ThinkSystem NVIDIA H100 PCIe Gen5 GPUs
Product Guide

The ThinkSystem NVIDIA H100 PCIe Gen5 GPU delivers unprecedented performance, scalability, and security for every workload. The GPUs use breakthrough innovations in the NVIDIA Hopper™ architecture to deliver industry-leading conversational AI, speeding up large language models by 30X over the previous generation.

The NVIDIA H100 GPU features fourth-generation Tensor Cores and the Transformer Engine with FP8 precision, further extending NVIDIA's market-leading AI leadership with up to 9X faster training and an incredible 30X inference speedup on large language models. For high-performance computing (HPC) applications, The GPUs triple the floating-point operations per second (FLOPS) of FP64 and add dynamic programming (DPX) instructions to deliver up to 7X higher performance.

The following figure shows the ThinkSystem NVIDIA H100 PCIe Gen5 GPU in the double-width PCIe adapter form factor.

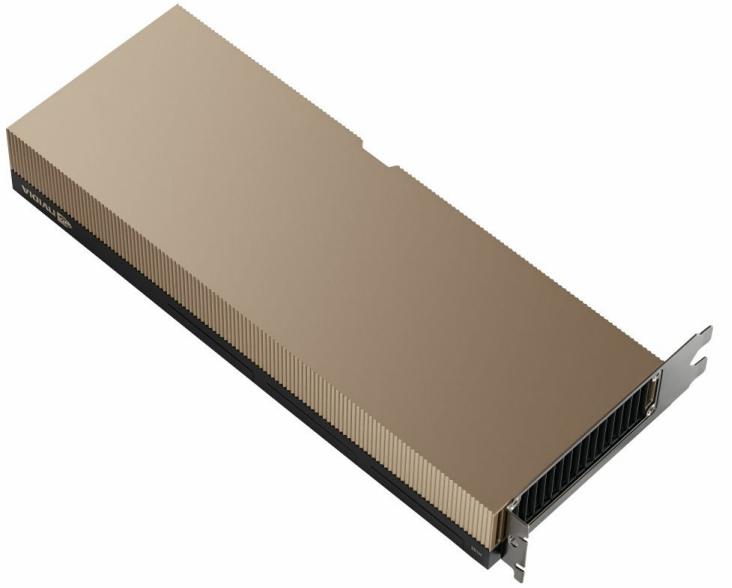


Figure 1. ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU

Did you know?

The NVIDIA H100 is available in both double-wide PCIe adapter form factor and in SXM form factor. The latter is used in Lenovo's Neptune direct-water-cooled ThinkSystem SD665-N V3 server for the ultimate in GPU performance and heat management.

The NVIDIA H100 NVL Tensor Core GPU is optimized for Large Language Model (LLM) Inferences, with its high compute density, high memory bandwidth, high energy efficiency, and unique NVLink architecture.

Part number information

The following table shows the part numbers for the ThinkSystem NVIDIA H100 PCIe Gen5 GPU.

Not available in China, Hong Kong and Macau : The H100 is not available in China, Hong Kong and Macau. For these markets, the H800 is avalable. See the NVIDIA H800 product guide for details, https://lenovopress.lenovo.com/LP1814

Table 1. Ordering information

Part number	Feature code	Description
Double-wide PCIe adapter form factor		
4X67A89325	BXAK	ThinkSystem NVIDIA H100 NVL 94GB PCIe Gen5 Passive GPU
4X67A82257	BR9U	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU
SXM form factor		
CTO only	BQQV	ThinkSystem NVIDIA H100 SXM5 700W 80G GPU Board
CTO only	BUBB	ThinkSystem NVIDIA H100 SXM5 700W 94G HBM2e GPU Board
NVLink bridge (for PCIe adapters only, not SXM)		
4X67A71309	BG3F	ThinkSystem NVIDIA Ampere NVLink 2-Slot Bridge (3 required per pair of GPUs)

The PCIe option part numbers includes the following:

- One GPU with full-height (3U) adapter bracket attached
- Documentation

The following figure shows the NVIDIA H100 SXM5 4-GPU Board installed in the ThinkSystem SD665-N V3 server

