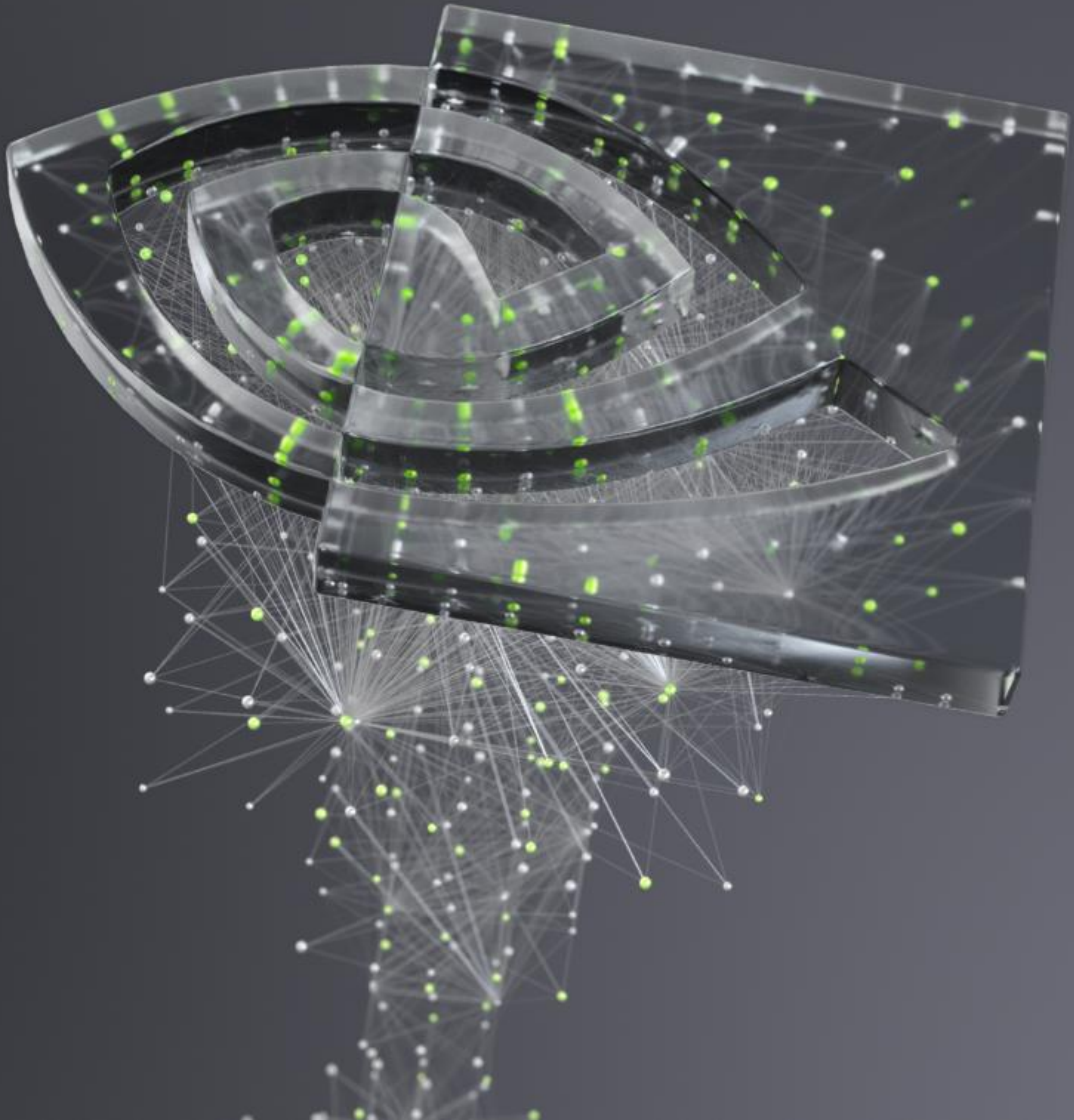


DGX A100

Charlie Boyle

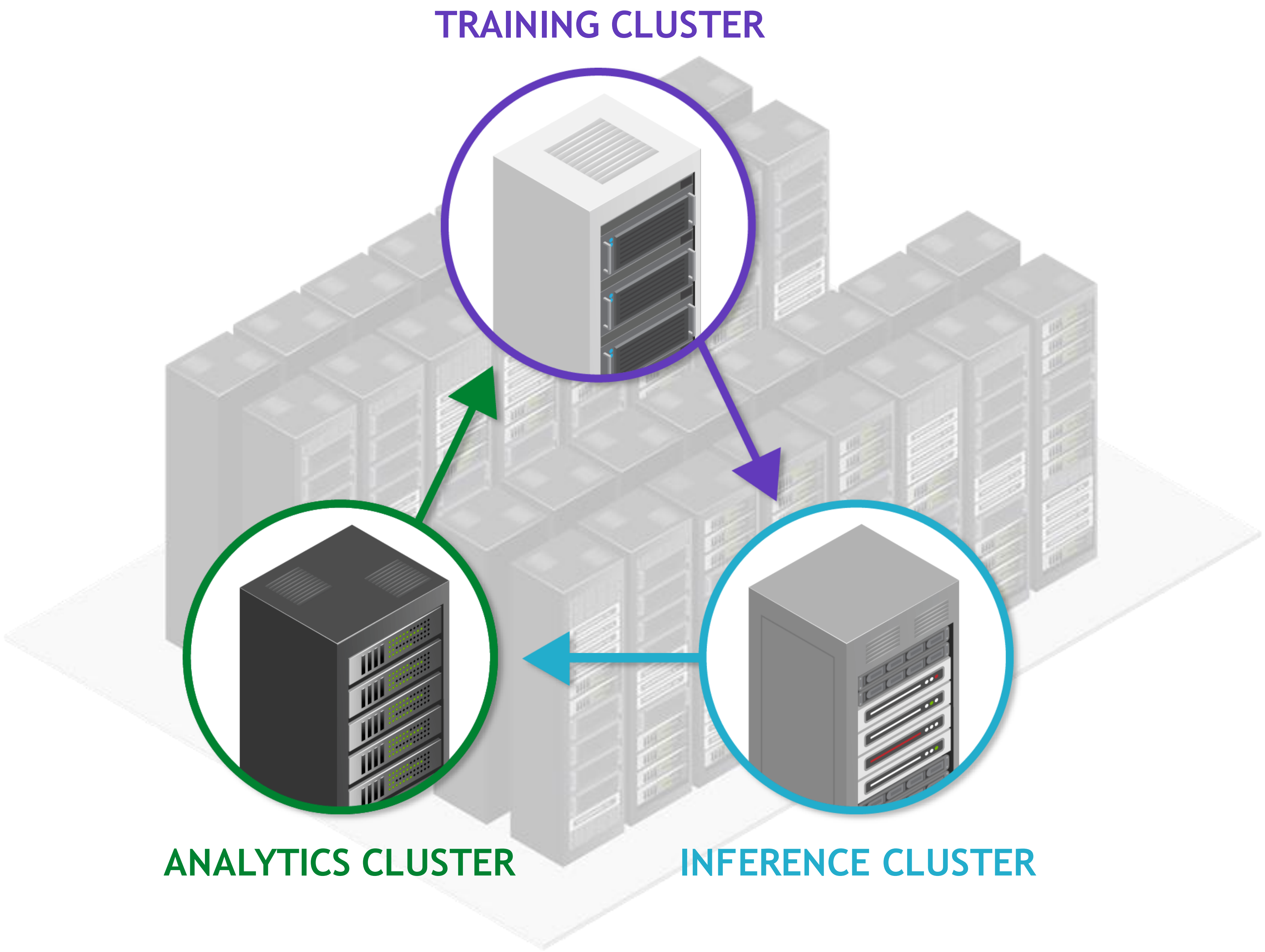
Chris Lamb

Rajeev Jayavant

OVERVIEW

SOLVING THE INFLEXIBILITY OF AI INFRASTRUCTURE

Not Optimized, Complex to Manage, Difficult to Scale Predictably



- Inflexible infrastructure silos that were never meant for the pace of AI
  - Constrained workload placement by system-level characteristics
  - Non-uniform performance across the data center
  - Unable to adapt to dynamic workload demands
