

NVIDIA DGX-2
THE WORLD'S MOST POWERFUL DEEP LEARNING SYSTEM FOR THE MOST COMPLEX AI CHALLENGES

현재의 AI와 Deep Learning의 요구를 충족시키기 위한 스케일링의 챌린지

딥 뉴럴 네트워크는 비즈니스 및 연구에서 가장 시급한 과제에 대응하여 크기와 복잡성이 급속히 증가하고 있습니다. 오늘날의 현대적 AI 작업 부하를 지원하는 데 필요한 계산 능력은 기존의 데이터 센터 아키텍처를 능가합니다. 개발자가 점점 커지는 가속화된 컴퓨팅 클러스터를 구축하고 데이터 센터 규모의 한계를 뛰어 넘으면서 모델 병렬을 사용하는 최신 기술은 GPU 간 대역폭의 한계와 충돌합니다. 세계를 변화시킬 수있는 더 빠른 통찰력을 얻기위한 장벽을 뛰어 넘기 위해 거의 무한한 AI 컴퓨팅규모를 제공하는 새로운 접근 방식이 필요합니다.

이전에는 불가능했던 훈련

점차 복잡해지는 AI는 전례없는 수준의 계산을 요구합니다. NVIDIA DGX–2 는 세계에서 가장 앞선 GPU 중 16 개를 탑재한 세계 최초의 2 페타플롭 시스템으로 이전에는 통용할 수 없었던 최신 학습 모델 유형을 가속화합니다. 획기적인 GPU 스케일을 사용하면 단일 노드에서 4 배 더 큰 모델을 학습 할 수 있습니다. 레거시 x86아키텍처와 비교할 때, DGX–2의 ResNet–50 교육 능력은 듀얼 Intel Xeon Gold CPU를 장착 한 300대의 서버에 270만 달러 이상이 필요합니다.

NVIDIA NVSwitch – 혁신적인 AI 네트워크 패브릭

최첨단 연구는 모델 병렬성을 활용할 자유를 요구하며 이전에는 볼 수 없었던 수준의 GPU 간 대역폭을 필요로합니다. NVIDIA는 이러한 요구를 해결하기 위해 NVSwitch를 개발했습니다. 다이얼업에서 초고속 광대역에 이르는 발전과 마찬가지로 NVSwitch는 오늘날의 미래를위한 네트워킹 패브릭을 제공합니다. NVIDIA DGX–2를 사용하면 기존 아키텍처의 한계로 인해 모델의 복잡성과 크기가 더 이상 제한받지 않습니다. DGX–2의 네트워킹 패브릭과 모델 병렬 트레이닝을 결합하여 이전 세대에 비해 24배 증가된 2.4TB/s의 이분법 대역폭을 제공합니다. 이 새로운 인터커넥트 "superhighway"는 16개의 GPU에 분산된 교육의 힘을 한 번에 얻을 수있는 모델 유형에 무한한 가능성을 제공합니다.

SYSTEM SPECIFICATIONS

GPUs	16X NVIDIA ® Tesla V100
GPU Memory	512GB total
Performance	2 petaFLOPS
NVIDIA CUDA ® Cores	81920
NVIDIA Tensor Cores	10240
NVSwitches	12
Maximum Power Usage	10 kW
CPU	Dual Intel Xeon Platinum 8168, 2.7 GHz, 24-cores
System Memory	1.5TB
Network	8X 100Gb/sec Infiniband/100GigE Dual 10/25Gb/sec Ethernet
Storage	OS: 2X 960GB NVME SSDs Internal Storage: 30TB (8X 3.84TB) NVME SSDs
Software	Ubuntu Linux OS See Software stack for details
System Weight	340 lbs (154.2 kgs)
System Dimensions	Height: 17.3 in (440.0 mm) Width: 19.0 in (482.3 mm) Length: 31.3 in (795.4 mm) - No Front Bezel 32.8 in (834.0 mm) - With Front Bezel
Operating Temperature Range	5°C to 35°C (41°F to 95°F)

