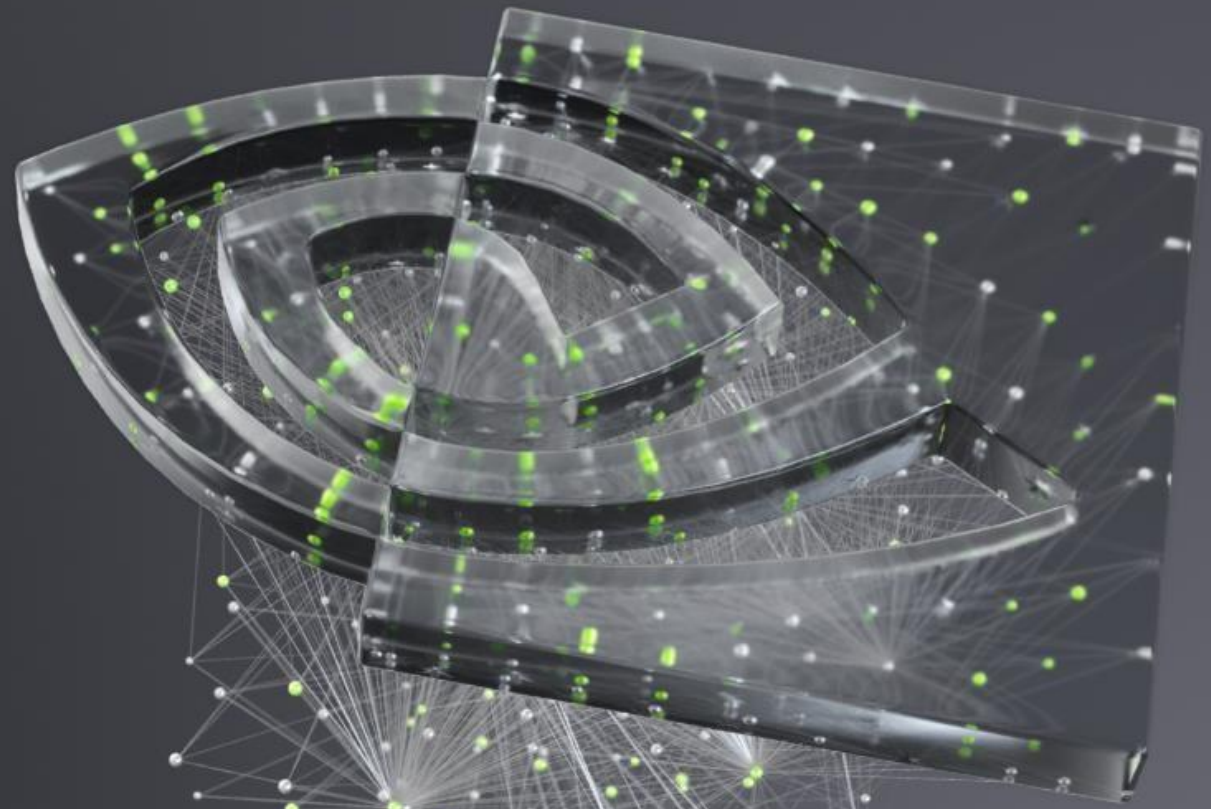
DGX A100 소개

발표자 : ㈜ 베이넥스 전우정 상무

NVIDIA Distributor Partner



NVIDIA

PURESTORAGE®

NVIDIA - Visual computing company

GeForce

Quadro

DATACENTER

TEGRA

AI & DATA SCIENCE IS THE KEY TO MODERN BUSINESS

Forecasting, Fraud Detection, Conversational AI, Recommendations, and More

RETAIL

Supply Chain & Inventory Management
Price Management / Markdown Optimization
Promotion Prioritization And Ad Targeting

TELECOM

Detect Network/Security Anomalies
Forecasting Network Performance
Network Resource Optimization (SON)

HEALTHCARE

Improve Clinical Care
Drive Operational Efficiency
Speed Up Drug Discovery

MANUFACTURING

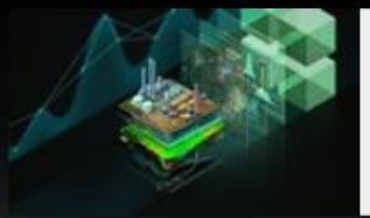
Remaining Useful Life Estimation
Failure Prediction
Demand Forecasting

FINANCIAL SERVICES

Claim Fraud
Customer Service Chatbots/Routing
Risk Evaluation

CONSUMER INTERNET

Ad Personalization
Click Through Rate Optimization
Churn Reduction



OIL & GAS

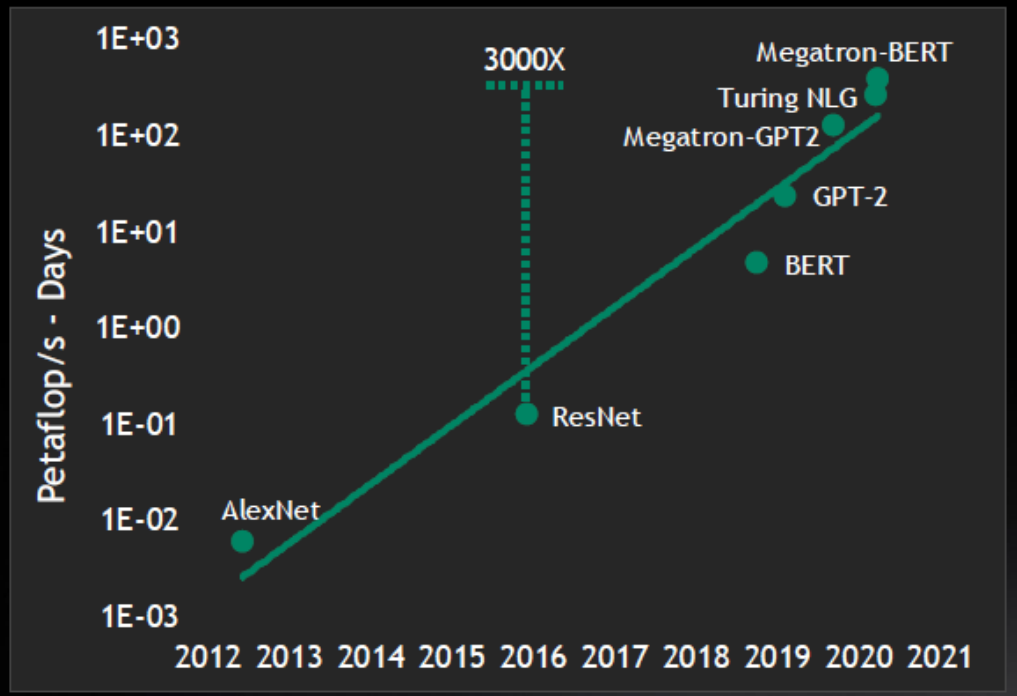
Sensor Data Tag Mapping
Anomaly Detection
Robust Fault Prediction

