NVIDIA DGX SuperPOD: Scalable Infrastructure for AI Leadership

NVIDIA DGX A100 System Reference Architecture

RA-09950-001 | May 2020

Document History

RA-09950-001

Version	Date	Authors	Description of Change
01	2020-05-14	Craig Tierney, Jeremy Rodriguez, Premal Savla, Robert Sohigian	Initial release

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 training today's state-of-the-art deep learning (DL) models and to fuel innovation well into the future. 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 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. This design introduces compute building blocks called scalable units (SU) allowing for a modular build out of the full DGX SuperPOD of 140 DGX A100 systems and scaling to hundreds of nodes. The DGX SuperPOD design includes Mellanox networking switches, DGX POD certified storage, and NVIDIA GPU CLOUD® (NGC) optimized software.

The DGX SuperPOD is a part of the NVIDIA SATURNV research and development platform that deploys over 1800 DGX systems. That knowledge fuels the ability of NVIDIA to innovate at an accelerated scale in AI for autonomous vehicles, natural language processing, robotics, graphics, high performance computing (HPC), and other domains.

The DGX SuperPOD can be purchased from select NVIDIA partners and deployed either on-premises or at DGX-Ready Data Center Colocation Partners around the world.

Contents

DGX SuperPOD with DGX A100 Systems 1
NVIDIA DGX A100 System 2
Features 3
Design Requirements 4
Compute Fabric 4
Storage Fabric 4
SuperPOD Architecture 5
Network Architecture 7
Compute Fabric 8
Storage Fabric 10
In-Band Management Network 12
Out-of-Band Management Network 12
Management Servers 13
AI Software Stack 15
DGX OS and POD Management 15
NVIDIA GPU Cloud (NGC) 17
NVIDIA GPU Operator for Kubernetes 18
CUDA-X and Magnum IO 19
Data Center Configurations 21
Power 21
Cooling 22
Rack Elevations 22
Scalable Unit Racks 23
Spine Rack 24
Storage Rack 25
Storage Requirements 27
Summary 29
Appendix A. Major Components vi

DGX SuperPOD with DGX A100 Systems

The compute needs 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 utilizes dozens, hundreds, or even thousands of GPUs for evaluating and optimizing different model configurations and parameters. In addition, organizations have multiple AI researchers that all need to train numerous models simultaneously. Systems at this massive scale may be new to AI researchers, but these installations have traditionally been the 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 a new perception of 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 in order to build fast, capable, and cost-efficient systems. Developing infrastructure requirements can often 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 which represent the overall needs of an organization is a time-consuming process.

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:

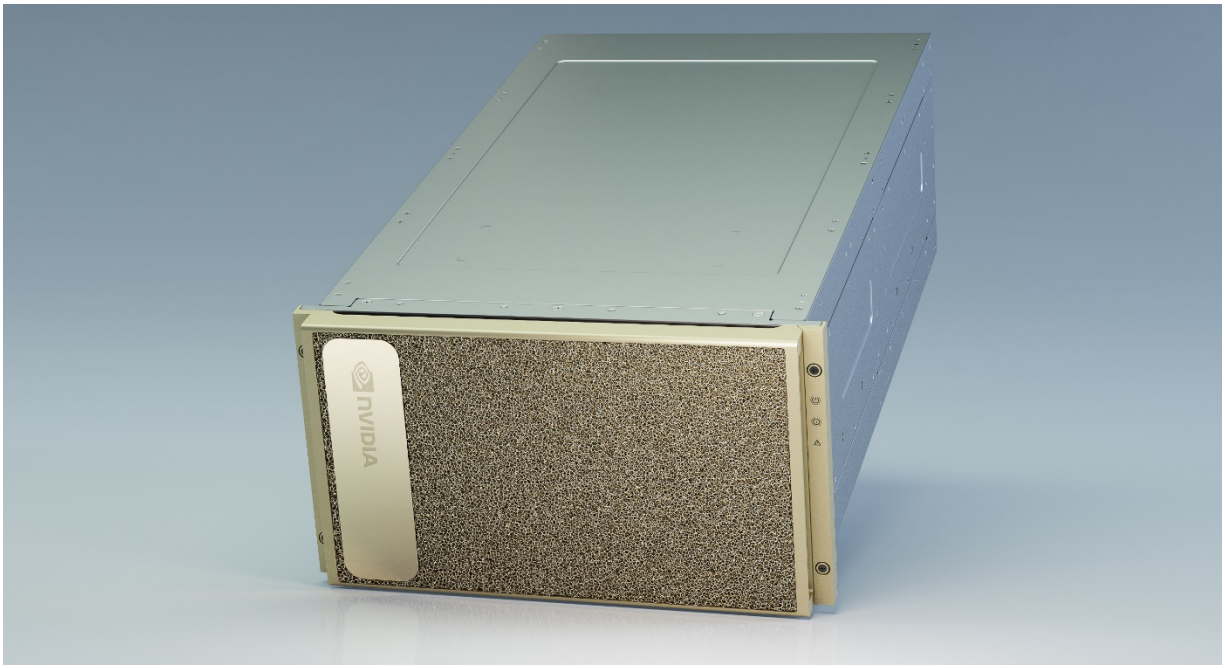
- Powerful GPU nodes with many GPUs, a large memory footprint, and fast connections between the GPUs for scale-up computing to support the variety of DL models in use.
- A low-latency, high-bandwidth, network interconnect designed with the capacity and topology to minimize bottlenecks.
- A storage hierarchy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

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

NVIDIA DGX A100 System

The DGX SuperPOD is an optimized system for multi-node DL and HPC. It consists of 140 DGX A100 systems (Figure 1), with a total of 1120 NVIDIA A100 GPUs. It is built using the NVIDIA DGX POD reference architecture and is configured to be a scalable and balanced system providing maximum performance.

