AI Inferencing on Intel CPU-Powered Lenovo Servers: Strategic CPU Selection and Implementation

Planning / Implementation

In this third and final part of our series, we focus on the real-world capacity of Lenovo ThinkSystem SR650 V3 servers when running a diverse set of Large Language Models (LLMs). Previous papers in the series are available at the following links:

- Part 1: AI Inferencing on Intel CPU-Powered Lenovo Servers: Demystifying LLM Architectures
- Part 2: AI Inferencing on Intel CPU-Powered Lenovo Servers: Accelerating LLM Performance with Intel Technology

In this paper, we leveraged on 2 Xeon CPU types:

- 4th Gen Intel Xeon 6426Y with 16 cores (mid-range CPU)
- 5th Gen Intel Xeon 8570 with 56 cores (high-end CPU)

Our objective is to determine the maximum batch size each model can handle while maintaining a next-token latency of 130 milliseconds (ms). This threshold is based on human reading speed—typically six words per second—which makes 130ms per token an optimal latency for smooth user consumption. Faster response times would provide diminishing returns, making this a practical benchmark for balancing performance and usability.

To achieve this, we tested nine LLMs across three categories—Encoder-Only, Decoder-Only, and Encoder-Decoder architectures—ranging from small to large model sizes. The nine LLMs we tested are listed in the following table.

Table 1. Tested LLMs

Model Type	**Model Name**	**Parameter Count**
Encoder-Only	DistilBERT	~66M
Encoder-Only	BERT Base	~110M
Encoder-Only	BERT Large	~340M
Decoder-Only	LLaMA 3.2 1B	~1B
Decoder-Only	LLaMA 3.2 3B	~3B
Decoder-Only	LLaMA 3.1 8B	~8B
Encoder-Decoder	T5-Base	~248M
Encoder-Decoder	T5-Large	~783M
Encoder-Decoder	Flan-T5-XL	~2.85B

By determining the maximum batch size that stays within the 130ms latency limit, we can estimate the queries per second (QPS) the server can handle, providing valuable insights into the real-world inference capacity of the ThinkSystem SR650 V3. By balancing batch size, latency, and QPS, enterprises can optimize their AI infrastructure for efficiency and cost-effectiveness.

Lenovo ThinkSystem SR650 V3: Performance, Scalability, and AI Readiness

We use Lenovo ThinkSystem SR650 V3 throughout this paper, Lenovo ThinkSystem SR650 V3 is a high-performance enterprise server built for AI inferencing, machine learning, and data-intensive workloads. Powered by Intel Xeon Scalable processors, it supports up to 8TB of DDR5 memory, PCIe Gen 5 connectivity, and NVMe SSDs, delivering high compute density, rapid data access, and scalability.

Optimized for AI acceleration with Intel Deep Learning Boost (DL Boost) and Advanced Matrix Extensions (AMX), it enhances efficiency for AI workloads while ensuring energy-efficient operations through advanced cooling solutions. With flexible configurations, businesses can tailor deployments to their workload demands, making it a cost-effective alternative to cloud-based AI inferencing.

Figure 1. Lenovo ThinkSystem SR650 V3

Evaluating Batch Size and Its Business Impact

Batch size is a critical factor in AI inferencing as it directly affects throughput and efficiency. In real-world deployments, businesses need to balance latency per query and maximum throughput to meet user demand efficiently. A larger batch size improves throughput by enabling more parallel processing, but it can also increase response times if not managed correctly. The goal of this study is to determine the maximum batch size for different LLMs while keeping next-token latency within 130ms, ensuring a smooth user experience.

The relationship between batch size (B) and queries per second (QPS) can be expressed as:

$$QPS = \frac{Batch\ Size}{Next\ Token\ Latency}$$

This equation helps businesses estimate how many simultaneous queries a ThinkSystem SR650 V3 server can handle efficiently. A server that can process higher batch sizes without exceeding the latency limit will be able to serve more requests at the same time, making it more cost-effective for AI deployments.

By benchmarking different LLMs with varying batch sizes, we provide a data-driven approach for enterprises to choose the right model and hardware configuration.