AUTOMOTIVE

Personalization & Intelligent Customer Interactions
Connected Vehicle Predictive Maintenance
Forecasting, Demand, & Capacity Planning

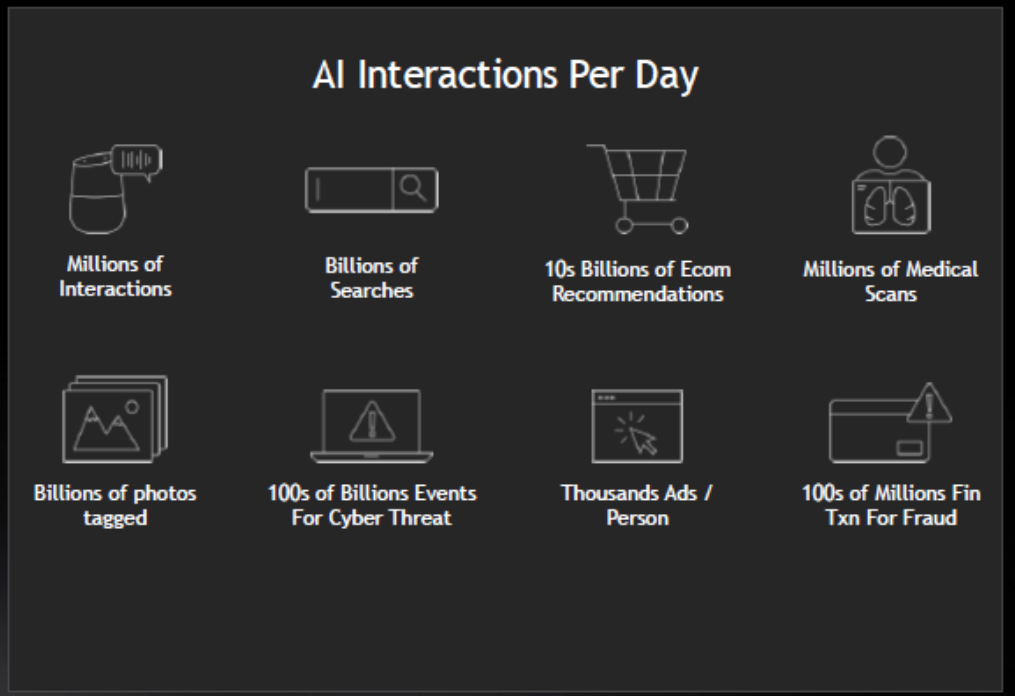
CHALLENGES: ACCELERATING BIG AND SMALL

AI Advances Demand Exponentially Higher Compute



3000X Higher Compute Required to Train Largest Models Since Volta

AI Applications Demand Distributed Pervasive Acceleration



Every AI Powered Interaction Needs Varying Amount of Compute

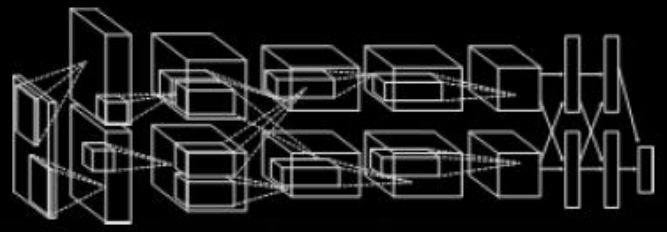
Source: OpenAI, NVIDIA

REIMAGINING THE GPU

Three Breakthroughs to Fuel the Next Era of Modern Accelerated Data Centers

20X

A GIANT LEAP IN PERFORMANCE



UNIFIED AI TRAINING AND INFERENCE ACCELERATION

1-50

SCALABILITY FOR THE ELASTIC DATACENTER

INTRODUCING NVIDIA A100 SXM AND NVIDIA A100 PCIE

Greatest Generational Leap - 20X Volta

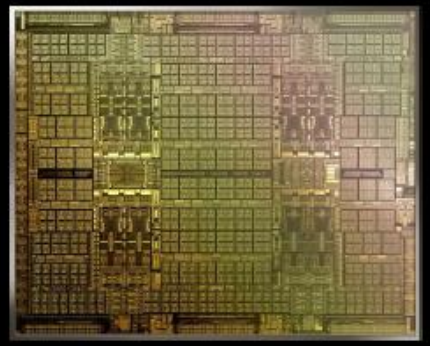
	Peak		Vs Volta
FP32 TRAINING	312	TFLOPS	20X
INT8 INFERENCE	1,248	TOPS	20X
FP64 HPC	19.5	TFLOPS	2.5X
MULTI INSTANCE GPU			7X GPUs

