ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU

Product Guide

The ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU based on the Ada Lovelace architecture, is a powerful universal GPU for the data center, delivering breakthrough multi-workload acceleration for large language model (LLM) inference and training, graphics, and video applications. As the premier platform for multi-modal generative AI, the L40S GPU provides end-to-end acceleration for inference, training, graphics, and video workflows to power the next generation of AI-enabled audio, speech, 2D, video, and 3D applications.

The following figure shows the ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU.

Figure 1. ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU

Did you know?

AI models are exploding in complexity and popularity with the disruption led by large language models (LLMs) such as ChatGPT and generative AI diffusion models. L40S's fourth-generation Tensor Cores with the Transformer Engine and new FP8 data format enable AI performance that exceeds the NVIDIA A100 Tensor Core GPUs for many AI training and inference workloads.

Part number information

The following table shows the ordering information for the NVIDIA L40S GPU.

Note: The NVIDIA L40S GPU is not available in the following markets: China, Hong Kong, Macau

Table 1. Ordering information

Part number	Feature code	Description
4X67A90669	BYFH	ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU

The option part number includes the following:

- One NVIDIA L40S GPU with full-height (3U) adapter bracket attached
- Documentation

Features

Generative AI is fueling transformative change, unlocking a new frontier of opportunities for enterprises across every industry. To transform with AI, enterprises need more compute resources, greater scale, and a broad set of capabilities to meet the demands of an ever-increasing set of diverse and complex workloads.

The NVIDIA L40S GPU is the most powerful universal GPU for the data center, delivering end-to-end acceleration for the next generation of AI-enabled applications—from generative AI and model training and inference to 3D graphics, rendering, and video applications.

Enterprises are looking to use mainstream infrastructure to satisfy their compute needs, but training state-of-the-art models requires massive compute capability. For LLM models, eight L40S's in mainstream servers bring up to 1.7X the training performance of an NVIDIA HGX™ A100 8-GPU system, giving enterprises fast time to solution with traditional infrastructure. When compared to the A100 80GB SXM for inference, the L40S delivers up to 1.2X more generative AI inference performance using StableDiffusion and up to 1.5X inference performance on popular networks, such as those included within the MLPerf benchmark.

Key use cases of the NVIDIA L40S GPU:

- **Generative AI**

  The AI, graphics, and media acceleration capabilities of the L40S GPU make it the premier platform for multi-modal generative AI pipelines. With powerful inferencing capabilities, combined with NVIDIA RTX™-accelerated ray tracing and dedicated encode and decode engines, the L40S accelerates AI-enabled audio, speech, 2D, video, and 3D generative AI applications.

  For image generative AI inference, the L40S GPU delivers more than 5X higher performance than the previous-generation NVIDIA A40 GPU and 1.2X more performance than the HGX A100. This breakthrough performance, combined with 48GB of memory capacity, makes the L40S GPU the ideal generative AI platform for high-quality images and immersive visual content.

- **LLM Inference and Training**

  Accelerate training, fine tuning, and inference workloads with powerful throughput and floating-point performance to build and deploy state-of-the-art AI models. Powerful NVIDIA-Certified Systems™ with eight L40 GPUs can train foundational models with up to 175 billion parameters to convergence and accelerate fine-tuning and retraining of existing large-scale models to adapt them for new tasks.

  Combining NVIDIA's full stack of inference serving software with the compute capabilities of the L40S provides a powerful platform for trained models ready for inference. With support for structural sparsity and a broad range of precisions, including TF32, INT8, and FP8, the L40S delivers over 1 petaFLOPS of inference operation performance, delivering actionable insights with speed and precision.

- **AI-Ready Development Platform with NVIDIA AI Enterprise**

  Enterprise adoption of AI is now mainstream and leading to an increased demand for skilled AI developers and data scientists. Organizations require a flexible, high-performance platform consisting of optimized hardware and software to maximize productivity and accelerate AI development.

  NVIDIA AI Enterprise is an end-to-end, enterprise-grade AI software platform that offers 100+ frameworks, pretrained models, and libraries to streamline development and deployment of production AI, including generative AI, computer vision, and speech AI. Optimized and certified for reliable performance, NVIDIA AI Enterprise, together with the L40S, provides a unified platform to develop applications once and deploy anywhere, reducing the risks involved with moving from pilot to production.

- **Rendering and 3D Graphics**

  Running professional 3D visualization applications with NVIDIA L40S enables creative professionals to iterate more, render faster, and unlock tremendous performance advantages that increase productivity and speed up project completion. The NVIDIA L40S's third-generation RT Cores and industry-leading 48GB of GDDR6 memory deliver up to 2X the real-time ray-tracing performance of the previous generation.

