The Business Value of CPU-Based Inference: Why Enterprises Need More Than One Inference Strategy

Article

GPUs remain essential for training and the most demanding inference and agentic AI workloads, but they are not the only path to production. Modern CPU platforms can support a wide range of enterprise AI applications while improving cost predictability, infrastructure utilization, and data control to:

- **Lower your long-term AI infrastructure cost.** For sustained LLM inference workloads, an on-prem Lenovo ThinkSystem SR650 V4 with Intel Xeon 6 can become more economical than cloud rental in as little as 4.5 months.
- **Reduce 5-year infrastructure spend by over $1M/$1.02M in net savings** over five years versus comparable cloud rental costs.
- **Offset infrastructure investment with utilization above 3.6 hours per day** can make on-prem more economical than renting cloud capacity.
- **Lower your cost per million tokens by approximately6.7× lower cost per million tokens** compared with the cloud helping make enterprise-scale AI usage more affordable.
- **Run practical enterprise LLM workloads without a GPU server.** Common use cases such as internal chatbots, knowledge assistants, document summarization, long-form generation, and RAG-style question answering can run effectively on CPU-only infrastructure.
- **Keep sensitive data under your control.** On-prem inference helps keep prompts, documents, embeddings, and generated responses inside your own environment, supporting data governance, compliance, and privacy requirements.
- **Avoid overbuying GPU capacity.** Not every LLM workload requires a high-cost GPU server. For smaller and mid-sized models, predictable internal workloads, and moderate-concurrency serving, Intel Xeon 6 CPUs provide a right-sized alternative.
- **Gain predictable capacity and predictable cost.** Instead of paying ongoing cloud rental fees, you can deploy dedicated inference capacity with clearer cost planning, better workload control, and reduced dependency on cloud availability or quota limits.
- **Support both batch and serving use cases.** The benchmark results show strong offline output throughput as well as practical online serving performance across short chat, standard chat, long generation, and RAG-like workloads.

Today, organizations face a different challenge: turning AI from experimental technology into a scalable business capability that's efficient, customer focused and faster time to revenue. This shift is forcing technology leaders to evaluate AI infrastructure through a broader lens. Performance remains important, but it is no longer the only consideration. CIOs, infrastructure teams, and AI platform leaders must also account for utilization rates, governance requirements, operational complexity, and the long-term cost of serving inference workloads at scale.

The reality is that not every AI application requires the highest-performance accelerator available. Predictable costs, reliable response times, and straightforward deployment models create opportunities

where CPU-based inference provides a practical and economically attractive alternative.

The Economics of Production AI

Cloud infrastructure has played a critical role in accelerating AI adoption. It provides rapid access to compute resources, simplifies experimentation, and allows organizations to deploy new applications without significant upfront investment.

However, the economics of AI can change dramatically once workloads become persistent. For example, a pilot project may serve a small group of users for a limited period of time while a production deployment may support thousands of users, generate millions of tokens per day, and run continuously throughout the year. Unpredictable and recurring consumption costs become a strategic consideration rather than a line-item expense.