A100 PCIe

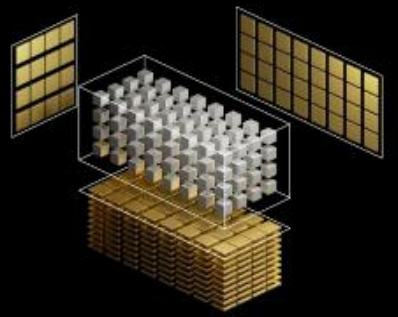
A100 SXM

54B XTOR | 826mm2 | TSMC 7N | 40GB Samsung HBM2

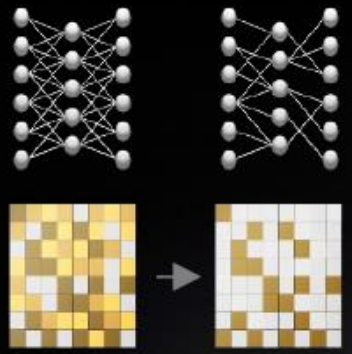
5 MIRACLES OF A100



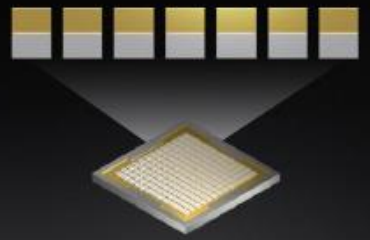
NVIDIA Ampere Architecture
World's Largest 7nm chip
54B XTORS, HBM2



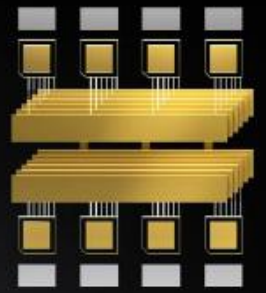
3rd Gen Tensor Cores
Faster, Flexible, Easier to use
20x AI Perf with TF32
2.5x HPC Perf



New Sparsity Acceleration
Harness Sparsity in AI Models
2x AI Performance



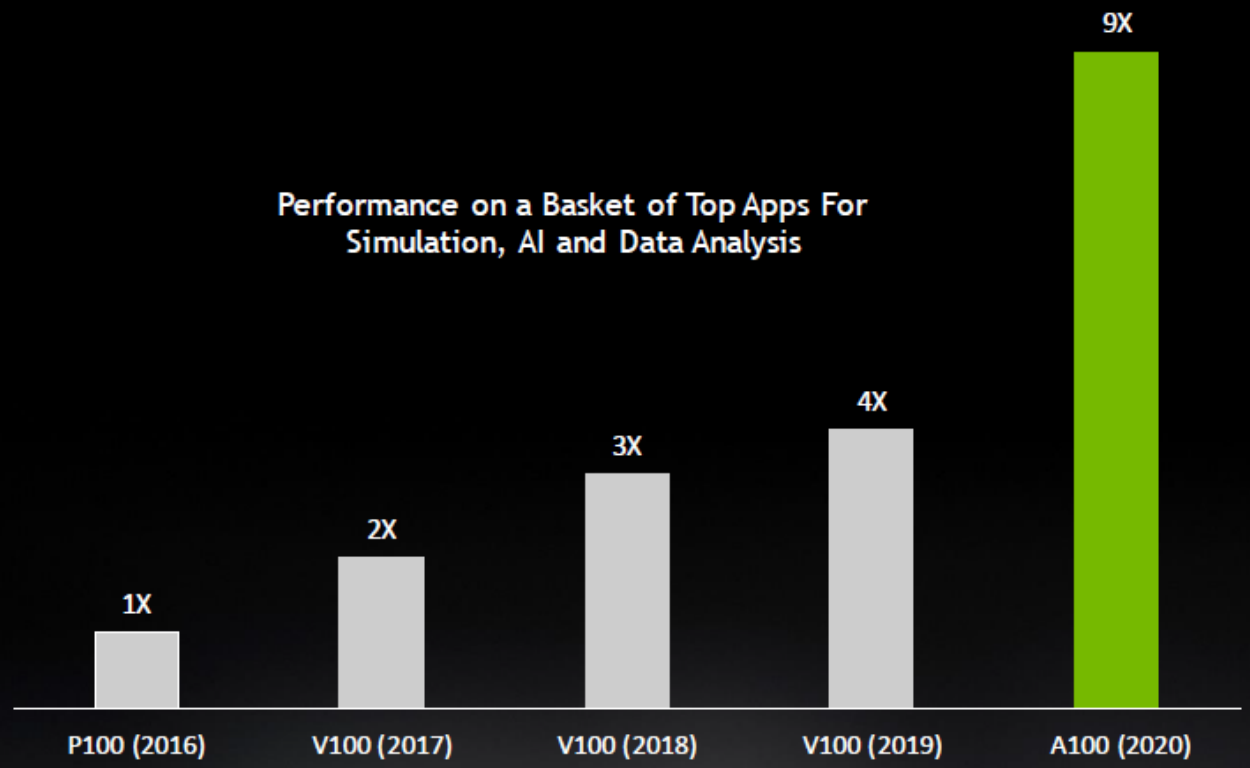
New Multi-Instance GPU
Optimal utilization with right sized GPU
7x Simultaneous Instances per GPU



3rd Gen NVLINK and NVSWITCH
Efficient Scaling to Enable Super GPU
2X More Bandwidth