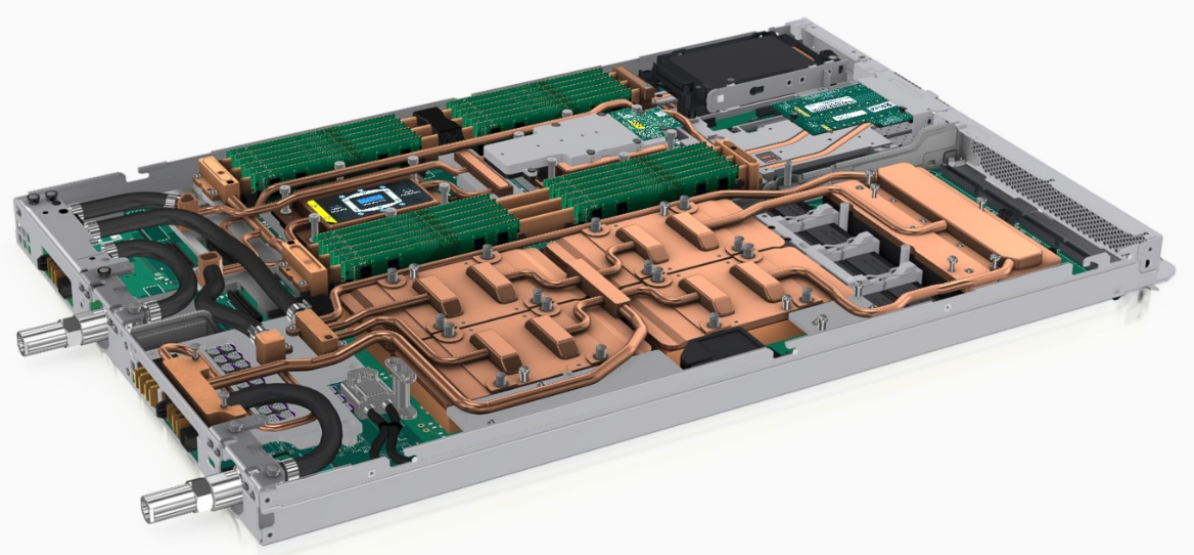


Figure 2. NVIDIA H100 SXM5 4-GPU Board in the ThinkSystem SD665-N V3 server

Features

The ThinkSystem NVIDIA H100 PCIe Gen5 GPU delivers high performance, scalability, and security for every workload. The GPU uses breakthrough innovations in the NVIDIA Hopper™ architecture to deliver industry-leading conversational AI, speeding up large language models (LLMs) by 30X over the previous generation.

The PCIe versions of the NVIDIA H100 GPUs include a five-year software subscription, with enterprise support, to the NVIDIA AI Enterprise software suite, simplifying AI adoption with the highest performance. This ensures organizations have access to the AI frameworks and tools they need to build accelerated AI workflows such as AI chatbots, recommendation engines, vision AI, and more.

The NVIDIA H100 GPU features fourth-generation Tensor Cores and the Transformer Engine with FP8 precision, further extending NVIDIA's market-leading AI leadership with up to 9X faster training and an incredible 30X inference speedup on large language models. For high-performance computing (HPC) applications, the GPU triples the floating-point operations per second (FLOPS) of FP64 and adds dynamic programming (DPX) instructions to deliver up to 7X higher performance. With second-generation Multi-Instance GPU (MIG), built-in NVIDIA confidential computing, and NVIDIA NVLink Switch System, the NVIDIA H100 GPU securely accelerates all workloads for every data center from enterprise to exascale.

Key features of the NVIDIA H100 GPU:

- NVIDIA H100 Tensor Core GPU

  Built with 80 billion transistors using a cutting-edge TSMC 4N process custom tailored for NVIDIA's accelerated compute needs, H100 is the world's most advanced chip ever built. It features major advances to accelerate AI, HPC, memory bandwidth, interconnect, and communication at data center scale.

- Transformer Engine

  The Transformer Engine uses software and Hopper Tensor Core technology designed to accelerate training for models built from the world's most important AI model building block, the transformer. Hopper Tensor Cores can apply mixed FP8 and FP16 precisions to dramatically accelerate AI calculations for transformers.

- NVLink Switch System

  The NVLink Switch System enables the scaling of multi-GPU input/output (IO) across multiple servers. The system delivers up to 9X higher bandwidth than InfiniBand HDR on the NVIDIA Ampere architecture.

- NVIDIA Confidential Computing

  NVIDIA Confidential Computing is a built-in security feature of Hopper that makes NVIDIA H100 the world's first accelerator with confidential computing capabilities. Users can protect the confidentiality and integrity of their data and applications in use while accessing the unsurpassed acceleration of H100 GPUs.

- Second-Generation Multi-Instance GPU (MIG)

  The Hopper architecture's second-generation MIG supports multi-tenant, multi-user configurations in virtualized environments, securely partitioning the GPU into isolated, right-size instances to maximize quality of service (QoS) for 7X more secured tenants.

- DPX Instructions

  Hopper's DPX instructions accelerate dynamic programming algorithms by 40X compared to CPUs and 7X compared to NVIDIA Ampere architecture GPUs. This leads to dramatically faster times in disease diagnosis, real-time routing optimizations, and graph analytics.

Technical specifications

The following table lists the GPU processing specifications and performance of the NVIDIA H100 GPU.

