Lenovo ThinkSystem SR650 V4 with Intel Xeon 6: Proven AI Performance in MLPerf 5.1
Article

MLPerf is the industry-standard benchmark suite from MLCommons that provides objective and comparable performance metrics that allow organizations to assess hardware capabilities under standardized conditions. This paper presents an interpretation of the MLPerf 5.1 benchmark results for the Lenovo ThinkSystem SR650 V4, a data center-grade server powered by Intel Xeon 6 processors.

Figure 1. Lenovo ThinkSystem SR650 V4

The Lenovo results demonstrate the balanced performance of the ThinkSystem SR650 V4 across multiple domains: generative AI (Llama-3.1 8B), speech-to-text (Whisper), recommendation engines (DLRMv2), computer vision (RetinaNet), and graph analytics (rGAT). Each model's results are analyzed in terms of throughput, latency, and practical fit for real-world use cases.

The SR650 V4 not only proves versatile across multimodal AI workloads but also delivers competitive global standings in MLPerf 5.1:

- 1st place on DLRMv2-99.9 Server
- 2nd place on Llama-3.1 8B Server
- 3rd place on Llama-3.1 8B Offline
- 3rd place on RetinaNet Server
- 3rd place on rGAT Offline

These achievements highlight Lenovo's ability to provide data center solutions that combine high throughput, predictable latency, and enterprise-ready scalability, reinforcing the ThinkSystem SR650 V4 as a competitive choice for AI deployments at scale.

Cross-Model Summary & Comparison

This section provides a side-by-side comparison of the different models benchmarked on Lenovo ThinkSystem platforms. It highlights throughput, latency, and fit for use cases to provide a holistic view of model suitability.

Table 1. Cross-Model Summary & Comparison

Model	Key Metric	Highlights	Best Fit Use Cases
Llama-3.1 8B	p99.9 e2e Latency ~15s, TTFT ~2s TPOT ~113ms	Strong for generative tasks with long context	Chatbots, tutoring, customer support
Whisper	18.57 samples/sec	High transcription throughput	Speech-to-text, meeting transcription
DLRMv2	p99.9 Latency ~114ms, Throughput >12k	Extremely low latency and high throughput	Ad ranking, recommendation engines
RetinaNet	Server FPS ~375, Offline ~452	Real-time capable with good accuracy	Object detection in video streams, surveillance
RGAT	Throughput ~13.6k samples/sec	Handles graph workloads efficiently	Knowledge graph queries, fraud detection

Overall, the ThinkSystem platforms deliver balanced performance across diverse AI workloads. Llama-3.1 excels in language generation, Whisper in transcription, DLRMv2 in recommendation, RetinaNet in vision, and RGAT in graph workloads—demonstrating versatility of the system.

Verified MLPerf score of v5.1 Inference closed Llama3.1-8B, RetinaNet, DLRMV2 Server and Offline, rGAT and Whisper Offline. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/, Sep 2nd, 2025, entry 5.1-0063. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See mlcommons.org for more information.

Llama-3.1 8B

This section summarizes MLPerf 5.1 benchmark results for Llama-3.1 8B. Results reflect both server (real-time) and offline (batch) performance.

- ThinkSystem SR650 V4 – 2nd place on Llama3.1-8b Server
- ThinkSystem SR650 V4 – 3rd place on Llama3.1-8b Offline

Since MLPerf uses the CNN/DailyMail dataset, the input and output length assumptions are critical for interpreting throughput and latency results.

Benchmark I/O configuration:

- Average input length: ~870 tokens (CNN/DailyMail article text)
- Maximum output length: 128 tokens (fixed by MLPerf harness)

The figure below shows the input length distribution.

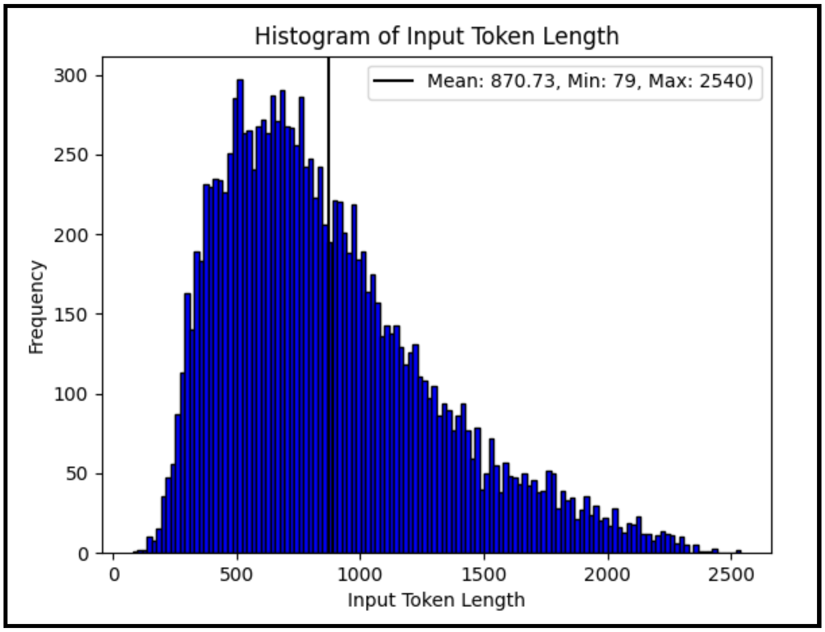

Figure 2. Llama-3.1 8B input length distribution

Server Use Case Sizing

The following table summarizes the various use cases with expected performance associated with them.

