NVIDIA DGX SuperPOD: Data Center Design

Reference Guide

Featuring NVIDIA DGX A100 Systems

DG-10265-001 | February 2023

Document History

DG-10265-001

Version	Date	Authors	Description of Change
01	2021-02-12	Ali Heydari, Dennis O'Brien, Jeremy Rodriguez, Premal Savla, Robert Sohigian, and Steven Hambruch	Initial release
02	2021-02-23	Alex Naderi, Jeremy Rodriguez, Premal Savla, and Robert Sohigian	Electrical updates
03	2021-03-26	Brian Forbes, Robert Sohigian, and Scott Ellis	Miscellaneous updates
04	2021-08-23	Dennis O'Brien, Premal Savla, Robert Sohigian, and Steven Hambruch	Cabling updates for NVIDIA Base Command™ Manager
05	2021-10-29	Premal Savla and Robert Sohigian	Updates to cabling counts and other minor changes.

Abstract

The NVIDIA DGX SuperPOD™ with NVIDIA DGX™ A100 systems is the next generation artificial (AI) supercomputing infrastructure, providing the computational power necessary to train today's state-of-the-art deep learning (DL) models and to fuel innovation well into the future. The 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 document provides guidelines for selecting the right data center to deploy a DGX SuperPOD: a reference architecture (RA) that is the result of codesign between DL scientists, application performance engineers, system architects, and data center architects to build a system capable of supporting the widest range of DL and HPC workloads.

This design introduces compute building blocks called scalable units (SU) enabling 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 in customer sites around the world, as well as being leveraged within the 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 can leverage the NVIDIA DGX SuperPOD Solution for Enterprise that 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.

Contents

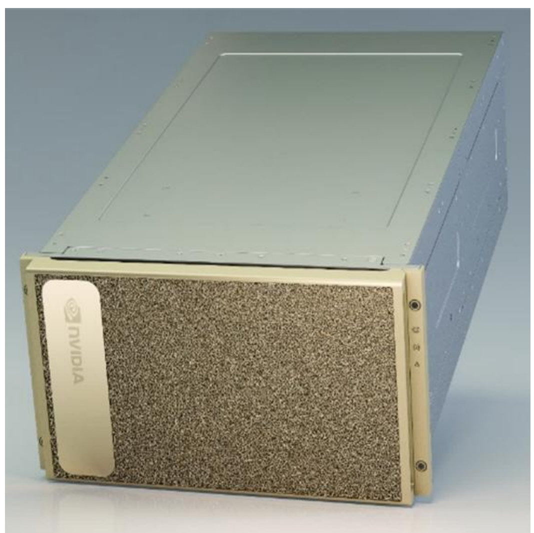
1. NVIDIA DGX A100 System Overview 1
2. Planning a Data Center Deployment 3
   2.1 Data Center Components Overview 4
3. Standard SU Elevations 5
   3.1 Rack Elevations 5
      3.1.1 Scalable Unit 6
      Four DGX A100 Systems Per Rack 7
   3.2 Two DGX A100 Systems Per Rack 8
4. Electrical Specifications 9
   4.1 Component Power Usage 9
   4.2 Data Center Power Configuration 10
   4.3 Power Redundancy 11
   4.4 Planning and Deploying Power Connections 11
   4.5 Rack Power Distribution Unit (rPDU) Selection 12
      4.5.1 Horizontal rPDU 12
      4.5.2 Vertical rPDU 13
   4.6 Phase Balancing 13
5. Rack Specifications 14
   5.1 Rack Design Considerations 14
   5.2 Rack Options 15
   5.3 Rack Mounting and Seismic Bracing 16
   5.4 Rack Selection versus Cable Lengths 16
      5.4.1 Rack Width Variations 17
   5.5 Server Mounting Requirements 17
   5.6 Device Installation 18
      5.6.1 Device Rail Kits 18
   5.7 Air Flow Management 19
      5.7.1 Equipment Orientation 19
   5.8 Static Weight and Point Load 20
6. InfiniBand Networking 21
7. Data Center Thermal Specifications 23
   7.1 Environmental and Thermal Guidelines 23
   7.2 Air Purity 24
8. Data Center Cooling 25
   8.1 Air Cooling Methods 27

8.1.1 Traditional Plenum Environment 27
8.1.2 Cold Aisle Containment 27
8.1.3 Hot Aisle Containment 28
8.2 Containment System Cable Paths 29
8.3 Air Containment 30
8.3.1 Rack Air Containment 30
8.4 Floor Air Containment 31
8.5 Open Space Containment 31
9. Summary 32
A. Scaling Deployments viii
A.1 Adjacent InfiniBand Deployments viii
A.2 Storage Growth viii
A.3 Spine Growth viii
B. Data Center Optimizations ix
B.1 CFD Modeling ix
B.2 Supplemental Cooling xii
B.2.1 In-Row Cooling xii
Active Rear Door Heat Exchanger xiii
C. Safety Considerations xiv
C.1 Noise xiv
C.2 Fire Protection xiv
C.3 Server Lifts xiv
D. Advanced Electrical Planning xvi
E. Industry Standards Reference xvii

1. NVIDIA DGX A100 System Overview

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



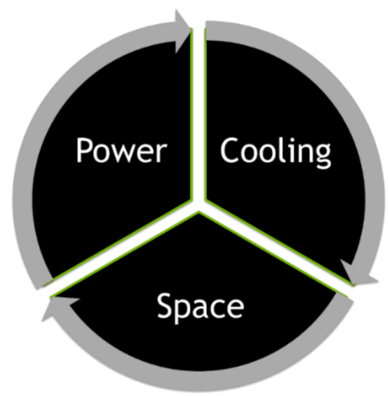
Specifications for the DGX A100 system that are integral to data center planning are shown in Table 1.

Table 1. DGX A100 system specifications

Specification	Value
System power usage	6.5 kW max
System weight	271 lb (123 kg)
Rack units	6

Height / width / length	10.4 in / 19.0 in max / 35.3 mm max (264.0 mm / 482.3 mm max / 897.1 mm max)
Operating temperature	41ºF to 86ºF (5ºC to 30ºC)
Cooling	Air
Airflow	840 CFM front-to-back @ 80% fan PWM | ∆ T (°C) à 13.9 C @ 80% fan PWM
Heat output	22,179 BTU/hr
Power consumption	6.5 kW (DC Side) / 6.6 kVA (AC side)

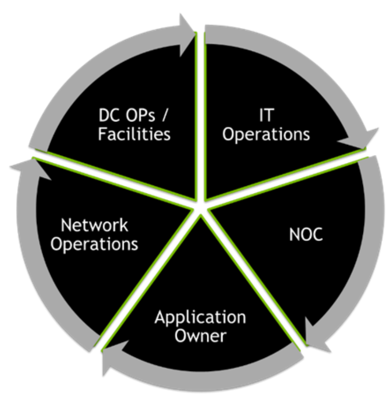
2. Planning a Data Center Deployment



Planning a DGX SuperPOD deployment requires the coordination and alignment of multiple teams within an organization. Impacted teams may include the application owners, end-users, data center operations, facilities teams, IT security, various information technology and networking teams, and network operations center support teams. Deployments can also impact third-party vendors who provide critical services to those teams. The teams must be aligned to ensure a well planned and executed installation of a DGX SuperPOD configuration.

Changes that occur in one technical domain can impact other domains. For example, if there is a data center site-level cooling constraint that limits rack density and causes the implementation to be distributed across a higher number of rack footprints, not only will that impact the data center facility's floor layout plan, but it will likely also impact the network design—due to necessary recalculation of cable lengths and possible latency-related performance impact. Good alignment, coordination, and communication among the various domain experts at every phase of the design and implementation of a DGX A100 system deployment will create the best result for a company.

Data centers have finite resources, each allocated to support a vital aspect of system operation. As in any economy, the limitation of resources creates scarcity in the face of demand. Therefore, it is usually necessary for organizations to optimize resource utilization in the data center environment.



Inefficient resource utilization can lead to significant direct and indirect costs to the organization.

The three main resource constraints in a data center environment are power, cooling, and space. These resource constraints are interrelated, such that excess demand in one resource domain can negatively impact other resource domains. For example, a cooling limitation that causes servers to be distributed across many racks rather than consolidated into fewer racks impacts space efficiency.

2.1 Data Center Components Overview

Data centers each have their own unique constraints regarding power, cooling, and space resources. NVIDIA has developed two different baseline configurations to address the two most common deployment patterns. However, should customization be necessary to address specific data center parameters in a particular deployment, NVIDIA can typically accommodate. It is very important to work with NVIDIA and communicate any data center design constraints, so that a performance optimized deployment can be achieved. Altering the deployment pattern without such consultation can lead to serious performance, operational, support, or scalability challenges.

NVIDIA has created scalable units (SU) which can be combined for large deployments. These deployment patterns are optimized for both performance and cost. Each SU consists of 20 DGX A100 systems plus associated InfiniBand leaf switch connectivity infrastructure. A full DGX SuperPOD has seven SUs (140 total DGX A100 systems) which are interconnected using a shared InfiniBand fabric.

The two baseline DGX SuperPOD configurations for these SUs are shown in Table 2.

Table 2. Baseline rack configurations for a single SU (20 node) DGX SuperPOD

DGX A100 Systems Per Rack	Compute Racks	Networking Racks	Total SU Power Requirement	Total Power Per Rack Footprint
2	10	1	139 kW	13 kW
4	5	1	139 kW	26 kW

In addition to these racks, every installation requires two racks for management, networking (InfiniBand spine and Ethernet switches) and storage infrastructure. Each of these two racks consumes up to 21 kW. These are common to all SUs within the DGX SuperPOD architecture.

The interleaved design of the InfiniBand fabric within the DGX SuperPOD architecture presents constraints regarding cable path distance. Therefore, the deployment patterns are modeled with careful attention to cable length.

3. Standard SU Elevations

This section covers the standard SU:

- Compute racks containing DGX A100 systems.
- Leaf rack for InfiniBand switching, as well as in-band and out-of-band Ethernet networking.
- Management rack.
- Data storage rack.

> **Note:** In many illustrations throughout this document, racks are shown without side panels to illustrate the devices more clearly within the rack. In actual practice, side panels should be used on all racks.

3.1 Rack Elevations

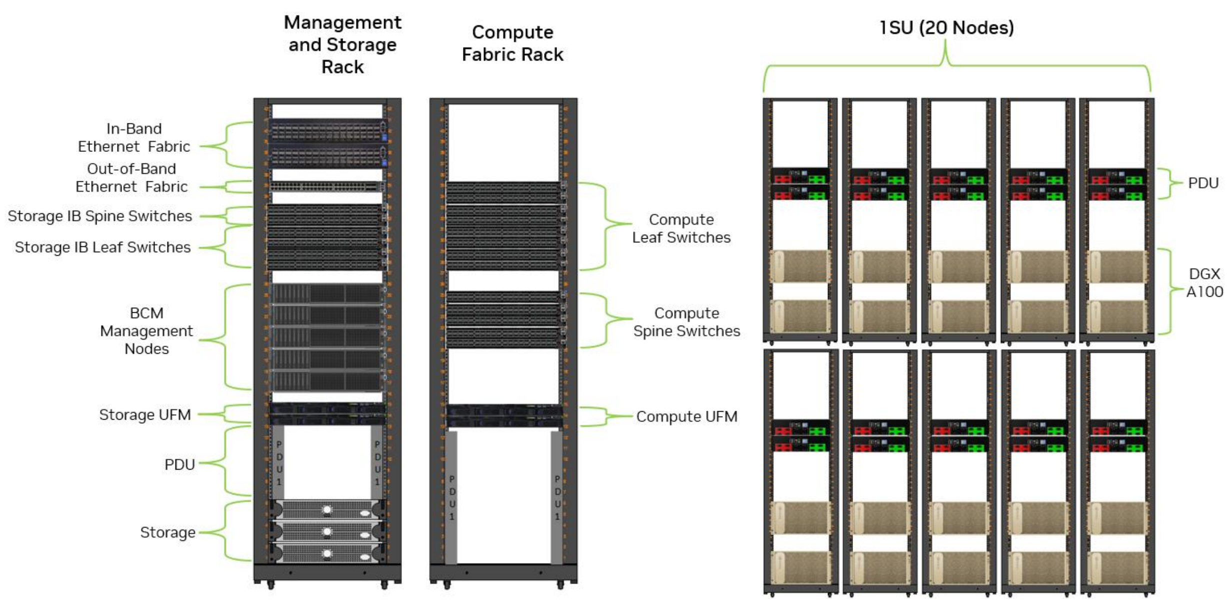
The reference design for an SU has five compute racks, each with four DGX A100 systems. Every SU also has a network and management rack, centrally located between the compute racks to minimize cable lengths. The upper-level spine switches, management nodes, core network, and storage components occupy 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 1,200 mm deep. The extra width and depth ensure that the 0U PDUs and InfiniBand 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 racks should support a minimum static load of 600 kg. Cable pathways should conform to TIA 942 standards.

