

On-Premise vs Cloud: Generative AI Total Cost of Ownership (2026 Edition)

Positioning Information

Executive Summary

As the artificial intelligence landscape transitions from the experimental fervor of 2023–2024 into the industrial-scale deployment phase of 2026, the economic models governing AI infrastructure are undergoing a radical realignment. The initial wave of Generative AI adoption was characterized by rapid prototyping on readily available cloud infrastructure, prioritizing speed of access over cost efficiency. However, the maturation of Large Language Models (LLMs), epitomized by the release of massive models exceeding 400 billion parameters and the agentic workflows they enable, has introduced a new paradigm of sustained, high-throughput inference that challenges the financial viability of public cloud "rent-seeking" models.

This 2026 update to the [2025 Generative AI TCO paper](#) rigorously re-evaluates the financial and technical trade-offs between on-premises infrastructure and cloud services. We incorporate the latest advancements in accelerated computing, specifically the NVIDIA Blackwell architecture (B200, B300), the widespread enterprise adoption of the H200 Tensor Core GPU, and the emergence of cost-efficient inference accelerators like the L40S. Furthermore, this report introduces "**Token Economics**", a granular, metric-driven framework analyzing the amortized cost-per-million-tokens—to provide a direct, apples-to-apples comparison between owning infrastructure and consuming intelligence via APIs.

Our analysis, based on a 5-year enterprise hardware lifecycle, identifies three critical market inflections in 2026:

1. **The Blackwell Efficiency Singularity:** The architectural leap from Hopper (H100) to Blackwell (B200/B300) improves inference throughput significantly, fundamentally altering TCO calculations by compressing the physical footprint required for massive models.
2. **Accelerated Breakeven Velocity:** For sustained inference workloads (utilization >20%), on-premises infrastructure now reaches a breakeven point against hyperscale cloud providers in as little as **4 months**, a significant compression from the 12–18 month cycles observed in previous generations.
3. **The Rise of Efficient Inference:** For the "workhorse" models in the 7B to 70B parameter range, the Lenovo ThinkSystem SR650a V4 (L40S) has emerged as a TCO champion, often outperforming H100-class cloud instances in price-performance for batch inference tasks.

The Generative AI Infrastructure Landscape in 2026

The operational reality of Generative AI has bifurcated into two distinct domains: the "Training Factory," characterized by massive, burst-oriented compute loads, and the "Inference Engine," characterized by persistent, latency-sensitive utility requirements. While the cloud remains a potent tool for the former, the latter drives the overwhelming majority of long-term enterprise costs.

From Hopper to Blackwell: The Hardware Evolution

The hardware landscape of 2026 is defined by the migration from NVIDIA's Hopper architecture (H100) to the Blackwell architecture (B200, B300). This is not merely a linear performance increase; it is a structural change in how AI compute is delivered, heavily influencing Total Cost of Ownership.

- **NVIDIA H100 & H200:** The H100 established the baseline for GenAI. The H200 refined this with HBM3e memory, delivering 141GB of capacity and 4.8 TB/s of bandwidth. This expansion was critical for alleviating the memory-bound bottlenecks of LLM inference, allowing larger models to reside on fewer GPUs.
- **NVIDIA B200:** The B200 features a dual-die design packing 208 billion transistors, offering up to 192GB of HBM3e memory and 8 TB/s of bandwidth. Most critically for TCO, it introduces FP4 precision support via the second-generation Transformer Engine.
- **NVIDIA B300 (Blackwell Ultra):** Designed for the "trillion-parameter era," the B300 expands memory capacity to 288GB HBM3e per GPU. Used in massive-context inference and training, it offers 1.5x the memory of the B200 to fit even larger models on a single node.
- **NVIDIA L40S:** A universal data center GPU based on the Ada Lovelace architecture. With 48GB of GDDR6 memory, it avoids the supply constraints of H-series GPUs and offers a highly cost-effective alternative for fine-tuning and inference of small-to-medium models.

The Shift to "Token Economics"

In 2026, the primary metric for AI success has evolved from "FLOPS" (Floating Point Operations Per Second) to "**Tokens Per Second per Dollar**" (TPS/\$). Organizations are moving away from measuring server uptime and towards measuring the cost efficiency of token generation. This shift necessitates a comparison not just of server hardware costs, but of the amortized cost of generating 1 million tokens on-premise versus the "retail price" of 1 million tokens from a cloud API.

The Lenovo Infrastructure Portfolio

To facilitate direct comparison, we analyze specific Lenovo ThinkSystem platforms engineered for these accelerators:

- **ThinkSystem SR675 V3:** A versatile 3U platform supporting mixed configurations of H100, H200, and L40S GPUs.
- **ThinkSystem SR680a V3:** An 8U air-cooled server optimized for maximum GPU density, supporting 8x H200 or B200 SXM GPUs.
- **ThinkSystem SR680a V4:** The flagship platform for 2026, designed for the B300 and Intel Xeon 6 processors. It features N+N power redundancy and advanced thermal headroom.
- **ThinkSystem SR650a V4:** A dense, cost-optimized 2U server supporting up to 4x L40S GPUs, targeting efficient inference at the edge or in space-constrained data centers.

Large Language Models - The New Workload

To accurately calculate TCO, one must first understand the physical resource consumption of the workload itself. Large Language Models (LLMs) impose unique constraints on infrastructure that differ significantly from traditional enterprise applications. For more information, you can refer to the [Lenovo LLM Sizing Guide](#).

