New 8-GPU AI Servers from Lenovo
Article

The new Lenovo ThinkSystem SR680a V3, SR685a V3, and SR780a V3 GPU systems deliver massive computational performance for Artificial Intelligence (AI), High-Performance Computing (HPC), and graphical and simulation workloads across various industries. The family of servers supports eight high-performance GPUs, either from AMD or NVIDIA, with planned support for GPUs from Intel.

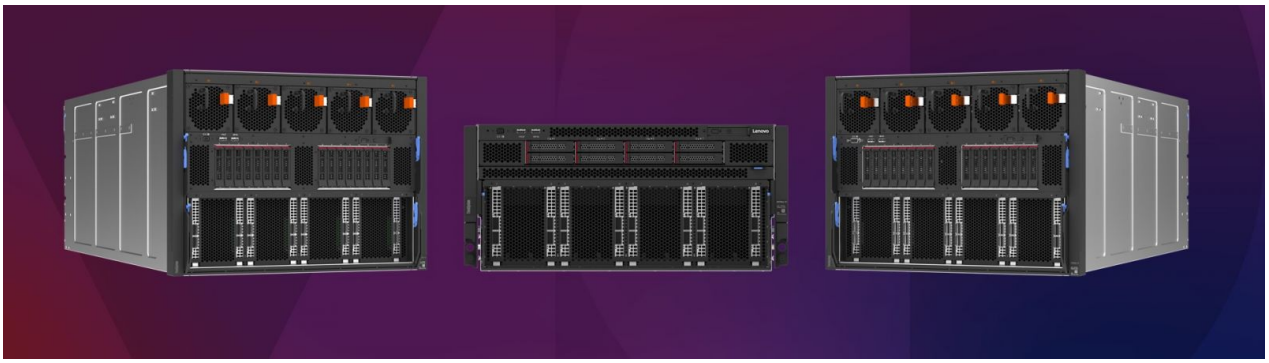

Figure 1. New Lenovo ThinkSystem servers that have 2 processors and 8 GPUs (left to right): SR680a V3, SR780a V3, and SR685a V3

The SR680a V3 and SR685a V3 servers feature two Intel® Xeon® and two AMD EPYC™ processors, respectively, and are 100% air-cooled servers suitable for most data centers. The SR780a V3 server features two Intel Xeon processors and uses a hybrid water-air cooling system to take advantage of the efficiencies of cooling the CPUs and GPUs using a direct water cooled infrastructure.

We also announced enhancements to our Lenovo AI Services Practice with new offerings such as AI Discover and new Fast Start services. We unveiled services to implement, adopt, and scale Generative AI solutions which customers can easily deploy and manage as a Service with Lenovo TruScale.

ThinkSystem SR680a V3

The ThinkSystem SR680a V3 offers a choice of acceleration platforms featuring NVIDIA H100 or H200 GPUs with planned support for AMD MI300X and other future GPUs. With high-speed interconnects between the GPUs, the system delivers unparalleled computational power for demanding AI and HPC workloads.

- 8U chassis with 100% air-cooling
- Two 5th Gen Intel® Xeon® Scalable processors
- 8 GPUs with high-speed interconnect, choice of:
  - NVIDIA H100 700W SXM5 GPUs with 80GB HBM3 GPU memory per GPU
  - NVIDIA H200 700W SXM5 GPUs with 141GB HBM3 GPU memory per GPU
  - AMD MI300X 750W OAM GPUs with 192GB HBM3 memory per GPU (planned)
- Up to 2TB memory with 32x 64GB RDIMMs
- PCIe 5.0 x16 interfaces to all GPUs and network adapters
- Support for high-speed networking up to 400 Gb/s, directly connected to the GPU complex
- Up to 16x high-speed 2.5” NVMe drives

Figure 2. ThinkSystem SR680a V3

For more information see the following resources:

- ThinkSystem SR680a V3 datasheet
- ThinkSystem SR680a V3 product web page
- Thinksystem SR680a V3 product guide (coming in April)
- Thinksystem SR680a V3 interactive 3D tour (coming in April)

ThinkSystem SR685a V3

The ThinkSystem SR685a V3 is an 8U2S rack server built for demanding AI and HPC workloads. It's accelerated for intense computing using industry-leading 4th Generation AMD EPYC™ Processors, interconnects with the fastest transfer rate with AMD Infinity Fabric or NVIDIA NVLink, and supports 8x of the latest GPUs. It has the computing power to tackle modeling, training, rendering, financial tech, scientific research, and more.