3.1.1 Scalable Unit

Each SU consists of 20 DGX A100 systems. The standard configuration (Figure 2) is with four DGX A100 systems per rack, coupled with compute and management 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 2. Single SU configuration with supporting components



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

The compute fabric rack 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.

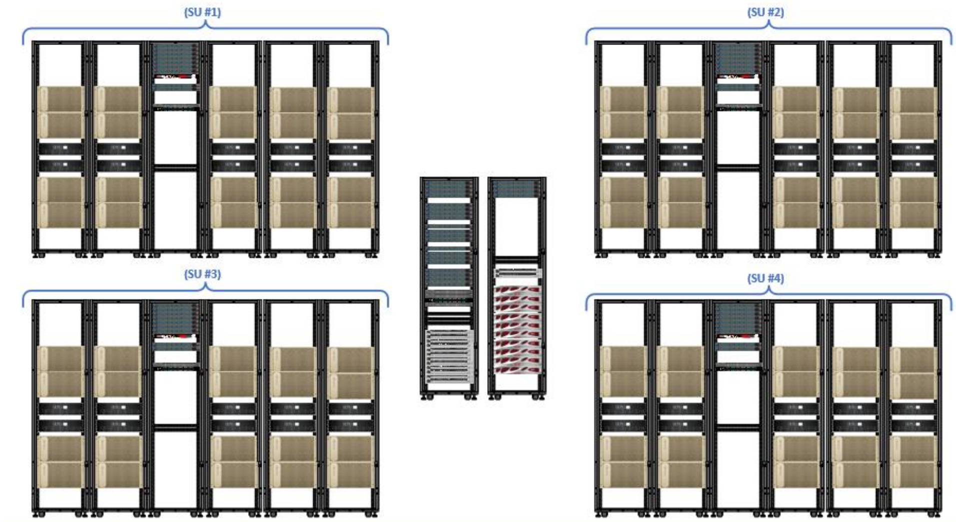
Descriptions of the management nodes and UFM appliances are in the *NVIDIA DGX SuperPOD: Scalable Infrastructure for AI Leadership Reference Architecture*.

> **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 constraints.

Four DGX A100 Systems Per Rack

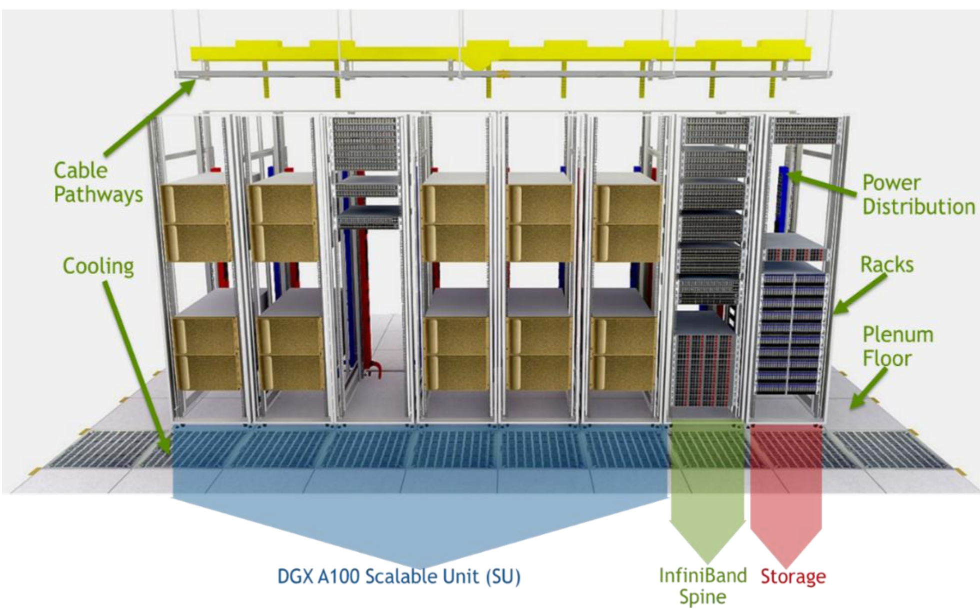
Figure 3 shows a configuration on four DGX A100 systems per rack.

Figure 3. Configuration with four DGX A100 systems per rack



Error! Reference source not found. shows a DGX SuperPOD design consisting of four SUs and the representation of the infrastructure footprint needed to support the solution.

Figure 4. 80 node configuration with four DGX systems per rack



3.2 Two DGX A100 Systems Per Rack

Error! Reference source not found. shows a configuration based on two DGX A100 systems per rack.

Figure 5. Configuration with two DGX A100 systems per rack

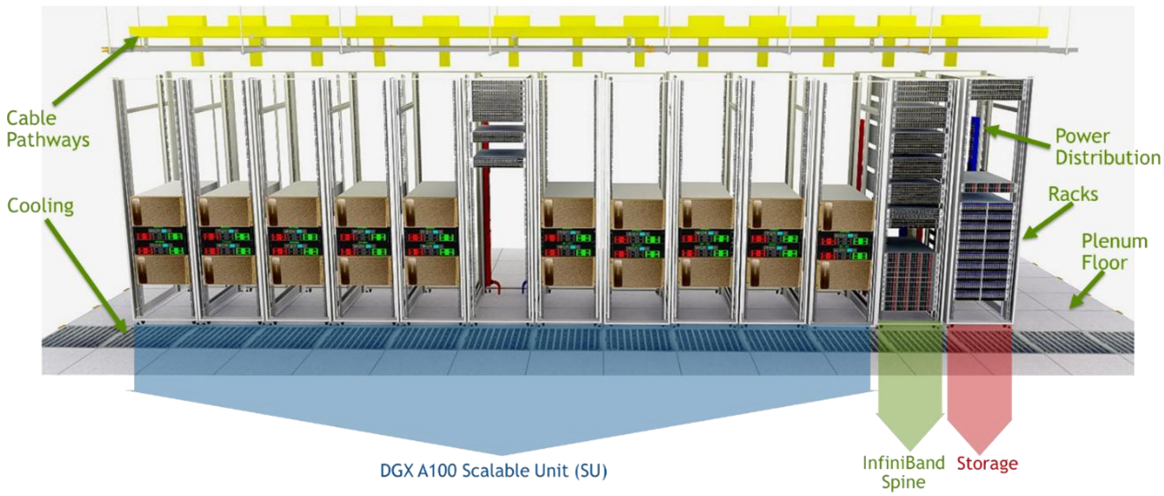
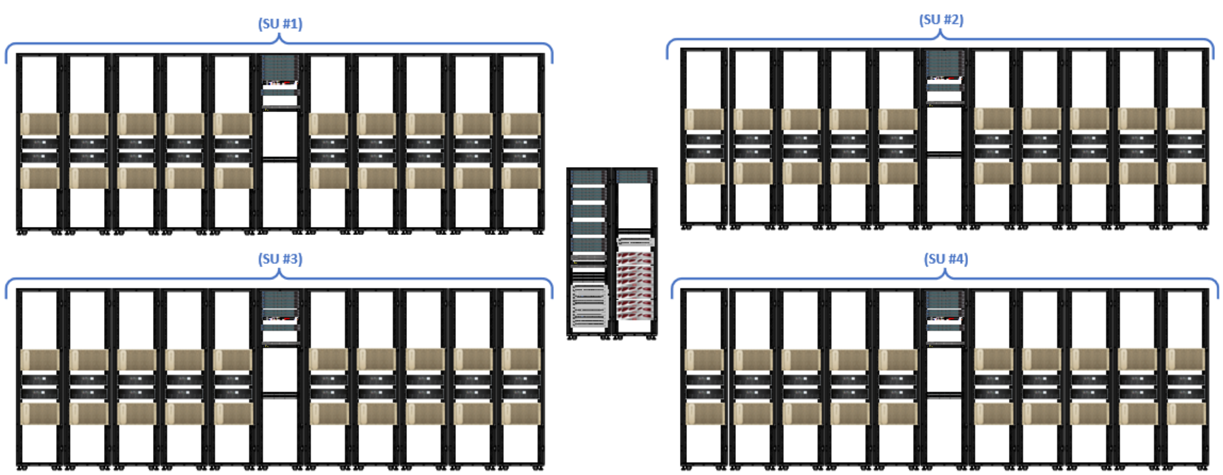


Figure 6 shows a DGX SuperPOD design consisting of four SUs and the representation of the infrastructure footprint needed to support the solution.

Figure 6. 80 node configuration with two DGX systems per rack



4. Electrical Specifications

This section details the power and electrical considerations for operating a DGX SuperPOD.

4.1 Component Power Usage

Table 3 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 information in Rack Elevations to compute the power per rack.

Table 3. Per component power usage

Equipment	Maximum Power
DGX A100 system	6.50 kW
Management nodes	0.60 kW
NVIDIA Quantum QM8790 switch	0.65 kW
NVIDIA SN34600C switch	0.50 kW
NVIDIA AS4610 switch	0.10 kW

Each SU requires 139 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 20 kW) is approximately 1 MW. The rack layouts can be altered to match the power distribution and per-rack cooling requirements for a specific data center.

4.2 Data Center Power Configuration

Redundant (2N) UPS/generator supported power circuits should be provisioned for each rack in a DGX A100 system deployment. Circuit capacity (as determined by voltage, phase count, and de-rated breaker rating) will determine the correct operating envelope for the total number of DGX A100 systems per rack. Table 4 provides an overview of typical circuit capacities in comparison with DGX A100 system maximum power requirements. The table shows that for higher rack density deployments, it may be necessary to provide more than one circuit from each redundant power source, depending on circuit sizing.