The Memory Wall

For inference, LLMs are predominantly memory-bound rather than compute-bound. The entire model architecture must be loaded into GPU memory (VRAM), and for every token generated, the model weights must be moved from memory to the compute units.

- **Model Sizing Rule of Thumb:** A model generally requires approximately 2 bytes per parameter in FP16/BF16 precision.
- **70B Parameter Model:** Requires ~140GB VRAM (FP16), barely fitting on a single H200 or necessitating multi-GPU tensor parallelism on smaller cards.
- **400B+ Parameter Model:** Requires ~800GB+ VRAM (FP16), mandating an 8-GPU cluster (HGX H100/H200/B200) just to load the model.

KV Cache Overhead

In addition to static weights, the Key-Value (KV) cache grows linearly with context length and batch size. For modern "long-context" models supporting 1M tokens, the KV cache can consume more memory than the model itself.

Quantization as a Cost Lever

Quantization reduces the precision of model weights (e.g., from 16-bit to 8-bit or 4-bit), significantly lowering memory requirements and increasing throughput without substantial accuracy loss.

Financial Framework and Metrics

This report utilizes a rigorous financial framework to compare disparate consumption models—CapEx-heavy on-premises vs. OpEx-heavy cloud.

Core TCO Metrics

- **CapEx (Capital Expenditure):** The upfront cost of purchasing servers, GPUs, networking (InfiniBand/Ethernet), and storage. For this analysis, we amortize CapEx over a standard 5-year enterprise hardware lifecycle. While, in real conditions a system will have a recovery value, for our comparison we donate the system away with \$0 recovery.
- **OpEx (Operational Expenditure):**
 - **On-Prem:** Maintenance, Electricity (kW/h), Cooling, and Datacenter Space.
 - **Cloud:** Hourly instance rates **only**. To demonstrate that Lenovo infrastructure remains superior even under the most charitable cloud assumptions, we explicitly **exclude** storage costs, data egress fees, and support plans from the cloud calculation.
- **TPS (Tokens Per Second):** The measure of throughput; defined as the number of output tokens produced by the model per second. This is the primary metric for measuring the speed of a model on a system.
- **Cost Per Million Tokens (\$/1M):** A normalized efficiency metric allowing direct comparison between buying hardware and buying API tokens.

Cloud Pricing Methodology

We compare Lenovo infrastructure against the three major hyperscalers (AWS, Azure, GCP). Prices are derived from listed rates for US regions where systems are available to rent (as of Dec 23, 2025).

Hardware and Instance Comparison (2026)

To perform a valid TCO analysis, we align Lenovo on-premises configurations with their nearest cloud equivalents.

Comparison Configurations

We compared configurations across different generations to provide a comprehensive overview. The table below lists the systems we compared.

Table 1. Comparison of different configurations used for analysis

| Configuration ID | Lenovo On-Prem Configuration | GPU Config | AWS Equivalent | Azure Equivalent | GCP Equivalent |
|-----------------------------------|-------------------------------------|-------------------|------------------|------------------|----------------|
| Config A (Hopper) | ThinkSystem SR680a V3 (or SR675 V3) | 8x H100 HBM 80GB | p5.48xlarge | ND96isr H100 v5 | a3-highgpu-8g |
| Config B (Hopper Refresh) | ThinkSystem SR680a V3 | 8x H200 HBM 141GB | p5en.48xlarge | ND96isr H200 v5 | a3-ultragpu-8g |
| Config C (Blackwell) | ThinkSystem SR680a V3 | 8x B200 SXM 192GB | p6-b200.48xlarge | ND GB200 v6 | a4-highgpu-8g |
| Config D (Blackwell Ultra) | ThinkSystem SR680a V4 | 8x B300 SXM 288GB | p6-b300.48xlarge | ND GB300 v6 | -- |
| Config E (Efficient) | ThinkSystem SR650a V4 | 4x L40S 48GB | g6e.24xlarge | -- | -- |

System Pricing

Using our Internal Configurator tool, we have created configs comparable to the cloud counterparts. Provided below are the usual sale price to the customer (as of January 15, 2026, in USD):

- **Config A (8x H100):** \$250,141.8 (phased out, no longer available to purchase)
- **Config B (8x H200):** \$277,897.75
- **Config C (8x B200):** \$338,495.75
- **Config D (8x B300):** \$461,567.5
- **Config E (4x L40S):** \$52,390.5

Comparative Pricing Analysis

This section establishes the baseline costs (As of December 23, 2025) used for the TCO calculations.

Cloud Instance Pricing (Hourly Rates)

All major cloud instances have the following systems available to rent on-demand or reserve an instance. We are using the publically available pricing that was available from official sources at the time of writing.