Table 2. Server Use Case Sizing

Use case	Output tokens (L)	p99.9 latency (s)	Sustained RPS	Concurrent sessions
Code completion / short replies	50	7.9	5.52	43
Chat assistant (concise)	80	11.3	3.44	38
RAG Q&A (typical)	120	15.8	2.29	36
Agent assist / support reply	150	19.2	1.84	35
Customer support (detailed)	200	24.9	1.37	34
Tutoring / explanations	300	36.2	0.91	33

The formulas to generate the above table are as follows:

- **P99.9 latency** = TTFT + (Output tokens L -1) * TPOT, where TTFT = 2.37s, TPOT = 113ms
- **Sustained RPS** = Sustained Token/s / Output Tokens(L), Where Sustained Tokens/s = 275.78 tokens/s
- **Concurrent Sessions** = P99.9 latency * Sustained RPS

Offline Use Case Sizing

The following table demonstrates the model capacity for each use cases of batch processing.

Table 3. Offline Use Case Sizing

Use case	Output tokens (L)	Items/sec	Items/hour	Items/8h	Items/24h
Code completion / short replies	50	15.54	55,933	447,465	1,342,395
Chat assistant (concise)	80	9.71	34,958	279,666	838,997
RAG Q&A (typical)	120	6.47	23,305	186,444	559,331
Agent assist / support reply	150	5.18	18,644	149,155	447,465
Customer support (detailed)	200	3.88	13,983	111,866	335,599
Tutoring / explanations	300	2.59	9,322	74,578	223,733

Key Takeaways – Llama3.1 8B

The key takeaways from these results are as follows:

- **Consistent SLA** – Predictable p99.9 latency, suitable for mission-critical apps.
- **Balanced Performance** – ~275 tok/s, TPOT = 113 ms, 33–43 concurrent users.
- **Scalable Use Cases** – Fast for short chats (~8s), practical for longer tasks (~36s).
- **Enterprise Efficiency** – Strong price-performance on Intel Xeon 6 ThinkSystem SR650 V4

Whisper

This section contains use-case tables and highlights for Whisper benchmark. Whisper is an advanced speech-to-text tool from OpenAI that can quickly turn spoken words into accurate written text. It works across many languages and is designed to handle real-world situations like different accents or background noise, making it a powerful solution for everyday transcription and translation needs.

Whisper Offline Use Case Sizing

The following table showcase the model performance under various use cases. We assume clips with an average length of 30-second.

Table 4. Whisper Offline Use Case Sizing

Use Case	Requirement	ThinkSystem Capability	Fit?
Live transcription (real-time)	≥1x RT (1 sec audio/sec)	~557 concurrent streams	□ Yes
Multi-stream transcription (broadcasts, meetings)	≥10 concurrent streams	~557 concurrent streams	□ Yes
Massive offline transcription (archive, call center logs)	1000+ hrs/day	~4,456.8 hr per 8h ~13,370.4 hr per 24h	□ Yes

The formula to generate the above table are as follows:

- **Concurrent Streams** = Sustained RPS * 30, where Sustained RPS = 18.57 sample/s

Key Takeaways – Whisper

The key takeaways from these results are as follows:

- **Real-Time Ready** – Supports ~557 concurrent 1× streams, ideal for live captioning and transcription.
- **High Offline Capacity** – Processes ~13,370 hours/day, fitting large-scale archives and compliance needs.

DLRMv2

This section contains use case tables, and highlights for DLRMv2. DLRMv2 (Deep Learning Recommendation Model v2) is a state-of-the-art model developed for powering recommendation systems, such as those used in e-commerce, ads, and content platforms. It efficiently processes both numerical and categorical data to deliver highly accurate, real-time personalized recommendations at scale.

- ThinkSystem SR650 V4 – 1st place on dlrm-v2-99.9 Server

DLRMv2 Use Case Sizing

The following table demonstrates the model capacity for each use cases of batch processing.

Table 5. DLRMv2 Use Case Sizing

Use case	Description	Latency Fit	Throughput Implication
E-commerce product recommendation	Suggest related items instantly when user views a product page.	<100 ms (✔) — fit with 114 ms p99.9. still practical in interactive setting	11.8K QPS supports tens of millions recommendations/day
News feed & content ranking	Rank posts, videos, or items for each session refresh.	<50–150 ms (✔)— fit with 114 ms p99.9	Supports tens of thousands of concurrent sessions
Personalized search (retail, media)	Tailor search results to user profile and catalog.	<50–150 ms (✔)— fit with 114 ms p99.9	Offline 11.9K samples/s enables billions of user-item pairs/day

The ThinkSystem V4 platform shows strong suitability for personalization workloads, covering both real-time (ads, e-commerce) and high-volume offline (catalog re-ranking) scenarios.

