Lenovo NVIDIA GB300 NVL72 Rack Scale AI

Product Guide

The Lenovo NVIDIA GB300 NVL72 is a rack scale solution built on the NVIDIA Blackwell architecture, which delivers groundbreaking advancements in accelerating computing, leading the new era of AI with optimized compute and increased memory.

The GB300 NVL72 is the NVIDIA flagship rack-scale AI offering. It is an end-to-end NVIDIA reference design adopted by the market.

The NVIDIA GB300 NVL72 is a platform designed for AI reasoning performance and efficiency, a scalable building block approach design, featuring 72 NVIDIA Blackwell Ultra GPUs and 36 Arm-based NVIDIA Grace CPUs in a hybrid cooling standard rack.

The GB300 NVL72 uses a hybrid cooling architecture: CPUs, GPUs, and NVSwitch components are liquid-cooled, while OSFP modules, storage drives, and PDB components are air-cooled.

The components of the AI rack solution are designed to be compliant with the NVIDIA MGX rack. The liquid cooling solution consists of a CDU, rear manifold, quick disconnects, and cold plates for the CPUs, GPUs, ConnectX-8 network adapters, and all NVSwitch components.

360° View | Full 3D Tour

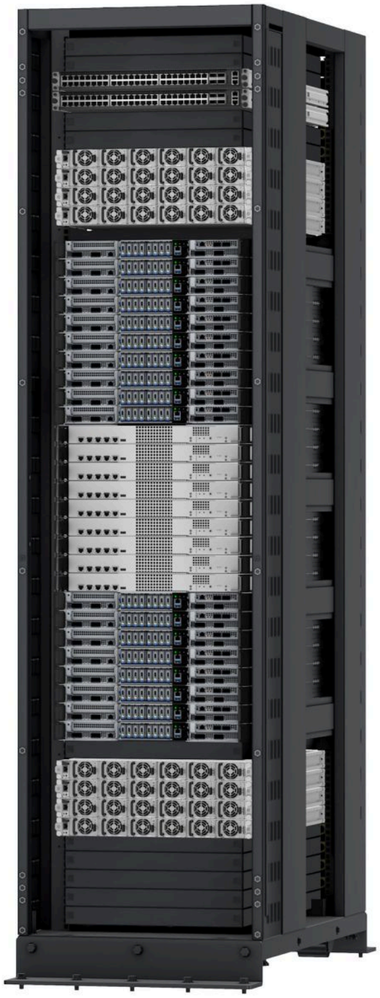

Figure 1. The Lenovo NVIDIA GB300 NVL72

Did you know?

The GB300 NVL72's integrated NVLink switch enables every GPU in the rack to communicate as part of one unified fabric. Each rack delivers 72 NVIDIA Blackwell GPUs interconnected through the 5th generation NVIDIA NVLink™ technology, forming a single high performance compute domain ideal for large-scale multimodal reasoning, high-performance inference and LLM training workloads.

Key features

The Lenovo NVIDIA GB300 NVL72 is a rack scale solution for multiple interconnected NVIDIA GPUs. The solution consists of 18 GPU Compute trays, 9 Switch trays and either 6 or 8 Power Shelves all connected via a rear cable cartridge backplane.

Turnkey AI Rack Design

Engineered for large-scale Artificial Intelligence (AI) and High Performance Computing (HPC), Lenovo NVIDIA GB300 NVL72 excels in intensive simulations and complex modeling. It is designed to handle modeling, training, simulation, rendering, financial tech, scientific research, technical computing, grid deployments, and analytics workloads in various fields such as research, life sciences, energy, engineering, and financial simulation.

The NVIDIA GB300 NVL72 by Lenovo features a mix of air and liquid cooling to meet its demanding power requirements. The NVIDIA GB300 NVL72 integrates seamlessly into a standard 19" rack cabinet. Each rack consumes 135 kW TDP; up to 155 kW peak depending on workload and EDP behavior. At the rack level, the split between air and liquid heat capture is about 10% to air and 90% to liquid.

This design ensures easy serviceability and extreme performance density, making the NVIDIA GB300 NVL72 the go-to choice for compute clusters of all sizes - from departmental/workgroup levels to the world’s most powerful supercomputers – from Exascale to Everyscale®.

Each rack delivers 72 NVIDIA Blackwell GPUs interconnected through the 5th generation NVIDIA NVLink™ technology, forming a single high performance compute domain ideal for large-scale multimodal reasoning, high-performance inference and LLM training workloads.

Direct water-cooled solutions are factory-integrated and are re-tested at the rack-level to ensure that a rack can be directly deployed at the customer site. This careful and consistent quality testing has been developed as a result of over a decade of experience designing and deploying DWC solutions to the very highest standards.

Scalability and performance

The Lenovo NVIDIA GB300 NVL72 is based on a GB300 rack infrastructure which offers the following features to boost performance and improve scalability:

- The NVIDIA GB300 NVL72 Rack Level GPU System
  - One rack with 1U Compute Trays, 72 GPUs with NVLINK with non-scalable (NVLink domain is contained within a single rack; no inter-rack NVLink scaling) NVSwitches.
  - Cooling solution is a hybrid of liquid and air-cooled components
  - Power shelves either 6 or 8. Quantity dependent on rack configuration and power redundancy requirement
- The 1U Compute Tray leverages the NVIDIA GB300 NVL72 architecture:
  - Two NVIDIA Grace processors
  - Four NVIDIA Blackwell SXM7 GPUs
  - Highspeed NVLink interconnects between GPUs
- Each compute tray has up to 2x PCIe Gen 5 x16 slots for high-speed networking, depending on the configuration.
- Each compute tray also supports up to 8x E1.S 3.84T drive bays, depending on the configuration, for internal storage and 1x 1.92T M.2 NVMe SSD for OS boot
- All drives are E1.S NVMe drives with PCIe Gen 5 x4 host interface, to maximize I/O performance in terms of throughput, bandwidth, and latency.
- The compute tray includes one 1Gb Ethernet onboard port for debug networking. One BMC card for system management (Out Of Band port supports for BMC management communication). High speed host networking can be added through the included PCIe slots.

- The compute tray offers PCI Express 5.0 (PCIe Gen 5) I/O expansion capabilities. A PCIe Gen 5 x16 slot provides 128 GB/s bandwidth, enough to support 800GbE via OSFP network connections (1x 1pt NVIDIA BlueField-3 adapter). 400/200GbE supported via breakout DAC/AOC cables. Two Gen 5 x16 connections (32 lanes total) provide the bandwidth needed for an 800 GbE connection using the ConnectX-8 8180 network adapter (2x 2port ConnectX-8 adapters).
- The 1U NVLink™ Switch Tray leverages the NVIDIA GB300 NVL72 architecture:
  - One GB300 NVL72 NS Switch Tray Base (non-scalable)
  - One GB200 NVL72 Reduntant Power Shelf Solution
- The NVIDIA SN2201 Management Switch leverages the NVIDIA GB300 NVL72 architecture:
  - One NVIDIA SN2201 48 port DC switch
  - One GB200 NVL72 Redundant Power Shelf Solution

Energy efficiency

The direct water cooled solution offers the following energy efficiency features to save energy, reduce operational costs, increase energy availability, and contribute to a green environment:

- Water cooling eliminates power that is drawn by cooling fans in the enclosure and dramatically reduces the required air movement in the server room, which also saves power in the data center due to the reduced need for air conditioning.
- Water chillers may not be required with a direct water cooled solution. Chillers are a major expense for most geographies and can be reduced or even eliminated because the water temperature can now be 10°C to 35°C in an air-cooled environment.
- Heat energy absorbed may be reused for heating buildings in the winter, or generating cold through Adsorption Chillers, for further operating expense savings.
- The processors and other microelectronics are run at lower temperatures because they are water cooled, which uses less power, and allows for higher performance.
- The power solution consists of 50V power shelves delivering power to the bus bar and complying with the standard MGX v1.1 rack specifications.

Manageability and security

The following powerful systems management features simplify local and remote management of the NVIDIA GB300 NVL72 system:

- Support for industry standard management protocols, IPMI 2.0, Redfish REST API, serial console via IPMI
- The NVIDIA GB300 NVL72 is enabled with Lenovo HPC & AI Software Stack, so, you can support multiple users and scale within a single cluster environment.
- Lenovo HPC & AI Software Stack provides our HPC customers you with a fully tested and supported open-source software stack to enable your administrators and users with for the most effective and environmentally sustainable consumption of Lenovo supercomputing capabilities.
- Our Confluent management system provides an interface designed to abstract the users from the complexity of HPC cluster orchestration and AI workloads management, making open-source HPC software consumable for every customer.
- Integrated Trusted Platform Module (TPM) 2.0 support enables advanced cryptographic functionality, such as digital signatures and remote attestation.
- Supports Secure Boot to ensure only a digitally signed operating system can be used.
- Industry-standard Advanced Encryption Standard (AES) NI support for faster, stronger encryption.
- For management: BMC Server Management chip (ASPEED AST2600) and HMC.
- Power shelves are located at top and bottom of the rack (33kW each).

- The Baseboard Management Controller (BMC) in the compute trays is used in conjunction with the Host Module Controller (HMC) to provide Out-of-band management.

Availability and serviceability

The GB300 Compute tray and GB300 NVL Switch Tray provide the following features to simplify serviceability and increase system uptime:

- Designed to run 24 hours a day, 7 days a week
- With hybrid cooling, water cooling and air cooled systems, some fans are required.
- Power Shelves to minimize downtime as current is shared between multiple shelves.
- Toolless cover removal on the nodes provides easy access to upgrades and serviceable parts such as adapters and drives.
- The system desired to support non-disruptive firmware upgrade for any programmable.
- FRUs (Field Replaceable Unit) that are supported by the design. FRUs are parts that can be field replaceable.
- Any cabling connectors desired to include a latch and a fool proof design to prevent users from plugging it in backwards.
- The external debug console port desired to have terminal access to each of the major components (CPU/OS, BMC) for debug and diagnostic purposes.
- There is a three-year customer replaceable unit and onsite limited warranty, with next business day 9x5 coverage. Optional warranty upgrades and extensions are available.

Components and connectors

The table below shows the main system components of the rack scale solution. This is to depict an overall hardware configurability, and capability, for the complex set of feature options that the system needs to be able to support in hardware. Not all configurable combinations, or feature options, are shown here; such as DIMM memory options, TPM, or storage options as examples. Also not shown are the TOR switches which are dependent on customer network topology configuration.

Table 1. Standard specifications NVIDIA GB300 NVL72 rack

Components	**Quantity**	**Description**
Rack	1	48U MGX Rack
GPU Compute Trays	18	NVL72: 1U trays with 2x GB300 HPM
NVL Switch Trays	9	NVIDIA product NVL72: 1U tray with no OSFP (non-scalable)
50V Power Shelf	6-8	Quantity dependent on rack configuration and customer power redundancy configuration
TOR Switches	2	MSN2201-CSMRC NVIDIA Spectrum based 1GbE/100GbE 1U Open Ethernet switch
Cable Cartridge Backplane*	4	Provided by NVIDIA, including mechanicals. May not be populated in rack depending on customer configuration
CDU	1	Various CDU Options
Liquid Cool Manifolds	2	Rear Left and Rear Right (Inlet and Outlet), top and bottom feed manifolds are available
Bus Bar	1	Supports single 1400A bus bar in the middle rear

* Is a collection of many high-speed NVLink channels providing the interconnection between every GPU to every NVSwitch Tray. The copper wires are protected by a chassis to ensure no damage to the connectors or wires within the cartridge.

The GB300 Compute trays, GB300 NVL Switch Trays, Power Shelves and Top of Rack switches are all integrated and rack mounted. Below is a high-level rack overview of the NVL72 rack configuration.

Note: Configuration choices are:

- Customer can choose 6 or 8 power shelves.
- Customer can choose either Top-Feed Manifold (CA40) or Bottom-Feed Manifold (C5RN)

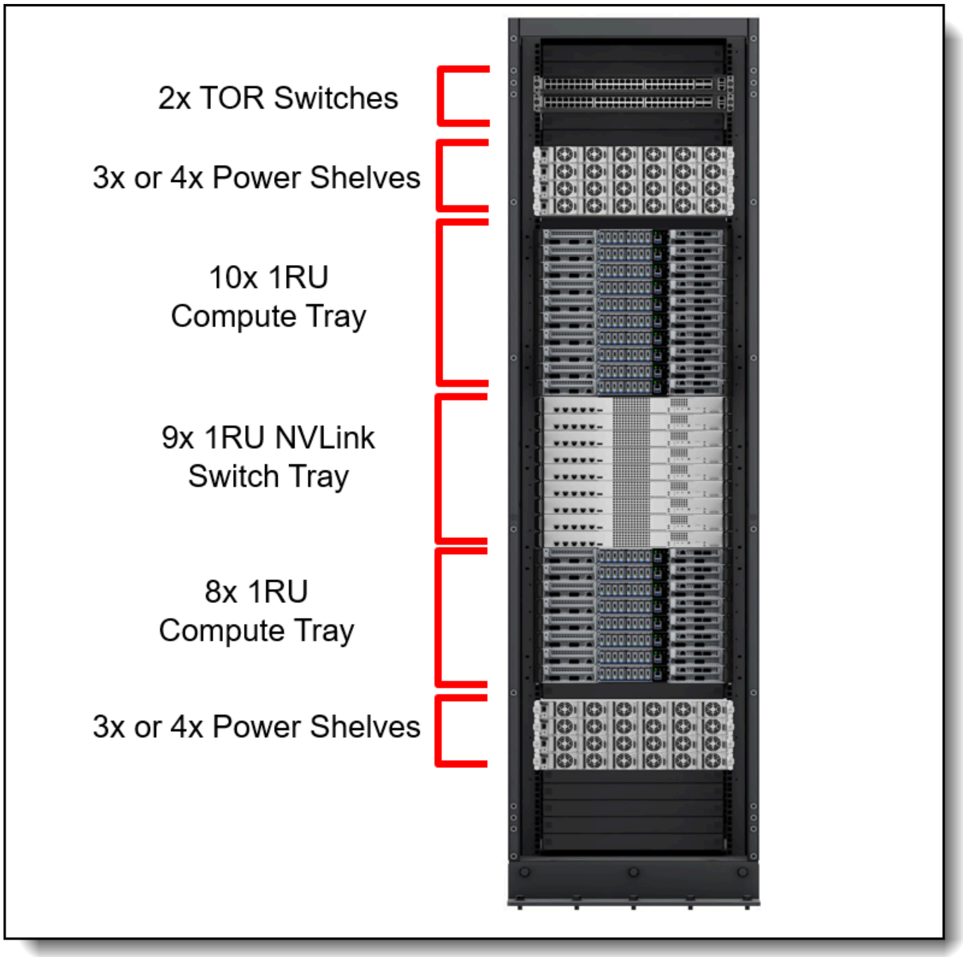

Figure 2. Front view of the Lenovo NVIDIA GB300 NVL72 rack

The front of the compute server is shown in the following figure.