  With these capabilities, artists and designers can work with complex geometry and high-resolution textures in real time to generate photorealistic designs and power full-fidelity creative workflows, from interactive rendering to virtual production.

- **NVIDIA Omniverse**

  NVIDIA Omniverse is a multi-GPU-enabled open platform for Universal Scene Description (USD)-based collaboration and real-time photorealistic simulation. The full-stack platform based on USD and NVIDIA RTX is the powerful culmination of NVIDIA's core graphics, compute, and AI technologies. NVIDIA L40S GPUs bring powerful AI and RTX capabilities to accelerate 3D content creation and industrial digitalization.

  For the most complex Omniverse workloads like extended reality (XR), multi-user design collaboration, and digital twins, the NVIDIA L40S enables ray-traced and path-traced rendering of materials, physically accurate simulations, and generation of photorealistic 3D synthetic data.

- **Streaming and Video Content**

  The NVIDIA L40S takes streaming and video content workloads to the next level, delivering breakthrough media acceleration capabilities with three video encode and three video decode engines. With the addition of AV1 encoding, the L40S delivers up to 2X the performance and improved TCO for broadcast streaming, video production, and transcription workloads.

- **Virtual Workstations**

  When combined with NVIDIA RTX Virtual Workstation (vWS) software, the NVIDIA L40S can be virtualized to deliver high-performance workstation instances to remote users for high-end design, AI, and compute workloads. With 48GB of GPU memory, the NVIDIA L40S with vWS enables flexible, work-from-anywhere solutions for GPU memory-intensive workloads.

Technical specifications

The NVIDIA L40S GPU has the following specifications:

- Form factor
  - PCIe Full Height Full Length adapter (4.4-in x 10.5-in), Double-width (dual slot)
  - NVIDIA Form Factor 5.5

- Host interface:
  - PCIe 4.0 x16
  - MSI-X interrupt messaging protocol (MSI not supported)
  - PCIe Lane Polarity Inversion and Lane Reversal
- Single Root I/O Virtualization (SR-IOV) support
  - 256 virtual functions (VFs)
  - ARI Forwarding
- Hardware Root of Trust
  - Secure boot
  - Secure firmware upgrade
  - Firmware rollback protection
  - Support for in-band firmware update disable (established after each GPU reset)
  - Secure application processor recovery

The following table lists the GPU processing specifications and performance of the NVIDIA L40S GPU.

Table 2. Specifications of the NVIDIA L40S GPU

Feature	**Specification**
GPU Architecture	NVIDIA Ada Lovelace
NVIDIA CUDA Parallel Processing Cores	18,176
NVIDIA Tensor Cores (4th gen)	568
NVIDIA RT Cores (3rd Gen)	142
Peak FP32 performance (non-Tensor)	91.6 TFLOPS
Peak FP16 Tensor performance	362.05 TFLOPS, 733 TFLOPS*
Peak Tensor Float 32 (TF32) performance	183 TFLOPS, 366 TFLOPS*
Peak Bfloat16 (BF16) Tensor performance	362.05 TFLOPS, 733 TFLOPS*
Peak FP8 Tensor performance	733 TFLOPS, 1466 TFLOPS*
Peak INT8 Integer Performance	733 TOPS, 1466 TOPS*
Peak INT4 Integer Performance	733 TOPS, 1466 TOPS*
RT Core performance	209 TFLOPS
GPU Memory	48 GB GDDR6
Memory Bandwidth	864 GB/s
ECC	Yes
NVIDIA NVLink	No support
System Interface	PCIe Gen 4, x16 lanes
Form Factor	PCIe full height/length, double width (10.5" x 4.4")
Multi-Instance GPU (MIG)	No support
Max Power Consumption	350 W
Thermal Solution	Passive
vGPU Software Support	NVIDIA vPC/vApps, NVIDIA RTX Virtual Workstation (vWS)
Display connectors	4x DisplayPort 1.4a (disabled by default**)

Feature	Specification
Max Simultaneous Displays	Up to four 5K Monitors at 60Hz per card or dual 8K displays @ 60Hz (requires DisplayPort 1.4 DSC); Each display port can support 4K at 120 Hz with 30-bit color
Graphics APIs	DirectX 12 Ultimate, Shader Model 6.6, OpenGL 4.6, Vulkan 1.3
Compute APIs	CUDA 12.0, Direct Compute, OpenCL 3.0

* With structural sparsity enabled
** To enable the DisplayPort ports, see https://developer.nvidia.com/displaymodeselector

Server support

The following tables list the ThinkSystem servers that are compatible.