Figure 1. DGX A100 system



Features

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

Table 1. DGX SuperPOD features

Component	Technology	Description
Compute Nodes	NVIDIA DGX A100 System	• 1120 DGX A100 SXM3 GPUs • 45.6 TB of HBM2 memory • 336 AI PFLOPS via Tensor Cores • 140 TB System RAM • 2.2 PB local NVMe • 600 GBps NVLink bandwidth per GPU • 4.8 TBps total NVSwitch bandwidth per node
Compute Fabric	QM8790 Mellanox Quantum™ HDR InfiniBand Smart Switch	Full fat-tree network built with eight connections per DGX A100 system
Storage Fabric	QM8790 Mellanox Quantum HDR InfiniBand Smart Switch	Fat-tree network with two connections per DGX A100 system
In-band Management Network	Mellanox SN3700C switch	One connection per DGX A100 system
Out-of-band Management Network	Mellanox AS4610 switch	One connection per DGX A100 system
Management Software	DeepOps DGX POD Management Software	Software tools for deployment and management of SuperPOD nodes, Kubernetes and Slurm
Key System Software	NVIDIA Magnum IO™ technology	Suite of library technologies that optimize GPU communication performance
	CUDA-X™	CUDA-X is a collection of libraries, tools, and technologies that maximize application performance on NVIDIA GPUs
User Runtime Environment	NVIDIA GPU Cloud (NGC)	NGC provides the best performance for all DL frameworks
	Slurm	Slurm is used for the orchestration and scheduling of multi-GPU and multi-node training jobs

Design Requirements

The DGX SuperPOD is designed to minimize system bottlenecks and maximize performance for the diverse nature of AI and HPC workloads. In order 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 which allows distributed DL applications to scale across hundreds of nodes.
- The ability to quickly deploy and update the system. Leveraging the reference architecture allows data center staff to develop a full solution with fewer design iterations.
- Key management services in high availability configurations required to monitor and manage the system. This includes system provisioning, fabric management, login, system monitoring and reporting, as well as workload management and orchestration.

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 full fat-tree topology, maximizing the network capability for the DGX A100 systems.
- Multiple DGX SuperPOD configurations can be connected to create systems with thousands of nodes.
- The fabric supports Adaptive Routing[1].
- The network is optimized for Mellanox Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ version 2[1].

Storage Fabric

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

- Provide single node bandwidth in excess of 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 share 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] To learn more about configurating these technologies, contact a representative from NVIDIA or Mellanox.

SuperPOD Architecture

The basic building block for the DGX SuperPOD is the scalable unit (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 can be well supported. A single SU is capable of 48 AI PFLOPS.

Figure 2. DGX A100 scalable unit

The DGX A100 systems have eight HDR (200 Gbps) InfiniBand host channel adapters (HCAs) for compute traffic. Each GPU has its own associated HCA. For the most efficient network, there are eight network planes, one for each HCA of the DGX A100 system that connect 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.

The SU has its own 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 additional space of a SU management rack or separate rack depending on the data center layout.

Details about SU are covered in the following sections.

Note: The DGX A100 system supports Multi Instance GPU (MIG) partitioning of each DGX A100 GPU. This feature can enhance the productivity of the DGX SuperPOD by:

- Providing AI research teams the ability to efficiently run thousands of smaller experiments in isolation for each other.
- Providing enhanced AI inference by supporting thousands of simultaneous inference processes.

MIG is an advanced feature that integrates with DGX POD or data center resource management and orchestration software. Use of this feature will be covered in supplemental DGX A100 documentation. Contact NVIDIA enterprise support for more information.

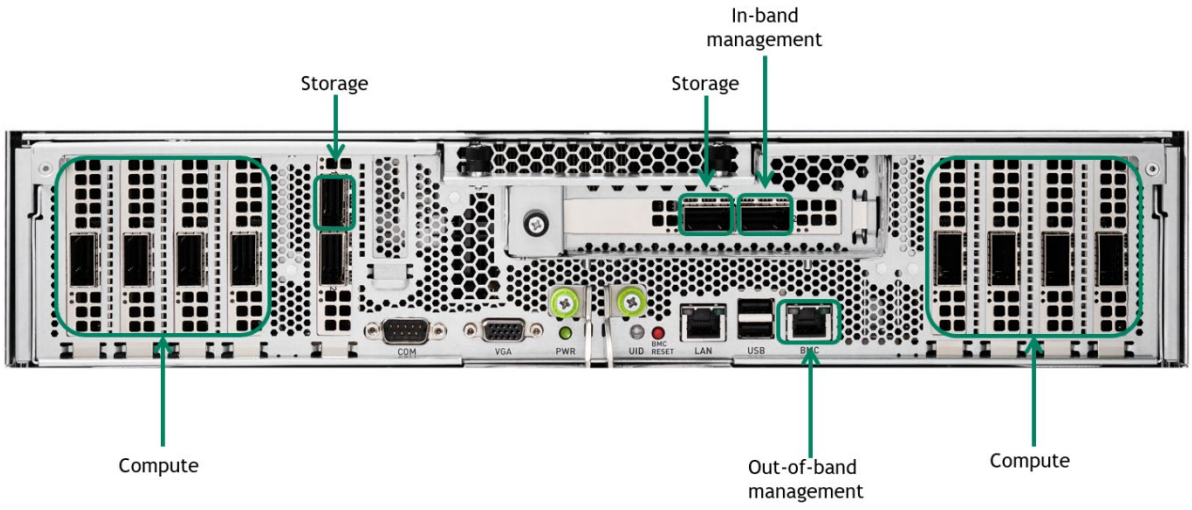
Network Architecture

The DGX SuperPOD has four networks:

- **Compute fabric.** Connects the eight Mellanox ConnectX®-6 HCAs from each DGX A100 through separate storage planes.
- **Storage fabric.** Uses two ports, one each from two dual-port Mellanox 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 BMC port of each DGX A100 system to an additional Ethernet switch.

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

Figure 3. Network connections for DGX A100 system



Note: In order to maximize bandwidth from storage and to achieve per -node performance of over 20 GBps, each DGX A100 system has an additional dual-port Mellanox ConnectX-6 HCA installed.

