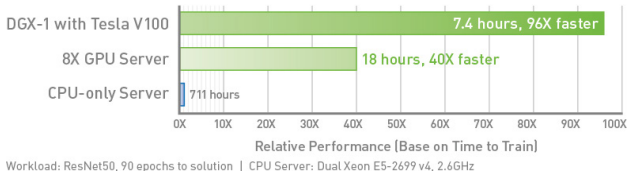
NVIDIA GPU CLOUD and Support

Access to NVIDIA's vast deep learning Knowledge, expertise and the latest Software updates

빠른 트레이닝과 새로운 혁신

Unparalleled Deep Learning Training Performance

NVIDIA DGX-1 Delivers 96X Faster Training



Workload: ResNet50, 90 epochs to solution | CPU Server: Dual Xeon E5-2699 v4, 2.6GHz

	Tesla V100 탑재	Tesla P100 탑재
TFLOPS (GPU FP16)	960	170
GPU 메모리	128GB 총 시스템	
CPU	Dual 20-Core Intel Xeon E5-2698 v4 2.2 GHz	
NVIDIA® CUDA® 코어	40,960	28,672
NVIDIA 텐서 코어 (V100 기준 시스템)	5,120	N/A
NVLink 대 PCIe의 가속 성능 비교	10배	5배
딥 러닝 트레이닝 가속	3배	1배
최대 소요 전력	3,200 W	
시스템 메모리	512 GB 2, 133 MHz DDR4 LRDIMM	
스토리지	4X 1.92 TB SSD RAID 0	
네트워크	Dual 10GbE, 최대 4 IB EDR	
소프트웨어	Ubuntu Linux Host OS 자세한 내용은 소프트웨어 스택 참조	
시스템 무게	134 lbs	
시스템 크기	866 D x 444 W x 131 H(mm)	
패키지 크기	1,180 D x 730 W x 284 H(mm)	
작동 온도	10 ~ 35 ℃	

세계 최초 딥 러닝 용 워크스테이션

NVIDIA® DGX Station™

4x Tesla V100 PCIe 16GB

20,480 CUDA cores / 2,560 Tensor Cores

FP16 : 480 TFLOPS / FP32 : 60 TFLOPS / FP64 : 30 TFLOPS

Xeon CPU / NVLink 2.0 / 1500W / Water Cooled

최첨단 AI 개발을 위한 개인용 슈퍼컴퓨터

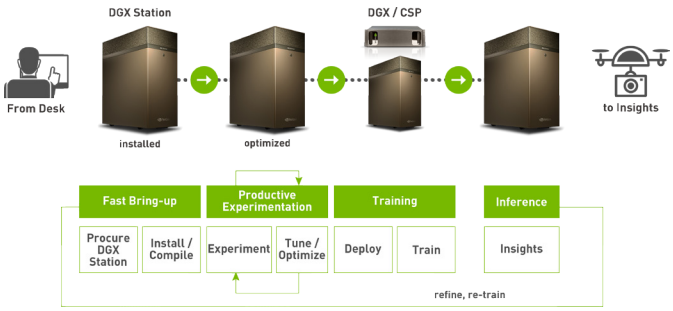
데이터 과학팀은 강력한 딥 러닝과 분석을 통해 통찰력을 얻고 더 빠르게 혁신하기 위해 컴퓨팅 성능을 활용합니다. 지금까지 AI 슈퍼컴퓨팅은 데이터 센터에 국한되어 대규모로 트레이닝 하기 전에 인공신경망을 테스트하기 위해 필요한 실험을 제한적으로 실행했습니다. 그러나 여기 AI 슈퍼컴퓨팅 성능을 아주 가까운 곳으로 가져오는 동시에 딥 러닝을 실험 할 수 있는 능력을 제공하는 솔루션이 있습니다.

AI의 신기원

이제는 400 CPU 컴퓨팅 용량의 워크스테이션을 간편하게 설치해 1/20 이하의 전력으로 사용 할 수 있습니다. 사무용으로 설계된 NVIDIA® DGX Station™은 경이로운 딥 러닝과 분석 성능을 제공하여 소음은 다른 워크스테이션의 1/10 수준에 불과합니다. 데이터 과학자들과 AI 연구자들은 최적화된 딥 러닝 소프트웨어에 액세스 할 수 있고 일반적인 분석 소프트웨어를 실행하는 워크스테이션을 통해 생산성을 즉시 향상 시킬 수 있습니다.

딥 러닝으로 더 빠르게 시작

DGX Station은 자체 딥 러닝 플랫폼 구축의 한계를 뛰어넘는 솔루션 입니다. 하드웨어와 소프트웨어의 조달, 통합 및 테스트를 위해 1개월 이상을 소비하고도 프레임워크, 라이브러리, 그리고 드라이버를 최적화하기 위해 추가적인 전문지식과 노력이 필요 할 수 있습니다. 트레이닝과 실험에 투자 할 수 있는 소중한 시간과 비용을 시스템 통합과 소프트웨어 엔지니어링에 써야 하는 것 입니다. NVIDIA® DGX Station™은 단 하루 만에 인공신경망을 트레이닝 할 수 있는 간소화된 플러그인과 구동 경험으로 AI 이니셔티브를 시작하도록 설계하였습니다.



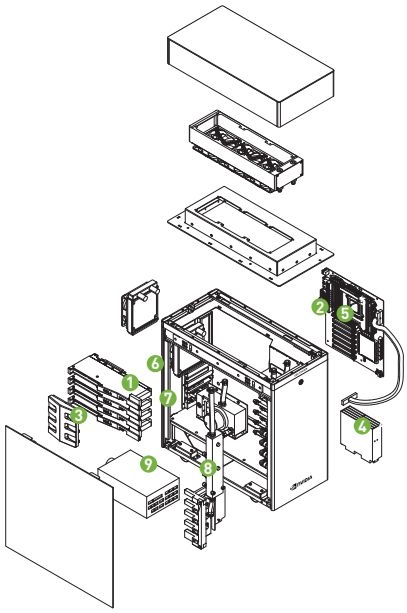
480 TFLOPS

Water-cooled performance–the only workstation built on 4 Tesla V100's

3X
the performance of today's fastest GPU workstations

30%
faster training over non-DGX stack solutions

5X
increase in I/O performance with 4-way NVLink vs. PCIe-connected GPU's



1. GPUs
4X NVIDIA Tesla® V100 16 GB/GPU
480 TFLOPS (FP16)
20,480 Total NVIDIA CUDA® Cores
2,560 Tensor Cores

2. SYSTEM MEMORY
256 GB LRDIMM DDR4

3. GPU INTERCONNECT
NVIDIA NVLink™,
Fully Connected 4-Way

4. STORAGE
Data: 3 x 1.92 TB SSD RAID 0
OS: 1 x 1.92 TB SSD

5. CPU
Intel Xeon E5-2698 v4
2.2 GHz 20-Core

6. NETWORKING
2X 10 GbE

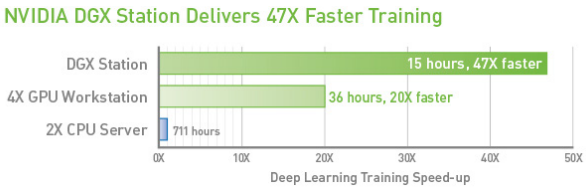
7. DISPLAYS
3X DisplayPort,
4K Resolution

