NVIDIA DGX SuperPOD: Scalable Infrastructure for AI Leadership

Reference Architecture

Featuring NVIDIA DGX A100 Systems

RA-09950-001 | October 2021

Document History

RA-09950-001

Version	Date	Authors	Description of Change
01	2020-05-14	Craig Tierney, Jeremy Rodriguez, Premal Savla, and Robert Sohigian	Initial release
02	2020-06-08	Craig Tierney, Premal Savla, Scott Ellis, and Robert Sohigian	Updated networking and other refinements
03	2020-06-20	Craig Tierney and Robert Sohigian	June 2020 TOP500 results and minor changes
04	2020-07-27	Ivan Goldwasser, Premal Savla, and Robert Sohigian	July 2020 MLPerf v0.7 results and minor changes
05	2020-11-16	Craig Tierney, Premal Savla, Scott Ellis, and Robert Sohigian	November 2020 TOP500/Green500 and networking updates.
06	2020-11-18	Craig Tierney, Premal Savla, Scott Ellis, and Robert Sohigian	Removed outdated information
07	2020-12-18	Craig Tierney, Michael Balint, Premal Savla, Scott Ellis, and Robert Sohigian	Updated software section and associated references
08	2021-02-12	Michael Balint and Premal Savla	Specified NVIDIA® Cumulus® Linux for Ethernet switch OS
09	2021-03-26	Craig Tierney and Robert Sohigian	Miscellaneous updates
10	2021-04-27	Michael Balint and Premal Savla	Bare metal for management, SU size updates
11	2021-06-05	Craig Tierney, Michael Balint, Premal Savla, and Robert Sohigian	Updates for NVIDIA Base Command™ Manager
12	2021-06-08	Craig Tierney Michael Balint, Scott Tracy, Premal Savla, and Robert Sohigian	More updates for Base Command Manager
13	2021-07-06	Sukwoo Kang, Cameron Foster, Alexey Gorbunov	Updates for In-band/Out-of-band management network
14	2021-07-16	Craig Tierney, Premal Savla, Sukwoo Kang, and Robert Sohigian	Updates for Base Command Manager and other areas
15	2021-10-12	Premal Savla and Robert Sohigian	Updates to UFM cable counts

Abstract

The NVIDIA DGX SuperPOD™ with NVIDIA DGX™ A100 systems is 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 future innovation. The DGX SuperPOD delivers groundbreaking performance, deploys in weeks as a fully integrated system, and is designed to solve the world's most challenging computational problems.

This DGX SuperPOD reference architecture (RA) is the result of codesign between DL scientists, application performance engineers, and system architects to build a system capable of supporting the widest range of DL workloads. Selene, a DGX SuperPOD used for research computing at NVIDIA, earned the sixth spot on the June 2021 TOP500 list and is the fastest commercial supercomputer. The University of Florida deployed a DGX SuperPOD that was ranked second on the June 2021 Green500 list. Also in June 2021, the DGX SuperPOD and NVIDIA A100 GPU set records for “At Scale'' and “Per Chip” performance among commercially available solutions across all benchmarks in MLPerf 1.0 Training[1]. The latest results are available on the NVIDIA Data Center Deep Learning Product Performance webpage.

This RA design introduces compute building blocks called scalable units (SU) allowing for the modular deployment of a full 140-node DGX SuperPOD, which can further scale to hundreds of nodes. The DGX SuperPOD design includes NVIDIA networking switches, software, storage, and NVIDIA NGC™ optimized applications.

The DGX SuperPOD RA has been deployed at customer sites around the world, as well as being leveraged within infrastructure that powers NVIDIA research and development in autonomous vehicles, natural language processing (NLP), robotics, graphics, HPC, and other domains. Organizations wanting to deploy their own supercomputing infrastructure should use the NVIDIA DGX SuperPOD Solution for Enterprise, which offers the DGX SuperPOD RA deployed in a turnkey infrastructure solution along with a full lifecycle of advanced services from planning to design to deployment to on-going optimization.

[1] Per-Accelerator Records with eight NVIDIA A100 GPUs per system: BERT: 1.0-1033 | DLRM: 1.0-1037 | Mask R-CNN: 1.0-1057 | Resnet50 v1.5: 1.0-1038 | SSD: 1.0-1038 | RNN-T: 1.0-1060 | 3D-Unet: 1.0-1053 | MiniGo: 1.0-1061. Max Scale Records using NVIDIA DGX A100 systems: BERT: 1.0-1077 | DLRM: 1.0-1067 | Mask R-CNN: 1.0-1070 | Resnet50 v1.5: 1.0-1076 | SSD: 1.0-1072 | RNN-T: 1.0-1074 | 3D-Unet: 1.0-1071 | MiniGo: 1.0-1075. MLPerf name and logo are trademarks. See www.mlcommons.org for more information.

Contents

DGX SuperPOD with DGX A100 Systems

The compute requirements of AI researchers continues to increase as the complexity of DL networks and training data grow exponentially. Training in the past has been limited to one or a few GPUs, often in workstations. Training today commonly uses dozens, hundreds, or even thousands of GPUs for evaluating and optimizing different model configurations and parameters. In addition, organizations have many AI researchers that must train numerous models simultaneously. Systems at this massive scale may be new to AI researchers, but these installations have been a hallmark of the world's most important research facilities and academia, fueling innovation that propels scientific endeavors of almost every kind.

The supercomputing world is evolving to fuel the next industrial revolution, which is driven by how massive computing resources can be brought together to solve mission critical business problems. NVIDIA is ushering in a new era in which enterprises can deploy world-record setting supercomputers using standardized components in weeks.

Designing and building scaled computing infrastructure for AI requires an understanding of the computing goals of AI researchers to build fast, capable, and cost-efficient systems. Developing infrastructure requirements can be difficult because the needs of research are often an ever-moving target and AI models, due to their proprietary nature, often cannot be shared with vendors. Additionally, crafting robust benchmarks that represent the overall needs of an organization is a time-consuming process. Going it alone, and designing a supercomputer is simply not a feasible option.