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

Table 2. DGX SuperPOD network connections

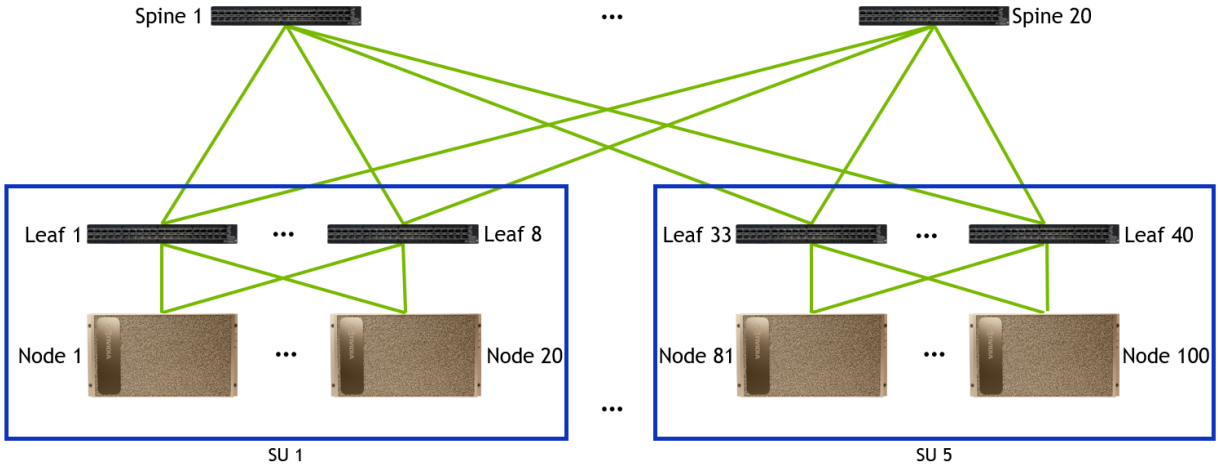
Component	InfiniBand		Ethernet	
	Compute	Storage	In-Band	Out-of-Band
DGX A100 Systems	1120	280	140	140
Management Servers[1]	Depends	Depends	28	14
Storage System[2]	Varies	Varies	Varies	Varies

1. Two ports are required for each fabric if an external subnet manager is used.
2. The number of storage system connections will depend on the system to be installed.

Compute Fabric

Each DGX A100 system has eight connections to the compute fabric (Figure 4). The 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.

Figure 4. Compute fabric topology for 100 node system



The full DGX SuperPOD is built by connecting seven SUs using three layers of switching. Building systems with 100 nodes or less is simpler as the third layer of switching is not required. Table 3 shows the switch and link count for different sized systems.

Table 3. Compute fabric switch and cable counts

Nodes	SUs	QM8790 Switches			Cables		
		Leaf	Spine	Core	Leaf	Spine	Core
10	½	8	2		80	80	
20 (Single SU)	1	8	4		160	160	
40	2	16	10		320	320	
80	4	32	20		640	640	
100	5	40	20		800	800	
140 (DGX SuperPOD)	7	56	56	28	1120	1120	560

The Mellanox Unified Fabric Manager (UFM) can be used for advanced fabric monitoring. While not required, it is strongly recommended two servers be put on each fabric, in place of a DGX A100 system, to run UFM in high availability mode.

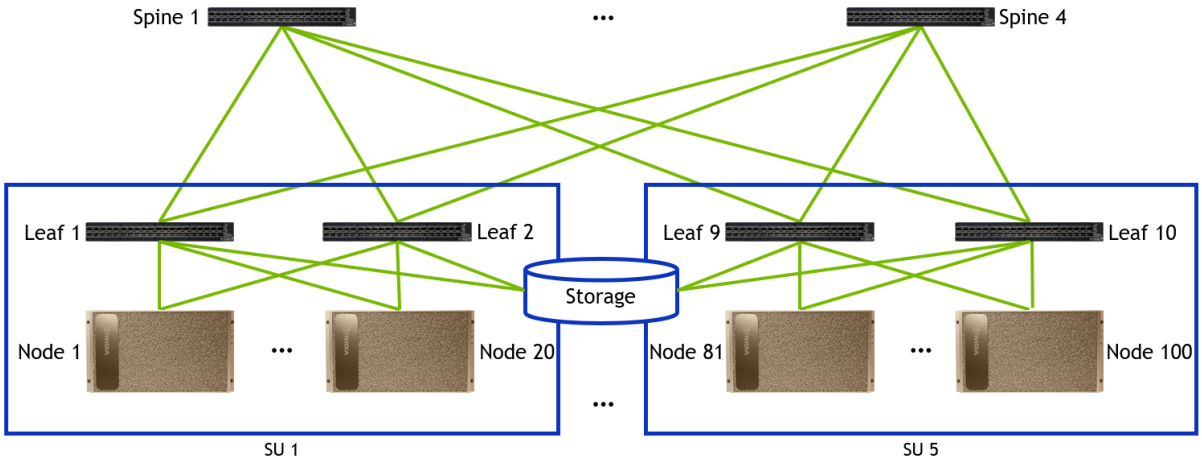
The compute fabric utilizes QM8790 Series Mellanox Quantum HDR 200 Gbps InfiniBand Smart Switches (Figure 5).

Figure 5. Mellanox QM8790 HDR 200Gbps InfiniBand Switch

Storage Fabric

The storage fabric employs an InfiniBand network fabric (Figure 6). An InfiniBand-based fabric is essential to maximize bandwidth since the I/O per-node requirements for the DGX SuperPOD is to 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 6. Storage fabric topology for 100 node system



The storage fabric also uses Mellanox QM8790 switches. The fabric is a tree with an approximately 3-to-2 subscription ratio. This network topology offers a good balance between performance and cost. In the design below, we are assuming that the storage servers require eight ports per SU. This may vary depending on the specific storage architecture and the specific storage performance requirements of a given deployment.

