NVIDIA DGX™ A100 Systems
NVIDIA Mellanox Quantum™ HDR
200 Gb/s InfiniBand Smart Switches
Petabytes of All-Flash Storage
Modular, with 20 Node Scalable Units
Optimized Software from NVIDIA NGC™
Deployable in Weeks, Instead of Months

UNIFIED AI INFRASTRUCTURE

Today's Workload
Tomorrow's Workload
Next Week's Workload

INFERENCE
ANALYTICS
TRAINING

GAME/hyphen.caseCHANGING PERFORMANCE FOR INNOVATORS

NVIDIA DGX SUPERPOD: UNLOCKING AI AT SCALE FOR EVERY ENTERPRISE

ACCELERATE AI, ACCELERATE ROI
DEPLOY FLEXIBLE AI INFRASTRUCTURE THAT REINS IN CAPEX

25 developers the performance equivalent of
6 dedicated 28-core dual CPU servers each

THE MOST POWERFUL TOOL FOR DATA SCIENCE TEAMS
GIVE EVERY DEVELOPER THE POWER TO EXPLORE

ACCELERATION FOR EVERY AI WORKLOAD
RUN DATA ANALYTICS, TRAINING, AND INFERENCE WORKLOADS ON THE SAME SYSTEM

Multi-Instance GPU (MIG)
in a single DGX A100 gives:

5 DGX A100 Systems for AI Training and Inference
5 DGX A100 Systems for AI Training and Inference

1/10th THE COST
1/20th THE POWER
1/25th THE SPACE

THE BUILDING BLOCK OF THE AI DATA CENTER
AI INFRASTRUCTURE REIMAGINED, OPTIMIZED, AND READY FOR ENTERPRISE AI

500 GB/s Peak Bi-directional Bandwidth
10x NVIDIA® Mellanox® ConnectX-6
200 Gb/s Network Interface
3.2X More Cores to Power the Most Intensive AI Jobs
Dual 64-Core AMD Rome CPUs
2 TB RAM
4.8 TB/s Bi-directional Bandwidth
2X More than Previous-Generation NVSwitch
6x NVIDIA NVSwitches™
50 GB/s Peak Bandwidth
2X Faster than Gen3 NVME SSDs
30 TB Gen4 NVME SSD
Up to 640 GB Total GPU Memory
12 NVIDIA NVLinks™ per GPU
600 GB/s GPU-to-GPU Bi-directional Bandwidth
8x NVIDIA A100 Tensor Core GPUs

TRILLION Transistors
1 KILOMETER of Traces
5 PETAFLOPS of AI Performance
1 MILLION Drill Holes
30 THOUSAND Components
50 THOUSAND Connector Pins

DGX A100 INFRASTRUCTURE IS AGILE
Run any workload on any system for maximized utilization.

TRADITIONAL INFRASTRUCTURE IS CONSTRAINED
Infrastructure silos starve AI workloads or waste capacity.

Optimal
Over
Under

CLUSTER 1: TRAINING
CLUSTER 2: ANALYTICS
CLUSTER 3: INFERENCE

Unlock AI in Your Organization Today
www.nvidia.com/DGX100