NVIDIA DGX SuperPOD: Instant Infrastructure for AI Leadership

Reference Architecture

RA-09720-001 | November 2019

Document History

RA-09720-001

Version	Date	Authors	Description of Change
001	2019-11-01	David Coppit, Angelica Lin, Alex Naderi, Jeremy Rodriguez, Robert Sohigian, and Craig Tierney	Initial release
002	2019-11-13	Robert Sohigian and Craig Tierney	Updates and corrections

Abstract

The NVIDIA DGX SuperPOD™ is a first-of-its-kind artificial intelligence (AI) supercomputing infrastructure that delivers groundbreaking performance, deploys in weeks as a fully integrated system, and is designed to solve the world's most challenging AI problems. Increasingly complex AI models and larger data sizes demand powerful supercomputers to support the iteration speed and time-to-train required to fuel innovation.

The DGX SuperPOD reference architecture is based on 64 DGX-2 systems, Mellanox InfiniBand networking, DGX POD certified storage, and NVIDIA® GPU Cloud (NGC) optimized software. The design also includes mechanical, power, and cooling options for both compute room air handler (CRAH) and rear door heat exchanger (RDHX) facilities.

With a modified DGX SuperPOD design, consisting of 96 DGX-2H systems, the following results were obtained:

- 9.4 petaFLOPS on the TOP500 HPL[1] benchmark, making it the 22nd world's fastest supercomputer
- Eight new MLPerf performance records[2]

The DGX SuperPOD can be purchased from select NVIDIA partners and deployed either on-premise or at DGX ready data center colocation partners around the world.

NVIDIA operates over 1500 DGX systems configured in multiple DGX PODs for our SATURNV deep learning (DL) research and development. This is imperative for NVIDIA to achieve innovation at an accelerated scale in AI for autonomous vehicles, robotics, graphics, high performance computing (HPC), and other domains.

---

[1] https://www.top500.org/system/179691

[2] MLPerf 0.6 submission information: Per accelerator comparison using reported performance for NVIDIA DGX-2H systems (16 Tesla V100 GPUs) compared to other submissions at same scale except for MiniGo where the NVIDIA DGX-1 systems (eight Tesla V100 GPUs) submission was used.

MLPerf ID Max Scale: Mask R-CNN: 0.6-23, GNMT: 0.6-26, MiniGo: 0.6-11 | MLPerf ID Per Accelerator: Mask R-CNN, SSD, GNMT, Transformer: all use 0.6-20, MiniGo: 0.6-10. See mlperf.org for more information.

Contents

NVIDIA DGX SuperPOD

The compute needs of AI researchers continue 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. Also, the most complex models require multiple GPUs to train faster or support larger configurations. In addition, organizations with multiple AI researchers need to train many models simultaneously, requiring extensive compute resources. 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 endeavor of almost every kind.

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

Designing and building computers at these scales 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 a large GPU cluster to achieve the best 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:

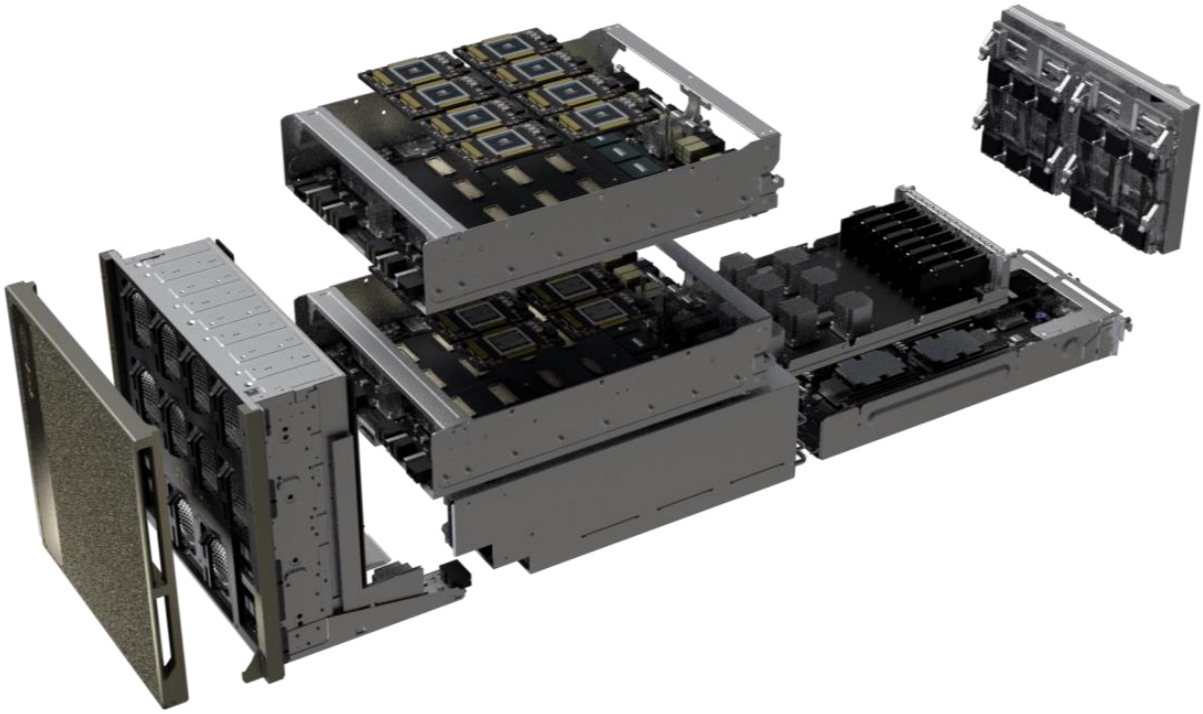
- A low-latency, high-bandwidth, network interconnect designed with the capacity and topology to minimize bottlenecks.
- A storage hierarchy that can provide maximum performance for the various dataset structure needs.