Performance Results, Analysis and Business Insights

In this section, we present the performance results, highlight our analysis, and delve into the business insights derived from these findings. By comparing large language models of different architectures and scales, we enable businesses to select the most suitable hardware for their specific use cases and requirements.

Results

To better guide hardware decisions, we present performance data grouped by LLM architecture—Encoder-only, Decoder-only, and Encoder-Decoder—highlighting how each type performs under latency-constrained conditions on Lenovo SR650 V3 servers.

- Encoder-Only Models: Efficient for Understanding and Classification
- Decoder-Only Models: Optimized for Generation and Dialogue
- Encoder-Decoder Models: Balanced Architecture for Complex Tasks

Encoder-Only Models: Efficient for Understanding and Classification

Encoder-only LLMs, like BERT, excel in tasks requiring deep understanding, classification, and retrieval rather than text generation. Businesses use them for search optimization, fraud detection, sentiment analysis, content moderation, recommendation systems, and legal/medical document processing. They offer higher accuracy, lower computational costs, and better security than generative models, making them ideal for industries like finance, healthcare, and e-commerce. Created to enhance contextual language understanding, these models improve decision-making, compliance, and operational efficiency while avoiding the risks of AI hallucinations.

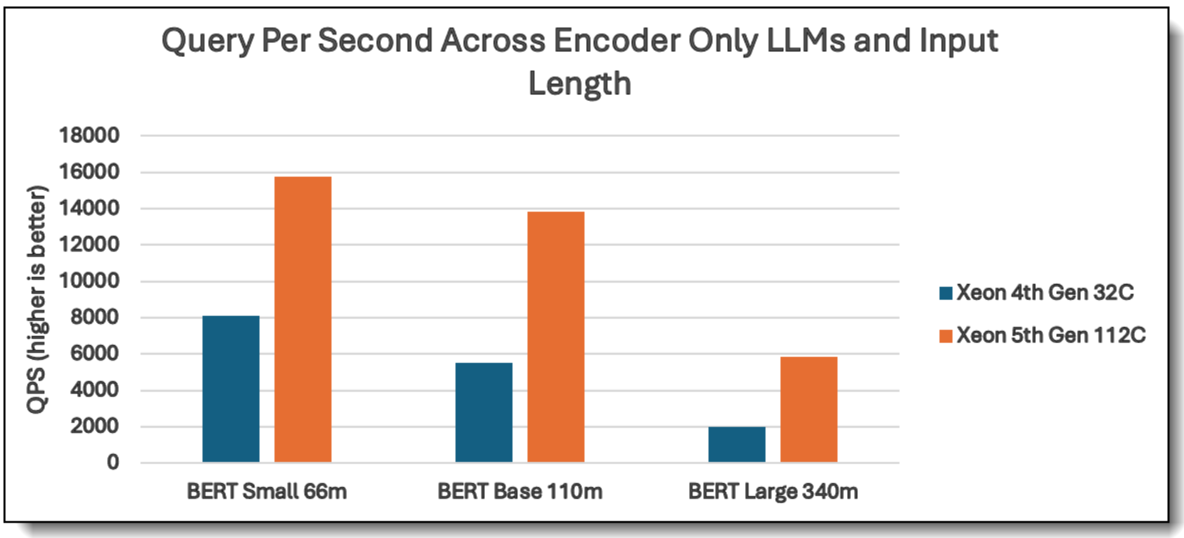

Figure 2. QPS of Encoder-only LLMs with 512 input length

Decoder-Only Models: Optimized for Generation and Dialogue

Decoder-only LLMs, like LLaMA, specialize in text generation, summarization, and conversational AI, making them ideal for chatbots, content creation, code generation, personalized recommendations, and creative writing. Businesses benefit by automating customer support, marketing content, document summarization, and AI-powered assistants, reducing costs and enhancing user engagement. Unlike encoder models, they generate human-like responses, making them useful for interactive applications, storytelling, and brainstorming. Developed to improve natural language generation (NLG), they help businesses scale personalized interactions, streamline workflows, and enhance creative productivity.