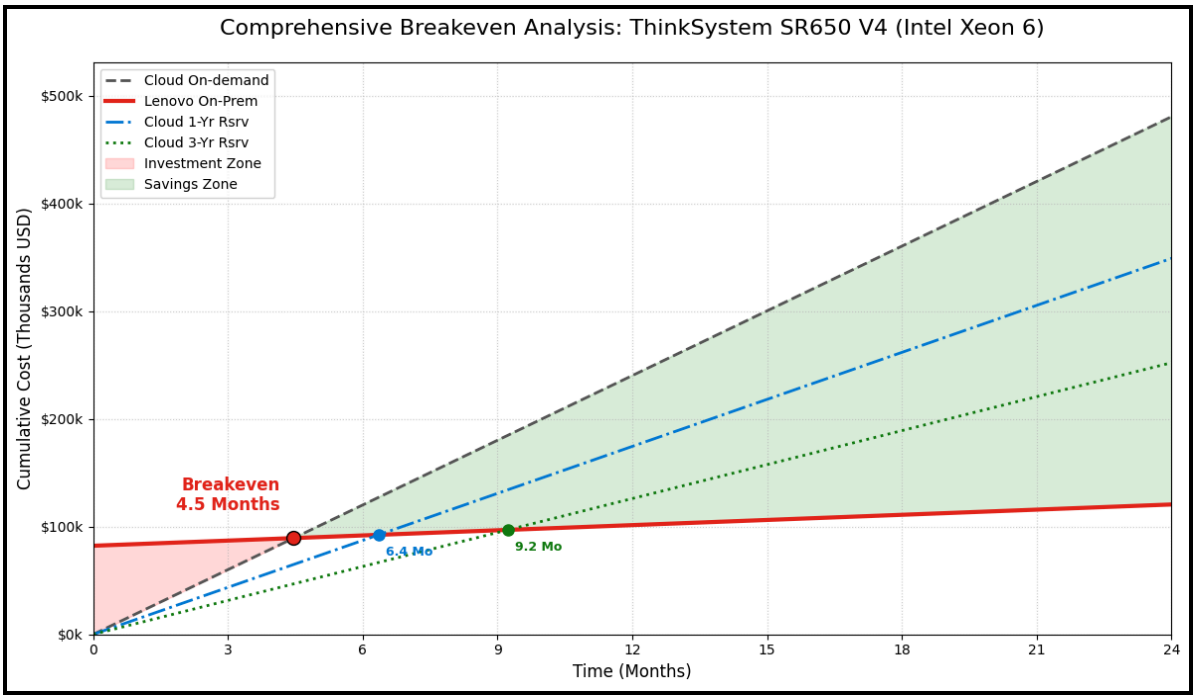

Figure 1. Using the on-prem system for more than 4.5 months becomes more economical than the cloud

Analysis of Lenovo ThinkSystem SR650 V4 servers, powered by Intel Xeon 6 processors indicates that organizations with sustained inference demand can achieve breakeven with comparable cloud infrastructure in less than five months. That's how quickly the economics of ownership can become favorable.

The financial impact becomes even more apparent when infrastructure is evaluated over its full operational lifecycle.

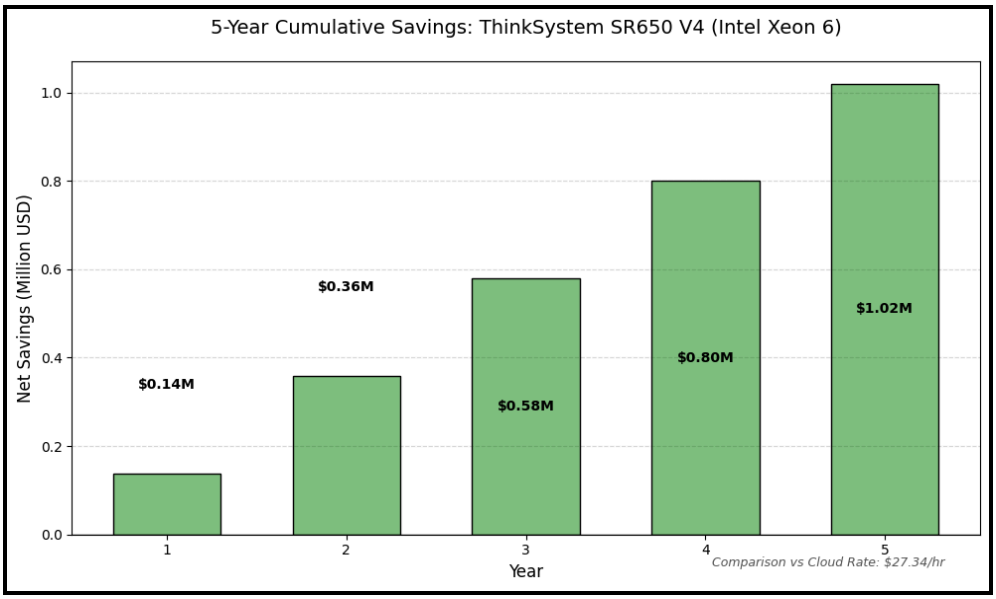

Figure 2. Over a 5 year lifecycle, enterprises end up saving over a million dollars compared to renting the cloud

After accounting for power, cooling, maintenance, and colocation costs, the modeled deployment generated approximately $1.02 million in net savings over five years compared with a cloud-based alternative.