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

Overview

The DGX SuperPOD is an optimized system for multi-node DL and HPC. It consists of 64 DGX-2 systems (Figure 1), with a total of 1,024 NVIDIA Tesla® V100 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-2 system



The DGX-2 system provides incredible performance for unprecedented training capability. Each DGX-2 system has 16 Tesla V100 GPUs connected with NVIDIA NVLink® technology and the NVIDIA NVSwitch™ AI network fabric. The fabric has 2.4 TB per second of bisection bandwidth which provides the necessary resources to support scaling the most complex AI models. This building block of the DGX SuperPOD enables training performance at a whole new level.

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

Table 1. DGX SuperPOD features

Component	Technology	Description
Compute Nodes	64 NVIDIA DGX-2 systems	1,024 Tesla V100 SXM3 GPUs 32 TB of HBM2 memory 128 AI petaFLOPS via Tensor Cores 96 TB System RAM 192 TB local NVMe
Compute Network	Mellanox CS7500 InfiniBand switch	648 EDR/100 Gbps ports per switch Eight connections per DGX-2 system
Storage Network	Mellanox CS7520 InfiniBand switch	216 EDR/100 Gbps ports InfiniBand Two connections per DGX-2 system
Management Networks	In-band: Mellanox SN3700C Out-of-band: Mellanox AS4610	Each DGX-2 system has Ethernet connections to both switches
Management Software	DeepOps DGX POD Management Software	Software tools for deployment and management of SuperPOD nodes, Kubernetes and Slurm.
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 design requirements for the DGX SuperPOD include:

- Design an infrastructure around the DGX-2 system to allow distributed DL applications to scale to hundreds of nodes. Use MLPerf benchmarks as a proxy for these applications.
- Design a compute fabric which optimizes the common communications patterns found in distributed DL applications and provides:
  - Full-fat tree connectivity between eight IB ports of every DGX-2 system in the cluster
  - Advanced traffic management
- Design a storage fabric which:
  - Scales to hundreds of ports
  - Provides single node bandwidth in excess of 10 GB/s
  - Leverages RDMA communications for the fastest, low-latency data movement.
  - Provides additional connectivity to share storage between the DGX SuperPOD and other resources in the data center
- Provide a hierarchical storage system which:
  - Minimizes time to stage data to local storage
  - Allows for training of DL models that require peak IO performance, exceeding 15 GB/s, and data sizes which exceed the local NVMe storage cache
  - Provides a large, cost-effective, LTS area for data that are not in active use
- Provide a user experience that allows management of complex multi-node and multi-job workflows
- Deploy and update the system quickly. Leveraging the reference architecture allows data center staff to develop a full solution with fewer design iterations.

Network Architecture

The DGX SuperPOD has four networks, a compute fabric, a storage fabric, an in-band management network, and an out-of-band management network.

The storage network uses two Mellanox ConnectX-5 NICs on the CPU baseboard of the DGX-2 system. One 100 Gbps port on each NIC can be configured for either InfiniBand or 100 Gbps Ethernet/RoCE.

The in-band management network uses the second 100 Gbps port on the first ConnectX NIC on the baseboard of the DGX-2 system. This link is connected to a separate 100 Gbps Mellanox Ethernet switch.

Finally, an out-of-band management network running at 1 Gbps connects the BMC port of each DGX-2 system to an additional Mellanox Ethernet switch.

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
64 DGX-2 Systems	512	128	64	64
4 Management Servers			8	4
Storage System[1]	X		X	X
1. The number of storage system connections will depend on the system to be installed.				

Compute Fabric

The high-performance compute fabric is a 100 Gbps/EDR InfiniBand-based network using the Mellanox CS7500 Director Switch (Figure 2). The CS7500 switch, which utilizes the ConnectX-5 architecture, provides a non-blocking fabric of up to 648 ports. Each DGX-2 system has eight connections to the compute fabric. Careful consideration was given to the fabric design to maximize performance for typical communications traffic of AI workloads, as well as providing some redundancy in the event of hardware failures and minimizing cost.

Note: Since the DGX SuperPOD was first built, Mellanox released the next-generation CS8500 director switch, based on the ConnextX-6 architecture, which supports up-to 800 HDR (200 Gb/s) ports or up-to 1600 HDR100 (100 Gb/s) ports. Please contact your NVIDIA representative for current status of ConnectX-6 support on the DGX SuperPOD.

Figure 2. Mellanox CS7500 director switch

The NVIDIA Collective Communications Library (NCCL) is the main communications library for deep learning. It uses communication rings and trees to optimize the performance of common collective communication operations used by deep learning applications.