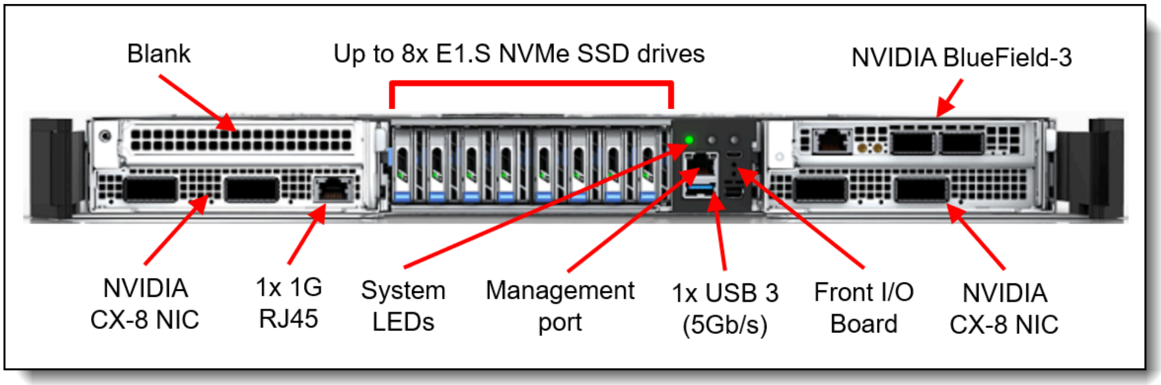

Figure 3. Front view of the GB300 Compute Tray

The following figure shows the main components on the rear of the Compute Tray configuration that includes Quick Disconnect (QD) ports and Busbar clip.

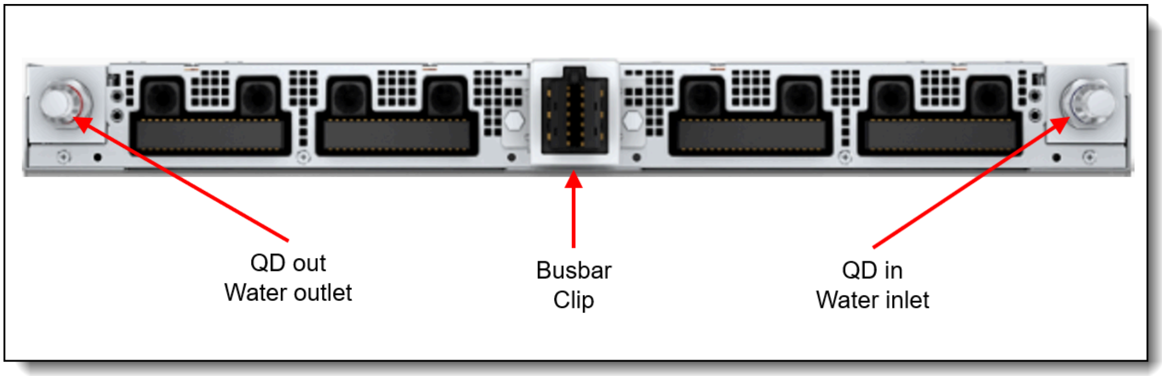

Figure 4. Rear view of the GB300 Compute Tray

The following figure shows key components internal to the Compute Tray.

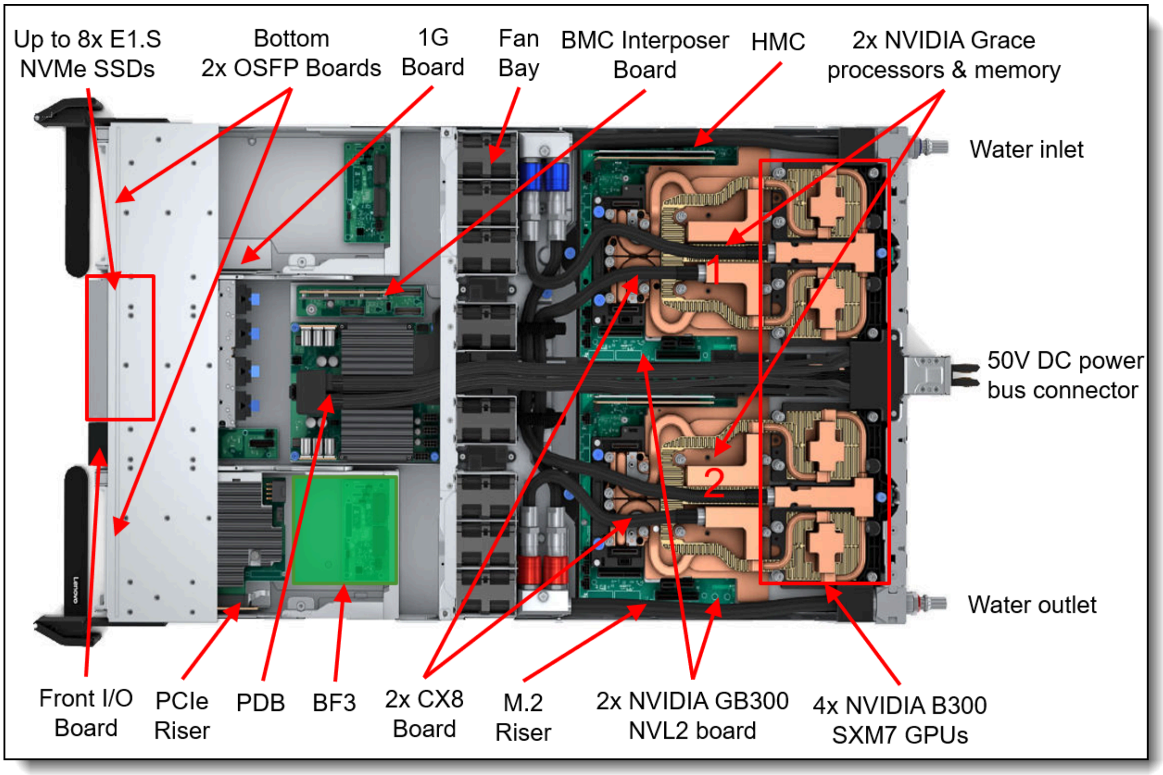

Figure 5. Inside view of the GB300 Compute in the water-cooled tray

The following figure shows key components internal GB300 NVL Switch Tray.

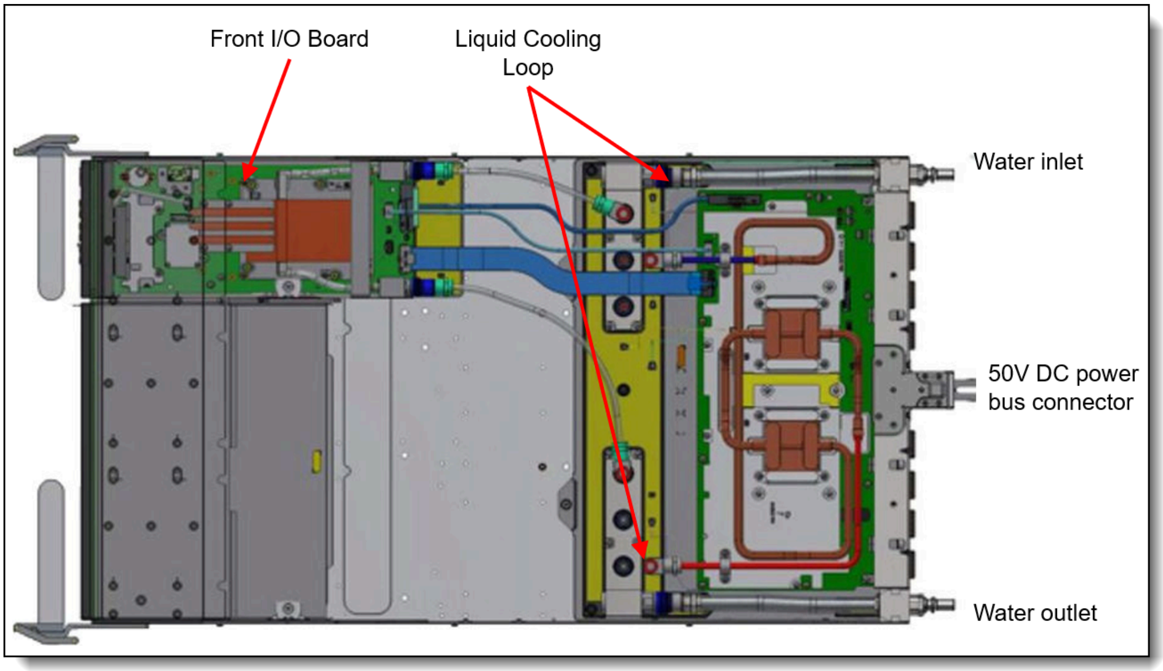

Figure 6. Internal view of the GB300 NVL Switch Tray

System architecture

The NVIDIA GB300 NVL72 is offered with the following integrated systems:

- GB300 Compute tray
- GB300 NVL Switch Tray
- SN2201 Management Switch
- Power Shelf

The following figure shows the architectural block diagram of the GB300 Compute Tray. This configuration includes 4x GPUs and 3x PCIe Gen5 x16 slots all directly connected to the CPUs. This configuration optimizes network connectivity while maximizing computational efficiency, ensuring cost-effectiveness in performance.

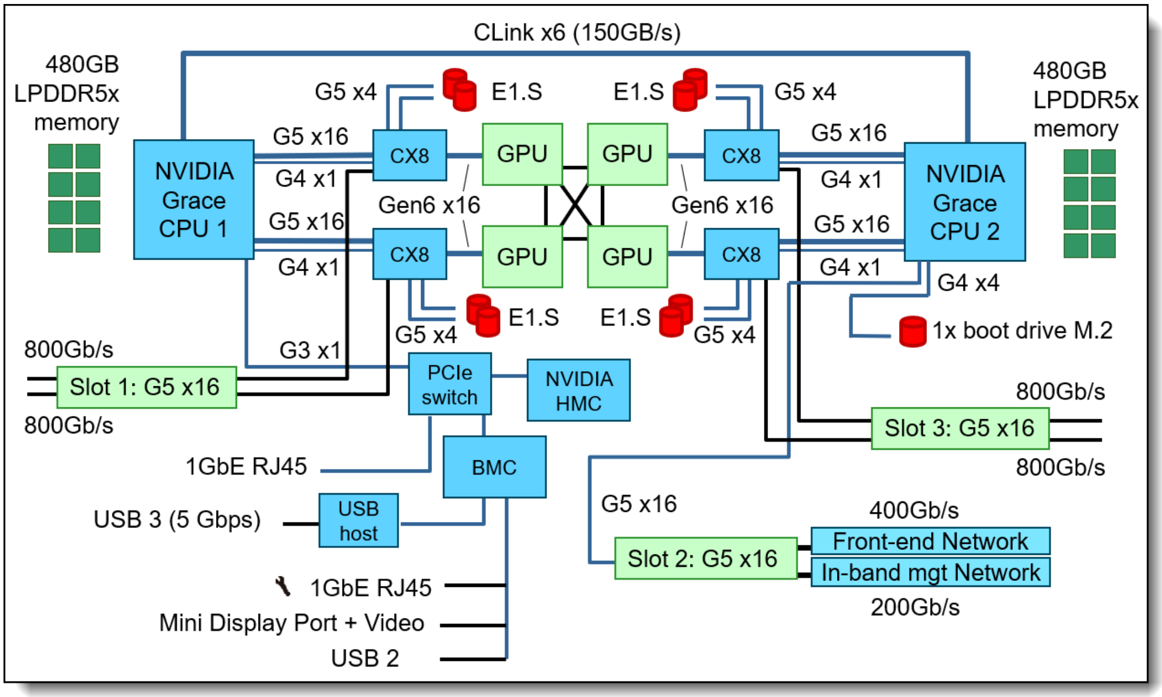

Figure 7. 1U GB300 Compute Tray system architectural block diagram

The following figure shows the architectural block diagram of the GB300 NVL Switch Tray. This switch interfaces by connecting the Compute Tray nodes, GPUs, 800Gb/s OSFP network interfaces and drives all together using ConnectX-8 switches. The AI Training configuration is optimized for scale-out GPU performance which provides advantages in large AI training workloads.

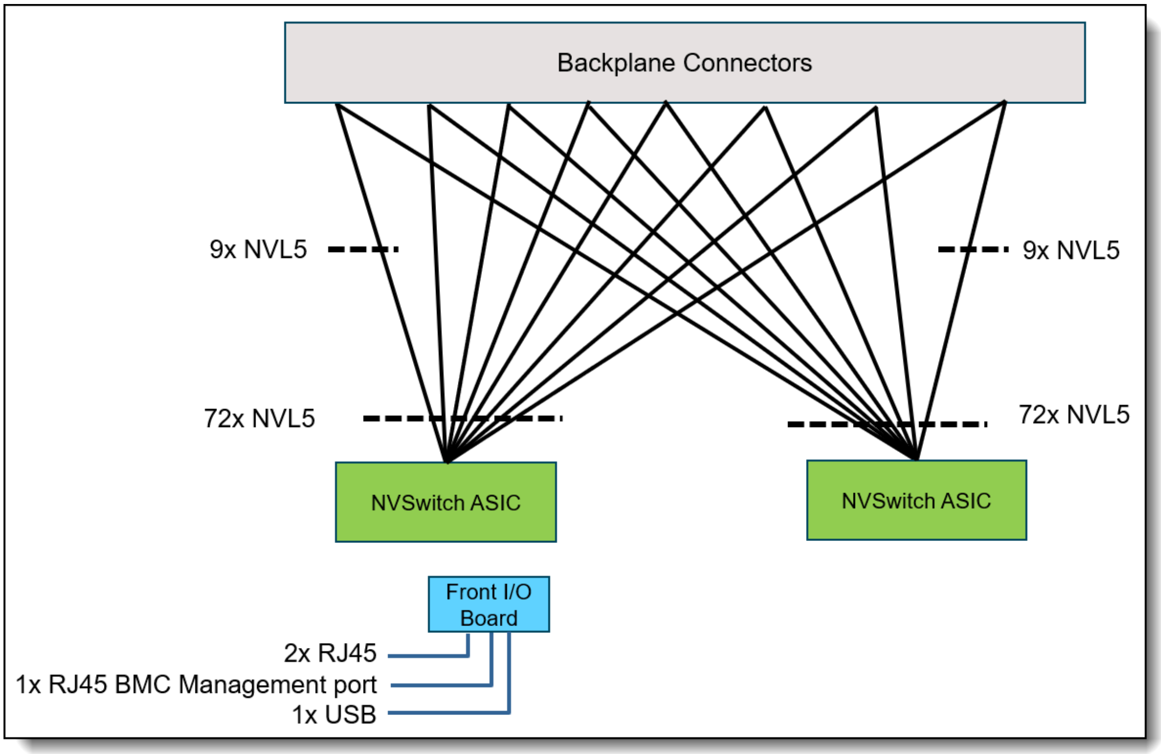

Figure 8. GB300 NVL Switch Tray system architectural block diagram

For more information about the two configurations, see the Configurations section.