These savings are significant because they can be reinvested into additional AI initiatives that focus on customer experience and faster time to revenue. Lower infrastructure costs create opportunities to support more concurrent users, deploy more applications, and expand AI adoption across business functions without proportionally increasing costs.

Another important finding is that ownership does not require continuous utilization to be economically viable. In fact, it's very simple.

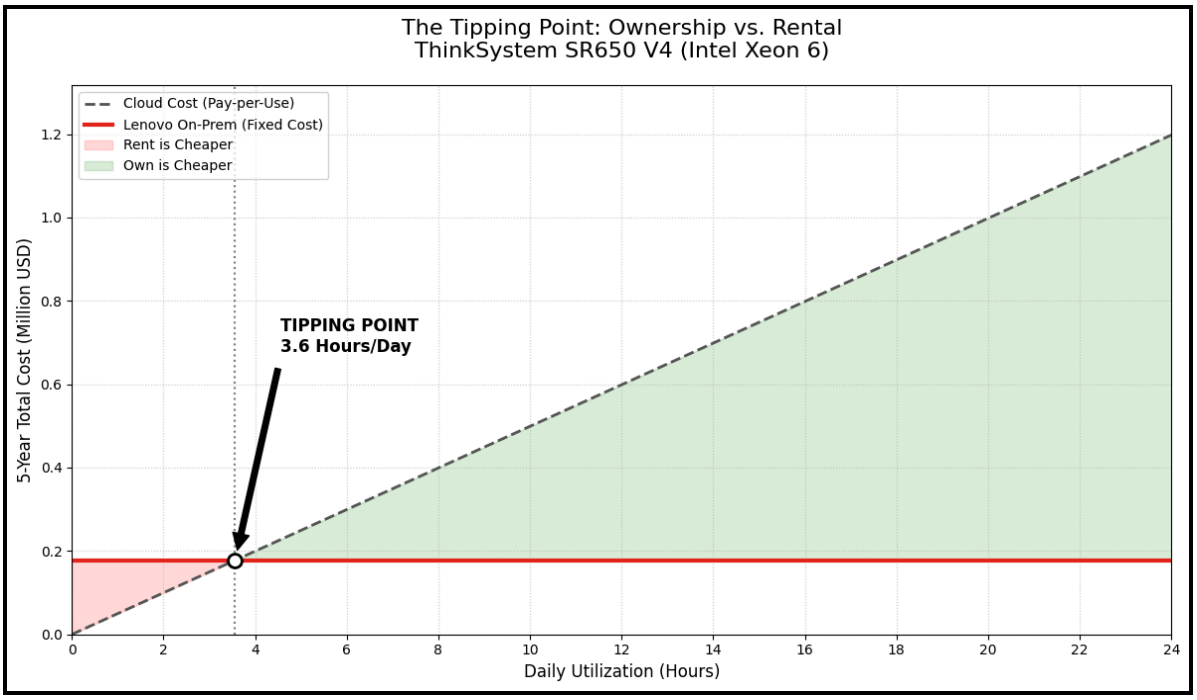

Figure 3. Running the server more than 3.6 hours a day makes on-prem more economical

In addition to potential cost savings, owned infrastructure provides predictable capacity and greater visibility into resource planning, reducing uncertainty associated with unpredictable, exponential cloud consumption.

Performance and Enterprise Workloads

Enterprise AI workloads differ significantly from large-scale public AI services. While consumer-facing applications often prioritize maximum throughput and concurrency, many enterprise deployments focus on employee productivity, knowledge access, and workflow automation.

Common examples include internal assistants, enterprise search, document summarization, content generation, and retrieval-augmented generation (RAG) applications. These workloads typically require responsive performance and consistent user experiences rather than extreme levels of throughput.

