Smarter AI for All: Lenovo and NVIDIA Advance AI Inference Efficiency with MLPerf 6.0

Article

Running today's most advanced AI models requires processing massive volumes of data with extreme efficiency. That performance depends on a tightly optimized combination of compute, networking, storage, and cooling—engineered for reliability, resilience, and scale.

Lenovo and NVIDIA have worked together to deliver trusted AI infrastructure for enterprise and research environments. In MLPerf Inference v6.0, Lenovo again demonstrated powerful performance in accelerated computing across a broad set of industry-standard workloads, including large language models (LLMs), and graph neural networks.

Systems submitted were the following:

- Lenovo ThinkSystem SR680a V4 with 8× NVIDIA HGX B300-SXM (8x NVIDIA Blackwell Ultra) 2.1TB)
- Lenovo ThinkSystem SR675i V3 with 8× NVIDIA RTX™ PRO 6000 Blackwell Server Edition

Figure 1. Lenovo ThinkSystem SR680a V4 includes 8× NVIDIA B300-SXM GPUs

Highlights of MLPerf Inference v6.0

The **Lenovo ThinkSystem SR680a V4** is next-gen, ultra-dense AI platform engineered for industrialized AI factories. Designed with Lenovo Neptune™ liquid-cooling technology, it delivers GPU-dense compute optimized for training, refining, and deploying large-scale AI models across a diverse range of workloads, proving itself with strong performance in LLMs and multimodal tasks.

The **Lenovo ThinkSystem SR675i V3** is the GPU-dense platform that is optimized for AI inference and

powerful enough for fine-tuning, and simulation across verticals, delivering exceptional parallelism, memory bandwidth, and throughput. GPUs receive data without disruption to maximize utilization. Engineered for AI factories at scale, enterprises can realize better cost/token, faster ROI, and accelerated training for larger models.

Benchmarks submitted:

- LLM llama2-70b
- GPT-OSS-120B
- Automatic Speech Recognition – whisper

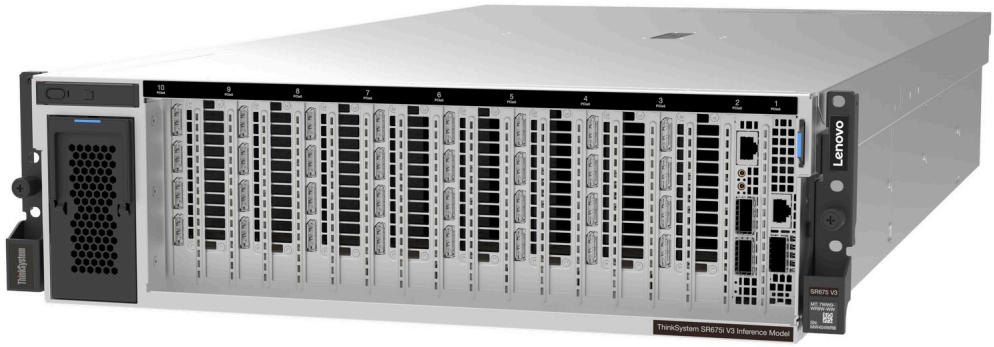

Figure 2. ThinkSystem SR675i V3 Inference Model includes 8× NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs

Empowering Businesses with Lenovo's AI Leadership

Lenovo's MLPerf benchmark results reinforce our continued focus on delivering high-performance, future-ready AI systems that meet the evolving needs of modern workloads.

Key differentiators include:

1. **Cross-Workload Excellence**: Lenovo systems consistently deliver strong results across LLM use cases.
2. **Scalable Infrastructure**: Built with the industry's first AI optimized inference server powered by NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, NVIDIA HGX B300 Blackwell Ultra for large-scale AI training, fine-tuning and inference. Lenovo's systems offer the performance and flexibility needed for AI projects of any size.
3. **Enterprise Integration**: With support for massive inference at scale, centralized training and Hybrid AI deployment, Lenovo empowers organizations to scale AI from research labs to production environments.

Looking Ahead: AI Innovation with Lenovo & NVIDIA

As AI workloads continue to grow in complexity and scale, Lenovo remains firmly committed to innovation. Our strong MLPerf Inference results reflect not only technical excellence but also a forward-thinking approach to solving tomorrow's AI challenges with real business outcomes.

With deep industry collaborations like NVIDIA and a focus on engineering co-designed agile, efficient infrastructure, Lenovo carves a leading path for breakthroughs in generative AI, LLMs, AI Agents, Recommender Systems, and more.

Conclusion

Lenovo's latest MLPerf Inference results reaffirm its leadership in advanced AI infrastructure. Powered by ThinkSystem SR675i V3 and SR680a V4 servers, Lenovo delivers powerful performance, scalability, and

unmatched reliability—whether training complex language models or enabling real-time inference for generative and agentic AI. For organizations ready to lead in AI, Lenovo Hybrid AI Advantage™ with NVIDIA provides a trusted foundation to build, train, and scale with confidence—unlocking innovation and accelerating Smarter AI.

Lenovo and NVIDIA are driving the future of AI with faster, Smarter AI for All.

Read the NVIDIA Extreme co-design delivers new mlperf inference records blog or

See more NVIDIA performance data on the Data Center Deep Learning Product Performance Hub and Performance Explorer pages.

Author

Traci Parker is the Worldwide Solutions Marketing Manager for Enterprise IT and AI at Lenovo. She specializes in hybrid cloud, infrastructure modernization and AI solutions. She has more than 15 years of experience as a Marketing Manager and Product Marketing Manager across high-tech, fin-tech and healthcare industries.

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- Hybrid AI Factory
- MLPerf Benchmark
- ThinkSystem SR675 V3 Server
- ThinkSystem SR680a 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, LP2413, 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/LP2413
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2413.

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
Neptune®
ThinkSystem®

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