Microsoft Azure:

Table 2. Azure Cloud Instance Pricing (Hourly Rates)

| Instance name | Comparable Config | GPU | On-Demand | 1 yr reserved | 3 yr reserved | 5 yr reserved |
|-----------------|-------------------|---------|-----------|---------------|---------------|---------------|
| ND96isr H100 v5 | A | 8x H100 | \$98.32 | \$62.92 | \$43.16 | \$39.32 |

Google Cloud Platform (GCP):

Table 3. GCP Cloud Instance Pricing (Hourly Rates)

| Machine type | Comparable Config | GPU | On demand | 1 yr reserved | 3 yr reserved |
|----------------|-------------------|---------|-----------|---------------|---------------|
| a3-ultragpu-8g | B | 8x H200 | 84.808 | 58.47 | 37.20 |
| a4-highgpu-8g | | 8x B200 | -- | 88.68 | 56.55 |

Amazon Web Services (AWS) EC2:

Table 4. AWS EC2 Cloud Instance Pricing (Hourly Rates)

| Instance name | Comparable Config | GPU | On-Demand |
|------------------|-------------------|---------|------------|
| p5.48xlarge | | 8x H100 | \$55.04 |
| p5en.48xlarge | | 8x H200 | \$63.296 |
| p6-b200.48xlarge | C | 8x B200 | \$113.9328 |
| p6-b300.48xlarge | D | 8x B300 | \$142.416 |
| g6e.24xlarge | E | 4x L40S | \$15.06559 |

On-Premises Operational Costs

For accurate TCO, we assume the following operational costs for an on-premises deployment:

- **Annual Maintenance:** 12% of system cost per year.
- **Electricity:** \$0.12 per kWh (US Commercial Average).
- **Cooling Costs:** \$0.18 per kWh for air cooled, \$0.09 per kWh for liquid cooled (on average).
- **Colocation:** ~\$1,500/month per rack for high-density power (Configs A-D). ~\$600/month per rack for standard-density (Config E)

Total Cost of Ownership (TCO) Analysis

In this section, we evaluate the financial performance of on-premise vs. cloud across three specific scenarios:

- [Case 1: Breakeven Point Analysis](#)
- [Case 2: Total Cost of System Over Time \(5-Year Lifecycle\)](#)
- [Case 3: Hourly Utilization Threshold](#)

Case 1: Breakeven Point Analysis

Scenario: Comparing the breakeven point of purchasing **Config A (8x H100)** versus renting the equivalent **Azure ND96isr H100 v5**.

- **Cloud Cost (Azure):**
 - On-Demand: \$98.32/hr
 - 1-Year Reserved: \$62.92/hr
 - 3-year Reserved: \$43.16/hr
 - 5-year Reserved: \$39.32/hr
- **On-Prem Cost (Lenovo Config A):**
 - CapEx: \$250,141.8
 - OpEx (Hourly): \$6.37 (\$3.42 amortized maintenance + \$0.87 for power and cooling + \$2.08 colocation)

Breakeven Calculation (vs. On-Demand):

$$250141.8 + 6.37x = 98.32x$$

$$91.95x = 250141.8$$

$$\Rightarrow x \approx 2,720 \text{ hours}$$

Result: ~3.7 months to break even.

Breakeven Calculation (vs. 1-Year Reserved):

$$250141.8 + 6.37x = 62.92x$$

$$56.55x = 250141.8$$

$$\Rightarrow x \approx 4,423 \text{ hours}$$

Result: ~6 months to break even. Even against a committed 1-year reserved instance, the Lenovo system pays for itself in half a year!

The following figure showcases the comparison between on-prem system and the cloud pricing for on-demand as the breakeven. We also show how changing to a 1-year reserved instance affects the breakeven period.

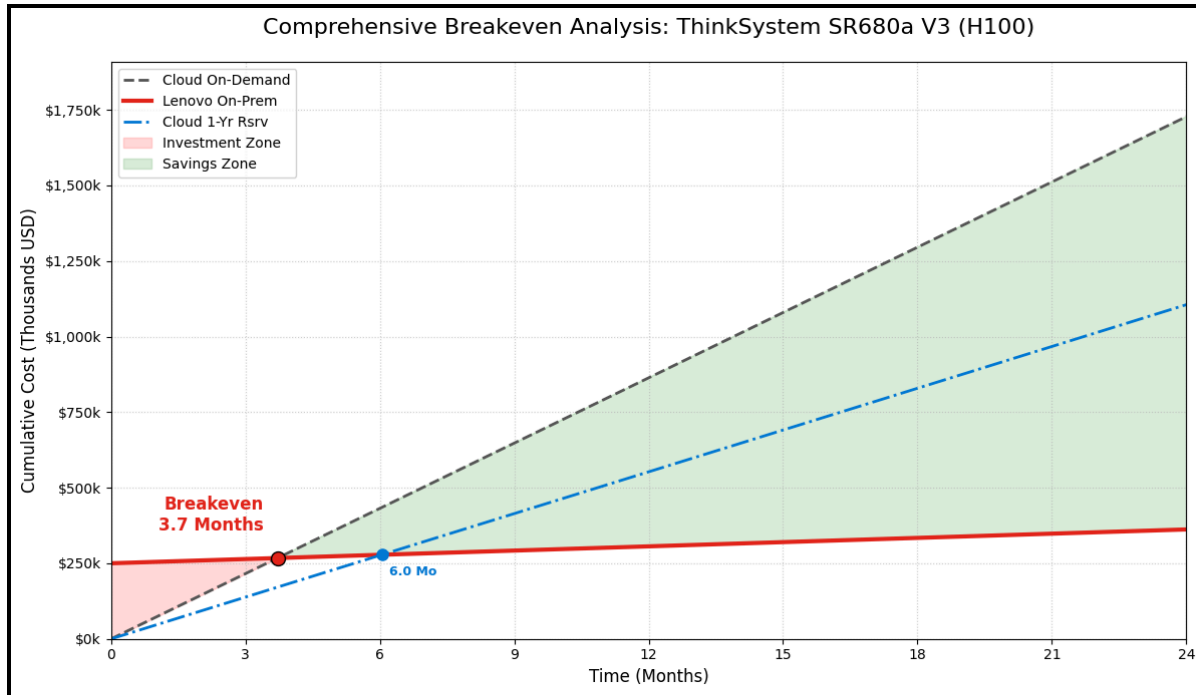


Figure 1: Breakeven Analysis for on-demand and 1-year reserved instance for 8x H100 on cloud vs on-prem

Breakeven Calculation (vs. 3-Year Reserved):

$$250141.8 + 6.37x = 43.16x$$

$$36.79x = 250141.8$$

$$\Rightarrow x \approx 6,800 \text{ hours}$$

Result:~9.3 months to break even. Considering heavy discounts on cloud after committing for 3-years, Lenovo system pays for itself much under a year.

