ThinkSystem NVIDIA HGX H200 141GB 700W 8-GPU Board

Product Guide

The NVIDIA H200 Tensor Core GPU supercharges generative AI and high-performance computing (HPC) workloads with game-changing performance and memory capabilities. H200 is the newest addition to NVIDIA's leading AI and high-performance data center GPU portfolio, bringing massive compute to data centers.

The NVIDIA H200 141GB 700W GPU is offered in the ThinkSystem SR680a V3 server, with eight SXM5 form factor GPU modules and NVIDIA® NVLink® Fabric to create an 8-FC (fully connected) NVLink topology per baseboard. Leveraging the power of H200 multi-precision Tensor Cores, an eight-way HGX H200 provides over 32 petaFLOPS of FP8 deep learning compute and over 1.1TB of aggregate HBM memory for the highest performance in generative AI and HPC applications.

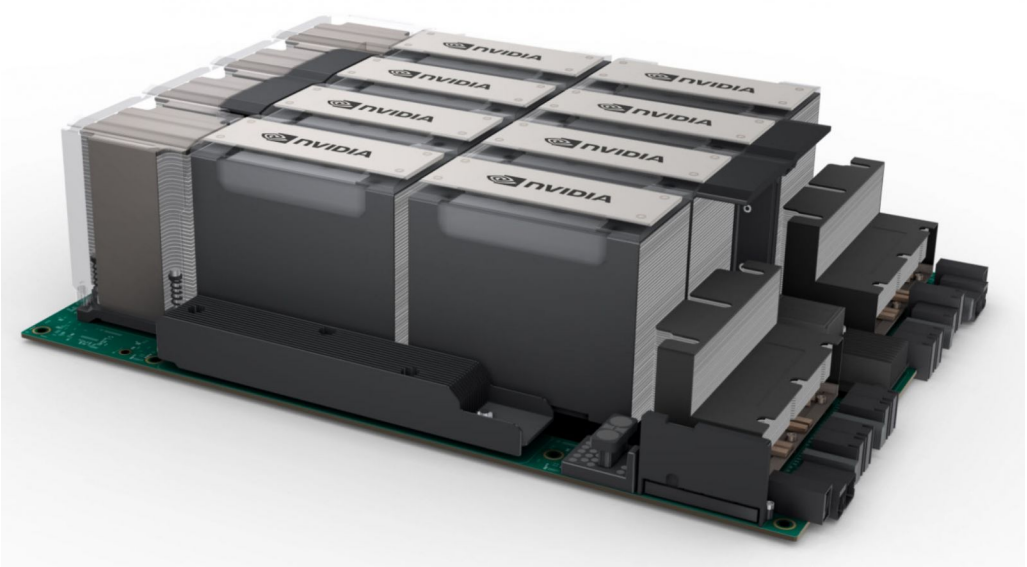

Figure 1. ThinkSystem NVIDIA HGX H200 141GB 700W 8-GPU Board in the ThinkSystem SR680a V3 server

Did you know?

To maximize compute performance, H200 is the world's first GPU with HBM3e memory with 4.8TB/s of memory bandwidth, a 1.4X increase over H100. H200 also expands GPU memory capacity nearly doubled to 141GB. The combination of faster and larger HBM memory accelerates performance of computationally intensive generative AI and HPC applications, while meeting the evolving demands of growing model sizes.

Part number information

The following table shows the part numbers for the 8-GPU board. Feature code C1HMcontains 8x H200 GPUs in the SXM form factor plus the NVLink high-speed interconnections.

Table 1. Ordering information

Part number	Feature code	Description
CTO only	C1HM	ThinkSystem NVIDIA HGX H200 141GB 700W 8-GPU Board

Features

The NVIDIA H200 Tensor Core GPU supercharges generative AI and HPC with game-changing performance and memory capabilities. As the first GPU with HBM3e, H200's faster, larger memory fuels the acceleration of generative AI and LLMs while advancing scientific computing for HPC workloads.

NVIDIA HGX™ H200, the world's leading AI computing platform, features the H200 GPU for the fastest performance. An eight-way HGX H200 provides over 32 petaflops of FP8 deep learning compute and 1.1TB of aggregate high-bandwidth memory for the highest performance in generative AI and HPC applications.

Key AI and HPC workload features:

- Unlock Insights With High-Performance LLM Inference
  In the ever-evolving landscape of AI, businesses rely on large language models to address a diverse range of inference needs. An AI inference accelerator must deliver the highest throughput at the lowest TCO when deployed at scale for a massive user base. H200 doubles inference performance compared to H100 when handling LLMs such as Llama2 70B.
- Optimize Generative AI Fine-Tuning Performance
  Large language models can be customized to specific business case needs with fine-tuning, low-rank adaptation (LoRA), or retrieval-augmented generation (RAG) methods. These methods bridge the gap between general pretrained results and task-specific solutions, making them essential tools for industry and research applications.