Table 4 shows the switch and link count for different sized systems.

Table 4. Storage fabric switch and cable counts

Nodes	SUs	Storage Ports	QM8790 Switches		Cables			Subscription Ratio
			Leaf	Spine	Leaf	Spine	Storage	
10	1/2	4	2	1	20	20	4	1:1
20	1	8	2	1	40	32	8	3:2
40	2	16	4	2	80	64	16	3:2
80	4	32	8	4	160	128	32	3:2
100	5	40	10	4	200	160	40	3:2
140	7	56	14	8	280	224	56	5:4

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 Kubernetes and to other services outside of the cluster such as the NGC registry, code repositories, and data sources.

The in-band network is built using 100 GbE Mellanox SN3700C switches (Figure 7). There are two uplinks from each switch to the data center core switch. Connectivity to external resources and to the internet are routed through the core data center switch.

Figure 7. Mellanox SN3700C 100 GbE Data Center Open Ethernet Switch

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 out-of-band management network is based on 1 GbE Mellanox AS4610 switches (Figure 8), running the DevOps-friendly Cumulus Linux network operating system. These switches are connected directly to the data center core switch. In addition, all Ethernet switches are connected via serial connections to existing Opengear 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.

Figure 8. Mellanox AS4610 1 GbE Data Center Open Ethernet Switch

Management Servers

The DGX SuperPOD requires several CPU based servers for management of the system. The services provided are:

- Fabric management of both the Compute and Storage Fabrics
- System Provisioning
- Login
- System Monitoring and Reporting
- NVIDIA GPU Cloud Mirror and Cache
- Workload Management and Orchestration

To provide the highest level of availability, each of these services are on their own servers and the services are run in pairs in a high availability configuration.

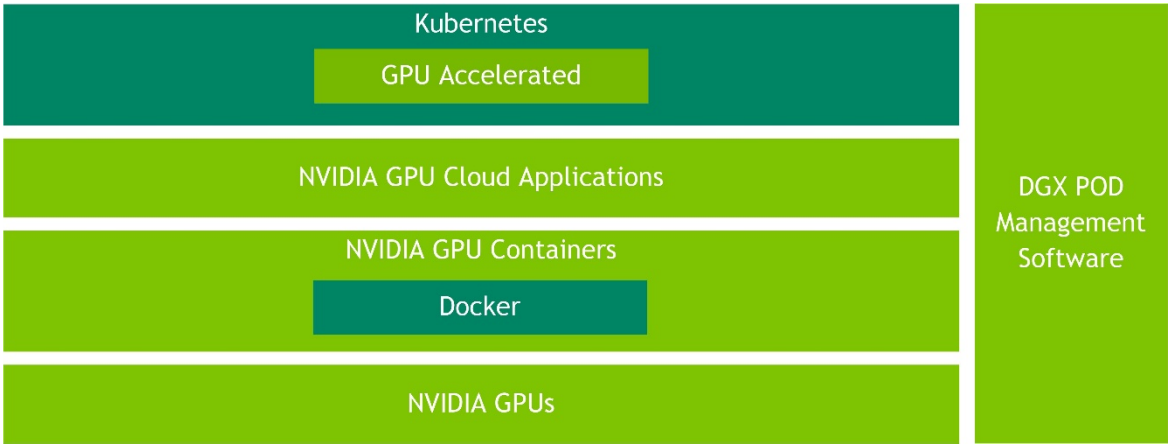
AI Software Stack

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 high availability. The software stack begins with the DGX Operating System (DGX OS), which is tuned and qualified for use on DGX A100 systems. The DGX SuperPOD contains a set of tools to manage the deployment, operation, and monitoring of the cluster. NGC is a key component of the DGX SuperPOD, providing the latest DL frameworks. NGC provides packaged, tested and optimized containers for quick deployment, ease of use, and the best performance on NVIDIA GPUs. Lastly, key tools like CUDA-X, Magnum IO, and RAPIDS provide developers the tools they need to maximize DL, HPC, and data science performance in multi-node environments.

DGX OS and POD Management

NVIDIA AI software (Figure 9) running on the DGX SuperPOD provides a high-performance DL training environment for large scale multi-user AI software development teams. In addition to the DGX OS, it contains cluster management, orchestration tools and workload schedulers (DGX POD management software), NVIDIA libraries and frameworks, and optimized containers from the NGC container registry. For additional functionality, the DGX POD management software includes third-party opensource tools recommended by NVIDIA which have been tested to work on DGX POD racks with the NVIDIA AI software stack. Support for these tools is available directly from third-party support structures.

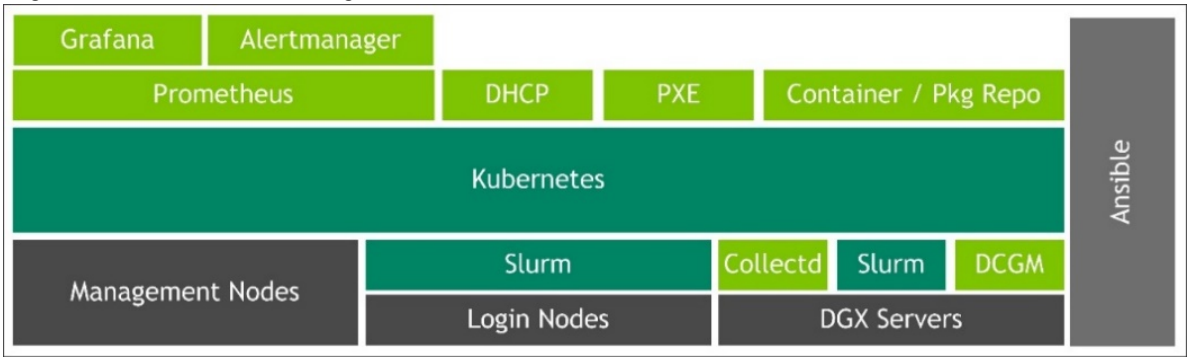
Figure 9. AI software stack



The foundation of the NVIDIA AI software stack is the DGX OS, built on an optimized version of the Ubuntu or RedHat Linux operating system and tuned specifically for the DGX hardware. The DGX OS software includes certified GPU drivers, a network software stack, pre-configured NFS caching, NVIDIA data center GPU management (DCGM) diagnostic tools, GPU-enabled container runtime, NVIDIA CUDA-X, and Magnum IO developer tools.