  - Constrained capacity planning

DGX A100: THE UNIVERSAL AI SYSTEM

One System for Every AI Workload

Performance meets utility – analytics, AI training and inference all in one

Integrated Access to Unmatched AI Expertise

Fastrack AI transformation with DGXpert know-how and experience

Game-changing Performance for Innovators

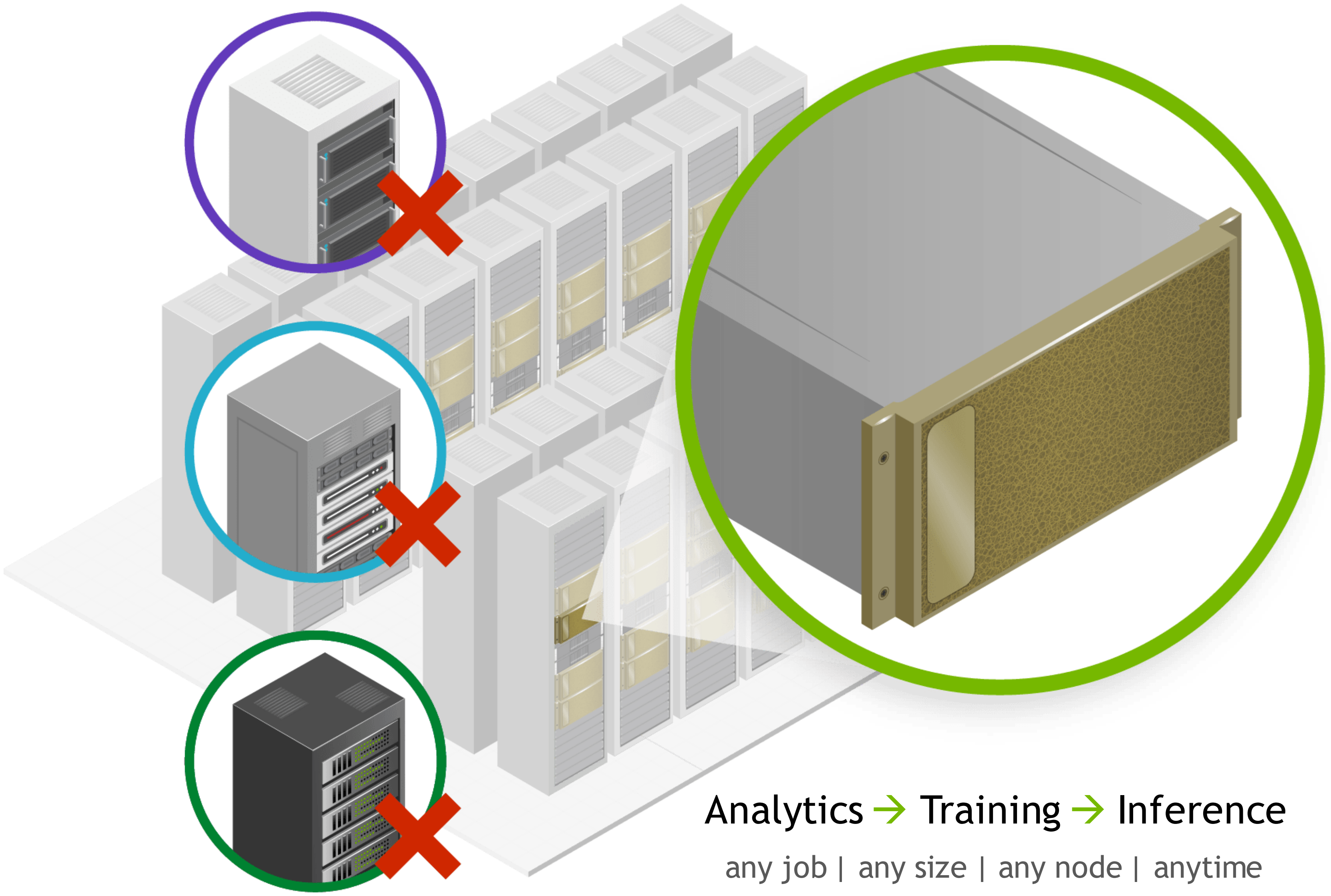
Fastest time-to-solution the world’s first 5 PFLOPS AI system, built on NVIDIA A100

