GPU Options for ThinkSystem Servers
Reference Information

Lenovo ThinkSystem servers support GPU technology from NVIDIA and AMD to accelerate different computing workloads, maximize performance for graphic design, virtualization, artificial intelligence and high performance computing applications in Lenovo servers.

- NVIDIA AI and Virtualization
- NVIDIA 3D Graphics
- Intel AI and Virtualization
- AMD AI and Virtualization
- Qualcomm AI

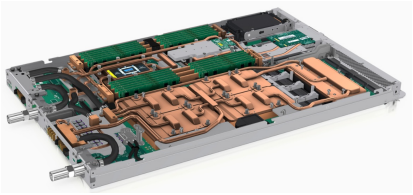
NVIDIA AI and Virtualization

In this section:

SXM GPUs

The following SXM GPUs from NVIDIA are offered for ThinkSystem servers.

ThinkSystem NVIDIA H100 SXM5

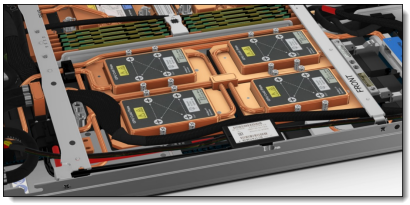


The ThinkSystem NVIDIA H100 PCIe Gen5 GPU delivers unprecedented performance, scalability, and security for every workload. The GPUs use breakthrough innovations in the NVIDIA Hopper™ architecture to deliver industry-leading conversational AI, speeding up large language models by 30X over the previous generation. The NVIDIA H100 is available in both double-wide PCIe adapter form factor and in SXM form factor. The H100 SXM5 GPU is used in Lenovo's Neptune direct-water-cooled ThinkSystem SD665-N V3 server for the ultimate in GPU performance and heat management.

Learn more:

- H100 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA A100 SXM



NVIDIA A100 Tensor Core GPUs delivers outstanding acceleration and flexibility to power the world's highest-performing elastic data centers for AI, data analytics, and HPC applications. As the engine of the NVIDIA data center platform, A100 provides up to 20X higher performance over V100 GPUs and can efficiently scale up to thousands of GPUs, or be partitioned into seven isolated GPU instances to accelerate workloads of all sizes. NVIDIA A100 is available in both double-wide PCIe adapter form factor and in SXM form factor. The A100 SXM GPU is used in Lenovo's Neptune direct-water-cooled ThinkSystem SD650-N V2 server for the ultimate in GPU performance and heat management.

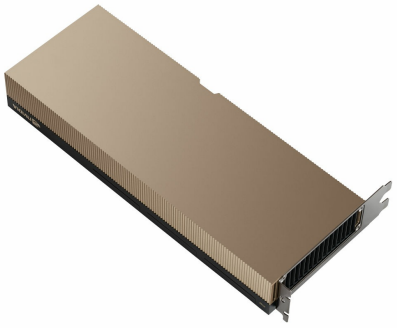
Learn more:

- A100 Product Guide
- ThinkSystem GPU summary

NVIDIA dual slot adapters

The following dual-slot (double-wide) GPUs from NVIDIA are offered for ThinkSystem and ThinkAgile servers.

ThinkSystem NVIDIA H100 & H100 NVL GPUs

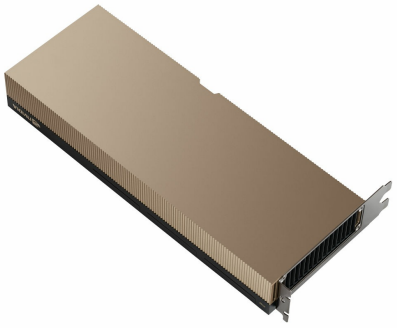


The ThinkSystem NVIDIA H100 PCIe Gen5 GPU delivers unprecedented performance, scalability, and security for every workload. The GPUs use breakthrough innovations in the NVIDIA Hopper™ architecture to deliver industry-leading conversational AI, speeding up large language models by 30X over the previous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

Learn more:

- H100 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA H800 & H800 NVL GPUs



The ThinkSystem NVIDIA H800 PCIe Gen5 GPU delivers high performance, scalability, and security for every workload. It uses breakthrough innovations in the NVIDIA Hopper architecture to deliver industry-leading conversational AI.

Note: The ThinkSystem NVIDIA H800 is only available in the China, Hong Kong and Macau markets.

Learn more:

- H800 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA L40S GPU

The ThinkSystem NVIDIA L40S 48GB PCIe Gen4 Passive GPU is a powerful universal GPU for the data center, delivering breakthrough multi-workload acceleration for Generative AI and large language model (LLM) inference and training, graphics, and video applications. 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.

