Enterprise AI Inference on Lenovo ThinkSystem V4
Solution Brief

The Lenovo ThinkSystem V4 servers, powered by Intel® Xeon® 6 (Granite Rapids) processors and Red Hat AI 3.4, delivers a validated and production ready enterprise AI inference platform that eliminates the GPU or nothing paradigm for the majority of real world AI workloads.

With the Q2 2026 commercial launch of Red Hat AI 3.4, the first release to treat Intel Xeon as a fully supported and first class inference target, enterprises can deploy high performance LLM serving, Retrieval Augmented Generation, and agentic AI pipelines on CPU infrastructure with zero code changes, up to 80% lower total cost of ownership compared to GPU alternatives for small language models at moderate concurrency, and the identical OpenShift AI operational experience whether running on Xeon or GPU nodes.

Built on Xeon 6 purpose built AI acceleration including Advanced Matrix Extensions for 3x higher BF16 and INT8 throughput, MRDIMM support for 37% greater memory bandwidth, and Intel TDX for confidential in memory data protection, and further enhanced by the llm-d KV cache aware scheduling framework that delivers up to 57x faster time to first token for multi turn agentic workloads, the ThinkSystem V4 servers provides a scalable and right sized foundation that organizations can deploy today and grow incrementally from CPU only inference through hybrid CPU plus GPU environments as workload demands increase.

3x **AI Throughput** Higher BF16 and INT8 throughput via Advanced Matrix Extensions vs. prior Xeon generation. No separate accelerator required.	**80%** **TCO Savings** Lower total cost of ownership for small language models at moderate concurrency compared to GPU-based alternatives.
57x **Faster TTFT** Improvement in time to first token for multi-turn agentic workloads using llm-d KV cache-aware scheduling.	**37%** **Memory Bandwidth** Greater memory bandwidth with Multiplexed Rank DIMM support, resolving KV cache bottlenecks in LLM inference.

Red Hat AI 3.4: Full Commercial Xeon Support

The Q2 2026 release of Red Hat AI 3.4 delivers the industry's first fully commercial, enterprise grade AI platform treating Intel Xeon as a first class inference target on par with GPU nodes. Customers running on Lenovo ThinkSystem V4 servers experience the identical Red Hat OpenShift AI deployment workflow on Xeon as on GPU, with no separate toolchain, no code changes, and full support for AVX2, AVX-512, and AMX instruction sets from day one.

- Unified control plane: auto-scaling, RBAC, audit logging, and model governance for Xeon and GPU nodes alike
- Quickstarts catalog: ready to run, production optimized examples for LLM serving, RAG, Agentic AI, AI Virtual Agents, Guardrail LLMs, and edge inference at docs.redhat.com/en/learn/ai-quickstarts
- Validated model catalog on Hugging Face: Llama 3.1 and 4, Granite 3.1 and 3.2, Qwen 2.5 and 3, Phi-4, Gemma 2, Mistral 7B, plus embedding models for RAG
- Phased adoption path: start with CPU only inference, progress to RAG pipelines, then scale to hybrid CPU plus GPU as concurrency demands grow

Intel Xeon 6 (Granite Rapids): Purpose Built for AI Inference

Intel Xeon 6 is the first Xeon generation explicitly architected for datacenter scale AI inference. Every capability below is built directly into the processor with no separate accelerator required and is fully exposed through the Red Hat AI 3.4 and vLLM software stack on the ThinkSystem V4 servers.

Capability	Impact on AI Inference
Advanced Matrix Extensions (AMX)	3x higher throughput for BF16 and INT8.[1] No separate accelerator needed. 2.8x faster TPOT for Llama 3.1 8B vs. AMD EPYC 9755.[2]
Priority Core Turbo (PCT)	17% frequency boost on designated cores for sequential processing, data validation, and GPU feeding in hybrid clusters.[3]
MRDIMM Memory Support	37% higher bandwidth vs. standard RDIMMs. Resolves KV cache bottlenecks that limit LLM inference throughput.[4]
Up to 128 cores and 500 MB LLC[1]	Massive on-die cache reduces memory latency for autoregressive token generation. Supports up to 96 concurrent users per dual-socket node.
Intel Trust Domain Extensions (TDX)	VM level hardware encryption isolates AI models and user data from the hypervisor. Enables confidential inference for regulated industries.[5]

Strategic Fit and Key Use Cases

For the dominant enterprise AI tier, small language models in the 7B to 20B parameter range at moderate concurrency of 10 to 50 users, Xeon 6 on the ThinkSystem V4 servers delivers performance within SLO at up to 80% lower TCO than GPU based alternatives. Four validated use cases drive the majority of enterprise deployments:

Retrieval Augmented Generation (RAG)

Embedding generation, vector search, and reranking are CPU friendly workloads. Xeon 6 PCIe 5.0 NVMe throughput accelerates document ingestion and retrieval, delivering 2.7x higher embedding throughput vs. prior Xeon generation. Validated models include Granite-embedding-278m-multilingual and all-MiniLM-L6-v2.