9X MORE PERFORMANCE IN 4 YEARS

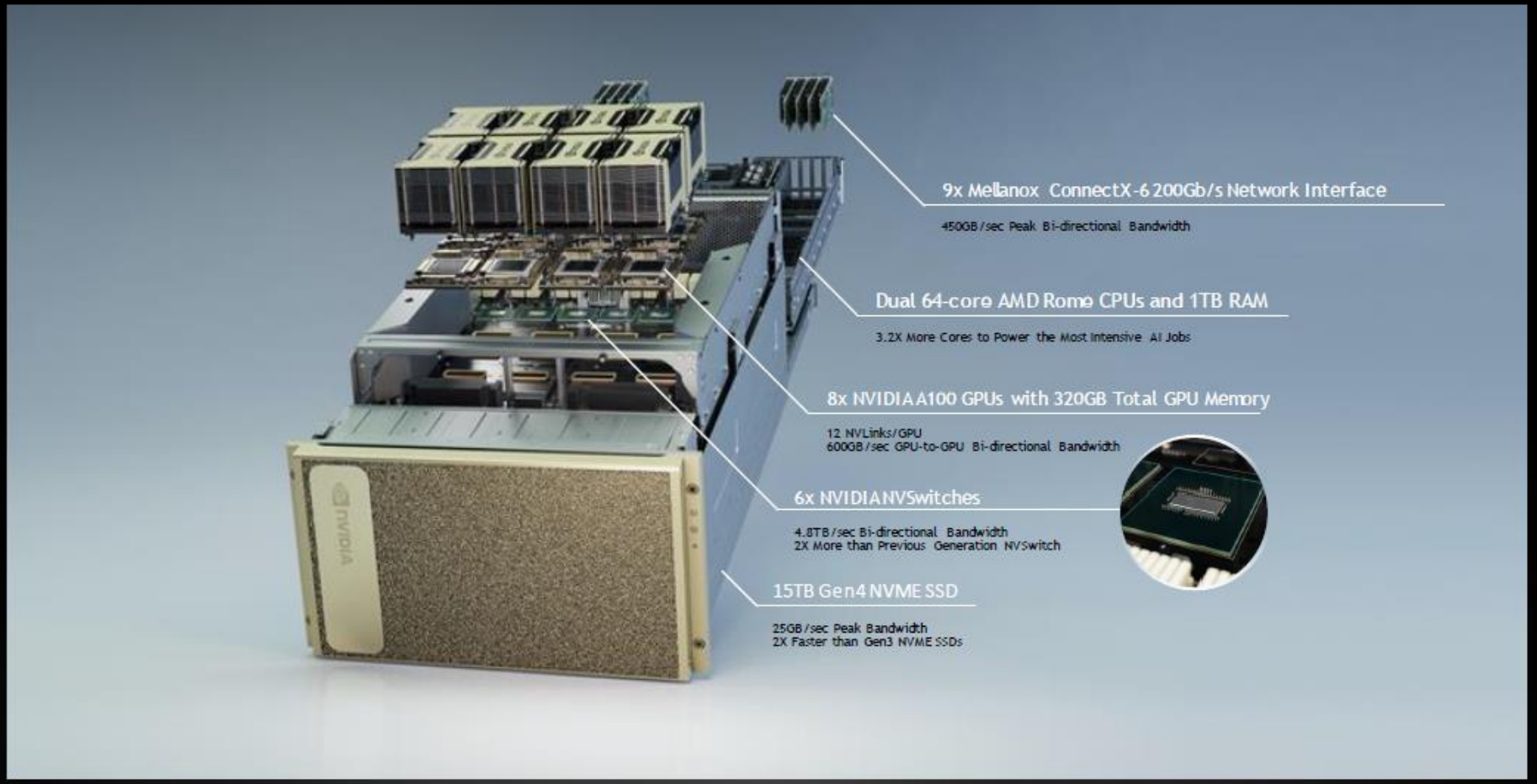
Beyond Moore's Law With Full Stack Innovation



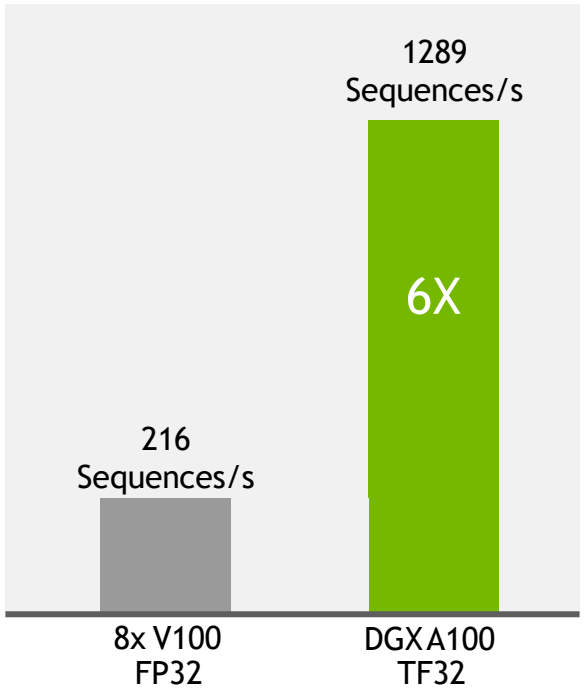
Geometric Mean of application speedups vs. P100 : Benchmark Application: Amber [PME-Cellulose_NVE], Chroma [szscl21_24_128], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT Large Fine Tuner), Quantum Espresso [AUSURF112-jR]; Random Forest FP32 [make_blobs (160000 x 64 : 10)], TensorFlow [ResNet-50], VASP 6 [Si Huge], |GPU node: with dual-socket CPUs with 4x P100, V100, or A100 GPUs.