Standard specifications - GB300 Compute tray

The following table lists the standard specifications of the Lenovo NVIDIA GB300 Compute tray.

Table 2. Standard specifications GB300 Compute tray

Components	Specification
Machine type	7DLZ - 3-year warranty
Form factor	GB300 Compute architecture system in a hybrid air/water-cooled compute tray, installed horizontally in a rack
Processor	Two NVIDIA Grace processors, with 72 Arm Neoverse V2 cores, core speeds of up to 3.1 GHz base frequency, and TDP rating of 300W. Supports PCIe 5.0 for high performance connectivity to network adapters and NVMe drives.
Chipset	None. Integrated into the processor.
GPUs	Four NVIDIA Blackwell B300 GPUs per tray, each with 186 GB HBM3e memory
CPU memory	Up to 480GB of LDPPR5 ECC memory per processor, memory operates at 4266 MHz
Memory maximum	960GB per Compute Tray
Memory protection	ECC with Single Error Correction (SEC), Double Error Detection (DED, parity, scrubbing, poisoning, and page off-lining
Disk drive bays	Each Compute Tray supports up to 8x EDSFF E1.S NVMe SSDs, installed in 8 internal drive bays (non-hot-swap). Only 4x 8x EDSFF E1.S NVMe SSDs populated in reference architecture. Each drive has a PCIe x4 host interface.
Maximum internal storage	61.4TB using 8x 7.68TB E1.S NVMe SSDs
Storage controllers	• Onboard NVMe ports (RAID provided by operating system, if desired) • Software RAID only (no hardware RAID controller)
Optical drive bays	None
Network interfaces	• Four OSFP (NVIDIA CX8) ○ CX8 I/O board 900-9X86E-00CX-SP0 – up to 800Gb/s
PCI Expansion slots	• One PCIe Gen5 x16 Slot (BlueField-3 DPU) • Two PCIe Gen5 x16 Slot (OSFP Boards) • Up to four PCIe Gen5 x4 E1.S SSDs
Power Supply	• Input Source: Input Power will be 50V from the bus bar (Power Shelf) • Input Range: 47.5-51.5V (50V nominal)
Cooling	Hybrid cooling: • The CPU, GPU, and HBM (GPU memory) are liquid cooled managed by a CDU, rear manifold, quick disconnects and cold plates • Front option components, Power Distribution Board, and M.2 are air cooled managed by 8x built-in fans
System LEDs	System identification, reset, and system power. Each power conversion station (PCS) has AC, DC, and error LEDs.

Components	**Specification**
Temperature	- ASHRAE standard A2 compliant - Operating air temperature: - 10°C to 35°C (50°F to 95°F) - Operating air temperature: - De-rated 1C per 300m (10,00ft) to 1,371m (4,500ft) above sea level - Maximum rate of change (°C /hr) ≤ 20 See Operating Environment for more information.
Fans	- Eight 1U Dual Rotor 40x56 Fans - Supports up to 34k/32.5k RPM
Electrical power	- Input source: 50V MGX rack power bus bar - Input range: 47.5V – 51.5V (50V nominal)
Power cords	Custom power cables for direct data center attachment, 1 dedicated (32A) or 1 shared (63A) per power module
External USB	- One USB 3.0 port - One USB 2.0 for debug use
Video	One video port (Mini Display Port). Maximum resolution is 1920x1200 32bpp at 60Hz.
Security features	Power-on password, administrator's password, Trusted Platform Module (TPM), supporting TPM 2.0.
BMC Management	- Embedded management based on the ASPEED AST2600 baseboard management controller (BMC) - One RJ-45 Connector (10/100/1000 Mbps RJ-45) - NVIDIA management software provides additional systems management functions from power monitoring to liquid leakage detection, tray and power conversation station level
Operating systems supported	Ubuntu is Supported & Certified. See the Operating system support section for details and specific versions.
Limited warranty	Three-year or five-year customer-replaceable unit and onsite limited warranty with 9x5 next business day (NBD).
Service and support	Optional service upgrades are available through Lenovo Services: 4-hour or 2-hour response time, 6-hour fix time, 1-year or 2-year warranty extension, software support for Lenovo hardware and some third-party applications.
Dimensions	Width: 438 mm (17 inches), height: 43.6 mm (1.72 inches), depth: 766 mm (30.16 inches)
Weight	29 kg (63.93 lbs)

Notes:

Onsite spares (compute nodes) would be needed to support an onsite Next Business Day Service Level target. 2% of the order recommended quantity. For Compute Tray base alternatives available for selection, see the Spare Compute Trays section.

Standard specifications - GB300 NVL Switch Tray

The following table lists the standard specifications of the Lenovo GB300 NVL Switch Tray.

Table 3. Standard specifications GB300 NVL Switch Tray

Components	Specification
Machine type	7DJY - 3-year warranty
Form factor	GB300 NVL Switch architecture system in a water-cooled compute tray, installed horizontally in a rack
I/O architecture	None integrated. The NVL72 configuration includes connections for all 72 GPUs with no front OSPF connections. Use top-of-rack networking switches.
Power Supply	• Input Source: Input Power will be 50V from the bus bar (Power Shelf) • Input Range: 47.5-51.5V (50V nominal)
Cooling	Direct water cooling supplied by water hoses connected to the rear of the enclosure.
System LEDs	System error, identification, status, and system power. Each power conversion station (PCS) has AC, DC, and error LEDs.
Systems management	The NVL Switch Tray includes connection for a COMe module board used for NVL Switch Tray management.
Temperature	• Operating water temperature: ○ 2°C to 50°C (35.6°F to 122°F) (ASHRAE W45 compliant) • Operating air temperature: ○ 5°C to 40°C (41°F to 104°F) (ASHRAE A3 compliant) See Operating Environment for more information.
Electrical power	3-phase 200V-480Vac
Power cords	Custom power cables for direct data center attachment, 1 dedicated (32A) or 1 shared (63A) per power module
Ports	DisplayPort , 1x USB port and 2x RJ-45 1GbE.
Limited warranty	Three-year or five-year customer-replaceable unit and onsite limited warranty with 9x5 next business day (NBD).
Service and support	Optional service upgrades are available through Lenovo Services: 4-hour or 2-hour response time, 6-hour fix time, 1-year or 2-year warranty extension, software support for Lenovo hardware and some third-party applications.
Dimensions	Width: 438 mm (17 inches), height: 43.6 mm (1.72 inches), depth: 766 mm (30.16 inches)
Weight	17 kg (37.5 lbs)

Models

The NVIDIA GB300 NVL72 rack is configured by using the configure-to-order (CTO) process with the Lenovo Cluster Solutions configurator (x-config) or in Lenovo Data Center Solution Configurator (DCSC).

The following table lists the base CTO models and base feature codes

Table 4. Base CTO models

Machine Type/Model	Feature code	Description
7DJVCTO2WW	C5RK	48U MGX Rack (Rack Base Non-redundant Power, 6 Shelves)
7DJVCTO2WW	C5RJ	48U MGX Rack (Rack Base Redundant Power, 8 Shelves)
7DLZCTO2WW	CAAX	GB300 Compute tray
7DJYCTO3WW	C63C	GB300 NVL Switch Tray
7DJWCTO1WW	C5RL	SN2201 Management Switch

The Lenovo NVIDIA GB300 NVL72 node is equipped with high-performance GPUs. It is RoHS complaint and meets all the environmental certifications.

Configurations

There are two primary configurations offered with the NVIDIA GB300 NVL72. Block diagrams for Compute Tray and Switch Tray leveraged in these configurations are shown in the System architecture section.

The following table lists the Bill of Materials for the **Non-redundant Power with 6 Power Shelves**

Table 5. Bill of Materials for Non-redundant Power with 6 Power Shelves

Feature code	Description	Quantity
7DJVCTO2WW	48U MGX Rack (Rack Base Non-redundant Power, 6 Shelves)	1
C5RK	Rack Base Non-redundant Power, 6 Shelves	1
C5RR	Power Shelf, 1U, 33kW	6
C5RN *	Manifold R&L, Bottom Feed	1
C5RT	NVL Power Whips Pair, Standard	3
7DJYCTO3WW	1U Compute Tray leverages the NVIDIA GB300 NVL72	18
CCCT	Nvidia GB300 NVL72 HPM 1CPU:2GPU PC	2
C7VH**	Samsung PM9D3a 3.84T MZTL63T8HFLT-00AW7 E1.S 9.5mm SSD without Latch	4
CA62	GB300 15mm 4-Bay E1.S BP	2
BTMB	PG8A0N/PG8C0N M.2 Riser	1
C58W	ThinkSystem M.2 SATA/x4 NVMe 2-Bay Adapter	1
CERX	Samsung PM9A3 1.92T MZ1L21T9HCLS-00A07 M.2 NVMe SSD	1
C58N	Nvidia BF3 B3240 900-9D3B6-00CN-PA0 2X400G FHHL QSFP11 DPU NIC	1
CCCS	Nvidia 900-9X86E-00CX-ST0 CX-8 IO Board Partner Cooled with 2 CX-8 NIC	1
C5CN	Nvidia 900-24764-0000-000 HMC SKU1 for GB300	1
CC57	GB300 PCIe Riser Cable	1
CCCU	GB300_Waterloop ASM Coldplate inner manifold	1
CC5E	GB300 1RU Busbar Cable	1
CC50	GB300 Latch E1S for one tray	1
7DJYCTO3WW	Lenovo NVIDIA GB300 NVLink Switch Tray	9
C63C	GB300 NVL72 NS Switch Tray Base	1
C6MW	GB300 NVL72 Redundant Power Shelf Solution	1
7DJWCTO2WW	NVIDIA SN2201 48 port DC switch	2
C5RL	NVIDIA SN2201 48 port DC switch	1
C6MV	GB300 NVL72 Non-Redunant Power Shelf Solution	1

* Either C5RN Bottom feed or CA40 Top feed
** Up to 8 drives are supported

The following table lists the Bill of Materials for **Redundant Power with 8 Power Shelves**

Table 6. Bill of Materials for Redundant Power with 8 Power Shelves

Feature code	Description	Quantity
7DJVCTO2WW	48U MGX Rack (Rack Base Non-redundant Power, 8 Shelves)	1
C5RJ	Rack Base Redundant Power, 8 Shelves	1
C5RR	Power Shelf, 1U, 33kW	8
C5RN *	Manifold R&L, Bottom Feed	1
C5RT	NVL Power Whips Pair, Standard	4

Feature code	Description	Quantity
7DJYCTO3WW	1U Compute Tray leverages the NVIDIA GB300 NVL72	18
CCCT	Nvidia GB300 NVL72 HPM 1CPU:2GPU PC	2
C7VH**	Samsung PM9D3a 3.84T MZTL63T8HFLT-00AW7 E1.S 9.5mm SSD without Latch	4
CA62	GB300 15mm 4-Bay E1.S BP	2
BTMB	PG8A0N/PG8C0N M.2 Riser	1
C58W	ThinkSystem M.2 SATA/x4 NVMe 2-Bay Adapter	1
CERX	Samsung PM9A3 1.92T MZ1L21T9HCLS-00A07 M.2 NVMe SSD	1
C58N	Nvidia BF3 B3240 900-9D3B6-00CN-PA0 2X400G FHHL QSFP11 DPU NIC	1
CCCS	Nvidia 900-9X86E-00CX-ST0 CX-8 IO Board Partner Cooled with 2 CX-8 NIC	1
C5CN	Nvidia 900-24764-0000-000 HMC SKU1 for GB300	1
CC57	GB300 PCIe Riser Cable	1
CCCU	GB300_Waterloop ASM Coldplate inner manifold	1
CC5E	GB300 1RU Busbar Cable	1
CC50	GB300 Latch E1S for one tray	1
7DJYCTO3WW	Lenovo NVIDIA GB300 NVLink Switch Tray	9
C63C	GB300 NVL72 NS Switch Tray Base	1
C6MW	GB300 NVL72 Redundant Power Shelf Solution	1
7DJWCTO2WW	NVIDIA SN2201 48 port DC switch	2
C5RL	NVIDIA SN2201 48 port DC switch	1
C6MV	GB300 NVL72 Non-Redunant Power Shelf Solution	1

* Either C5RN Bottom feed or CA40 Top feed
** Up to 8 drives are supported
For more information about the, see the Lenovo NVIDIA GB300 NVL72 datasheet.

The following table lists the Configurators used to create the above configurations.

Table 7. Method to configure the NVIDIA GB300 NVL72

Configuration	Config Tool	Description
NVL72	DCSC	• General Purpose Mode ○ Select Servers and Large AI Optimized and then select NVIDIA GB300 NVL72 • Deployment Ready Solutions ○ Select AI, then select NVIDIA GB300 NVL72
NVL72	x-Config	• Edit an existing rack • Select the Lenovo 48U MGX Rack

Configurator tips:

- The following Compute Tray Base alternatives are available for selection:
  - For Rack Installed compute Trays: Select Lenovo Feature Code CABF - 1U 4GPU GB300 DWC Front IO
  - For Spare compute Trays with individual node packaging: Select Lenovo Feature Code CD96 - 1U 4GPU GB300 DWC Front IO w Package - Mainly for customer who would like to order for replacement or as a spare.
  - For Service Spare compute Trays with individual node packaging: Select Lenovo Feature Code CCQD - 1U 4GPU GB300 DWC Front IO w/o E1.S, w Package" – For Service, for repair purpose.

Spare Compute Trays

To guarantee a consistent next business day response objective and minimize potential downtime, we strongly recommend the purchase of on-site spare Compute Trays**.** Lenovo recommends a minimum spare-to-production ratio of 2%, however the ideal number of spares depends on the scale of your deployment. Contact Lenovo Services for tailored guidance on the appropriate number of spares.