Table 3. Server support (Part 1 of 4)

Part Number	Description	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)	SR680a V3 (7DHE)	SR685a V3 (7DHC)	ST50 V3 (7DF4 / 7DF3)	ST250 V3 (7DCF / 7DCE)	SR250 V3 (7DCM / 7DCL)
4X67A90669	ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU	N	3	N	3	N	N	3	N	N	N	N	N	N	8	8	N	N	N	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)
4X67A90669	ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

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)
4X67A90669	ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU	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)
4X67A90669	ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU	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 L40S 48GB PCIe Gen4 Passive GPU, 4X67A90669

Operating systems	SR650 V3 (4th Gen Xeon)	SR650 V3 (5th Gen Xeon)	SR655 V3	SR665 V3	SR675 V3	SR670 V2
Microsoft Windows 10	Y	Y	Y	Y	N	N
Microsoft Windows 11	Y	Y	Y	Y	N	N
Microsoft Windows Server 2019	Y	Y	Y	Y	Y	Y
Microsoft Windows Server 2022	Y	Y	Y	Y	Y	Y
Red Hat Enterprise Linux 8.6	Y	N	Y	Y	Y	Y
Red Hat Enterprise Linux 8.7	Y	N	Y	Y	Y	Y
Red Hat Enterprise Linux 8.8	Y	Y	Y	Y	N	Y
Red Hat Enterprise Linux 8.9	Y	Y	Y	Y	N	Y
Red Hat Enterprise Linux 9.0	Y	N	Y	Y	Y	Y
Red Hat Enterprise Linux 9.1	Y	N	Y	Y	Y	Y
Red Hat Enterprise Linux 9.2	Y	Y	Y	Y	N	Y
Red Hat Enterprise Linux 9.3	Y	Y	Y	Y	N	Y
SUSE Linux Enterprise Server 15 SP4	Y	N	Y	Y	Y	Y
SUSE Linux Enterprise Server 15 SP5	Y	Y	Y	Y	N	Y
Ubuntu 18.04.5 LTS	N	N	N	N	N	Y
Ubuntu 20.04.5 LTS	N	N	Y	Y	Y	N
Ubuntu 22.04 LTS	Y	N	Y	Y	Y	Y
VMware vSphere Hypervisor (ESXi) 8.0 U1	Y	N	Y	Y	Y	Y
VMware vSphere Hypervisor (ESXi) 8.0 U2	Y	Y	Y	Y	Y	Y

NVIDIA GPU software

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

- NVIDIA vGPU Software (vApps, vPC, RTX vWS)
- NVIDIA Omniverse Software (OVE)
- NVIDIA AI Enterprise Software
- NVIDIA HPC Compiler Software

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 SUMS ONLY 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		

Part number	Feature code 7S02CTO1WW	Description
7S020009WW	B1MV	NVIDIA vPC SUMS 5Yr ONLY, 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 SUMS ONLY 5Yr, 1 CCU
7S02000GWW	B1N2	NVIDIA RTX vWS Subsc Lic 1Yr 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 SUMS ONLY 5Y, 1CCU
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 Omniverse Software (OVE)

NVIDIA Omniverse™ Enterprise is an end-to-end collaboration and simulation platform that fundamentally transforms complex design workflows, creating a more harmonious environment for creative teams.

NVIDIA and Lenovo offer a robust, scalable solution for deploying Omniverse Enterprise, accommodating a wide range of professional needs. This document details the critical components, deployment options, and support available, ensuring an efficient and effective Omniverse experience.

Deployment options cater to varying team sizes and workloads. Using Lenovo NVIDIA-Certified Systems™ and Lenovo OVX nodes which are meticulously designed to manage scale and complexity, ensures optimal performance for Omniverse tasks.

Deployment options include:

- Workstations: NVIDIA-Certified Workstations with RTX 6000 Ada GPUs for desktop environments.
- Data Center Solutions: Deployment with Lenovo OVX nodes or NVIDIA-Certified Servers equipped with L40, L40S or A40 GPUs for centralized, high-capacity needs.