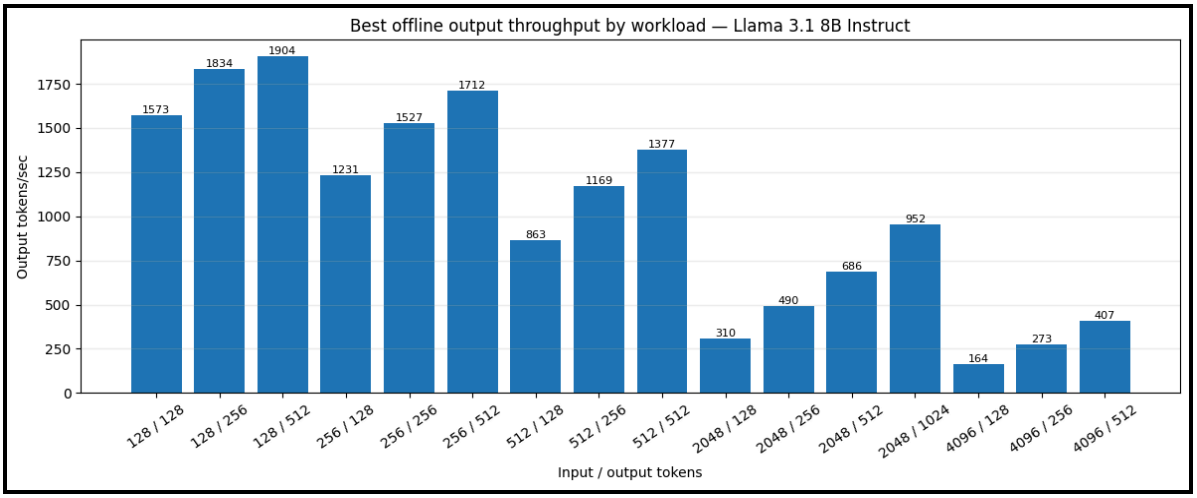

Figure 4. CPU based inference can support production workloads without increasing latency

Throughput and latency across short chat, standard chat, long-generation, and RAG-style workloads. The benchmark test set uses a Server Latency Agreement of TTFT < 1 second and TPOT < 100 ms.

Benchmark testing demonstrated responsive serving performance across representative enterprise use cases. Results indicate that CPU-based inference can effectively support many production workloads while meeting latency expectations commonly associated with business applications.

This distinction is important. The question for most organizations is not whether CPUs can replace GPUs in every scenario. Rather, it is whether a given workload requires accelerator-class performance to deliver business value. For many enterprise applications, the answer may be no.

Why Token Economics Matter

As AI usage expands, token generation becomes one of the most important measures of infrastructure efficiency.

Every prompt, retrieval request, summary, and generated response contribute to overall operating costs. At enterprise scale, small differences in cost per token can translate into substantial differences in annual spending.