INTRODUCING DGX A100

The Universal AI System - Data Analytics, Training and Inference



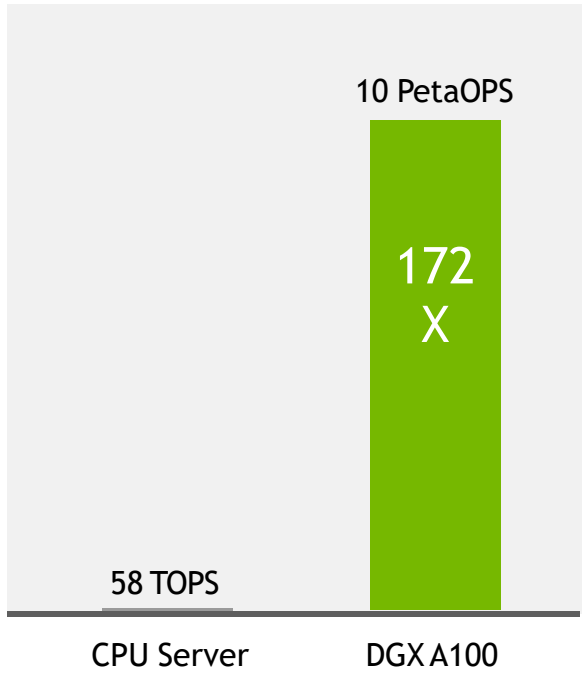
DGX A100 PERFORMANCE



Training

NLP: BERT-Large

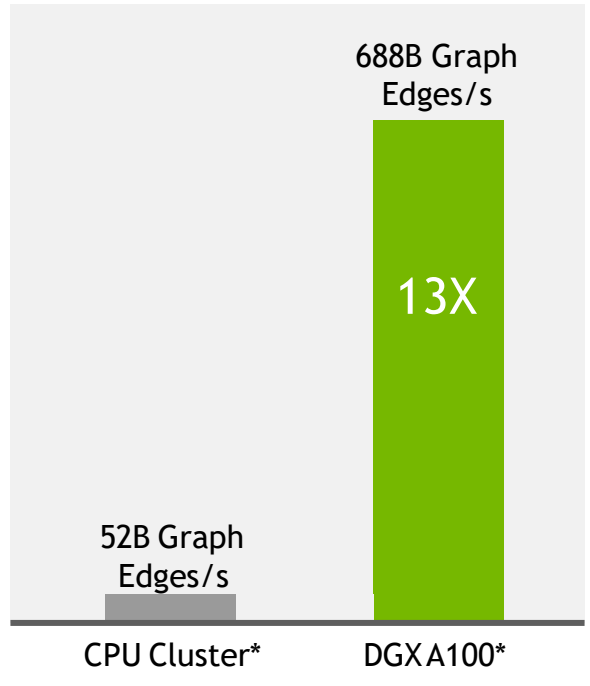
BERT Pre-Training Throughput using PyTorch including (2/3)Phase 1 and (1/3)Phase 2 | Phase 1 Seq Len = 128, Phase 2 Seq Len = 512 V100: DGX-1 Server with 8x V100 using FP32 precision DGX A100: DGX A100 with 8x A100 using TF32 precision



Inference

Peak Compute

CPU Server: 2x Intel Platinum 8280 using INT8
DGX A100: DGX A100 with 8x A100 using INT8 with Structural Sparsity



Analytics

PageRank

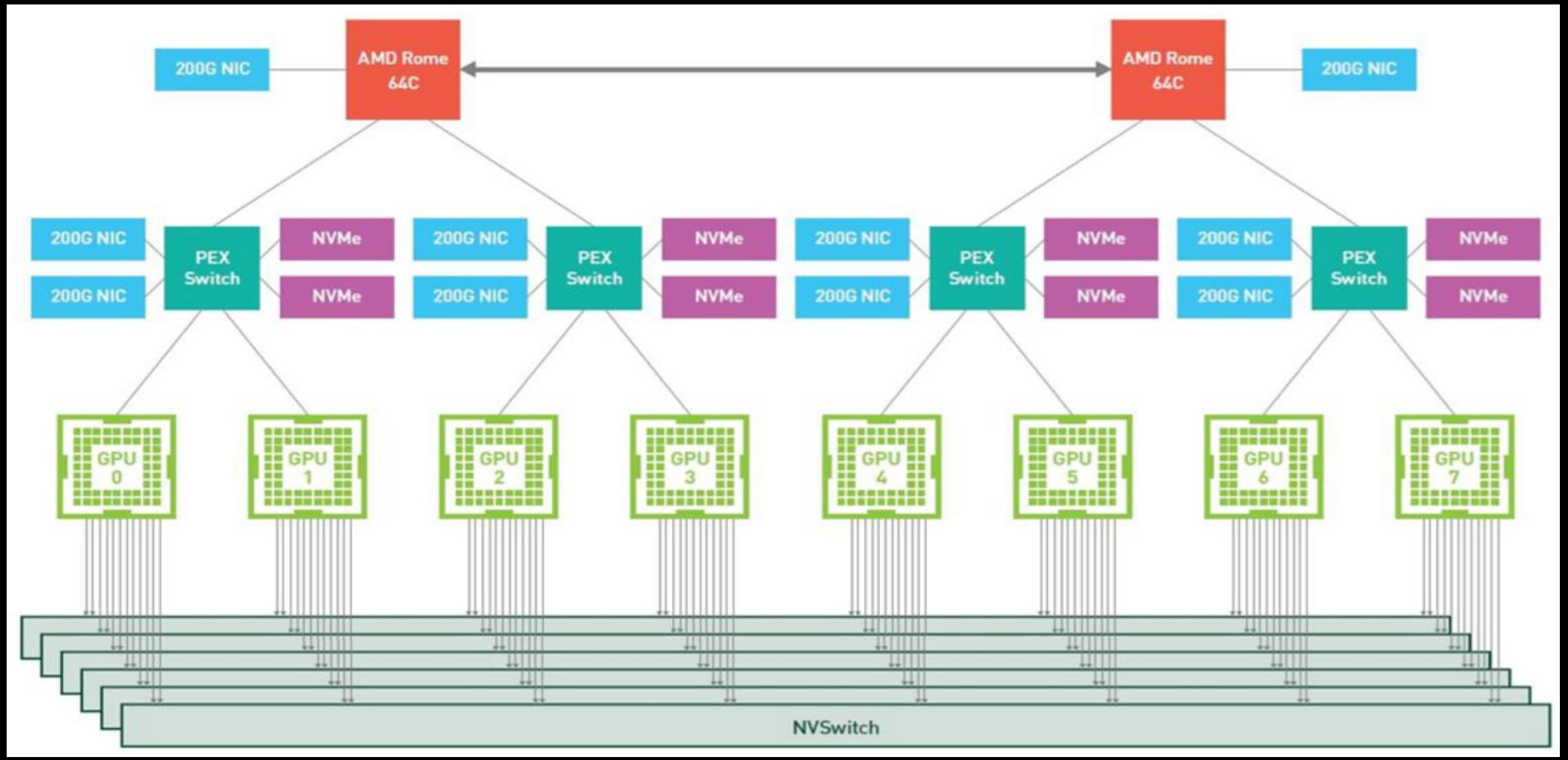
3000x CPU Servers vs. 4x DGX A100
Published Common Crawl Data Set: 128B Edges, 2.6TB Graph