NVIDIA Omniverse Enterprise includes the following components and features:

- Platform Components: Kit, Connect, Nucleus, Simulation, RTX Renderer.
- Foundation Applications: USD Composer, USD Presenter.
- Omniverse Extensions: Connect Sample & SDK.
- Integrated Development Environment (IDE)
- Nucleus Configuration: Workstation, Enterprise Nucleus Server (supports up to 8 editors per scene); Self-Service Public Cloud Hosting using Containers.
- Omniverse Farm: Supports batch workloads up to 8 GPUs.
- Enterprise Services: Authentication (SSO/SSL), Navigator Microservice, Large File Transfer, User Accounts SAML/Account Directory.
- User Interface: Workstation & IT Managed Launcher.

- Support: NVIDIA Enterprise Support.
- Deployment Scenarios: Desktop to Data Center: Workstation deployment for building and designing, with options for physical or virtual desktops. For batch tasks, rendering, and SDG workloads that require headless compute, Lenovo OVX nodes are recommended.

The following part numbers are for a subscription license which is active for a fixed period as noted in the description. The license is for a named user which means the license is for named authorized users who may not re-assign or share the license with any other person.

Table 9. NVIDIA Omniverse Software (OVE)

Part number	Feature 7S02CTO1WW	Description
7S02003ZWW	SCX0	NVIDIA Omniverse Enterprise Subscription per GPU, 1 Year
7S020042WW	SCX3	NVIDIA Omniverse Enterprise Subscription per GPU, 3 Years
7S020041WW	SCX2	NVIDIA Omniverse Enterprise Subscription per GPU, INC, 1 Year
7S020040WW	SCX1	NVIDIA Omniverse Enterprise Subscription per GPU, EDU, 1 Year
7S020043WW	SCX4	NVIDIA Omniverse Enterprise Subscription per GPU, EDU, 3 Years

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 10. 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

Features	Supported in NVIDIA AI Enterprise
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 11. 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
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 12. 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

Part number	Feature code 7S09CTO6WW	Description
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 GPU option part number does not ship with auxiliary power cables. Cables are server-specific due to length requirements and the connector on the server end of the cable. 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.

Tip: The names of the cable options below may only include the H100 or L40 GPU, however these cables are also supported with the L40S.

Table 13. Auxiliary power cables for NVIDIA L40S GPU

Auxiliary power cable needed with the SR650 V3, SR665 V3
400mm 16-pin (2x6+4) cable **Feature**: BRWK **SBB**: SBB7A66338 **Option**: ● SR650 V3: 4X67A82883, ThinkSystem SR650 V3 GPU Full Length Thermal Option Kit* ● SR665 V3: 4X67A85856, ThinkSystem SR665 V3 GPU Full Length Thermal Option Kit* **Base**: SC17B33047 **FRU**: 03KM846 * The option part number is for the thermal kit and includes other components needed to install the GPU. See the relevant server product guide for details.
Auxiliary power cable needed with the SR675 V3
235mm 16-pin (2x6+4) cable **Feature**: BSD2 **SBB**: SBB7A65299 **Option**: 4X97A84510, ThinkSystem SR675 V3 Supplemental Power Cable for H100 GPU Option **Base**: SC17B39301 **FRU**: 03LE554
Auxiliary power cable needed with the SR670 V2
215mm 16-pin (2x6+4) cable **Feature**: BRWL **SBB**: SBB7A66339 **Option**: 4X97A85027, ThinkSystem SR670 V2 H100/L40 GPU Option Power Cable **Base**: SC17B33046 **FRU**: 03KM845

Regulatory approvals

The NVIDIA L40S GPU has the following regulatory approvals:

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

Operating environment

The NVIDIA L40S 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 to 85% (5 to 93% short term*)
  - Storage: 5 to 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. **Introduction to Artificial Intelligence**
   2024-08-02 | 11 minutes | Employees and Partners

   IMPORTANT: If you receive the following error message:
   "There is an issue with this slide content. Please contact your administrator", please change your VPN location setting and try again. We are actively working on fixing this issue. Thank you for your understanding!

   This NVIDIA course aims to answer questions such as:

   • What is AI?
   • Why are enterprises so interested in it?
   • How does AI happen?
   • Why are GPUs so important for it?
   • What does a good AI solution look like?

   Course Objectives:

   By the end of this training, you should be able to:
   1. Describe AI on a high level and list a few common enterprise use cases
   2. List how enterprises benefit from AI
   3. Distinguish between Training and Inference
   4. Say how GPUs address known bottlenecks in a typical AI pipeline
   5. Tell a customer why NVIDIA's AI solutions are well-respected in the market

   Published: 2024-08-02
   Length: 11 minutes
   Employee link: Grow@Lenovo
   Partner link: Lenovo Partner Learning
   Course code: DAINVD104r2

2. **GPU Fundamentals**
   2024-08-02 | 10 minutes | Employees and Partners

   IMPORTANT: If you receive the following error message:
   "There is an issue with this slide content. Please contact your administrator",
   please change your VPN location setting and try again. We are actively working on fixing this issue.
   Thank you for your understanding.