Table 2. Specifications of the NVIDIA H100 GPU

Feature	H100 NVL 94GB PCIe adapter	H100 80GB PCIe adapter	H100 80GB SXM board	H100 94GB SXM board
GPU Architecture	NVIDIA Hopper	NVIDIA Hopper	NVIDIA Hopper	NVIDIA Hopper
NVIDIA Tensor Cores	TBD	528 fourth-generation Tensor Cores per GPU	528 fourth-generation Tensor Cores per GPU	528 fourth-generation Tensor Cores per GPU
NVIDIA CUDA Cores (shading units)	TBD	18,432 FP32 CUDA Cores per GPU	18,432 FP32 CUDA Cores per GPU	18,432 FP32 CUDA Cores per GPU
Peak FP64 performance	34 TFLOPS	26 TFLOPS	34 TFLOPS	34 TFLOPS
Peak FP64 Tensor Core performance	67 TFLOPS	51 TFLOPS	67 TFLOPS	67 TFLOPS
Peak FP32 performance	67 TFLOPS	51 TFLOPS	67 TFLOPS	67 TFLOPS
Peak Tensor Float 32 (TF32) performance	990 TFLOPS*	756 TFLOPS*	989 TFLOPS*	989 TFLOPS*
Peak FP16 performance	1,980 TFOPS*	1,513 TFLOPS*	1,979 TFLOPS*	1,979 TFLOPS*
Peak Bfloat16 (BF16) performance	1,980 TFOPS*	1,513 TFLOPS*	1,979 TFLOPS*	1,979 TFLOPS*
Peak FP8 performance	3,960 TFOPS*	3,026 TFLOPS*		
INT8 Integer Performance	3,960 TOPS*	3,026 TOPS*	3,958 TOPS*	3,958 TOPS*
GPU Memory	94 GB HBM3	80 GB HBM2e	80GB board (feature BQQV): 80 GB HBM3	94GB board (feature BUBB): 90GB HBM2e
Memory Bandwidth	3.9 TB/s	2 TB/sec	80GB board (feature BQQV): 3.35 TB/sec	94GB board (feature BUBB): 2.4 TB/sec
ECC	Yes	Yes	Yes	Yes
Interconnect Bandwidth	NVLink: 600 GB/sec PCIe Gen5: 128 GB/sec	NVLink: 600 GB/sec PCIe Gen5: 128 GB/sec	NVLink: 900 GB/sec PCIe Gen5: 128 GB/sec	NVLink: 900 GB/sec PCIe Gen5: 128 GB/sec
System Interface	PCIe Gen 5.0, x16 lanes	PCIe Gen 5.0, x16 lanes	PCIe Gen 5.0, x16 lanes	PCIe Gen 5.0, x16 lanes
Form Factor	PCIe full height/length, double width	PCIe full height/length, double width	SXM5	SXM5
NVLink support	Yes; 3 NVLink Bridge supported per pair of GPUs (all 3 required)		Yes, integrated	
Multi-Instance GPU (MIG)	Up to 7 GPU instances, 12GB each	Up to 7 GPU instances, 10GB each	Up to 7 GPU instances, 10GB each	Up to 7 GPU instances, 10GB each
Max Power Consumption	400 W	350 W	700 W	700 W
Thermal Solution	Passive	Passive	Water cooled	Water cooled

Feature	H100 NVL 94GB PCIe adapter	H100 80GB PCIe adapter	H100 80GB SXM board	H100 94GB SXM board
Compute APIs	CUDA, DirectCompute, OpenCL, OpenACC	CUDA, DirectCompute, OpenCL, OpenACC	CUDA, DirectCompute, OpenCL, OpenACC	CUDA, DirectCompute, OpenCL, OpenACC

* With structural sparsity enabled

Server support

The following tables list the ThinkSystem servers that are compatible.

NVLink server support: The NVLink Ampere bridge is supported with additional NVIDIA A-series and H-series GPUs. As a result, there are additional servers listed as supporting the bridge that don't support the H100 GPU.

Table 3. Server support (Part 1 of 4)

Part Number	Description	2S AMD V3				2S Intel V3			4S 8S Intel V3			Multi Node			GPU Rich		1S V3	
		SR635 V3 (7D9H / 7D9G)	SR655 V3 (7D9F / 7D9E)	SR645 V3 (7D9D / 7D9C)	SR665 V3 (7D9B / 7D9A)	ST650 V3 (7D7B / 7D7A)	SR630 V3 (7D72 / 7D73)	SR650 V3 (7D75 / 7D76)	SR850 V3 (7D97 / 7D96)	SR860 V3 (7D94 / 7D93)	SR950 V3 (7DC5 / 7DC4)	SD535 V3 (7DD8 / 7DD1)	SD530 V3 (7DDA / 7DD3)	SD550 V3 (7DD9 / 7DD2)	SR670 V2 (7Z22 / 7Z23)	SR675 V3 (7D9Q / 7D9R)	ST250 V3 (7DCF / 7DCE)	SR250 V3 (7DCM / 7DCL)
Double-wide PCIe adapter form factor																		
4X67A89325	ThinkSystem NVIDIA H100 NVL 94GB PCIe Gen5 Passive GPU	N	N	N	N	N	N	N	N	N	N	N	N	N	N	8	N	N
4X67A82257	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU	N	3	N	3	N	N	3	2	4	N	N	N	N	8	8	N	N
SXM form factor																		
BQQV	ThinkSystem NVIDIA H100 SXM5 700W 80G HBM3 GPU Board	N	N	N	N	N	N	N	N	N	N	N	N	N	N	1	N	N
BUBB	ThinkSystem NVIDIA H100 SXM5 700W 94G HBM2e GPU Board	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
NVLink bridge (for PCIe adapters only, not SXM; order 3 per pair of GPUs)																		
4X67A71309	ThinkSystem NVIDIA Ampere NVLink 2-Slot Bridge	N	N	N	N	N	N	N	N	N	N	N	N	N	Y	Y	N	N

Table 4. Server support (Part 2 of 4)

Part Number	Description	Edge					Super Computing					1S Intel V2			2S Intel V2		
		SE350 (7Z46 / 7D1X)	SE350 V2 (7DA9)	SE360 V2 (7DAM)	SE450 (7D8T)	SE455 V3 (7DBY)	SD665 V3 (7D9P)	SD665-N V3 (7DAZ)	SD650 V3 (7D7M)	SD650-I V3 (7D7L)	SD650-N V3 (7D7N)	ST50 V2 (7D8K / 7D8J)	ST250 V2 (7D8G / 7D8F)	SR250 V2 (7D7R / 7D7Q)	ST650 V2 (7Z75 / 7Z74)	SR630 V2 (7Z70 / 7Z71)	SR650 V2 (7Z72 / 7Z73)
Double-wide PCIe adapter form factor																	
4X67A89325	ThinkSystem NVIDIA H100 NVL 94GB PCIe Gen5 Passive GPU	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
4X67A82257	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	3
SXM form factor																	
BQQV	ThinkSystem NVIDIA H100 SXM5 700W 80G HBM3 GPU Board	N	N	N	N	N	N	1[1]	N	N	1	N	N	N	N	N	N
BUBB	ThinkSystem NVIDIA H100 SXM5 700W 94G HBM2e GPU Board	N	N	N	N	N	N	1[1]	N	N	1	N	N	N	N	N	N
NVLink bridge (for PCIe adapters only, not SXM; order 3 per pair of GPUs)																	
4X67A71309	ThinkSystem NVIDIA Ampere NVLink 2-Slot Bridge	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