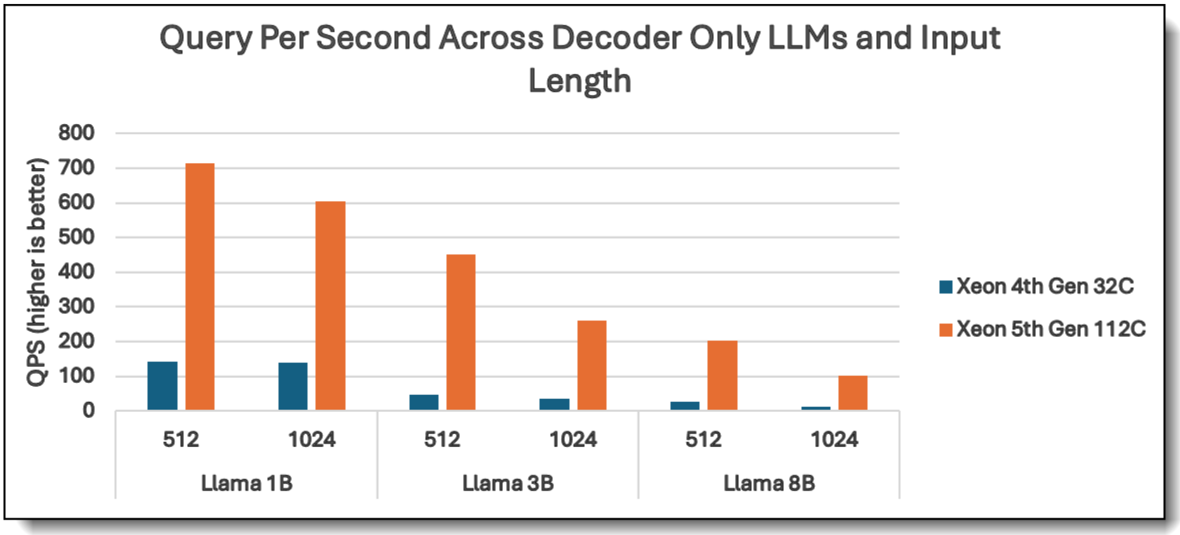

Figure 3. QPS of Decoder-only LLMs with 512 and 1024 input length

Encoder-Decoder Models: Balanced Architecture for Complex Tasks

Encoder-decoder LLMs, like T5, combine deep language understanding (encoder) with text generation (decoder), making them ideal for machine translation, text summarization, data-to-text generation, question answering, and automated report writing. Businesses benefit by automating document processing, multilingual communication, content generation, and knowledge extraction, improving efficiency and reducing manual effort. Unlike encoder-only models, they not only analyze but also generate meaningful text, and unlike decoder-only models, they ensure more context-aware and structured outputs. Created to handle complex NLP tasks requiring both comprehension and generation, these models drive automation, personalization, and multilingual accessibility across industries.