8. COOLING
Water-Cooled

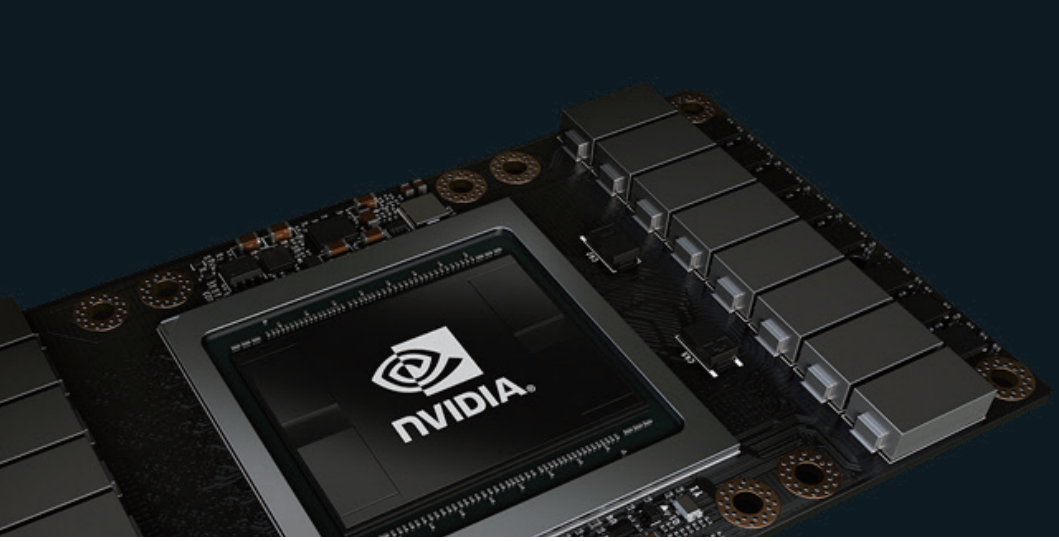
9. POWER
1500 W

빠른 트레이닝과 새로운 혁신

Unparalleled Deep Learning Training Performance



GPUs	4X Tesla V100
TFLOPS(GPU FP16)	480
GPU 메모리	64GB 총 시스템
NVIDIA 텐서 코어	2,560
NVIDIA® CUDA® 코어	20,480
CPU	Intel Xeon E5-2698 v4 2.2GHz (20-Core)
시스템 메모리	256GB LRDIMM DDR4
스토리지	데이터 : 3X 1.92TB SSD RAID 0 OS : 1X 1.92TB SSD
네트워크	이중 10Gb LAN
디스플레이	3X DisplayPort, 4K 해상도
음향	< 35dB
시스템 무게	88lbs/40kg
시스템 크기	518 D x 256 W x 639 H(mm)
최대 소요 전력	1,500 W
작동 온도	10 ~ 30 ℃
소프트웨어	Ubuntu Desktop Linux OS DGX 권장 GPU Driver CUDA Toolkit



가장 빠르고 생산적인 GPU
for Deep Learning & HPC

NVIDIA® Tesla®

서버 용 TESLA 데이터센터 GPU

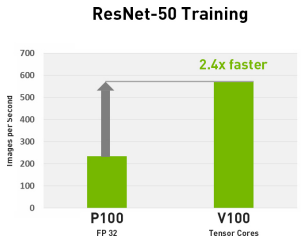
NVIDIA® Tesla® GPU를 사용하여 높은 사양을 요구하는 HPC와 하이퍼스케일 데이터센터 워크로드를 가속화하십시오. 데이터 과학자와 연구자는 이제 에너지 탐사에서 딥 러닝에 이르는 다양한 응용 분야에서 기존의 CPU가 지원했던 수준보다 훨씬 더 빠르게 페타바이트급 데이터 주문을 파싱 및 이전보다 훨씬 빠른 속도로 더욱 큰 규모의 시뮬레이션을 실행 가능합니다.

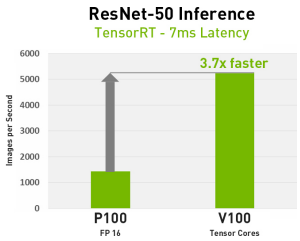
Tesla V100 SXM2
NVLink 지원하는 Socket 타입의 GPGPU

Tesla V100 PCIe
PCIe 타입의 GPGPU, Deep Learning, HPC 용

NVIDIA® CUDA® Cores	5,120	
Double-Precision Performance	7.8 TFLOPS	7 TFLOPS
Single-Precision Performance	15.7 TFLOPS	14 TFLOPS
Tensor Performance	120 TFLOPS	112 TFLOPS
Interconnect Bandwidth	300 GB/sec NVLink	32 GB/sec
GPU Memory	16 GB HBM2	
Memory Bandwidth	900 GB/sec	

VOLTA : A GIANT LEAP FOR DEEP LEARNING

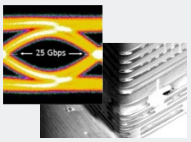




Volta Architecture

Most Productive GPU

Improved NVLink & HBM2

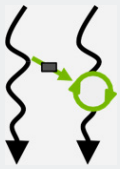


Efficient Bandwidth

Volta MPS

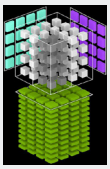
Inference Utilization

Improved SIMT Model



New Algorithms

Tensor Core



120 Programmable TFLOPS Deep Learning

GPU 성능 비교(P100 vs M40 vs K40)