1. Contains 4 separate GPUs connected via high-speed interconnects

Table 5. Server support (Part 3 of 4)

Part Number	Description	AMD V1					Dense V2				4S V2		8S	4S V1			1S Intel V1			
		SR635 (7Y98 / 7Y99)	SR655 (7Y00 / 7Z01)	SR655 Client OS	SR645 (7D2Y / 7D2X)	SR665 (7D2W / 7D2V)	SD630 V2 (7D1K)	SD650 V2 (7D1M)	SD650-N V2 (7D1N)	SN550 V2 (7Z69)	SR850 V2 (7D31 / 7D32)	SR860 V2 (7Z59 / 7Z60)	SR950 (7X11 / 7X12)	SR850 (7X18 / 7X19)	SR850P (7D2F / 2D2G)	SR860 (7X69 / 7X70)	ST50 (7Y48 / 7Y50)	ST250 (7Y45 / 7Y46)	SR150 (7Y54)	SR250 (7Y52 / 7Y51)
Double-wide PCIe adapter form factor																				
4X67A89325	ThinkSystem NVIDIA H100 NVL 94GB PCIe Gen5 Passive GPU	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
4X67A82257	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU	N	N	N	N	3	N	N	N	N	N	N	N	N	N	N	N	N	N	N
SXM form factor																				
BQQV	ThinkSystem NVIDIA H100 SXM5 700W 80G HBM3 GPU Board	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
BUBB	ThinkSystem NVIDIA H100 SXM5 700W 94G HBM2e GPU Board	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
NVLink bridge (for PCIe adapters only, not SXM; order 3 per pair of GPUs)																				
4X67A71309	ThinkSystem NVIDIA Ampere NVLink 2-Slot Bridge	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Table 6. Server support (Part 4 of 4)

Part Number	Description	2S Intel V1								Dense V1			
		ST550 (7X09 / 7X10)	SR530 (7X07 / 7X08)	SR550 (7X03 / 7X04)	SR570 (7Y02 / 7Y03)	SR590 (7X98 / 7X99)	SR630 (7X01 / 7X02)	SR650 (7X05 / 7X06)	SR670 (7Y36 / 7Y37)	SD530 (7X21)	SD650 (7X58)	SN550 (7X16)	SN850 (7X15)
Double-wide PCIe adapter form factor													
4X67A89325	ThinkSystem NVIDIA H100 NVL 94GB PCIe Gen5 Passive GPU	N	N	N	N	N	N	N	N	N	N	N	N
4X67A82257	ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU	N	N	N	N	N	N	N	N	N	N	N	N
SXM form factor													
BQQV	ThinkSystem NVIDIA H100 SXM5 700W 80G HBM3 GPU Board	N	N	N	N	N	N	N	N	N	N	N	N
BUBB	ThinkSystem NVIDIA H100 SXM5 700W 94G HBM2e GPU Board	N	N	N	N	N	N	N	N	N	N	N	N
NVLink bridge (for PCIe adapters only, not SXM; order 3 per pair of GPUs)													
4X67A71309	ThinkSystem NVIDIA Ampere NVLink 2-Slot Bridge	N	N	N	N	N	N	N	N	N	N	N	N

Operating system support

The following table lists the supported operating systems.

Tip: These tables are automatically generated based on data from Lenovo ServerProven.

Table 7. Operating system support for ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU, 4X67A82257