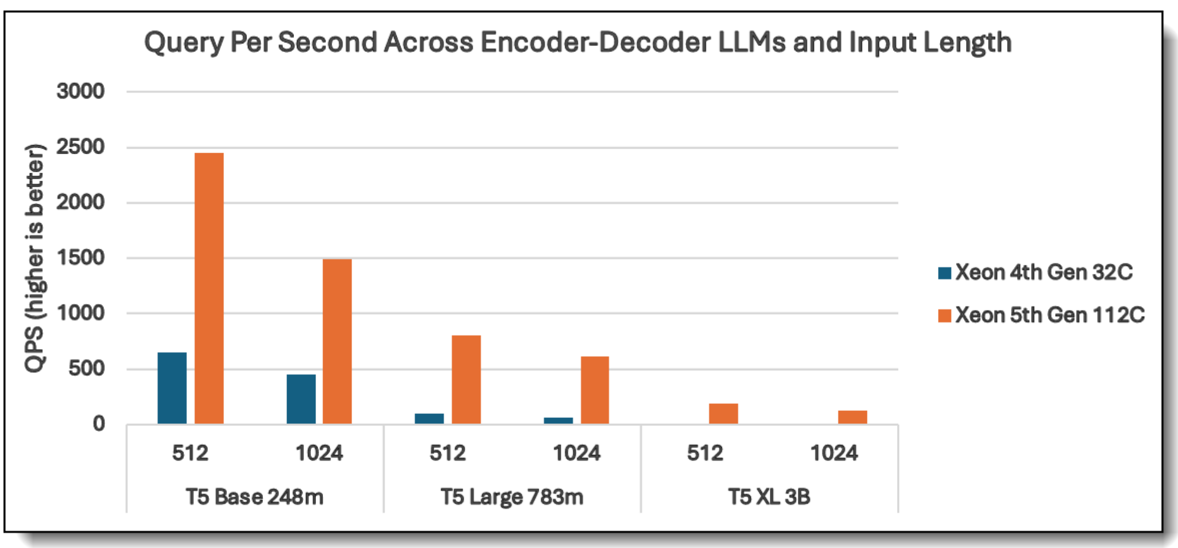

Figure 4. QPS of Encoder-Decoder LLMs with 512 and 1024 input length

Analysis

This section analyzes how different LLM architectures—encoder-only, decoder-only, and encoder-decoder—scale across 4th and 5th Gen Intel Xeon CPUs, highlighting their computational demands, throughput behavior, and suitability for various AI tasks.

1. 5th Gen Xeon delivers up to **7.8x** faster inference, making it ideal for scaling AI workloads

   Across all models, 5th Gen Xeon outperforms 4th Gen Xeon with speedups ranging from 1.9x to 7.8x, ensuring faster response times and better real-time processing. LLaMA 8B sees the highest improvement, increasing from 26 (4th Gen Xeon) to ~200 (5th Gen Xeon), a 7.8x boost, showing that even larger models run significantly faster on 5th Gen Xeon.

2. BERT models maintain strong performance, even on 4th Gen Xeon

   BERT models exhibit high throughput, making them a cost-effective choice for NLP tasks such as search ranking, text classification, and fraud detection. BERT 66M delivers ~8100 throughput on 4th Gen Xeon and ~16,000 on 5th Gen Xeon, a 1.94x improvement, showing that even on older hardware, BERT remains efficient. Even BERT 340M on 4th Gen Xeon achieves ~2000 throughput, proving scalability while remaining computationally manageable.

3. LLaMA and T5 models benefit from hardware optimization despite being computationally intensive

   LLaMA models, while demanding, still see significant improvements on 5th Gen Xeon. LLaMA 3B (512 input) achieves a 9.4x improvement (48 to ~450), ensuring faster processing for AI assistants and language modeling tasks. T5 models, though encoder-decoder based, show strong scaling. T5 Base 248M (512 input) achieves a 3.75x boost, (~650 to ~2400), making summarization and translation tasks more efficient. Even the largest T5 XL 3B model improves from 12 to ~200 (16x boost!), proving 5th Gen Xeon is well-equipped for complex generative AI workloads.

Business Insights

Businesses leveraging AI models for natural language processing (NLP), content generation, and real-time inference need to make data-driven hardware decisions for cost-effectiveness, scalability, and efficiency. Based on the performance insights of BERT, LLaMA, and T5 models on 4th Gen Xeon and 5th Gen Xeon, here are key actionable recommendations:

1. **Upgrade to 5th Gen Xeon for large-scale AI applications and high-throughput needs:**

   Businesses deploying large transformer models (LLaMA 8B, T5 XL 3B) should prioritize 5th Gen Xeon, as it delivers up to 7.8x higher inference throughput compared to 4th Gen Xeon.

2. **Optimize model selection based on business use cases and computational budget:**

   BERT models perform efficiently even on 4th Gen Xeon, making them a great option for businesses needing text classification, search optimization, and sentiment analysis without significant hardware investment.

3. **Manage computational costs by balancing input length and model size:**

   Longer input lengths significantly impact performance, particularly for decoder-based models (LLaMA, T5). For models with fewer than 500m parameters, 4th Gen Xeon CPU could be sufficient to meet the business requirements, while for models with more than 1B parameters, high end Xeon CPUs are recommended. Businesses should optimize their AI pipelines to truncate, compress, or pre-process data to maintain efficiency.