NVIDIA DGX A100 SYSTEM SPECS

App Focus Components	
GPUs	8x NVIDIA A100 Tensor Core GPUs
GPU Memory	320GB /640GB Total
NVIDIA NVSwitch	6
Performance	5 petaFLOPS AI 10 petaOPS INT8
CPU	Dual AMD Rome, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)
System Memory	1TB
Networking	9x Mellanox ConnectX-6 VPI HDR InfiniBand/200GigE 10th Dual-port ConnectX-6 optional
Storage	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 15TB (4x 3.84TB) U.2 NVME drives

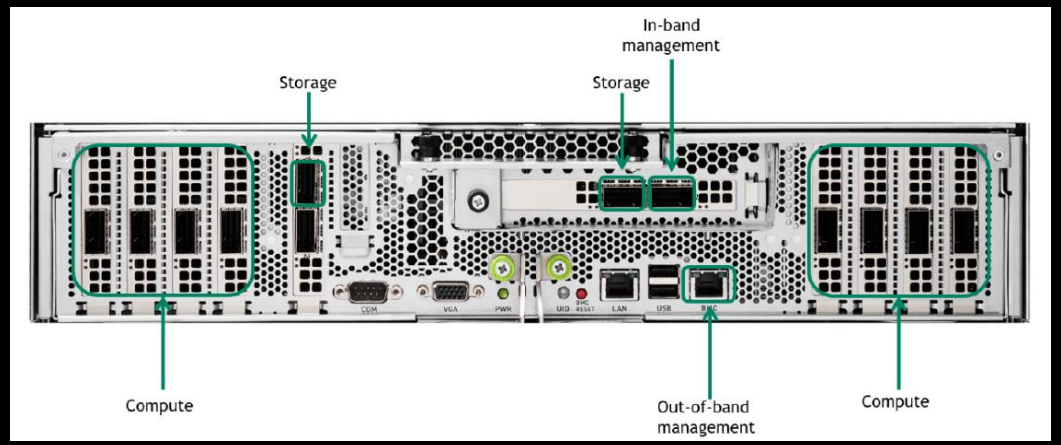
Power and Physical Dimensions	
System Power Usage	6.5 kW Max
System Weight	271 lbs (123 kgs)
System Dimensions	6 Rack Units (RU) Height: 10.4 in (264.0 mm) Width: 19.0 in (482.3 mm) Max Length: 35.3 in (897.1 mm) Max
Operating Temperature	5°C to 30°C (41°F to 86°F)
Cooling	Air

NVIDIA DGX A100 내부 구조



DGX A100 network

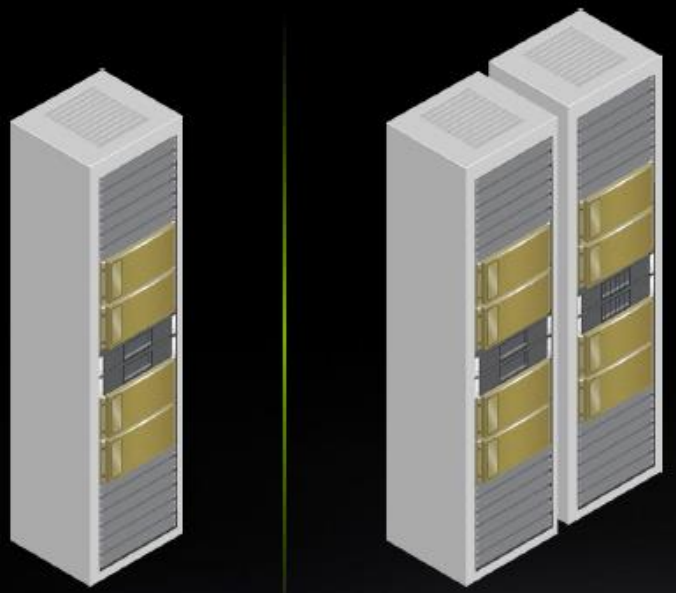
Highest Network Throughput for Data and Clustering



- **Compute fabric.** Connects the eight NVIDIA Mellanox ConnectX®-6 HCAs from each DGX A100 through separate network planes.
- **Storage fabric.** Uses two ports, one each from two dual-port ConnectX-6 HCAs connected through the CPU.
- **In-band management.** Uses a 100 Gbps port on the DGX A100 system to connect to a dedicated Ethernet switch.
- **Out-of-band management.** Connects the baseboard management controller (BMC) port of each DGX A100 system to an additional Ethernet switch.

