Accelerate Agentic AI with Lenovo ThinkSystem SR650 V4 and Intel Xeon 6 Processors

Solution Brief

Deliver up to 2.4x higher AI throughput at significantly lower infrastructure cost with fewer servers running on Red Hat OpenShift.

Enterprise IT leaders are under increasing pressure to support rapidly expanding AI workloads while keeping infrastructure costs under control. As organizations move beyond experimentation and into production-scale deployment of Generative AI (GenAI), the infrastructure decisions they make today will define their competitive position for years to come.

Agentic AI is a class of AI systems capable of autonomously executing multi-step workflows, reasoning over documents, and generating actionable insights. It represents one of the most transformative and compute-intensive enterprise use cases. Workloads such as automated document summarization, contract analysis, and intelligent process automation demand both high throughput and consistent responsiveness as concurrent user loads scale across the organization.

This solution brief examines how the Lenovo ThinkSystem SR650 V4, powered by Intel Xeon 6 processors and deployed on a Red Hat OpenShift cluster, delivers a validated, high-performance platform for agentic AI at enterprise scale. The findings are drawn from rigorous performance testing of end-to-end document summarization workloads - a real-world proxy for the types of GenAI tasks organizations are deploying today.

The Challenge

Agentic AI is transforming enterprise operations by enabling the automation of complex workflows and generating actionable insights that improve productivity and accelerate innovation.

However, deploying GenAI at scale often presents significant infrastructure challenges:

- Infrastructure efficiency - scaling AI workloads without proportionally scaling hardware spend
- Cost optimization - managing rising expenses as compute demands fluctuate
- Operational complexity - maintaining reliable support for production AI applications
- Performance consistency - ensuring responsiveness as concurrent user loads increase

Enterprises need a platform that delivers superior AI throughput with fewer servers - reducing cost, complexity, and the carbon footprint of production AI deployments.

Organizations evaluating on-premises AI infrastructure face a difficult tradeoff: deploying too little hardware risks performance degradation during peak loads, while over-provisioning drives up capital and operational expenditure. Previous-generation server platforms often required significant scale-out to meet throughput targets, multiplying rack space, power draw, cooling costs, and management complexity.

The Solution

Lenovo ThinkSystem SR650 V4 servers with Intel Xeon 6 processors deliver a purpose-built platform for enterprise AI workloads — combining cutting-edge silicon performance with Lenovo's proven server engineering.

Figure 1. Lenovo ThinkSystem SR650 V4

Deployed as a Red Hat OpenShift cluster, this solution enables organizations to run agentic document summarization and other GenAI workflows with outstanding efficiency. Intel's latest generation Xeon 6 processors deliver significant gains through:

- Faster cores with higher clock speeds for demanding AI inference tasks
- Larger cache capacity and increased memory bandwidth for large model throughput
- DDR5 support and high-bandwidth memory capabilities
- Improved performance-per-watt to control power and cooling costs

Performance Results

In validated testing, two Lenovo ThinkSystem SR650 V4 servers with Intel Xeon 6 processors consistently outperformed four SR650 V3 servers with 5th Gen Intel Xeon processors across every user load tested from 32 to 128 concurrent users. This means IT organizations can achieve superior AI performance while cutting server count in half.

Table 1. Test Configurations

Specification	Detail
Workload	Agentic document summarization (end-to-end)
Platform (New)	2x Lenovo ThinkSystem SR650 V4 - Intel Xeon 6 processors
Platform (Previous)	4x Lenovo ThinkSystem SR650 V3 - 5th Gen Intel Xeon Platinum
Deployment	Red Hat OpenShift cluster
Concurrent Users Tested	32, 64, 96, and 128 simultaneous users
Metric	Normalized throughput (requests per second)

The figure below shows normalized throughput (requests per second) for an agentic document summarization workload on a Red Hat OpenShift cluster, as concurrent users requesting summaries scale from 32 to 128.