Breakeven Calculation (vs. 5-Year Reserved):

$$250141.8 + 6.37x = 39.32x$$

$$32.95x = 250141.8$$

$$\Rightarrow x \approx 7,591 \text{ hours}$$

Result:~10.4 months to break even. So, you will be saving money for nearly 50 months after the breakeven!

Similar to Figure 1, the figure below showcases the breakeven comparison but for 3-year reserved instance and showcases the tiny change when renting a 5-year reserved instance. Still, breakeven is achieved in under a year!

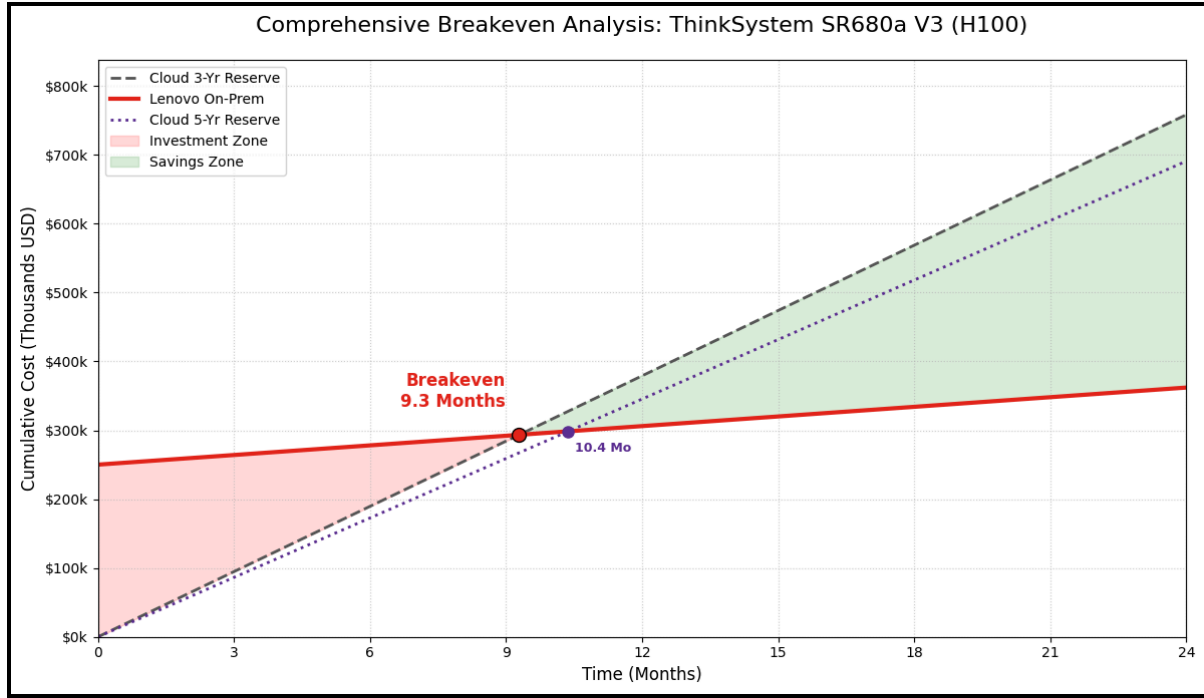


Figure 2: Breakeven Analysis for 3-year and 5-year reserved instance for 8x H100 on cloud vs on-prem

Case 2: Total Cost of System Over Time (5-Year Lifecycle)

Scenario: A 5-year comparison of the flagship **Config D (8x B300)** versus the **AWS p6-b300.48xlarge**.

- **Cloud Cost (AWS p6-b300):**
 - Hourly Rate: \$142.42
 - Total 5-Year Cost (24/7): $142.42 \times 24 \times 365 \times 5 = \$6,238,000$
- **On-Prem Cost (Lenovo Config D):**
 - CapEx: \$461,567.5
 - OpEx (Hourly) : ~\$12.6 (\$6.32 maintenance + \$4.2 power and cooling + \$2.08 colocation)
 - Total 5-Year Cost: $461567.5 + (12.6 \times 43,800) = \$1,013,447$

Total Savings: \$5,224,552 (83.8% Savings!) Over a 5-year lifecycle, owning the B300 infrastructure saves over \$5.2 million per server compared to the hourly cloud rate.

In the figure below, we can see the yearly (cumulative) savings by owning a Lenovo ThinkSystem compared to renting on cloud.