Unmatched Data Center Scalability

Build leadership-class infrastructure that scales to keep ahead of demand

ONE SYSTEM FOR ALL AI INFRASTRUCTURE

AI Infrastructure Re-Imagined, Optimized, and Ready for Enterprise AI-at-Scale



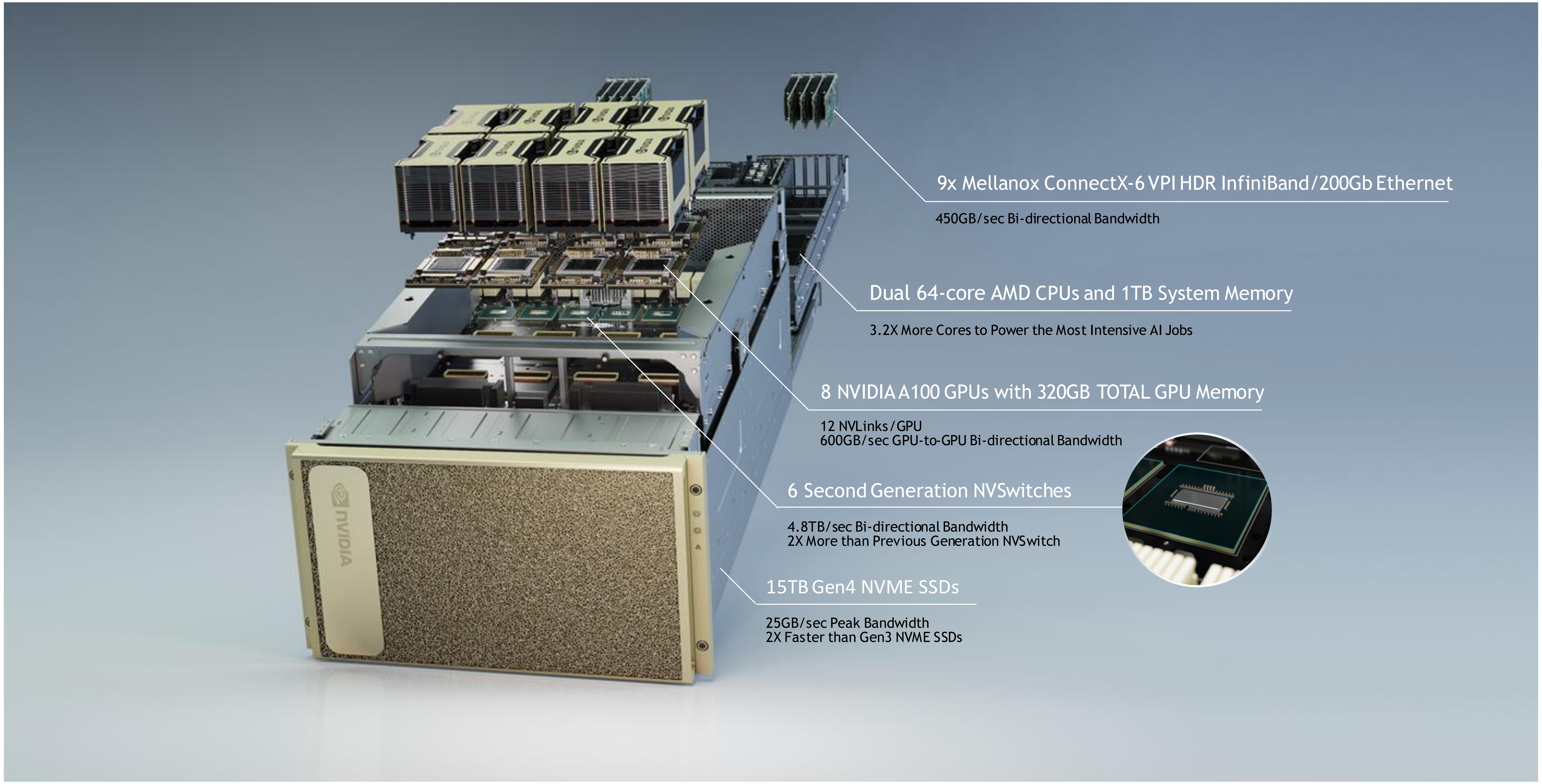
Analytics → Training → Inference

any job | any size | any node | anytime

Flexible AI infrastructure that adapts to the pace of enterprise

- One universal building block for the AI data center
- Uniform, consistent performance across the data center
- Any workload on any node - any time
- Limitless capacity planning with predictably great performance with scale