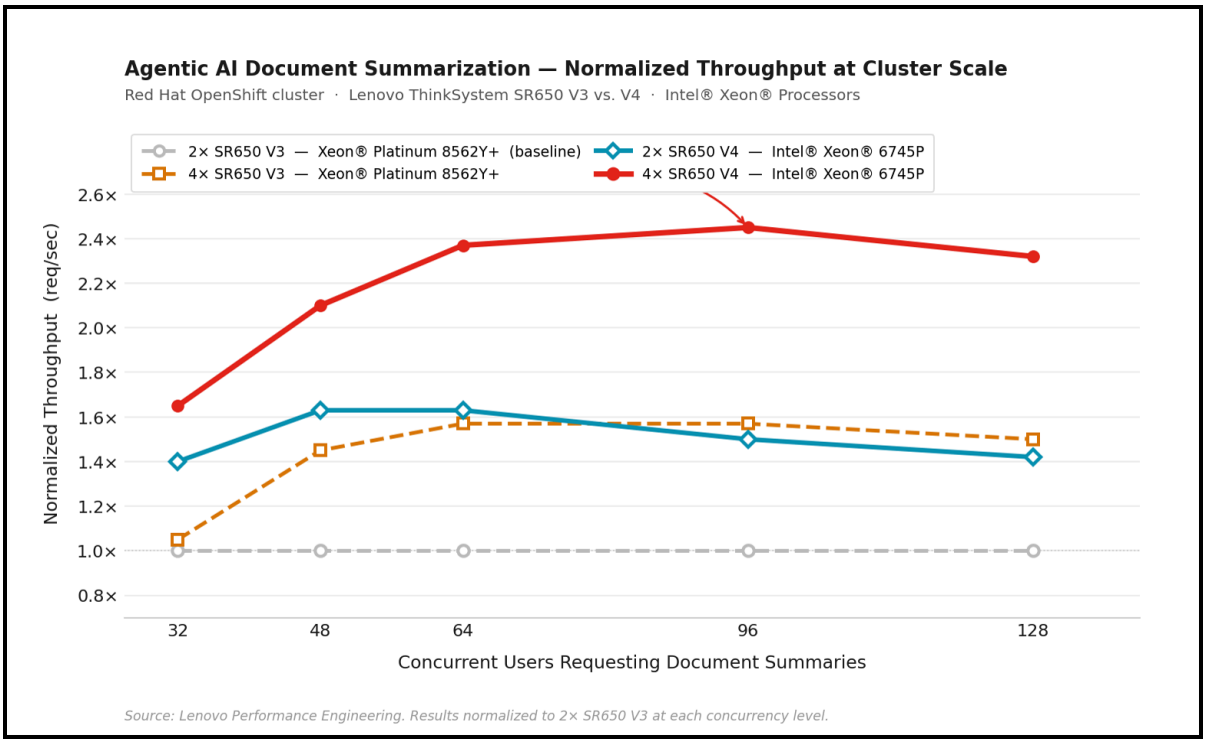

Figure 2. Normalized throughput for end-to-end document summarization at cluster scale.

Throughput is normalized against a baseline of two SR650 V3 servers with 5th Gen Intel Xeon Platinum 8562Y+ processors (1.0x). Four SR650 V3 servers with the same processors reach roughly 1.6x throughput at peak. By comparison, two Lenovo ThinkSystem SR650 V4 servers with Intel Xeon 6745P processors deliver up to ~1.7x throughput, and four SR650 V4 servers scale to as much as 2.4x — demonstrating substantial throughput gains and improved cost efficiency per server with the new platform.

Key Findings

Two SR650 V4 servers with Intel Xeon 6 processors consistently outperformed four SR650 V3 servers across every user load tested, from 32 to 128 concurrent users:

Two SR650 V4 servers with Intel Xeon 6 outperformed **four SR650 V3 servers** at every load point tested - delivering 2.4x higher throughput with half the hardware footprint.

- Up to 2.4x higher normalized throughput (requests per second) for end-to-end document summarization at cluster scale
- Superior performance achieved with 50% fewer physical servers - 2 SR650 V4 units versus 4 SR650 V3 units
- Consistent throughput advantage maintained across all concurrent user loads tested (32-128 users)
- Significantly lower infrastructure cost for each unit of AI processing capacity delivered.

Why Lenovo ThinkSystem SR650 V4

The Lenovo ThinkSystem SR650 V4 is a purpose-engineered 2U rack server that combines Intel's latest silicon advances with Lenovo's proven server engineering - creating a platform optimized for the demands of enterprise AI at scale.

