Lenovo ThinkSystem SR650 V4 Shows Scalable Enterprise AI Performance in MLPerf 6.0

Article

The latest MLPerf 6.0 results show that the Lenovo ThinkSystem SR650 V4 continues to be a strong platform for enterprise AI inference. Using a 2-socket configuration with Intel Xeon 6787P processors, Lenovo submitted results on three important workloads:

- Llama 3.1 8B for generative AI
- Whisper for speech AI
- rGAT for graph AI

Together, these benchmarks highlight balanced performance across some of the most relevant AI use cases in business today.

The SR650 V4 is a 2U, 2-socket platform based on Intel Xeon 6700- and 6500-series processors with DDR5 memory support, making it well suited for customers who want to scale AI on familiar enterprise infrastructure.

Figure 1. Lenovo ThinkSystem SR650 V4

MLPerf 6.0 highlights on Lenovo ThinkSystem SR650 V4

In MLPerf 6.0, Lenovo focused on three workloads that align closely with real enterprise AI demand: generative AI, speech AI, and graph AI. Running on the Lenovo ThinkSystem SR650 V4 with 2x Intel Xeon 6787P processors, the results show that the platform can deliver competitive performance across these very different inference scenarios.

For **Llama 3.1 8B**, the ThinkSystem SR650 V4 delivered strong results in both interactive and batch-style testing. In the **Server** scenario, the system achieved **281.78 tokens per second**, which placed Lenovo in **2nd place** among the compared results. In the **Offline** scenario, it delivered **775.98 tokens per second**, good for **3rd place**. The submission also recorded a **p99.9 end-to-end latency of about 14 seconds**, **time to first token of about 2.5 seconds**, and **time per output token of about 103 milliseconds**. Together, these results show that the platform can support both responsive user-facing generative AI and higher-throughput inference environments.

For **Whisper**, the ThinkSystem SR650 V4 achieved **18.57 samples per second** in Lenovo's benchmark summary. In the competitive comparison, the system reached **1,411.42 samples per second** in the Offline test, placing **3rd**. This highlights the platform's ability to handle speech AI workloads such as transcription, meeting summarization, and call analytics efficiently.

For **rGAT**, Lenovo measured throughput of approximately **13.6K samples per second**. In the competitive table, the ThinkSystem SR650 V4 posted **13,570.30 samples per second** in the Offline category, also placing **3rd**. This demonstrates strong capability for graph-based AI workloads such as fraud detection, relationship analysis, and recommendation use cases.

Taken together, these MLPerf 6.0 results show that the ThinkSystem SR650 V4 is not just tuned for a single benchmark. It delivers a balanced profile across language, speech, and graph inference, which is exactly what many enterprises need as they expand AI into production.

Why these MLPerf 6.0 results matter

Enterprise AI is moving from pilots into production. As organizations deploy more AI services, they need systems that can deliver not only raw throughput, but also predictable response times, operational simplicity, and the flexibility to support multiple workload types on a common platform. That is why MLPerf matters: it provides a standardized way to evaluate real AI workloads under comparable conditions. MLCommons describes MLPerf Inference as an industry-standard benchmark suite for measuring machine learning system performance on representative workloads.

The ThinkSystem SR650 V4 fits well into this shift. It gives enterprises a mainstream data center server that can support modern AI inferencing without requiring every deployment to be accelerator centric. For organizations that want to run language, speech, and graph inference within existing operational models, that is a meaningful business advantage. This interpretation is partly inferred from the benchmark coverage and the SR650 V4 platform design.

Built on a proven enterprise platform

Our earlier MLPerf 5.1 paper positioned the ThinkSystem SR650 V4 as a data center-grade server for CPU-only inferencing across diverse AI workloads, including Llama 3.1 8B, Whisper, DLRMv2, RetinaNet, and rGAT. That paper also described the benchmarked configuration as using Intel Xeon 6787P processors, 1024 GB of DDR5-6400 memory, and Ubuntu 24.04.2 LTS.

Reusing that hardware foundation in the MLPerf 6.0 discussion helps show continuity: the same server family continues to deliver practical AI performance as benchmark focus shifts toward newer production workloads.

Conclusion

The MLPerf 6.0 results strengthen the case for the Lenovo ThinkSystem SR650 V4 as a versatile platform for enterprise AI inference. Across Llama 3.1 8B, Whisper, and rGAT, the server demonstrated competitive performance on workloads that map directly to high-value business use cases in generative AI, speech AI, and graph AI.

For organizations looking to deploy AI on proven infrastructure, the ThinkSystem SR650 V4 offers a practical balance of scalability, flexibility, and performance. Rather than being optimized for just one narrow workload, it shows the kind of breadth that enterprises need as AI becomes a standard part of production IT.

System Configuration and Software Environment

The following table lists the server configuration.

Table 1. System Configuration and Software Environment

Component	**Specification**
Platform	Lenovo ThinkSystem SR650 V4
CPU Model	Intel Xeon 6787P
Architecture	x86_64
Microarchitecture	GNR_X2
Base Frequency	2.0GHz
All-core Maximum Frequency	3.2GHz
Maximum Frequency	3.8GHz
L1d Cache	8.1 MiB (172 instances)
L1i Cache	10.8 MiB (172 instances)
L2 Cache	344 MiB (172 instances)
L3 Cache	336 MiB
L3 per Core	3.907 MiB
Installed Memory	1024GB (16x64GB DDR5 6400MT/s [6400MT/s])
Operating system	Ubuntu 24.04.2 LTS
Kernel	6.11.0-25-generic
Python3	Python 3.12.3
OpenSSL	OpenSSL 3.0.13 30 Jan 2024

Author

Kelvin He is an AI Data Scientist at Lenovo. He is a seasoned AI and data science professional specializing in building machine learning frameworks and AI-driven solutions. Kelvin is experienced in leading end-to-end model development, with a focus on turning business challenges into data-driven strategies. He is passionate about AI benchmarks, optimization techniques, and LLM applications, enabling businesses to make informed technology decisions.

Related product families

Product families related to this document are the following:

- MLPerf Benchmark
- ThinkSystem SR650 V4 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, LP2414, was created or updated on April 1, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2414
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2414.

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®
ThinkSystem®

The following terms are trademarks of other companies:

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

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