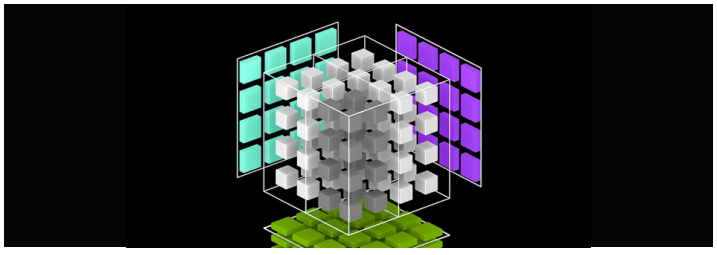
	P100	M40	K40
Double-Precision TFLOPS	5.3	0.2	1.4
Single-Precision TFLOPS	10.6	7.0	4.3
Half-Precision TFLOPS	21.2	NA	NA
GPU Memory	16 GB	12GB, 24GB	12GB
Memory Bandwidth(GB/s)	720	288	288

P100 vs V100 성능 비교

	P100	V100	Ratio
Training acceleration	10.6 TFLOPS	120 TFLOPS	12x
Inference acceleration	21.2 TFLOPS	120 TFLOPS	6x
FP64/FP32	5.3/10.6 TFLOPS	7.8/15.7 TFLOPS	1.5x
HBM2 Bandwidth	720 GB/s	900 GB/s	1.2x
NVLink Bandwidth	160 GB/s	300 GB/s	1.9x
L2 Cache	4 MB	6 MB	1.5x
L1 Caches	1.3 MB	10 MB	7.7x

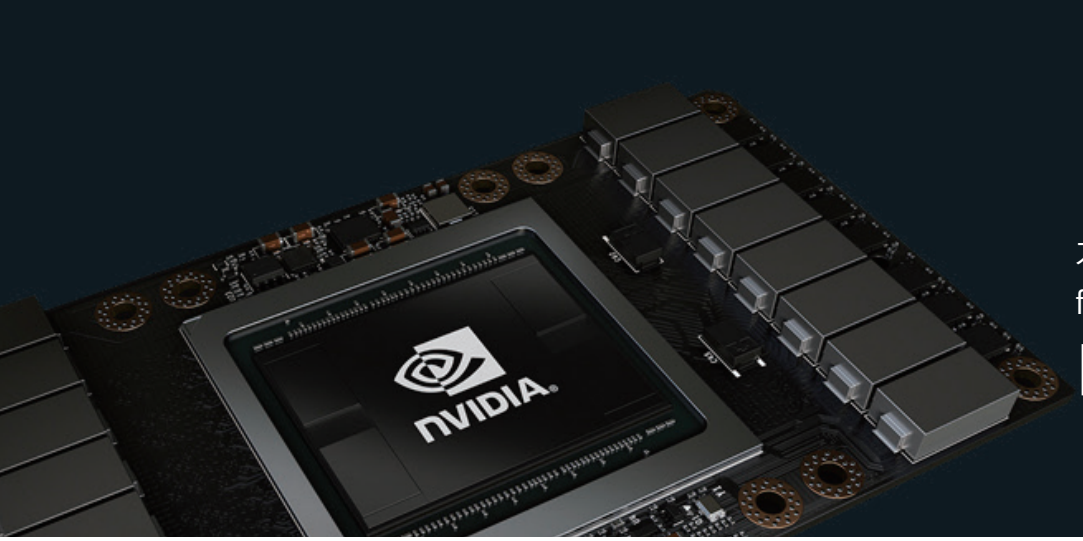
VOLTA TENSOR CORE

딥 러닝에 가장 최적화된 프레임워크, 라이브러리 지원



NVIDIA cuDNN, cuBLAS, TensorRT

오직 NVIDIA GPU에서만!



가장 빠르고 생산적인 GPU
for Deep Learning & HPC

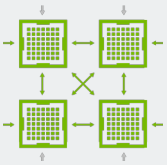
NVIDIA® Tesla®

NVIDIA TESLA®	 Tesla V100 for SXM2 NVLink 2.0 지원 딥 러닝 트레이닝, HPC 용	 Tesla V100 for PCIe NVLink 2.0 지원 딥 러닝 트레이닝, HPC 용	 Tesla P100 for PCIe (12GB/16GB) 딥 러닝 트레이닝, HPC 용	 Tesla P4 딥 러닝 트레이닝, 추론 용	 Tesla P40 딥 러닝 트레이닝, 추론 용
NVIDIA® CUDA® Cores	5,120	5,120	3584	2560	3840
Double-Precision Performance	7.8 TFLOPS	7 TFLOPS	4.7 TFLOPS	-	-
Single-Precision Performance	15.7 TFLOPS	14 TFLOPS	9.3 TFLOPS	5.5 TFLOPS	12 TFLOPS
Half-Precision Performance	120 TFLOPS	112 TFLOPS	18.7 TFLOPS	11 TFLOPS	24 TFLOPS
GPU Memory	16 GB	16 GB	16GB or 12GB	8 GB	24 GB
Memory Bandwidth	900 GB/sec	900 GB/sec	720 GB/s or 540 GB/s	192 GB/s	346 GB/s

Pascal Architecture

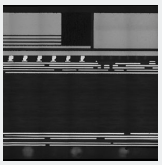
Highest Compute Performance

NVLink



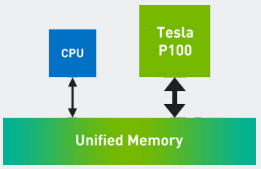
GPU Interconnect for Maximum Scalability

HBM2 Stacked Memory



Unifying Compute & Memory in Single Package

Page Migration Engine



Simple Parallel Programming with 512 TB of Virtual Memory

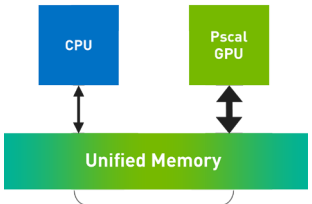
CUDA

CUDA는 GPU 가속 어플리케이션을 위한 가장 강력한 소프트웨어 플랫폼입니다. 빠른 GPU 가속 라이브러리, 유연한 스레드 관리를 위한 새로운 프로그래밍 모델, 컴파일러 및 개발자 도구 개선을 포함합니다. CUDA를 사용하면 NVIDIA GPU 응용 프로그램의 속도와 확장 성을 향상시킬 수 있습니다.

CUDA 9은 Volta GPU에서 AI 및 HPC 애플리케이션을 가속화하는 새로운 알고리즘과 최적화를 제공합니다.