- 8U chassis with 100% air-cooling
- Two 4th Gen AMD EPYC processors
- 8 GPUs with high-speed interconnect, choice of:
  - AMD MI300X 750W OAM GPUs with 192GB HBM3 memory per GPU
  - NVIDIA H100 700W SXM5 GPUs with 80GB HBM3 GPU memory per GPU
  - NVIDIA H200 700W SXM5 GPUs with 141GB HBM3 GPU memory per GPU (planned)
- Up to 2.25TB memory with 32x 96GB RDIMMs
- PCIe 5.0 x16 interfaces to all GPUs and network adapters
- Support for high-speed networking up to 400 Gb/s, directly connected to the GPU complex
- Up to 16x high-speed 2.5" NVMe drives

Figure 3. ThinkSystem SR685a V3

For more information see the following resources:

- ThinkSystem SR685a V3 datasheet
- ThinkSystem SR685a V3 product web page
- Thinksystem SR685a V3 product guide (coming in April)
- Thinksystem SR685a V3 interactive 3D tour (coming in April)

ThinkSystem SR780a V3

The ThinkSystem SR780a V3 harnesses the might of eight NVIDIA H200 Tensor Core GPUs, paired with two 5th Gen Intel® Xeon® Scalable processors and 32 DDR5 DIMMs. Through Lenovo Neptune™ liquid cooling, this system achieves the extensive computational prowess crucial for handling demanding AI and HPC workloads. Housed in a compact 5U design, it seamlessly fits into a standard 19-inch rack with a water manifold attached at the rear of the rack.

- 5U chassis with up 80% water-cooling (CPUs & GPUs), 20% air-cooling (other components)
- Two 5th Gen Intel® Xeon® Scalable processors
- 8 GPUs with high-speed interconnect, choice of:
  - NVIDIA H100 700W SXM5 GPUs with 80GB HBM3 GPU memory per GPU
  - NVIDIA H200 700W SXM5 GPUs with 141GB HBM3 GPU memory per GPU
  - NVIDIA B200 SXM6 GPUs with 175GB HBM3 GPU memory per GPU (planned)
  - AMD MI300X 750W OAM GPUs with 192GB HBM3 memory per GPU (planned)
- Up to 2TB memory with 32x 64GB RDIMMs
- PCIe 5.0 x16 interfaces to all GPUs and network adapters
- Support for high-speed networking up to 400 Gb/s, directly connected to the GPU complex
- Up to 8x high-speed 2.5” NVMe drives

Figure 4. ThinkSystem SR780a V3

For more information see the following resources:

- ThinkSystem SR780a V3 datasheet
- ThinkSystem SR780a V3 product web page
- Thinksystem SR780a V3 product guide (planned for later in 2024)
- Thinksystem SR780a V3 interactive 3D tour (planned for later in 2024)

Seller training courses

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

1. **FY24Q4 AI Update**
   2024-03-19 | 10 minutes | Employees and Partners

   This Quick Hit covers the Lenovo and NVIDIA collaboration to deliver the power of Generative AI to every enterprise using new hybrid AI solutions. These solutions include new servers and new services for CSPs and enterprise customers. There are three Lenovo HG Series servers which are based on the NVIDIA MGX modular reference design, and two Lenovo ThinkSystem servers. All use the latest NVIDIA GPUs to meet the demanding workloads generated by AI processing.

   Published: 2024-03-19
   Length: 10 minutes
   Employee link: Grow@Lenovo
   Partner link: Lenovo Partner Learning
   Course code: SXXW2528a
2. **FY25Q1 AI Server Update**
   2024-03-19 | 15 minutes | Employees and Partners

   This Quick Hit focuses on the new Lenovo servers that form part of the Lenovo and NVIDIA collaboration to deliver the power of Generative AI to every enterprise using new hybrid AI solutions. There are three Lenovo HG Series servers which are based on the NVIDIA MGX modular reference design, and three Lenovo ThinkSystem servers. All use the latest NVIDIA GPUs to meet the demanding workloads generated by AI processing.

   Published: 2024-03-19
   Length: 15 minutes
   Employee link: Grow@Lenovo
   Partner link: Lenovo Partner Learning
   Course code: SXXW2528r2a

For more information

For more information about these servers, contact your Lenovo sales representative or use the Contact Us button on one of the product web pages listed above. Preorders for the SR680a V3 and SR685a V3 are planned in March with general orders planned to open towards the end April. Orders for the SR780a V3 are planned for later in 2024.

Related product families

Product families related to this document are the following:

- ThinkSystem SR680a V3 Server
- ThinkSystem SR685a V3 Server
- ThinkSystem SR780a V3 Server

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, LP1921, was created or updated on March 18, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1921
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1921.

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®
Lenovo Neptune®
ThinkSystem®

The following terms are trademarks of other companies:

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

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