Storage Fabric

The storage fabric is a 100 Gbps/EDR InfiniBand-base fabric using the Mellanox CS7520 Director Switch (Figure 3). The CS7520 switch provides 216 ports. The Director switch was selected for its ease of deployment and reduced cable complexity. Separating the storage traffic on its own fabric removes the congestion that could reduce application performance, and removes the need to purchase a larger switch to support both compute and storage communication.

Figure 3. Mellanox CS7520 director switch

Since the I/O requirements for the DGX SuperPOD exceed 15 GB/s, an InfiniBand-based fabric was essential to minimize latency and overhead of communications. With a substantial investment in the compute fabric, there is little additional management overhead for using the same technology for storage. High bandwidth requirements with advanced fabric management features such as congestion control and adaptive routing also benefit the storage fabric.

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 NVIDIA GPU Cloud registry, code repositories, and data sources.

Each DGX-2 system has one link to the In-Band Ethernet network and management nodes have two links. The In-band network is built using Mellanox SN3700C switches (Figure 4) running at 100 Gbps. 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 4. Mellanox SN3700C switch

Out-of-Band Management Network

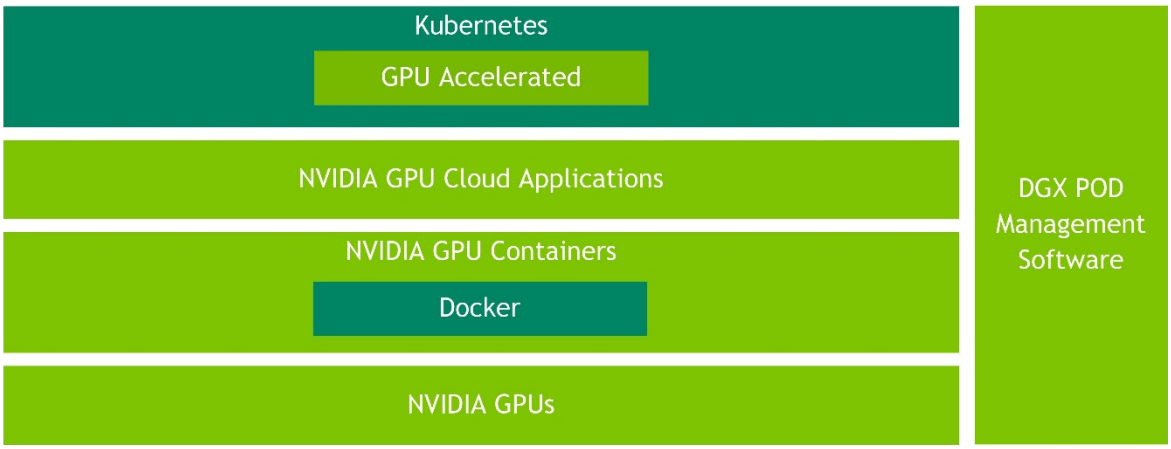
The out-of-band network is used for system management via the BMC and provides connectivity to manage all networking equipment. Out-of-band management is critical to the operation of the cluster by providing low usage paths that ensure management traffic does not conflict with other cluster services. The out-of-band management network is based on 1 Gbps Mellanox AS4610 switches (Figure 5). 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 5. Mellanox AS4610 switch

AI Software Stack

NVIDIA AI software (Figure 6) running on the DGX SuperPOD provides a high-performance DL training environment for large scale multi-user AI software development teams. It includes the DGX operating system (DGX OS), 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 open-source tools recommended by NVIDIA which have been tested to work on DGX POD racks with the NVIDIA AI software stack. Support for these tools can be obtained directly through third-party support structures.

Figure 6. AI software stack



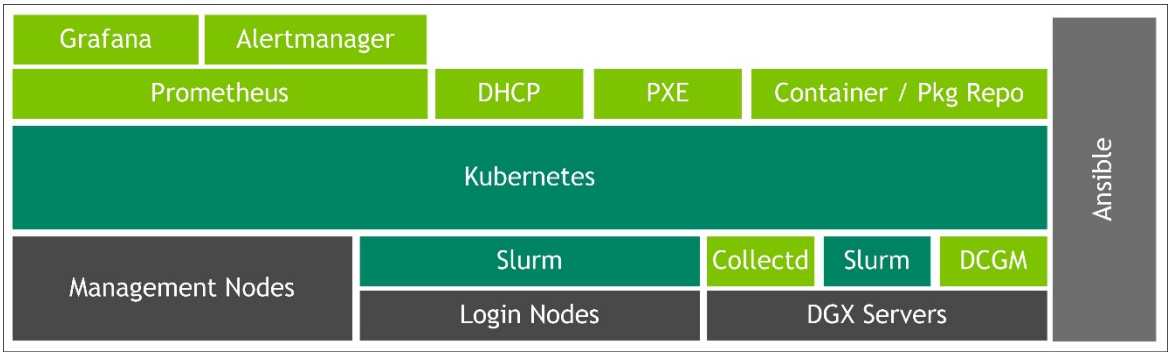
The foundation of the NVIDIA AI software stack is the DGX OS, built on an optimized version of the Ubuntu Linux operating system and tuned specifically for the DGX hardware. The DGX OS software includes certified GPU drivers, a network software stack, pre-configured NFS caching, NVIDIA data center GPU management (DCGM) diagnostic tools, GPU-enabled container runtime, NVIDIA CUDA® SDK, cuDNN, NCCL and other NVIDIA libraries, and support for NVIDIA GPUDirect™ technology.