It takes more than just many GPU nodes to achieve optimal performance across a variety of model types. To build a flexible system capable of running a multitude of DL applications at scale, organizations need a well-balanced system, which at a minimum incorporates:

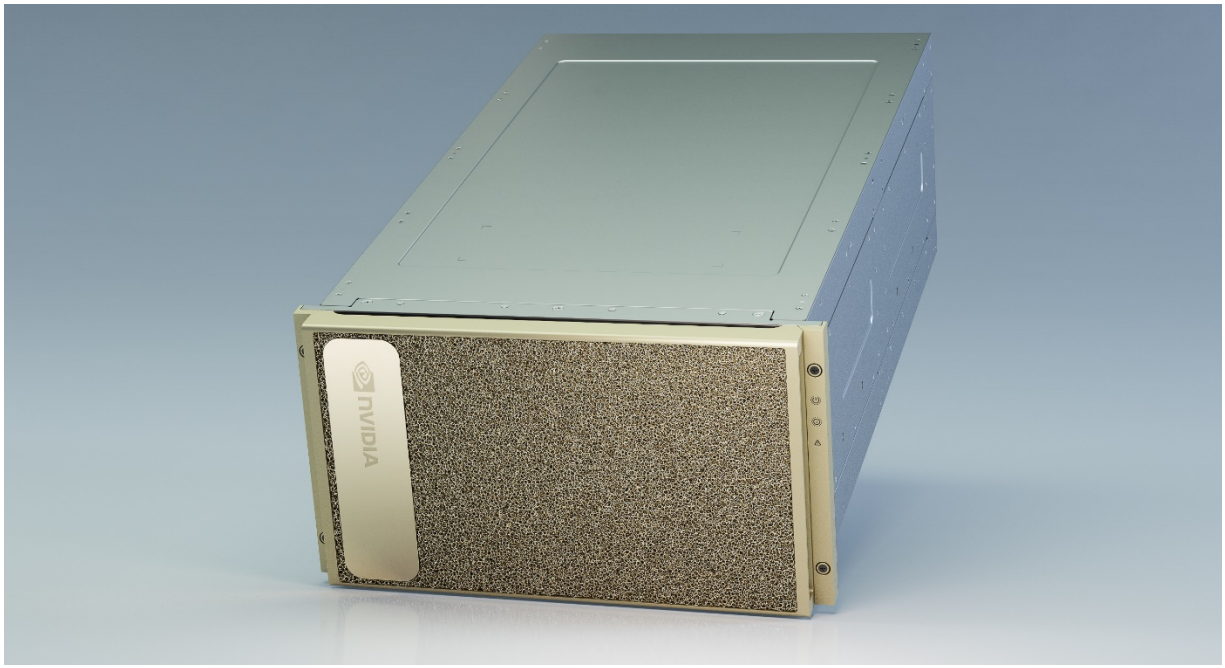
- Powerful nodes with many GPUs, a large memory footprint, and fast connections between the GPUs for computing to support the variety of DL models in use.
- A low-latency, high-bandwidth, HDR InfiniBand interconnect designed with the capacity and topology to minimize bottlenecks.
- A storage hierarchy that can provide maximum performance for the needs of various dataset structures.

These requirements, weighed with cost considerations to maximize overall value, can be met with the design presented in this paper—the NVIDIA DGX SuperPOD.

NVIDIA DGX A100 System

The NVIDIA DGX A100 system (Figure 1) is the universal system for all AI workloads, offering unprecedented compute density, performance, and flexibility in the world's first 5 petaFLOPS AI system.

Figure 1. DGX A100 system



Features

The features of the DGX SuperPOD are described in Table 1.

Table 1. 140-node DGX SuperPOD features

Component	Technology	Description
Compute nodes	NVIDIA DGX A100 System with eight 80 GB GPUs	• DGX operating system (DGX OS) • 1120 DGX A100 SXM4 GPUs • 89.6 TB of HBM2 memory • 336 AI PFLOPS using Tensor Cores • 280 TB System RAM • 4.4 PB local NVMe • 600 GBps NVIDIA NVLink® bandwidth per GPU • 4.8 TBps total NVIDIA NVSwitch™ bandwidth per node
Compute fabric	NVIDIA Quantum™ QM8790 HDR 200 Gb/s InfiniBand Smart Switch	Full fat-tree network built with eight connections per DGX A100 system
Storage fabric		Fat-tree network with two connections per DGX A100 system
Compute/Storage Fabric Management	NVIDIA® Unified Fabric Manager (NVIDIA UFM®) Enterprise	NVIDIA UFM combines enhanced, real-time network telemetry with AI-powered cyber intelligence and analytics to manage scale-out InfiniBand data centers.
In-band management network	NVIDIA Spectrum™ SN4600C switch running NVIDIA Cumulus® Linux	Two connections per DGX A100 system
Out-of-band management network	NVIDIA AS4610 switch running Cumulus Linux	One connection per DGX A100 system
Management System	NVIDIA Base Command™ Manager	Software tools for deployment and management of SuperPOD nodes and resources
DGX SuperPOD software stack	NVIDIA Magnum IO™ Technology	Suite of library technologies that optimize GPU communication performance
	NVIDIA CUDA-X™ Technology	A collection of libraries, tools, and technologies that maximize application performance on NVIDIA GPUs
User runtime environment	NVIDIA NGC	Containerized DL and HPC applications, optimized for performance
	Slurm	Orchestration and scheduling of multi-GPU and multinode jobs

Design Requirements

The DGX SuperPOD is designed to minimize system bottlenecks and maximize performance for the diverse nature of AI and HPC workloads. To do so, this design provides:

- A modular architecture constructed from SUs. Multiple SUs are connected to create one system that supports many users running diverse AI workloads simultaneously.
- A hardware and software infrastructure built around the DGX SuperPOD that enables distributed DL applications to scale across hundreds of nodes optimally without virtualization.
- The ability to quickly deploy and update the system. Leveraging the RA enables data center staff to develop a full solution with fewer design iterations.
- Management and monitoring services configured for high availability (HA).

Compute Fabric

The compute fabric must be capable of scaling from hundreds to thousands of nodes while maximizing performance of DL communication patterns. To make this possible:

- SUs are connected in a rail optimized, full fat-tree topology, maximizing the InfiniBand network capability for the DGX A100 systems.
- Multiple DGX SuperPOD clusters can be connected to create even larger systems with thousands of nodes.
- The fabric supports adaptive routing[1].

Storage Fabric

The storage fabric must provide high-throughput access to shared storage. The described storage fabric should:

- Provide single node bandwidth more than 40 GBps.
- Maximize storage access performance from a single SU.
- Leverage remote direct memory access (RDMA) communications for the fastest, low-latency data movement.
- Provide additional connectivity to shared storage between the DGX SuperPOD and other resources in the data center.
- Allow for training of DL models that require peak I/O performance, exceeding 16 GBps (2 GBps per GPU) directly from remote storage.

[1] Contact a representative from NVIDIA to learn more about configuring these technologies.