   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. We will also talk about the value GPUs bring to modern-day enterprise computing.

   Course Objectives:

   By the end of this training, you should be able to:
   1. Distinguish between serial and parallel processing
   2. Explain what a GPU is and what it does at a high level
   3. Articulate the value of GPU computing for enterprises
   4. List three typical GPU-accelerated workloads and a few uses cases
   5. Recommend the appropriate NVIDIA GPU for its corresponding enterprise computing workloads

   Published: 2024-08-02
   Length: 10 minutes
   Employee link: Grow@Lenovo
   Partner link: Lenovo Partner Learning
   Course code: DAINVD103r2

3. **Key NVIDIA Use Cases for Industry Verticals**
   2024-08-02 | 32 minutes | Employees and Partners

   IMPORTANT: If you receive the following error message:
   "There is an issue with this slide content. Please contact your administrator",
   please change your VPN location setting and try again. We are actively working on fixing this issue.
   Thank you for your understanding.

   In this NVIDIA course, you will learn about key AI use cases driving innovation and change across Automotive, Financial Services, Energy, Healthcare, Higher Education, Manufacturing, Retail and Telco industries.

   Course Objectives:
   By the end of this training, you should be able to:
   1. Discuss common AI use cases across a broad range of industry verticals
   2. Explain how NVIDIA's AI software stack speeds up time to production for AI projects in multiple industry verticals

   Published: 2024-08-02
   Length: 32 minutes
   Employee link: Grow@Lenovo
   Partner link: Lenovo Partner Learning
   Course code: DAINVD108

4. **Generative AI Overview**
   2024-08-02 | 17 minutes | Employees and Partners

   IMPORTANT: If you receive the following error message:
   "There is an issue with this slide content. Please contact your administrator", please change your VPN location setting and try again. We are actively working on fixing this issue. Thank you for your understanding!

   Since ChatGPTs debut in November of 2022, it has become clear that Generative AI has the potential to revolutionize many aspects of our personal and professional lives. This NVIDIA course aims to answer questions such as:

   • What are the Generative AI market trends?
   • What is generative AI and how does it work?

   Course Objectives:

   By the end of this training, you should be able to:
   1. Discuss the Generative AI market trends and the challenges in this space with your customers.
   2. Explain what Generative AI is and how the technology works to help enterprises to unlock new opportunities for the business.
   3. Present a high-level overview of the steps involved in building a Generative AI application.

   Published: 2024-08-02
   Length: 17 minutes
   Employee link: Grow@Lenovo
   Partner link: Lenovo Partner Learning
   Course code: DAINVD106r2

5. **Retrieval Augmented Generation**
   2024-08-02 | 15 minutes | Employees and Partners

   IMPORTANT: If you receive the following error message:
   "There is an issue with this slide content. Please contact your administrator", please change your VPN location setting and try again. We are actively working on fixing this issue. Thank you for your understanding!

   In this NVIDIA course, Dave Barry, Senior Solutions Architect, talks about a technique known as Retrieval Augmented Generation (RAG). It is a powerful tool for enhancing the accuracy and reliability of Generative AI models with facts fetched from external sources.

   This course requires prior knowledge of Generative AI concepts, such as the difference between model training and inference. Please refer to relevant courses within this curriculum.

   Course Objectives:

   By the end of this training, you should be able to:
   1. Explain the limitations of large language models to customers
   2. Articulate the value of RAG to enterprises
   3. Demo an NVIDIA RAG workflow with a video
   4. Drive TCO conversations using an authentic use case

   Published: 2024-08-02
   Length: 15 minutes
   Employee link: Grow@Lenovo
   Partner link: Lenovo Partner Learning
   Course code: DAINVD107

6. **AI Industry Use Cases & Solutions**
   2024-08-02 | 25 minutes | Employees and Partners

   IMPORTANT: If you receive the following error message:
   "There is an issue with this slide content. Please contact your administrator", please change your VPN location setting and try again. We are actively working on fixing this issue. Thank you for your understanding!

   This NVIDIA course aims to answer the question:

   • How does NVIDIA bring AI solutions to market with and through the partner ecosystem?