Operating systems	SR650 V3 (4th Gen Xeon)	SR650 V3 (5th Gen Xeon)	SR655 V3	SR665 V3	SR675 V3	SR850 V3	SR860 V3	SR650 V2	SR670 V2	SR665
Microsoft Windows 10	N	Y	Y	Y	N	N	N	N	N	N
Microsoft Windows 11	N	Y	Y	Y	N	N	N	N	N	N
Microsoft Windows Server 2019	Y	Y	Y	Y	Y	Y [2]	Y [2]	Y	Y	Y [3]
Microsoft Windows Server 2022	Y	Y	Y [1]	Y	Y	Y	Y	Y	Y	Y [3]
Red Hat Enterprise Linux 7.9	N	N	N	N	N	N	N	Y	Y	N
Red Hat Enterprise Linux 8.3	N	N	N	N	N	N	N	Y	Y	Y [3]
Red Hat Enterprise Linux 8.4	N	N	N	N	N	N	N	Y	Y	Y [3]
Red Hat Enterprise Linux 8.5	N	N	N	N	N	N	N	Y	Y	Y [3]
Red Hat Enterprise Linux 8.6	Y	N	Y	Y	Y	Y	Y	Y	Y	Y [3]
Red Hat Enterprise Linux 8.7	Y	N	Y	Y	Y	Y	Y	Y	Y	Y [3]
Red Hat Enterprise Linux 8.8	Y	Y	Y	Y	N	Y	Y	Y	Y	Y [3]
Red Hat Enterprise Linux 9.0	Y	N	Y	Y	Y	Y	Y	Y	Y	Y [3]
Red Hat Enterprise Linux 9.1	Y	N	Y	Y	Y	Y	Y	Y	Y	Y [3]
Red Hat Enterprise Linux 9.2	Y	Y	Y	Y	N	Y	Y	Y	Y	Y [3]
SUSE Linux Enterprise Server 15 SP3	N	N	N	N	N	N	N	Y	Y	Y [3]
SUSE Linux Enterprise Server 15 SP4	Y	N	Y [1]	Y	Y	Y	Y	Y	Y	Y [3]
SUSE Linux Enterprise Server 15 SP5	Y	Y	Y	Y	N	Y	Y	Y	Y	Y [3]
Ubuntu 18.04.5 LTS	N	N	N	N	N	N	N	Y	Y	N
Ubuntu 20.04 LTS	N	N	N	N	N	N	N	Y	N	N
Ubuntu 20.04.5 LTS	N	N	Y	Y	Y	Y	Y	N	N	N
Ubuntu 22.04 LTS	Y	N	Y [1]	Y	Y	Y	Y	Y	Y	Y [3]
VMware vSphere Hypervisor (ESXi) 7.0 U3	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y [3]
VMware vSphere Hypervisor (ESXi) 8.0	Y	N	Y [1]	Y	N	N	N	Y	Y	Y [3]
VMware vSphere Hypervisor (ESXi) 8.0 U1	Y	N	Y	Y	Y	Y	Y	Y	Y	Y [3]
VMware vSphere Hypervisor (ESXi) 8.0 U2	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y [3]

[1] For limitation, please refer Support Tip TT1064

[2] For limitation, please refer Support Tip TT1591

[3] HW is not supported with EPYC 7002 processors.

NVIDIA GPU software

This section lists the NVIDIA software that is available from Lenovo.

- NVIDIA vGPU Software (vApps, vPC, RTX vWS, and vCS)
- NVIDIA AI Enterprise Software
- NVIDIA HPC Compiler Software

The PCIe adapter H100 GPUs include a five-year software subscription, including enterprise support, to the NVIDIA AI Enterprise software suite:

- ThinkSystem NVIDIA H100 NVL 94GB PCIe Gen5 Passive GPU, 4X67A89325
- ThinkSystem NVIDIA H100 80GB PCIe Gen5 Passive GPU, 4X67A82257

This license is equivalent to part number 7S02001HWW listed in the NVIDIA AI Enterprise Software section below.

To activate the NVIDIA AI Enterprise license, see the following page:
https://www.nvidia.com/en-us/data-center/activate-license/

SXM GPUs: The NVIDIA AI Enterprise software suite is not included with the SXM H100 GPUs and will need to ordered separately if needed.

NVIDIA vGPU Software (vApps, vPC, RTX vWS)

Lenovo offers the following virtualization software for NVIDIA GPUs:

- **Virtual Applications (vApps)**

  For organizations deploying Citrix XenApp, VMware Horizon RDSH or other RDSH solutions. Designed to deliver PC Windows applications at full performance. NVIDIA Virtual Applications allows users to access any Windows application at full performance on any device, anywhere. This edition is suited for users who would like to virtualize applications using XenApp or other RDSH solutions. Windows Server hosted RDSH desktops are also supported by vApps.

- **Virtual PC (vPC)**

  This product is ideal for users who want a virtual desktop but need great user experience leveraging PC Windows® applications, browsers and high-definition video. NVIDIA Virtual PC delivers a native experience to users in a virtual environment, allowing them to run all their PC applications at full performance.

- **NVIDIA RTX Virtual Workstation (RTX vWS)**

  NVIDIA RTX vWS is the only virtual workstation that supports NVIDIA RTX technology, bringing advanced features like ray tracing, AI-denoising, and Deep Learning Super Sampling (DLSS) to a virtual environment. Supporting the latest generation of NVIDIA GPUs unlocks the best performance possible, so designers and engineers can create their best work faster. IT can virtualize any application from the data center with an experience that is indistinguishable from a physical workstation — enabling workstation performance from any device.

The following license types are offered:

- **Perpetual license**

  A non-expiring, permanent software license that can be used on a perpetual basis without the need to renew. Each Lenovo part number includes a fixed number of years of Support, Upgrade and Maintenance (SUMS).

- **Annual subscription**

  A software license that is active for a fixed period as defined by the terms of the subscription license, typically yearly. The subscription includes Support, Upgrade and Maintenance (SUMS) for the duration of the license term.

- **Concurrent User (CCU)**

A method of counting licenses based on active user VMs. If the VM is active and the NVIDIA vGPU software is running, then this counts as one CCU. A vGPU CCU is independent of the connection to the VM.

The following table lists the ordering part numbers and feature codes.