Table 4. Power requirements and circuit sizing matrix

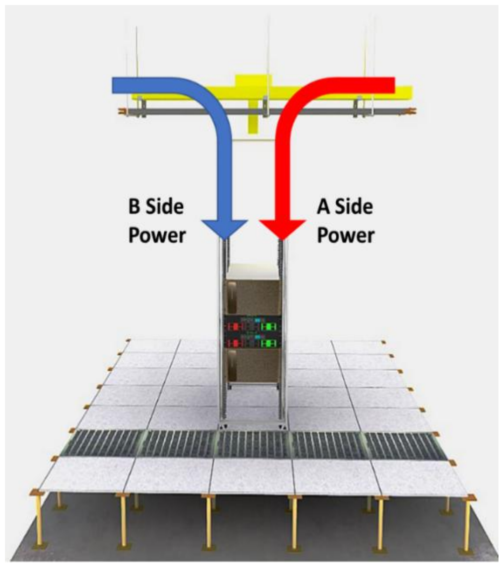
Number of DGX A100 Systems Supported	Circuit Sizing					
	Phase	Volts	Amps	80% Breaker De-Rating	kVA	Circuit kW
0	1	208	30	Yes	5.0	4.7
1	1	208	60	Yes	10.0	9.5
1	1	230	32	No	7.4	7.0
2	1	230	63	No	14.5	13.8
2	3	208	60	Yes	17.3	16.4
2	3	415	30	Yes	17.3	16.4
3	3	415	30	No	21.6	20.5
3	3	400	32	No	22.2	21.1
4	3	415	60	Yes	34.5	32.8

Note: See Table 14 for a detailed table of power requirements and circuit sizing.

4.3 Power Redundancy

At the rack level, each rack will minimally have two rack power distribution units (rPDUs) of equal type and sizing, whose power whips are connected to independent data center PDUs (Figure 7). This provides 2N power redundancy or increased capacity to the rack . Data center racks shall be earthed/bonded to the facility ground system. rPDU features must include overall power monitoring (discrete outlet monitoring is not required), remote switching of outlets, and temperature monitoring.

Figure 7. Redundant power concept



4.4 Planning and Deploying Power Connections

Follow these best practice guidelines when connecting AC power to the racks and systems:

- Power provisioning should be completed with the data center before connecting power to the rPDUs and system deployment.
- Power connections should be labeled to indicate the source of power (PDU #) and the specific circuit breaker numbers used within each PDU.
- Power cables should be color coded to help achieve and ensure proper redundancy.
- An electrician or facilities representative should verify that the AC voltage and total kW supplied are within specification at each of the floor-mounted PDUs and individual circuits (i.e., power drops) that feed the racks.
- An electrician or facilities representative should perform AC verification testing at each rack by turning off the individual circuit breakers feeding each rack power strip to verify that power redundancy has been achieved in each rack.

4.5 Rack Power Distribution Unit (rPDU) Selection

In addition to provisioning optimal power circuits, selecting appropriate rPDUs is critical. When single circuits are being provisioned, vertical (zero U) rPDUs may be selected. Horizontal rPDUs will typically be required when multiple circuits are being provisioned to each rack.

The DGX SuperPOD design requires these rPDU features:

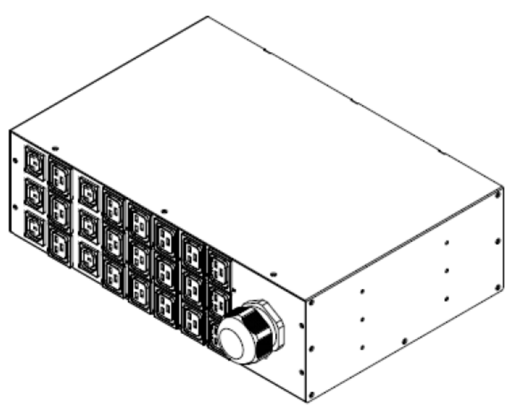
- Integrated smart module.
- Network interface.
- SNMP access with a publicly available MIB and a recommended polling rate of least once per second.
- REST API interface.
- Ports for temperature and sensor probes.
- Locking receptacles.
- rPDU-level metering.
- Remote outlet switching per receptacle.

The DGX SuperPOD can be extremely demanding on power requirements. To assist data center operators with this requirement, NVIDIA has worked with PDU OEMs to provide several rPDU options.

4.5.1 Horizontal rPDU

The Raritan PX3-5878I2R-P1Q2R1A15D5 rPDU (Figure 8) supports the power demands of DGX A100 systems.

Figure 8. Horizontal rPDU



Basic specifications are:

- Unit Dimensions (W*D*H): 440 mm × 293 mm × 132 mm; 17.3" × 11.5" × 5.2".
- Receptacles: (6) C13, 12A (18) C19, 16A.

4.5.2 Vertical rPDU

The Raritan PX3-5747V-V2 rPDU (Figure 9) supports the power demands of DGX A100 systems.

Figure 9. Vertical rPDU

Basic specifications are:

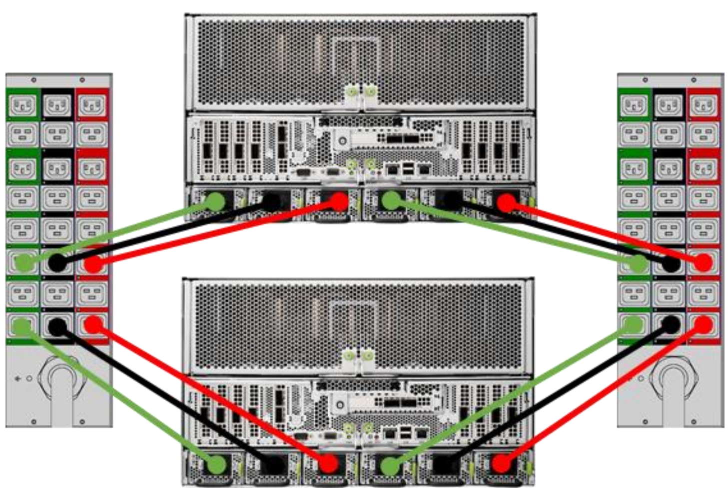
- Unit Dimensions (W*D*H): 109mm × 80mm × 1981mm; 4.3" × 3.1" × 78.0".
- Receptacles: (3) C13, 12A (18) C19, 16A.

> **Note:** Due to its height, this rPDU requires at least a 48U rack.

4.6 Phase Balancing

Ensure that power draw across the phases of a three-phase circuit is as balanced as possible (Figure 10). Each power cord of each server should be connected to a different *leg* (or phase) on the rPDU. For single phase rPDUs that include onboard circuit breakers, the power cords would be balanced across those onboard circuits as well. The onboard metering function of the rPDU provides an indication of power draw per phase or circuit, to assist in this.

Figure 10. Phase balancing



5. Rack Specifications

Racks must conform to EIA-310 standards for enclosed racks with 19" EIA mounting. Racks must be at least 24" × 40" (600 mm × 1,000 mm) in size, and at least 42U tall. NVIDIA recommends 30” × 48” × 52U (700 mm × 1,200 mm) racks. Racks shall provide a total of at least 96 in^2 (619 cm^2) of brush grommet protected cable ingress and egress access in the rear of the rack top, preferably through two openings.

IT racks come in a variety of sizes and are often designed for specific purposes. Each rack OEM follows specific minimum EIA-310 standards to ensure that industry standard devices fit properly. But OEMs will enhance the EIA-310 standard with their own unique designs and features that allow them to stand out in the marketplace. These features may include:

- Air management.
- Cable management.
- Modular subcomponents.
- Removable components.
- Proprietary accessories.
- Supplemental security devices.
- Custom manufacturing, company logos, etc.

5.1 Rack Design Considerations

Some design considerations when choosing the depth, width, and height of racks are:

- Rack depth:
  - The DGX A100 system has a minimum depth of 36 inches.
  - A rack depth of 48 inches provides the required depth for vertical rPDUs and cable management.
- Rack width:
  - Many data center operators standardize on a rack width of 24 inches (600 mm) to align with the width of floor tiles, and to maximize the number of rack footprints per row.
  - For DGX A100 system InfiniBand deployments, a rack width of 24 inches will present challenges for the leaf and spine racks that require a significant amount

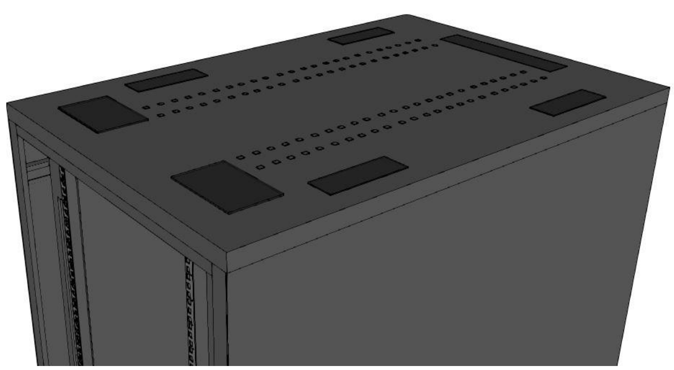
of InfiniBand cable termination. 32-inch (800 mm) racks should be considered for leaf and spine racks.
  - Wider racks should be considered when using 0U rPDUs because cables can get in the way of servicing DGX A100 systems and other components.
- Rack height:
  - Standard rack heights are 42U, 48U, and 52U.
  - A taller rack might seem more advantageous to mount more equipment within the rack footprint. But this also must be evaluated against the power and cooling available to the rack.
  - Some data centers may have pre-engineered cable trays, power conduits/busways, or aisle containment systems that may limit rack height options.

5.2 Rack Options

When ordering racks the following options and considerations are:

- Rack doors. Unless mandated by facility operations or security policy, front and rear doors are not required. Eliminating the front and rear doors helps improve airflow and reduces cost.
- Side panels. Side panels should be added and installed to maximize airflow management.
- Grounding and bonding kit. This kit is often an optional item and may not come with the rack. Proper grounding and bonding are essential.
- Blanking panels. Install blanking panels in each unused RU position to prevent exhaust air recirculation.
- Rack top options and accessories. Many rack manufacturers offer a range of rack top options and accessories to facilitate cooling management and cable routing management (Figure 11).

Figure 11. Rack top view with cable pass-though openings



5.3 Rack Mounting and Seismic Bracing

For the safety of staff members and equipment, all racks shall be fastened to the floor surface in accordance with manufacturer recommendations. This may involve bolting the rack using flanges or mount points designed for that purpose.

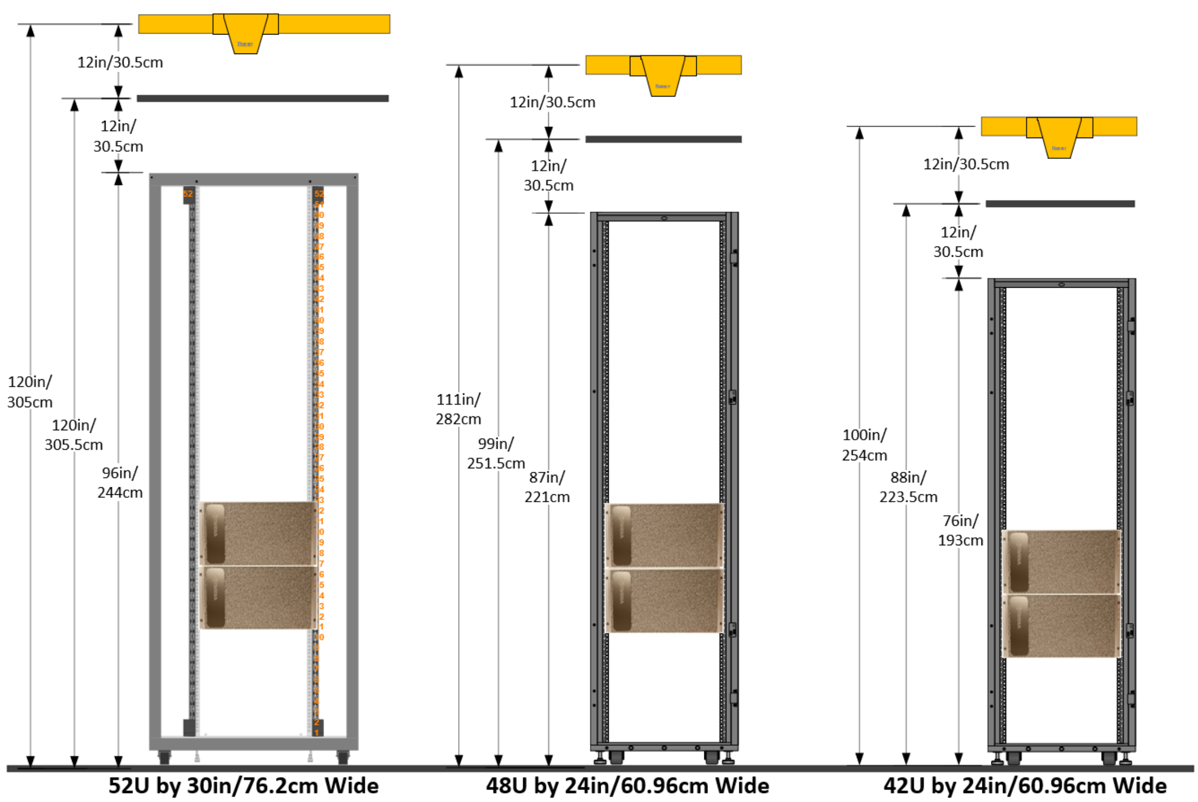
Additionally, wherever mandated by local Authorities Having Jurisdiction (AHJs), racks mounted on raised floor surfaces may require seismic bracing to be installed in the subfloor space. Seismic bracing should be designed and installed by qualified licensed structural engineers specializing in seismic engineering.