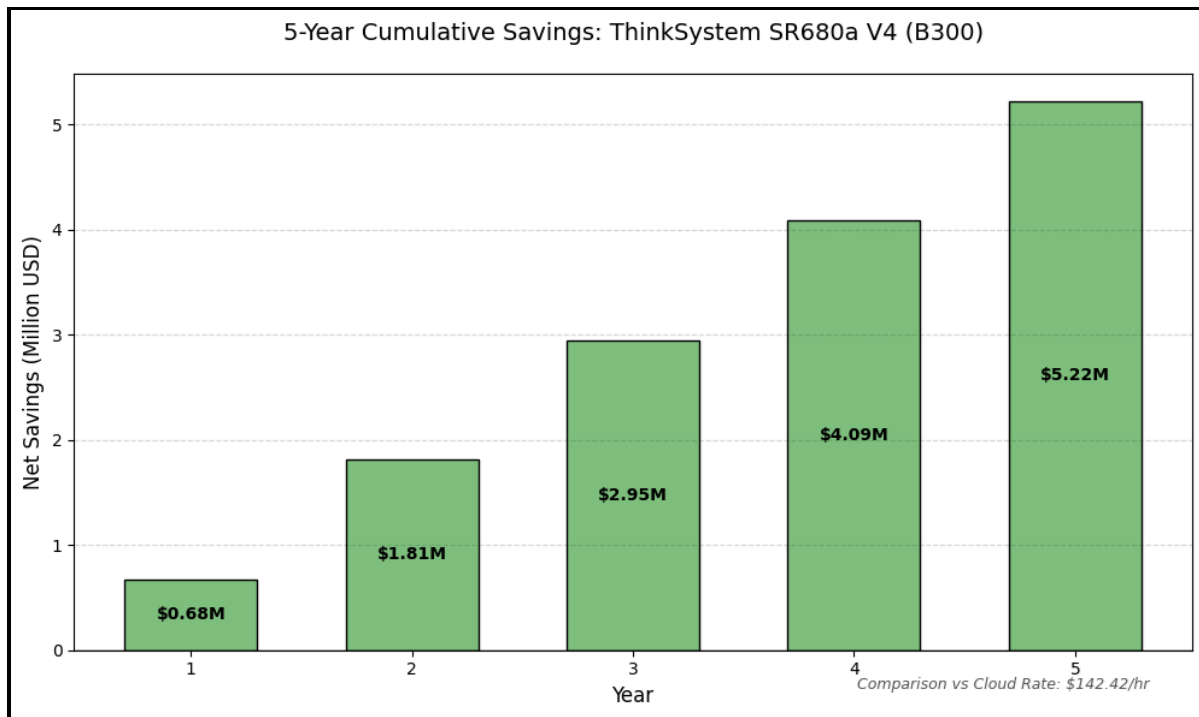


Figure 3: Cumulative Savings on 8x B300 system over a life cycle

Case 3: Hourly Utilization Threshold

Scenario: Comparing **Config B (8x H200)** against **GCP a3-ultragpu-8g**. At what daily utilization does owning become cheaper than renting?

- **Cloud Rate (GCP On-Demand):** \$84.81/hr
- **On-Prem Cost (Config B):**
 - CapEx: \$277,897.75
 - OpEx (Hourly): \$8.79 (\$3.8 maintenance + \$2.91 power and cooling + \$2.08 colocation)
 - Total 5-Year On-Prem Cost (Always On): 277898 + (8.79 × 43,800) = 662,900

Calculation:

$$\begin{aligned} \text{Total OnPrem Cost} &= \text{Cloud Hourly} \times \text{Hours/Day} \times 365 \times 5 \\ 662,900 &= 84.81 \times H \times 1825 \\ 662,900 &= 154,778 \times H \\ \Rightarrow H &\approx 4.3 \text{ hours/day} \end{aligned}$$

Insight: If the system is used for just **4.3 hours per day**, purchasing the Lenovo SR680a V3 is more economical than renting from Google Cloud over a 5-year period.

In the following figure, we see the total savings or loss associated with utilization and the threshold which becomes the tipping point.