- NVIDIA® CUDA®의 새로운 기능을 사용하여 심층적인 학습을 위한 이미지 확대 알고리즘 개발
- cuBLAS의 새로운 API를 사용하여 Volta Tensor 코어에서 일괄 처리 된 신경망 처리 및 시퀀스 모델링 작업 실행
- cuFFT에서 새로운 딥 러닝 모델을 가진 다중 GPU 시스템에서 보다 큰 2D 및 3D FFT 문제를 이전보다 효율적으로 해결 가능
- 새로운 코어 최적화 지원으로 최대 12배 빠른 커널 실행

CUDA 8 이상

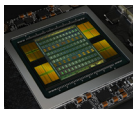


GPU Memory Size 이상 할당 가능

INTRODUCING CUDA 9

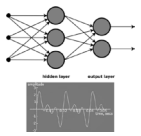
BUILT FOR VOLTA

- Tesla V100
- New GPU Architecture
- Tensor Cores
- NVLink
- Independent Thread Scheduling



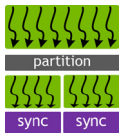
FASTER LIBRARIES

- cuBLAS for Deep Learning
- NPP for Image Processing
- cuFFT for Signal Processing



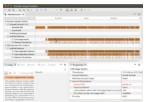
COOPERATIVE THREAD GROUPS

- Flexible Thread Groups
- Efficient Parallel Algorithms
- Synchronize Across Thread
- Blocks in a Single GPU or Multi-GPUs

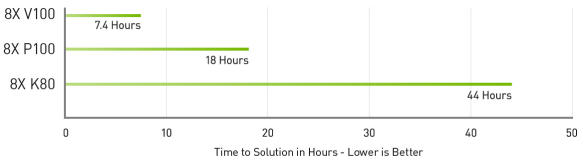


DEVELOPER TOOLS & PLATFORM UPDATES

- Faster Compile Times
- Unified Memory Profiling
- NVLink Visualization
- New OS and Compiler Support



Deep Learning Training in One Workday



Server Config: Dual Xeon E5-2699 v4, 2.6GHz | 8x Tesla K80, Tesla P100 or Tesla V100 | V100 performance measured on pre-production hardware. | ResNet-50 Training on Microsoft Cognitive Toolkit for 90 Epochs with 1.28M ImageNet dataset

Introducing
Quadro GV100
Tesla V100 32GB

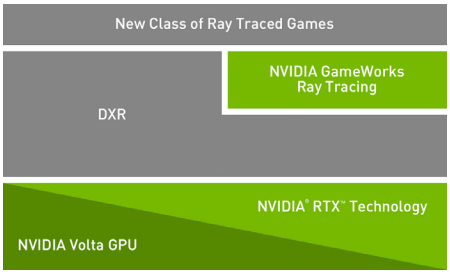
NVIDIA QUADRO	 GV 100
CUDA® Cores	5120
Display Connectors	DP 1.4(4), DVI-D DL(1), Stereo
FP 32	14.8 TFLOPS
Features	32K Texture & Render Processing / Quadro MOSAIC, NVLink, SLI
Power	250W
Memory	32GB HBM2

NVIDIA TESLA	 V 100
CUDA® Cores	5120
FP 32	14.8 TFLOPS
FP 16	112 TFLOPS
Features	NVLink 2.0, CUDA, NVIDIA GPU Cloud
Power	250W
Memory	32GB HBM2

NVIDIA RTX를 소개합니다.

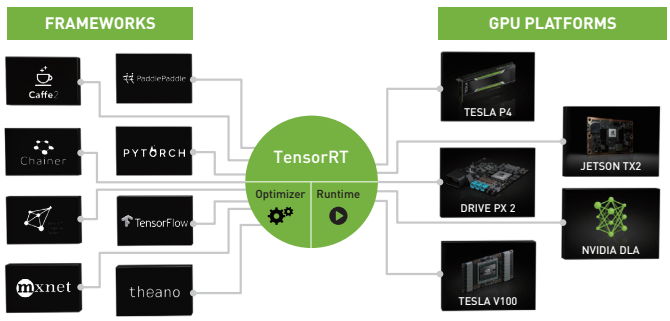
수십 년간의 알고리즘 개발과 GPU의 발전은 컴퓨터 그래픽을 한차원 더 높은 레벨로 이끌었습니다.

지난 10년간의 그래픽 알고리즘과 GPU의 Volta 아키텍처로 인해 Ray Tracing 기술이 매우 최적화 되었습니다.
Partnered with Microsoft와의 협업으로 DirectX에서 Raytracing을 API로 사용할 수 있게 되었습니다.

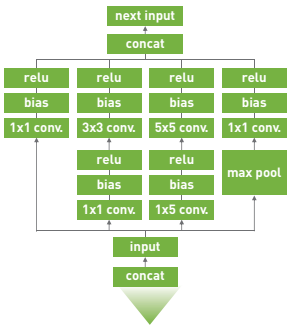


NVIDIA TensorRT

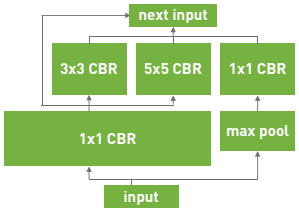
추론에 가장 최적화된 NVIDIA Software 입니다. 트레이닝 된 모델을 가볍게 만들고 실제 환경에서 트레이닝한 모델과 비슷한 성능을 낼 수 있게 도와줍니다.



최적화되지 않은 네트워크



TensorRT로 최적화된 네트워크

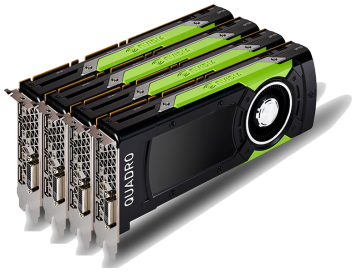


세계에서 가장 강력한 비주얼 컴퓨팅 플랫폼

NVIDIA® Quadro®

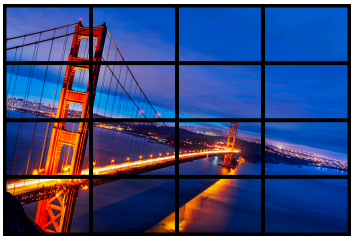
Quadro Sync
16개의 멀티 디스플레이 지원 가능

Quadro Sync는 전문적인 사용자에게 전례 없는 전력 효율과 성능, 그리고 기능들을 제공하는 세계에서 가장 진보된 전문 그래픽 솔루션입니다.



