Power-Efficient LLM Inference with Intel Xeon 6 Processors

Planning / Implementation

The default assumption in enterprise AI planning is that meaningful LLM inference requires GPU acceleration. For large frontier models served to thousands of concurrent users, that assumption often holds, however a substantial share of enterprise AI workloads including coding assistants and internal RAG assistants utilize small-to-midsize language models in the 7B-8B parameter range and run at far more modest concurrency, typically in the range of single-digit to low double-digit simultaneous users per node.

For this workload tier, the GPU-or-nothing framing imposes unnecessary cost, procurement lead time, and operational complexity. Modern server-class CPUs, purpose-built with AI acceleration in mind, are increasingly capable of serving this tier directly without a discrete accelerator, without a separate toolchain, and without code changes to the inference stack.

This paper tests that proposition directly. We benchmarked Meta Llama 3.1-8B on Lenovo ThinkSystem servers equipped with three Intel Xeon 6 SKUs of varying core count and thermal design power (TDP), measuring total throughput, per-user throughput, time-to-first-token, and the central focus of this analysis: power efficiency in output tokens per kilowatt-hour. The goal is to give enterprise architects a data-backed view of where Xeon 6 CPU-only inference fits, what tradeoffs to expect at scale, and how to select the right processor configuration for a given concurrency target.

Intel Xeon 6 for Power Efficient LLM Inferencing

Intel Xeon 6 (codenamed Granite Rapids) is built with datacenter-scale AI inference as a first-order design goal. Advanced Matrix Extensions (AMX) accelerate the matrix multiplication operations that dominate transformer inference, while higher core counts and large L3 caches reduce memory-access latency during autoregressive token generation. The three SKUs evaluated in this study span a wide range of core count and TDP, allowing the benchmark to isolate how processor scale affects both raw throughput and power efficiency.

Table 1. Processor configurations evaluated in this study

Processor	Cores	TDP	L3 Cache
Intel Xeon 6972P	96	500 W	432 MB
Intel Xeon 6960P	72	500 W	288 MB
Intel Xeon 6732P	32	350 W	144 MB

All three processors were evaluated using an identical software stack and workload definition, ensuring that observed differences in throughput, latency, and power efficiency are attributable to the processor configuration rather than environmental variation.

Table 2. Test environment and methodology summary

Parameter	Detail
Model	Lenovo ThinkSystem SC750 V4 (6972P & 6960P) and SR630 V4 (6732P)
CPUs Tested	Intel Xeon 6972P Intel Xeon 6960P Intel Xeon 6732P
OS	Rocky 9.6
Inference FW	vLLM (0.21.0+cpu), OpenAI-compatible API endpoint
Model	Meta Llama 3 8B
Precision	BF16
Concurrency	1, 2, 4, 8, 16, 32, 64, 128 simultaneous users

Benchmark Results and Discussion

This section presents the full benchmark results across the 1-128 user concurrency sweep for all three Xeon 6 SKUs. We examine the data through two complementary lenses: user experience, captured by output throughput/latency metrics, and system-level efficiency, captured by output tokens per kilowatt-hour. Together these views show how concurrency reshapes the tradeoff between individual responsiveness and aggregate power efficiency, and how that tradeoff varies with specific choice of SKU.

Per-User Throughput and Time-to-First-Token

The following figure plots per-user output throughput (tokens per second per user, solid lines) against time-to-first-token (TTFT, dashed lines) as concurrency scales from 1 to 128 simultaneous users. The pattern is consistent across all three SKUs and is characteristic of shared inference infrastructure generally, whether CPU- or GPU-based: as more users share the same compute resources, per-user throughput declines and queuing increases TTFT.