What's Next: On Premises or Cloud?

As businesses increasingly adopt AI-driven solutions, a critical decision arises: deploy AI models on-premises or leverage cloud-based infrastructure? This choice impacts cost, scalability, security, and operational efficiency, making it essential to align deployment strategies with business objectives.

The table below introduces a general guidance how to choose between on-premises VS cloud.

Table 2. What's Next: On Premises or Cloud?

Factor	On-Premises	Cloud
Upfront Cost	High – Requires purchasing servers, GPUs, storage, and networking.	Low – No hardware investment; pay-as-you-go pricing.
Ongoing Cost	Lower in the long run, but includes maintenance, power, and IT staffing.	Higher over time due to recurring cloud fees, storage, and data transfer costs.
Scalability	Limited – Must manually upgrade hardware.	Instant – Can scale up/down based on demand.
Performance & Latency	Low latency for real-time AI processing. Best for local data inference.	Can experience latency issues due to network dependency.
Security & Compliance	Full data control – Best for finance, healthcare, and government.	Relies on cloud provider's security – May not meet strict compliance needs.
IT Management	Requires in-house IT for server maintenance, updates, and monitoring.	Managed by cloud provider – Reduces IT workload.
Best Use Case	Long-term, stable AI workloads where cost savings are realized over time.	Short-term, scalable, and high-performance AI workloads requiring on-demand compute power.

On-premises deployment offers greater control, data privacy, and long-term cost efficiency, while cloud deployment provides scalability, ease of maintenance, and access to high-performance computing resources. Here we are introducing the costs that associated with two strategies.

More details: We plan to provide a more detailed guidance on how to choose between on-prem and cloud in a future paper.

Now, let's provide an example with numbers. When evaluating the cost-effectiveness of deploying AI workloads on-premises using a Lenovo ThinkSystem SR650 V3 server versus utilizing cloud services, it's essential to consider both initial investments and ongoing operational expenses.

Using the SR650 V3 with 4th Gen Xeon CPU as an example: (these are estimated cost and we only consider the upfront and ongoing cost; others are minor cost such as maintenance cost, or cost associated with downtime, data backup these hidden cost which are hard to quantify.)

- On Premises Cost:
    - Initial Investment on the server: Approximately $12,000 with 2 Intel Xeon Gold 6426Y, 512GB (16x32GB DDR5 5600 MT/s) RAM etc.
    - Ongoing Operational Expenses: Rough Estimate at $500 per month, the exact amount varies a lot among SMB to large enterprise.
- Cloud Cost:
    - Select 2* r7i.8xlarge with 32 vCPU and 256 GB RAM. The monthly cost is around $1400 for a 3 year saving plan.

The following figure shows these costs.

