Scaling Enterprise AI: High-Density CPU Inferencing with Lenovo ThinkSystem SR650 V4 and Intel Xeon 6

Solution Brief

Abstract

The Lenovo ThinkSystem SR650 V4, powered by Intel® Xeon® 6 processors, provides a scalable and cost-effective foundation for enterprise generative AI. Engineered to meet the performance demands of real-time AI workloads, the platform supports approximately 96 to 110 concurrent users per server while maintaining response times below 100 milliseconds. With sustained throughput exceeding 1,000 tokens per second and consistent performance across both bare-metal and containerized environments, the SR650 V4 enables enterprises to deploy high-density, CPU-only AI inferencing solutions that deliver fast, reliable, and responsive user experiences for business-critical applications.

The Enterprise AI Challenge

Generative AI is rapidly reshaping enterprise operations, enabling use cases such as intelligent customer engagement, content generation, and advanced analytics. As organizations move from experimentation to production-scale deployments, they face increasing pressure to support larger numbers of concurrent users while maintaining low-latency, real-time responsiveness. At the same time, infrastructure costs, power consumption, and operational complexity must be carefully controlled. Modern enterprise platforms must therefore deliver significantly higher performance and efficiency without relying exclusively on costly accelerator-based architectures.

Solution Architecture

This solution is built on the Lenovo ThinkSystem SR650 V4, purpose-designed to address the computational and operational demands of modern AI inferencing workloads.

Lenovo ThinkSystem SR650 V4
The SR650 V4 offers a flexible and high-density design, supporting up to 40 2.5-inch or 20 3.5-inch hot-swap drives. Advanced thermal technologies, including the Lenovo Neptune™ Processor Direct Water-Cooling module, combined with intelligent power and systems management through Lenovo XClarity™, enable sustained performance under demanding AI workloads.

Intel Xeon 6 Processors
Intel Xeon 6 processors are optimized for AI inference through built-in Intel® Advanced Matrix Extensions (AMX), accelerating matrix operations without the need for discrete GPUs. Enhanced memory bandwidth and capacity support the high-throughput, low-latency requirements of production generative AI environments.

Optimized for Right-Sized Models up to 20B Parameters
The architecture is tuned for efficient CPU-only inferencing of right-sized models, such as Meta Llama 3.1 8B. This approach enables high user density while avoiding the cost, power, and operational overhead associated with GPU-based infrastructures.

Performance Methodology and Results

Testing utilized the vLLM serving framework with Meta LLAMA 3.1 8B models, focusing on a strict Service Level Agreement (SLA) of **100ms or less time per output token (TPOT)**.

Metric	Intel Xeon 6745P (32 Cores)	Intel Xeon 6767P (64 Cores)
Peak Throughput	934 tokens/sec	1,089 tokens/sec
Concurrent User Capacity	96 prompts	110 prompts
Bare Metal vs. RHOS Variance	< 4%	Negligible

Table 1 Comparison: SR650 V4 with Xeon 6745P (32 cores) vs. Xeon 6767P (64 cores)

The Intel Xeon 6767P delivers up to 15 percent higher concurrent user capacity compared to the 6745P while sustaining strong performance at larger input and output lengths. Even at 1024 and 2048 token workloads, the platform supports more than 50 concurrent users at throughput exceeding 500 tokens per second.

Enterprise Value Proposition

Infrastructure Efficiency: Supporting up to 110 concurrent users per server reduces capital expenditures, data center footprint, and overall infrastructure complexity.

Deployment Flexibility: Minimal performance differences between bare-metal and Red Hat OpenShift deployments allow organizations to align AI deployments with existing security, orchestration, and operational policies.

Lower Total Cost of Ownership: High per-server performance, energy-efficient processors, and simplified management reduce power consumption and operational costs while improving utilization.

Why Lenovo and Intel

The performance advantages of this solution are driven by deep co-engineering between Lenovo and Intel. Joint optimization of memory subsystems, I/O architecture, and thermal design ensures that Intel Xeon 6 processor capabilities, including AMX acceleration, are fully realized. Lenovo Neptune liquid cooling further enables sustained performance under continuous AI workloads, delivering higher throughput and efficiency than generic server implementations.

Conclusion

The Lenovo ThinkSystem SR650 V4, powered by Intel Xeon 6 processors, provides a high-performance foundation for scaling generative AI in the enterprise. In independent testing using Meta Llama 3.1 8B models, the platform achieved breakthrough throughput exceeding 1,000 tokens per second and supported between 96 and 110 concurrent users while maintaining a sub-100ms response time. This performance

remains consistent across diverse deployment models, showing less than a 4% variance between bare-metal installations and Red Hat OpenShift containerized environments.

By leveraging built-in Intel Advanced Matrix Extensions (AMX) for AI acceleration, the SR650 V4 enables high-density inferencing models up to 20B parameters without the need for dedicated GPUs. This CPU-centric approach significantly reduces infrastructure complexity and total cost of ownership. Ultimately, the solution allows organizations to serve more users with fewer servers, providing a future-ready platform that scales reliably with evolving AI demands.

Bill of Materials

Part Number	Product Description	Qty
7DGDCTO1WW	Server: ThinkSystem SR650 V4-3yr Base Warranty	1
C3QK	ThinkSystem SR650 V4 24x2.5" Chassis	1
CARA	Intel Xeon 6745P 32C	2
C3QR	ThinkSystem 2U V4 Performance Heatsink	2
C0U9	ThinkSystem 32GB TruDDR5 6400MHz (1Rx4) RDIMM	16
BGM1	ThinkSystem RAID 940-8i 4GB Flash PCIe Gen4 12Gb Adapter for U.3	1
B9XC	Controller 1 HW RAID Array 1 RAID 0	1
C2BW	ThinkSystem 2.5" U.3 7500 MAX 3.2TB Mixed Use NVMe PCIe 4.0 x4 HS SSD	8
C3RU	ThinkSystem 2U V4 8x2.5" AnyBay Backplane	1
C0JJ	ThinkSystem M.2 RAID B540p-2HS SATA/NVMe Adapter	1
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
B5T1	ThinkSystem Broadcom 5719 1GbE RJ45 4-port OCP Ethernet Adapter	1
BN2T	ThinkSystem Broadcom 57414 10/25GbE SFP28 2-Port OCP Ethernet Adapter	1
BK1J	ThinkSystem Broadcom 57508 100GbE QSFP56 2-Port PCIe 4 Ethernet Adapter	1
C0UB	ThinkSystem 2700W 230V Platinum CRPS Hot-Swap Power Supply v2.4	2
C3RD	ThinkSystem 2U 6056 20K Performance Fan Module	6
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1

Table 2. Bill of Materials

Why Lenovo

Lenovo is a US$70 billion revenue Fortune Global 500 company serving customers in 180 markets around the world. Focused on a bold vision to deliver smarter technology for all, we are developing world-changing technologies that power (through devices and infrastructure) and empower (through solutions, services and software) millions of customers every day.

For More Information

To learn more about this Lenovo solution contact your Lenovo Business
https://www.lenovo.com/au/en/c/servers-storage/servers/racks/

References:

Lenovo ThinkSystem SR650 V4:
https://www.lenovo.com/au/en/p/servers-storage/servers/racks/lenovo-thinksystem-sr650-v4/len21ts0042

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.

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 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.