The DGX POD management software (Figure 7) 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.

The DGX POD management software leverages the Ansible configuration management tool. Ansible roles are used to install Kubernetes on the management nodes, install additional software on the login and DGX systems, configure user accounts, configure external storage connections, install Kubernetes and Slurm schedulers, as well as perform day-to-day maintenance tasks such as new software installation, software updates, and GPU driver upgrades.

DGX POD monitoring utilizes Prometheus for server data collection and storage in a time-series database. Cluster-wide alerts are configured with Alertmanager, and system metrics are displayed using the Grafana web interface. 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.

Figure 7. DGX POD management software



Users access the system via a login node, which is one of the management nodes. The login node provides a user environment to edit code, access external code repositories, compile code as needed, and build containers. Training jobs are managed via Slurm. Training jobs run in containers, but the system also supports bare metal jobs.

Kubernetes runs management services on management nodes. Slurm runs user workloads and is installed on the login node as well as the DGX systems. Slurm provides advanced HPC-style batch scheduling features including multi-node scheduling and backfill.

The software management stack and documentation are available as an open-source project on GitHub.

User workloads on the DGX POD primarily utilize containers from NGC (Figure 8), which provides researchers and data scientists with easy access to a comprehensive catalog of GPU-optimized software for DL, HPC applications, and HPC visualization that take full advantage of the GPUs. The NGC container registry includes NVIDIA tuned, tested, certified, and maintained containers for the top DL frameworks such as TensorFlow, PyTorch, and MXNet. NGC also has third-party managed HPC application containers, and NVIDIA HPC visualization containers.

Figure 8. NGC overview

.

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.

In this section we describe two possible deployments:

- Computer room air handlers (CRAH) with contained cold air aisles
- Rear-door heat exchangers (RDHX) in a flooded-air room

Table 3 provides 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. The total power draw will be approximately 670 kW.

Table 3. Power usage

Equipment	Maximum Power
DGX-2 system	10 kW
Management/login node	.6 kW
Typical storage appliance	12 kW
Mellanox CS7500 director switch	13.5 kW
Mellanox CS7520 director switch	5 kW
Mellanox SN3700C switch	.5 kW
Mellanox AS4610 switch	.1 kW

For both designs, the data center should maintain cooling to ASHRAE TC9.9 2015 recommended thermal guidelines. The data center dewpoint temperature must be tightly controlled (i.e. WSHP) to avoid condensation.

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.

Compute Room Air Handler (CRAH)

The CRAH layout consists of 34 racks arranged in two clusters of contained air (Figure 9). Each cluster of 17 racks has a central area that is sealed to prevent air leakage. CRAHs on either end supply cool air under the raised floor and upward through perforated tile into the enclosed cold aisle. Air discharges through to the back of the racks, where it is taken into the CRAHs for conditioning.

Most of the racks are compute racks that contain two DGX-2 systems, and possibly an ethernet switch. The InfiniBand racks are centrally located to minimize cable lengths and to ease fiber cabling. The "Compute InfiniBand" rack contains the larger InfiniBand switch that is used for the compute traffic, as well as the management nodes. The "Storage InfiniBand + Storage" rack contains the smaller InfiniBand switch for storage traffic, as well as the storage appliance. Ethernet TORs and utility CPU nodes are spread throughout the racks.

The racks themselves are 48U tall, 700 mm wide, and 1200 mm deep. The extra width and depth are to ensure that the 0U PDUs and cabling do not interfere with maintenance of the DGX-2 systems. Replacing a GPU tray can be very challenging in smaller racks. The cabinets should support at least 600 kg of static load. Cable pathways should conform to TIA 942 standards.

Each rack has two 0U PDUs. Since the maximum rack power does not exceed 21 kW, 415V, 48A, three-phase PDUs can be used. This provides 34.5 kVA of redundant power.

Figure 9. Floor plan for CRAH layout

The CRAH layout consists of four rows of racks, arranged in two "closets". Figure 10 and Figure 11 illustrate the rack elevations of the closets, as viewed from the front of the racks. Each rack has two 0U PDUs. The equipment avoids the lowest 4U positions, as server lifts do not go that low, and as the velocity of air leaving the perforated tile can cause cooling issues in the lowest rack unit positions.

The four racks on either side form a unit of eight DGX-2 systems, with in-band and out-of-band Ethernet top-of-rack switches. 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. The compute InfiniBand switch includes management nodes for administrative tasks. The rack with the InfiniBand storage switch also has space for storage equipment.