Intel's latest generation Xeon 6 processors deliver significant architectural gains that translate directly into AI inference performance:

- Faster cores with higher clock speeds purpose-built for demanding AI inference tasks
- Larger cache capacity and increased memory bandwidth to sustain throughput for large language model inferencing
- DDR5 and high-bandwidth memory support enabling faster data movement between CPU and memory
- Improved performance-per-watt to control power and cooling costs at data center scale

Advanced Cooling Options

For customers seeking greater efficiency than traditional air cooling, the Lenovo Processor Neptune Core Module offers an open-loop liquid cooling alternative. By removing processor heat more effectively, it reduces reliance on server fans and data center air conditioning, lowering energy consumption and operating costs. Depending on the configuration, organizations can achieve up to 63% fan power savings per node and nearly 25% rack-level power savings.

Simplified Infrastructure Management

Lenovo XClarity complements the SR650 V4 with a unified infrastructure management platform that leverages AI-driven monitoring and automation to reduce downtime and streamline operations. With centralized visibility, Zero Trust security, support for open standards, and flexible cloud or on-premises deployment options, XClarity enables organizations to manage and scale their infrastructure with greater efficiency, security, and operational simplicity.

The following table lists the specifications of the ThinkSystem SR650 V4.

Table 2. SR650 V4 specifications

Specification	Detail
Form Factor	2U rack server
Processors	Up to 2x Intel Xeon 6 processors, up to 86 cores and 350W TDP
Memory	Up to 32x TruDDR5 3DS/RDIMMs up to 6400MHz; up to 16x MRDIMMs up to 8000MHz; max 8TB
Storage	Up to 24x 2.5-inch NVMe, or 32x E3.S NVMe, or 12x 3.5-inch SAS/SATA; 2x M.2 boot
Expansion	Up to 10x PCIe Gen5 slots + 2x OCP 3.0 slots
GPUs	Up to 10x single-width or 2x double-width GPUs
Cooling	Lenovo Neptune Core and Air - optional liquid cooling for high-TDP configs
Power	Dual redundant AC & DC; Platinum & Titanium efficiency ratings
Management	Lenovo XClarity Controller with AI-predictive analytics and remote access
OS Support	Red Hat, SUSE, Microsoft, VMware

Key Solution Benefits

The key benefits of the SR650 V4 platform include the following:

- **Superior Throughput**

  2.4x higher AI request throughput versus prior-generation platform, sustained across all concurrency levels tested

- **Fewer Servers**

  Achieve production AI throughput targets with 50% fewer physical servers, reducing rack space, cabling, and per-node management overhead

- **Lower Total Cost**

  Significantly reduced infrastructure cost per unit of AI throughput; fewer servers means lower CapEx, power, and cooling spend

- **Scalability**

  Performance advantage holds and grows as concurrent user loads scale from 32 to 128+ users - validated on Red Hat OpenShift

- **Sustainability**

  Lenovo Neptune cooling and ENERGY STAR-compliant designs enable up to 3:1 rack consolidation and reduced PUE impact

- **Enterprise Management**

  XClarity One provides centralized control, SSD failure prediction, and federated directory for simplified operations

- **Future-Ready**

  PCIe Gen5, CXL 2.0, MRDIMM, and up to 8TB memory support protect infrastructure investments against emerging AI workload requirements

Business Outcomes

Organizations that deploy agentic AI on the ThinkSystem SR650 V4 realize measurable improvements across three dimensions: performance, economics, and operational efficiency.

- **Higher Employee Productivity**

  Automating document-intensive workflows - summarization, contract review, compliance checking - frees knowledge workers from time-consuming manual tasks. Requests that previously required 15-30 minutes of human effort are completed in seconds at scale.

- **Faster Time-to-Insight**

  With 2.4x higher throughput, AI inference results are returned faster across all concurrent users. Teams acting on AI-generated insights move more quickly from analysis to decision.

- **Reduced Infrastructure CapEx**

  Achieving the same or better throughput with 50% fewer servers translates directly to lower hardware acquisition costs. A deployment that previously required 4 servers requires only 2, cutting procurement, rack, and cabling costs in half.