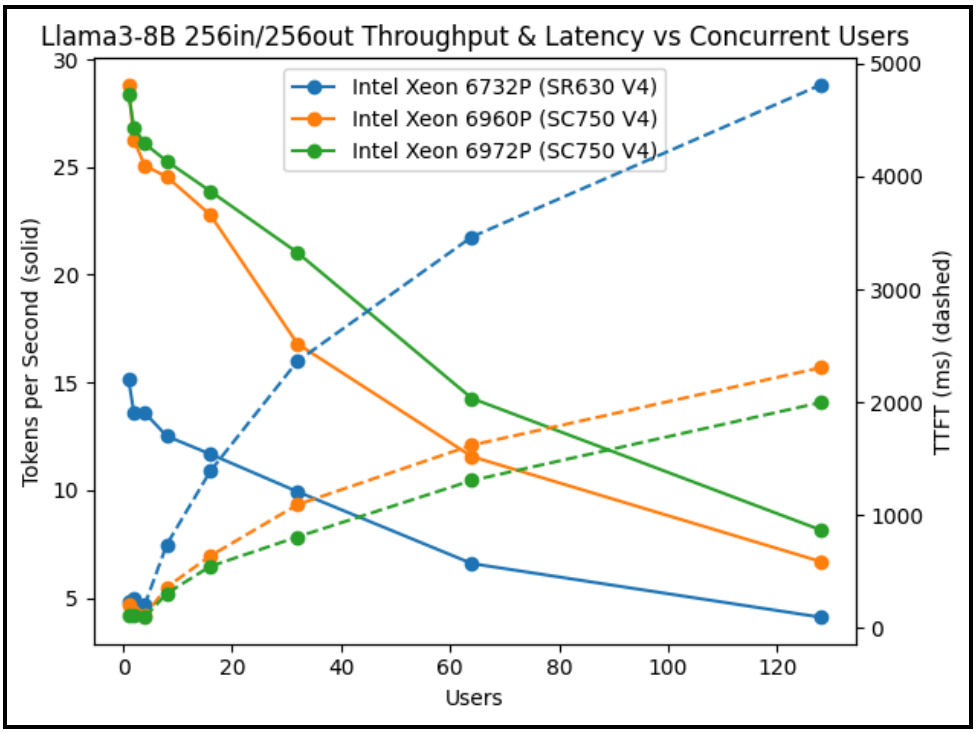

Figure 1. Per-User Throughput and Time-to-First-Token

At low concurrency (1-8 users), the 6960P and 6972P in particular deliver strong, comparable single-user responsiveness approaching 30 tokens/s/user with sub-200ms TTFT. This regime matters most for single-user or lightly shared departmental assistants, and Xeon 6 performs well across the board regardless of SKU.

As concurrency rises into the 16-32 user range, the higher core-count processors begin to separate from the 32-core 6732P. The 6972P and 6960P sustain noticeably higher per-user throughput and lower TTFT than the 6732P at the same load, reflecting their larger core counts and cache capacity absorbing concurrent request scheduling more gracefully. By 64 and 128 users, the gap is substantial: the 6732P's per-user throughput falls toward single digits with TTFT climbing toward roughly 4.8 seconds at 128 users, while the 6972P and 6960P maintain meaningfully better per-user responsiveness, with TTFT in the 1.9-2.0 second range at the same load.

This is the expected and honest tradeoff of shared inference infrastructure: **higher concurrency improves aggregate system efficiency but reduces the responsiveness experienced by each individual user.** Enterprise architects should treat the concurrency level at which per-user TTFT crosses their application's acceptable latency threshold as the practical ceiling for a given SKU and select core count accordingly. For latency-sensitive interactive use cases above roughly 32 concurrent users per node, the higher core-count 6972P and 6960P are the stronger fit; for lighter-loaded or batch-tolerant scenarios, the 6732P remains a capable and lower-TDP option.

Power Efficiency: Output Tokens per kiloWatt-hour

The following figure isolates the metric most relevant to this paper's central thesis: total system power efficiency, measured as output tokens per kilowatt-hour, across the same concurrency sweep. Unlike per-user throughput, this metric improves consistently as concurrency increases for all three processors, because a larger share of each kWhr of fixed processor power draw is converted into useful aggregate token output as more requests are batched and scheduled concurrently.