Figure 10. Rack elevations for CRAH layout (closet 1)

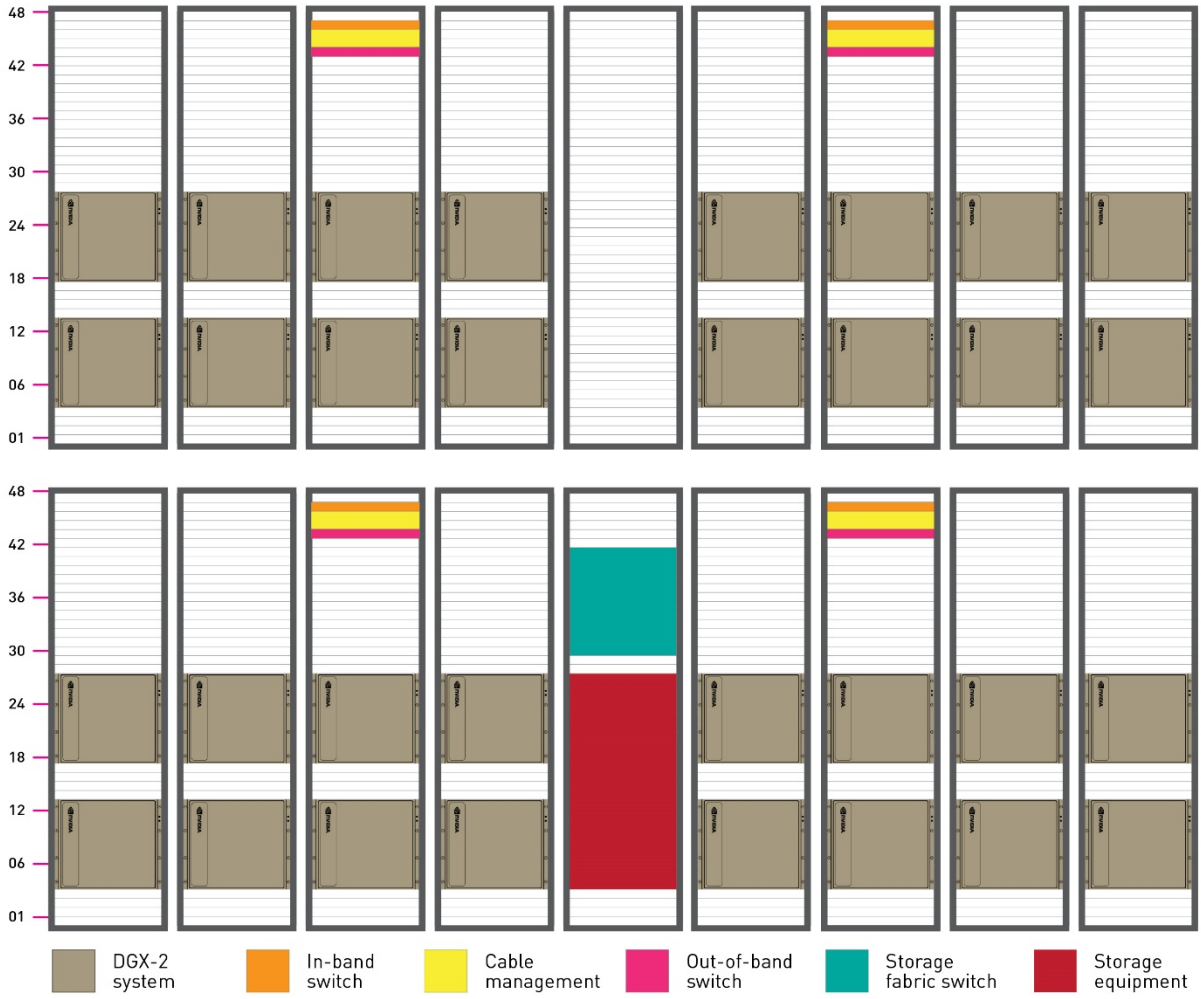
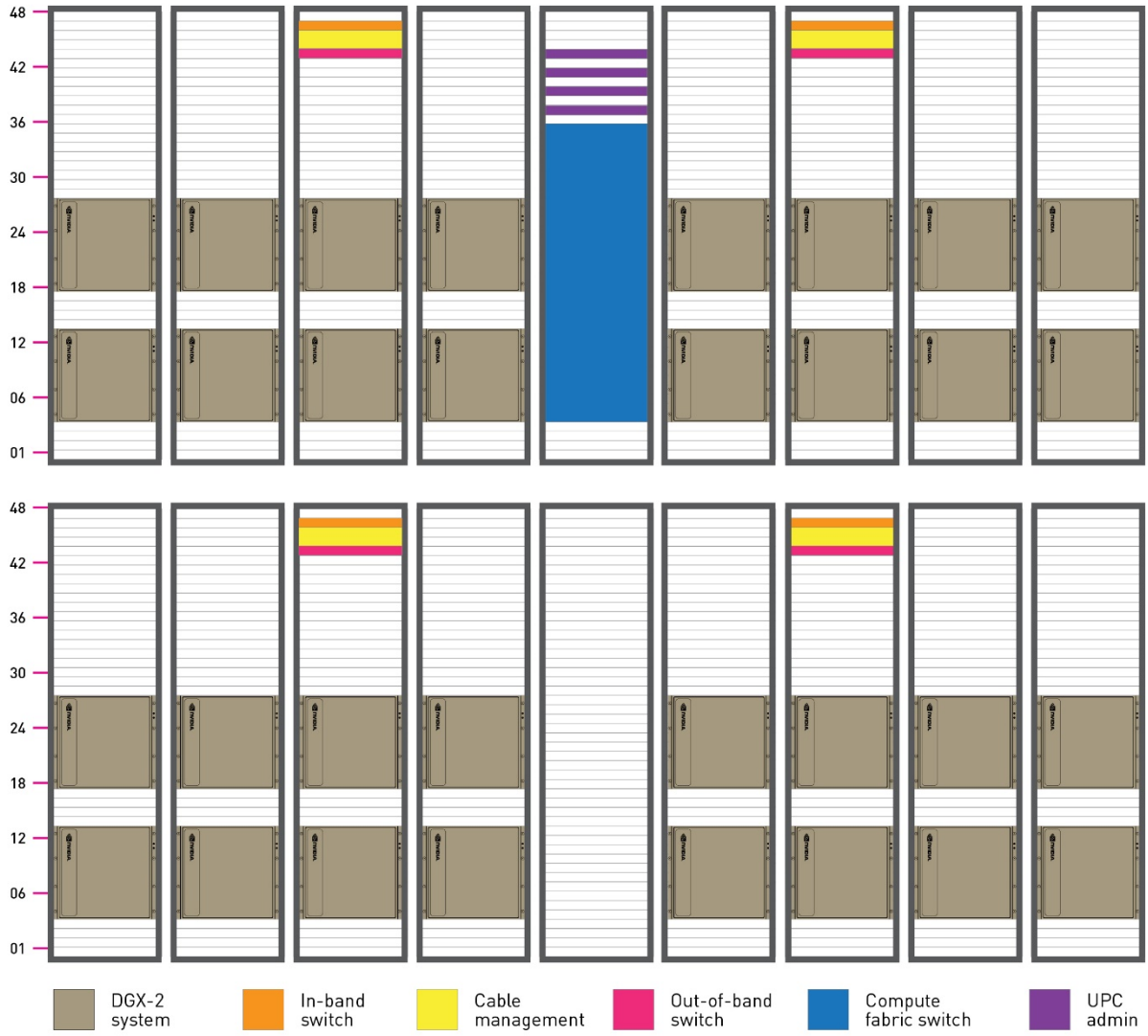