Quadro GP100*4

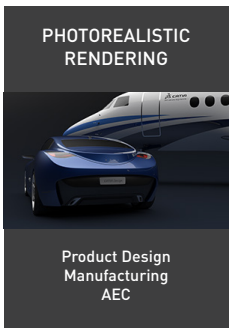
Quadro Sync II

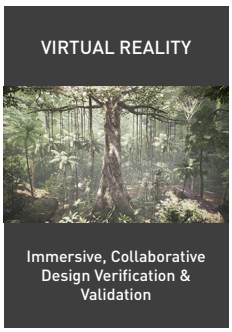


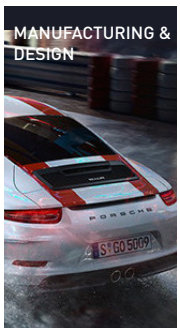
16 Displays

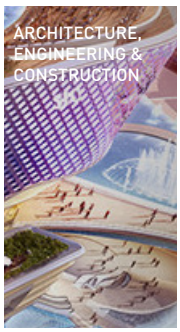
3D Rendering, 시뮬레이션

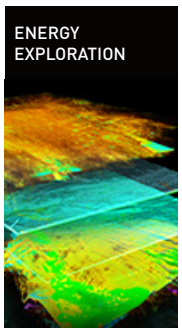
글로벌 시장에서 경쟁력을 관리하는 것은 모든 규모의 제조업체에게 중요한 과제입니다. 혁신적인 제품과 파격적인 아키텍처를 만드는 경우 모두 NVIDIA의 전문 그래픽은 워크플로를 최적화하고 디자인 애플리케이션의 성능을 강화하는 데 유용합니다.











NVIDIA® Quadro®는 VR 및 VR 디자인을 지원합니다.

VR에서의 탁월한 시각화와 협업

가상 현실(VR)의 꿈이 NVIDIA 하드웨어 및 소프트웨어 기술을 통해 실현되고 있습니다.
그리고 전문가의 작업 과정에서 그 혁신이 일어나고 있습니다.

NVIDIA VRREADY

NVIDIA VRWorks™
A Giant Step towards Full Presence

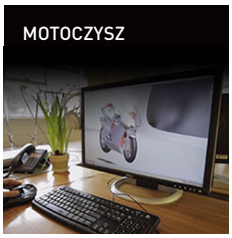
	SIGHT	SOUND	TOUCH	BEHAVIOR
API	VRWORKS GRAPHICS	VRWORKS AUDIO	PHYSX	PHYSX
ENGINE	MULTI-PROJECTION	OPTIX		

세계에서 가장 강력한 비주얼 컴퓨팅 플랫폼

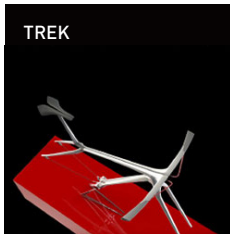
NVIDIA® Quadro®

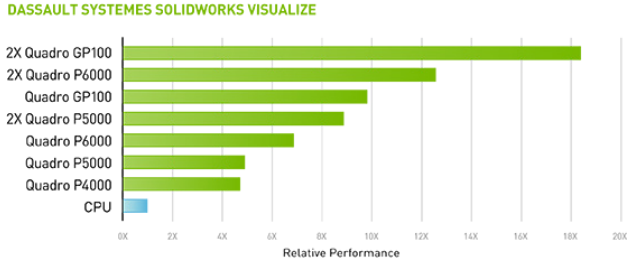
NVIDIA QUADRO	 P1000	 P2000	 P4000	 P5000	 P6000	 GP100
CUDA® Cores	640	1024	1792	2560	3840	3584
Display Connectors	mDP(4)	DP 1.4 (4)	DP 1.4 (4)	DP 1.4 (4) +DVI-D DL(1)	DP 1.4 (4) +DVI-D (1)	DP 1.4 (4) +DVI-D (1)
FP 32	1.894 TFLOPS	3.0 TFLOPS	5.3 TFLOPS	8.9 TFLOPS	12 TFLOPS	10.3 TFLOPS
Features	5K Resolution	Mosaic, Iray & Metal Ray	Quadro Sync II, Mosaic, Iray & Metal Ray	Quadro Sync II, Mosaic, Iray & Metal Ray	Quadro Sync II, Mosaic, Iray & Metal Ray	Quadro Sync II, Mosaic, Iray & Metal Ray, NVLink Supported
Power	47W	75W	105W	180W	250W	235W
Memory	4GB GDDR5X	5GB GDDR5X	8GB GDDR5X	16GB GDDR5X	24GB GDDR5X	16GB HBM2