The following Compute Tray Base alternatives are available for selection:

Table 8. On-site spare Compute Trays

Base selection	Feature Code	Description
Rack Installed compute Trays	CABF	1U 4GPU GB300 DWC Front IO
Spare compute Trays with individual node packaging	CD96	1U 4GPU GB300 DWC Front IO w Package - Mainly for customer who would like to order for replacement or as a spare
Service Spare compute Trays with individual node packaging	CCQD	1U 4GPU GB300 DWC Front IO w/o E1.S, w Package – For Service, for repair purpose

NVIDIA GB300 NVL72 HPM

The NVIDIA GB300 NVL72 leverages two NVIDIA GB300 NVL72 host processor motherboards (HPMs) per tray that integrate processors, memory and GPU accelerators in a powerful package.

- Processors
- Memory
- GPU accelerators

Processors

Each of the NVIDIA GB300 NVL72 HPMs include one NVIDIA Grace processor, as part of the NVIDIA GB300 NVL72 module. The processors are standard in all configurations of the NVIDIA GB300 NVL72.

Each processor has the following features:

- 72 Arm Neoverse V2 cores with 4x 128b SVE2
- 3.1 GHz base frequency
- 3.0 GHz all-core SIMD frequency
- L1 cache: 64KB i-cache + 64KB d-cache
- L2 cache: 1MB per core
- L3 cache: 114MB
- 64 PCIe Gen5 lanes (implemented as 8x x8)
- 300 GB/s bi-directional NVLink-C2C connectivity between processors
- 450 GB/s bi-directional NVLink connectivity to two GPUs each
- NVIDIA Scalable Coherency Fabric (SCF) with 3.2 terabytes/s bisection bandwidth interconnecting Cores, Memory and IO
- 300W TDP (includes power for memory)

The NVIDIA Grace processor is integrated using Ball Grid Array (BGA) technology and is soldered onto the NVIDIA GB300 NVL72 HPM board. To replace the CPU, it is necessary to replace the entire module.

Memory

The NVIDIA GB300 NVL72 HPM boards incorporate processor memory via LPDDR5x modules.

Features of the memory subsystem are as follows:

- 960GB of LPDDR5X ECC memory in a fixed configuration, 480GB per processor
- Implemented using 8x 128GB modules (4 per CPU)
- Peak memory bandwidth per CPU: Up to 384 GB/s
- Memory clock per CPU: 4266MT/s

The Single Error Correction (SEC) and Double Error Detection (DED) algorithms as part of the ECC are designed to mitigate multi-bit aliasing and correct errors effectively. Furthermore, features such as parity, scrubbing, poisoning, and page off-lining are implemented to enhance Reliability, Availability, and Serviceability (RAS).

GPU accelerators

The NVIDIA GB300 NVL72 HPM boards support four onboard NVIDIA B300 GPUs. The following table lists the specifications for each GPU and for the full package. Measured performance will vary by application.

The following table lists the specifications for each GPU.

Table 9. GPU specifications

Specification	**Each NVIDIA B300 GPU**
FP64	40 teraFLOPS
FP64 Tensor Core	40 teraFLOPS
FP32	80 teraFLOPS
TF32 Tensor Core	2,500 teraFLOPS*
BFLOAT16 Tensor	5,000 teraFLOPS*
FP16 Tensor Core	5,000 teraFLOPS*
FP8 Tensor Core	10,000 teraFLOPS*
FP4 Tensor Core	20,000 teraFLOPS*
INT8 Tensor Core	10,000 TOPS*
GPU Memory	279 GB HBM3e per GPU
GPU Memory Bandwidth	8 TB/s
GPU Memory ECC	Supported (enabled by default)
Multi-Instance GPUs (MIG)	Up to 7 Instances
Total Graphics Power (TGP)	1400W
Interconnect	NVLink between GPUs: 1.8 TB/s (full mesh) NVLink to CPUs: 450GB/s per GPU PCIe Gen6 x16: 256GB/s per GPU

* With structural sparsity enabled

The following figure shows how the GPUs interconnect.

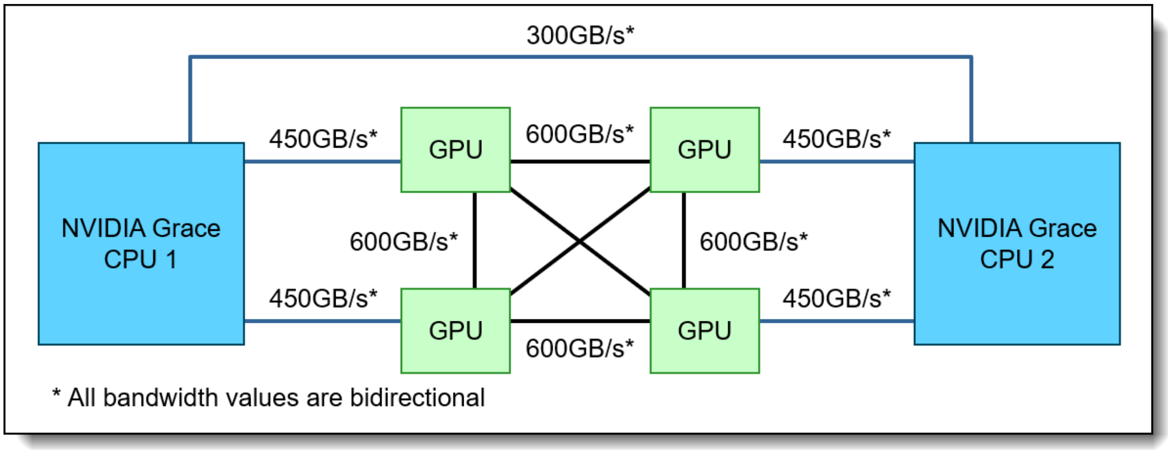

Figure 9. Compute Tray NVLink topology

Internal storage

The NVIDIA GB300 NVL72 node supports the following drives.

- Up to 8x E1.S drives for data storage.
  - These are front accessible.
- 1x M.2 NVMe drive for OS boot
  - This is an internal drive - not front accessible and non-hot-swap.

The following images shows drives that come standard in the Compute Tray and their location.

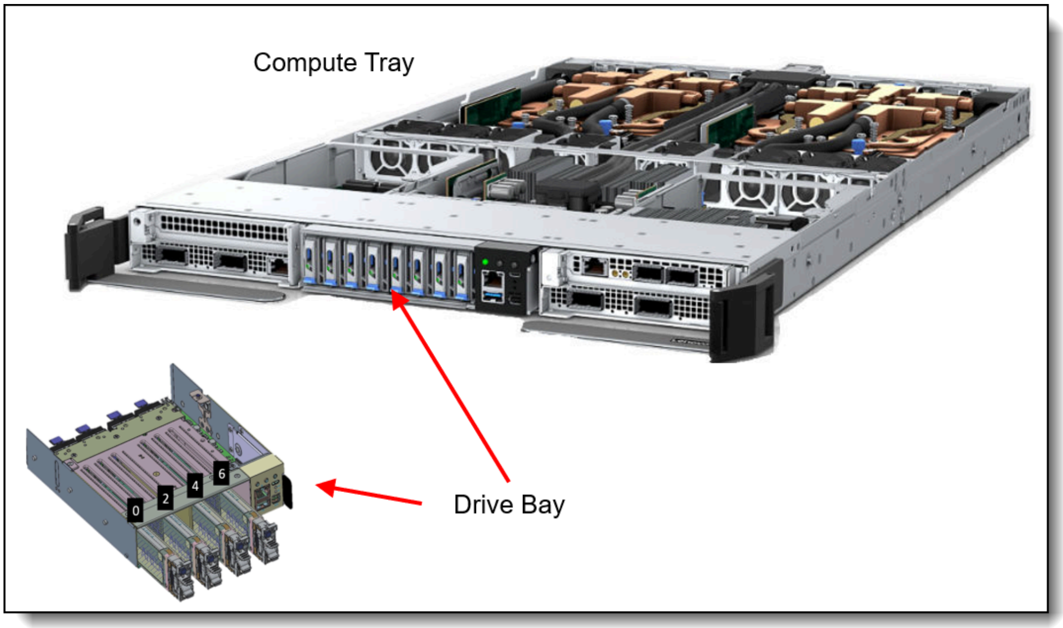

Figure 10. NVIDIA GB300 NVL72 internal drive bays

Each of the 8 drive bays supports either of the following:

- E1.S 7.68TB drives
- E1.S 3.84 TB drives

Configuration notes:

- E1.S installation rule:
  - 1. Default: 4 x E1.S SSDs, install bays 0, 2, 4, 6
  - 2. If select 8 x E1.S SSDs, install bays 0, 1, 2, 3, 4, 5, 6, 7
  - Dummy Panels required to fill in unused bay slots
- The Compute Tray supports E1.S NVMe SSDs for data and 1× M.2 NVMe SSD for OS boot
- The drives are connected to onboard controllers; RAID functionality is provided using the operating system features
- Boot drive is the same as other drives but use switched PCIe and may have lower performance. The M.2 riser kit (feature C58W) is needed. The M.2 riser kit is not field upgradable.
- The same PCIe lanes can be used either for PCIe adapters or for NVMe drives.
- NVMe drives are connected to CPUs as follows:
  - Drives 0-3 are data storage drives connected to CX8 / CPU 1
  - Drives 4-7 are data storage drives connected to CX8 / CPU 2

The feature codes to select the appropriate storage cage are listed in the following table.

Tip: To configure boot drives, specify feature C58W plus feature CERX or C43X (1x boot drive). See the table below. The boot drive kit is only required when 8 performance and 1 boot drives are installed.

Table 10. Drive mounting kits

Feature code	Description	Max qty	Purpose
C58W	PG8A0N/PG8C0N M.2 Riser	1	Enables the use of boot drive
CA62	GB300 15mm 4-Bay E1.S BP	2	Backplane board
CERX	Samsung PM9A3 1.92T MZ1L21T9HCLS-00A07 M.2 NVMe SSD	1	M.2 NVMe drive for OS boot
C43X	Micron 7450PRO 1.92T MTFDKBG1T9TFR-1BC15ABYY M.2 NVMe SSD	1	M.2 NVMe drive for OS boot

Controllers for internal storage

The drives of the NVIDIA GB300 NVL72 are connected to integrated NVMe storage controllers.

If desired, RAID functionality is provided by the installed operating system (software RAID only)

Internal drive options

The following table lists the drive options for internal storage of the Compute Tray or server.

Table 11. E1.S EDSFF drives

Part number	Feature code	Description	Max Qty
E1.S trayless SSDs - PCIe 5.0 NVMe			
CTO only	CFY3	Solidigm PS1010 7.68T SB5PHU7X076TNV1 E1.S 9.5mm NVMe SSD without Tray	8
CTO only	CG6E	Kioxia XD8 3.84T KXD8DRJ93T84 E1.S 15mm NVMe SSD without Tray	8
CTO only	CA5C	Micron 9550 3.84T MTFDLCE3T8THA-1BK1DABYY E1.S 15mm NVMe SSD without Tray	8
E1.S without latch SSDs - PCIe 5.0 NVMe			
CTO only	CFY2	Micron 9550 7.68T MTFDLCE7T6THA-1BK1DABYY E1.S 15mm SSD without Latch	8
CTO only	C7VJ	Sandisk SN861 3.84T SDS6A7638PKP8X7 E1.S 15mm SSD without Latch	8
CTO only	C7VH	Samsung PM9D3a 3.84T MZTL63T8HFLT-00AW7 E1.S 9.5mm SSD without Latch	8
M.2 NVMe SSDs - PCIe 4.0 NVMe			
CTO only	C43X	Micron 7450PRO 1.92T MTFDKBG1T9TFR-1BC15ABYY M.2 NVMe SSD	1
CTO only	CERX	Samsung PM9A3 1.92T MZ1L21T9HCLS-00A07 M.2 NVMe SSD	1

I/O expansion

The NVIDIA GB300 NVL72 node supports up to 3x front-accessible slots, depending on the configuration:

- Slots 1, 2: PCIe 5.0 x16 slots connected to CPU 1
- Slots 3: PCIe 5.0 x16 slot connected to CPU 2

The location of the slots is shown in the following figure.

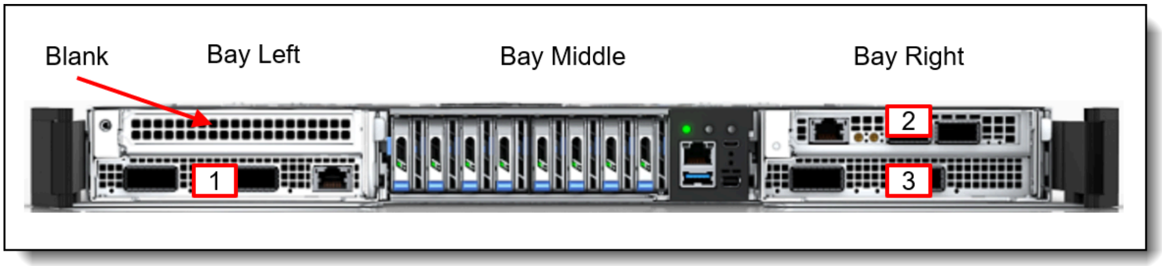

Figure 11. NVIDIA GB300 NVL72 PCIe slots

Slot availability is based on the configuration selected. Derived parts come standard within the 7DLZCTO2WW GB300 Compute tray BOM.

PCIe adapters and data drives share the same PCIe connections to the processors. The following table lists the supported data and boot drives.

Table 12. Data and boot drives

Front adapter slots	Data drives (E1.S)	Boot drives (E1.S)
1-2-3 adapters	4 - 8 data drives*	1 boot drive

* 4x E1.S drives populated by default in the Compute Tray

OSFP 800Gb ports

The GB300 Compute tray configuration includes two OSFP interface boards (feature CCCS), each providing two OSFP ports. Supported connections are each up to 800 Gb/s (a total of 3200 Gb/s with 4x ports).

Supported cables and transceivers follow ConnectX-8 OSFP compatibility matrix.

For more information, including the supported transceivers and cables, see the adapter product guide:

- ConnectX-8: https://lenovopress.lenovo.com/lp2163-thinksystem-nvidia-connectx-8-8180-800gbs-osfp-pcie-gen6-x16-adapter

Network adapters

The NVIDIA GB300 NVL72 supports up to 3 network adapters installed in the front PCIe slots. The following table lists the supported adapters.