   Course Objectives:

   By the end of this training, you should be able to:
   1. Think of solutions in terms of an industry and use case approach
   2. Develop solutions that address the industry-specific challenges (with FSI as the illustrative model)
   3. Engage customers with their conversations and advance deals with stakeholder's concerns in mind
   4. Replicate NVIDIA's best practices and ecosystem engagement strategies appropriately

   Published: 2024-08-02
   Length: 25 minutes
   Employee link: Grow@Lenovo
   Partner link: Lenovo Partner Learning
   Course code: DAINVD105r2
7. **Partner Technical Webinar - NVIDIA Smart Spaces**
   2024-07-24 | 60 minutes | Employees and Partners

   In this 60-minute replay, Alex Pazos, NVIDIA BDM for Smart Spaces, reviewed the NVIDIA AI for Smart Spaces framework and use cases. Alex reviewed the Metropolus Framework and the Smart Spaces ecosystem. Then he reviewed several use cases including sports stadiums, warehouses, airports, and roadways.

   Published: 2024-07-24
   Length: 60 minutes
   Employee link: Grow@Lenovo
   Partner link: Lenovo Partner Learning
   Course code: 071924
8. **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.

   Published: 2024-07-01
   Length: 25 minutes
   Employee link: Grow@Lenovo
   Partner link: Lenovo Partner Learning
   Course code: DNVIS102

9. **Think AI Weekly: Lenovo AI PCs & AI Workstations**
   2024-05-23 | 60 minutes | Employees Only

   Join Mike Leach, Sr. Manager, Workstations Solutions and Pooja Sathe, Director Commercial AI PCs as they discuss why Lenovo AI Developer Workstations and AI PCs are the most powerful, where they fit into the device to cloud ecosystem, and this week's Microsoft announcement, Copilot+PC

   Published: 2024-05-23
   Length: 60 minutes
   Employee link: Grow@Lenovo
   Course code: DTAIW105

10. **VTT Cloud Architecture: NVidia Using Cloud for GPUs and AI**
    2024-05-22 | 60 minutes | Employees Only

    Join JD Dupont, NVIDIA Head of Americas Sales, Lenovo partnership and Veer Mehta, NVIDIA Solution Architect on an interactive discussion about cloud to edge, designing cloud Solutions with Nvidia GPUs and minimizing private\hybrid cloud OPEX with GPUs. Discover how you can use what is done at big public cloud providers for your customers. We will also walk through use cases and see a demo you can use to help your customers.

    Published: 2024-05-22
    Length: 60 minutes
    Employee link: Grow@Lenovo
    Course code: DVCLD212

11. **Partner Technical Webinar - Nvidia Update**
    2024-05-13 | 60 minutes | Employees and Partners

    In this 60-minute replay, Veer Mehta, Nvidia Solutions Architect gave an Nvidia AI update for Lenovo. Veer reviewed the highlights from the Nvidia GTC. He also reviewed the Nvidia hardware and software offerings that Lenovo sells.

    Published: 2024-05-13
    Length: 60 minutes
    Employee link: Grow@Lenovo
    Partner link: Lenovo Partner Learning
    Course code: 051024

12. **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

13. **NVIDIA L40S GPU Overview and Business Use Case**
    2023-10-12 | 60 minutes | Employees Only

    Welcome to the NVIDIA L40S GPU Overview and Business Use Case course. This course offers a closer look at the L40S GPU, featuring a webinar presented by Brad Davidson from NVIDIA. Throughout this course, we delve deep into the L40S GPU's capabilities, provide situational use cases, guide you on effectively positioning the L40S in various scenarios, and facilitate a meaningful comparison between the L40S and DGX systems.

    Completing this course will enable you to:
    • Describe the basics of NVIDIA L40S
    • Discuss how NVIDIA L40S delivers level performance for AI
    • Discuss generative AI and omniverse

    Published: 2023-10-12
    Length: 60 minutes
    Employee link: Grow@Lenovo
    Course code: DAINVD102

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:
  http://www.lenovo.com/us/en/serverproven
- Lenovo Reference Architecture for Generative AI Based on Large Language Models (LLMs)
  https://lenovopress.lenovo.com/lp1798-reference-architecture-for-generative-ai-based-on-large-language-models
- NVIDIA L40S product page:
  https://www.nvidia.com/en-us/data-center/l40s/

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, LP1812, was created or updated on May 24, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1812
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1812.

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:

AMD is a trademark of Advanced Micro Devices, Inc.

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®, DirectX®, 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.