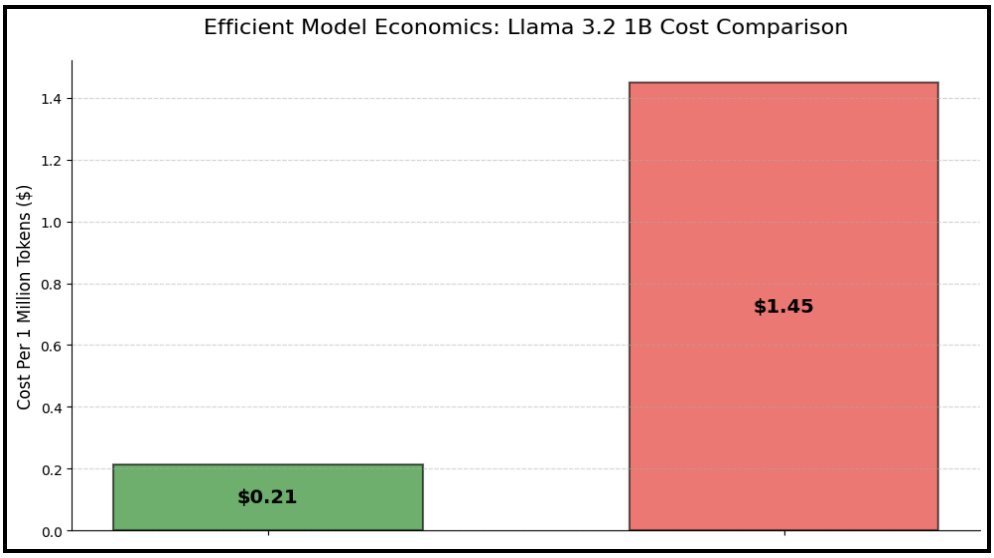

Figure 5. Lenovo versus cloud for CPU-based inference delivers approximately 6.7× lower cost per million tokens than the cloud comparison

Lower token-generation costs can help organizations move beyond limited deployments and support broader adoption across departments. Instead of restricting AI access to a small number of teams, enterprises can deploy AI more widely while maintaining financial discipline.

Governance, Security, and Data Control

Cost and performance are only part of the infrastructure decision.

Many organizations operate within environments that require strict governance controls, regulatory compliance, or enhanced protection of intellectual property. As AI systems interact with proprietary documents and business-critical information, questions about data handling become increasingly important.

Running inference on-premises allows prompts, source content, and generated outputs to remain within the organization's environment. This can simplify governance processes and provide greater control over security policies, access controls, and data management practices.

Building a balanced Hybrid AI Infrastructure includes GPU-accelerated systems and CPU-based platforms. The most successful organizations are increasingly adopting a workload-centric approach, matching infrastructure resources to business needs efficiently.

GPUs remain the preferred platform for model training and highly demanding inference environments. Cloud services provide flexibility and rapid access to new capabilities. CPU-based infrastructure can support a wide range of production AI applications where economics, governance, and operational simplicity are primary considerations.

Organizations can scale AI strategically, allocating accelerator resources where they deliver the greatest value while using CPU infrastructure to support broader deployment across the enterprise.

Conclusion

As enterprise AI matures, infrastructure decisions are becoming business decisions. For many the question is not whether GPUs are powerful; it is whether every LLM workload truly needs GPU infrastructure. Intel Xeon 6 CPU-based inference gives you a practical path to deploy AI where cost, data control, and operational simplicity matter as much as raw accelerator performance. CPU-only inference can deliver the performance needed while avoiding the cost and complexity of dedicated GPU servers.

On-prem CPU inference also helps address one of the biggest enterprise AI concerns: control. By running models inside your own environment, you can keep sensitive data, business documents, prompts, and generate outputs closer to your governance and security policies. For the right use cases, Intel Xeon 6 offers a balanced approach: strong LLM serving capability, lower token economics, simpler infrastructure planning, and a clear alternative to expensive GPU or cloud-only inference strategies.

Combining practical performance with favorable economics and greater control over enterprise data means your IT investments can be used more efficiently and for innovations that drive better customer experience and business outcomes, including revenue.

CPU-based inference is not a replacement for cloud services or GPU infrastructure. Instead, it represents an additional deployment option that can help organizations optimize costs, improve governance, and extend AI capabilities across the business.

Scaling AI successfully means aligning infrastructure choices with workload requirements, and a balance of performance, economics, and long-term operational sustainability that drives revenue.

For more informtion, the following Lenovo Reference Architecture, Lenovo Hybrid AI Platform 201 for Enterprise RAG With Red Hat AI Enterprise.

Authors

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.

Sachin Gopal Wani is a Staff Data Scientist at Lenovo, working on end-to-end Machine Learning (ML) applications for varying customers. He has published articles on the sizing guide and provides sizing information for customers. Sachin holds extensive experience in AI solutions including LLMs, and Computer Vision. He graduated from Rtugers University as a gold medalist specializing in ML and has secured the J.N. Tata Scholarship.

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- Processors

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, LP2460, 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/LP2460
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2460.

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:

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.