GAME-CHANGING PERFORMANCE FOR INNOVATORS



9x Mellanox ConnectX-6 VPI HDR InfiniBand/200Gb Ethernet

450GB/sec Bi-directional Bandwidth

Dual 64-core AMD CPUs and 1TB System Memory

3.2X More Cores to Power the Most Intensive AI Jobs

8 NVIDIA A100 GPUs with 320GB TOTAL GPU Memory

12 NVLinks/GPU
600GB/sec GPU-to-GPU Bi-directional Bandwidth

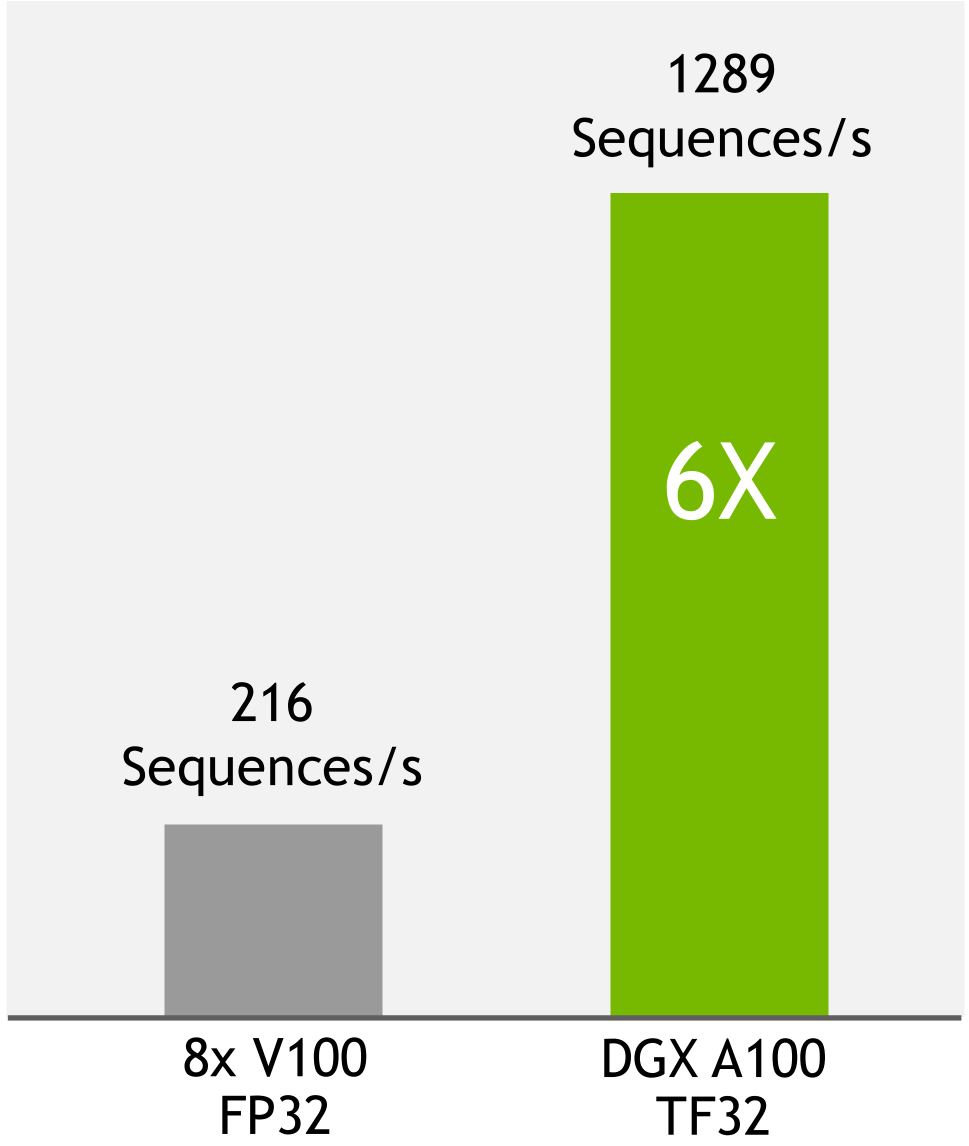
6 Second Generation NVSwitches

4.8TB/sec Bi-directional Bandwidth
2X More than Previous Generation NVSwitch

15TB Gen4 NVME SSDs

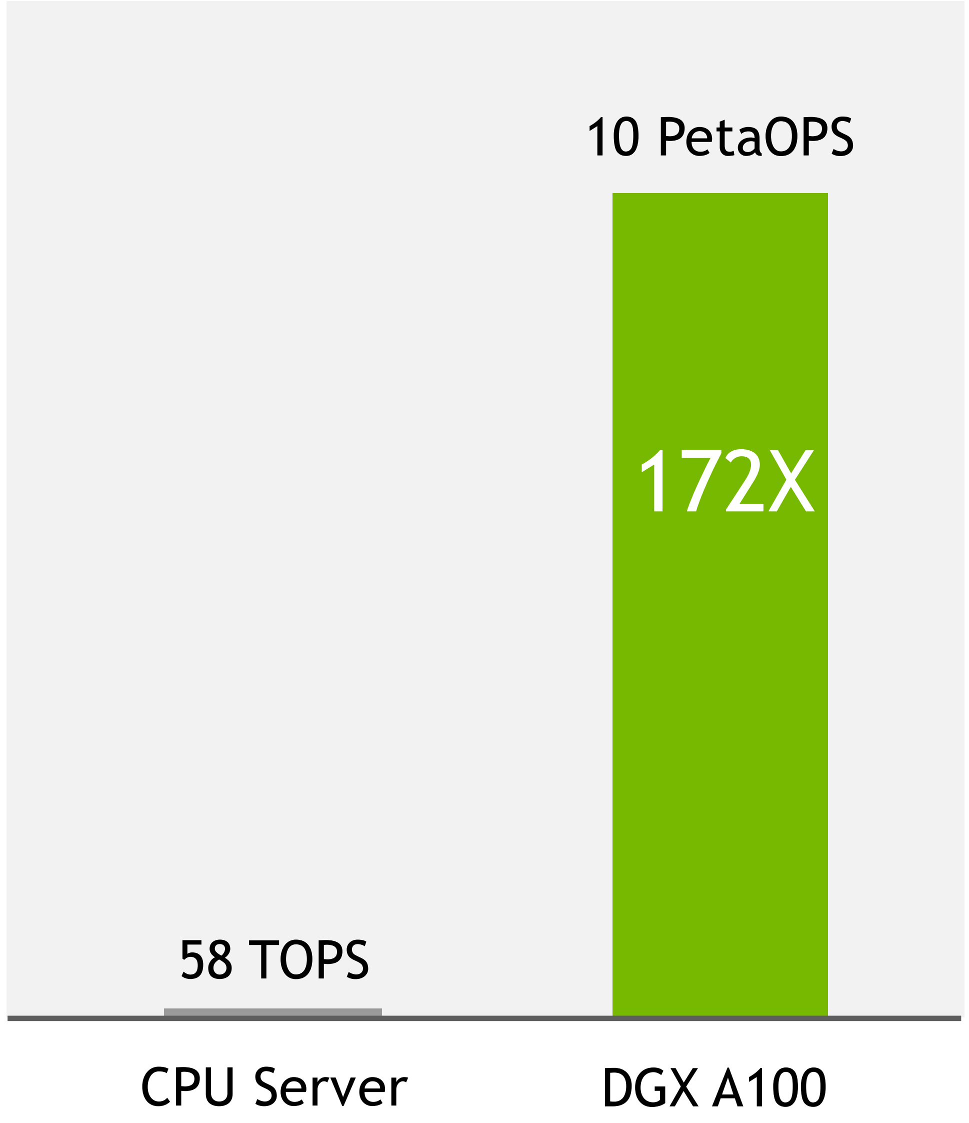
25GB/sec Peak Bandwidth
2X Faster than Gen3 NVME SSDs