Figure 11. Rack elevations for CRAH layout (closet 2)



Rear Door Heat Exchanger (RDHX)

The RDHX layout consists of two rows of 24 racks (Figure 12). Each rack has a rear-door heat exchanger. InfiniBand racks are located centrally to minimize cable lengths and ease cabling. Each compute rack has three DGX-2 systems. As with the CRAH layout, Ethernet switches are distributed through the racks. Unlike the CRAH layout, there is no air containment, with all the racks in a flooded-air room.

The racks are 48U tall, 700 mm wide, and 1200 mm deep, and support closed-coupled cooling with active rear door heat exchangers. Special attention to the sealing around the rear door heat exchanger is needed to ensure proper airflow. Compared to the CRAH layout, these cabinets must support 1200 kg of static load. The extra load is due to the rear door heat exchangers and water, which weigh approximately 300 kg, and due to the additional DGX-2 systems, which weighs approximately 150 kg.

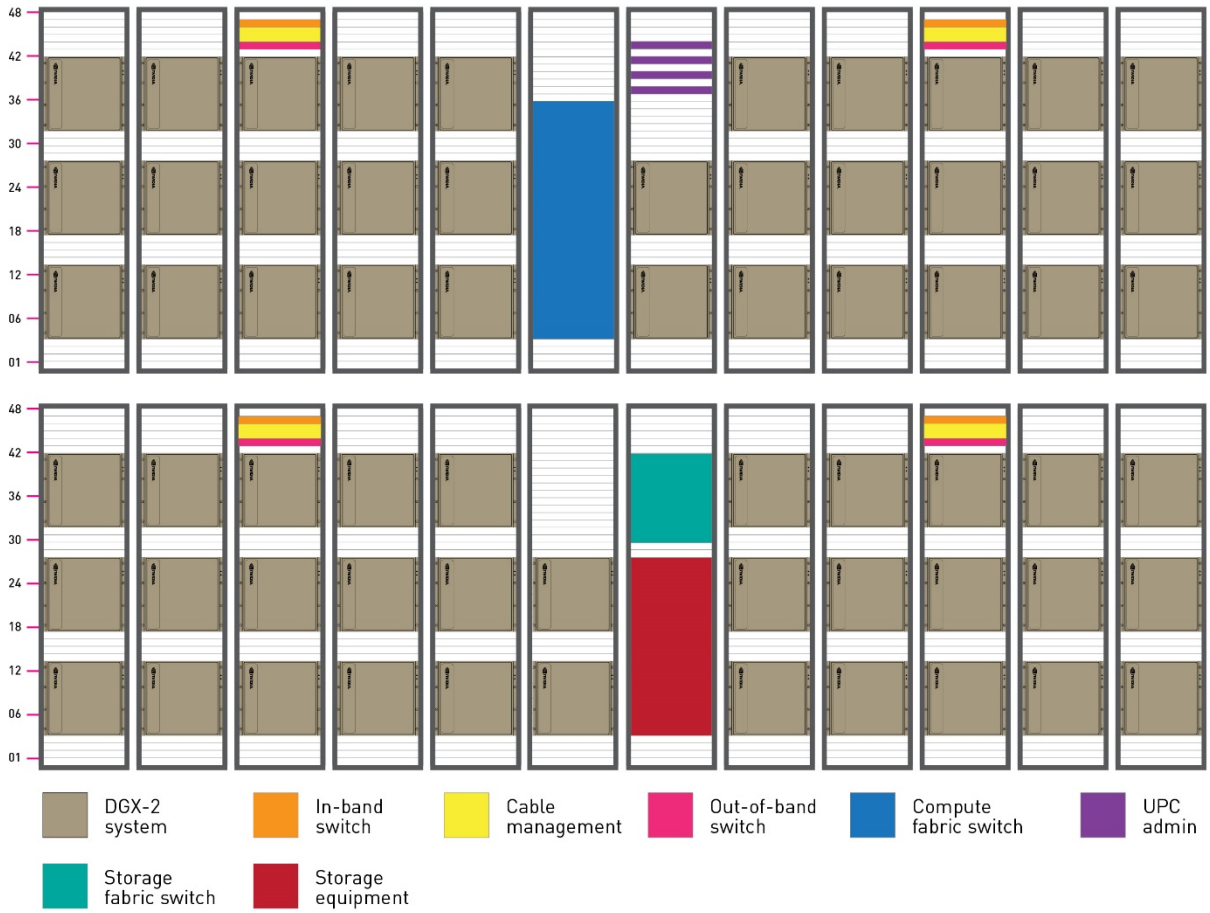
Each rack has two 0U PDUs. Since the maximum rack power does not exceed 31 kW, 415V, 48A, three-phase PDUs can be used. This provides 34.5 kVA of redundant power.