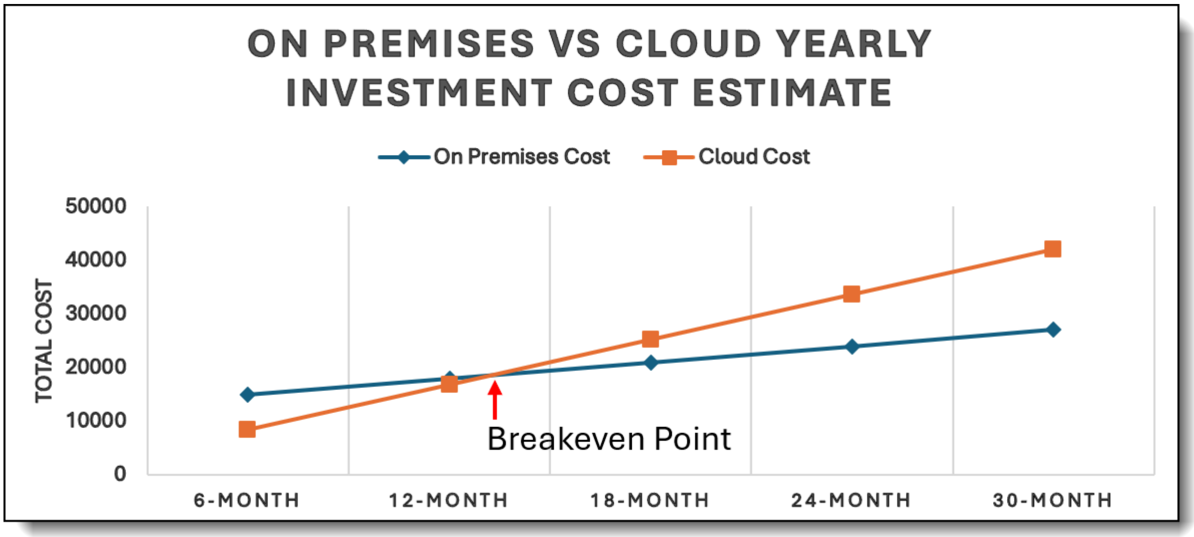

Figure 5. Estimated yearly cost for on-premises VS cloud

We notice that the breakeven point is around year 3. This means that after **13 months**, the total expenditure on AWS will exceed the upfront cost of an on-premises server.

Final Takeaway

- If AI workloads run continuously for over **13 months**, on-premises hardware investment is more cost-effective. In addition, it will be saving approximately **$15,000** over a 3-year period.
- If workload flexibility, no maintenance, and scalability are priorities, AWS Cloud may be a better option despite the higher long-term cost.

Conclusion

Choosing the right hardware and deployment strategy is critical for businesses optimizing AI workloads. Based on the performance, cost, and scalability analysis, here's a clear decision framework for selecting between 4th Gen Xeon vs. 5th Gen Xeon CPUs and on-premises vs. cloud deployment.

1. When to Choose 4th Gen Xeon CPUs?
    - Best for cost-effective, stable AI workloads that don't require maximum throughput.
    - Ideal for smaller AI models like BERT-base, moderate-sized inference tasks, and batch processing applications.
    - Suitable for businesses with budget constraints where power efficiency is a priority over peak performance.
2. When to Upgrade to 5th Gen Xeon CPUs?
    - Recommended for high-performance AI applications, especially LLaMA, T5, and large-scale transformer models.
    - If AI workloads require real-time inference, faster response times, and handling large input sequences, 5th Gen Xeon offers 2-7.8x higher throughput.
    - Best for businesses deploying AI-driven chatbots, recommendation engines, fraud detection, and high-speed NLP applications.
3. When to Choose On-Premises AI Deployment?
    - Best for long-term cost savings (AI workloads running continuously for 2 or 3+ years).
    - Recommended for strict data privacy & regulatory compliance (e.g., finance, healthcare, government).
    - Suitable for real-time AI inference where low latency and data processing speed are critical.
    - Ideal for organizations with stable, predictable AI workloads that justify an upfront hardware investment.
4. When to Use Cloud for AI Deployment?
    - Best for short-term, flexible, and scalable AI workloads where on-demand resources are needed.
    - Cost-effective for AI projects running less than 1.5 years, avoiding high upfront investments.
    - Ideal for training large-scale AI models, distributed computing, and experimenting with different architectures.
    - Reduces IT management overhead, making it a preferred choice for businesses with limited infrastructure expertise.

Final Recommendation

For businesses running long-term AI workloads, prioritizing cost savings, and requiring strict data security, on-premises deployment with 5th Gen Xeon is the optimal choice. However, for organizations needing large variable scalability, high-performance cloud computing, and low upfront capital costs, cloud-based solutions remain the best fit. By aligning AI workload characteristics with hardware and deployment strategies, businesses can maximize efficiency, performance, and cost-effectiveness.

Test configurations

The following table lists the configurations used in our tests.