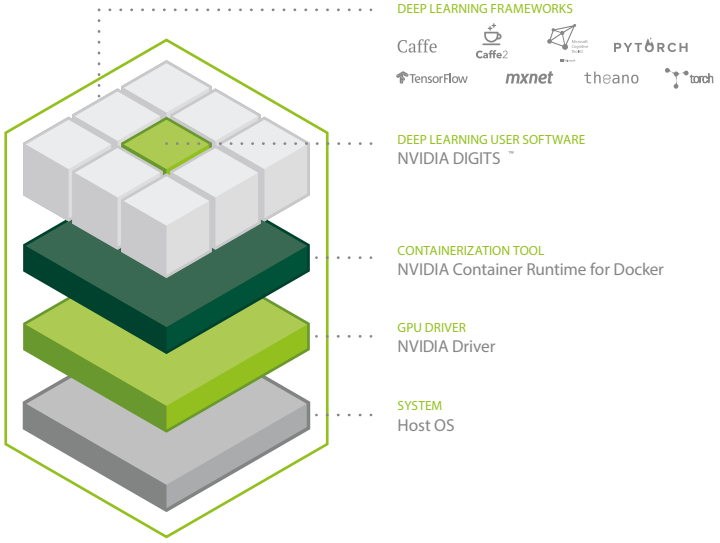
완전히 새로운 수준의 AI Scale

현대 기업은 비즈니스 요구 사항에 따라 AI 파워를 신속하게 배포해야하며, 비용이나 복잡성을 높이지 않고 AI를 수평 확장해야합니다. 우리는 DGX-2를 구축했으며 가속화된 배치, 운영 단순화를 가능하게하는 DGX 소프트웨어로 전원을 공급합니다. DGX-2는 가상화 지원과 함께 AI 확장에 대한 가장 빠른 경로를 제공하는 즉시 사용 가능한 솔루션을 제공하여 개인 엔터프라이 급 AI 클라우드를 구축 할 수 있습니다. 이제 비즈니스는 가속화 된 컴퓨팅 자원을 함께 묶는 데 필요한 일부 네트워크 인프라로 손쉽게 확장 할 수 있는 솔루션에서 제한없는 AI 성능을 활용할 수 있습니다. 가속된 배포 모델과 간편한 규모의 아키텍처를 통해 팀은 통찰력을 높이고 인프라 구축 시간을 단축 할 수 있습니다.

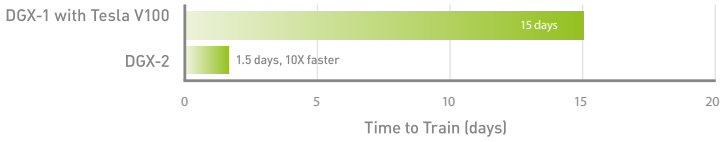
엔터프라이즈급 AI 인프라

AI 플랫폼이 귀사 비즈니스에 핵심적인 것이라면 RAS (Reliability, Availability and Serviceability)를 염두에두고 설계된 AI 플랫폼이 필요합니다. DGX-2는 엄격한 24 시간 AI 작업을 위해 구축 된 엔터프라이즈 급이며, 계획되지 않은 다운 타임을 줄이고 서비스 가능성을 능률화하며 작동 지속성을 유지하기 위해 RAS 용으로 특수 제작되었습니다. 튜닝 및 최적화 시간을 줄이고 발견에 더 많은 시간을 집중하십시오. NVIDIA의 엔터프라이즈 급 지원은 시간이 많이 걸리는 하드웨어 및 오픈 소스 소프트웨어 문제 해결 작업에서 당신을 보호합니다. 모든 DGX 시스템에서 소프트웨어, 도구 및 NVIDIA 전문 지식이 포함된 통합 솔루션을 사용하여 신속하게 시작하고, 빠르게 학습하고, 빠르게 유지하십시오.

NVIDIA DGX Software Stack



NVIDIA DGX-2 Delivers 10 Times Faster Deep Learning Training



Workload: FairSeq, 55 epochs to solution. PyTorch training performance.

KOLON BENIT

NVIDIA PREFERRED PARTNER

NVIDIA PREFERRED PARTNER **코오롱베니트㈜ NVIDIA 팀**

13837 경기도 과천시 코오롱로 11, 9층 (별양동, 코오롱타워)
Tel. 02-3677-7078 / E-mail linuxer@kolon.com / 홈페이지 : www.kolonbenit.com

자세한 정보는 www.nvidia.com/dgx-station에서 보실 수 있습니다.

© 2018 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, and Iray are trademarks and/or registered trademarks of NVIDIA Corporation. All company and product names are trademarks or registered trademarks of the respective owners with which they are associated. Features, pricing, availability, and specifications are all subject to change without notice. APR 18

NVIDIA