Other Considerations

This RA is not a complete blueprint for the installation and operation of the DGX SuperPOD. Standard IT best-practices should be applied.

Areas to consider are:

- Local site integration. user authentication, NFS, Firewall, NTP, and so on.
- Additional HA and redundancy features. DNS round-robin or load balancing expected to be provided by local infrastructure.
- Information assurance/security. File scanning, access controls.
- Backup/restore and disaster recovery.

The DGX SuperPOD RA does not provide specific guidance on these items. However, if you have questions about these topics, or others not explicitly discussed in the RA, contact NVIDIA.

SuperPOD Architecture

The basic building block for the DGX SuperPOD is the SU, which consists of 20 DGX A100 systems (Figure 2). This size optimizes both performance and cost while still minimizing system bottlenecks so that complex workloads are well supported. A single SU is capable of 48 AI PFLOPS.

Figure 2. DGX A100 SU

Note: The rack elevations in this paper are based on a DGX SuperPOD deployed at NVIDIA. The number of nodes per rack can be modified as needed based on local data center restraints.

The DGX A100 systems have eight HDR (200 Gbps) InfiniBand host channel adapters (HCAs) for compute traffic. Each pair of GPUs has a pair of associated HCAs. For the most efficient network, there are eight network planes, one for each HCA of the DGX A100 system that connects using eight leaf switches, one per plane. The planes are interconnected at the second level of the network through spine switches. Each SU has full bisection bandwidth to ensure maximum application flexibility.

Each SU has a dedicated management rack. The leaf switches are centralized in the management rack. Other equipment for the DGX SuperPOD, such as the second-level spine switches or management servers, could be in the empty space of a SU management rack or separate rack depending on the data center layout.

Details about SUs are covered in the following sections.

Network Architecture

The DGX SuperPOD has four networks:

- Compute fabric. Connects the eight NVIDIA HDR 200 Gb/s ConnectX®-6 HCAs from each DGX A100 system 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 two 100 Gbps ports on the DGX A100 system to connect to dedicated Ethernet switches.
- Out-of-band management. Connects the baseboard management controller (BMC) port of each DGX A100 system to additional Ethernet switches.

Network connections to the DGX A100 system are shown in Figure 3.

Figure 3. Network connections for DGX A100 system

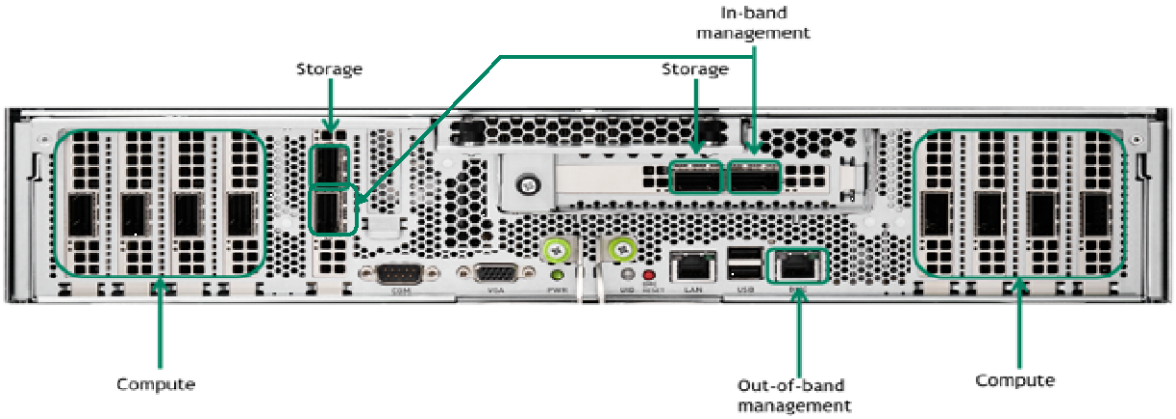


Table 2 shows an overview of the connections, with details provided in the following sections.

Table 2. Number of DGX SuperPOD network connections

Component	InfiniBand		Ethernet	
	Compute	Storage	In-Band	Out-of-Band
DGX A100 systems	1120	280	280	140
Management servers	0	4	26	13
Storage system[1]	Varies	Varies	Varies	Varies
UFM appliances	2	2	2	2
1. The number of storage system connections will depend on the system to be installed.				

Compute Fabric

The compute fabric design maximizes performance for typical communications traffic of AI workloads, as well as providing some redundancy in the event of hardware failures and minimizing cost.

The InfiniBand switches are categorized as:

- Leaf. There are eight leaf switches for each SU. The DGX A100 systems in the SU have a connection to each leaf switch. The fabric is rail-optimized, meaning that all the same HCAs from each system are connected to the same leaf switch. This rail-optimized organization is critical for maximizing DL training performance.
- Spine group (SG). An SG of ten QM8790 switches is used to optimize the fabric. Eight SGs are required as there are eight InfiniBand modules per DGX A100 system.
- Core group (CG). A CG of fourteen QM8790 switches is used to connect the SGs. There are two CGs for a 140-node deployment.

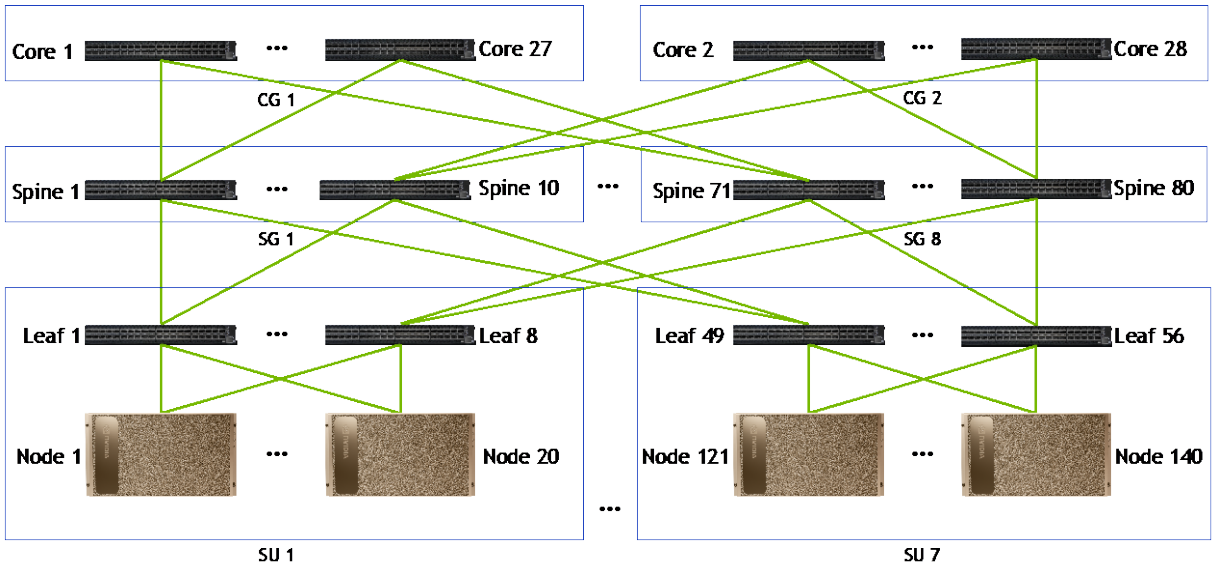
As shown in Figure 4, the first leaf switch from each SU connects to each switch in SG1, the second leaf switch from each SU connects to each switch in SG2, and so on.