In this design, each CDU provides up to 400 kW of closed coupled cooling per 12 cabinets. Mechanical cooling should be N+1 redundant. For this design, the data center must have a leak detection and control system to avoid damage to equipment if there is a leak.

Figure 12. Floor plan for RDHX layout

Figure 13 shows the rack elevation for the RDHX layout. The contents of the racks are like that of the CRAH layout. One difference is that most of the compute racks have three rather than two DGX-2 systems. The two sides of each row are also asymmetric, with one extra cabinet with two DGX-2 system. As with the CRAH layout, the sides in each row can be built out in phases.

Figure 13. Rack elevations for RDHX layout



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 needs to be designed considering the hierarchy of different storage technologies, either persistent or non-persistent, to balance the needs of performance, capacity, and cost.

Table 4 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 4. DGX SuperPOD storage and caching hierarchy

Storage Hierarchy Level	Technology	Total Capacity	Read Performance
RAM	DDR4	1.5 TB per node	> 100 GB/s
Internal Storage	NVMe RAID	30 TB per node	> 25 GB/s
High-Speed Storage	Generic	Varies depending on specific needs	Required: Aggregate system read > 32 GB/s Aggregate system write > 16 GB/s Single-Node read > 5 GB/s Desired: Single-Node 1 GB/s read per GPU (16 GB/s)

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 to 1.5 TB on a DGX-2 system and that capacity must be shared with the operating system, application, and other system processes. The local storage on the DGX-2 system provides 30 TB of very fast NVMe (and can be upgraded to 60TB if required). While the local storage is fast, it isn't practical to manage a dynamic environment with local disk alone.

The high-speed storage provides a shared view of your organization's data to all nodes. It needs to be optimized for small, random IO 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 be support both efficient multi-threaded reads and writes from a single system but most of the DL workloads will be read dominant.

Datasets today that are 30 TB are still considered large, but we see use cases in automotive and other computer vision tasks where 1080p images are used for training and in some cases are uncompressed. Datasets in these formats can easily exceed 30 TB in size. In these cases, we see a need for 1 GB/s per GPU for read performance.

The metrics above assume variety of workloads, datasets, and needs for training locally and directly from the high-speed storage system. It's 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 platforms and DGX SuperPOD. A list of these partners is available here.

The storage hierarchy can be extended further. Long term storage (LTS) is often important to preserve data and historical results. LTS could as large as tens or hundreds of petabytes. Data from this tier is accessed infrequently, and store and recall performance isn't as critical. Solutions for this can be cost-optimized differently than the high-speed storage discussed above. Solutions could be based on slower spinning disk, use S3 compatible object storage technologies, or even use the cloud.

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-2 system, marks a major milestone in the evolution of supercomputing, offering a solution that any enterprise can acquire and deploy to access massive computing power to propel business innovation. 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's backed by 24x7 enterprise-grade support.

Appendix A. Major Components

This section outlines the major components of the two POD designs. The section covers the CRAH design and the second one covers that RDHX design Cables are listed in separate tables.

Note: Part numbers for the active optical cables indicate the 15m length. It is possible to substitute shorter lengths, but it is not advised to substitute direct-attached copper cables for the optical cables.

Compute Room Air Handler (CRAH)

Major components for the CRAH configuration are listed in Table 5, and associated tables are listed in Table 6.