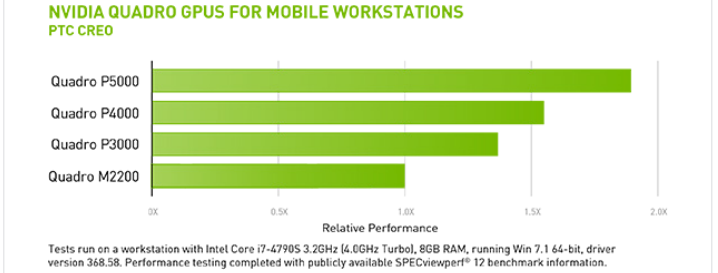
성공 사례



NVIDIA® mental ray®





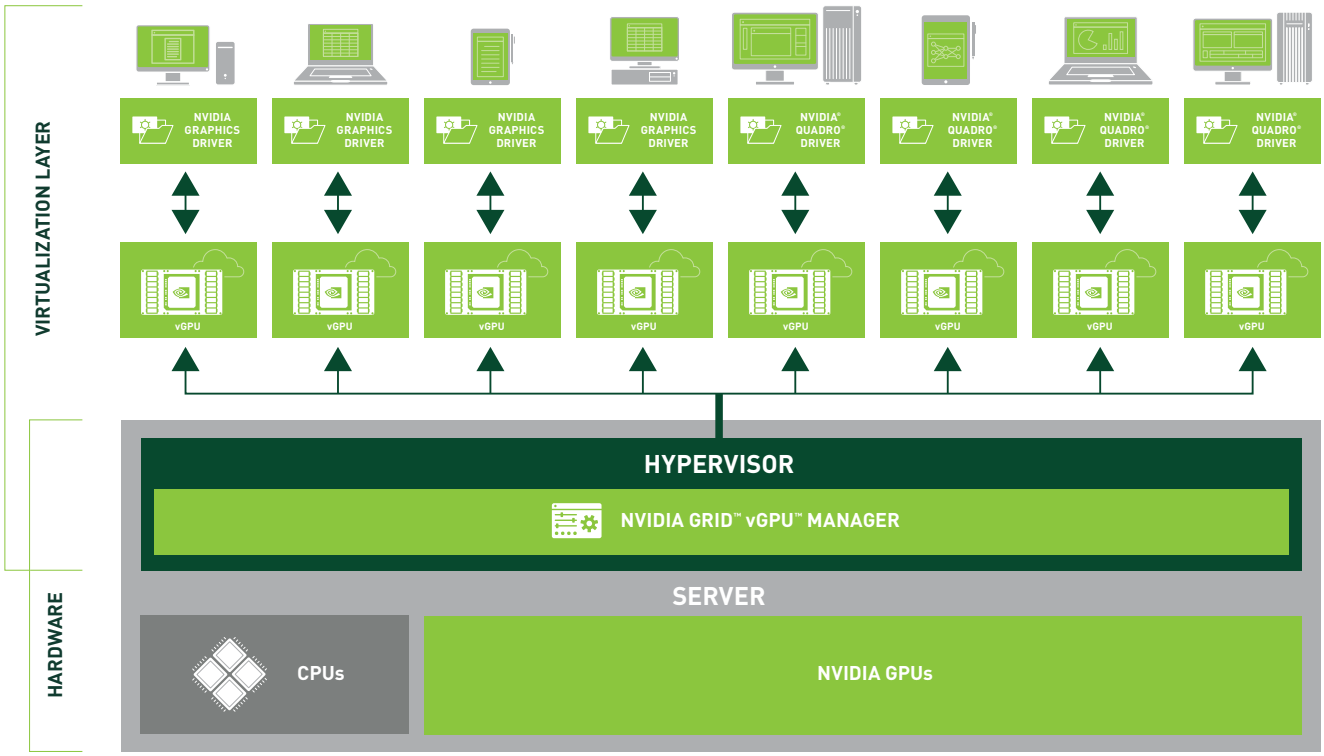


NVIDIA GRID™ 가상 GPU 기술

NVIDIA GRID™ 는 여러 가상 데스크톱 및 애플리케이션 인스턴스에서 가상 GPU(vGPU)를 공유할 때 사용되는 기술 중에서도 업계에서 가장 진보된 기술입니다. 이제 NVIDIA 데이터 센터 GPU의 능력을 전부 활용하여 어디서나, 어느 기기에나 탁월한 가상 그래픽 경험을 제공할 수 있습니다. NVIDIA GRID™ 플랫폼에서는 최고 수준의 성능, 유연성, 관리 기능, 보안을 통해 모든 가상 작업 흐름에 적절한 수준의 사용자 경험을 제공합니다.

커스텀 그래픽 프로파일

IT 관리자는 각 사용자에게 최적의 그래픽 자원을 할당하고 커스텀 그래픽 프로파일을 제공할 수 있습니다. 모든 가상 데스크톱은 실제 데스크톱처럼 전용 그래픽 메모리가 있기 때문에 항상 최고의 성능으로 애플리케이션을 시작하고 실행할 수 있습니다. NVIDIA GRID™ 에서는 수십 명의 사용자가 각각의 실제 GPU를 공유하므로 사용 가능한 GPU 리소스를 최적 균형 상태로 가상 시스템에 할당할 수 있습니다.



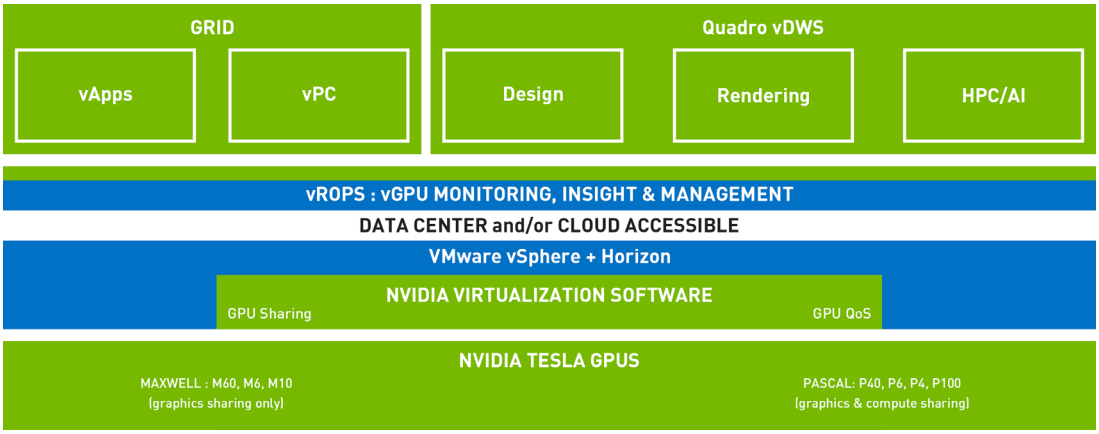
NVIDIA GRID™ Software Edition

NVIDIA GRID™ 는 세가지 에디션이 있으므로 실제요구에 따라 각 사용자 기반에 적절한 수준의 리소스를 할당할 수 있습니다.

XenApp 또는 기타 RDSH 솔루션을 배포하는 조직에 적합 PC Windows 애플리케이션을 최대 성능으로 제공하도록 설계	가상 데스크톱을 원하지만 PC애플리케이션, 브라우저, 고화질 비디오를 활용한 최상의 사용자 경험도 원하는 사용자에게 적합	어디서나, 어느 기기에서나 최상의 성능으로 원격 전문 그래픽 애플리케이션을 사용하려는 사용자에게 적합

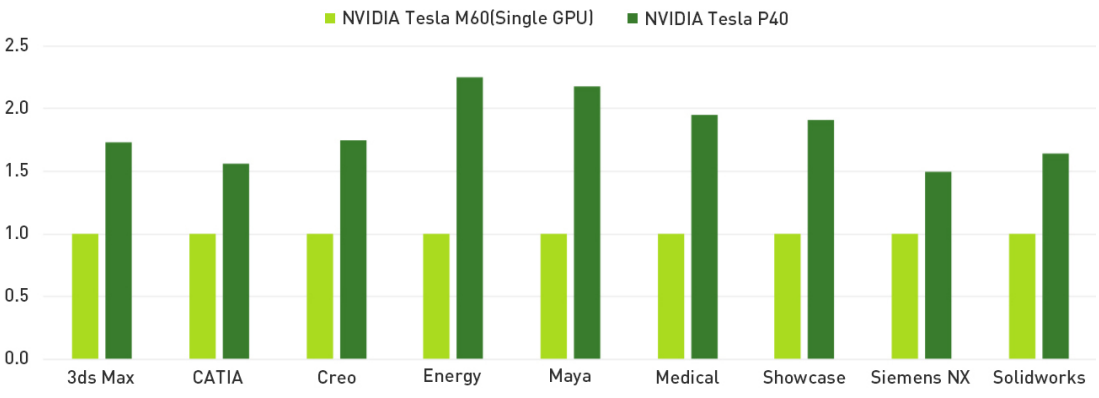
NVIDIA GRID™ GPU 가상화 플랫폼

Industry Leading Virtualization Platform



높은 성능의 어플리케이션을 위한 향상된 성능

Tesla M60 보다 2배 이상 뛰어난 Tesla P40



GPU Throughput of a P40 Compared to a Single M60 GPU. The maximum throughput per GPU compares the overall performance of a GPU that can be shared across multiple virtual machines. The score differs to a single SPEC ViewPert 12.1 score because the GPU is only consistently and fully utilized with multiple virtual machines.