- **Lower Operational Costs**

  Fewer physical servers mean reduced power consumption, cooling load, and software licensing fees. Lenovo Neptune thermal efficiency and ENERGY STAR compliance further reduce per-server energy costs.

- **Improved IT Agility**

  A smaller, higher-performing cluster is easier to manage, patch, and scale. XClarity One provides AI-predictive management, reducing mean time to resolution and lowering IT labor costs.

- **Faster ROI on AI Investment**

  By consolidating AI infrastructure onto fewer, more capable servers, organizations accelerate payback on their AI platform investment - freeing savings to fund additional AI use cases.

- **Future-Proof Platform**

  PCIe Gen5, CXL 2.0, MRDIMM support, and up to 8TB memory capacity accommodate next-generation AI models and expanding workload requirements without a full infrastructure refresh.

Customer Use Cases

The ThinkSystem SR650 V4 is validated for production-scale agentic AI deployments across industries where document intelligence, workflow automation, and real-time AI reasoning deliver measurable business value.

From financial services to healthcare to manufacturing, the SR650 V4 delivers the throughput, scalability, and on-premises security posture that enterprise agentic AI demands.

Table 3. Customer Use Cases

Industry/ Use Case	**How SR650 V4 Delivers Value**
Financial Services Automated Contract & Document Review	Banks, insurers, and asset managers autonomously review, summarize, and flag risk clauses across thousands of contracts and filings daily - reducing review cycles from days to minutes with 50% fewer servers.
Healthcare Clinical Documentation & Records Summarization	Synthesize patient records, surface relevant clinical history, and generate structured care-team summaries at scale - on-premises to meet data sovereignty and HIPAA compliance requirements.
Legal Intelligent Discovery & Case Preparation	Rapidly surface relevant documents, generate case summaries, and identify precedents across large document sets - supporting 128+ concurrent users without performance degradation under deadline pressure.
Manufacturing & Supply Chain Operational Intelligence	Automate extraction of key terms from supplier contracts, technical specs, and regulatory documents; flag supply chain risks and generate operational summaries while keeping sensitive IP on-premises.
Public Sector Policy Research & Regulatory Compliance	Deploy agentic AI within air-gapped or controlled network environments to summarize policy documents, procurement records, and regulatory filings without exposing government data to public cloud.
Cross-Industry Enterprise Knowledge Management	Create intelligent interfaces over internal knowledge repositories - product docs, HR policies, runbooks - giving employees instant, authoritative summaries and reducing time spent searching.

Conclusion

The results are clear: the Lenovo ThinkSystem SR650 V4 with Intel Xeon 6 processors delivers outstanding throughput and cost efficiency for agentic AI workloads at enterprise scale. By achieving up to 2.4x higher throughput with half the server footprint compared to the previous generation, organizations can scale AI deployments more effectively, reduce infrastructure costs, and bring transformative agentic AI capabilities to production faster. Deployed on Red Hat OpenShift, the SR650 V4 provides a validated, enterprise-grade foundation for the full spectrum of GenAI workloads - from document summarization and contract analysis today, to more complex autonomous agent workflows as AI capabilities continue to evolve.

For organizations evaluating infrastructure investments for AI at scale, the ThinkSystem SR650 V4 platform represents a compelling path to achieving more with fewer resources: higher performance, lower cost, and a simpler operational footprint - delivered by the trusted combination of Lenovo engineering and Intel Xeon 6 silicon.

Ready to accelerate your agentic AI deployment? Contact your Lenovo representative or go to the ThinkSystem SR650 V4 product web page to learn more about the server.

Author

Chris Honoré is a Solutions Product Manager at Lenovo with deep expertise in datacenter products and solution offerings. He has a strong background in consulting and solution development, helping customers design and support on-premises and hybrid environments. Chris has spent the past 15 years with IBM and Lenovo, specializing in x86 server and data center solutions. Prior to that, he built two decades of experience in the telecommunications industry, serving in both technical and business leadership roles.

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, LP2455, was created or updated on June 24, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2455
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2455.

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®
Neptune®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

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

Microsoft is a trademark of Microsoft Corporation in the United States, other countries, or both.

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