Table 13. Network adapters

Part Number	Feature code	Description	Max Qty
CTO only	CCCS	NVIDIA 900-9X86E-00CX-ST0 CX-8 IO Board Partner Cooled with 2 CX-8 NIC	2
CTO only	C58N	NVIDIA BF3 B3240 900-9D3B6-00CN-PA0 2X400G FHHL QSFP11 DPU NIC	1
CTO only	C5CN	NVIDIA 900-24764-0000-000 HMC SKU1 for GB300	1

For more information, including the supported transceivers and cables, see the adapter product guide:

- ConnectX-8: https://lenovopress.lenovo.com/lp2163-thinksystem-nvidia-connectx-8-8180-800gbs-osfp-pcie-gen6-x16-adapter

Cooling

The Lenovo NVIDIA GB300 NVL72 offering supports the two most common fluids used the de-ionized water and PG25.

The customer can choose the coolant they want to use:

- Deionized (DI) water (recommended); Rack ships without water
- PG25 (glycol-based)

One of the most notable features of the Lenovo NVIDIA GB300 NVL72 offering is direct water cooling. Direct water cooling (DWC) is achieved by circulating the cooling water directly through cold plates that contact the GPUs, CPU and memory in the Compute Trays and through cold plates in the NVL Switch Trays.

The NVIDIA GB300 NVL72 by Lenovo features a mix of air and liquid cooling to meet its demanding power requirements. Each rack consumes 135 kW TDP; up to 155 kW peak depending on workload and EDP behavior. At the rack level, the split between air and liquid heat capture is about 10% to air and 90% to liquid.

GPU Compute Tray contains a hybrid cooling solution. The CPU, GPU, and HBM (GPU memory) are liquid cooled while the front option components, Power Distribution Board, and M.2 are air cooled. Therefore, while the air cooled component temperatures drive fans speeds, the liquid cooled components do not change fluid flow as a function of component temperature.

There are several components' temperatures monitored to properly drive the eight system fan pairs. They are OSFP transceivers, QSFP transceivers, BlueField-3, Power Distribution Board components, ambient air, BMC, HMC, M.2, and E1.S components.

One of the main advantages of direct water cooling is the water can be relatively warm and still be effective because water conducts heat much more effectively than air. Depending on the environmentals like water and air temperature, effectively 90% of the heat can be removed by water cooling; in configurations that stay slightly below that, the rest can be easily managed by a standard computer room air conditioner.

Allowable inlet temperatures for the water can be as high as 45°C (113°F) with the NVIDIA GB300 NVL72 for real-world applications. In most climates, water-side economizers can supply water at temperatures below 45°C for most of the year. This ability allows the data center chilled water system to be bypassed thus saving energy because the chiller is the most significant energy consumer in the data center. Typical economizer systems, such as dry-coolers, use only a fraction of the energy that is required by chillers, which produce 6-10°C (43-50°F) water. The facility energy savings are the largest component of the total energy savings that are realized when the NVIDIA GB300 NVL72 is deployed.

The advantages of the use of water cooling over air cooling result from water's higher specific heat capacity, density, and thermal conductivity. These features allow water to transmit heat over greater distances with much less volumetric flow and reduced temperature difference as compared to air.

For cooling IT equipment, this heat transfer capability is its primary advantage. Water has a tremendously increased ability to transport heat away from its source to a secondary cooling surface, which allows for large, more optimally designed radiators or heat exchangers rather than small, inefficient fins that are mounted on or near a heat source, such as a CPU.

The NVIDIA GB300 NVL72 offering uses the benefits of water by distributing it directly to the highest heat generating node subsystem components. That energy savings results from the removal of the system fans and the lower operating temp of the direct water-cooled system components.

The direct energy savings at the enclosure level, combined with the potential for significant facility energy savings, makes the NVIDIA GB300 NVL72 an excellent choice for customers that are burdened by high energy costs or with a sustainability mandate.

The liquid cooling solution consists of a coolant distribution unit (CDU), rear manifold, Disconnect Quick

Connect (QD)s, and cold plates for the CPUs, GPUs, CX-8 NICs and all NVL Switch components.

The following table lists the reference design cooling percentages.

Table 14. NVIDIA GB300 NVL72 reference

Subsystem	**Cooling Method**	**% of Heat Removed**
Compute Trays	Hybrid (primarily liquid, some air)	Part of overall 90 % liquid / 10 % air
NVL Switch Trays	Fully liquid-cooled	Included in the 90 % liquid
Rack Ancillary Systems	Air-cooled	~10 %
Total Rack Cooling Split	~90 % liquid / ~10 % air	N/A

Water connections

Water connections to the MGX Rack are provided using hoses that connect directly to the coolant distribution unit (CDU), inlet and return, either an in-rack CDU or in-row CDU. As shown in the following figure, hose lengths required depend on the placement of the Compute Trays and GB300 NVL Switch Trays in the rack cabinet.

The midplane in the chassis routes the water via QuickConnects to each of the Compute Tray, and each of the NVL Switch Trays.

The GB300 offers two options for the rack level liquid cooling manifold: bottom feed and top feed. Both options are shown below.

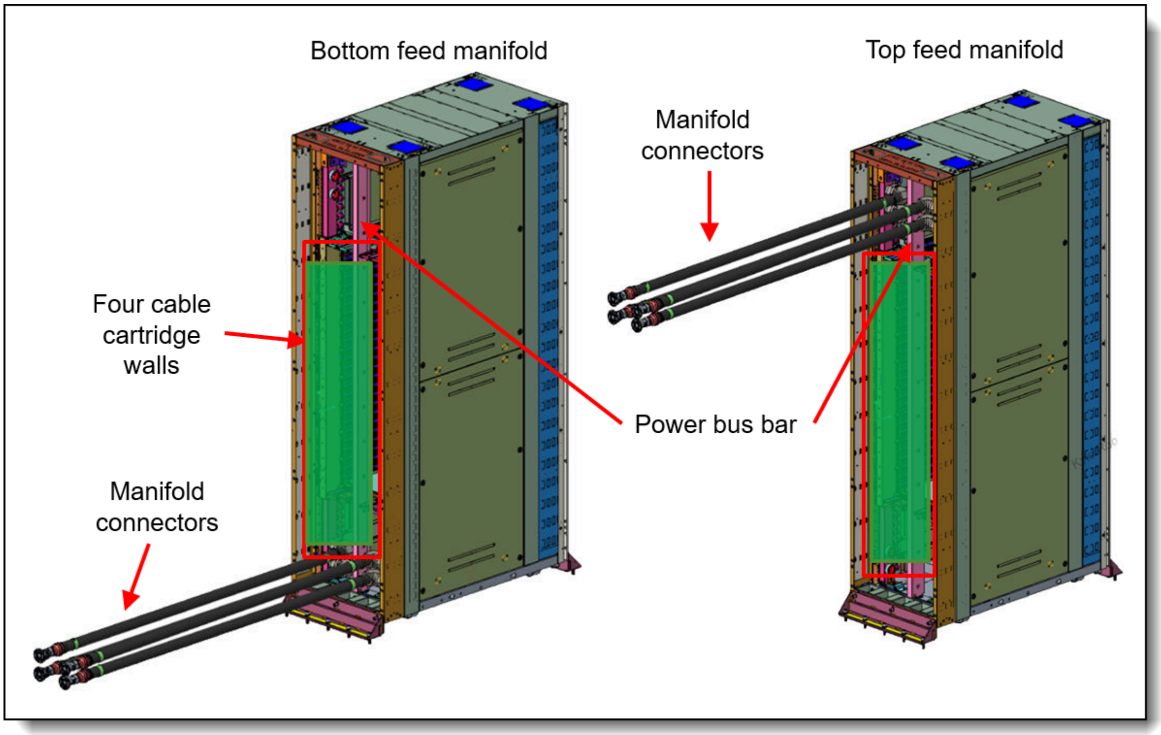

Figure 12. Water connections to the MGX Rack cabinet

Water connections are made up of the following:

- Short hoses permanently attached to the rear of the rack cabinet
- Intermediate hoses, orderable as three different pairs of hoses, based on the location of the components in the rack.
- Connection to the CDU through additional hoses or valves on the data center water loop. Lenovo Datacenter Services can provide the end-to-end enablement for your water infrastructure needs.

For additional information, see the Cooling section.

To support the onsite setup for the direct water-cooled solution, a service Kit is available providing a flow meter, bleed hose, pressure gauge and vent valve.

Table 15. Service kit

Part number	Service kit FRU	Description
4XH7B14786	03PV563	• The equipment shall be prepared and provided by the customer • L2 PN: STL7C32889

Manifold information is listed in the following table.

Table 16. Manifold information

Part number	Feature code	Description
CTO only	CA40	Manifold ASM, Top Feed
CTO only	C5RN	Manifold R&L, Bottom Feed

The liquid cooling manifold information is listed in the following table.

Table 17. Liquid Cooling Manifold

Components	Specification
Ownership	Vendor
System Support	Per NVIDIA reference
Universal	Per NVIDIA reference
Hybrid Tray Support	N/A
Material	Manifold shall be made from 304L or 316L austenitic stainless steel
Serviceability	Serviceable from rear of the rack
Mounting	Per NVIDIA reference
QD Type	UQDB04
Bleed Valve	Manifold should employ UQD04 on top of manifold for bleed valve port
Hose Connection	Manifold shall use tri-clamp connectors for hose adaption
QD Type to CDU	1" FD83 BSPP
Static Pressure at QD	30psi
Form Factor	Per NVIDIA reference

For details on Lenovo Neptune direct water-cooling standards, see the following document:

Lenovo Neptune Direct Water-Cooling Standards https://lenovopress.lenovo.com/lp2018-lenovo-neptune-direct-water-cooling-standards

Power shelf

The Lenovo NVIDIA GB300 NVL72 supports up to 6 or 8 power shelves. The use of water-cooled power conversion stations enables an even greater amount of heat can be removed from the data center using water instead of air-conditioning.

The Power Shelves supply internal system power to a 50V busbar. This innovative design merges power conversion, rectification, and distribution into these power units, a departure from traditional setups that require separate units, including separate power supplies, resulting in best-in-class efficiency.

The following table lists specifications for Power Shelf.

Table 18. Standard specifications Power shelf

Components	Specification
Type / Form Factor	1U 19” MGX Rack Compliant
Output Power	• 33 kW ○ 6x 33 kW power shelves (198 kW) ○ 8x 33 kW power shelves (264 kW)
Output Voltage	50VDC
Output Current	Per Power Shelf Specification
AC Feed	Per Power Shelf Specification
PSUs	6x 5.5kW hot-swappable single phase PSU modules
Input Voltage	Wide input voltage range of 200-277VAC (346-480 VAC WYE 5 wires)
AC input options	• Single phase or three-phase input (208/240/277 Vac nominal) • Single input connector
Efficiency	Minimum 97.5% PSUs Peak 96.5% @230Vac
Management	Support for PMI or PMC
Current Sharing	PSU balancing via Active Sharing or Droop methods
Power Whip	• 1x 60A IEC60309 whip per power shelf ○ 7 pin 60A AC input connector • Power whip should be secured to the back of the rack • Considerations should be made for rack shipment

The following table lists Bus Bar specifications.

Table 19. Bus Bar specifications

Components	Specification
Type	MGX Rack Compliant
Current Rating	1400A

The following table lists other rack components.

Table 20. Other rack components

Components	Specification
Rack Leak Detection Module	• Standalone 3rd party leak detection modules that connect into the OOB Management Switch for rack manager • Several CDU models possess external port for leak detection to work in conjunction with a rack leak pan • Other vendors (i.e. Rittal) provide off the shelf options for Rack Managers • Per NVIDIA leak detection strategy and customer requirement
Dummy Panels	Required to fill in unused rack slots
Rack adapters	Adapters to mount 19” IT gear in ORV3 rack
Rack Leak Tray	Per NVIDIA reference
Rack Manager	Optional rack manager developed by Lenovo to manage power, leak detection and OOB management

The following image shows the power shelf.

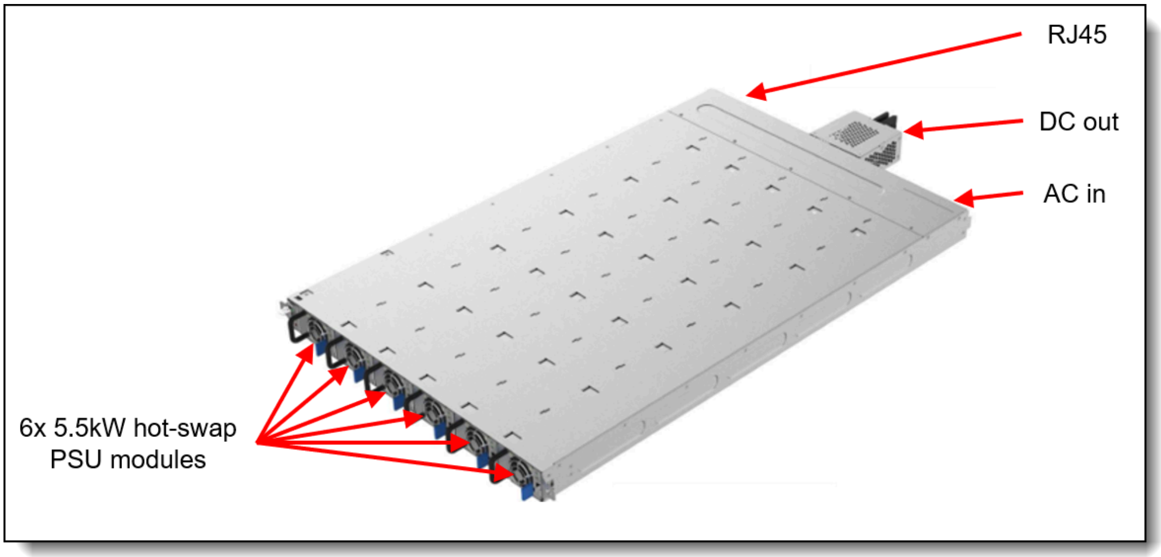

Figure 13. Power shelf

The following table lists the supported power cables.

Table 21. Power shelf cables

Part Number	Feature code	Description
4L67B04869	CCZ9	3.5m 1U 33kW Power Whip
4L67B04867	CCZ7	1.5m 1U 33kW Power Whip

System Management

The NVIDIA management is the BMC board option which provides advanced control, monitoring, and alerting functions. The BMC is used in conjunction with the Host Module Controller (HMC) to provide OOB management. The HMC interposer board is within the GPU Compute Tray, which connects the HMC board and the BMC board to the GB300 HPM via cables.

The HMC and BMC work together to provide the out-of-band manageability for the tray.