A third layer of switching is required to complete the fat-tree topology:

- Odd switches from each of the eight SG fans out to each switch in CG1 (odd).
- Even switches from each of the eight SG fans out to each switch in CG2 (even).

For a full 140-node DGX SuperPOD, the design is rail optimized through both the leaf and spine levels—each InfiniBand HCA on a DGX A100 system is connected to its own fat tree topology. In all DGX SuperPOD deployments, the UFM appliance will be connected to two predetermined spine switch ports.

Figure 4. Compute fabric topology for a 140-node DGX SuperPOD



As shown in Figure 5, building DGX SuperPOD configurations with fewer than 80 nodes is simpler as the third layer of switching is not required.

Figure 5. Compute fabric topology for 80-node DGX SuperPOD

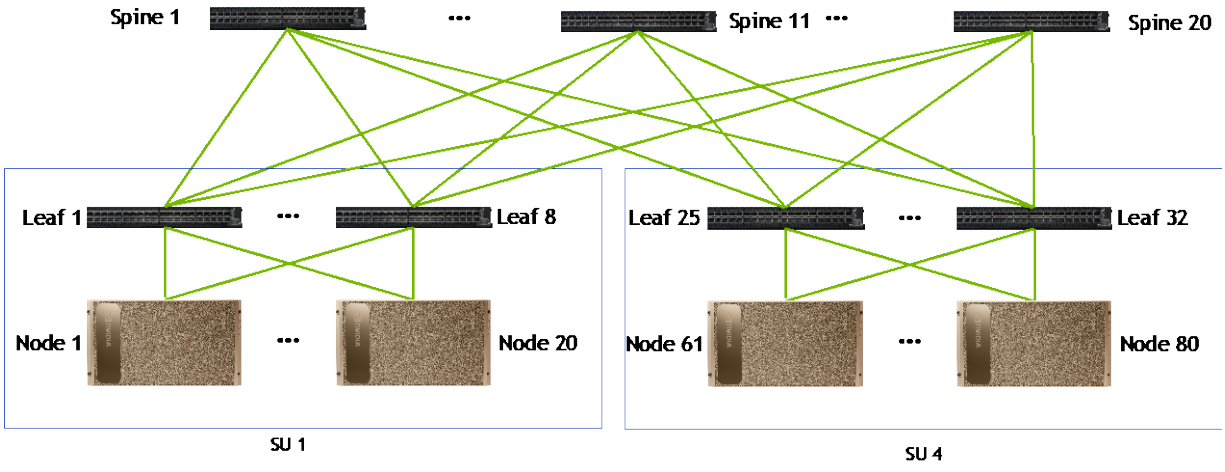


Table 3 shows the switch and cable count for different-sized systems.

Table 3. Compute fabric switch and cable counts

Nodes	SUs	QM8790 Switches			Cables		
		Leaf	Spine	Core	Leaf	Spine[1]	Core
20 (Single SU)	1	8	5		160	164	
40	2	16	10		320	324	
60	3	24	20		480	484	
80	4	32	20		640	644	
120	6	48	80	24	960	964	960
140 (DGX SuperPOD)	7	56	80	28	1120	1124	1120
1. UFM Appliance is connected to two different spine switches.							

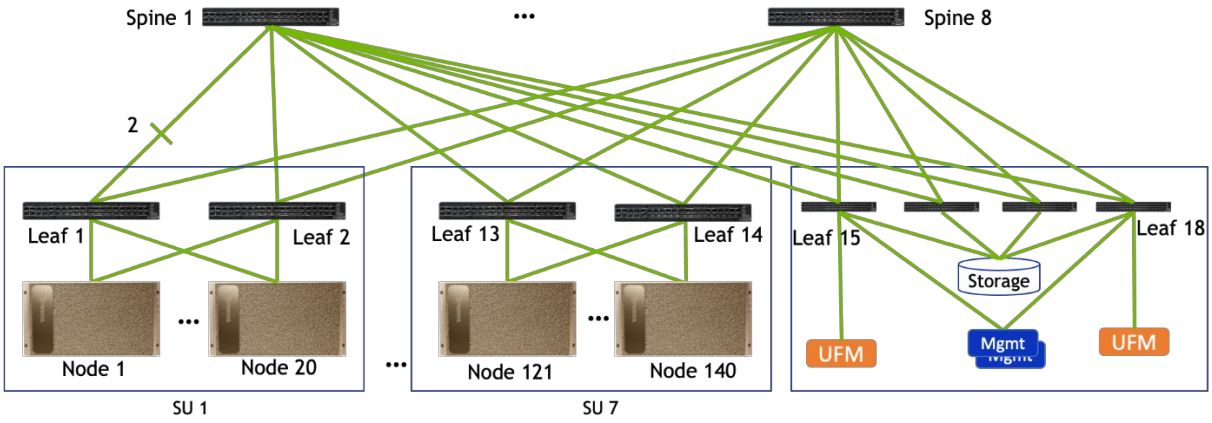
The compute fabric uses NVIDIA Quantum QM8790 switches (Figure 6).

Figure 6. QM8790 switch

Storage Fabric

The storage fabric employs an InfiniBand network fabric that is essential to maximum bandwidth (Figure 7). This is because the I/O per-node for the DGX SuperPOD must exceed 40 GBps. High-bandwidth requirements with advanced fabric management features, such as congestion control and adaptive routing, provide significant benefits for the storage fabric.