5.4 Rack Selection versus Cable Lengths

The rack height and width combined with the overhead cable tray specification and layout will affect cable lengths. Cable length in InfiniBand networks is a critical performance factor. Point-to-point cable runs should be limited to a maximum of 100 feet (30 meters). Because of possible performance issues, cable length should be a primary design criterion when selecting rack dimensions and designing overhead cable routing apparatus.

Heights and widths of 42U, 48U, and 52U racks are shown in Figure 12.

Figure 12. Rack height dimensions

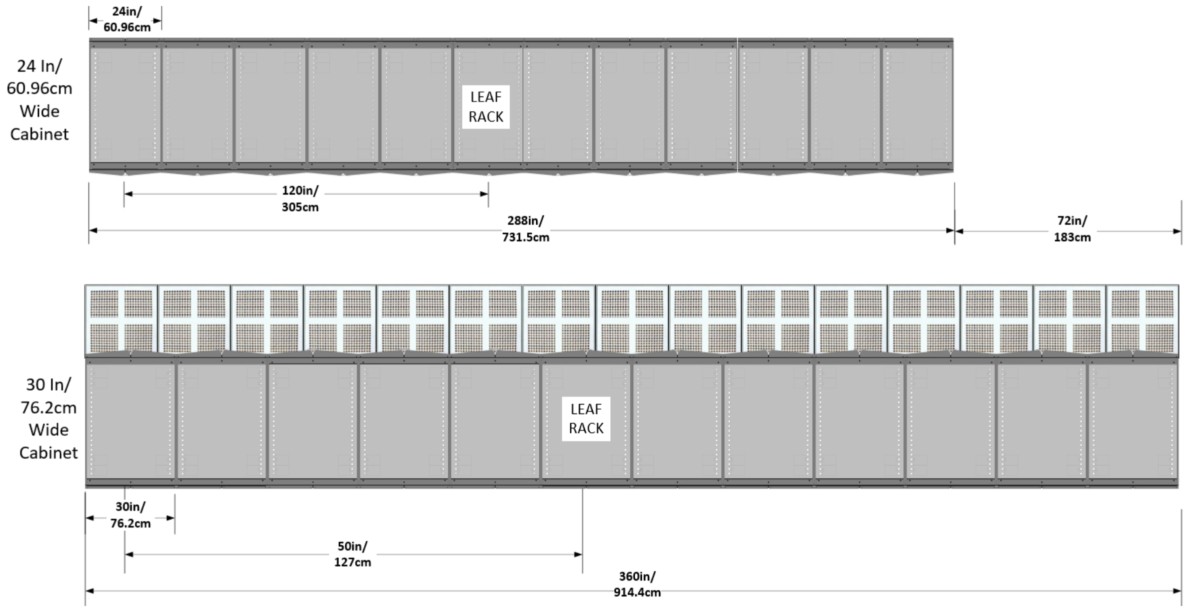


5.4.1 Rack Width Variations

When selecting wider racks, attention should be given to the impact on cable lengths and row width. While wider racks are beneficial for servicing equipment, it may be necessary to reduce the number of racks per row so that maximum row width and optimum cable lengths can be maintained.

Racks widths of 24 in (60.96 cm) and 30 in (76.2 cm) are shown in Figure 13.

Figure 13. Row width based on rack width



5.5 Server Mounting Requirements

DGX SuperPOD equipment is designed to be mounted in traditional IT server racks that conform to the EIA-310-D standard, which among other factors specifies the parameters in Table 5.

Table 5. Server mounting requirements

Parameter	Value
Vertical hole spacing	Vertical hole spacing is defined as a repeating pattern of holes within one RU of 1.75". The hole spacing alternates at 1/2" - 5/8" - 5/8" and repeats.[1] The start and stop of the "U" space is in the middle of the 1/2" spaced holes
Horizontal spacing	The horizontal spacing of the vertical rows of holes is specified at 18 5/16" (18.312) (465.1 mm)
Rack opening	The space in the rack where the equipment is placed is specified as a minimum of 17.72" (450 mm) wide
Front panel width	The total width of the front face of the equipment (with its rack mounting brackets) is 19" (482.6 mm)
1. Many manufacturers use equipment mounting slots instead of holes to allow for variations in this dimension.	

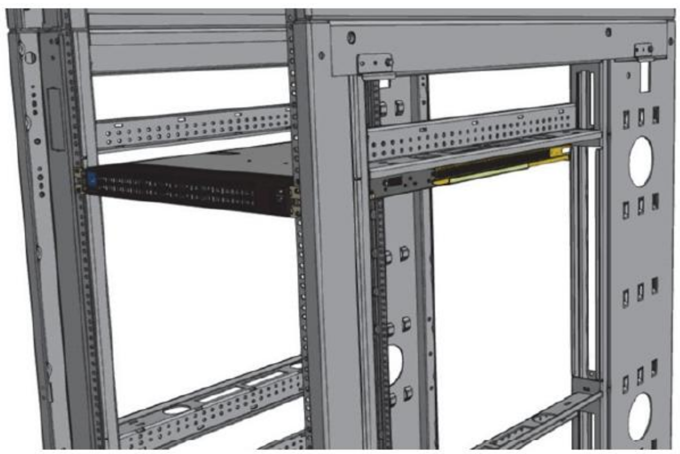
5.6 Device Installation

Another area requiring design attention is how and where the equipment will be mounted in the rack. This first involves setting the proper distance for the rack rails. This will set the horizontal orientation within the rack.

5.6.1 Device Rail Kits

Some DGX SuperPOD devices are installed using rail kits (Figure 14). A rail kit is a bracket specifically designed to support the full weight of the device in the rack, by spanning from the front rack rail to the rear rack rail. A rail kit may be either a stationary shelf type bracket, or a bracket that allows for the device to be pulled out of the rack enclosure on retractable struts. These rail kits have an extension range of 28-32 inches. Wherever provided, rail kits must be used in accordance with manufacturer recommendations.

Figure 14. Device rail kit



5.7 Air Flow Management

Air flow management is an essential part of any data center deployment. Improper air flow management will lead to energy inefficiencies and possible damage to IT equipment.

5.7.1 Equipment Orientation

Some IT devices have the option of selecting what direction the airflow will pass through a device. This choice must be made during BOM selection. In most cases, air flow is determined by the power supply fans and is not field modifiable. Figure 15 shows a side view of a front-facing network switch. Air flows from front (network ports) panel inlets to rear (power connector) panel outlets.

Figure 15. Front facing switch

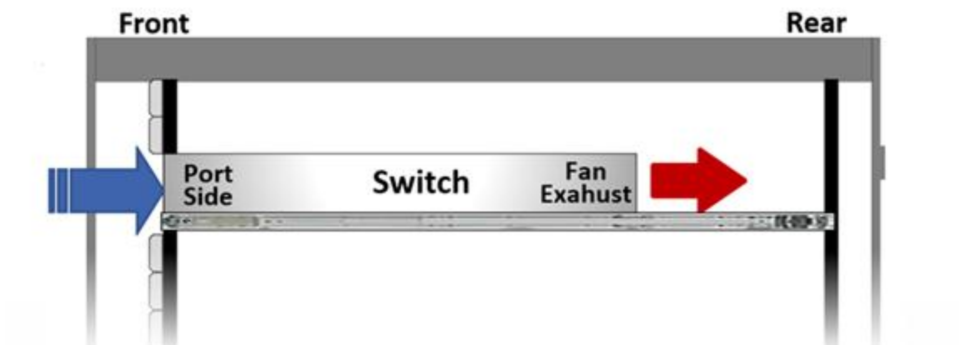
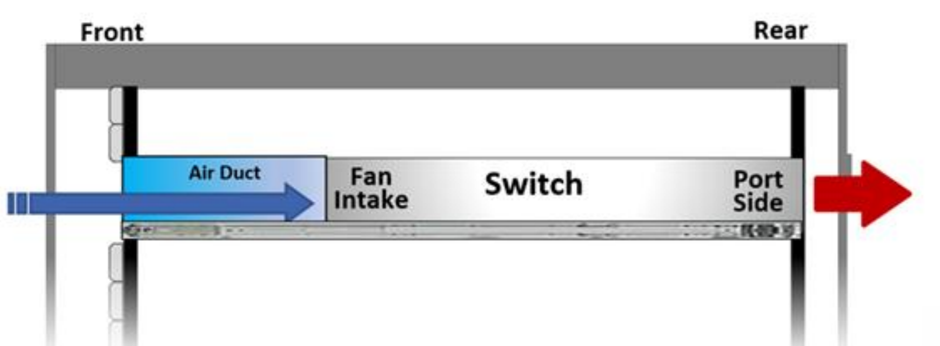


Figure 16 shows a side view of a rear-facing switch. Air flows from rear (power connector) panel inlets to front (network ports) panel outlets.

Figure 16. Rear facing switch



5.8 Static Weight and Point Load

A typical IT rack has an average unloaded weight of 350 pounds (158 Kg). An individual DGX A100 system weighs 271 pounds (123 Kg). In addition to the servers and the racks themselves, peripheral devices such as rPDUs, blanking panels, cable management apparatus, environmental sensors, and cabling add additional weight. Ensure that all flooring structures (including subfloor structures and slabs wherever applicable) are engineered to support the combined weight of the equipment racks they must support. It is also important to ensure that all ingress and egress pathways between the loading dock and the server room floor are engineered to support the combined weight of the equipment plus any conveyances used to move the equipment to the rack location. Refer to **Error! Reference source not found.** to plan weight loads for structured cabling.

Table 6 lists the estimated (rounded) weight of different rack profiles. These are only general estimates. The exact weight of a specific loaded rack would depend on the unloaded weight of the actual rack model that was selected, as well as any extraneous peripheral components and cabling.

Table 6. Weight profiles of DGX A100 system racks

DGX A100 Systems per Rack	Total Rack Weight	Point Load
1	750 lb (318 kg)	175 lb (79 kg)
2	1,000 lb (453 kg)	250 lb (113 kg)
3	1,250 lb (567 kg)	313 lb (142 kg)
4	1,500 lb (680 kg)	375 lb (170 kg)
5	1,750 lb (794 kg)	438 lb (199 kg)

6. InfiniBand Networking

This section provides a high-level overview of the infrastructure needed to support InfiniBand networks.

Table 7 shows the number of InfiniBand cables required based on the number of Compute nodes being deployed.