Topics in this section:

- Local management
- NVIDIA AI Enterprise

Local management

The following figure shows the ports and LEDs on the front of the GB300 Compute tray.

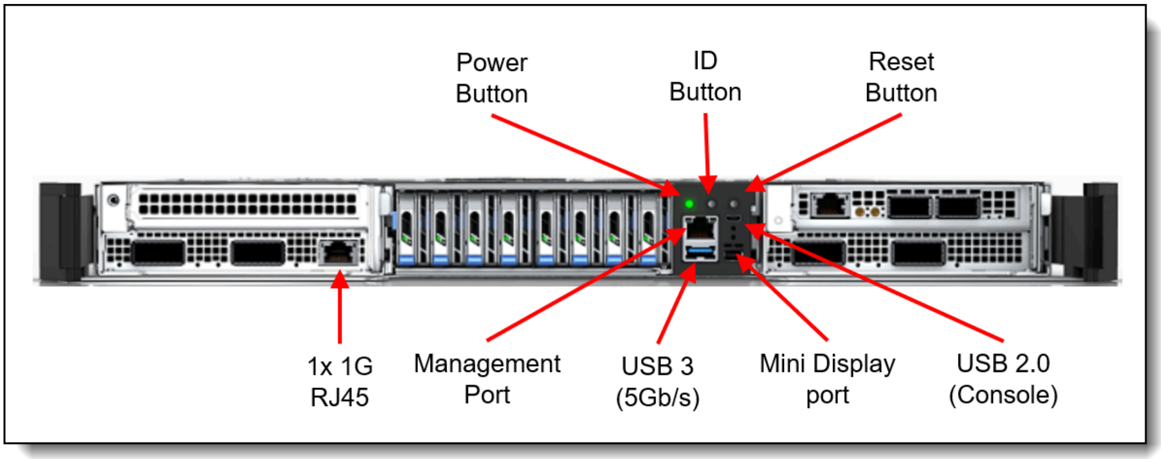

Figure 14. NVIDIA GB300 NVL72 Front operator panel

The LEDs are as follows:

- Identification (ID) LED (blue)to identify a specific server to local service engineers
- Error LED (yellow): Indicates if there is a system error.
- Drive LED: Indicates activity on any drive

Additionally, the following table lists the features of the BMC Server Management.

Table 22. BMC Server Management

Component	Description
BMC Chip	ASPEED 2600
Management	BMC Management Board
Firmware Update	Web / Redfish / Tool OOB update
Telemetry	• Temperature, voltage, power, current of CPU/GPU/HSC/HW monitor chip/PCIe Card • Fan PWM and tachometer • BMC SEL • CPER Log
Reference Solution	NVIDIA Design Kit, AMI GB300 code base
Dedicate/Share MGMT Port	Dedicate MGMT port only
Firmware Codebase	AMI based
Feature Support	• KVM and virtual media- Network Connectivity for Management • Thermal and Fan Management- IPMI (SSIF) to the host CPU • Redfish Host Interface • UART/SOL support • Run Power Control • FW update/recovery for SBIOS/VBIOS/FPGA/BMC • Get boot progress code • Show FRU information

Component	Description
Security	Microchip EROT (CEC1736-S0-I/2ZW-PROTO2)
Failover	The BMC supports boot failover/recovery using primary and secondary SPI flash.

The following table lists the features of the SBIOS.

Table 23. SBIOS (Grace BIOS)

Component	Description
Type	UEFI
Boot Support	USB/ PXE/ HDD
TCM Support (PRC)	Not Supported
TPM support (ROW)	Yes
ACPI	ACPI 6.5 compatible
Firmware Codebase	AMI Based
BIOS Features	The BIOS supports boot failover/recovery using primary and secondary SPI flash

The following table lists the supported CAT6 cables.

Table 24. Local management cabling

Part Number	Feature code	Description
4X97B04866	CCYY	2.5m Green Cat6 Cable
4X97B04864	CCYW	2m Green Cat6 Cable

NVIDIA AI Enterprise

The NVIDIA GB300 NVL72 is designed for NVIDIA AI Enterprise, which is a comprehensive suite of artificial intelligence and data analytics software designed for optimized development and deployment in enterprise settings.

NVIDIA AI Enterprise includes workload and infrastructure management software known as Base Command Manager. This software provisions the AI environment, incorporating the components such as the Operating System, Kubernetes (K8S), GPU Operator, and Network Operator to manage the AI workloads.

Additionally, NVIDIA AI Enterprise provides access to ready-to-use open-sourced containers and frameworks from NVIDIA like NVIDIA NeMo, NVIDIA RAPIDS, NVIDIA TAO Toolkit, NVIDIA TensorRT and NVIDIA Triton Inference Server.

- **NVIDIA NeMo** is an end-to-end framework for building, customizing, and deploying enterprise-grade generative AI models; NeMo lets organizations easily customize pretrained foundation models from NVIDIA and select community models for domain-specific use cases.
- **NVIDIA RAPIDS** is an open-source suite of GPU-accelerated data science and AI libraries with APIs that match the most popular open-source data tools. It accelerates performance by orders of magnitude at scale across data pipelines.
- **NVIDIA TAO Toolkit** simplifies model creation, training, and optimization with TensorFlow and PyTorch and it enables creating custom, production-ready AI models by fine-tuning NVIDIA pretrained models and large training datasets.
- **NVIDIA TensorRT**, an SDK for high-performance deep learning inference, includes a deep learning inference optimizer and runtime that delivers low latency and high throughput for inference applications. TensorRT is built on the NVIDIA CUDA parallel programming model and enables you to optimize inference using techniques such as quantization, layer and tensor fusion, kernel tuning, and others on NVIDIA GPUs. https://developer.nvidia.com/tensorrt-getting-started
- **NVIDIA TensorRT-LLM** is an open-source library that accelerates and optimizes inference performance of the latest large language models (LLMs). TensorRT-LLM wraps TensorRT's deep learning compiler and includes optimized kernels from FasterTransformer, pre- and post-processing, and multi-GPU and multi-node communication. https://developer.nvidia.com/tensorrt
- **NVIDIA Triton Inference Server** optimizes the deployment of AI models at scale and in production for both neural networks and tree-based models on GPUs.

It also provides full access to the NVIDIA NGC catalogue, a collection of tested enterprise software, services and tools supporting end-to-end AI and digital twin workflows and can be integrated with MLOps platforms such as ClearML, Domino Data Lab, Run:ai, UbiOps, and Weights & Biases.

Finally, NVIDIA AI Enterprise introduced NVIDIA Inference Microservices (NIM), a set of performance-optimized, portable microservices designed to accelerate and simplify the deployment of AI models. Those containerized GPU-accelerated pretrained, fine-tuned, and customized models are ideally suited to be self-hosted and deployed on the NVIDIA GB300 NVL72.

Security

Topics in this section:

- Security features

Security features

The server offers the following electronic security features:

- Support for Platform Firmware Resiliency (PFR) hardware Root of Trust (RoT)
- Firmware signature processes compliant with FIPS and NIST requirements
- Administrator and power-on password
- Integrated Trusted Platform Module (TPM) supporting TPM 2.0

The server is NIST SP 800-147B compliant.

The following table lists the security options for the NVIDIA GB300 NVL72.

Operating system support

The NVIDIA GB300 NVL72 supports the following operating systems:

- Ubuntu 22.04 with NVIDIA optimized HWE kernel (kopt)

Physical and electrical specifications

The Lenovo 48U MGX Rack has the following dimensions:

- Width: 600 mm (23.6 inches)
- Height: 2,294 mm (90.3 inches)
- Depth: 1,068 mm (42.05 inches) - Expands to 1,200 mm with extension frame
- Weight: 185 kg (407.8 lbs)

GB300 Compute trays are installed onto the rack. Each GB300 Compute tray has the following dimensions:

- Width: 438 mm (17.3 inches)
- Height: 43.6 mm (1.72 inches)
- Depth: 766 mm (29.9 inches) (799 mm, including the water connections at the rear of the server)
- Weight: 29 kg (63.93 lbs)

The GB300 NVL Switch Tray has the following overall physical dimensions, excluding components that extend outside the standard chassis, such as EIA flanges, front security bezel (if any), and power conversion station handles:

- Width: 438 mm (17.3 inches)
- Height: 43.6 mm (1.72 inches)
- Depth: 770 mm (30.3 inches)
- Weight: 17 kg (37.5 lbs)

The NVIDIA SN2201 48 port DC switch has the following dimensions:

- Width: 438 mm (17.3 inches)
- Height: 43.6 mm (1.72 inches)
- Depth: 781 mm (28.3 inches)
- Weight: 11.39 kg (25.1 lbs)

The Power Shelf has the following dimensions:

- Width: 448 mm (17.6 inches)
- Height: 43.6 mm (1.72 inches)
- Depth: 718.5 mm (28.3 inches)

The following table lists the detailed dimensions. See the figure below for the definition of each dimension.

Table 25. Detailed dimensions

Dimension	Description
483 mm	X_a = Width, to the outsides of the front EIA flanges
448 mm	X_b = Width, to the rack rail mating surfaces
540 mm	X_c = Width, to the outer most chassis body feature
572 mm	Y_a = Height, from the bottom of chassis to the top of the chassis
1062 mm	Z_a = Depth, from the rack flange mating surface to the rearmost I/O port surface
1076 mm	Z_b = Depth, from the rack flange mating surface to the rearmost feature of the chassis body
1114 mm	Z_c = Depth, from the rack flange mating surface to the rearmost feature such as power supply handle
226 mm	Z_d = Depth, from the forwardmost feature on front of EIA flange to the rack flange mating surface
226 mm	Z_e = Depth, from the front of security bezel (if applicable) or forwardmost feature to the rack flange mating surface

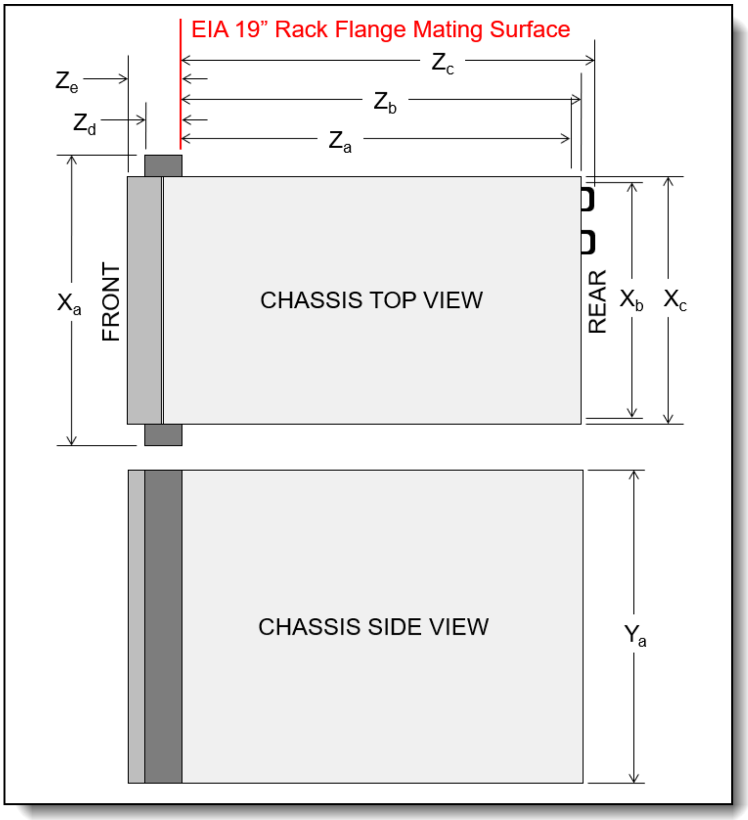

Figure 15. Enclosure dimensions

Operating environment

The GB300 Compute tray or server trays and GB300 NVL Switch Trays are supported in the following environment.

Topics in this section:

- Air temperature and humidity
- Water requirements
- Particulate contamination

Air temperature and humidity

Air temperature/humidity requirements:

- ASHRAE standard A2
  - Temperature operating:
    - Meets ASHRAE standard A2 (10°C to 35°C)
    - Maximum rate of change (°C /hr) ≤ 20
    - Operating temperature de-rated 1C per 300m (10,00ft) to 1,371m (4,500ft) above sea level
  - Humidity operating:
    - 20% - 80%, non-condensing
    - Humidity transition rate < 10%/hr
- Acoustics idle/working mode:
  - Fan control from BMC with multiple temperature input
  - Sound pressure level (LpAm): up to 82.9 dB
  - Estimated that the rack will operate around 92dBA from rack rear, typical ambient and rack exercised
- Altitude:
  - Supports 0 – 5,000 ft (0 - 1,524m)
- Temperature Non-operating: w/o PKG
  - -40C to 70C (Rack shipped with pressurized nitrogen)
  - Non-operating with Di-H20: 5-70C
  - Non-operating with PG25: 10-70C

Water requirements

For details on the NVIDIA GB300 NVL72 direct water-cooling standards, see the following table:

Table 26. Temperature, LPM and psi values

Fluid Supply Temperature	Required Rack Flow	Rack Pressure Drop
25°C	59 LPM	2.3 psi
30°C	71 LPM	3.2 psi
35°C	89 LPM	4.9 psi
40°C	119 LPM	8.5 psi
45°C	177 LPM	18.4 psi

Standard specifications of the Lenovo GB300 NVL Switch Tray.

- Operating water temperature: 2°C to 50°C (35.6°F to 122°F) (ASHRAE W45 compliant)

Particulate contamination

Airborne particulates (including metal flakes or particles) and reactive gases acting alone or in combination with other environmental factors such as humidity or temperature might damage the system that might cause the system to malfunction or stop working altogether.