Agentic AI with llm-d

Sequential agent loops are latency bound rather than throughput bound. The llm-d framework, included in Red Hat AI 3.4, uses KV cache aware scheduling to route multi-turn requests to pods where the prior conversation cache already resides, eliminating redundant prefill computation and delivering up to 57x faster time to first token on cache hit.

Confidential AI Inference

Intel TDX on the ThinkSystem V4 servers isolates AI models and user queries within hardware encrypted Trust Domains, protecting them from unauthorized access including at the hypervisor layer. For hybrid confidential GPU inference, TDX combines with the Nvidia encrypted bounce buffer approach to protect data transfers between the confidential VM and GPU. This pattern requires OpenShift Sandboxed Containers, Red Hat build of Trustee for attestation, and Kata Containers for TDX-protected VMs. Expect modest performance tradeoffs due to encryption and attestation overhead, which should be evaluated during design. Critical for healthcare, financial services, and sovereign AI deployments requiring data-in-use protection.

Hybrid CPU plus GPU Inference

The ThinkSystem V4 servers supports PCIe 5.0 GPU expansion, enabling Xeon to handle orchestration, small model serving under 20B parameters, and guardrail workloads while GPU nodes serve larger models. A semantic router via llm-d dynamically directs requests to the optimal backend, improving overall GPU utilization and reducing queue contention.

Reference Architecture and Ecosystem

The validated architecture deploys Red Hat OpenShift Container Platform and Red Hat OpenShift AI on Lenovo ThinkSystem V4 servers nodes with Intel Xeon 6, supporting three incremental deployment tiers: CPU only inference for immediate value, hybrid CPU plus GPU for scale, and Intel TDX confidential computing for security sensitive workloads.

Component	Detail
Lenovo ThinkSystem V4 servers	2U dual-socket, up to 128 cores per socket, up to 8TB DDR5 with MRDIMM, 32x PCIe 5.0 NVMe, optional GPU expansion, Intel TDX, XCC3 management
Red Hat OpenShift AI 3.4	Q2 2026 GA. Full Xeon AMX support. Unified lifecycle management, auto-scaling, RBAC, model governance.
vLLM with CPU AMX backend	Upstream merged CPU support. Zero code changes for standard LLM inference workloads.
llm-d Inference Scheduler	KV cache aware distributed scheduling. 57x TTFT improvement for multi-turn agentic workloads.
Red Hat OpenShift Data Foundation	Persistent storage for RAG vector stores, model registry, and embedding databases.
Intel TDX plus Red Hat Trustee	Confidential computing attestation. Protects model weights and user data in memory.

Conclusion

The convergence of Lenovo ThinkSystem V4 servers hardware, Intel Xeon 6 processor capabilities, and Red Hat AI 3.4 platform support represents a decisive step toward democratizing production AI across the enterprise. Organizations no longer need to wait for GPU availability, absorb GPU level infrastructure costs, or accept operational complexity to run meaningful AI workloads at scale. With validated performance across small language models, RAG pipelines, agentic workflows, and confidential computing use cases, this solution provides an immediate, globally available, and cost effective path to production AI that grows with the business. The phased adoption model ensures that investments made today in CPU based inference infrastructure remain relevant and expandable as workload demands evolve toward hybrid CPU and GPU environments. For enterprise teams prioritizing speed to value, total cost of ownership, data sovereignty, and operational simplicity, the Lenovo ThinkSystem V4 servers with Red Hat AI 3.4 and Intel Xeon 6 is the enterprise AI platform built for where AI is going, not just where it has been.

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 sales representative or authorized channel partner to:
https://www.lenovo.com/au/en/c/servers-storage/servers/racks/

- Lenovo Representative: Lenovo sales representative or authorized channel partner

References:

- Lenovo ThinkSystem SR650 V4:
  - Lenovo ThinkSystem SR650 V4
- [1] Intel Corporation. "Intel® Xeon® 6 Product Brief." Intel, 2024.
- [2] Intel Corporation. "Accelerating vLLM Inference: Intel Xeon 6 Processor Advantage over AMD EPYC." Intel Community Blog, 2024.
- [3] Intel Newsroom. "New Intel Xeon 6 CPUs to Maximize GPU-Accelerated AI Performance." Intel Corporation, 2025.
- [4] Intel Newsroom. "New Ultrafast Memory Boosts Intel Data Center Chips." Intel Corporation, 2024.
- [5] Intel Corporation. "Intel® Trust Domain Extensions (Intel® TDX) Overview." Intel, 2024.
  - Veeam Ready Repository – Primary Target. Backup Target - Disk iSCSI
- [6] Red Hat and Intel. "AI on Intel® Xeon® Processors with Red Hat OpenShift AI — Reference Architecture for Scalable Inference and Agentic AI on Intel® Xeon® 6." Red Hat / Intel Reference Guide, May 2026.

Authors

Chris Honore 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 datacenter solutions. Prior to that, he built two decades of experience in the telecommunications industry, serving in both technical and business leadership roles.

Related product families

Product families related to this document are the following:

- AI Servers
- 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, LP2435, was created or updated on May 11, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2435
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2435.

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:

AMD and AMD EPYC™ are trademarks of Advanced Micro Devices, Inc.

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.