Learn more:

- L40S Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA L40 GPU

The ThinkSystem NVIDIA L40 48GB PCIe Gen4 Passive GPU delivers unprecedented visual computing performance for the data center and provides revolutionary neural graphics, compute, and AI capabilities to accelerate the most demanding visual computing workloads. The NVIDIA L40, based on the NVIDIA Ada Lovelace GPU architecture features new generation RT cores and Tensor cores, delivering in combination over a petaflop of inferencing performance. These new features are combined with the latest generation CUDA Cores and 48GB of graphics memory to accelerate visual computing workloads from high-performance virtual workstation instances to large-scale digital twins in NVIDIA Omniverse.

Learn more:

- L40 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA A100 GPU

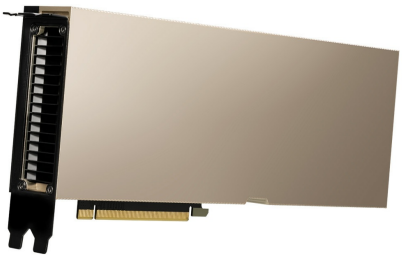
The NVIDIA A100 Tensor Core GPU delivers acceleration at every scale for AI, data analytics, and HPC to tackle the world's toughest computing challenges. As the engine of the NVIDIA data center platform, A100 can efficiently scale up to thousands of GPUs or, using new Multi-Instance GPU (MIG) technology, can be partitioned into seven isolated GPU instances to accelerate workloads of all sizes. A100's third-generation Tensor Core technology now accelerates more levels of precision for diverse workloads, speeding time to insight as well as time to market.

Learn more:

- A100 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA A800 GPU

The NVIDIA A800 Tensor Core GPU delivers outstanding acceleration and flexibility to power the highest-performing elastic data centers for AI, data analytics, and HPC applications. As the engine of the NVIDIA data center platform, A800 provide up to significantly higher performance over V100 GPUs and can efficiently scale up to thousands of GPUs, or be partitioned into seven isolated GPU instances to accelerate workloads of all sizes.



Note: The ThinkSystem NVIDIA H800 is only available in the China, Hong Kong and Macau markets.

Learn more:

- A800 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA A30 GPU

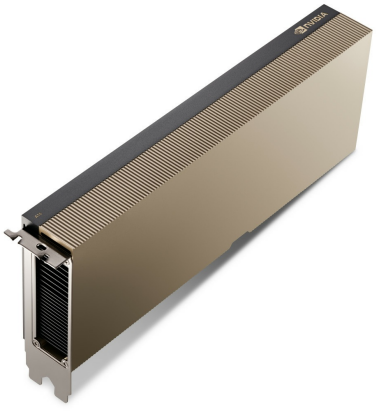
The NVIDIA A30 offers versatile compute acceleration for mainstream enterprise servers. With NVIDIA Ampere architecture Tensor Cores and Multi-Instance GPU (MIG), it delivers speedups securely across diverse workloads, including AI inference at scale and HPC applications. The A30 combines fast memory bandwidth and low-power consumption in a PCIe form factor to enable an elastic data center and delivers maximum value for enterprises.

Learn more:

- A30 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA A16 GPU

Take remote work to the next level with NVIDIA A16. Combined with NVIDIA Virtual PC (vPC) or NVIDIA RTX Virtual Workstation (vWS) software, the A16 enables virtual desktops and workstations with the power and performance to tackle any project from anywhere. Purpose-built for high-density, graphics-rich virtual desktop infrastructure (VDI) and leveraging the NVIDIA Ampere architecture, A16 provides double the user density versus the previous generation, while ensuring the best possible user experience.



Learn more:

- A16 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA Tesla M10 GPU

ThinkSystem NVIDIA Tesla M10 GPU accelerator works with NVIDIA GRID software to provide the industry's highest user density for virtualized desktops and applications. It supports 64 desktops per board and 128 desktops per server, giving your business the power to deliver great experiences to all of your employees at an affordable cost.

Learn more:

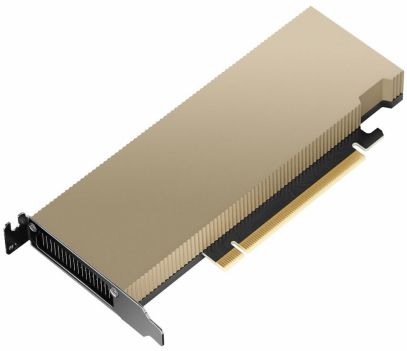
- ThinkSystem GPU summary

NVIDIA single-slot adapters

The following single-slot (single-wide) GPUs from NVIDIA are offered for ThinkSystem, ThinkEdge and ThinkAgile servers.