NVIDIA GRID™ Support with vRealize Operations for Horizon

Entire Stack Monitoring

Insights into Users, Apps and Infrastructure

Single Pane of Glass

Monitor both Horizon and XenApp Stacks

Right-Sized Resources

Utilization Metrics and Management

End-User Viewpoints

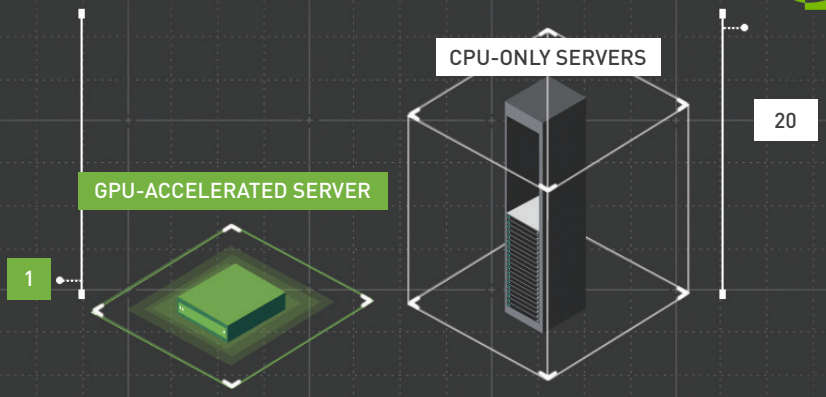
Optimize Performance and Meet SLAs

GPU Support

Monitor GPU status

Powered by NVIDIA

GPU Servers

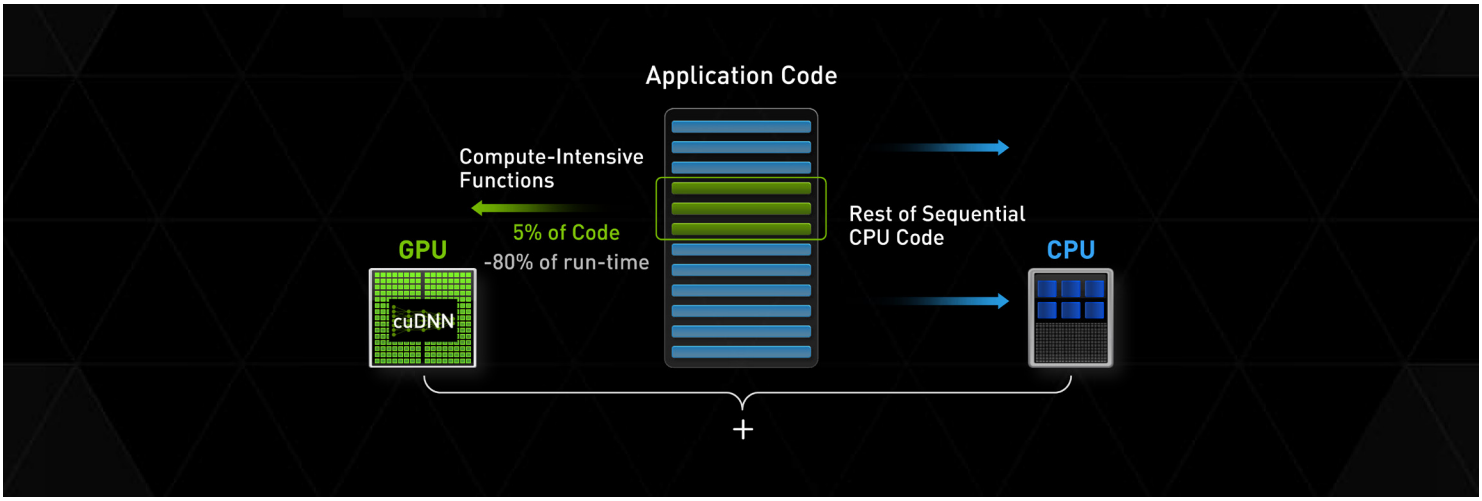


GPU 가속 컴퓨팅이란?

GPU 가속 컴퓨팅은 그래픽 처리 장치(GPU)와 CPU를 함께 이용하여 과학, 분석, 공학, 소비자 및 기업 애플리케이션의 처리속도를 높이는 것을 말합니다. NVIDIA에 의해 2007년도에 개척된 GPU 가속은 현재 전세계 정부 산하의 연구소, 대학교, 대기업 및 중소기업의 에너지 효율적인 데이터센터에 사용되고 있습니다. GPU는 자동차부터 휴대폰, 태블릿, 드론 및 로봇까지 아우르는 다양한 플랫폼상의 애플리케이션을 가속합니다.

GPU가 애플리케이션을 가속하는 방법

어플리케이션의 연산 집약적인 부분을 GPU로 넘기고 나머지 코드만을 CPU에서 처리하는 GPU 가속 컴퓨팅은 전례 없이 강력한 어플리케이션 성능을 제공합니다. 사용자 입장에서는 어플리케이션의 속도가 놀라울 정도로 빨라 졌음을 느낄 수 있습니다.

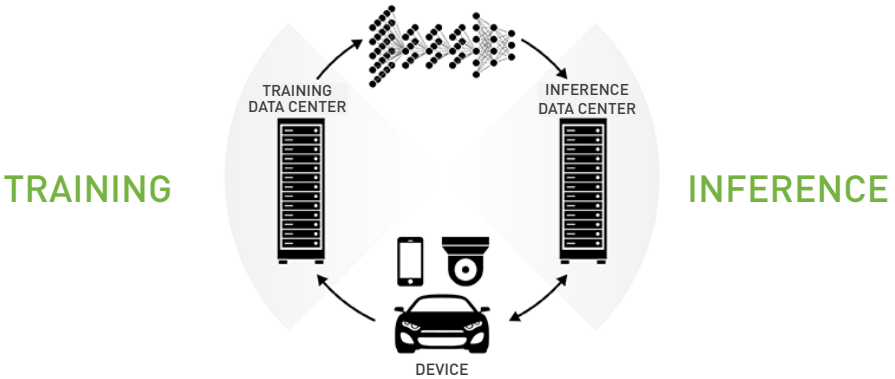


BayNex의 GPU Server는

NVIDIA® Tesla® GPU를 사용하여 높은 사양을 요구하는 HPC와 하이퍼스케일 데이터센터 워크로드를 가속화 할 수 있습니다. 데이터 과학자와 연구자는 다양한 응용 분야에서 기존의 CPU가 지원했던 수준보다 훨씬 더 빠르게 페타바이트급 데이터 주문을 파싱 할 수 있습니다. 그 뿐만 아니라 BayNex Server는 Tesla 가속기를 활용하여 훨씬 빠른 속도로 더욱 큰 규모의 시뮬레이션을 실행하는 데 필요한 성능을 제공합니다.