Figure 7. Storage fabric topology for 140-node system



The storage fabric also uses QM8790 switches. The storage fabric slightly oversubscribed, typically 5:4, for the leaf switches connecting the DGX A100 systems. This network topology offers a good

balance between performance and cost. The design detailed in Table 4 is based on storage servers requiring eight ports per SU. This may vary depending on the storage architecture and the storage performance requirements of a given deployment.

Table 4. Storage fabric counts

Nodes	SUs	Storage Ports	QM8790 Switches		Cables		
			Leaf	Spine	To-Node	To-Storage[1]	Spine
20	1	24	4	2	40	36	64
40	2	40	6	4	80	52	96
60	3	40	8	4	120	52	128
80	4	56	12	8	160	68	192
120	6	80	16	8	240	92	256
140	7	80	18	8	280	92	288
1. Includes connection to management servers and UFM.							

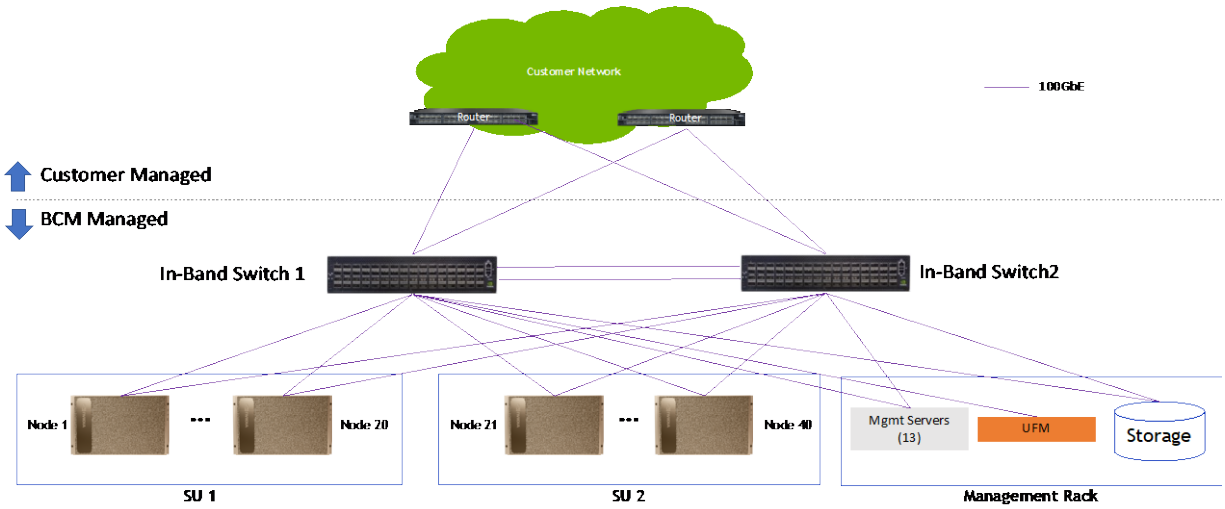
In-Band Management Network

The in-band Ethernet network has several important functions:

- Connects all the services that manage the cluster.
- Enables access to the home filesystem and storage pool.
- Provides connectivity for in-cluster services such as Slurm, and to other services outside of the cluster such as the NGC registry, code repositories, and data sources.

The in-band Ethernet design maximizes the stability, performance, and ease of management by using Underlay network, as well as providing some redundancy in the event of hardware failures. As shown in Figure 8.

Figure 8. In-band Ethernet topology for one or two SUs



There are two uplinks from each switch to a data center distribution router/switch. Connectivity to external resources and to the Internet are routed through the core router/switch. It is recommended that connectivity to the core data center router/switches follow data center best practices for connectivity. Additional routing and firewall equipment may be required and must be considered for deployment.

For three or more SUs, the second layer of switching (spine levels) is required to interconnect all leaf switches in a full-mesh topology.

The in-band network is built using NVIDIA SN4600 100 GbE switches (Figure 9) running Cumulus Linux.

Figure 9. SN4600 switch

Table 5 shows the switch count for different-sized systems.

Table 5. In-band Ethernet switch counts

Nodes	SUs	Leaf Switches	Spine Switches
20 (Single SU)	1	2	0
40	2	2	0
60	3	4	2
80	4	4	2
120	6	6	2
140 (DGX SuperPOD)	7	6	2

Out-of-Band Management Network

The out-of-band Ethernet network is used for system management via the BMC and provides connectivity to manage all networking equipment. Out-of-band management is critical to the operation of the cluster by providing low usage paths that ensure management traffic does not conflict with other cluster services.

The uplinks of the out-of-band Ethernet switches can be connected to in-band leaf switches or connected to the Customer out-of-band network. All Ethernet switches are connected through serial connections to existing console servers in the data center. These connections provide a means of last-resort-connectivity to the switches in the event of a network failure.

The out-of-band management network is based on NVIDIA AS4610 1 GbE switches (Figure 10). As with the in-band switches, it also runs Cumulus Linux.

Figure 10. AS4610 switch

Table 6 shows the switch count for different-sized systems.

Table 6. Out-of-band Ethernet switch counts

Nodes	SUs	Leaf Switches
20 (Single SU)	1	2
40	2	3
60	3	4
80	4	5
120	6	8
140 (DGX SuperPOD)	7	10

Storage Architecture

In the DGX SuperPOD, there are many types of data that require storage and retrieval. The majority of access—in both time and quantity—is of the data used to training and validate models, initialize simulations, and visualize results. The size of this can vary between organizations from Terabytes, to Petabytes, or even Exabytes. Protecting data like this is often done via replication to other storage tiers.

Source code, container definition scripts, data processing scripts, and other critical files need to be stored. These files, which can change frequently, are protected through more traditional techniques such as snapshots and backup/restore systems. For providing an optimized platform for critical files, a highly available NFS appliance should be used with a DGX SuperPOD configuration. The appliance will host the home filesystem for the users and provide a place for administrators to store any data necessary for managing the system. To limit dependencies and maximize system reliability, the appliance should not be connected to external components.

Storage and access of model data for training of DL models is unique due to its performance and access requirements. Training performance can be limited by the rate at which data can be read and reread from storage. The key to performance is the ability to read data multiple times. The closer the data are cached to the GPU, the faster they can be read. Storage architecture and design must consider the hierarchy of different storage technologies, either persistent or nonpersistent, to balance the needs of performance, capacity, and cost.

Table 7 documents the storage caching hierarchy. Depending on data size and performance needs, each tier of the hierarchy can be leveraged to maximize application performance.