The following specifications indicate the limits of particulates that the system can tolerate:

- Reactive gases:
  - The copper reactivity level shall be less than 200 Angstroms per month (Å/month)
  - The silver reactivity level shall be less than 200 Å/month

- Airborne particulates:
  - The room air should be continuously filtered with MERV 8 filters.
  - Air entering a data center should be filtered with MERV 11 or preferably MERV 13 filters.
  - The deliquescent relative humidity of the particulate contamination should be more than 60% RH
  - Environment must be free of zinc whiskers

For additional information, see the Specifications section of the documentation for the server, available from the Lenovo Documents site, https://pubs.lenovo.com/

Regulatory compliance

The NVIDIA GB300 NVL72 conforms to the following standards:

- ANSI/UL 62368-1
- IEC 62368-1 (CB Certificate and CB Test Report)
- CSA C22.2 No. 62368-1
- Mexico NOM-019
- CE, UKCA Mark (EN55032 Class A, EN62368-1, EN55035, EN61000-3-11, EN61000-3-12, (EU) 2019/424, and EN IEC 63000 (RoHS))
- FCC - Verified to comply with Part 15 of the FCC Rules, Class A
- Canada ICES-003, issue 7, Class A
- CISPR 32, Class A, CISPR 35
- Japan VCCI, Class A
- Taiwan BSMI CNS15936, Class A; CNS15598-1; Section 5 of CNS15663
- Australia/New Zealand AS/NZS CISPR 32, Class A; AS/NZS 62368.1
- SGS, VOC Emission
- Japanese Energy-Saving Act
- EU2019/424 Energy Related Product (ErP Lot9)

Warranty upgrades and post-warranty support

The server and enclosure have the following warranty:

- Lenovo GB300 Compute tray (7DLZ) - 3-year warranty
- Lenovo GB300 NVL Switch Tray (7DJY) - 3-year warranty
- SN2201 Mgmt Switch (7DJW) - 3-year warranty
- Power Shelf, 1U, 33kW (C5RR) - 3-year warranty

Our global network of regional support centers offers consistent, local-language support enabling you to vary response times and level of service* to match the criticality of your support needs:

- **Standard Next Business Day** – Best choice for non-essential systems requiring simple maintenance.
- **Premier Next Business Day** – Best choice for essential systems requiring technical expertise from senior-level Lenovo engineers.
- **Premier 24x7 4-Hour Response** – Best choice for systems where maximum uptime is critical.
- **Premier Enhanced Storage Support 24x7 4-Hour Response** – Best choice for storage systems where maximum uptime is critical.

For more information, consult the brochure Lenovo Operational Support Services for Data Centers Services.

* Some service levels may not be available in all markets. Contact your sales representative for more information.

Services

Lenovo Data Center Services empower you at every stage of your IT lifecycle. From expert advisory and strategic planning to seamless deployment and ongoing support, we ensure your infrastructure is built for success. Our comprehensive services accelerate time to value, minimize downtime, and free your IT staff to focus on driving innovation and business growth.

> **Note**: Some service options may not be available in all markets or regions. For more information, go to https://lenovolocator.com/. For information about Lenovo service upgrade offerings that are available in your region, contact your local Lenovo sales representative or business partner.

In this section:

Lenovo Advisory Services

Lenovo Advisory Services simplify the planning process, enabling customers to build future-proofed strategies in as little as six weeks. Consultants provide guidance on projects including VM migration, storage, backup and recovery, and cost management to accelerate time to value, improve cost efficiency, and build a flexibly scalable foundation.

- **Assessment Services**

  An Assessment helps solve your IT challenges through an onsite, multi-day session with a Lenovo technology expert. We perform a tools-based assessment which provides a comprehensive and thorough review of a company's environment and technology systems. In addition to the technology based functional requirements, the consultant also discusses and records the non-functional business requirements, challenges, and constraints. Assessments help organizations like yours, no matter how large or small, get a better return on your IT investment and overcome challenges in the ever-changing technology landscape.

- **Design Services**

  Professional Services consultants perform infrastructure design and implementation planning to support your strategy. The high-level architectures provided by the assessment service are turned into low level designs and wiring diagrams, which are reviewed and approved prior to implementation. The implementation plan will demonstrate an outcome-based proposal to provide business capabilities through infrastructure with a risk-mitigated project plan.

Lenovo Plan & Design Services

Unlock faster time to market with our tailored, strategic design workshops to align solution approaches with your business goals and technical requirements. Leverage our deep solution expertise and end-to-end delivery partnership to meet your goals efficiently and effectively.

Lenovo Deployment, Migration, and Configuration Services

Optimize your IT operations by shifting labor-intensive functions to Lenovo's skilled technicians for seamless on-site or remote deployment, configuration, and migration. Enjoy peace of mind, faster time to value, and comprehensive knowledge sharing with your IT staff, backed by our best-practice methodology.

- **Deployment Services for Storage and ThinkAgile**

  A comprehensive range of remote and onsite options tailored specifically for your business needs to ensure your storage and ThinkAgile hardware are fully operational from the start.

- **Hardware Installation Services**

  A full-range, comprehensive setup for your hardware, including unpacking, inspecting, and positioning components to ensure your equipment is operational and error-free for the most seamless and efficient installation experience, so you can quickly benefit from your investments.

- **DM/DG File Migration Services**

  Take the burden of file migration from your IT's shoulders. Our experts will align your requirements and business objectives to the migration plans while coordinating with your team to plan and safely execute the data migration to your storage platforms.

- **DM/DG/DE Health Check Services**

  Our experts perform proactive checks of your Firmware and system health to ensure your machines are operating at peak and optimal efficiency to maximize up-time, avoid system failures, ensure the security of IT solutions and simplify maintenance.

- **Factory Integrated Services**

  A suite of value-added offerings provided during the manufacturing phase of a server or storage system that reduces time to value. These services aim at improving your hardware deployment experience and enhance the quality of a standard configuration before it arrives at your facility.

Lenovo Support Services

In addition to response time options for hardware parts, repairs, and labor, Lenovo offers a wide array of additional support services to ensure your business is positioned for success and longevity. Our goal is to reduce your capital outlays, mitigate your IT risks, and accelerate your time to productivity.

- **Premier Support for Data Centers**

  Your direct line to the solution that promises the best, most comprehensive level of support to help you fully unlock the potential of your data center.

- **Premier Enhanced Storage Support (PESS)**

  Gain all the benefits of Premier Support for Data Centers, adding dedicated storage specialists and resources to elevate your storage support experience to the next level.

- **Committed Service Repair (CSR)**

  Our commitment to ensuring the fastest, most seamless resolution times for mission-critical systems that require immediate attention to ensure minimal downtime and risk for your business. This service is only available for machines under the Premier 4-Hour Response SLA.

- **Multivendor Support Services (MVS)**

  Your single point of accountability for resolution support across vast range of leading Server, Storage, and Networking OEMs, allowing you to manage all your supported infrastructure devices seamlessly from a single source.

- **Keep Your Drive (KYD)**

  Protect sensitive data and maintain compliance with corporate retention and disposal policies to ensure your data is always under your control, regardless of the number of drives that are installed in your Lenovo server.

- **Technical Account Manager (TAM)**

  Your single point of contact to expedite service requests, provide status updates, and furnish reports to track incidents over time, ensuring smooth operations and optimized performance as your business grows.

- **Enterprise Software Support (ESS)**

  Gain comprehensive, single-source, and global support for a wide range of server operating systems and Microsoft server applications.

For more information, consult the brochure Lenovo Operational Support Services for Data Centers.

Lenovo Managed Services

Achieve peak efficiency, high security, and minimal disruption with Lenovo's always-on Managed Services. Our real-time monitoring, 24x7 incident response, and problem resolution ensure your infrastructure operates seamlessly. With quarterly health checks for ongoing optimization and innovation, Lenovo's remote active monitoring boosts end-user experience and productivity by keeping your data center's hardware performing at its best.

Lenovo Managed Services provides continuous 24x7 remote monitoring (plus 24x7 call center availability) and proactive management of your data center using state-of-the-art tools, systems, and practices by a team of highly skilled and experienced Lenovo services professionals.

Quarterly reviews check error logs, verify firmware & OS device driver levels, and software as needed. We’ll also maintain records of latest patches, critical updates, and firmware levels, to ensure you systems are providing business value through optimized performance.

Lenovo Sustainability Services

- **Asset Recovery Services**

  Lenovo Asset Recovery Services (ARS) provides a secure, seamless solution for managing end-of-life IT assets, ensuring data is safely sanitized while contributing to a more circular IT lifecycle. By maximizing the reuse or responsible recycling of devices, ARS helps businesses meet sustainability goals while recovering potential value from their retired equipment. For more information, see the Asset Recovery Services offering page.

- **CO2 Offset Services**

  Lenovo’s CO2 Offset Services offer a simple and transparent way for businesses to take tangible action on their IT footprint. By integrating CO2 offsets directly into device purchases, customers can easily support verified climate projects and track their contributions, making meaningful progress toward their sustainability goals without added complexity.

- **Lenovo Certified Refurbished**

  Lenovo Certified Refurbished offers a cost-effective way to support IT circularity without compromising on quality and performance. Each device undergoes rigorous testing and certification, ensuring reliable performance and extending its lifecycle. With Lenovo’s trusted certification, you gain peace of mind while making a more sustainable IT choice.

- **Data Center Power and Cooling Services**

  The Data Center Infrastructure team will provide solution design and implementation services to support the power and cooling needs of the multi-node chassis and multi-rack solutions. This includes designing for various levels of power redundancy and integration into the customer power infrastructure. The Infrastructure team will work with site engineers to design an effective cooling strategy based on facility constraints or customer goals and optimize a cooling solution to ensure high efficiency and availability. The Infrastructure team will provide the detailed solution design and complete integration of the cooling solution into the customer data center. In addition, the Infrastructure team will provide rack and chassis level commissioning and stand-up of the water-cooled solution which includes setting and tuning of the flow rates based on water temperature and heat recovery targets. Lastly, the Infrastructure team will provide cooling solution optimization and performance validation to ensure the highest overall operational efficiency of the solution.

Additional Lenovo Essential Equipment Service

For safety reasons, it is highly recommended to use the Genie® Lift™ GL™-8 due to the server's weight. Lenovo offers the lift with add on fixtures, see Genie Material Lift for more information. If no lift is available onsite, customers must move the machine to an accessible, powered area before the technician arrives and handle reinstallation.

Lenovo strongly advises configuring a complete solution for GB300 and SC-Systems with the lift tool through options.

Essential Equipment that must be available at the customer site ahead of each Service event.

Note: All components listed below must be provided by the customer

- Service Lift: PN: 4XF7B02087
- Air compressor:
  - Max Pressure: 150 PSI
  - Air Flow: 1.41 SCFM at 0 PSI
  - Accuracy: +/- 1 PSI with auto-pressure check
  - Hose: 36-inch hose with all-brass chuck
  - Smart Features: Auto-shutoff at target PSI and 4 programmable memory presets
  - System: Cordless - Battery operated - with charger and battery
- Fill charger:
  - Flow Rate: 480 Gallons Per Hour (GPH)
  - Connections: 3/4" Brass (Standard garden hose threads)
  - Lift/Head: 18 ft. max suction lift; 75 ft. max head height
  - Priming: Self-priming (No manual priming required)
  - Protection: Auto-shutoff when water stops flowing (prevents motor damage)
  - System: Cordless - Battery operated - with charger and battery
- 2x 30L Labeled tanks for new and discarded liquid coolant
- Spill kit
- PG25 or NALCO® TCS101 or HPCCL2000 (DI water premixed with antimicrobial agent)
- Service kit (FRU PN: 03PV563/L2 PN: STL7C32889)
  - For Self-maintainer: see Service kit tablet for more information.

Notes:

Onsite spares (compute nodes) would be needed to support an onsite Next Business Day Service Level target. 2% of the order recommended quantity. For Compute Tray base alternatives available for selection, see the Spare Compute Trays section.

Rack cabinets

The GB300 Compute trays, GB300 NVL Switch Trays, SN2201 Management Switches and Power Shelves are supported in the following racks:

- 7DJVCTO2WW Lenovo 48U MGX Rack

Genie Material Lift

Considering the weight of the trays in the server, an onsite material lift is required to allow service by a single person. If you do not already have a material lift available, Lenovo offers the Genie Lift GL-8 material lift as configurable option to the rack cabinets. Ordering information is listed in the following table.

Note: If neither the Genie Lift GL-8 nor the ServerLift SL-350x is available onsite when onsite service is required, the customer will be responsible for getting the system to a suitable work surface (with access to power) prior to service technician arrival and returning the system to the rack when service is complete prior to service technician departure.

Table 27. Genie Lift GL-8 ordering information

Part number	Description
4XF7B02087	Genie Lift GL-8 (Standard Base) Material Lift Option Kit • Genie GL-8 material lift • Load platform • Foot-release brake

Lenovo Financial Services

Why wait to obtain the technology you need now? No payments for 90 days and predictable, low monthly payments make it easy to budget for your Lenovo solution.

- **Flexible**

  Our in-depth knowledge of the products, services and various market segments allows us to offer greater flexibility in structures, documentation and end of lease options.

- **100% Solution Financing**

  Financing your entire solution including hardware, software, and services, ensures more predictability in your project planning with fixed, manageable payments and low monthly payments.

- **Device as a Service (DaaS)**

  Leverage latest technology to advance your business. Customized solutions aligned to your needs. Flexibility to add equipment to support growth. Protect your technology with Lenovo's Premier Support service.

- **24/7 Asset management**

  Manage your financed solutions with electronic access to your lease documents, payment histories, invoices and asset information.

- **Fair Market Value (FMV) and $1 Purchase Option Leases**

  Maximize your purchasing power with our lowest cost option. An FMV lease offers lower monthly payments than loans or lease-to-own financing. Think of an FMV lease as a rental. You have the flexibility at the end of the lease term to return the equipment, continue leasing it, or purchase it for the fair market value. In a $1 Out Purchase Option lease, you own the equipment. It is a good option when you are confident you will use the equipment for an extended period beyond the finance term. Both lease types have merits depending on your needs. We can help you determine which option will best meet your technological and budgetary goals.

Ask your Lenovo Financial Services representative about this promotion and how to submit a credit application. For the majority of credit applicants, we have enough information to deliver an instant decision and send a notification within minutes.

Seller training courses

The following sales training courses are offered for employees and partners (login required). Courses are listed in date order.

1. **FY26Q4 Solutions Launch GTC Launch Quick Hit**
   2026-03-12 | 10 minutes | Employees and Partners

   This Quick Hit covers the Lenovo Hybrid AI Advantage - from personal AI to gigawatt scale AI factories. Lenovo is uniquely positioned to meet your customers' needs in all AI environments, no matter how small or large. The combination of hardware, software, services, and solutions is unmatched in the industry.

   Published: 2026-03-12
   Length: 10 minutes

   Start the training:
   Employee link: Grow@Lenovo
   Partner link: Lenovo 360 Learning Center

   Course code: DAIQ101

2. **HPC VTT: Unlocking Hybrid HPC + AI_ Slinky Bridge for Unified GPU Workloads**
   2026-03-10 | 62 minutes | Employees Only

   View this session to hear from our speakers, Aurelien Ortiz, Software Architect
   HPC & AI, Lenovo and Nick Ihli, Sr Product Manager System Software – Slurm, SchedMD, as they explain how Slinky helps with unified GPU workloads. Topics include:
   Why Slinky? Bridging HPC and Cloud-Native AI Workloads
   How Slinky Bridge Works: Architecture, Components, and Flow
   Deployment Requirements & How to Position Slinky

   Published: 2026-03-10
   Length: 62 minutes

   Start the training:
   Employee link: Grow@Lenovo

   Course code: DVHPC231

3. **Edge VTT - NVIDIA Robotics Platform**
   2026-01-08 | 67 minutes | Employees Only

   In this session we feature speakers from both NVIDIA and Lenovo. Attendees will learn about NVIDIA's platform stack for Robotics and what Lenovo is doing in the field of robotics.