Table 3. Test configuration

Hardware Configuration	**System 1**	**System 2**
Server	Lenovo ThinkSystem SR650 V3	Lenovo ThinkSystem SR650 V3
Processor	Intel Xeon Gold 6426Y	Intel Xeon Platinum 8570
Microarchitecture	SPR	EMR_XCC
Sockets	2	2
Cores per Socket	16	56
Hyperthreading	Intel Hyper-Threading Technology Enabled	Intel Hyper-Threading Technology Enabled
CPUs	32	112
Turbo	Intel Turbo Boost Technology Enabled	Intel Turbo Boost Technology Enabled
Base Frequency	1.8GHz	2.1GHz
All-core Maximum Frequency	2.8GHz	4.0GHz
Maximum Frequency	3.0GHz	3.0GHz
NUMA Nodes	2	2
Installed Memory	512GB (16x32GB DDR5 5600 MT/s [5600 MT/s])	1024GB (16x64GB DDR5 5600 MT/s [5600 MT/s])
NIC	2x Ethernet Controller E810-XXV for SFP 4x I350 Gigabit Network Connection 2x Ethernet Controller E810-C for QSFP	2x Ethernet Controller E810-XXV for SFP 4x I350 Gigabit Network Connection 2x Ethernet Controller E810-C for QSFP
Disk	1x 29.3G ProductCodeM 16x 7T Micron_7450_MTFDKCC7T6TFR 1x 894.3G Micron_7450_MTFDKBA960TFR	4x 2.9T Micron_7450_MTFDKCB3T2TFS 2x 11.6T INTEL SSDPF2KE128T1O 2x 447.1G MTFDDAV480TGA-1BC16A 03KH112D7B08305LEN
BIOS	ESE128C-3.30	ESE128C-3.30
Microcode	0x2b000620	0x21000283
OS	Ubuntu 22.04.5 LTS	Ubuntu 24.04.2 LTS
Kernel	5.15.0-131-generic	6.8.0-52-generic

References

For more information, see these resources:

- Liu, B. Comparative analysis of encoder-only, decoder-only, and encoder-decoder language models. College of Liberal Arts & Sciences, University of Illinois Urbana-Champaign, Champaign, IL. https://www.scitepress.org/Papers/2024/128298/128298.pdf
- Accelerating RAG Pipelines for Enterprise LLM Applications using OpenVINO on the Lenovo ThinkSystem SR650 V3 with 5th Gen Intel Xeon Scalable Processors. By Rodrigo Escobar, Abirami Prabhakaran, David Ellison, Ajay Dholakia, Mishali Naik. https://lenovopress.lenovo.com/lp2025-accelerating-rag-pipelines-for-llms-using-openvino-on-sr650-v3
- Intel Extension for PyTorch Github: https://github.com/intel/intel-extension-for-pytorch
- Intel Advanced Matrix Extensions (Intel AMX) Overview. https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/advanced-matrix-extensions/overview.html
- Lenovo ThinkSystem SR650 V3 Server Product Guide. https://lenovopress.lenovo.com/lp1601-thinksystem-sr650-v3-server
- Lenovo ThinkSystem SR650 V3 Configure and Buy: https://www.lenovo.com/us/en/p/servers-storage/servers/racks/thinksystem-sr650-v3/len21ts0013

Author

Kelvin He is an AI Data Scientist at Lenovo. He is a seasoned AI and data science professional specializing in building machine learning frameworks and AI-driven solutions. Kelvin is experienced in leading end-to-end model development, with a focus on turning business challenges into data-driven strategies. He is passionate about AI benchmarks, optimization techniques, and LLM applications, enabling businesses to make informed technology decisions.

David Ellison is the Chief Data Scientist for Lenovo ISG. Through Lenovo's US and European AI Discover Centers, he leads a team that uses cutting-edge AI techniques to deliver solutions for external customers while internally supporting the overall AI strategy for the Worldwide Infrastructure Solutions Group. Before joining Lenovo, he ran an international scientific analysis and equipment company and worked as a Data Scientist for the US Postal Service. Previous to that, he received a PhD in Biomedical Engineering from Johns Hopkins University. He has numerous publications in top tier journals including two in the Proceedings of the National Academy of the Sciences.

Tanisha Khurana is an AI Data Scientist at Lenovo ISG with over 5 years of experience developing machine learning solutions. She focuses on end-to-end AI development and deployment across diverse industries with expertise in vision-based applications and a growing focus on large language models.

Related product families

Product families related to this document are the following:

- Artificial Intelligence

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, LP2204, was created or updated on April 17, 2025.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2204
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2204.

Trademarks