Table 7. DGX SuperPOD storage and caching hierarchy

Storage Hierarchy Level	Technology	Total Capacity	Performance
RAM	DDR4	2 TB	> 200 GBps per node
Internal storage	NVMe	30 TB	> 55 GBps per node
High-speed storage	Varies	Varies depending on specific needs	Depends

Caching data in local RAM provides the best performance for reads. This caching is transparent after the data are read from the filesystem. However, the size of RAM is limited and less cost effective than other storage and memory technologies. Local NVMe storage is a more cost-effective way to provide caching close to the GPUs. However, manually replicating datasets to the local disk can be tedious. While there are ways to leverage local disks automatically (for example, `cachefilesd` for NFS filesystems), not every network filesystem provides a method to do so.

The performance of the high-speed storage in Table 5 is not provided because the answer greatly depends on the types of models that are being trained and the size of the dataset.

In Table 8, the storage performance requirements have been broken into three categories; good, better, and best.

Table 8. Storage performance requirements

Performance Level	Work Description	Dataset Size
Good	Natural language processing (NLP)	Most all datasets fit in cache
Better	Image processing with compressed images, ImageNet/ResNet-50	Many to most datasets can fit within the local node's cache
Best	Training with 1080p, 4K, or uncompressed images, offline inference, ETL	Datasets are too large to fit into cache, massive first epoch I/O requirements, workflows that only read the dataset once

Based on these descriptions, Table 9 provides guidelines for the necessary performance.

Table 9. Guidelines for storage performance

Performance Characteristic[1]	Good (GBps)	Better (GBps)	Best (GBps)
140 node aggregate system read	50	140	450
140 node aggregate system write	20	50	225
Single SU aggregate system read	6	20	65
Single SU aggregate system write	2	6	20
Single node read	2	4	20
Single node write	1	2	5

1. Achieving these performance characteristics may often require the use of optimized file formats, such as reading from data formats such as TFRecord, RecordIO, or LMDB.

High-speed storage provides a shared view of an organization's data to all nodes. It needs to be optimized for small, random I/O patterns, and provide high peak node performance and high aggregate filesystem performance to meet the variety of workloads an organization may encounter. High-speed storage should support both efficient multithreaded reads and writes from a single system, but most DL workloads will be read-dominant.

Use cases in automotive and other computer vision-related tasks, where 1080p images are used for training (and in some cases are uncompressed) involve datasets that easily exceed 30 TB in size. In these cases, 2 GBps per GPU for read performance is needed.

The preceding metrics assume a variety of workloads, datasets, and needs for training locally and directly from the high-speed storage system. It is best to characterize workloads and organizational needs before finalizing performance and capacity requirements.

NVIDIA has several partners with whom we collaborate to validate storage solutions for the DGX systems and DGX SuperPOD. For more information about these partners, see the Reference Architectures section on the NVIDIA DGX POD page.

Note: As datasets get larger, they may no longer fit in cache on the local system. Pairing large datasets that do not fit in cache with very fast GPUs can create a situation where it is difficult to achieve maximum training performance. NVIDIA GPUDirect Storage® (GDS) provides a way to read data from the remote filesystem or local NVMe directly into GPU memory providing higher sustained I/O performance with lower latency.

Using the storage fabric on the DGX SuperPOD, a GDS-enabled application should be able to read data at over 20 GBps directly into the GPUs.

Management Architecture

The DGX SuperPOD requires UFM appliances and CPU-based servers for system management that is deployed in a redundant configuration. The functions of these servers are described in Table 10.

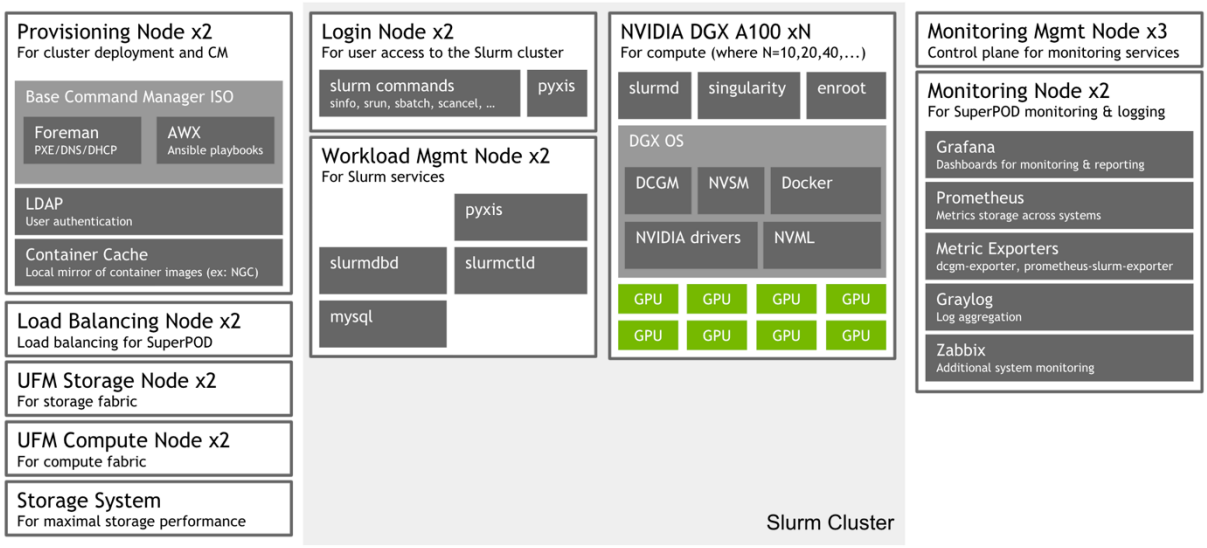
Table 10. DGX SuperPOD management servers

Server Functions	Quantity	Services Running on this Node
Provisioning, LDAP, and container cache	2	• System provisioning to manage deployment of operating system (OS) images • LDAP to offer centralized user authentication. • Container Cache locally mirrors or caches external container registries
Load balancing	2	Load balancing for BGP and data vlan routing
UFM storage	2	NVIDIA UFM for the Storage Fabric, HA pair
UFM compute	2	NVIDIA UFM for the Compute Fabric, HA pair
Login	2	Initial access point to system resources for the user
Workload mgmt	2	Resource management and scheduling (Slurm)
Monitoring mgmt	3	Management control plane for monitoring services
Monitoring	2	Gather and display cluster-wide system information

DGX SuperPOD Software Stack

The value of the DGX SuperPOD architecture extends well beyond its hardware. The DGX SuperPOD is a complete system providing all the major components for system management, job management, and optimizing workloads to ensure quick deployment, ease of use, and HA (Figure 11). Workloads are run bare metal, without virtualization, to gain maximal performance out of the system.