   During this session we will dive into NVIDIA's three-computer stack for Physical AI. Our speaker will explore libraries and workflows to develop, train, simulate, deploy, operate, and optimize AI robot systems and software. This session will cover the basics of the technical platform, how to get started and case studies from some NVIDIA's ecosystem.

   Objectives:
   Discuss acceleration libraries
   Describe simulation workflows
   List foundational models for robotics

   Tags: Artificial Intelligence (AI), Sales, Software Platforms, Technical Sales

   Published: 2026-01-08
   Length: 67 minutes

   Start the training:
   Employee link: Grow@Lenovo

   Course code: DVEDG223

4. **Lenovo Unlocks the Power of NIM: Overview, Industry Use Cases and Lenovo Services**
   2025-10-16 | 55 minutes | Employees and Partners

   Join us for an insightful session with our Lenovo speakers, Farah Toos and Dinesh Tripathi where we'll explore the transformative potential of NVIDIA Inference Microservices (NIM). This webinar will provide a comprehensive overview of NIMs, highlighting how it streamlines operations, enhances scalability, and drives innovation across industries.

   Discover real-world use cases in sectors such as healthcare, manufacturing, retail, and finance, and learn how Lenovo's portfolio of services—including deployment, optimization, and lifecycle support—can help your customers maximize the value of their infrastructure investments.
   Whether you're engaging with enterprise clients or mid-market opportunities, this session will equip you with the knowledge and tools to position Lenovo's NIM solutions effectively and drive impactful conversations.

   Key Takeaways:
   - Understand the core capabilities and benefits of NIM
   - Explore industry-specific applications and success stories
   - Learn how Lenovo services complement and enhance NIM deployments
   - Gain selling strategies and resources to support customer engagements

   Tags: Artificial Intelligence (AI), NVIDIA,Services,Technical Sales, ThinkAgile, ThinkSystem

   Published: 2025-10-16
   Length: 55 minutes

   Start the training:
   Employee link: Grow@Lenovo
   Partner link: Lenovo 360 Learning Center

   Course code: DVCLD228

5. **Partner Technical Webinar - NVIDIA Software**
   2025-07-21 | 60 minutes | Employees and Partners

   In this 60-minute replay, Carlos Huescas, Lenovo, and Sandeep Brahmarouthu and Rob Magno of NVIDIA, presented the key software offerings of NVDIA AI Enterprise (NVAIE) and Run:ai, including a demo of Run:ai.

   Tags: Artificial Intelligence (AI)

   Published: 2025-07-21
   Length: 60 minutes

   Start the training:
   Employee link: Grow@Lenovo
   Partner link: Lenovo 360 Learning Center

   Course code: JUL1825
6. **Partner Technical Webinar - AI Vertical Spotlight Pt 2**
   2025-07-08 | 60 minutes | Employees and Partners

   In this 60-minute replay, we concluded the AI Vertical Spotlight (Pt 2) with our final two speakers. Peter Orban, AI Business Development Manager, discussed Financial and Banking, while Eric Skomra, Public Sector & Spaces AI Technologist, provided insights on State, Local, Education (SLED), and Smart Spaces.

   Tags: Artificial Intelligence (AI)

   Published: 2025-07-08
   Length: 60 minutes

   Start the training:
   Employee link: Grow@Lenovo
   Partner link: Lenovo 360 Learning Center

   Course code: JUN2725

7. **AI VTT: NVIDIA Run:ai**
   2025-07-02 | 75 minutes | Employees Only

   NVIDIA Run:ai is a GPU orchestration and optimization platform designed to help organizations maximize their GPU compute resources for AI workloads. It accelerates AI development, reduces costs, and improves AI development cycles by enabling dynamic allocation and scheduling of GPU resources, as well as workload submission and sharing. Essentially, it provides a centralized interface to manage AI compute infrastructure, making it easier for AI teams to access and utilize GPUs effectively.

   Join Carlos Huescas from Lenovo, Sandeep Brahmarouthu and Robert Magno from NVIDIA as they discuss NVIDIA Run:ai. Topics include:
   •What is Run:ai and its capabilities?
   •Customer segmentation for Run:ai
   •How to order, part numbers and licensing
   •Demo of Run:ai

   Tags: Artificial Intelligence (AI), NVIDIA

   Published: 2025-07-02
   Length: 75 minutes

   Start the training:
   Employee link: Grow@Lenovo

   Course code: DVAI218
8. **Partner Technical Webinar - Enterprise AI Team Intro and Vertical Spotlight Pt1**
   2025-06-17 | 60 minutes | Employees and Partners

   In this 60-minute replay, John Encizo introduced his new Enterprise AI Team. Part 1 covered three verticals: Retail with Allen Holmes, Manufacturing with Jason Hamp, and Healthcare with Janna Templin.

   Tags: Artificial Intelligence (AI)

   Published: 2025-06-17
   Length: 60 minutes

   Start the training:
   Employee link: Grow@Lenovo
   Partner link: Lenovo 360 Learning Center

   Course code: JUN1325

9. **VTT Edge: Understanding Visual AI Agents with NVIDIA June 2025**
   2025-06-16 | 60 minutes | Employees and Partners

   Join our guest speakers from NVIDIA as they discuss what's behind the scenes of visual AI Agents for Smart Cities, Smart Spaces and Manufacturing. Explore the modular approach to building a workforce of AI Agents. Topics include:

   • Sensors which feed the AI Agents
   • How AI agents improve safety and prevent accidents in Smart Spaces
   • Demo: Modular development of AI Agents

   Tags: Artificial Intelligence (AI), Technical Sales, NVIDIA

   Published: 2025-06-16
   Length: 60 minutes

   Start the training:
   Employee link: Grow@Lenovo
   Partner link: Lenovo 360 Learning Center

   Course code: DVEDG221

10. **Lenovo Cloud Architecture VTT: Supercharge Your Enterprise AI with NVIDIA AI Enterprise on Lenovo Hybrid AI Platform**
    2025-04-17 | 75 minutes | Employees and Partners

    Join us for an in-depth webinar with Justin King, Principal Product Marketing Manager for Enterprise AI exploring the power of NVIDIA AI Enterprise, delivering Generative and Agentic AI outcomes deployed with Lenovo Hybrid AI platform environments.
    In today's data-driven landscape, AI is evolving at high speed, with new techniques delivering more accurate responses. Enterprises are seeking not just an understanding but also how they can achieve AI-driven business outcomes.
    With this, the demand for secure, scalable, and high-performing AI operations-and the skills to deliver them-is top of mind for many. Learn how NVIDIA AI Enterprise, a comprehensive software suite optimized for NVIDIA GPUs, provides the tools and frameworks, including NVIDIA NIM, NeMo, and Blueprints, to accelerate AI development and deployment while reducing risk-all within the control and security of your Lenovo customer's hybrid AI environment.

    Tags: Artificial Intelligence (AI), Cloud, Data Management, Nvidia, Technical Sales

    Published: 2025-04-17
    Length: 75 minutes

    Start the training:
    Employee link: Grow@Lenovo
    Partner link: Lenovo 360 Learning Center

    Course code: DVCLD221

11. **AI VTT: GTC Update and The Lenovo LLM Sizing Guide**
    2025-03-12 | 86 minutes | Employees Only

    Please view this session that is two parts. Part one is Robert Daigle, Director, Global AI Solutions and Hande Sahin-Bahceci, AI Solutions Marketing Leader explaining the upcoming announcements for NVIDIA GTC. Part Two is Sachin Wani, AI Data Scientist explaining the Lenovo LLM Sizing Guide with these topics:

    • Minimum GPU requirements for fine-tuning/training and inference
    • Gathering requirements for the customer's use case
    • LLMs from a technical perspective

    Tags: Artificial Intelligence (AI), Technical Sales

    Published: 2025-03-12
    Length: 86 minutes

    Start the training:
    Employee link: Grow@Lenovo

    Course code: DVAI214

12. **VTT AI: Components of the AI Stack and Where Lenovo Sits November 2024**
    2024-11-26 | 75 minutes | Employees Only

    Join Per Ljungstrom, Lenovo Principal TC EMEA, as he explores AI concepts where innovations meet simplified predefined solutions which deploy at scale. Topics for this session include:
    • Associating software with the ground level of hardware
    • Attach NVIDIA AI Enterprise, Microsoft, Tiber AI Stacks and more
    • AI at the Edge and the complete solution
    • What to consider when talking AI Stack with your customer

    Tags: Artificial Intelligence (AI), Cloud, Technical Sales, Technology solutions, ThinkEdge

    Published: 2024-11-26
    Length: 75 minutes

    Start the training:
    Employee link: Grow@Lenovo

    Course code: DVAI210

13. **VTT AI: NVIDIA OVX**
    2024-10-23 | 55 minutes | Employees and Partners

    Please join this session as Steven Puzio, Global Sales Leader of NVIDIA Omniverse speaks to us about these topics:

    • OVX use cases
    • Target customers
    • OVX reference architectures
    • Parts, pieces and technical details

    Tags: Artificial Intelligence (AI), Nvidia

    Published: 2024-10-23
    Length: 55 minutes

    Start the training:
    Employee link: Grow@Lenovo
    Partner link: Lenovo 360 Learning Center

    Course code: DVAI209

14. **Think AI Weekly: Ride the NVIDIA Wave for AI**
    2024-10-07 | 60 minutes | Employees Only

    In this session, a panel including speakers from NVIDIA, Lenovo IDG and Lenovo ISG address the topics:
    •Leveraging AI workstations to start an AI journey
    •Leading an ISG sale with NVIDIA AI Enterprise
    •NVIDIA sales tools available for Lenovo sellers
    •NVIDIA training on grow@lenovo and more

    Tags: Artificial Intelligence (AI), Nvidia

    Published: 2024-10-07
    Length: 60 minutes

    Start the training:
    Employee link: Grow@Lenovo

    Course code: DTAIW121

15. **Lenovo VTT Cloud Architecture - Unlock Gen AI with VMware Private AI Foundation with NVIDIA**
    2024-07-16 | 60 minutes | Employees and Partners

    In today's rapidly evolving digital landscape, businesses are hungry for the transformative power of Artificial Intelligence (AI). They see AI as the key to streamlining operations and unlocking exciting new opportunities. However, widespread adoption has been hampered by concerns surrounding privacy, the complexity of implementation, and the hefty costs associated with deploying and managing AI solutions at an enterprise level.
    Join Chris Gully and Baker Hull, Solutions Architects from VMware by Broadcom, as they discuss how Lenovo, NVIDIA, and VMware By Broadcom are partnering to deliver a private, secure, scalable, and flexible AI infrastructure solution that helps enterprise customers build and deploy AI workloads within their own private cloud infrastructure, ensure the control of sensitive data and compliance with regulatory requirements, ultimately driving faster time to value and achieving their AI objectives.

    Tags: Artificial Intelligence (AI), Cloud, Nvidia, ThinkAgile, VMware

    Published: 2024-07-16
    Length: 60 minutes

    Start the training:
    Employee link: Grow@Lenovo
    Partner link: Lenovo 360 Learning Center

    Course code: DVCLD214

16. **Guidance for Selling NVIDIA Products at Lenovo for ISG**
    2024-07-01 | 25 minutes | Employees and Partners

    This course gives key talking points about the Lenovo and NVIDIA partnership in the Data Center. Details are included on where to find the products that are included in the partnership and what to do if NVIDIA products are needed that are not included in the partnership. Contact information is included if help is needed in choosing which product is best for your customer. At the end of this session sellers should be able to explain the Lenovo and NVIDIA partnership, describe the products Lenovo can sell through the partnership with NVIDIA, help a customer purchase other NVIDIA product, and get assistance with choosing NVIDIA products to fit customer needs.

    Tags: Artificial Intelligence (AI), Nvidia

    Published: 2024-07-01
    Length: 25 minutes

    Start the training:
    Employee link: Grow@Lenovo
    Partner link: Lenovo 360 Learning Center

    Course code: DNVIS102

17. **NVIDIA AI Solutions and Market Trends**
2023-10-12 | 55 minutes | Employees Only

The purpose of this course is to help the learner recognized AI Market and trends. Also, explain NVIDIA's Computing platform, and discuss its importance for the market.

Course Objectives:
Recognize AI Trends
Explain NVIDIA Computing Platform
Discuss Industry Verticals Marketing

Tags: Artificial Intelligence (AI), Nvidia, Sales

Published: 2023-10-12
Length: 55 minutes

Start the training:
Employee link: Grow@Lenovo

Course code: DAINVD101

Related publications and links

For more information, see these resources:

- Lenovo NVIDIA GB300 NVL72 server product page
https://www.lenovo.com/us/en/servers-storage/?IPromoID=LEN781264
- Lenovo NVIDIA GB300 NVL72 datasheet
- https://lenovopress.lenovo.com/datasheet/DS0207
- Interactive 3D Tour of the Lenovo NVIDIA GB300 NVL72:
- https://lenovopress.lenovo.com/datasheet/LP2381
- Lenovo NVIDIA GB300 NVL72 drivers and support
https://datacentersupport.lenovo.com/us/en/products/solutions-and-software/aiinfrastructure/gb300nvl72/7djv/downloads/driver-list/
- ServerProven hardware compatibility:
https://serverproven.lenovo.com/
- Data Center Solution Configurator (DCSC)
https://dcsc.lenovo.com
- Lenovo Cluster solutions configurator (x-config)
https://lesc.lenovo.com/products/hardware/configurator/worldwide/bhui/asit/index.html

Related product families

Product families related to this document are the following:

- AI Servers
- Artificial Intelligence
- Rack Scale AI

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP2357, was created or updated on March 16, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2357
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2357.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
from Exascale to Everyscale®
Neptune®
ServerProven®
ThinkAgile®
ThinkSystem®

The following terms are trademarks of other companies:

AMD is a trademark of Advanced Micro Devices, Inc.

Intel®, the Intel logo and Intel Core® are trademarks of Intel Corporation or its subsidiaries.

Microsoft® and Windows® are trademarks of Microsoft Corporation in the United States, other countries, or both.

IBM® is a trademark of IBM in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.