Table 5. Major components of the DGX SuperPOD (CRAH)

Count	Recommended Model	Component
Racks		
32	NVIDPD05	Rack (AFCO)
Nodes		
4	SMC SSG-6019P-ACR12L	CPU Nodes
64	NVIDIA DGX-2 systems	GPU Nodes
1	No Recommendation	High-Speed Storage
1	No Recommendation	Long-Term Storage
Ethernet Network		
8	Mellanox SN3700C	In-Band TOR
8	Mellanox AS4610	Out-of-Band TOR
InfiniBand Fabric		
1	Mellanox CS7500	Compute Director Switch Chassis
1	Mellanox SUP-CS7500-3S	Compute Director Switch Support and Warranty
2	Mellanox MMB7500	Compute Director Switch Management Module
18	Mellanox MSB7570-E	Compute Director Switch 36-port spine blade
18	Mellanox MSB7560-E	Compute Director Switch 36-port leaf blade
1	Mellanox CS7520	Storage Director Switch Chassis
1	Mellanox SUP-CS7520-3S	Storage Director Switch Support and Warranty
2	Mellanox MMB7500	Storage Director Switch Management Module
6	Mellanox MSB7570-E	Storage Director Switch 36-port spine blade
6	Mellanox MSB7560-E	Storage Director Switch 36-port leaf blade
Power and Cooling		
68	Raritan PX3-5747V-V2	PDUs
4	No recommendation	CRAH Units

Table 6. Cables for the DGX SuperPOD (CRAH)

Count	Recommended Model	Component
In-Band Ethernet Cables		
64	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (DGX-2 system)
8	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (CPU)
TBD[1]	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (Storage)
16	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (2 uplinks per TOR)
Out-of-Band Ethernet Cables		
64	No Recommendation	Cat5 cable (DGX-2 system)
4	No Recommendation	Cat5 cable (CPU)
TBD[1]	No Recommendation	Cat5 cable (Storage)
68	No Recommendation	Cat5 cable (PDUs)
16	Mellanox MC3309130-xxx	10 GbE passive copper cable SFP+ (2 uplinks per TOR)
Compute InfiniBand Cables		
512	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (DGX-2 system)
Storage InfiniBand Cables		
128	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (DGX-2 system)
TBD[1]	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (Storage)
1. Count required depends on specific storage selected		

Rear Door Heat Exchanger (RDHX)

Major components for the RDHX configuration are listed in Table 7, and associated tables are listed in Table 8.

Table 7. Major components of the DGX SuperPOD (RDHX)

Count	Recommended Model	Description
Racks		
24	Vertiv E48721	Rack
Nodes		
4	SMC SSG-6019P-ACR12L	CPU Nodes
64	NVIDIA DGX-2 systems	GPU Nodes
1	No Recommendation	High-Speed Storage
1	No Recommendation	Long-Term Storage
Ethernet Network		
4	Mellanox SN3700C	In-Band TOR
4	Mellanox AS4610	Out-of-Band TOR
InfiniBand Fabric		
1	Mellanox CS7500	Compute Director Switch Chassis
1	Mellanox SUP-CS7500-3S	Compute Director Switch Support and Warranty
2	Mellanox MMB7500	Compute Director Switch Management Module
18	Mellanox MSB7570-E	Compute Director Switch 36-port spine blade
18	Mellanox MSB7560-E	Compute Director Switch 36-port leaf blade
1	Mellanox CS7520	Storage Director Switch Chassis
1	Mellanox SUP-CS7520-3S	Storage Director Switch Support and Warranty
2	Mellanox MMB7500	Storage Director Switch Management Module
6	Mellanox MSB7570-E	Storage Director Switch 36-port spine blade
6	Mellanox MSB7560-E	Storage Director Switch 36-port leaf blade
Power and Cooling		
48	Raritan PX3-5747V-V2	PDUs
4	No Recommendation	CDUs

Table 8. Cables for the DGX SuperPOD (RDHX)

Count	Recommended Model	Description
In-Band Ethernet Cables		
64	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (DGX-2 system)
8	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (CPU)
TBD[1]	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (Storage)
8	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (2 uplinks per TOR)
Out-of-Band Ethernet Cables		
64	No Recommendation	Cat5 cable (DGX-2 system)
4	No Recommendation	Cat5 cable (CPU)
TBD[1]	No Recommendation	Cat5 cable (Storage)
48	No Recommendation	Cat5 cable (PDUs)
8	Mellanox MC3309130-xxx	10 GbE passive copper cable SFP+ (2 uplinks per TOR)
Compute InfiniBand Cables		
512	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (DGX-2 system)
Storage InfiniBand Cables		
128	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (DGX-2 system)
TBD[1]	Mellanox 930-20000-0007-000	100 GbE QSFP to QSFP AOC (Storage)
1. Count required depends on specific storage selected		

Notice

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation ("NVIDIA") does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. 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 publication supersedes and replaces all other specifications for the product that may have been previously supplied.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification 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. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this specification.

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 these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application 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 specification. NVIDIA does not accept any 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 specification, 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 specification. 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 specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

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. 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 NVIDIA terms and conditions of sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA DGX SuperPOD, NVIDIA DGX POD, NVSwitch, NVLink, Tesla, and GPU Direct 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