Figure 11. DGX SuperPOD software stack



The software stack begins with DGX OS, which is tuned and qualified for use on DGX A100 systems. DGX OS is built on an optimized version of Ubuntu 20.04 and includes certified GPU drivers, a network software stack, preconfigured NFS caching, NVIDIA data center GPU management (DCGM) diagnostic tools, and GPU-enabled container runtime. This means DGX OS is ready to run applications using NVIDIA CUDA-X, and Magnum IO developer tools, either natively or in containers.

NGC is also a key component of the DGX SuperPOD, providing the latest DL, ML, and HPC applications. NGC provides packaged, tested, and optimized containers for quick deployment, ease of use, and the best performance on NVIDIA GPUs.

Helpful NVIDIA software libraries and tools, such as CUDA-X, Magnum IO, and NVIDIA RAPIDS™, provide developers the tools they need to maximize DL, HPC, and data science performance in multinode environments.

System management of the DGX SuperPOD is handled by NVIDIA Base Command Manager. Base Command Manager is a cluster manager which bootstraps the bring up of DGX SuperPOD through

initial OS provisioning, InfiniBand fabric configuration, and deployment of Slurm for job scheduling. Base Command Manager also maintains the state of the DGX SuperPOD, keeping track of what software is running on the cluster and how it is configured. System Administrators can use Base Command Manager to manage user access, conduct system and software upgrades, and perform other common administrative tasks. Centralized monitoring and logging are also provided using Base Command Manager.

Base Command Manager leverages AWX as a central interface to run Ansible playbooks for common cluster tasks. After initial bootstrap of the cluster, system administrators can use this interface to:

- Deploy/redeploy cluster software components Foreman, Slurm, Prometheus, etc.).
- Perform health checks.
- Upgrade existing firmware and software on all hosts in the cluster.
- Benchmark the system.

The Ansible scripts and associated variables additionally serve as a cluster snapshot of the software on the cluster.

Foreman and NetBox provide services for network configuration (DHCP), a central place to track all DGX hosts, and fully automated DGX OS software provisioning over the network (PXE). The DGX OS software can be automatically reinstalled on demand through the Foreman GUI/CLI.

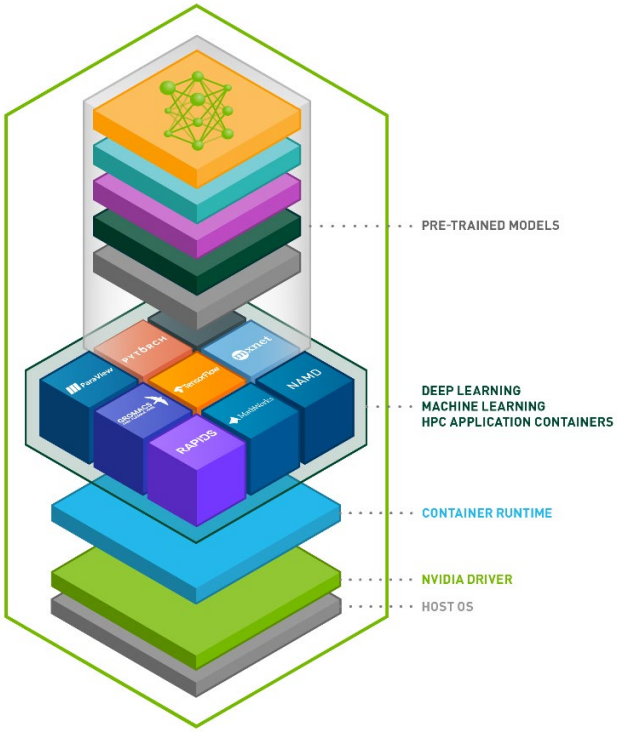
Using a job scheduler to allow users of the DGX SuperPOD to manage their workloads on the cluster is highly recommended. Base Command Manager deploys Slurm by default and examples exist for running interactive jobs and multinode jobs.

NVIDIA DGX system metrics are exported, aggregated, and stored in Prometheus. This data is then reported through Grafana. Alertmanager can use the collected data and send automated alerts as needed.

NVIDIA NGC

NGC (Figure 12) provides a range of options that meet the needs of data scientists, developers, and researchers with various levels of AI expertise. These users can quickly deploy AI frameworks with containers, get a head start with pretrained models or model training scripts, and use domain-specific workflows and Helm charts for the fastest AI implementations, giving them faster time-to-solution.

Figure 12. NGC components



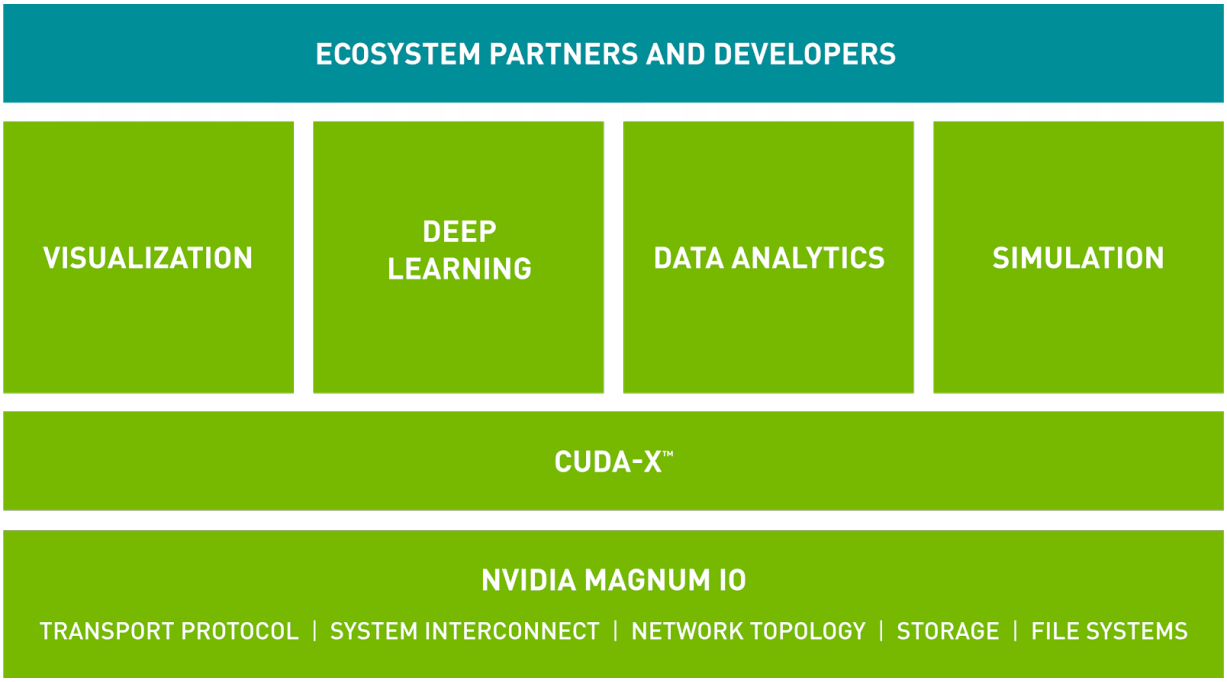
Spanning AI, data science, and HPC, the container registry on NGC features an extensive range of GPU-accelerated software for NVIDIA GPUs. The NGC hosts containers for the top AI and data science software. Containers are tuned, tested, and optimized by NVIDIA. Other containers for additional HPC applications and data analytics are fully tested and made available by NVIDIA as well. NGC containers provide powerful and easy-to-deploy software proven to deliver the fastest results, allowing users to build solutions from a tested framework, with complete control.