RetinaNet

This section contains use case tables and highlights for RetinaNet. RetinaNet is a deep learning model designed for object detection, capable of identifying and locating multiple objects within an image. Known for balancing speed and accuracy, it introduced the innovative “focal loss” technique, which makes it especially effective at detecting smaller or less frequent objects in real-world scenarios. It has been widely adopted in industries such as security, retail, healthcare, and autonomous driving, where reliable object detection is essential for video surveillance, inventory monitoring, medical imaging, and self-driving perception systems.

- ThinkSystem SR650 V4 – 3rd place on RetinaNet Server

RetinaNet Server Use Case Sizing

The following table show case the model fitness of live streaming use cases.

Table 6. RetinaNet Server Use Case Sizing

Use case	Target FPS	p99.9 latency (ms)	Fit Status
CCTV monitoring (low frame rate)	1 FPS	121 ms	□ Fit
Traffic camera (medium frame rate)	5 FPS	121 ms	□ Fit

RetinaNet Offline Use Case Sizing

The following table show case the model fitness of batch processing use cases.

Table 7. RetinaNet Offline Use Case Sizing

Use case	Scale	Throughput (images/s)	Frames per hour	Frames per day
City-wide traffic video archive	Large (24/7 cameras)	468.7	1,687,320	40,495,680
Retail chain CCTV backlog	Medium (hundreds of stores)	468.7	1,687,320	40,495,680
Warehouse incident review	Smaller (dozens of cameras)	468.7	1,687,320	40,495,680

The offline throughput shows that the ThinkSystem server can process over 40.5m images per day per server, enabling massive-scale video backlog analysis, incident detection, and compliance audit workloads.

RGAT

This section contains use case tables and highlights for RGAT. rGAT (relational Graph Attention Network) is a graph neural network model designed to capture relationships in complex, structured data by applying attention mechanisms across nodes and edges. This makes it especially powerful for tasks that require understanding connections, such as fraud detection, recommendation systems, and knowledge graph reasoning. It has been widely used in industries like finance, e-commerce, and social media, where uncovering hidden patterns and relationships is critical for decision-making and risk management.

- ThinkSystem SR650 V4 – 3rd place on rGAT Offline

RGAT Use Case Sizing (Offline)

Table 8. RGAT Use Case Sizing (Offline)

Use Case	Typical Requirement (Throughput)	RGAT Measured (Throughput)	Fit?
Fraud Detection (banking, payments)	≥ 5k txn/sec	13.6k	□ Fit
Recommendation Graphs (e-commerce, social)	≥ 10k items/sec	13.6k	□ Fit
Drug Discovery / Molecule Analysis	≥ 1k molecules/sec	13.6k	□ Fit
Knowledge Graph Completion	≥ 5k queries/sec	13.6k	□ Fit

The Lenovo system achieved 13.6k samples/sec offline throughput on the rGAT benchmark in MLPerf 5.1. This performance comfortably exceeds the throughput requirements across diverse real-world graph AI use cases such as fraud detection, recommendation systems, drug discovery, and knowledge graph completion.

Note: In MLPerf 5.1, Lenovo’s system achieved 13.6k samples/sec on rGAT in the offline benchmark. This demonstrates strong throughput capacity, comfortably above the requirements of common industry workloads like fraud detection (≥5k txn/sec) and recommendation graphs (≥10k items/sec). However, since MLPerf offline mode does not evaluate end-to-end latency, these results should be viewed as throughput potential under batch processing conditions. Real-time latency compliance requires additional validation.

Summary

The MLPerf 5.1 results for Lenovo ThinkSystem SR650 V4 with Intel Xeon 6 CPUs demonstrate a strong balance of throughput, latency, and efficiency across a wide range of AI models:

- Llama-3.1 8B provides consistent performance for generative tasks with predictable latencies
- Whisper delivers real-time transcription readiness at scale
- DLRMv2 achieves extremely low-latency, high-throughput personalized recommendations
- RetinaNet supports both live object detection and massive offline video analysis
- rGAT comfortably exceeds throughput requirements for graph-based workloads such as fraud detection and knowledge graph completion.

These outcomes reinforce that Lenovo's ThinkSystem platforms are not optimized for just one workload but can meet the demands of multimodal AI use cases. Importantly, while offline throughput results demonstrate impressive processing capacity, latency-sensitive scenarios require careful interpretation and, in some cases, additional validation in production environments.

Overall, the findings establish Lenovo ThinkSystem SR650 V4 as a versatile, enterprise-ready platform that can scale AI workloads efficiently while maintaining competitive price-performance ratios.

System Configuration and Software Environment

The following table lists the server configuration.

Table 9. 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:

- 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, LP2304, was created or updated on September 28, 2025.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2304
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2304.

Trademarks