Table 8. NVIDIA vGPU Software

Part number	Feature code 7S02CTO1WW	Description
NVIDIA vApps		
7S020003WW	B1MP	NVIDIA vApps Perpetual License and SUMS 5Yr, 1 CCU
7S020004WW	B1MQ	NVIDIA vApps Subscription License 1 Year, 1 CCU
7S020005WW	B1MR	NVIDIA vApps Subscription License 3 Years, 1 CCU
7S02003DWW	S832	NVIDIA vApps Subscription License 4 Years, 1 CCU
7S02003EWW	S833	NVIDIA vApps Subscription License 5 Years, 1 CCU
NVIDIA vPC		
7S020009WW	B1MV	NVIDIA vPC Perpetual License and SUMS 5Yr, 1 CCU
7S02000AWW	B1MW	NVIDIA vPC Subscription License 1 Year, 1 CCU
7S02000BWW	B1MX	NVIDIA vPC Subscription License 3 Years, 1 CCU
7S02003FWW	S834	NVIDIA vPC Subscription License 4 Years, 1 CCU
7S02003GWW	S835	NVIDIA vPC Subscription License 5 Years, 1 CCU
NVIDIA RTX vWS		
7S02000FWW	B1N1	NVIDIA RTX vWS Perpetual License and SUMS 5Yr, 1 CCU
7S02000GWW	B1N2	NVIDIA RTX vWS Subscription License 1 Year, 1 CCU
7S02000HWW	B1N3	NVIDIA RTX vWS Subscription License 3 Years, 1 CCU
7S02000XWW	S6YJ	NVIDIA RTX vWS Subscription License 4 Years, 1 CCU
7S02000YWW	S6YK	NVIDIA RTX vWS Subscription License 5 Years, 1 CCU
7S02000LWW	B1N6	NVIDIA RTX vWS EDU Perpetual License and SUMS 5Yr, 1 CCU
7S02000MWW	B1N7	NVIDIA RTX vWS EDU Subscription License 1 Year, 1 CCU
7S02000NWW	B1N8	NVIDIA RTX vWS EDU Subscription License 3 Years, 1 CCU
7S02003BWW	S830	NVIDIA RTX vWS EDU Subscription License 4 Years, 1 CCU
7S02003CWW	S831	NVIDIA RTX vWS EDU Subscription License 5 Years, 1 CCU

NVIDIA AI Enterprise Software

Lenovo offers the NVIDIA AI Enterprise (NVAIE) cloud-native enterprise software. NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data analytics software, optimized, certified, and supported by NVIDIA to run on VMware vSphere and bare-metal with NVIDIA-Certified Systems™. It includes key enabling technologies from NVIDIA for rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud.

NVIDIA AI Enterprise is licensed on a per-GPU basis. NVIDIA AI Enterprise products can be purchased as either a perpetual license with support services, or as an annual or multi-year subscription.

- The perpetual license provides the right to use the NVIDIA AI Enterprise software indefinitely, with no expiration. NVIDIA AI Enterprise with perpetual licenses must be purchased in conjunction with one-year, three-year, or five-year support services. A one-year support service is also available for renewals.
- The subscription offerings are an affordable option to allow IT departments to better manage the flexibility of license volumes. NVIDIA AI Enterprise software products with subscription includes support services for the duration of the software's subscription license

The features of NVIDIA AI Enterprise Software are listed in the following table.

Table 9. Features of NVIDIA AI Enterprise Software (NVAIE)

Features	**Supported in NVIDIA AI Enterprise**
Per GPU Licensing	Yes
Compute Virtualization	Supported
Windows Guest OS Support	No support
Linux Guest OS Support	Supported
Maximum Displays	1
Maximum Resolution	4096 x 2160 (4K)
OpenGL and Vulkan	In-situ Graphics only
CUDA and OpenCL Support	Supported
ECC and Page Retirement	Supported
MIG GPU Support	Supported
Multi-vGPU	Supported
NVIDIA GPUDirect	Supported
Peer-to-Peer over NVLink	Supported
GPU Pass Through Support	Supported
Baremetal Support	Supported
AI and Data Science applications and Frameworks	Supported
Cloud Native ready	Supported

Note: Maximum 10 concurrent VMs per product license

The following table lists the ordering part numbers and feature codes.

Table 10. NVIDIA AI Enterprise Software (NVAIE)

Part number	**Feature code 7S02CTO1WW**	**Description**
AI Enterprise Perpetual License		
7S02001BWW	S6YY	NVIDIA AI Enterprise Perpetual License and Support per GPU, 5 Years
7S02001EWW	S6Z1	NVIDIA AI Enterprise Perpetual License and Support per GPU, EDU, 5 Years
AI Enterprise Subscription License		
7S02001FWW	S6Z2	NVIDIA AI Enterprise Subscription License and Support per GPU, 1 Year
7S02001GWW	S6Z3	NVIDIA AI Enterprise Subscription License and Support per GPU, 3 Years
7S02001HWW	S6Z4	NVIDIA AI Enterprise Subscription License and Support per GPU, 5 Years
7S02001JWW	S6Z5	NVIDIA AI Enterprise Subscription License and Support per GPU, EDU, 1 Year
7S02001KWW	S6Z6	NVIDIA AI Enterprise Subscription License and Support per GPU, EDU, 3 Years

Part number	Feature code 7S02CTO1WW	Description
7S02001LWW	S6Z7	NVIDIA AI Enterprise Subscription License and Support per GPU, EDU, 5 Years

Find more information in the NVIDIA AI Enterprise Sizing Guide.