Table 7. Compute fabric InfiniBand cable counts

Nodes	SUs	QM8790 Switches			Cables		
		Leaf	Spine	Core	Leaf	Spine1	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. The UFM appliance is connected to two different spine switches.

Table 8 shows the number of InfiniBand cables required based on the number of storage nodes being deployed.

Table 8. Storage fabric InfiniBand cable 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.

There is a large concentration of cables in the central Leaf rack as well as in the supporting cable pathways overhead. Preplanning for the cabling infrastructure is a critical part of the design process.

Areas of analysis include:

> Top of rack cable penetration access.

> Cable tray types.

> Cable tray width and depth.

> Cable tray fill ratios.

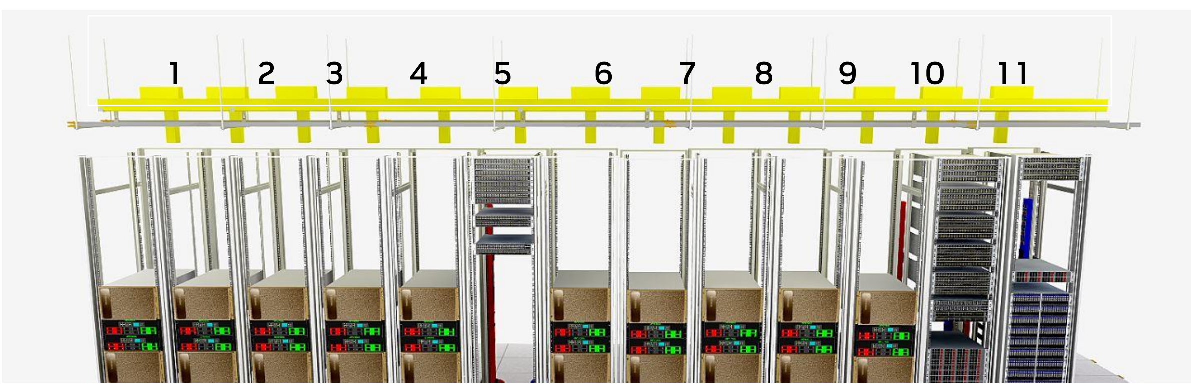
> Structural load/weight planning.

> Cable management devices.

All cable management and pathway apparatus should be selected, installed, and maintained in accordance with applicable industry standards including ANSI/TIA, BICSI, and NEMA. Further, it should be installed in compliance with all locally applicable fire safety and electrical codes, such as National Electric Code (NEC) in the United States and Mexico, The Canadian Electric Code in Canada, The International Electrotechnical Commission (IEC) in Europe, The British Standard (BS) 7671 in the United Kingdom, and NF C 15-100 in France.

Figure 17 shows the inbound SU cabling to the leaf rack.

Figure 17. Inbound cable count aggregating in pathway toward leaf rack



Error! Not a valid bookmark self-reference. gives the cable counts per each rack.

Table 9. Inbound cable count aggregation

Cable Function	Rack Number										
	1	2	3	4	5	6	7	8	9	10	11
InfiniBand compute leaf	16	32	40	48	56	112	56	48	40	32	16
Infiniband storage leaf	4	8	12	16	20	40	20	16	12	8	4
Out-of-band management (RJ45)	2	4	6	8	10	20	10	8	6	4	2
In-band management (fiber)	4	8	12	16	20	40	20	16	12	8	4

There are approximately 400 cables handling inbound and outbound traffic at the leaf rack, 212 inbound and 192 outbound to the management racks.

7. Data Center Thermal Specifications

Data centers are classified based on environmental and thermal capabilities. This section provides guidelines and industry standards that are recommended for a DGX SuperPOD deployment.

7.1 Environmental and Thermal Guidelines

Table 10 describes ASHRAE standard temperature and humidity requirements for the efficient cooling of racks. Data centers must satisfy either the recommended or allowable class A1 requirements to maintain appropriate levels of cooling for DGX A100 system deployments.

Table 10. ASHRAE 2015 thermal guidelines

Class	Dry-Bulb Temperature		Humidity Range Non-Condensing	Maximum Dew Point	
	F	C		F	C
Recommended (Suitable for all 4 classes)					
A1 to A4	**64.4°-80.6°**	**18°-27°**	**15.8°F DP and 59°F DP and 60% RH** **-9°C DP to 15°C DP and 60% RH**		
Allowable					
A1	**59°-89.6°**	**15°-32°**	**20% to 80% RH**	**62.6°**	**17°**
A2	50°-95°	10°-35°	20% to 80% RH	69.8°	21°
A3	41°-104°	5°-40°	10.4°F DP and 8% RH to 85% RH -12°C DP and 8% RH to 85% RH	75.2°	24°
A4	41°-113°	5°-45°	10.4°F DP and 8% RH to 90% RH -12°C DP and 8% RH to 90% RH	75.2°	24°
B	41°-95°	5°-35°	8% RH to 80% RH	82.4°	28°
C	41°-104°	5°-40°	8% RH to 80% RH	82.4°	28°

7.2 Air Purity

Source air contamination (such as smoke, dust, pollution, or other type of contamination) must be mitigated through filtration. Table 11 describes the ISO 14644-1 maximum allowable particle sizes and concentrations for different classes of air cleanliness. NVIDIA recommends Class-8 Standard with maximum particle concentrations not to exceed 3,520,000 @ 0.5 µm/m^3 for no longer than 15 minutes. Depending on the contamination level of the source air, filtration types ranging from ASHRAE standard 52.2 Minimum Efficiency Reporting Value (MERV) 8 to MERV 13 (or the equivalent EN779-2012 counterpart) may be required. Many traditional, closed-loop, air circulation environments require only MERV 8 filtration, while data centers that use outside air may require higher filtration levels. Using higher filtration levels than required to meet desired air cleanliness standards is not cost efficient and can impact nominal airflow.

Table 11. ISO 14644-1 standard for air cleanliness classifications

Class	Particle Size[1]					
	> 0.1 µm	> 0.2 µm	> 0.3 µm	> 0.5 µm	> 1 µm	> 5 µm
1	10	2				
2	100	24	10	4		
3	1,000	237	102	35	8	
4	10,0000	2,370	1,020	352	83	
5	100,000	23,700	10,200	3,520	832	29
6	1,000,000	237,000	102,000	35,200	8,320	293
7				352,000	83,200	2,930
8				**3,520,000**	**832,000**	**29,300**
9				35,200,000	8,320,000	293,000

1. Uncertainties related to the measurement process require that data with no more than three significant figures be used in determining the classification level.

Table 12 provides ASHRAE guidance for MERV filter types.

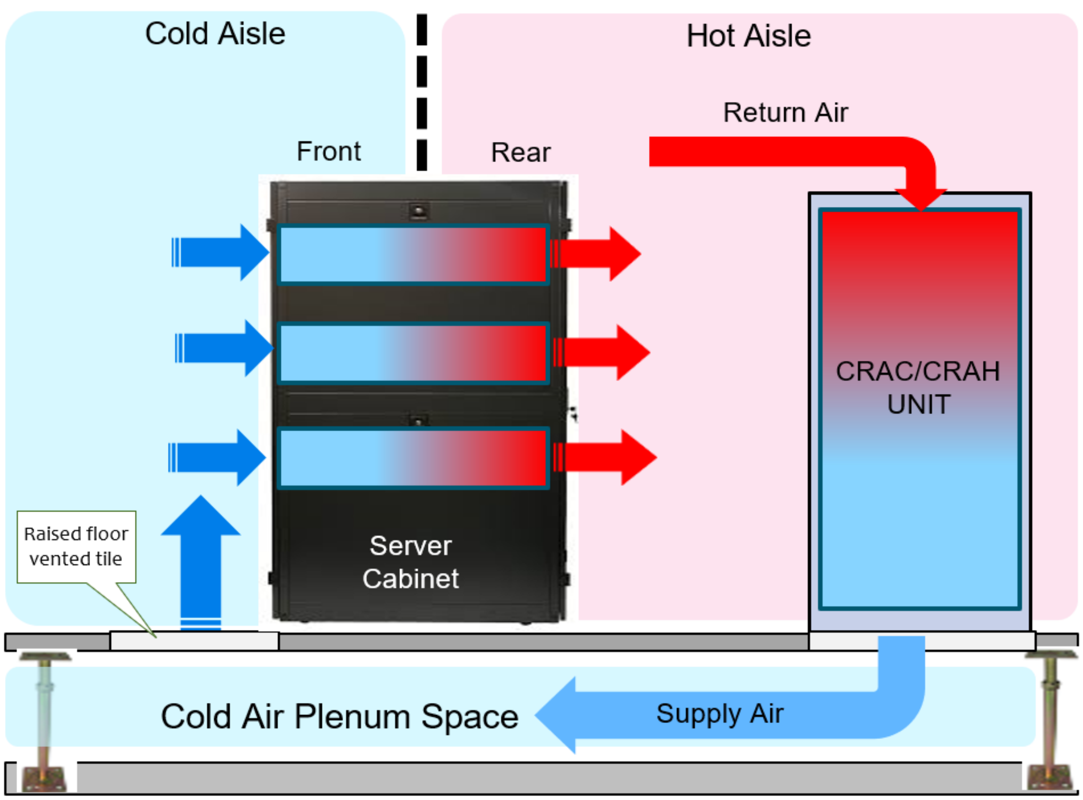
Table 12. Comparison of filter types

ASHRAE Standard 52.2-2007 MERV	EN 779-2012
MERV 1, 2, 3, 4	G1, G2
MERV 5	G3
MERV 6, 7, 8	G4
MERV 8, 9, 10	M5
MERV 9, 10, 11, 12, 13	M6
MERV 13, 14	F7
MERV 14, 15	F8
MERV 16	F9, E10, E11, E12, H13, H14, U15, U16

8. Data Center Cooling

All data centers operate on the same foundational laws of thermodynamics, regardless of the cooling apparatus in use. As they relate to data center cooling, the relevant thermodynamic principles can be simplified by stating that when objects with different amounts of intrinsic heat interact, the temperature of those objects will reach equilibrium through the transfer of heat from the higher temperature object to the lower temperature object, until both objects reach the same temperature (Figure 18). This is true regardless of the differing states of matter of the various objects (solid, liquid, gas).

Figure 18. Basic cooling cycle

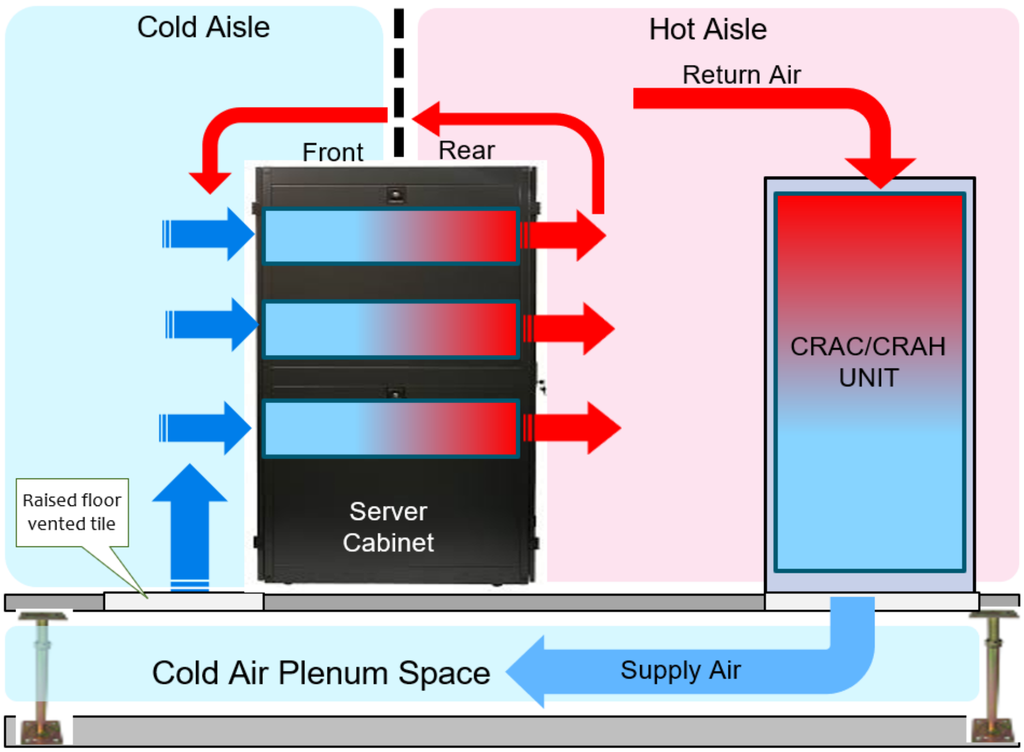


It is important to understand that the heat exchange potential in any cooling system is largely controlled by the temperature differential between the cooling media (air, chilled water, glycol, etc.) and the object to be cooled (the internal components of a heat source, such as a server or switch). This temperature difference is known as Delta T.

