The next generation artificial intelligence (AI) supercomputing infrastructure, providing the computational power necessary to train today’s state-of-the-art deep learning (DL) models and to fuel innovation well into the future. The AMAX NVIDIA DGX SuperPOD delivers groundbreaking performance and is designed to solve the world’s most challenging computational problems.

**Simplified AI**
Predictable performance, capacity and scaling

**Tested and Validated**
Reference architecture for the enterprise

**Fully Configured**
Solution ready to deploy and install in minutes

**Networking**
- 4x NVIDIA Mellanox QM8700
  - 2 Storage Network Switches
  - 2 Compute Network Switches

**Storage and Data Management**
- 1x DDN A1400X

**Compute**
- 2x NVIDIA's DGX A100 Systems

**Cluster Management**
- Bright Cluster Manager

**Support & Services**
- AMAX Professional Services

**DGX SUPERPOD SOLUTION FOR ENTERPRISE**
- 100-700 PFLOPS AI system
- 20-140 NVIDIA DGX A100 systems with NVIDIA BlueField DPUs
- 1-10PB high-performance storage
- 200Gbps NVIDIA networking fabric
Solving the Challenge of Large-Scale, Multi-Node AI Infrastructure

NVIDIA DGX SuperPOD is designed to tackle the most important challenges of AI at scale, delivering unmatched levels of multi-system training. Traditional large compute clusters are constrained by the complexity of scaling inter-GPU communications as configurations become larger and computation is parallelized over more and more nodes. This results in diminishing performance returns. DGX SuperPOD solves this scaling problem by optimizing every component in the system for the unique demands of multi-node AI infrastructure.

High-Performance Infrastructure in a Single Solution—Optimized for AI

NVIDIA DGX SuperPOD brings together a design-optimized combination of AI computing, network fabric, storage, and software. Its compute foundation is built on NVIDIA DGX™ A100, the universal system for all AI workloads, which provides unprecedented compute density, performance, and flexibility. NVIDIA DGX A100 systems, available with up to 640 gigabytes (GB) of total GPU memory each, feature the world’s most advanced accelerator, the NVIDIA A100 Tensor Core GPU, enabling enterprises to consolidate training, inference, and analytics in a unified, easy-to-deploy AI infrastructure.

NVIDIA DGX™ A100

System Features
- 3rd Gen Integrated AI System
- Uniform system for End-to-End Data Science & AI
- System memory: 1TB
- 6U Rackmount

GPU
- 8x NVIDIA Tensor Core A100
- Total GPU memory: 320GB or 640GB

Processor
2x AMD EPYC, 64 cores

Networking
- 8x Single-port Mellanox ConnectX-6 VPI
- 1x Dual-port Mellanox ConnectX-6 VPI
- 10/25/50/100/200 Gb/s Ethernet

Storage
15TB (4x 3.84TB) U.2 NMVe

Performance
5 PLFOPS of AI performance

DDN AI400X

System Features
- High performance GPU-optimized parallel file system
- Sequential read performance up to 48GB/s
- Sequential write performance up to 34GB/s
- Up to 3M IOPS per appliance

Controller Host Ports per Appliance
8x EDR/HDR100 InfiniBand or 100GbE

Drive Support
2.5’’ dual port NVMe drives
- 32TB, 64TB, 128TB, 256TB usable capacity configurations

Standard Software Features
- High performance parallel file system
- LUN mapping and masking, intelligent write striping, read QoS
- Port zoning detection
- Data integrity check/correction

Mellanox QM8700

Performance
- 40x HDR 200Gb/s ports in a 1U switch
- 880x HDR 100 Gb/s ports
- 16Tb/s aggregate switch throughput
- Sub-130ns switch latency

Optimized Design
- 1+1 redundant & hot swappable power
- N+1 redundant & hot-swappable fans
- 80 gold+ and energy star certified power supplies

Advanced Design
- Adaptive routing
- Congestion control
- Collective offloads
- VL mapping (VL2VL)

Bright Computing

- Deploying easily
- Automating the cluster build process and pre-checks everything
- Supporting heterogeneous environments, bringing all your apps and environments under one management platform; supports NGC containers
- Providing comprehensive monitoring
- Integrating with the public and private clouds
- Managing accelerators (GPUs, FPGAs, and IPUs)
- Optimizing the use of cluster resources
- Configuring and deploying HPC workload managers and Kubernetes
- Including two powerful user interfaces—command line and web-based graphical

• 3rd Gen Integrated AI System
• Unified system for End-to-End Data Science & AI
• System memory: 1TB
• 6U Rackmount

Copyright © 2022 AMAX. NVIDIA logo is a trademark or registered trademark of NVIDIA Corporation. DDN is a trademark or registered trademark of DataDirect Networks. Bright Computing is a trademark or registered trademark of Bright Computing, Inc. All trademarks are the properties of their respective owners. All rights reserved.