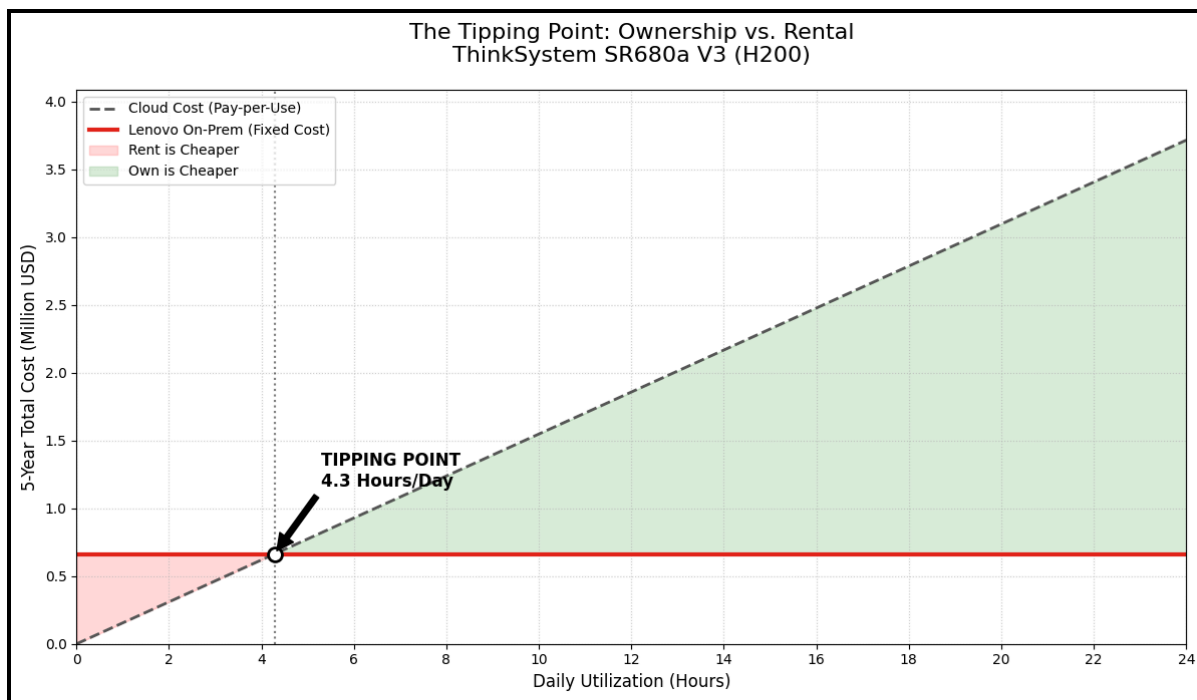


Figure 4: Utilization Threshold for on-prem systems to breakeven

Token Economics

In this section, we apply the financial data to the actual throughput capabilities of these systems. We utilize MLPerf Server inference speeds to establish a realistic baseline for output token generation.

- [MLPerf Inference Benchmarks \(Output Tokens/Sec\)](#)
- [Cost Per Million Tokens Analysis](#)

MLPerf Inference Benchmarks (Output Tokens/Sec)

MLPerf benchmarks serve as a crucial impartial platform to compare systems and their throughput on different GenAI models among others. The following table provides a comprehensive overview with results taken from MLPerf Server Inference (v5.0/v5.1) .

Table 5. Inference speeds of models across different configurations

| Config | Llama 70B -99 FP32 | Llama 3.1 405B FP16 | Mixtral-8x7b FP16 | GPT-J -99 FP32 |
|----------------------------|--------------------|---------------------|-------------------|----------------|
| Config A (8x H100) | 30,576.50 | 277.13 | 53,299.3 | 20,548.9 |
| Config B (8x H200) | 32,955.55 | 291.38 (mix) | 61,195.7 | 21,551.25 |
| GCP a3-ultra (H200) | 31,930.23 | 276.43 | 59,100.1 | 20,501.8 |
| Config C (8x B200) | 99,159.27 | 1,249.04 | 128,284.3 | -- |
| Config D (8x B300)* | 101,608 | 1,360 | -- | -- |
| Config E (4x L40S) | 1,469.19 | -- | -- | 3,096.45 |

* These speeds represent raw output token throughput from MLPerf Inference Benchmarking. Actual system pipelines may introduce latency or overhead that affects these speeds, but the relative performance delta between systems remains consistent.

The following figure showcases the generational leap from Hopper to Blackwell resulting in more than 3x speedup in throughput for the same model.

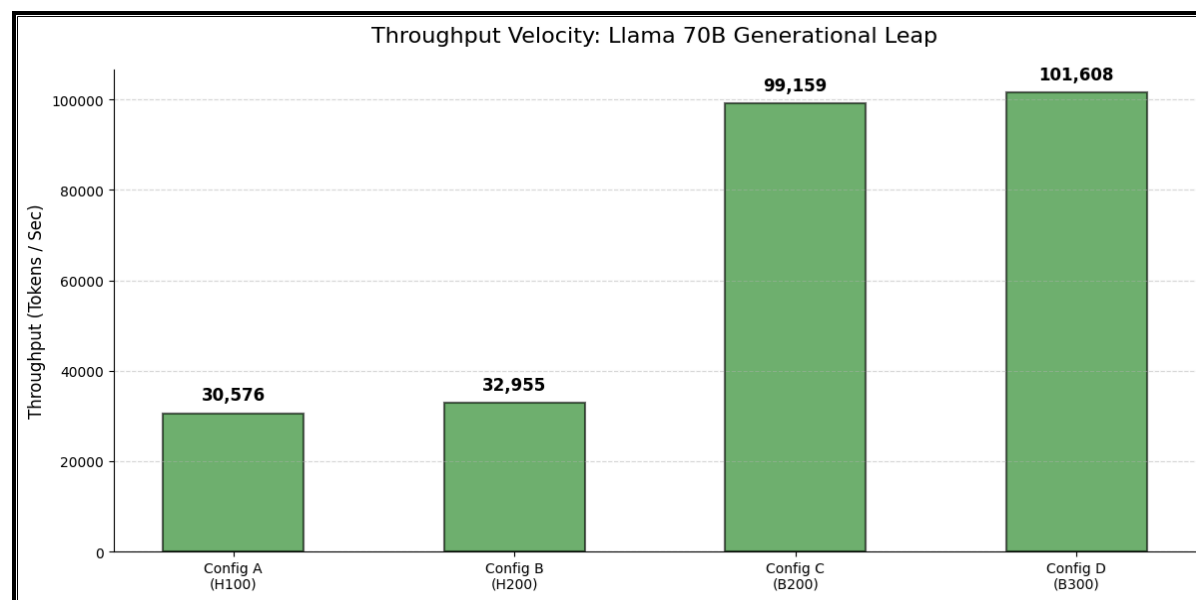


Figure 5. Throughput speeds compared across generations for 70B model

Cost Per Million Tokens Analysis

Using the 5-year amortized costs calculated in the [Total Cost of Ownership \(TCO\) Analysis](#) section, we can determine the cost to generate 1 million tokens for specific high-value models.

