Lenovo Leads the Pack in MLPerf Inference: Datacenter 5.0 Benchmarks with Industry-Best AI Performance

Article

The release of MLPerf Inference: Datacenter 5.0 benchmark suite by MLCommons marks another significant milestone in the evaluation of machine learning performance. By measuring how fast systems can process inputs and produce results using a trained model, these comprehensive benchmarks provide insights into datacenter AI capabilities, focusing on inference across diverse hardware configurations. We are proud to participate and are excited to share our outstanding achievements from the latest Lenovo ThinkSystem portfolio including SR650a V4, SR680a V3 and the SR780a V3.

With **groundbreaking** performance results, Lenovo shines as a global leader in AI infrastructure. In the latest MLPerf Inference benchmarks, three of Lenovo's powerful ThinkSystem servers—SR650a V4, SR680a V3, and SR780a V3—dominated across a wide spectrum of AI workloads. These systems, equipped with cutting-edge NVIDIA GPUs, delivered top-tier results in critical benchmarks using models such as GPT-J, Llama 2, Mixtral, Stable Diffusion, ResNet50, and more. Notably, the ThinkSystem SR650a V4 emerged as the top-performing general-purpose AI inferencing system, delivering unmatched versatility for both datacenter and colocation AI deployments.

Highlights from MLPerf Inference: Data Center Performance

Key highlights from the MLPerf results include benchmarks using these servers:

- SR650a V4 - 2U server with for 4x NVIDIA H100 NVL PCIe GPUs
- SR680a V4 - 8U air-cooled server with 8x NVIDIA H200 SXM GPUs
- SR780a V4 - 5U water-cooled server with 8x NVIDIA H200 SXM GPUs

ThinkSystem SR650a V4 with 4x NVIDIA H100-NVL-94GB

Tested with GPT-J, Mixtral, Stable Diffusion and others takes **1st place** as general purpose for AI inferencing systems and use cases. Key tests included:

- **Vision Tasks:** ResNet50 and RetinaNet (both server & offline)
- **Medical Imaging:** 3D-UNet (99 and 99.9 offline)
- **NLP & Generative AI:** GPT-J (99.9 both server & offline; 99 offline)
- **Diffusion Models:** Stable Diffusion XL (both server & offline)
- **Mixture of Experts:** Mixtral-8x7B (both server & offline)

Its consistent top-tier performance across diverse use cases demonstrates its versatility and efficiency, especially for enterprises seeking robust inferencing capabilities.

Figure 1. Lenovo ThinkSystem SR650a V4

ThinkSystem SR680a V3 with 8x NVIDIA H200-SXM-141GB

One of Lenovo's high-performance computing servers excelled in large language model workloads, taking **1st place** in:

- **LLMs:** Llama 2-70B (99 and 99.9 offline)
- **Vision:** ResNet50 offline

Designed to maximize GPU performance, SR680a achieves **2nd and 3rd places** in interactive LLM inferencing, Stable Diffusion XL, Mixtral-8x7B, and RGAT. This balance of top-end performance and flexibility makes the SR680a V3 ideal for demanding AI-heavy enterprises and research environments.

Figure 2. Lenovo ThinkSystem SR680a V3

ThinkSystem SR780a V3 with 8x NVIDIA H200-SXM-141GB

With dual 5th Gen Intel ® Xeon® Scalable processors, the ThinkSystem SR780a V3 offers the performance needed for compute-demanding AI and HPC workloads, proudly achieves first place as most advanced AI and HPC Inferencing systems, the SR780a V3 also impressed across a wide range of benchmarks, with 1st place finishes in:

- **NLP & LLMs:** GPT-J and Llama 2-70B (interactive offline)
- **Mixture of Experts:** Mixtral-8x7B offline
- **Graph Neural Networks:** RGAT offline

Strongly performing for vision and diffusion model workloads, achieving 2nd and 3rd place, underlining its strength as a general-purpose inferencing powerhouse for advanced AI applications.

Figure 3. Lenovo ThinkSystem SR780a V3

Lenovo Hybrid AI Advantage with NVIDIA

Lenovo Hybrid AI Advantage™ with NVIDIA help organizations improve productivity, increase agility, and innovate with trust through standardized and accelerated development and deployment of AI use case solutions. Lenovo Hybrid AI Advantage bring the power of Lenovo AI library and validated, tested hybrid AI factories (hybrid AI platforms, workstations, servers, storage, network, software, models, services, partner ecosystem) to the enterprises.

The hybrid AI factory is designed to support hybrid deployments at the Edge, data centers, Colos, and business locations with cloud integration. It offers flexibility of model, infrastructure choice, enables a wide range of AI applications, agentic and machine learning workflows, and real-time data analysis. Lenovo’s hybrid AI platforms power the hybrid AI factory, and they can scale from a single server with just four GPUs as starter environment to a rack scalable unit (SU) as a turnkey infrastructure solution with partner technology choice.

Future-focused AI with Lenovo

As generative AI, large language models, and inferencing continue to evolve, Lenovo is an established industry leader. These benchmark-topping results not only validate current capabilities but also demonstrate our commitment to AI readiness today and in the future—whether in the datacenter or cloud. Our partnership ecosystem with industry leaders like NVIDIA Lenovo is building the foundation for a more intelligent, connected, and autonomous future.

Lenovo’s leadership in the latest MLPerf Inference results showcases more than just benchmark dominance—it emphasizes our vision for AI that is inclusive, scalable, and enterprise-ready. The ThinkSystem SR650a V4, SR680a V3, and SR780a V3 exemplify Lenovo’s commitment to delivering innovative industry leading infrastructure for every stage of the AI journey. Whether you're developing the next breakthrough in generative AI or deploying edge inferencing at scale, Lenovo offers future-focused solutions you can trust.

For more information

For more information, see the following resources:

- Explore Lenovo AI solutions:
  https://www.lenovo.com/us/en/servers-storage/solutions/ai/
- MLCommons®, the open engineering consortium and leading force behind MLPerf, has now released new results for MLPerf benchmark suites:
  - Benchmark results: https://mlcommons.org/benchmarks/training/
  - Latest news about MLCommons: https://mlcommons.org/news-blog

Author

Aaron Gilbert is the Worldwide AI Solutions Marketing at Lenovo . He specializes in Lenovo Hybrid AI Platform with NVIDIA.

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- MLPerf Benchmark

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, LP2240, was created or updated on June 23, 2025.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2240
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2240.

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 Hybrid AI Advantage
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.