For air cooled environments, several factors affect Delta T in addition to basic temperature. These include relative humidity, surface contamination, and perhaps most importantly, the temperature rise of the cooled supply air in transit from the cooling apparatus to the server, as it absorbs heat from the surfaces it contacts and the ambient temperature of the common air in the space. This heat rise is a potential source of inefficiency in the cooling cycle.

The most common source of this heat rise is the mixing of cooled supply air with the heated exhaust air from the server racks. Figure 19 depicts hot air exhaust exiting the server rack and recirculating to the front of the rack.

Figure 19. Cooling cycle with exhaust air recirculation



To illustrate the effect of this air mixing, consider a scenario where all the exhaust air is mixed with all the supply air. If incoming 60° F supply air is mixed with 90°F exhaust air on its journey to the server, the resulting supply air at the server will be 75°F. Because there is a direct Delta T correlation between supply air temperature and exhaust air temperature, the exhaust air temperature will now increase by that same differential, to 105°F, which will then further heat the inbound supply air, which will further increase the exhaust air temperature, potentially resulting in a destructive cycle known as a thermal runaway condition. In more common practice, only a portion of the exhaust air mixes with the supply air, resulting in symptoms such as hot spots at various locations in the data center room. In an air-cooled data center, it is vital that steps are taken to prevent the mixing of cooled supply air and heated exhaust air.

8.1 Air Cooling Methods

This section covers various cooling methods used in the data center. As part of DGX SuperPOD deployment, air cooling capabilities help in optimizing rack planning for the final deployment.

8.1.1 Traditional Plenum Environment

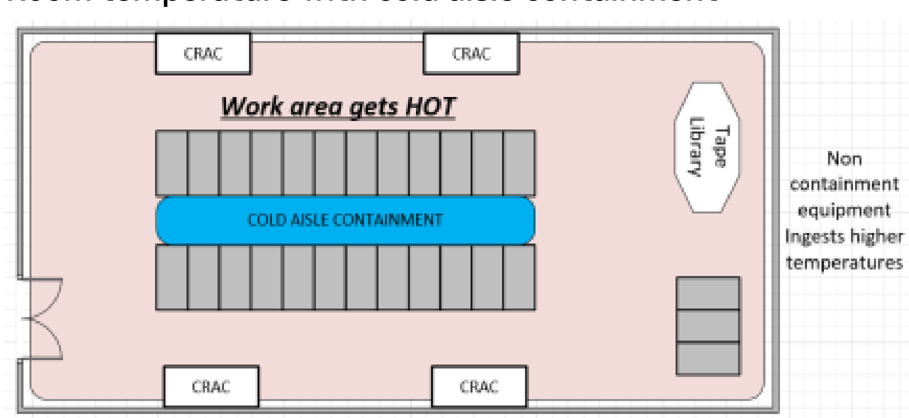
In a traditional plenum environment, a raised flooring system acts as a pressurized plenum to deliver cooled supply air through vented floor tiles at the server rack locations. Air handler units are placed around the perimeter of the room (and sometimes in the center areas) and are configured for downflow operation, blowing air into the subfloor space. Server racks are arranged in rows and oriented in a hot aisle and cold aisle layout. The fronts of two adjacent rows face one another, with several vented tiles between them supplying cooled air to both rows. Conversely, the backs of two adjacent rows also face one another, and heated exhaust air from both typically flows into vented tiles in a plenum drop ceiling where it flows back to the air intakes on the air handlers.

Traditional plenum environments are especially susceptible to exhaust air recirculation as described in the preceding text and are also susceptible to *hot spots* (areas that lack sufficient airflow), due to pressure differentials in the subfloor space caused by air movement patterns and airflow obstructions.

8.1.2 Cold Aisle Containment

Cold aisle containment is a common method of ensuring maximum efficiency of supply air delivery (Figure 20). By partitioning off the cold aisle, this method ensures that all cooled supply air is available for IT equipment and that no air recirculation takes place. Consistent positive pressure must be maintained within the aisle and the aperture and flow rate of the vented tiles must supply more than 100% of the required airflow demand. This is because the containment prevents any air from being drawn in from other areas of the room to make up for any deficit within the aisle itself.

Figure 20. Room temperature with cold aisle containment



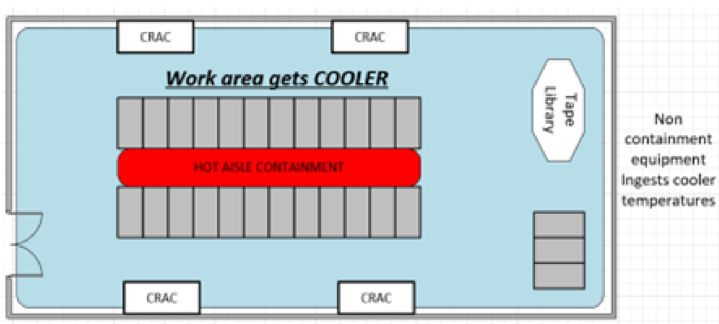
In a cold aisle containment configuration, the common areas of the data center are naturally part of the hot aisles, and thus the temperature in those common areas may be less comfortable for workers. This is also a consideration for data centers that have equipment that is not housed in traditional server racks, such as tape libraries or other mainframe peripherals.

> **Note:** Legacy mainframe data centers that have been updated with cold aisle containment commonly have mainframe components on the data center floor. Many of those devices are up-fed directly from the underfloor plenum without relying on vented tiles. PDUs, busways, and other power distribution components also can be outside the containment. As the common area ambient temp rises due to cold aisle containment use, the increased ambient room temperature can negatively affect the performance of those devices.

8.1.3 Hot Aisle Containment

Like cold aisle containment, hot aisle containment prevents air recirculation by segregating the hot exhaust air from the cooled supply air (Figure 21).

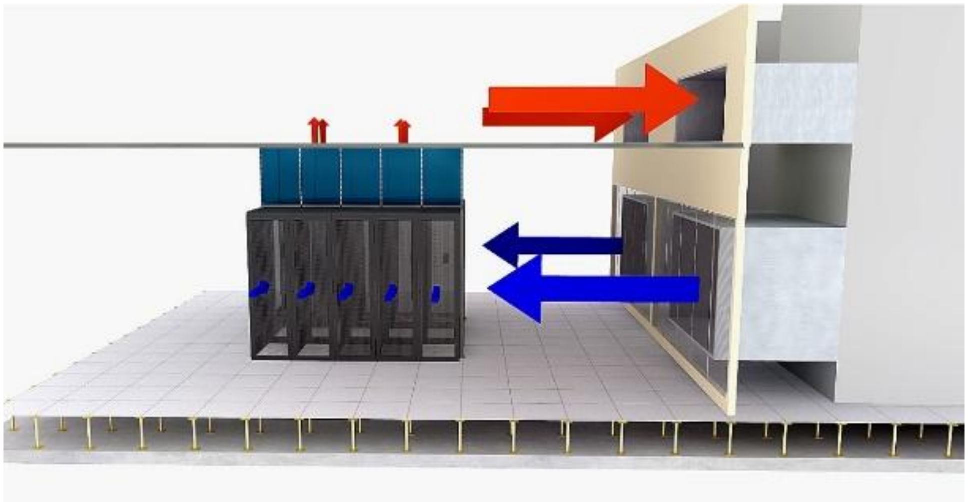
Figure 21. Room temperature with hot aisle containment.



With hot aisle containment, hot spots and other airflow-related challenges are less likely to be impactful, because the exhaust air is directed out of the environment, and the entire room is a potential source of cooled supply air. In a hot aisle containment environment, temperatures in the hot aisle can be very high. Workers should not remain in the hot aisle for extended periods of time. Power distribution components in the hot aisle may require derating.

Figure 22 provides an example of cold air entering the data center and hot air exiting from the aisle containment above the racks by a separated plenum.

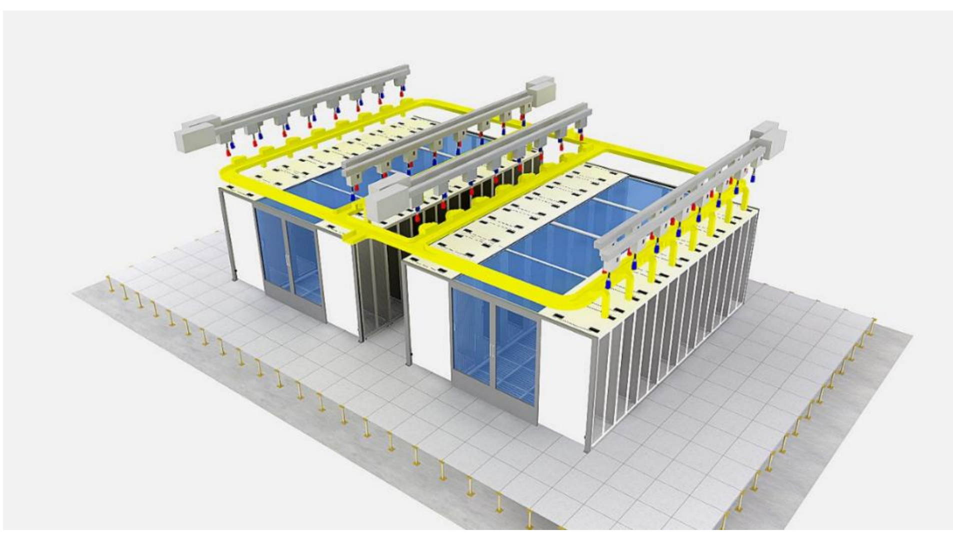
Figure 22. Hot aisle containment with *air wall* style supply air delivery



8.2 Containment System Cable Paths

Depending on the data center containment design and local fire/safety code, *shortest path* cable pathways might not be possible to install directly over the containment system. The yellow fiber optic trays in Figure 23 show an example of this.