The DGX POD management software (Figure 10) is composed of various services running on the Kubernetes container orchestration framework for fault tolerance and high availability. Services are provided for network configuration (DHCP) and fully-automated DGX OS software provisioning over the network (PXE). The DGX OS software can be automatically re-installed on demand by the DGX POD management software.

Figure 10. DGX POD management software



The DGX POD management software DeepOps provides the tools to provision, deploy, manage and monitor the DGX SuperPOD. Services are hosted in Kubernetes containers for fault tolerance and high availability. The DGX POD management software leverages the Ansible configuration tool to install and configure all the tools and packages needed to run the system. System data collected by Prometheus is reported through Grafana. Alertmanager can use the collected data and send automated alerts as needed.

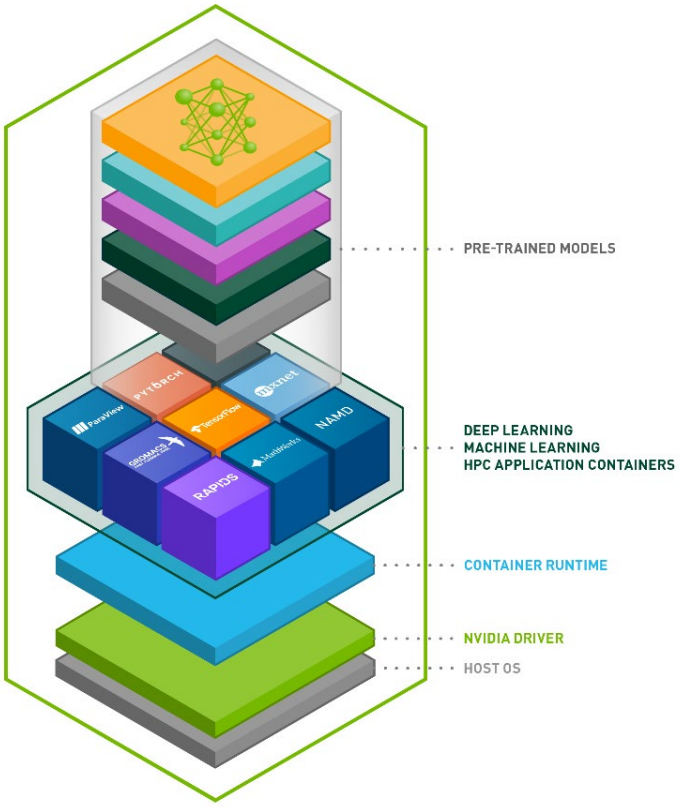
For sites required to operate in an air-gapped environment or needing additional on-premises services, a local container registry mirroring NGC containers, as well as Ubuntu and Python package mirrors, can be run on the Kubernetes management layer to provide services to the cluster.

The DGX POD management software can deploy Slurm or Kubernetes as the orchestration and workload manager. Slurm is often the best choice for scheduling training jobs in a shared multi-user, multi-node environment where advanced scheduling features such as job priorities, backfill, and accounting are required. Kubernetes is often the best choice in environments where GPU processes run as a service, such as inference, large use of interactive workloads through Jupyter notebooks, and where there is value to having the same environment as is often used at the edge of an organization's datacenter.

NVIDIA GPU Cloud (NGC)

The NGC (Figure 11) 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 pre-trained 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 11. 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 deep learning 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 pre-trained models for a variety of common AI tasks that are optimized for NVIDIA Tensor Core GPUs, and can be easily re-trained by updating just a few layers, saving valuable time.

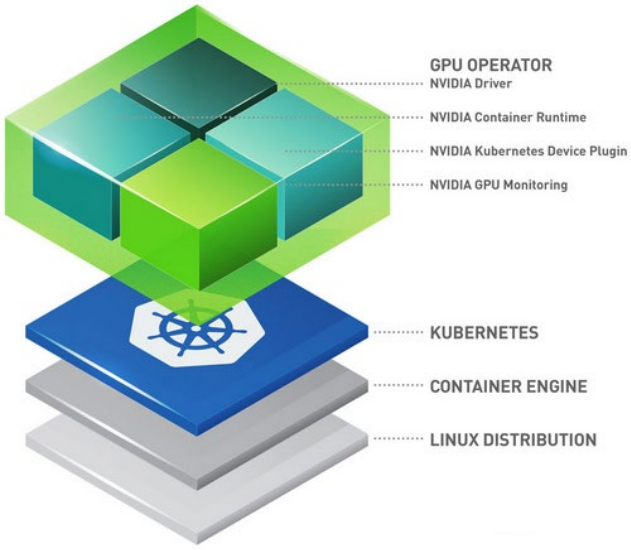
NVIDIA GPU Operator for Kubernetes

NGC containers can be run quite effectively within Kubernetes. To leverage the power of Kubernetes it must be tightly coupled to the hardware it manages. Kubernetes provides access to special hardware resources such as NVIDIA GPUs, NICs, InfiniBand adapters and other devices through the device plugin framework. However, configuring and managing nodes with these hardware resources requires configuration of multiple software components such as drivers, container runtimes or other libraries which are difficult and prone to errors.

The Operator Framework within Kubernetes takes operational business logic and allows the creation of an automated framework for the deployment of applications within Kubernetes using standard Kubernetes APIs and kubectl.

The NVIDIA GPU Operator (Figure 12) introduced here is based on the operator framework and automates the management of all NVIDIA software components needed to provision GPUs within Kubernetes. NVIDIA, Red Hat, and others in the community have collaborated on creating the GPU Operator.