DGX A100 PERFORMANCE



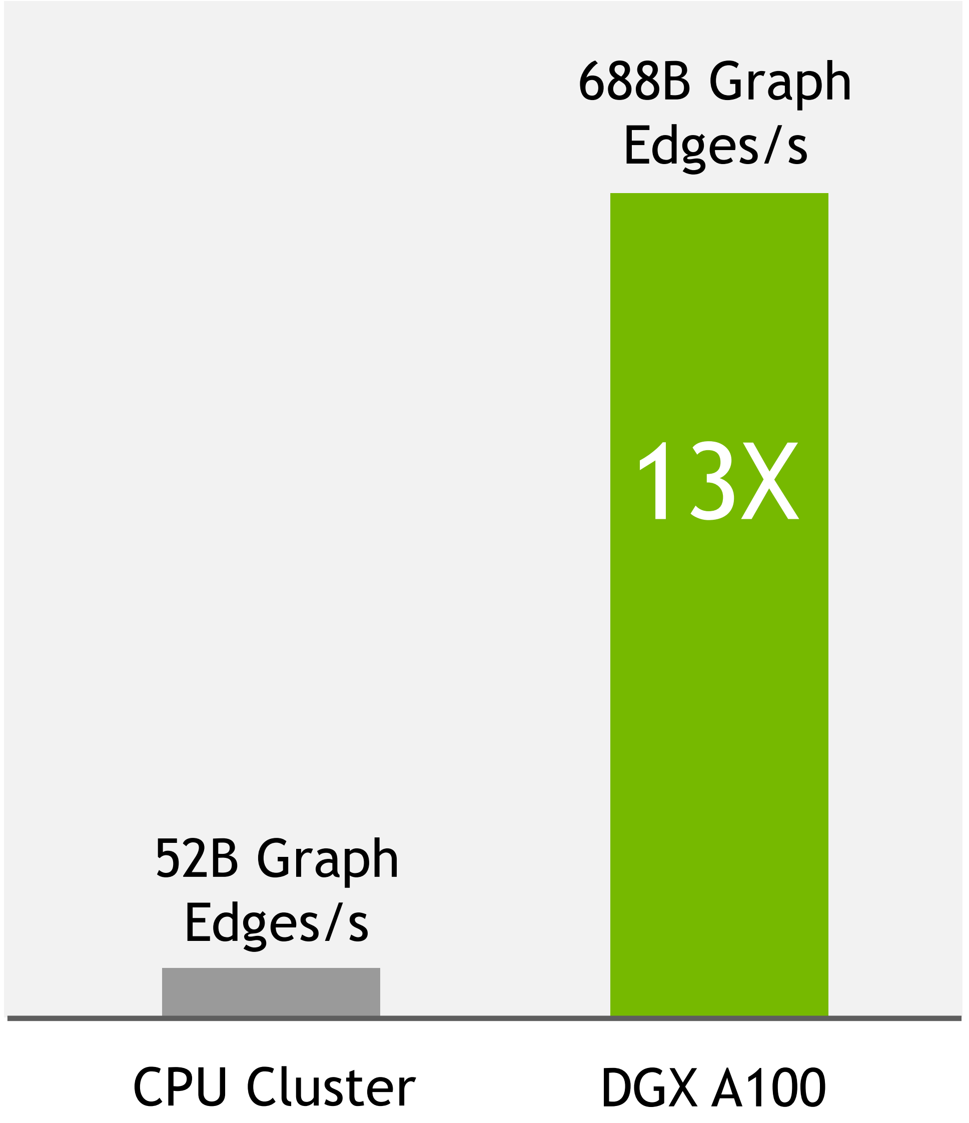
Training
NLP: BERT-Large

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 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