Scenario A: Llama 70B Inference (Config A vs Azure H100)

- **Lenovo Config A (8x H100):**
 - Hourly Cost (Amortized): **\$12.08**
 - Speed: **30,576 tokens/sec**
 - Cost per 1M Tokens: $1M/30,576 * 12.08/3600 = \mathbf{\$0.11}$
- **Azure H100 (On-Demand):**
 - Hourly Rate: **\$98.32**
 - Speed: **30,576 tokens/sec** (Assumed parity)
 - Cost per 1M Tokens: **\$0.89**
- **Result: 8x Cheaper On-Prem**

Scenario B: Proprietary API Comparison

- **GPT-5 mini API:** ~\$2.00 / 1M output tokens
- **Lenovo Config A (70B Model):** \$0.11 / 1M output tokens (**18x Cheaper**)

The results in the figure below show consistently that on-prem systems come out to be the cheapest for serving models compared to the cloud counterparts. Also, the cheapest frontier model still is orders of magnitude costlier than in-house, privately hosted models.

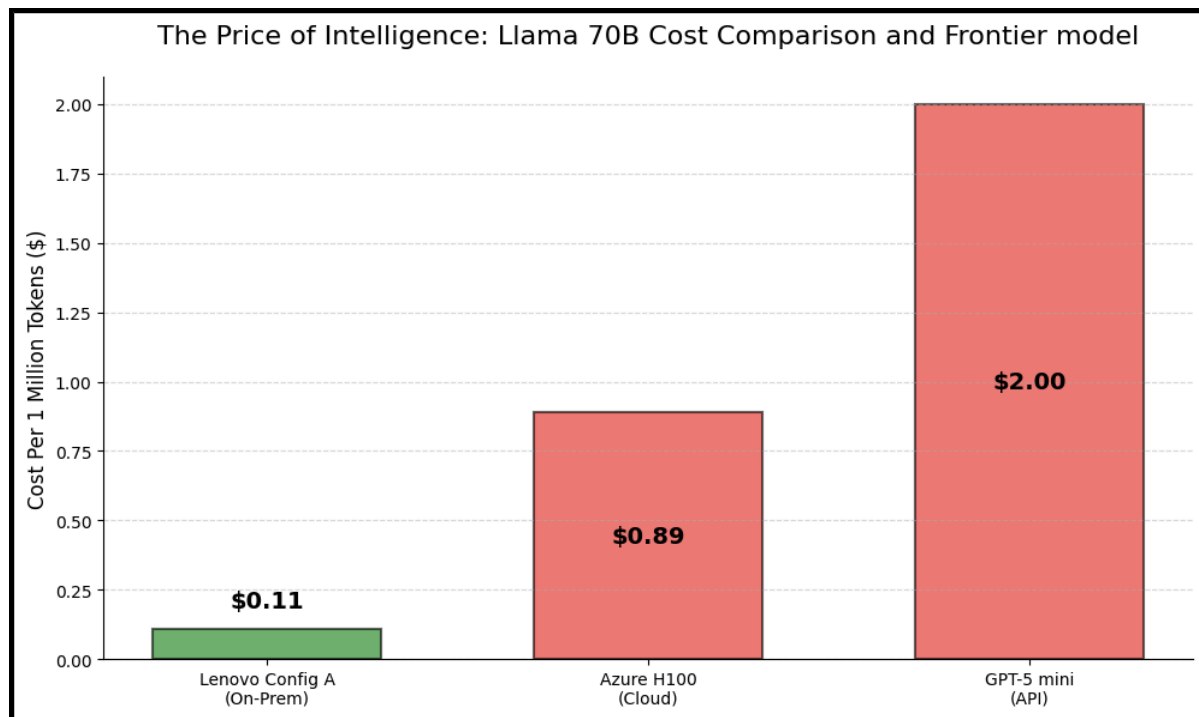


Figure 6: Cost per million token comparison between on-prem vs cloud vs frontier model

Scenario C: Llama 3.1 405B Inference (Config D vs AWS B300)

- **Lenovo Config D (8x B300):**
 - Hourly Cost (Amortized): **\$23.19**
 - Speed: **1,360 tokens/sec**
 - Cost per 1M Tokens: $1M/1360 * 23.19/3600 = \mathbf{\$4.74}$

- **AWS B300 (On-Demand):**
 - Hourly Rate: **\$142.42**
 - Speed: **1,360 tokens/sec**
 - Cost per 1M Tokens: **\$29.09**
- **Result: 84% Cheaper On-Prem**

In the figure below, we see the extreme difference in the cost for serving models, making Lenovo ThinkSystems an easier choice under constant workloads.