Figure 12. NVIDIA GPU Operator



The GPU Operator provides four key functions:

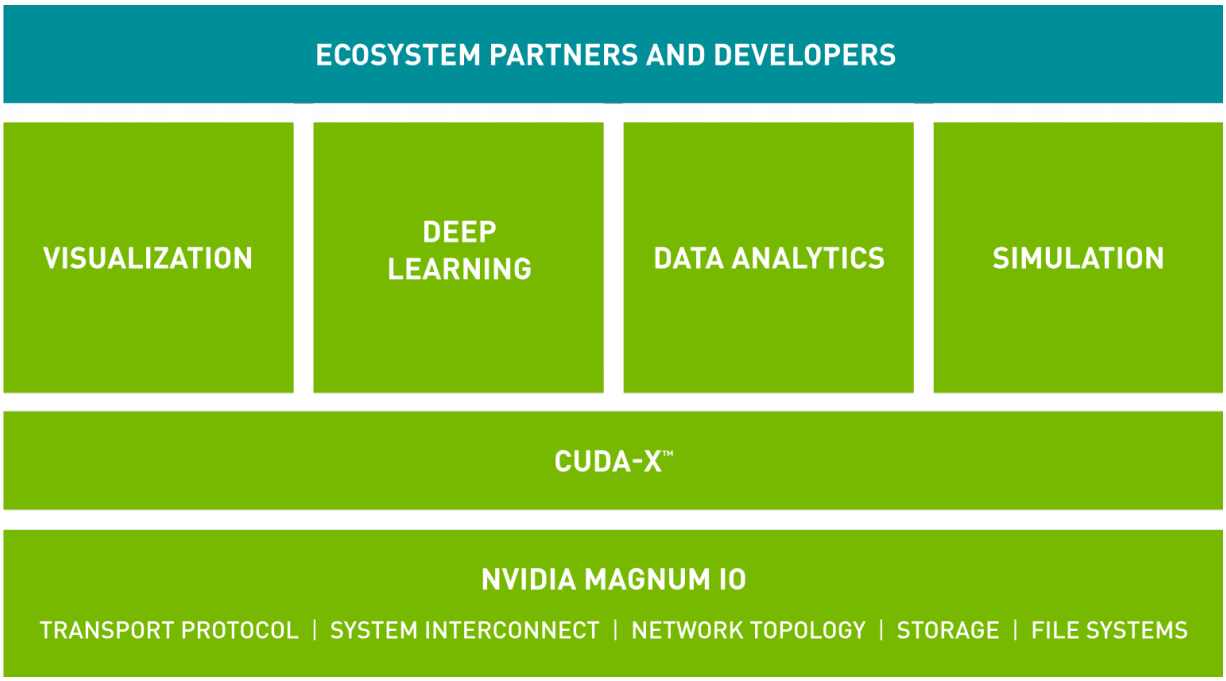
1. The NVIDIA driver runs as a containerized workload so that it can be managed by Kubernetes.
2. The NVIDA container runtime simplifies the integration of NGC containers into the Kubernetes environment.
3. The NVIDIA device plugin connects Kubernetes to the GPU specific functions.
4. GPU monitoring is provided so fine-grained usage data can be collected for real-time monitoring and historical analysis.

To simplify the deployment of the GPU operator itself, NVIDIA provides a Helm chart. The versions of the software components that are deployed by the operator (e.g. driver, device plugin) can be customized by the user with templates in the Helm chart. The operator then uses the template values to provision the desired versions of the software on the node. This provides a level of parameterization to the user.

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 CUDA® technology, is a collection of libraries, tools, and technologies that deliver dramatically higher performance than alternatives across multiple application domains—from artificial intelligence to high performance computing. Magnum IO is a suite of software to help AI, data scientists and high-performance computing 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 and NCCL, as well as Open MPI and 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.

Data Center Configurations

When deploying a DGX SuperPOD, due to the high-power consumption and corresponding cooling needs, server weight, and multiple networking cables per server, additional care and preparation is needed for a successful deployment. As with all IT equipment installation, it is important to work with the data center facilities team to ensure the environmental requirements can be met.

Power

Table 5 lists the maximum power consumed by the various components of the DGX SuperPOD. Some components such as management nodes and storage are estimated, as they depend on the chosen solution. These power values can be used with the rack elevations below to compute the power per rack.

Table 5. Per component power usage

Equipment	Maximum Power
DGX A100 system	6.5 kW
Management nodes	0.6 kW
Mellanox QM98790 InfiniBand switch	0.65 kW
Mellanox SN3700C switch	0.5 kW
Mellanox AS4610 switch	0.1 kW

Each SU requires approximately 137 kW. The maximum power draw for a single rack is 26 kW. The total power required for the full DGX SuperPOD including storage (assumed at 20kW) is approximately 1 MW. The rack layouts can be altered to match the power distribution and per-rack cooling needs for a specific data center.

Cooling

The compute room air handlers (CRAH) supply cool air under the raised floor and upward through perforated tile into the enclosed cold aisle. Air is discharged through the back of racks, where it is returned to CRAHs for conditioning. Alternate cooling solutions can be substituted as needed per data center.

The data center should maintain cooling to ASHRAE TC9.9 2015 recommended thermal guidelines.

Blanking panels must be installed wherever possible. All switches must have the correct fan flow direction and be mounted flush with front of the rack. Switch ports should face the rear of the rack to avoid cables returning into the rack from the front. These measures are needed to ensure proper airflow through the rack.

Rack Elevations

The reference design for the SU has four DGX A100 systems per rack. Each SU also has a management and management rack, centrally located between each of the compute racks to minimize cable lengths. The upper level spine switches, management nodes, core network, and storage components occupy a separate set of racks that are central to all the SU racks. The density of the SU racks, and the overall layout of the system can be modified depending on the specific requirements of the data center.