Figure 23. Data center containment with overhead infrastructure and pathways



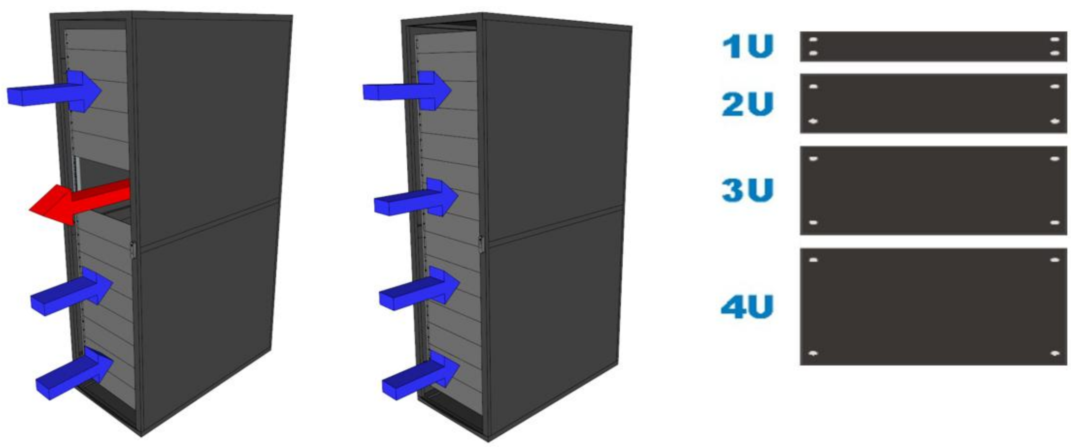
8.3 Air Containment

Recirculation and bypass airflow (cooled supply air passing through openings and gaps in the server rack instead of passing through the servers themselves) within server racks can also have an impact on the overall cooling efficiency of the rack and the data center. Numerous practices and products exist to remediate recirculation and bypass airflow.

8.3.1 Rack Air Containment

Blanking panels fill unused RU spaces in the rack to prevent recirculation and bypass airflow (Figure 24). They are required to assure proper airflow at the servers.

Figure 24. Blanking panels



In addition to filling open RUs with blanking panels, all other openings and gaps within the server rack (Figure 25) should be filled with appropriated blank plates, gaskets, brush grommets, or other method to maintain aisle containment integrity.

Figure 25. Locations of potential airflow bypass

8.4 Floor Air Containment

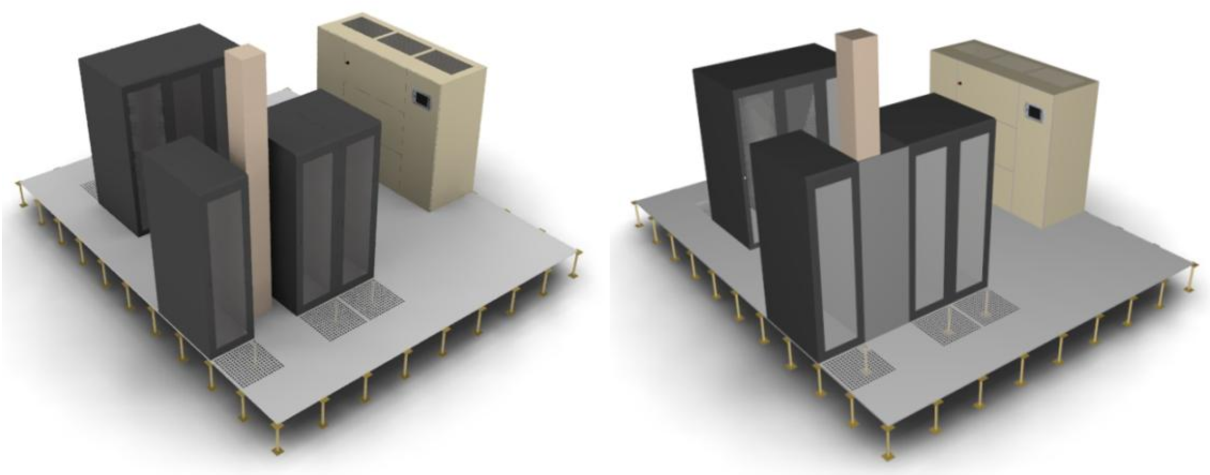
Penetrations in the floor tiles, including cutouts around physical obstructions such as columns and walls, should also be sealed using brush grommets, caps, blank plates, or gaskets (Figure 26). Using the proper floor grommet and air block is also an essential good installation practice.

Figure 26. Various floor grommets

8.5 Open Space Containment

There may be an instance where a rack row in the data center is disrupted by a support pillar. In all such situations, it is recommended that the gap between racks be partitioned off to maintain containment integrity (Figure 27).

Figure 27. Correcting a gap caused by a physical obstruction



9. 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 is backed by 24x7 enterprise-grade support.

A. Scaling Deployments

This section covers how to allocate racks for future expansions.

A.1 Adjacent InfiniBand Deployments

The InfiniBand spine rack is used to connect additional SUs. Consideration should be given to allocate unused rack space, power, and cooling additional NVIDIA InfiniBand spine switches will be added in the future.

A.2 Storage Growth

Consider the following when reserving capacity for data storage:

> Empty rack space.
> PDU receptacle.
> Power capacity.
> Cooling capacity.
> InfiniBand switch port capacity.
> InfiniBand cable lengths.

A.3 Spine Growth

Consider the following when reserving capacity for spine growth:

> Empty rack space.
> PDU receptacles.
> Power capacity.
> Cooling capacity.
> InfiniBand switch port capacity.
> InfiniBand cable lengths.
> Management port capacity.

B. Data Center Optimizations

This section covers thermal optimizations that help running an efficient data center. Computation Fluid Dynamic (CFD) Modeling and Supplemental Cooling considerations are presented in this section.

B.1 CFD Modeling

CFD modeling is a recommended practice in the design stage of the data center facility, and also as a tool to model the predicted impact of any major deployments that present significant cooling demand. Since precision is necessary to obtain tolerably accurate models, it may be necessary to work with outside consultants or the colocation provider for expertise in CFD modeling if in-house tools and expertise are not available. These models can reveal potential cooling challenges and can also be used to help find out the effectiveness of any proposed solutions before implementation.

Figure 28 provides an example of a CFD report where a different number of hot-aisle widths measured in floor tiles are being analyzed in parallel with cooling failures and different Supply Air Temperatures.

Figure 28. Sample CFD report

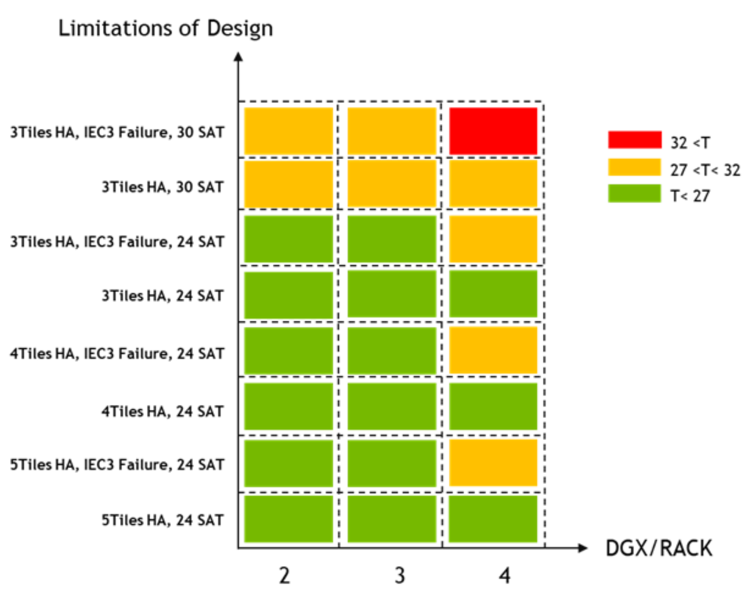


Table 13 details the findings shown in Figure 28.

Table 13. Example CFD analysis report

Case Design	Power (MW)	SAT (˚C)	Perforated Tiles per HA	IT Needed Airflow (m3/s)	Cooling Airflow (m3/s)	Rack worst mean Inlet T (˚C)	ASHRAE Temp Compliance
1	0.876	24	5	67.8	172	24.3	Works
2	0.876	24	5	67.8	158	24.3	Works
3	0.876	24	4	67.8	172	24.4	Works
4	0.876	24	4	67.8	158	24.5	Works
5	0.876	24	3	67.8	172	24.3	Works
6	0.876	24	3	67.8	158	24.4	Works
7	0.876	30	3	67.8	172	30.3	Marginal
8	0.876	30	3	67.8	158	30.4	Marginal
9	1.234	24	5	100	172	24.3	Works
10	1.234	24	5	100	158	24.4	Works
11	1.234	24	4	100	172	24.6	Works
12	1.234	24	4	100	158	24.7	Works
13	1.234	24	3	100	172	24.5	Works
14	1.234	24	3	100	158	24.6	Works
15	1.234	30	3	100	172	30.5	Marginal
16	1.234	30	3	100	158	30.6	Marginal
17	1.752	24	5	132	172	25.2	Works
18	1.752	24	5	132	158	29.5	Marginal
19	1.752	24	4	132	172	24.8	Works
20	1.752	24	4	132	158	31.5	Marginal
21	1.752	24	3	132	172	24.7	Works
22	1.752	24	3	132	158	29.2	Marginal
23	1.752	30	3	132	172	30.7	Marginal
24	1.752	30	3	132	158	35.2	Failed

A temperature map provides a visual representation of temperature by using colors. The following images are samples from an analysis of an SU.

Figure 29 illustrates the mean inlet air temperature based on numerous data center data parameters.

Figure 29. Mean inlet temperature

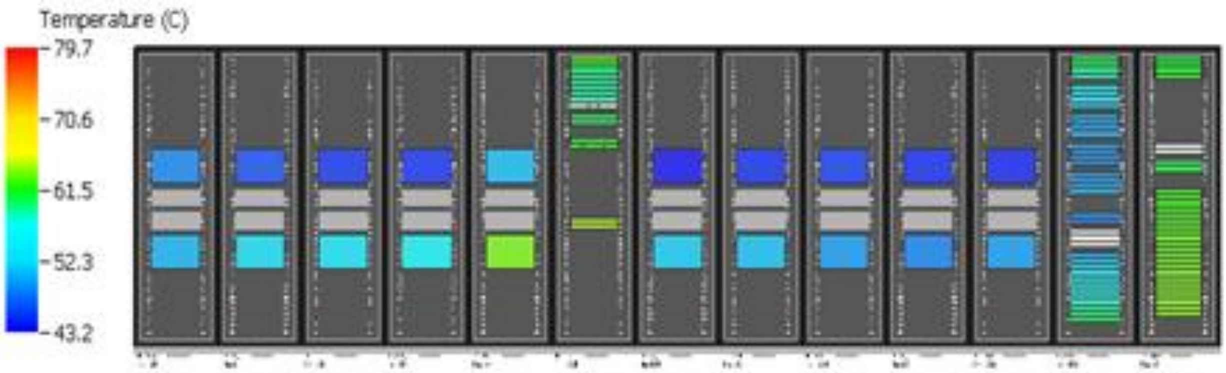
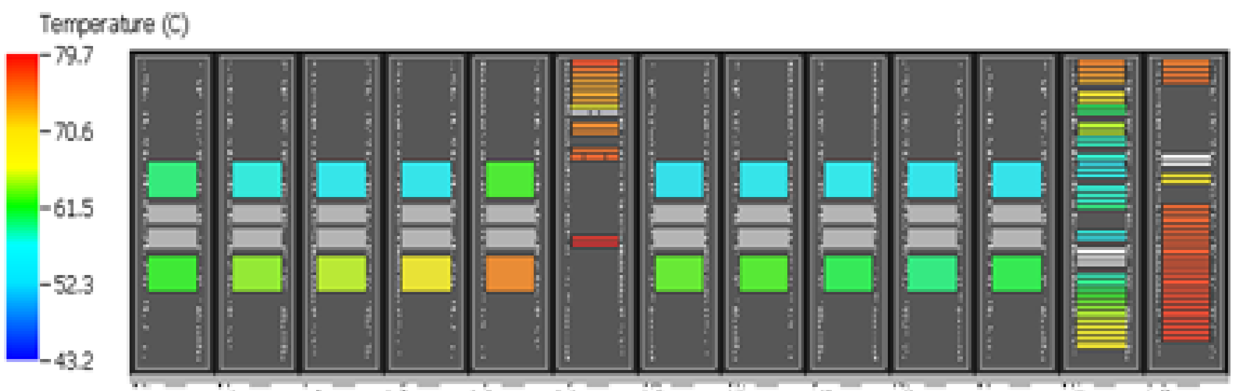


Figure 30 depicts a rear view of the rack row and the estimated mean exhaust temperatures for each device.