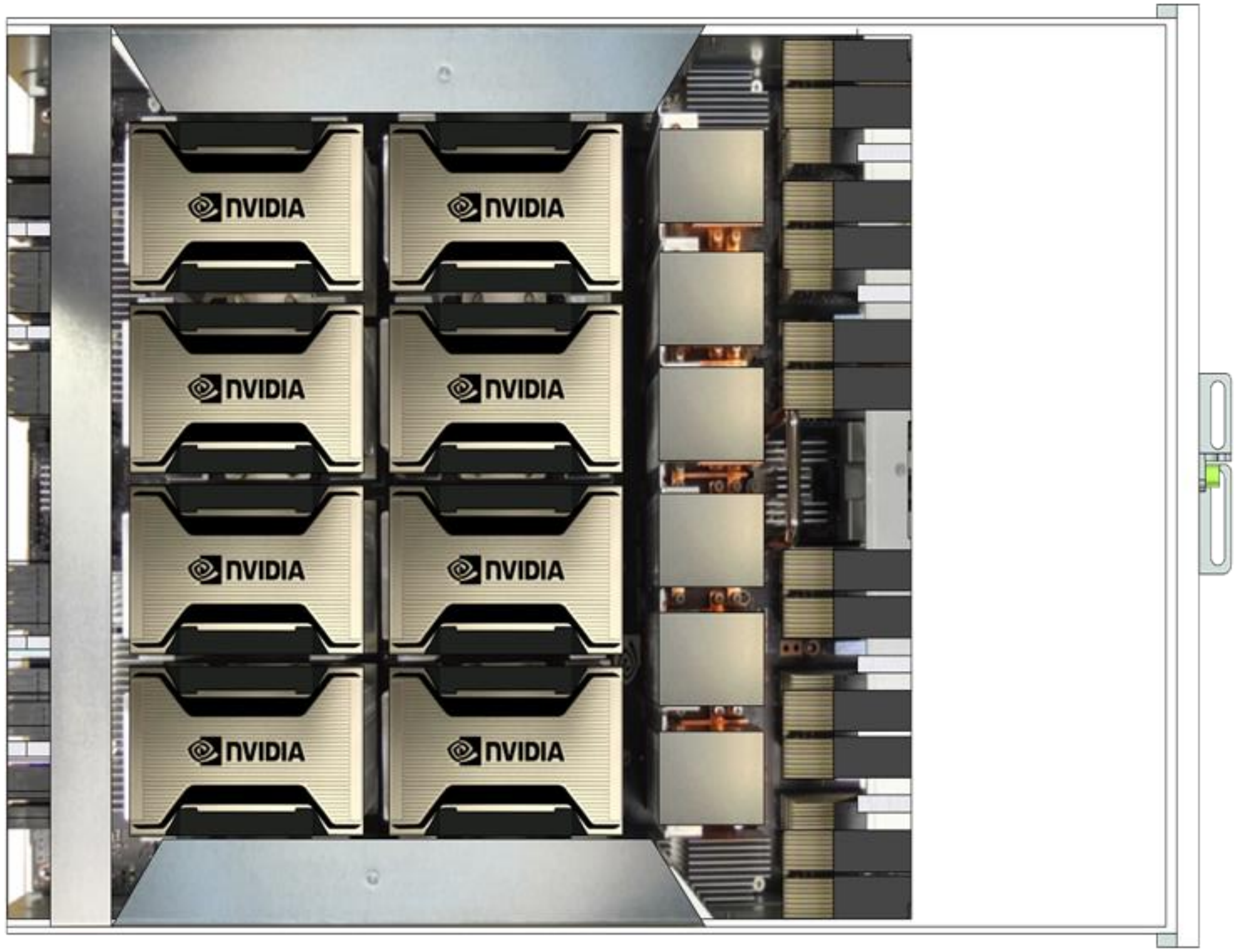
Analytics
PageRank

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

NEW FEATURES

DGX A100: NEW A100 GPUS AND 2X FASTER NVSWITCH

5 PetaFLOPS AI Performance



Eight new A100 Tensor Core GPUs/320GB total HBM2

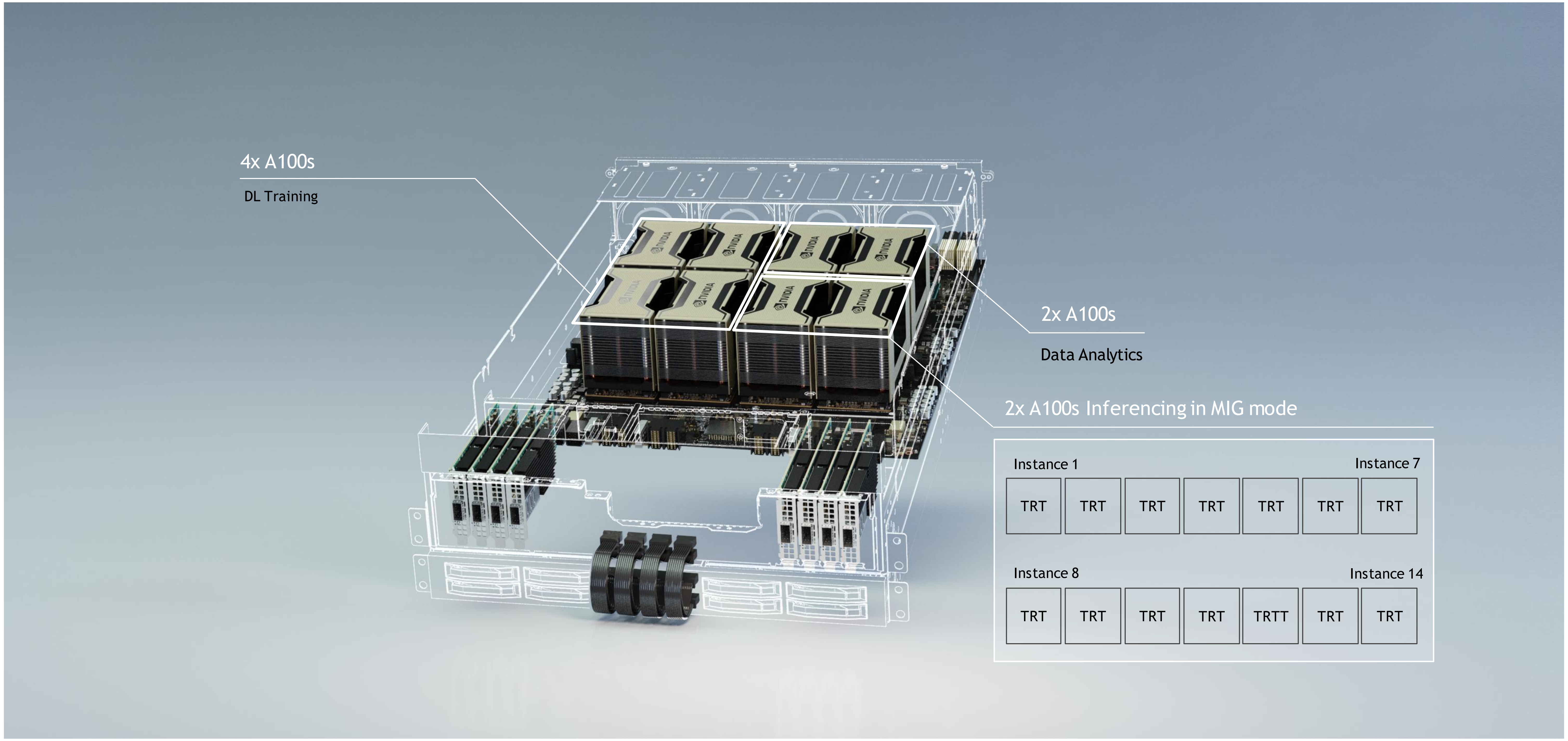
- Twelve NVLinks per GPU, 2x more than V100
- 600GB/s bi-directional bandwidth between any GPU pair
- ~10X PCIe Gen4 bandwidth with next-gen NVLink

All GPUs fully connected with six next-gen NVSwitch

- 4.8TB/s bi-directional bandwidth
- In one second we could transfer 426 hours of HD video
- Download HD video to 80K smartphones simultaneously

CONSOLIDATING DIFFERENT WORKLOADS ON DGX A100

One Platform for Training, Inference and Data Analytics



4x A100s

DL Training

2x A100s

Data Analytics

2x A100s Inferencing in MIG mode

Instance 1 Instance 7

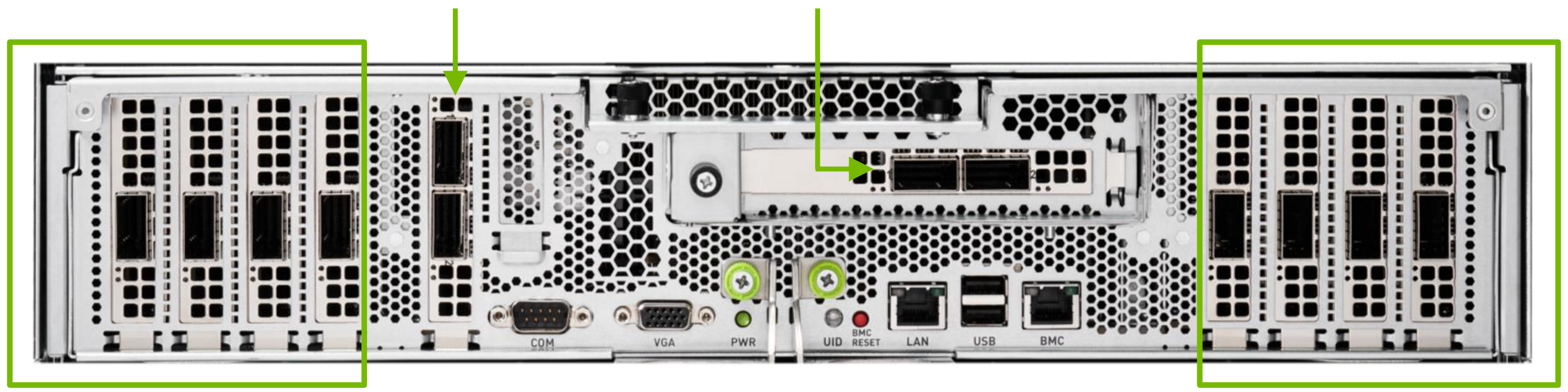
TRT TRT TRT TRT TRT TRT TRT

Instance 8 Instance 14

TRT TRT TRT TRT TRTT TRT TRT

UNMATCHED SCALABILITY WITH MELLANOX NETWORKING

Highest Network Throughput for Data and Clustering

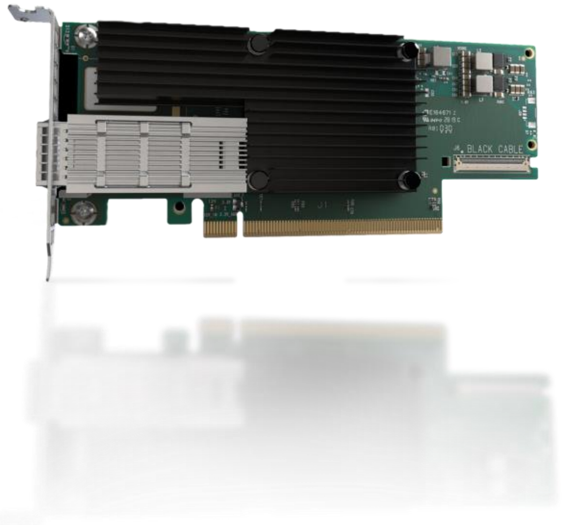


Storage Networking

Cluster Networking

Cluster Networking

Single-port CX-6 NIC



For clustering networking:

- Eight Mellanox single-port ConnectX-6
- Supporting HDR/HDR100/EDR InfiniBand default or 200GigE
- 450GB/sec total peak bandwidth

For data/storage networking:

- One Mellanox dual-port ConnectX-6
  - Supporting: 200/100/50/40/25/10Gb Ethernet default or HDR/HDR100/EDR InfiniBand
- One optional Dual-Port CX-6 available as add-on

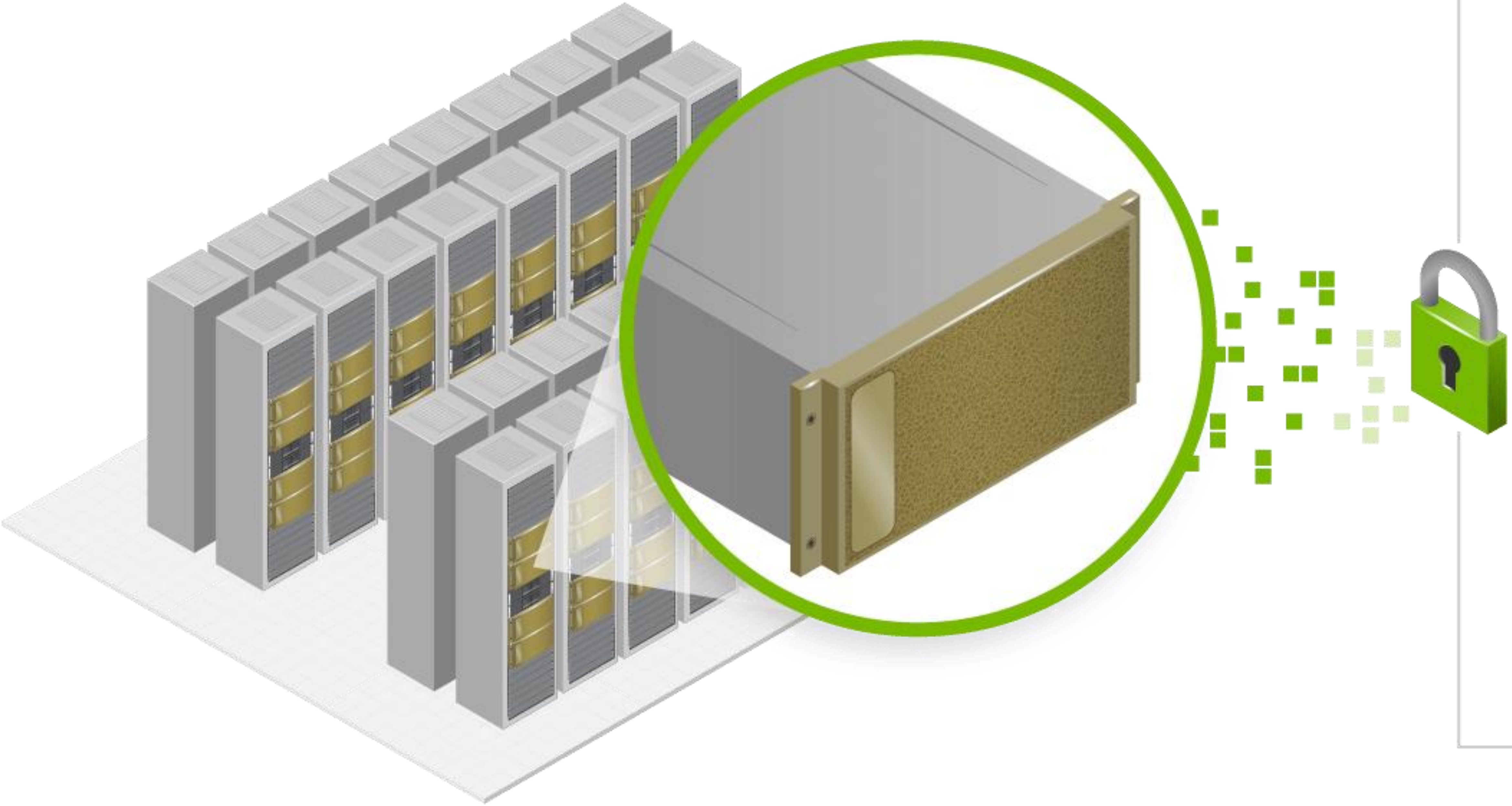
All I/O now PCIe Gen4, 2x performance increase over Gen3

Scale up multiple DGX A100 nodes with Mellanox Quantum Switch, the world’s smartest network switch

THE WORLD’S MOST SECURE AI SYSTEM FOR ENTERPRISE

Built-In Security: Multi-layered Defense for AI Infrastructure

DGX A100 delivers the most robust security posture for your AI enterprise

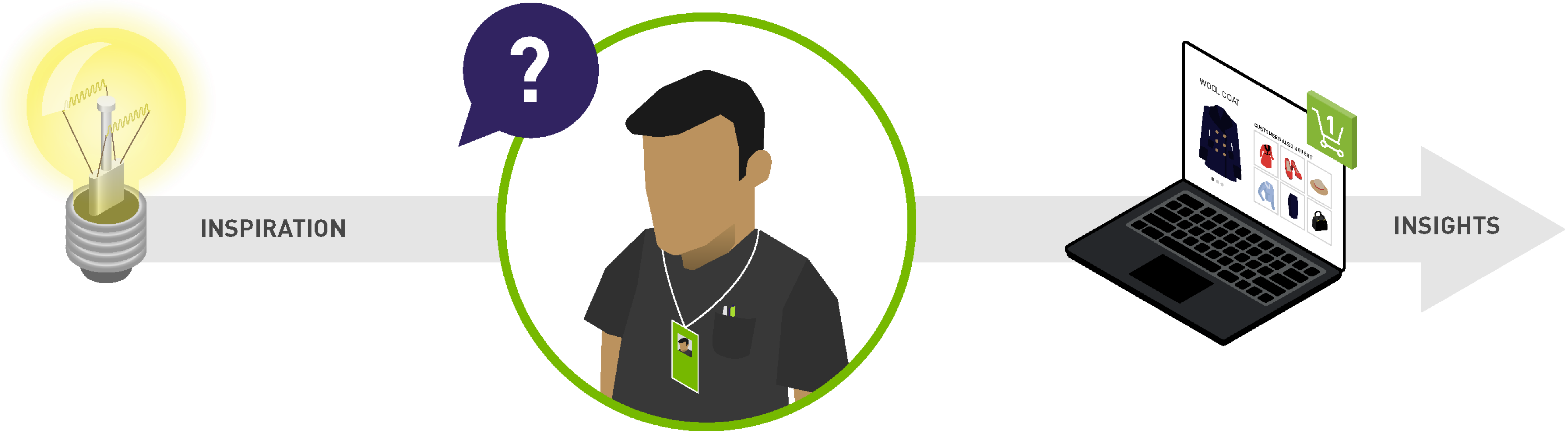


- Secure boot
- Self-Encrypted Drives (SED) to protect data at rest
- GPU Board
- CPU Board
- BMC