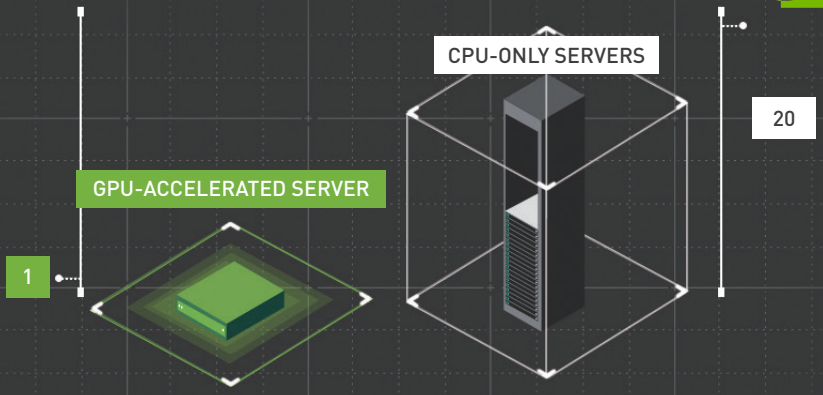
점점 더 복잡해지는 모델을 더욱 빠르게 트레이닝하는 것은 데이터 과학자의 생산성을 향상하고 AI 서비스를 보다 빠르게 제공하는데 매우 중요합니다. NVIDIA® Tesla® P100/V100을 탑재한 서버는 딥 러닝 트레이닝 시간을 몇 개월에서 몇 시간으로 단축할 수 있습니다.

추론은 트레이닝을 마친 신경망이 실질적으로 역할을 수행하는 영역입니다. 이미지, 음성, 비주얼 및 동영상 검색과 같은 새로운 데이터 요소가 등장함에 따라 추론은 수많은 AI 서비스의 중심에서 그에 대한 대답과 추천을 제공합니다. 단 한 개의 Tesla GPU를 장착한 서버가 싱글 소켓 CPU로만 구성된 서버보다 27배 높은 추론 처리량을 제공하여 비용을 대폭 절감할 수 있습니다.



Powered by NVIDIA

GPU Servers



NVIDIA GPU 전용서버

새로 나온 Intel Skylake 기반의 CPU 와 NVIDIA GPU의 결합은 이전의 Xeon CPU보다 약 30-50% 증가된 성능을 자랑합니다.

BayNex 서버에서는 그래픽, HPC 및 머신 러닝(Deep Learning) 분야에 활용되는 다양한 GPU 옵션을 제공합니다. 업무 용도에 따라 다양한 GPU 중 적합한 모델을 선택할 수 있으며, 시스템에 따라 여러 개의 GPU 어댑터를 장착하도록 설계되었습니다.

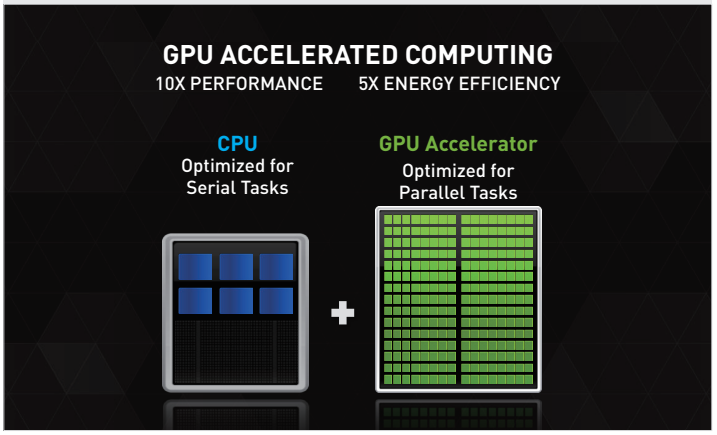
GPU 서버의 장점

- 고객 맞춤형 서버가능 (다른 Appliance 제품과 달리 내부구성 변경가능)
- NVIDIA의 최신 아키텍처 적용서버 – Volta 아키텍처 사용가능
- 투자대비 효용성 증가
- Tech Support by NVIDIA
- 다른 Appliance와 달리 즉시 공급이 가능, 엔비디아 파트너사의 24/7 지원가능

NVIDIA Library를 사용하는 어플리케이션 및 기관

NAMD, VASP, TOTAL, GROMACS, LAMMPS, Gaussian, SKA, HOOMD-blue, MATLAB, Desmond, COSMO, Amber

GPU 는 수 천개의 코어로 CPU 대비 연산집약적 병렬컴퓨팅에서 수십~수백배의 성능을 자랑합니다.



GPU Server Line up

POWERED BY NVIDIA




	BN280	BN480	BN880	BN882
Form Factor	2U	2U	2U	2U
CPU	2* Skylake TDP up to 165W	2* Skylake TDP up to 165W	2* Skylake TDP up to 165W	2* Skylake TDP up to 165W
Chipset	Intel® C620 series chipset	Intel® C620 series chipset	Intel® C620 series chipset	Intel® C620 series chipset
Memory	24 DIMM slots	24 DIMM slots	16× DDR4 DIMM per node	16× DDR4 DIMM per node
GPU	PCIe Max. 2ea	PCIe Max. 4ea	PCIe Max. 8ea	SXM2 8ea
	V100 / P40	V100 / P40	V100 / P40	V100
Power	2 * 500w/800w/1600w	2 * 500w/800w/1200w/1600w	2 * 3000w PSU 80plus Titanium	2 * 3000w PSU 80plus Titanium

| NVIDIA 한국총판

BayNex

(주)베이넥스

서울시 영등포구 당산로 41길 11 당산 SK V1센터 West동 1212호

TEL : 02)785-9977 | FAX : 02)785-7447

http://www.baynex.co.kr

| NVIDIA 공식파트너

iit (주)한국인프라 Infra Information Technology Co., Ltd.

(주)한국인프라

서울특별시 강남구 삼성로 150 (대치동,극동교회빌딩 3층)

상담문의 : 02)6204-5024 | jysuh@krinfra.co.kr

http://www.krinfra.co.kr