NGC offers step-by-step instructions and scripts for creating DL models, with sample performance and accuracy metrics to compare your results. These scripts provide expert guidance on building DL models for image classification, language translation, text-to-speech and more. Data scientists can quickly build performance-optimized models by easily adjusting hyperparameters. In addition, NGC offers pretrained models for a variety of common AI tasks that are optimized for NVIDIA Tensor Core GPUs and can be easily retrained by updating just a few layers, saving valuable time.

CUDA-X and Magnum IO

NVIDIA has two key software suites for optimizing application performance, CUDA-X and Magnum IO. CUDA-X, built on top of NVIDIA CUDA® technology, is a collection of libraries, tools, and technologies that deliver dramatically higher performance than alternatives across multiple application domains—from AI to high performance computing (HPC). Magnum IO is a suite of software to help AI developers, data scientists, and HPC researchers process massive amounts of data in minutes, rather than hours. NGC containers are enabled with both CUDA-X and Magnum IO as shown in Figure 13.

Figure 13. I/O optimized stack



At the heart of Magnum IO is GPUDirect technology, which provides a path for data to bypass CPUs and travel on "open highways" offered by GPUs, storage, and networking devices. Compatible with a wide range of communications interconnects and APIs—including NVIDIA NVLink® technology and NCCL, as well as OpenMPI and NVIDIA UCX®. Its newest element is GPUDirect Storage, which enables researchers to bypass CPUs when accessing storage and quickly access data files for simulation, analysis, or visualization.

Summary

AI is transforming our planet and every facet of life as we know it, fueled by the next generation of leading-edge research. Organizations that want to lead in an AI-powered world know that the race is on to tackle the most complex AI models that demand unprecedented scale. Our biggest challenges can only be answered with groundbreaking research that requires supercomputing power on an unmatched scale. Organizations that are ready to lead need to attract the world's best AI talent to fuel innovation and the leadership-class supercomputing infrastructure that can get them there now, not months from now.

The NVIDIA DGX SuperPOD, based on the DGX A100 system, marks a major milestone in the evolution of supercomputing, offering a scalable solution that any enterprise can acquire and deploy to access massive computing power to propel business innovation. Enterprises can start small from a single SU of 20 nodes and grow to hundreds of nodes. The DGX SuperPOD simplifies the design, deployment, and operationalization of massive AI infrastructure with a validated RA that is offered as a turnkey solution through NVIDIA value-added resellers. Now, every enterprise can scale AI to address their most important challenges with a proven approach that has 24x7 enterprise-grade support.

Appendix A. Major Components

Major components for the DGX SuperPOD configuration are listed in Table 11. These are representative of the configuration and must be finalized based on actual design.

Table 11. Major components of the 140 Node DGX SuperPOD

Count	Component	Recommended Model
Racks		
50	Rack (Legrand)	NVIDPD13
Nodes		
140	GPU Nodes	NVIDIA DGX A100 systems
4	UFM Appliance	NVIDIA Unified Fabric Manager Appliance
9	Management Servers	1U, AMD 7402P (1x24C), 256G, OS (2x480GB M.2 or SATA/SAS SSD in RAID1), 2TB NVME storage, 4x HDR200 VPI Ports, TPM 2.0
4	Management Servers	AMD EPYC 7742 2S (2x 64C), 256GB, OS (2x480GB M.2 or SATA/SAS SSD in RAID1), NVME 16TB (raw), 4x HDR200 VPI Ports, TPM 2.0
Varies	High-Speed Storage	See Storage Architecture
Ethernet Network		
8	In-Band Management	NVIDIA SN4600 switch with Cumulus Linux
10	Out-of-Band Management	NVIDIA AS4610 switch with Cumulus Linux
Compute InfiniBand Fabric		
166	Fabric Switches	NVIDIA Quantum QM8790 switch
Storage InfiniBand Fabric		
22	Fabric Switches	NVIDIA Quantum QM8790 switch
PDUs		
70	Rack PDUs	Raritan PX3-5878I2R-P1Q2R1A15D5
18	Rack PDUs	Raritan PX3-5747V-V2

Associated cables are listed in Table 12.

Table 12. Estimate of cables required for a 140-node DGX SuperPOD

Count	Component	Connection	Recommended Model
In-Band Ethernet Cables			
280	100 GbE QSFP to QSFP AOC	DGX A100 system	NVIDIA 930-20000-0007-000
34		Management nodes	
Varies		Storage	
Varies		Core DC	
Out-of-Band Ethernet Cables			
140	Cat6 cable	DGX A100 systems	Standard Cat6
17		Management nodes	
Varies		Storage	
88		PDUs	
14	10 GbE passive copper cable SFP+	Two uplinks per SU	NVIDIA MC3309130-xxx
Compute InfiniBand Cables[1]			
3374	200 Gbps QSFP56	DGX A100 systems, spine, and core	NVIDIA MF1S00-HxxxE
Storage InfiniBand Cables[1]			
568	200 Gbps QSFP56	DGX A100 systems and spine	NVIDIA MF1S00-HxxxE
56[2]		Storage	
16		Management nodes	

1. Part number will depend on exact cable lengths needed based on data center requirements.
2. Count and cable type required depend on specific storage selected.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure that the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA DGX, NVIDIA DGX SuperPOD, NVIDIA NGC, NVIDIA Quantum, CUDA, and CUDA-X are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051

http://www.nvidia.com