NVIDIA HPC Compiler Software

Table 11. NVIDIA HPC Compiler

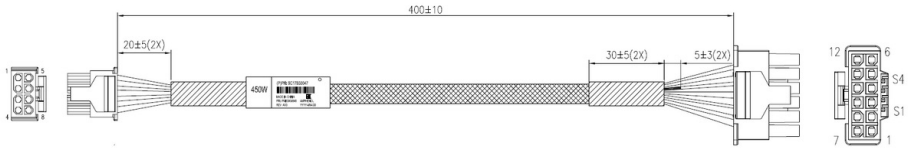
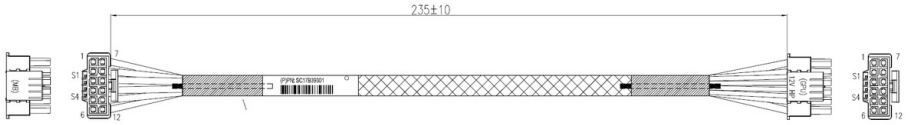
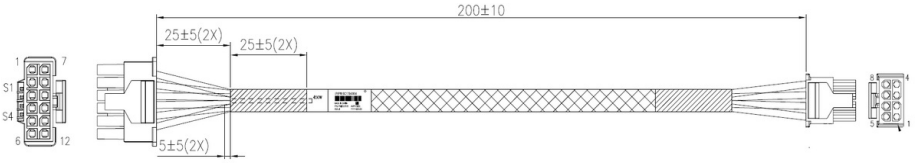
Part number	Feature code 7S09CTO6WW	Description
HPC Compiler Support Services		
7S090014WW	S924	NVIDIA HPC Compiler Support Services, 1 Year
7S090015WW	S925	NVIDIA HPC Compiler Support Services, 3 Years
7S09002GWW	S9UQ	NVIDIA HPC Compiler Support Services, 5 Years
7S090016WW	S926	NVIDIA HPC Compiler Support Services, EDU, 1 Year
7S090017WW	S927	NVIDIA HPC Compiler Support Services, EDU, 3 Years
7S09002HWW	S9UR	NVIDIA HPC Compiler Support Services, EDU, 5 Years
7S090018WW	S928	NVIDIA HPC Compiler Support Services - Additional Contact, 1 Year
7S09002JWW	S9US	NVIDIA HPC Compiler Support Services - Additional Contact, 3 Years
7S09002KWW	S9UT	NVIDIA HPC Compiler Support Services - Additional Contact, 5 Years
7S090019WW	S929	NVIDIA HPC Compiler Support Services - Additional Contact, EDU, 1 Year
7S09002LWW	S9UU	NVIDIA HPC Compiler Support Services - Additional Contact, EDU, 3 Years
7S09002MWW	S9UV	NVIDIA HPC Compiler Support Services - Additional Contact, EDU, 5 Years
HPC Compiler Premier Support Services		
7S09001AWW	S92A	NVIDIA HPC Compiler Premier Support Services, 1 Year
7S09002NWW	S9UW	NVIDIA HPC Compiler Premier Support Services, 3 Years
7S09002PWW	S9UX	NVIDIA HPC Compiler Premier Support Services, 5 Years
7S09001BWW	S92B	NVIDIA HPC Compiler Premier Support Services, EDU, 1 Year
7S09002QWW	S9UY	NVIDIA HPC Compiler Premier Support Services, EDU, 3 Years
7S09002RWW	S9UZ	NVIDIA HPC Compiler Premier Support Services, EDU, 5 Years
7S09001CWW	S92C	NVIDIA HPC Compiler Premier Support Services - Additional Contact, 1 Year
7S09002SWW	S9V0	NVIDIA HPC Compiler Premier Support Services - Additional Contact, 3 Years
7S09002TWW	S9V1	NVIDIA HPC Compiler Premier Support Services - Additional Contact, 5 Years
7S09001DWW	S92D	NVIDIA HPC Compiler Premier Support Services - Additional Contact, EDU, 1 Year
7S09002UWW	S9V2	NVIDIA HPC Compiler Premier Support Services - Additional Contact, EDU, 3 Years
7S09002VWW	S9V3	NVIDIA HPC Compiler Premier Support Services - Additional Contact, EDU, 5 Years

Auxiliary power cables

The power cables needed for the H100 SXM GPUs are included with the supported servers.

The H100 PCIe GPU option part number does not ship with auxiliary power cables. Cables are server-specific due to length requirements. For CTO orders, auxiliary power cables are derived by the configurator. For field upgrades, cables will need to be ordered separately as listed in the table below.

Table 12. Auxiliary power cables for H100

Auxiliary power cable needed with the SR650 V3, SR655 V3, SR665 V3, SR665, SR650 V2
400mm 16-pin (2x6+4) cable **Option**: SR665: 4X97A85028, ThinkSystem 400mm 2x6+4 GPU Power Cable SR650 V2: 4X97A85028, ThinkSystem 400mm 2x6+4 GPU Power Cable SR650 V3: 4X67A82883, ThinkSystem SR650 V3 GPU Full Length Thermal Option Kit* SR655 V3: 4X67A86438, ThinkSystem SR655 V3 GPU Enablement Kit* SR665 V3: 4X67A85856, ThinkSystem SR665 V3 GPU Full Length Thermal Option Kit* **Feature**: BRWK **SBB**: SBB7A66338 **Base**: SC17B33047 **FRU**: 03KM846  * The option part numbers are for thermal kits and include other components needed to install the GPU. See the SR650 V3 product guide or SR655 V3 product guide or SR665 V3 product guide for details.
Auxiliary power cable needed with the SR675 V3
235mm 16-pin (2x6+4) cable **Option**: 4X97A84510, ThinkSystem SR675 V3 Supplemental Power Cable for H100 GPU Option **Feature**: BSD2 **SBB**: SBB7A65299 **Base**: SC17B39301 **FRU**: 03LE554 
Auxiliary power cable needed with the SR850 V3, SR860 V3
200mm 16-pin (2x6+4) cable **Option**: 4X97A88016, ThinkSystem SR850 V3/SR860 V3 H100 GPU Power Cable Option Kit **Feature**: BW28 **SBB**: SBB7A72759 **Base**: SC17B40604 **FRU**: 03LF915 
Auxiliary power cable needed with the SR670 V2
215mm 16-pin (2x6+4) cable **Option**: 4X97A85027, ThinkSystem SR670 V2 H100/L40 GPU Option Power Cable **Feature**: BRWL **SBB**: SBB7A66339 **Base**: SC17B33046 **FRU**: 03KM845 