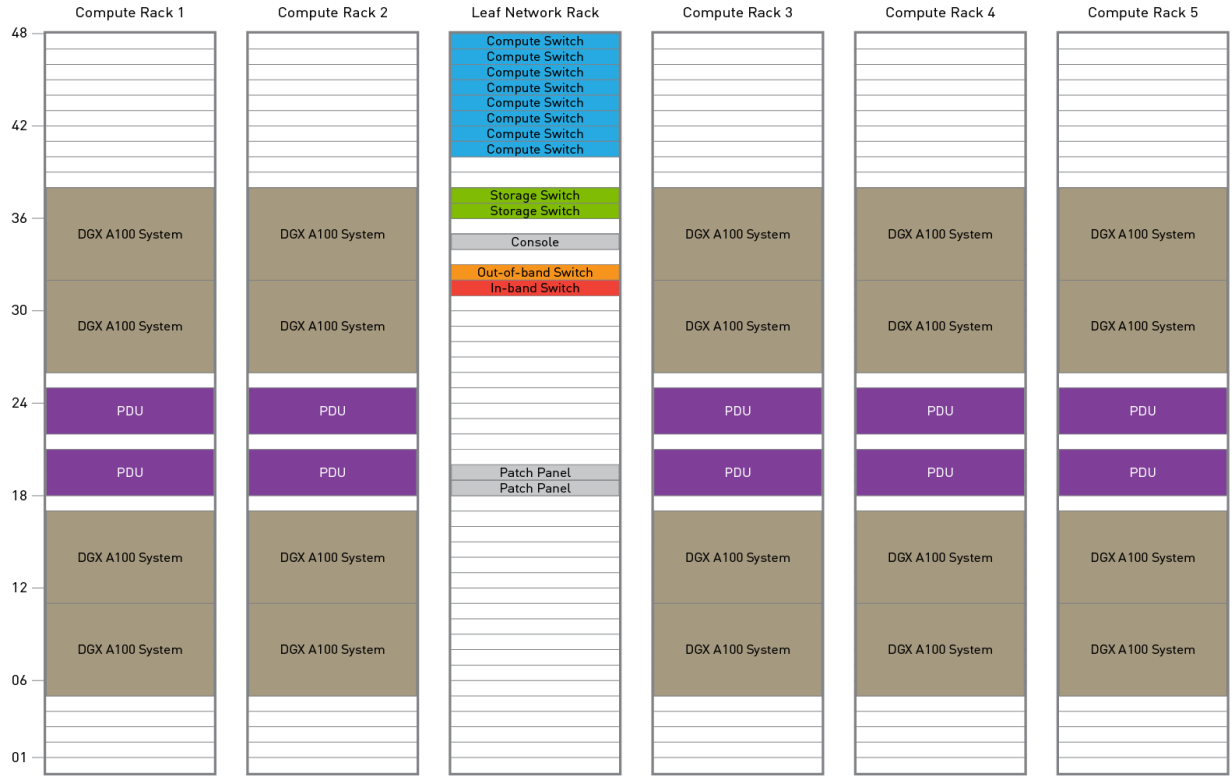
Most racks are compute racks that contain four DGX A100 systems. The InfiniBand racks are located to minimize cable lengths and to ease fiber cabling. Recommended rack size is 48U tall, 700 mm wide, and 1200 mm deep. The extra width and depth ensure that the 0U PDUs and IB cabling can be accommodated without interfering with maintenance of the DGX A100 systems. Replacing a GPU tray can be very challenging in smaller racks. The cabinets should support a minimum static load of 600 kg. Cable pathways should conform to TIA 942 standards.

Scalable Unit Racks

The layout consists of five interconnected SUs (Figure 14). The rack elevations of the clusters are illustrated as viewed from the front of the racks. In-rack equipment locations are illustrated for raised-floor data centers. Equipment position can be adjusted per data center cooling requirements.

Each SU consists of 20 DGX A100 systems distributed across six racks. Each rack of four DGX A100 systems has two 3U PDUs. One rack is dedicated to the leaf switches for compute and storage fabrics. Leaf racks also include ethernet and console devices. Although these units can be built out incrementally in phases, preparing the cabling in advance avoids more expensive incremental cabling work during later expansion phases.

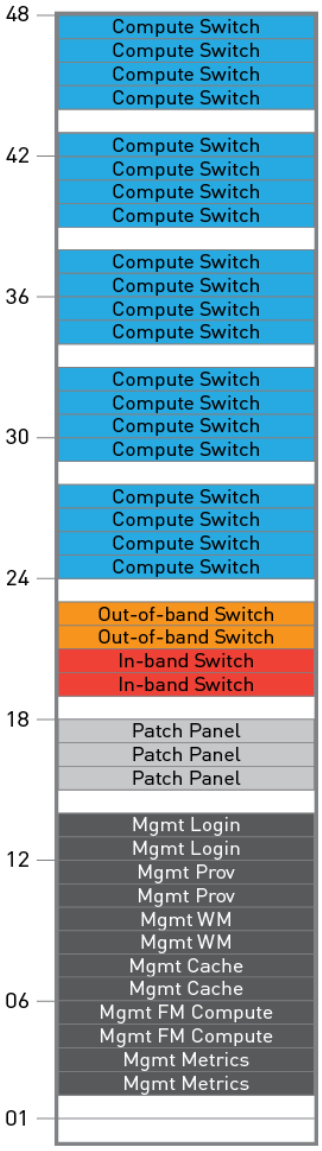
Figure 14. SU rack elevations



Spine Rack

One spine rack (Figure 15) contains the compute InfiniBand spine switches, management nodes for administrative tasks and ethernet connectivity with in-band and out-of-band Ethernet top-of-rack switches.