ThinkSystem NVIDIA L4 GPU

The ThinkSystem NVIDIA L4 24GB PCIe Gen4 Passive GPU delivers universal acceleration and energy efficiency for video, AI, virtual workstations, and graphics in the enterprise, in the cloud, and at the edge. With NVIDIA's AI platform and full-stack approach, L4 is optimized for video and inference at scale for a broad range of AI applications to deliver the best in personalized experiences.



Learn more:

- L4 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA A10 GPU

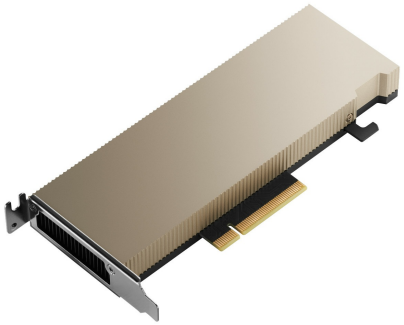
The NVIDIA A10 Tensor Core GPU, combined with NVIDIA RTX Virtual Workstation (vWS) software, brings mainstream graphics and video with AI services to mainstream enterprise servers, delivering the solutions that designers, engineers, artists, and scientists need to meet today's challenges. Built on the latest NVIDIA Ampere architecture, the A10 combines second-generation RT Cores, third-generation Tensor Cores, and new streaming microprocessors with 24 GB of GDDR6 memory for versatile graphics, rendering, AI, and compute performance. From virtual workstations, accessible anywhere in the world, to render nodes to the data centers running a variety of workloads, A10 is built to deliver optimal performance in a single-wide, full-height, full-length PCIe form factor.

Learn more:

- A10 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA A2 GPU

The NVIDIA A2 Tensor Core GPU provides entry-level inference with low power, a small footprint, and high performance for NVIDIA AI at the edge. Featuring a low-profile PCIe Gen4 card and a low 40-60W configurable thermal design power (TDP) capability, the A2 brings versatile inference acceleration to any server for deployment at scale.



Learn more:

- A2 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA Tesla T4 GPU

The NVIDIA Tesla T4 GPU supports diverse cloud workloads, including high-performance computing, deep learning training and inference, machine learning, data analytics, and graphics. Based on the new NVIDIA Turing Architecture and packaged in an energy-efficient 70-watt, small PCIe form factor, Tesla T4 is optimized for scale-out computing environments with its multi-precision Turing Tensor Cores and new RT Cores.

Learn more:

- ThinkSystem GPU summary

NVIDIA 3D Graphics

In this section:

NVIDIA dual-slot graphics adapters

The following dual-slot (double-wide) GPUs from NVIDIA are offered for ThinkSystem and ThinkAgile servers.

ThinkSystem NVIDIA A40 GPU

The NVIDIA A40 is a powerful data center GPU for visual computing, delivering high performance and capabilities to professionals for graphics-based workloads such as ray traced rendering, high-performance virtual workstations, simulation, 3D design, VR, and virtual production. The A40 GPU is a graphics-based virtualization solution for designers, engineers, scientists, and creatives that need this performance from anywhere in the world.

Learn more:

- A40 Product Guide
- ThinkSystem GPU summary

ThinkSystem NVIDIA Quadro RTX A6000 GPU

Unlock the next generation of revolutionary designs, scientific breakthroughs, and immersive entertainment with the NVIDIA RTX A6000, the world's most powerful visual computing GPU. With cutting-edge performance and features, the RTX A6000 lets you work at the speed of inspiration—to tackle the urgent needs of today and meet the rapidly evolving, compute-intensive tasks of tomorrow.

Learn more:

- ThinkSystem GPU summary

ThinkSystem NVIDIA RTX A4500 GPU

Based on the groundbreaking NVIDIA Ampere Architecture graphics processing unit (GPU), NVIDIA RTX A4500 delivers hardware-accelerated ray tracing, revolutionary AI features, advanced shading, and powerful simulation capabilities to creative professionals. With a graphics memory footprint of 20 GB of GDDR6 memory, the A4500 GPU enables the most graphics-intensive applications run with the highest level of user experience, even with largest of data sets.

Learn more:

- ThinkSystem GPU summary

ThinkSystem NVIDIA RTX A2500 GPU

The NVIDIA RTX A2000 brings the power of NVIDIA RTX technology, realtime ray tracing, AI-accelerated compute, and high-performance graphics to more professionals. Built on the NVIDIA Ampere architecture, the VR ready RTX A2000 combines 26 second-generation RT Cores, 104 third-generation Tensor Cores, and 3,328 next-generation CUDA cores and 6 or 12GB of GDDR6 graphics memory with error correction code (ECC) support for error free computing. The RTX A2000 GPU features a power-efficient low profile, dual-slot PCIe form factor, and the RTX A2000 12GB doubles memory for even larger models and datasets. Design bigger, render faster, and work smarter than ever before with RTX A2000 GPUs.