RACK-SCALE INFRASTRUCTURE

Building an AI Center of Excellence with DGX POD Built on DGX A100



4-node
DGX POD

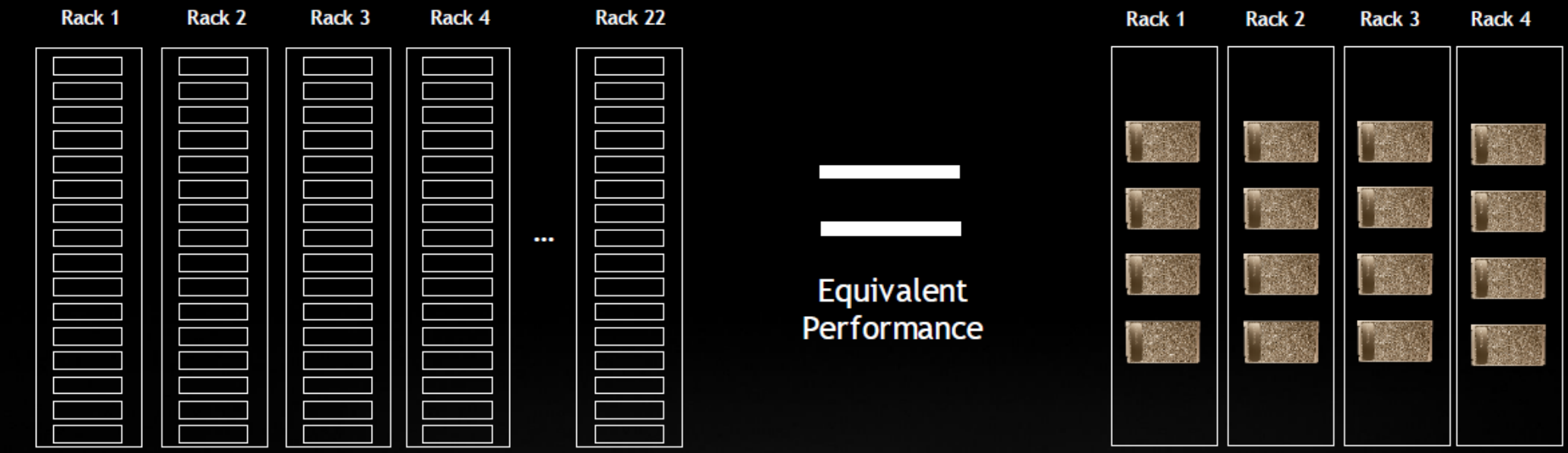
8-node
DGX POD

- DGX POD more attainable than ever with DGX A100
- Experience a faster start with building flexible AI infrastructure
- Proven architectures, with leading storage partners
- Up to 40 PFLOPS computing power in just 2 racks
- 700 PFLOPS of power to train the previously impossible

Complete AI infrastructure solutions:
DGX, storage, networking, services, software

NVIDIA SHATTERS BIG DATA ANALYTICS BENCHMARK

19.5X Faster TPCx-BB Performance Results on DGX A100 with RAPIDS



16 Servers / Rack

350 CPU Servers
$23M | 22 Racks | 300 kW

1/7th Cost
1/3rd Power

16 NVIDIA DGX A100 Systems
$3.3M | 4 Racks |100 kW

Performance: CPU = 4.7 hr, DGX A100 = 14.5 min (19.5x faster); After normalizing performance across CPU and GPU clusters -> Cost: CPU = $23M, DGX A100 = $3.3M (1/7th the cost); Power: CPU = 298kW, DGX A100 = 104kW (1/3rd the power); Space: CPU = 22 racks, DGX A100 = 4 racks (less than 1/5th the space)

THE SELENE SUPERCOMPUTER

NVIDIA's DGX SuperPOD Deployment

#7 on TOP500 (27.6 PetaFLOPS HPL)

#2 on Green500 (20.5 GigaFLOPS/watt)

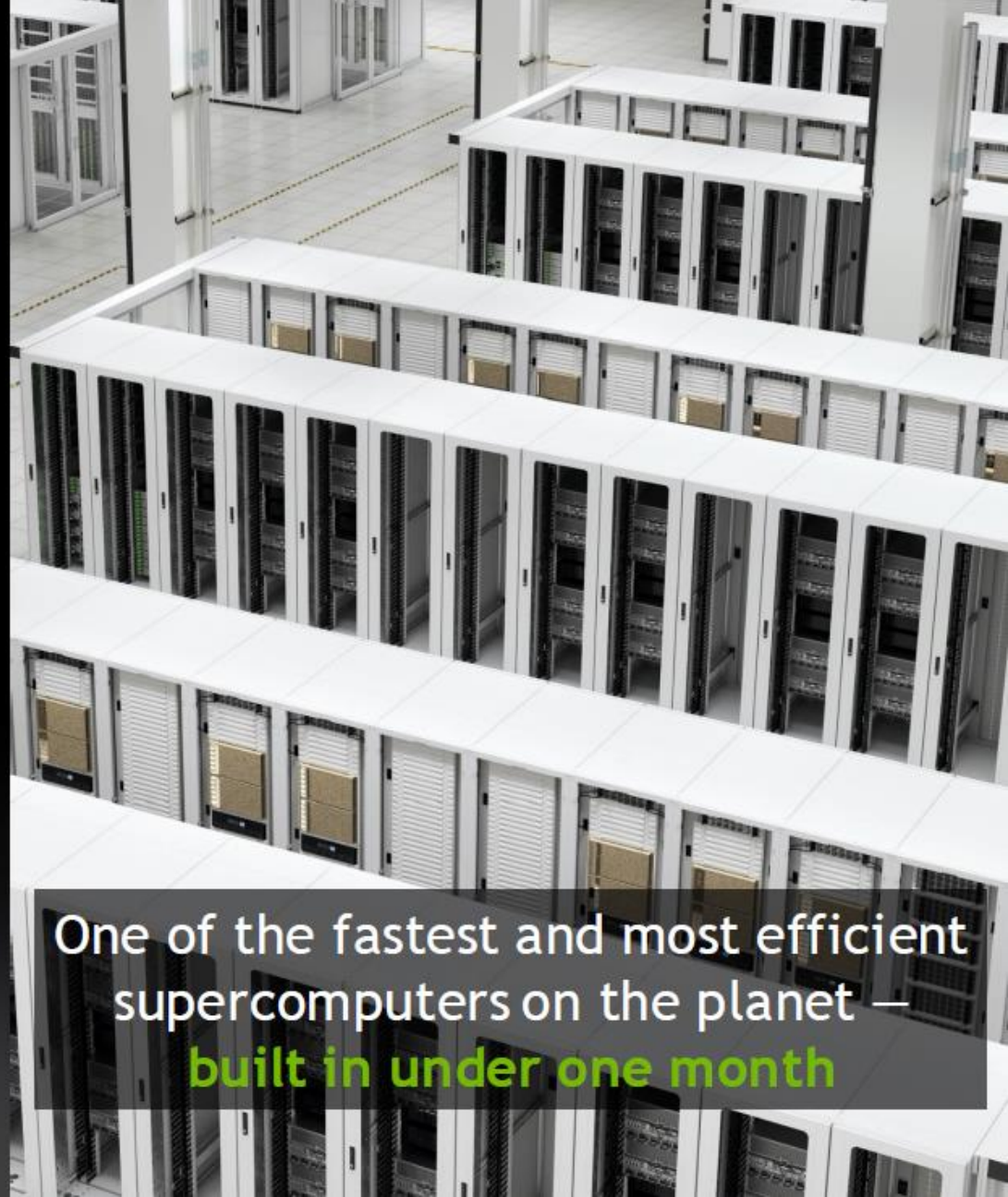
Fastest industrial system in the U.S. — 1+ ExaFLOPS AI