GPU Board, CPU Board, BMC: Secure Update of Firmware

INTRODUCING: NVIDIA DGXpert

With Every DGX system - Your Trusted Navigator in AI Transformation



14,000+
AI-Fluent Experts

DESIGN | PLAN | BUILD | TEST | DEPLOY | OPERATE | MONITOR

With you every step of the way - Included with every DGX system

DGX: DELIVERING AI FOR BUSINESS

Backed by 1000’s of Data Scientists, Engineers & SATURNV

Lifecycle Services →

Plan AI

- System Sizing
- Network Design
- Secure AI Guidance

Deploy

- HPL System testing
- Cluster Tools Setup
- System Runbook

Optimize

- DLI for new features
- App Code Reviews
- Technology Upgrades

Optimized Software →

- DL SW with TF32 Acceleration
- HPC SW Optimized with Tensor Core
- Data Analytics SW with RAPIDS

Highest Performance Systems →

Mellanox TECHNOLOGIES

NVIDIA DGX SUPERPOD WITH DGX A100

Unmatched data center scalability – deployed in under 3 weeks

Leadership-class AI infrastructure

- The blueprint for AI power and scale using DGX A100
- Infused with the expertise of NVIDIA’s AI practitioners
- Designed to solve the previously unsolvable
- Configurations start at 20 systems

NVIDIA DGX SuperPOD deployed in SATURNV

- 1,120 A100 GPUs
- 140 DGX A100 Systems
- 170 Mellanox 200G HDR switches
- 4 PB of high-performance storage
- 700 PFLOPS of power to train the previously impossible

Panel Discussion

Christopher Lamb, Vice President, Compute Software, NVIDIA
Rajeev Jayavant, Vice President, GPU Systems Engineering, NVIDIA

MORE THAN A SERVER – NVIDIA’S COMMITMENT TO DELIVERING AI SUCCESS

Backed by a Global Team of DGXperts

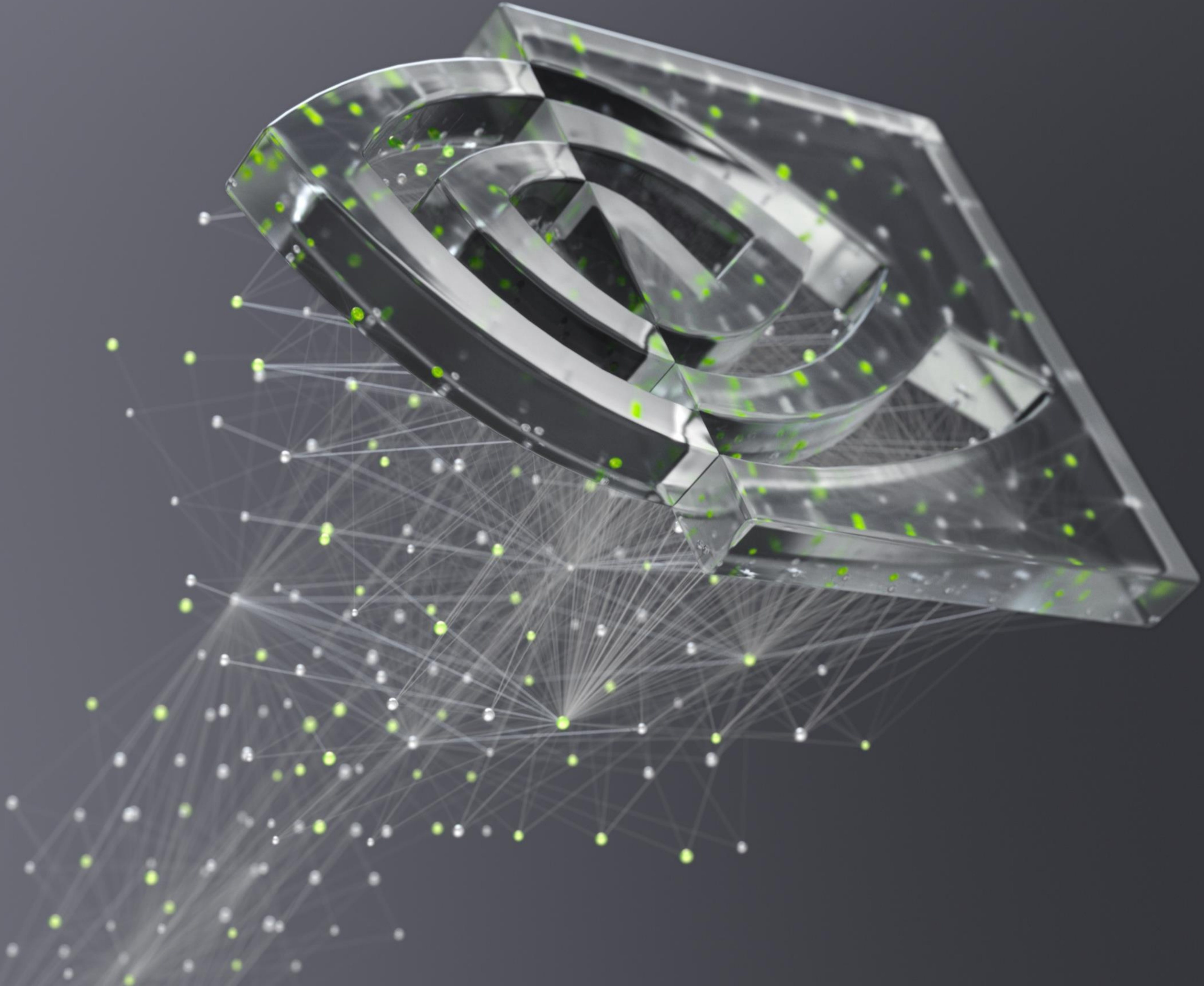
- 14,000+ of “AI-fluent” practitioners with a decade of experience
- Backed by SATURNV – world’s largest DGX proving ground

Fully-Optimized

- Full-stack solution, optimized at every layer: data, algorithms, models + compute, storage, networking, and more

Field-Proven

- Thousands of deployed AI systems and customers



NVIDIA