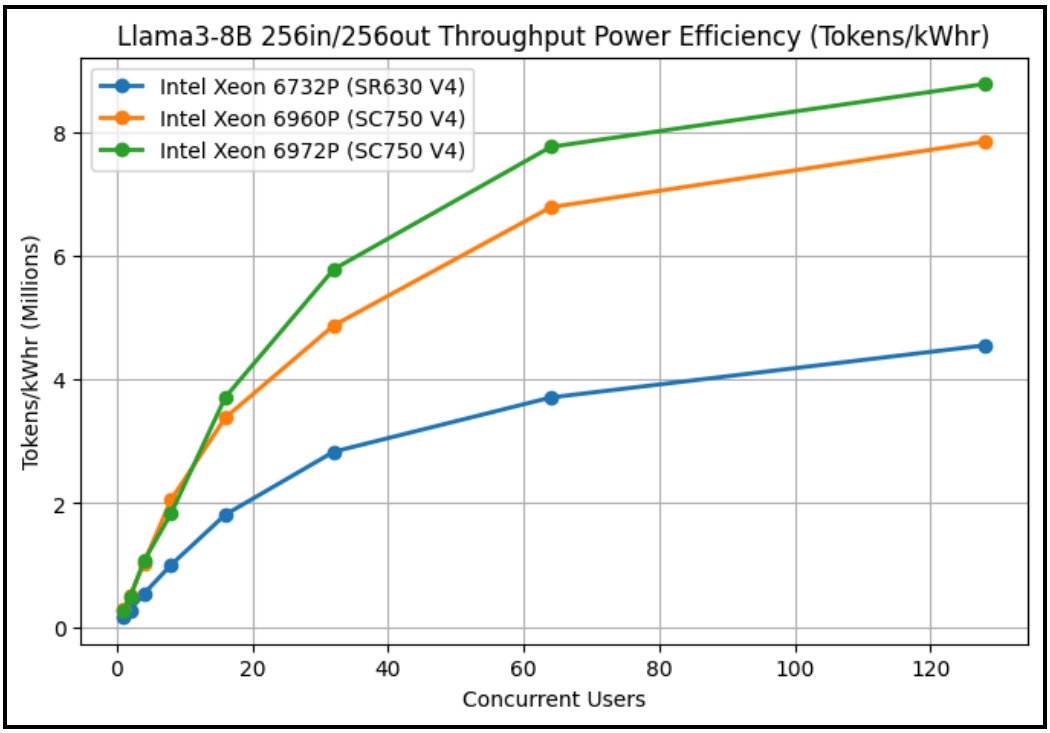

Figure 2. Power Efficiency: Output Tokens per kiloWatt-hour

At low concurrency (1-8 users), all three SKUs are similarly efficient, but as concurrency scales past 16 users, the higher core-count processors pull decisively ahead. By 64 concurrent users, the 6972P reaches approximately 7 million tokens/kWhr and the 6960P approximately 5.7 million tokens/kWhr, compared to roughly 3.2 million tokens/kWhr for the 6732P. At 128 users, the gap persists: the 6972P leads at 8 million tokens/kWhr, the 6960P close behind at roughly 6.5 million tokens/kWhr, and the 6732P trailing at 4 million tokens/kWhr, meaning the 96-core 6972P delivers close to double the power efficiency of the 32-core 6732P at the highest tested concurrency, despite drawing more peak power per socket.

Strategic Fit and Deployment Guidance

Taken together, the throughput, latency, and power efficiency data point to clear, workload-specific guidance for matching Xeon 6 SKU to deployment profile:

- **Low concurrency, latency-sensitive** (1-16 users): All three SKUs deliver strong per-user responsiveness and comparable efficiency. The 32-core 6732P, at 350W TDP, is the most power-conscious choice for departmental deployments that will not scale far past this range.
- **Moderate concurrency** (16–64 users): The 6960P offers a strong balance of per-user responsiveness and power efficiency, with meaningfully better scaling than the 6732P at a TDP in line with the 6972P.
- **High concurrency** (64-128+ users): The 6972P delivers the best aggregate throughput and the highest measured power efficiency, making it the preferred choice when a single node must serve a larger user population or higher-throughput batch and RAG workloads.
- **Latency-critical interactive applications at any concurrency**: Architects should size around the per-user TTFT curve in Figure 1 directly, since acceptable latency, not raw throughput, is typically the binding constraint for interactive copilots and chat assistants.

These results reinforce that CPU-only inference is not a uniform fallback option but a tunable design space: core count and concurrency target can be matched deliberately, the same way GPU model and quantity are matched to workload today. For enterprises running small-to-midsize language models at moderate concurrency, the dominant tier of real-world enterprise AI deployments, Lenovo ThinkSystem

servers with Intel Xeon 6 provide a validated, power-efficient, and immediately deployable alternative to GPU infrastructure, with a clear and predictable upgrade path across the Xeon 6 SKU stack as concurrency demands grow.

ThinkSystem SC750 V4 and SR630 V4

Our test environments were based on two Lenovo ThinkSystem servers, the SC750 V4 and SR630 V4.