Regulatory approvals

The NVIDIA H100 GPU has the following regulatory approvals:

- RCM
- BSMI
- CE
- FCC
- ICES
- KCC
- cUL, UL
- VCCI

Operating environment

The NVIDIA H100 GPU has the following operating characteristics:

- Ambient temperature
  - Operational: 0°C to 50°C (-5°C to 55°C for short term*)
  - Storage: -40°C to 75°C
- Relative humidity:
  - Operational: 5-85% (5-93% short term*)
  - Storage: 5-95%

* A period not more than 96 hours consecutive, not to exceed 15 days per year.

Warranty

One year limited warranty. When installed in a Lenovo server, the GPU assumes the server's base warranty and any warranty upgrades.

Seller training courses

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

1. **Generative AI Overview Foundational**
   2024-02-16 | 17 minutes | Employees Only

   It seems the whole world is excited about Generative AI, and while some of it is just hype, it has become clear that Generative AI has the potential to revolutionize many aspects of our personal and professional lives. In this brief NVIDIA course, we'll explore one aspect of the Generative AI excitement, the value you get from Generative AI technology. We will discuss what Generative AI is, how it works, and how enterprises are planning to use this technology.

   By the end of this course, you will be able to discuss the Generative AI market trends and the challenges in this space with your customers. And you will be able to explain what Generative AI is and how the technology works to help enterprises unlock new opportunities for business.

   Published: 2024-02-16
   Length: 17 minutes
   Employee link: Grow@Lenovo
   Course code: DAINVD106

2. **Industry Use Cases in Modern Computing Foundational**
   2024-02-16 | 9 minutes | Employees Only

   As GPU powered computing continues to improve exponentially, applications that were once science fiction are becoming best practice. This is an introductory NVIDIA course that explores some exciting industry focused use cases that are providing companies with faster time to insight, productivity at scale and a great ROI.

   By the end of this course, you will be able to explain how companies in a few key industry verticals are benefiting from a variety of accelerated compute use cases.

   Published: 2024-02-16
   Length: 9 minutes
   Employee link: Grow@Lenovo
   Course code: DAINVD105
3. **Introduction to Artificial Intelligence Foundational**
   2024-02-16 | 10 minutes | Employees Only

   This NVIDIA course aims to answer questions such as, what is AI and why are enterprises so interested in it? and how does AI happen, why are GPUs so important for it, and what does a good AI solution look like?

   By the end of this training, you should be able to describe AI and relate it to some common enterprise use cases. You'll know the difference between training and inference and be able to visualize a typical AI workflow. More importantly, you'll understand the difficulties of traditional CPU-based AI and appreciate why businesses would benefit greatly by adopting GPU-accelerated workflows. Finally, you'll also understand what features contribute to an awesome AI solution and why customers respect and enjoy NVIDIA's solutions.

   Published: 2024-02-16
   Length: 10 minutes
   Employee link: Grow@Lenovo
   Course code: DAINVD104
4. **GPU Fundamentals Foundational**
   2024-02-16 | 10 minutes | Employees Only

   This NVIDIA course introduces you to two devices that a computer typically uses to process information, the CPU and the GPU. We'll discuss their differences and look at how the GPU overcomes the limitations of the CPU. Once you understand the power and advantages of GPU processing, we will talk about the value GPUs bring to modern-day enterprise computing.

   By the end of this course, you should know the difference between serial and parallel processing. You will be able to explain what a GPU is in very simple terms and explain the value that GPUs bring to enterprises. Additionally, you'll become familiar with the typical GPU-accelerated enterprise workloads and list one or two use cases under them. By the time you exit this course, you should be able to target various GPU-accelerated computing opportunities with the right NVIDIA GPU.

   Published: 2024-02-16
   Length: 10 minutes
   Employee link: Grow@Lenovo
   Course code: DAINVD103

5. **Partner Technical Webinar – NVidia**
   2023-12-11 | 60 minutes | Employees and Partners

   In this 60-minute replay, Brad Davidson of Nvidia will help us recognize AI Trends, and Discuss Industry Verticals Marketing.

   Published: 2023-12-11
   Length: 60 minutes
   Employee link: Grow@Lenovo
   Partner link: Lenovo Partner Learning
   Course code: 120823

Related publications

For more information, refer to these documents:

- ThinkSystem and ThinkAgile GPU Summary:
  https://lenovopress.lenovo.com/lp0768-thinksystem-thinkagile-gpu-summary
- ServerProven compatibility:
  https://serverproven.lenovo.com/
- NVIDIA H100 product page:
  https://www.nvidia.com/en-us/data-center/h100/
- NVIDIA Hopper Architecture page
  https://www.nvidia.com/en-us/data-center/technologies/hopper-architecture/

Related product families

Product families related to this document are the following:

- GPU adapters

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, LP1732, was created or updated on November 8, 2023.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1732
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1732.

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®
ServerProven®
ThinkAgile®
ThinkSystem®

The following terms are trademarks of other companies:

Intel® and Xeon® are trademarks of Intel Corporation or its subsidiaries.

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

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

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