Learn more:

- ThinkSystem GPU summary

NVIDIA single-slot graphics adapters

The following single-slot (single-wide) GPUs from NVIDIA are offered for ThinkSystem and ThinkAgile servers.

ThinkSystem NVIDIA Quadro RTX T1000 GPU

The NVIDIA T1000, built on the NVIDIA Turing GPU architecture, is a powerful, low profile solution that delivers the full-size features, performance and capabilities required by demanding professional applications in a compact graphics card. Featuring 896 CUDA cores and 8GB of GDDR6 memory, the T1000 enables professionals to tackle multi-app workflows, from 3D modeling to video editing. Support for up to four 5K displays gives you the expansive visual workspace to view your work in stunning detail.

Learn more:

- ThinkSystem GPU summary

ThinkSystem NVIDIA Quadro RTX T400 GPU

The NVIDIA T400, built on the NVIDIA Turing GPU architecture, delivers amazing performance and capabilities to power a range of professional workflows. The RTX T400 GPU features 384 CUDA cores and 2GB of GDDR6 memory, and has native support for up to three 5K displays.

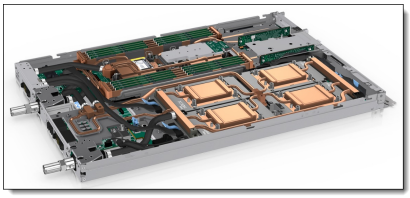
Learn more:

- ThinkSystem GPU summary

Intel AI and Virtualization

Intel Max Series 1550 GPU

The Intel Max Series 1550 GPUs are optimized for machine learning and high-performance computing applications while also containing media decode and encode engines to support certain media analytics use cases. These highly specialized GPUs are enabled with Intel's OneAPI, an open cross-architecture programming model. Intel Max Series 1550 GPUs are used in Lenovo's Neptune direct-water-cooled ThinkSystem SD650-I V3 server for the ultimate in GPU performance and heat management.

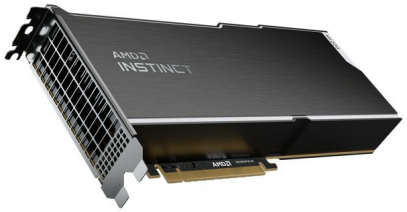


Learn more:

- ThinkSystem GPU summary

AMD AI and Virtualization

ThinkSystem AMD Instinct MI210 Accelerator



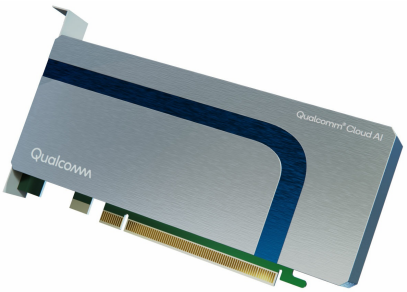
The ThinkSystem AMD Instinct MI210 Accelerator is a compute workhorse optimized for accelerating single precision and double-precision HPC-class system. The accelerator can also be deployed for training large scale machine intelligence workloads. The accelerator's powerful compute engine, new matrix math FP64 cores and advanced memory architecture, combined with AMD's ROCm open software platform and ecosystem, provides a powerful, flexible heterogeneous compute solution that is designed to help datacenter designers meet the challenges of a new era of compute.

Learn more:

- ThinkSystem GPU summary

Qualcomm AI

ThinkSystem Qualcomm Cloud AI 100 Accelerator



The Qualcomm Cloud AI 100 is designed for AI inference acceleration, and addresses the unique requirements in the cloud, including power efficiency, scale, process node advancements, and signal processing. The AI 100 enables data centers to run inference on the edge cloud faster and more efficiently. Qualcomm Cloud AI 100 is designed to be a leading solution for datacenters who increasingly rely on infrastructure at the edge-cloud. The ThinkSystem Qualcomm Cloud AI 100 accelerator is offered on ThinkEdge servers to enable customers to deploy AI workloads at the edge of their network. The AI 100 supports over 150 neural networks across multiple categories, including image classification, object detection, semantic segmentation, and natural language processing.

Learn more:

- Cloud AI 100 Product Guide
- ThinkSystem GPU summary

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, LP0767, was created or updated on September 28, 2023.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP0767
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP0767.

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®
ThinkAgile®
ThinkEdge®
ThinkSystem®

The following terms are trademarks of other companies:

Intel® is a trademark of Intel Corporation or its subsidiaries.

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