Built from the NVIDIA DGX SuperPOD Reference Architecture

- 2nd generation SuperPOD Reference Architecture
- NVIDIA DGX A100 and NVIDIA Mellanox InfiniBand network
- The blueprint for AI power and scale, infused with the knowhow from NVIDIA's decade plus of AI experience

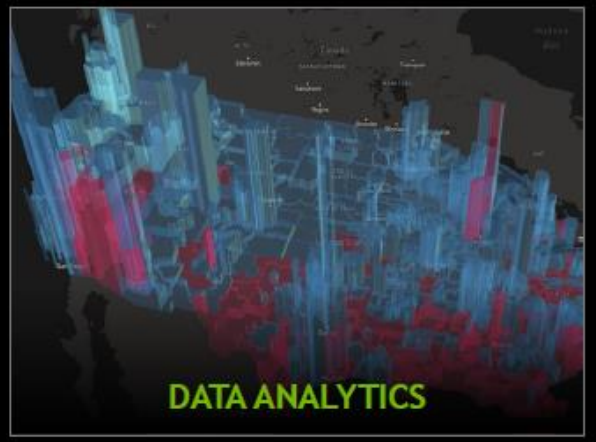
Configuration:

- 2,240 NVIDIA A100 Tensor Core GPUs
- 280 NVIDIA DGX A100 systems
- 494 Mellanox 200G HDR InfiniBand switches
- 7 PB of all-flash storage



NVIDIA ACCELERATED COMPUTING PLATFORM

Accelerating Modern Enterprise Workloads



SPARK3.0 | RAPIDS | more

cuDF | cuML | cuGRAPH

TensorFlow | PyTorch | more

cuDNN| TensorRT | NCCL

NAMD | GROMACS | +700 More

cuBLAS| cuFFT | cuSOLVER

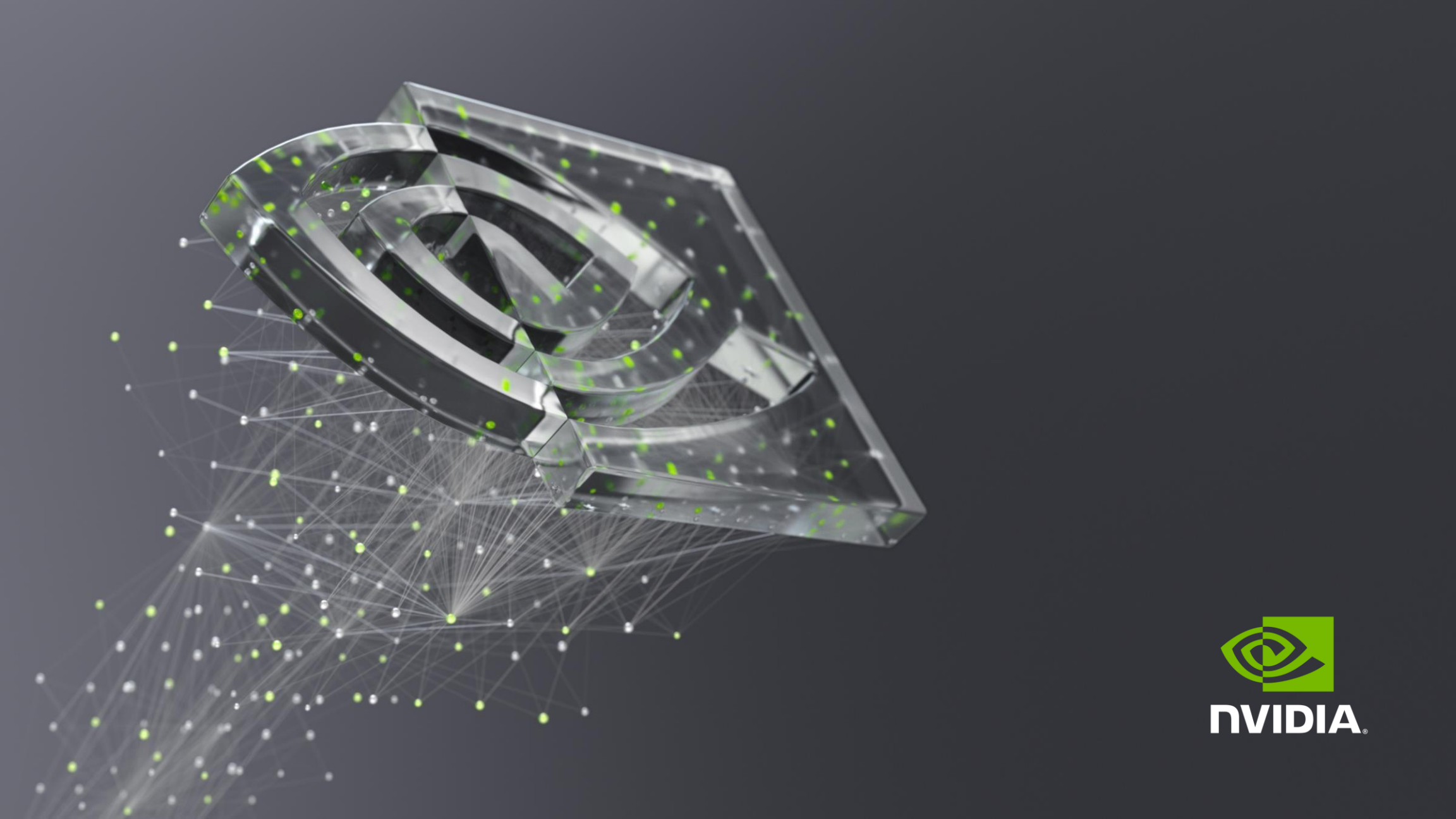
Desktop Development

Data Center Solutions

Accelerated Edge

Supercomputers

GPU-Accelerated Cloud


NVIDIA