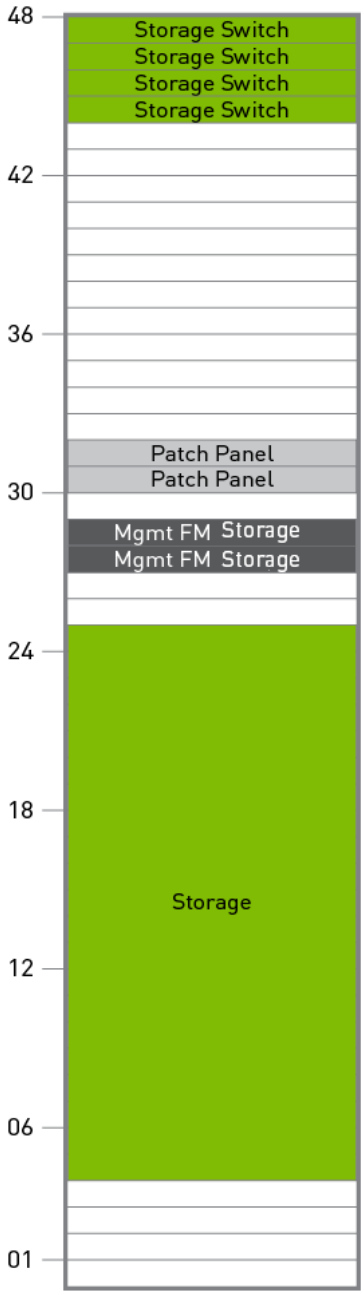
Figure 15. Spine rack elevations



Storage Rack

The storage rack (Figure 16) contains the storage InfiniBand spine switches and management nodes for administrative tasks. Ethernet connectivity is patched to the spine rack.

Figure 16. Storage rack elevations



Storage Requirements

Training performance can be limited by the rate at which data can be read and re-read 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 non-persistent, to balance the needs of performance, capacity, and cost.

Table 6 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 6. DGX SuperPOD storage and caching hierarchy

Storage Hierarchy Level	Technology	Total Capacity	Read Performance
RAM	DDR4	1 TB per node (upgradeable to 2 TB)	> 200 GBps
Internal Storage	NVMe	15 TB per node (upgradable to 30 TB)	> 55 GBps
High-Speed Storage	Varies	Varies depending on specific needs	Required: • Aggregate system read > 100 GBps • Aggregate system write > 50 GBps • Single-Node read > 6 GBps • Single-Node write > 2 GBps Desired: • Single-Node 2 GBps read per GPU (16 GBps)

Caching data in local RAM provides the best performance for reads. This caching is transparent once 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 (e.g. cachefilesd for NFS filesystems), not every network filesystem provides a method to do so.

High-speed storage provides a shared view of your 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 multi-threaded reads and writes from a single system but most of DL workloads will be read-dominant.

30 TB datasets are still considered large. 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. There is a need for 2 GBps per GPU for read performance in these cases.

The metrics above assume a variety of workloads, datasets, and needs for training locally and directly from the high-speed storage system. It is best to characterize your own workloads and 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. A list of these partners is available here.

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. GPUDirect Storage 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 nearly 30 GBps directly into the GPUs. If this performance is not enough for demanding applications, the storage could be connected to the compute fabric to achieve even greater performance. While the number of compute nodes would be reduced, this GPUDirect Storage-optimized version of the DGX SuperPOD could achieve ingest performance exceeding 100 GBps

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 Scalable Unit 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 reference architecture that is offered as a turnkey solution through our value-added resellers. Now, every enterprise can scale AI to address their most important challenges with a proven approach that is backed by 24x7 enterprise-grade support.

Appendix A. Major Components

Major components for the DGX SuperPOD configuration are listed in Table 7.

Table 7. Major components of the DGX SuperPOD

Count	Component	Recommended Model
Racks		
44	Rack (AFCO)	NVIDPD13
Nodes		
140	GPU Nodes	NVIDIA DGX A100 systems[1]
14	Management Servers	Dual Socket server, mid-bin processor, 16+ cores per CPU, 256+ GB, 512 GB RAID1 SSD, 2x HDR IB, 2x 10 GbE
Varies	High-Speed Storage	See Storage Requirements
Ethernet Network		
9	In-Band Management	Mellanox SN3700C
9	Out-of-Band Management	Mellanox AS4610
InfiniBand Fabric		
162	Fabric Switches	Mellanox QM8790
PDUs		
70	Rack PDUs	Raritan PX3-5878I2R-P1Q2R1A15D5
18	Rack PDUs	Raritan PX3-5747V-V2
1. A support plan of at least three years is recommended. Contact NVIDIA enterprise support for more information.		

Associated cables are listed in Table 8.

Table 8. Cables required for the DGX SuperPOD

Count	Component	Connection	Recommended Model
In-Band Ethernet Cables			
140	100 GbE QSFP to QSFP AOC	DGX A100 system	Mellanox 930-20000-0007-000
28	100 GbE QSFP to QSFP AOC	Management nodes	Mellanox 930-20000-0007-000
Varies	100 GbE QSFP to QSFP AOC	Storage	Mellanox 930-20000-0007-000
14	100 GbE QSFP to QSFP AOC	Two uplinks per SU	Mellanox 930-20000-0007-000
Out-of-Band Ethernet Cables			
140	Cat5 cable	DGX A100 systems	No Recommendation
14	Cat5 cable	Management nodes	No Recommendation
Varies	Cat5 cable	Storage	No Recommendation
88	Cat5 cable	PDUs	No Recommendation
14	10 GbE passive copper cable SFP+	Two uplinks per SU	Mellanox MC3309130-xxx
Compute InfiniBand Cables[1]			
2800	200 Gbps QSFP56	DGX A100 systems, spine, and core	Mellanox MF1S00-HxxxE
Storage InfiniBand Cables[1]			
504	200 Gbps QSFP56	DGX A100 systems and spine	Mellanox MF1S00-HxxxE
56[2]	200 Gbps QSFP56	Storage	Mellanox MF1S00-HxxxE
2	200 Gbps QSFP56	Management nodes	Mellanox MF1S00-HxxxE

1. Part number will depend on exact cable lengths needed based on data center requirements
2. Count required depends 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 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 GPU CLOUD, 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

© 2020 NVIDIA Corporation. All rights reserved.

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com