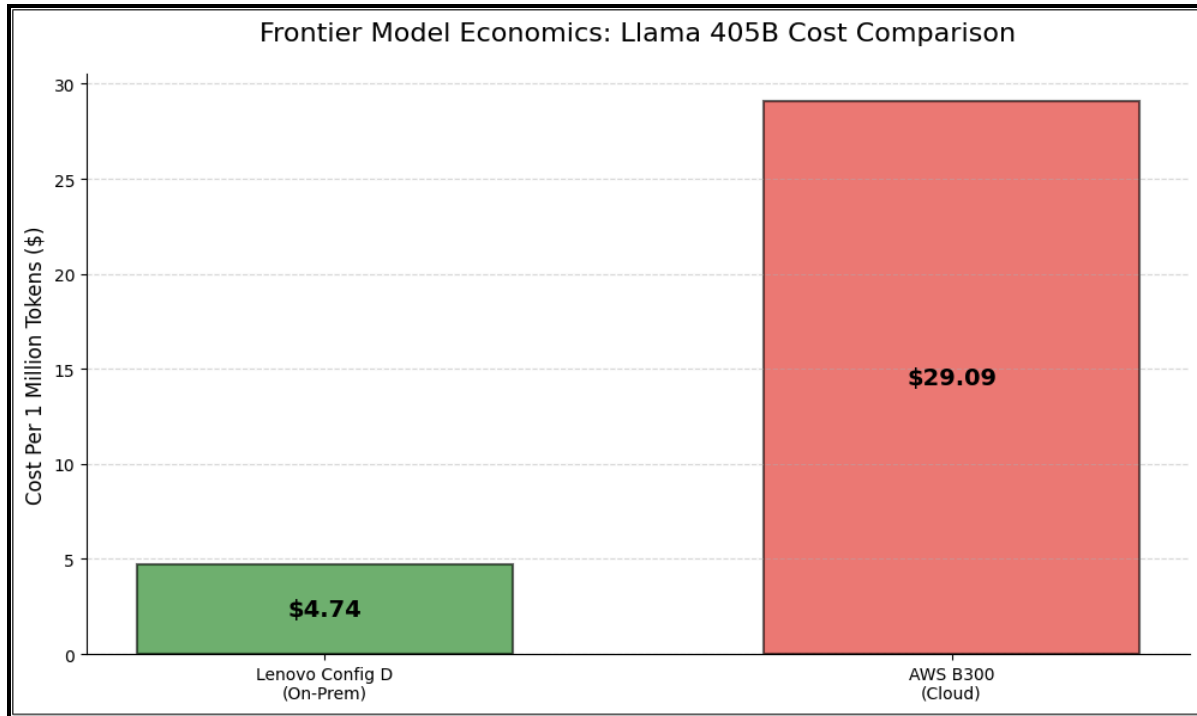


Figure 7: Large 405B model tokens per million comparison

The Accelerated Computing Ecosystem

As we approach 2026, the ecosystem for accelerated computing has matured into a diverse landscape. While the current market is heavily defined by the transition from NVIDIA Hopper to Blackwell architectures, the principles of high-performance computing remain universal.

High-Density Computing (Config C & D) The Lenovo ThinkSystem SR680a V4 represents the pinnacle of density. Whether utilizing NVIDIA B300 or future high-performance accelerators, these systems are designed for the massive thermal envelopes (1000W+ per chip) required by frontier model training.

Versatile Enterprise AI (Config A & B) Systems like the ThinkSystem SR680a V3 serve as the backbone of enterprise AI. By supporting industry-standard accelerators like the H100 and H200, they offer a balance of memory bandwidth and compute capability that is ideal for fine-tuning and serving models in the 70B parameter class.

Efficient Edge Inference (Config E) The ThinkSystem SR650a V4 demonstrates that effective AI does not always require the most expensive hardware. By utilizing efficient accelerators like the L40S, organizations can deploy competent inference nodes for smaller models (8B-70B) at the edge.

Sustainable AI & Green Computing

As AI data center power consumption is projected to double by 2028, efficiency is no longer optional—it is operational.

- **Liquid Cooling Efficiency:** Lenovo's Neptune™ liquid cooling technology (available in the SR680a V4) significantly reduces the PUE (Power Usage Effectiveness) from the industry average of 1.5 down to 1.1. This reduction in cooling overhead directly lowers the "fully burdened" kWh cost used in our calculations, potentially improving TCO by an additional 10-15%.
- **Carbon Impact:** By running dedicated, high-utilization clusters on-premise, organizations can schedule training jobs during off-peak hours when the grid is greener and cheaper, a flexibility often penalized by "on-demand" cloud pricing models.

Conclusion

As we move through 2026, the economic case for on-premises Generative AI infrastructure has solidified. The era of "cloud-first" for all AI workloads is over. While the cloud remains essential for bursty training and experimentation, the **Total Cost of Ownership analysis decisively favors on-premises infrastructure for sustained inference and fine-tuning workloads.**

- **Breakeven:** Achieved in **<4 months** for high-utilization environments against On-Demand pricing.
- **Token Cost:** Self-hosting on Lenovo hardware offers an **8x cost advantage** per million tokens compared to Cloud IaaS, and up to **18x** compared to frontier Model-as-a-Service APIs.
- **Lifecycle:** Over a standard **5-year lifecycle**, the savings per server can exceed **\$5 million**, freeing up massive capital for further innovation.

For enterprises committed to AI as a core competitive advantage, the transition from renting intelligence to owning the factory is not just a technical evolution, it is a financial imperative.

Authors

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 Rutgers University as a gold medalist specializing in ML and has secured the J.N. Tata Scholarship.

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.

Jarrett Upton is the AI Center of Excellence Lab Manager, where he leads the deployment and demonstration of cutting-edge AI solutions. With a strong focus on enabling customer proof-of-concept testing and validating independent software vendor (ISV) integrations. Jarrett plays a pivotal role in accelerating enterprise adoption of AI technologies. He oversees the design and operations of Lenovo's AI Lab, driving innovation and collaboration across product teams, partners, and clients.

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.

© Copyright Lenovo 2026. All rights reserved.

This document, LP2368, was created or updated on February 4, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
<https://lenovopress.lenovo.com/LP2368>
- Send your comments in an e-mail to:
comments@lenovopress.com

This document is available online at <https://lenovopress.lenovo.com/LP2368>.

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.

Microsoft® and Azure® are trademarks 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.