  NVIDIA H200's Transformer Engine and fourth-generation Tensor Cores speed up fine-tuning by 5.5X over A100 GPUs. This performance increase allows enterprises and AI practitioners to quickly optimize and deploy generative AI to benefit their business. Compared to fully training foundation models from scratch, fine-tuning offers better energy efficiency and the fastest access to customized solutions needed to grow business.
- Industry-Leading Generative AI Training
  The era of generative AI has arrived, and it requires billion-parameter models to take on the paradigm shift in business operations and customer experiences.

  NVIDIA H200 GPUs feature the Transformer Engine with FP8 precision, which provides up to 5X faster training over A100 GPUs for large language models such as GPT-3 175B. The combination of fourth-generation NVLink, which offers 900GB/s of GPU-to-GPU interconnect, PCIe Gen5, and NVIDIA Magnum IO™ software, delivers efficient scalability from small enterprise to massive unified computing clusters of GPUs. These infrastructure advances, working in tandem with the NVIDIA AI Enterprise software suite, make the NVIDIA H200 the most powerful end-to-end generative AI and HPC data center platform.
- Supercharged High-Performance Computing
  Memory bandwidth is crucial for high-performance computing applications, as it enables faster data transfer and reduces complex processing bottlenecks. For memory-intensive HPC applications like simulations, scientific research, and artificial intelligence, H200's higher memory bandwidth ensures that data can be accessed and manipulated efficiently, leading to up to a 110X faster time to results.

The NVIDIA data center platform consistently delivers performance gains beyond Moore's Law. And H200's breakthrough AI capabilities further amplify the power of HPC+AI to accelerate time to discovery for scientists and researchers working on solving the world's most important challenges.

- Reduced Energy and TCO
  In a world where energy conservation and sustainability are top of mind, the concerns of business leaders and enterprises have evolved. Enter accelerated computing, a leader in energy efficiency and TCO, particularly for workloads that thrive on acceleration, such as HPC and generative AI.

  With the introduction of H200, energy efficiency and TCO reach new levels. This cutting-edge technology offers unparalleled performance, all within the same power profile as H100. AI factories and at-scale supercomputing systems that are not only faster but also more eco-friendly deliver an economic edge that propels the AI and scientific community forward. For at-scale deployments, H200 systems provide 5X more energy savings and 4X better cost of ownership savings over the NVIDIA Ampere architecture generation.

Technical specifications

The following table lists the NVIDIA H200 GPU specifications.

Table 2. GPU specifications

Specification	**NVIDIA H200**
Form Factor	SXM5
FP64	34 teraFLOPS
FP64 Tensor Core	67 teraFLOPS
FP32	67 teraFLOPS
TF32 Tensor Core	495 / 989 teraFLOPS*
BFLOAT16 Tensor	990 / 1,979 teraFLOPS*
FP16 Tensor Core	990 / 1,979 teraFLOPS*
FP8 Tensor Core	1,979 / 3,958 teraFLOPS*
INT8 Tensor Core	1,979 / 3,958 TOPS*
GPU Memory	141 GB HBM3e
GPU Memory Bandwidth	4.8 TB/s
Total Graphics Power (TGP) or Continuous Electrical Design Point (EDPc)	700W
Multi-Instance GPUs	Up to 7 MIGS @ 16.5 GB
Interconnect	NVLink: 900 GB/s PCIe Gen5: 128 GB/s

* Without / with structural sparsity enabled

Server support

The following tables list the ThinkSystem servers that are compatible.

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)	SR680a V3 (7DHE)	SR685a V3 (7DHC)	ST250 V3 (7DCF / 7DCE)	SR250 V3 (7DCM / 7DCL)
C1HM	ThinkSystem NVIDIA HGX H200 141GB 700W 8-GPU Board	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	1[1]	N	N	N

1. Contains 8 separate GPUs connected via high-speed interconnects

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)
C1HM	ThinkSystem NVIDIA HGX H200 141GB 700W 8-GPU Board	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)
C1HM	ThinkSystem NVIDIA HGX H200 141GB 700W 8-GPU Board	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)
C1HM	ThinkSystem NVIDIA HGX H200 141GB 700W 8-GPU Board	N	N	N	N	N	N	N	N	N	N	N	N

Operating system support

Operating system support is based on that of the supported servers. See the SR680a V3 server product guide for details: https://lenovopress.lenovo.com/lp1909-thinksystem-sr680a-v3-server

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

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 7. 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 8. 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 9. 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

Part number	Feature code 7S02CTO1WW	Description
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 10. 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

Part number	Feature code 7S09CTO6WW	Description
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

Regulatory approvals

The NVIDIA H200 GPU has the following regulatory approvals:

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

Warranty

The NVIDIA H200 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 H200 product page:
  https://www.nvidia.com/en-us/data-center/h200/
- NVIDIA Hopper Architecture page
  https://www.nvidia.com/en-us/data-center/technologies/hopper-architecture/
- ThinkSystem SR680a V3 product guide
  https://lenovopress.lenovo.com/lp1909-thinksystem-sr680a-v3-server

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, LP1944, was created or updated on April 23, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1944
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1944.

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® is a trademark of Intel Corporation or its subsidiaries.

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

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.