The **ThinkSystem SC750 V4 Neptune Server** is a high-performance dual-socket server based on the sixth generation Lenovo Neptune direct water cooling platform. Engineered for large-scale cloud infrastructures and High Performance Computing (HPC), Lenovo ThinkSystem SC750 V4 Neptune excels in intensive simulations and complex modeling. It is designed to handle technical computing, grid deployments, and analytics workloads in various fields such as research, life sciences, energy, engineering, and financial simulation.

With two Intel Xeon 6900 ("Granite Rapids AP") or Xeon 6900+ ("Clearwater Forest AP") processors, including the 6972P and 6960P, the ThinkSystem SC750 V4 server combines the latest high-performance Intel processors and Lenovo's market-leading water-cooling solution, which results in extreme performance in dense packaging.

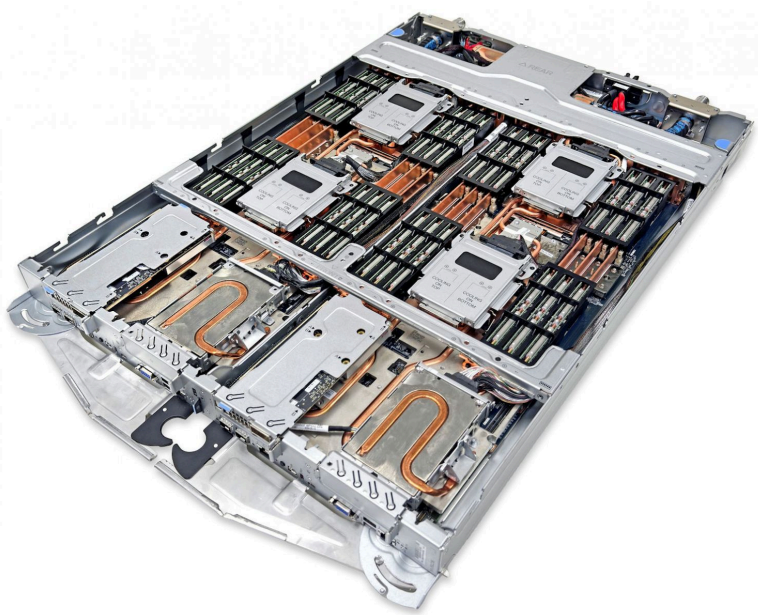

Figure 3. Two dual-socket ThinkSystem SC750 V4 nodes installed in a single water-cooled tray

The **ThinkSystem SR630 V4 Server** is an ideal 2-socket 1U rack server for customers that need industry-leading reliability, management, and security, as well as maximizing performance and flexibility for future growth. The SR630 V4 is based on two Intel Xeon 6 processors including the Xeon 6732P ("Granite Rapids SP") processor.

Combining performance and flexibility, the SR630 V4 server is a great choice for enterprises of all sizes. The server offers a broad selection of drive and slot configurations and offers numerous high performance features. Outstanding reliability, availability, and serviceability (RAS) and high-efficiency design can improve your business environment and can help save operational costs.

Figure 4. Lenovo ThinkSystem SR630 V4

Conclusion

This paper set out to demonstrate how far CPU-only LLM inference can carry real enterprise workloads and do so with particular attention to power efficiency. The data shows that Intel Xeon 6 processors on Lenovo ThinkSystem servers deliver consistent, predictable, and increasingly efficient inference performance as concurrency scales, with output token power efficiency improving greatly with the number of concurrent users across all three SKUs tested.

Higher core-count Xeon 6 processors translate directly into higher power efficiency at moderate-to-high concurrency, while all three SKUs provide strong single-user responsiveness at the low end of the concurrency range. The tradeoff between per-user responsiveness and aggregate efficiency at high concurrency is real and is documented transparently here so that architects can plan around it.

For the enterprise AI tier defined by small-to-midsize language models at moderate concurrency, Lenovo ThinkSystem servers powered by Intel Xeon 6 offer a validated, right-sized, and power-efficient path to production AI available today, without GPU procurement lead times, and with a clear scaling path as workload demands grow.

Author

Eric Page is an AI Engineer at Lenovo. He has 6 years of practical experience developing Machine Learning solutions for various applications ranging from weather-forecasting to pose-estimation. He enjoys solving practical problems using data and AI/ML.

Related product families

Product families related to this document are the following:

- Artificial Intelligence
- Intel Alliance
- Processors
- ThinkSystem SC750 V4 Server
- ThinkSystem SR630 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, LP2463, was created or updated on July 2, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2463
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2463.

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®

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.