Figure 30. Mean exhaust temperature



B.2 Supplemental Cooling

Supplemental cooling devices, such as in-row coolers, overhead coolers, active rear door heat exchangers, etc. may be in use in the data center environment. If any such system will be present in a row that houses DGX A100 systems, it is advisable to consult with NVIDIA to assure the system design for that data center is adequate and does not adversely impact DGX performance.

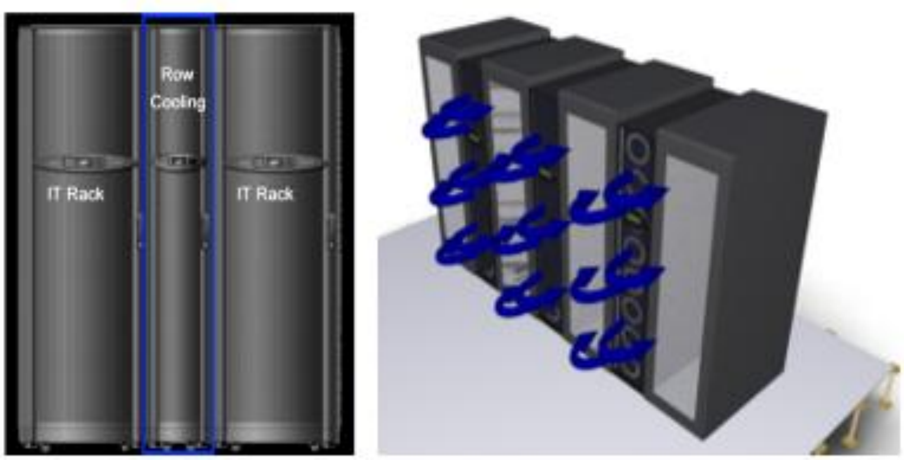
B.2.1 In-Row Cooling

Depending on the situation, in-row cooling may be required to supplement the data center facility cooling system (Figure 31).

In-row cooling system:

> Mounted between two racks within a row with one or more can be installed per row.
> Rack-level cooling that removes and neutralizes hot air from racks.
> Requires a chilled water or refrigerant source and a connection to a remote chiller system.

Figure 31. In-row cooling



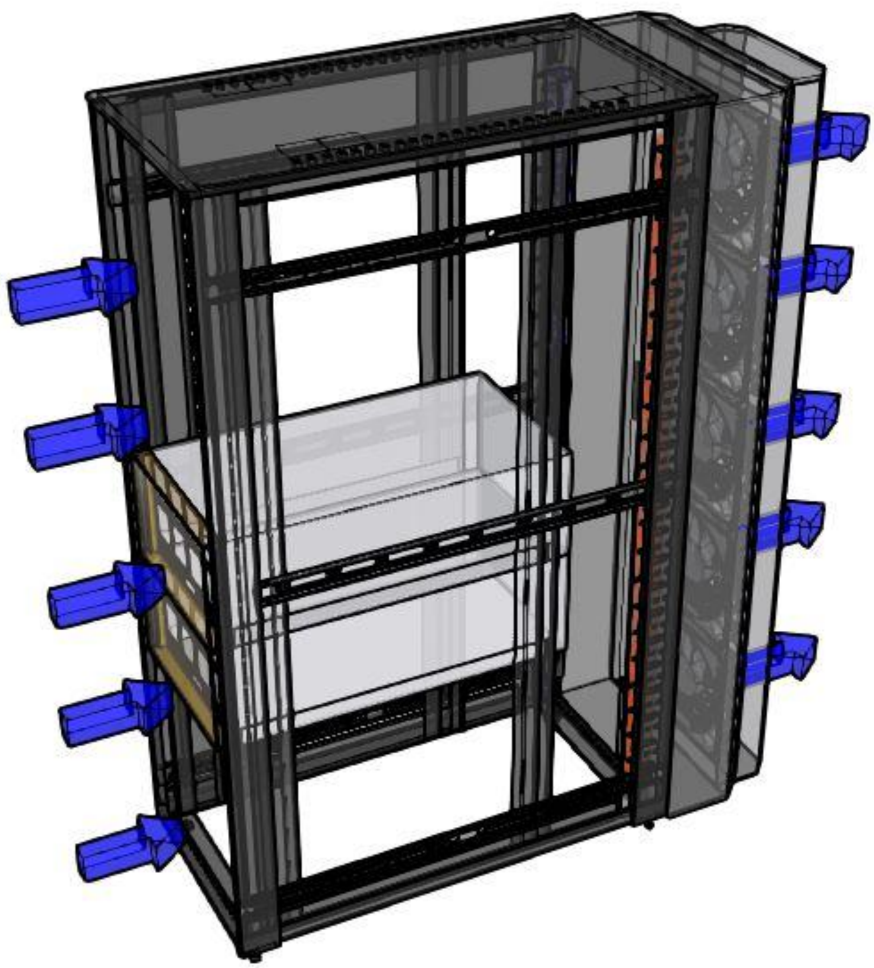
Active Rear Door Heat Exchanger

A rear door heat exchanger is an optional way to proactively reduce the air temperature as it exits the rear of a server rack (Figure 32).

These features include:

- Mounted on rear door of rack.
- Condition hot air and return it to the room at a better suitable temperature for struggling CRAC units.
- Installed on racks and does not require floor space.
- Most solutions require a chilled water source and a connection to a remote chiller system.

Figure 32. Rear door heat exchanger



C. Safety Considerations

This section covers safety considerations regarding noise, fire protection, and handling the DGX A100 systems.

> ! **CAUTION:** Due to their weight and size, high requirements, high heat rejection, and the loudness of their fans, operator SAFETY is of major importance. Personal protective equipment (PPE) must be used, and safety procedures must always be observed when working on or near DGX A100 systems.

C.1 Noise

DGX A100 systems can generate noise levels between 70 dBA and 105 dBA. It is strongly recommended that persons working in the data center wear hearing protection suitable to these noise levels. This may require two or more hearing protection devices, such as ear plugs (with an NRR 29 Rating) combined with earmuffs (with an NRR 27 rating), be worn simultaneously.

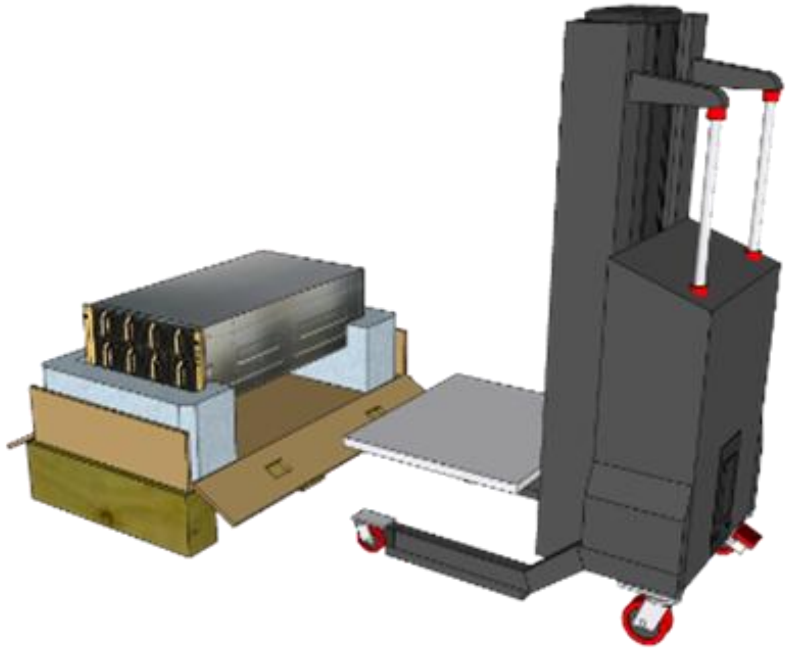
C.2 Fire Protection

Fire detection, suppression, and prevention systems are recommended in the data center to protect data center staff and infrastructure—including DGX A100 system deployments. These systems are the responsibility of the data center owner/operator and should be designed, installed, and maintained in accordance with the laws and regulations of local authorities having jurisdiction (AHJs) and insurance carriers. No recommendation in this document should be construed to override the requirements of any AHJ.

C.3 Server Lifts

Due to the heavy weight of the DGX A100 system, a server lift (Figure 33) must be used to install and remove the devices from the racks, and to transport them to the rack location. Use the lift in compliance with all safety precautions and protocols as specified by the manufacturer. Ensure that the server lift is rated for a minimum of 350 pounds of weight. Ensure proper clearance below any overhead obstructions before lifting the device.

Figure 33. Server lift



D. Advanced Electrical Planning

Table 14 provides additional detail to power requirements. It also reveals the amount of *stranded* power that is wasted by overprovisioning.

Table 14. Power requirements and circuit sizing matrix

Circuit Sizing						Power Requirements DGX A100 Systems				
Phase	Volts	Amps	80% Breaker De-Rating	kVA	Circuit kW	1 (6.5)	2 (13)	3 (19.5)	4 (26)	5 (32.5)
1	208	30	Yes	5.0	4.7	-1.8	-8.3	-14.8	-21.3	-27.8
1	208	60	Yes	10.0	9.5	3.0	-3.5	-10.0	-16.5	-23.0
1	230	32	No	7.4	7.0	0.5	-6.0	-12.5	-19.0	-25.5
1	230	63	No	14.5	13.8	7.3	0.8	-5.7	-12.2	-18.7
3	208	60	Yes	17.3	16.4	9.9	3.4	-3.1	-9.6	-16.1
3	415	30	Yes	17.3	16.4	9.9	3.4	-3.1	-9.6	-16.1
3	415	30	No	21.6	20.5	14.0	7.5	1.0	-5.5	-12.0
3	400	32	No	22.2	21.1	14.6	8.1	1.6	-4.9	-11.4
3	415	60	Yes	34.5	32.8	26.3	19.8	13.3	6.8	0.3

E. Industry Standards Reference

Industry standards applicable to this document are:

- Data Center:
  - BICSI 002, Data Center Design and Implementation Best Practices
  - ANSI/TIA-569-C (2012) Telecommunications Pathways and Spaces
  - Telecommunication Distribution Methods Manual BICSI, Eleventh Edition 2006
- Administration:
  - ANSI/TIA-606-B (2012) Administration Standard for Commercial Telecommunications Infrastructure
- Cable Standards:
  - ANSI/TIA-568.0-D, Generic Telecommunications Cabling for Customer Premises, 2015
  - ANSI/TIA-568.1-D, Commercial Building Telecommunications Infrastructure Standard, 2015
  - ANSI/TIA-568-C.2, Balanced Twisted-Pair Telecommunication Cabling and Components Standard, published 2009
  - ANSI/TIA-568-C.3, Optical Fiber Cabling Components Standard, published 2008, plus errata issued in October 2008.
  - TIA-569-B (2004; Amd 1 2009) Commercial Building Standard for Telecommunications Pathways and Spaces
- Cable Trays:
  - NEMA Standards Publication VE 2-2013 Cable Tray Installation Guidelines
  - NEC ARTICLE 392 - CABLE TRAY
- Cable Labeling:
  - ANSI/TIA-606-B Cable Labeling Standard

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

http://www.nvidia.com