On-Premise vs Cloud: Generative AI Total Cost of Ownership

Positioning Information

In recent years, Generative AI, including Large Language Models (LLMs) and vision models, has emerged as a transformative technology in artificial intelligence, driving innovation across industries. However, deploying these models—whether for training, fine-tuning, or inference—poses significant computational challenges.

The scale of data in GenAI is staggering. Models like Llama 3.1, trained on over 15 trillion tokens using a custom-built GPU cluster with 39.3 million GPU hours, illustrate the immense computational demands. Such training can be prohibitively expensive when relying on cloud services. Hypothetically, running on AWS P5 instance H100 system will run you over $483 M in cloud costs ignoring the storage requirements of the training data. Organizations must carefully evaluate deployment strategies, weighing the total cost of ownership (TCO) of on-premises infrastructure against cloud services.

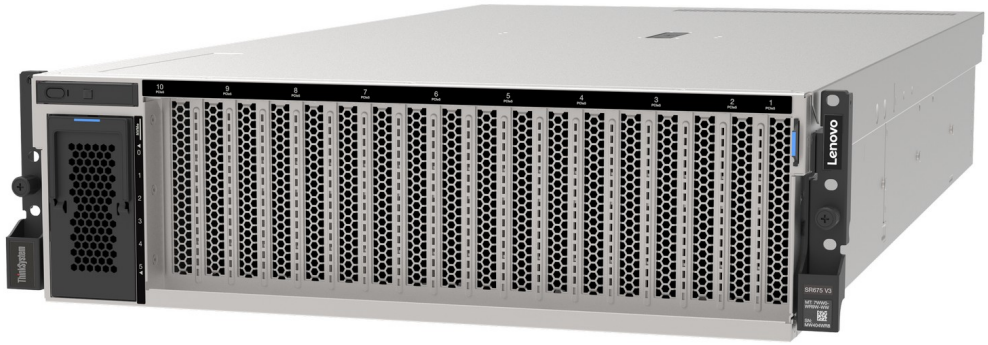

Figure 1. Lenovo ThinkSystem SR675 V3 with support for eight double-wide GPUs is an ideal on-prem Generative AI server

GenAI models typically operate in two key phases: **training** and **inference**. Training involves processing massive datasets—often measured in tens of trillions of tokens—requiring substantial compute resources over long periods. Inference, though less compute-intensive per request, demands continuous, low-latency responses at scale, especially as user demand grows. For both prolonged training and persistent inference at high throughput, **on-premises infrastructure offers significant advantages**. The fixed nature of capital expenditure (CapEx), combined with optimized utilization of dedicated GPUs, makes on-prem a more cost-efficient option over time. In contrast, cloud costs scale linearly with usage, making them ideal for short-term or burst workloads but economically inefficient for sustained GenAI operations.

Component Cost Analysis: On-Prem vs. Cloud Deployment

When deciding between on-premises and cloud deployment for Generative AI models, understanding the associated costs is crucial. Each approach comes with unique financial implications that can significantly impact a company's budget and operational efficiency.

Table 1. Component Cost Analysis of On-premise v Cloud

Cost Element	On-Premises	Cloud
Capital Expenditure (CapEx)	High initial investment (servers, GPUs, etc.)	None; operationalized as pay-as-you-go
Operational Expenditure (OpEx)	Power, cooling, staffing, facility overhead	Ongoing subscription, storage, bandwidth
Software & Licensing	Purchased outright or annually	Included in service fees or usage-based
Scalability	Limited to physical capacity or added rent/ leasing costs	Virtually unlimited and elastic
Data Privacy & Security	Full control, customizable, Data stays within owned infrastructure	Data handled off-site; subject to provider policies.
Refresh Cycle & Depreciation	3–5-year lifecycle, refresh required	Abstracted, continuous upgrades by provider

On-Premises Deployment Costs

On-premises infrastructure involves high upfront costs for servers, GPUs, storage, and physical infrastructure like cooling and space. These systems must be maintained regularly, and ongoing expenses such as electricity and software licensing add to the total cost of ownership. However, this model offers cost predictability and efficiency over time, especially for stable, long-term AI workloads where infrastructure utilization is consistently high. The costs include:

- GPU instance cost: This includes the cost of the GPU instance itself, which can vary depending on the type and quantity of GPUs required.
- Memory and storage costs: The cost of memory and storage required to support the GPU instance should also be considered.
- Power and cooling costs: The power and cooling requirements of the GPU instance can impact the overall cost of ownership.

Scope of Cost Analysis

While total cost of ownership can include numerous factors—such as operating system and application licensing, patching, networking costs, IT staffing, network variability across regions, and software stack maintenance—for the purpose of this whitepaper, we limit our analysis to infrastructure costs (primarily compute hardware) and power and cooling expenses. This simplification allows for a focused comparison between on-premises and cloud deployments under typical high-throughput AI workloads, though it is important to note that real-world deployments may incur additional costs outside the scope of this analysis.

A more comprehensive cost breakdown, incorporating these additional variables, is planned to be included in future updates to this paper.

Cloud Deployment Costs

Cloud infrastructure offers flexibility through pay-as-you-go pricing and managed services, reducing upfront capital expenditure. However, costs can quickly escalate due to data transfer, storage, retrieval, and usage-based pricing. Long-term commitments provide discounts but limit agility, while variable workloads and vendor lock-in complicate cost forecasting. For dynamic or short-term workloads, cloud remains advantageous, but for sustained usage, it often proves more expensive than on-prem as we will show in our analysis.

Scope of Cost Analysis

For this analysis, we compare the total cost of ownership for leading cloud providers—AWS, Google Cloud Platform (GCP), and Microsoft Azure—against Lenovo's on-premises infrastructure, focusing specifically on server acquisition, power consumption, and cooling. To maintain consistency and simplicity, we exclude ancillary cloud costs such as those associated with managed services (e.g., AWS Bedrock), data storage (AWS EBS or S3), or data transfer. Instead, we base our comparison on the hourly compute pricing of EC2-like instances. We evaluate both on-demand pricing and discounted rates through 1-year, and 3-year savings plans and directly compare these to the hourly cost of operating on-premises infrastructure.

Considerations

Both on-premises and cloud deployments present distinct cost profiles that require careful consideration. On-premises solutions demand significant upfront investments and ongoing operational expenses, while cloud deployment offers flexibility but introduces risks of unexpected charges, vendor lock-in and long-term commitments. On-prem also offers greater control over sensitive data, as all storage and processing remain within the organization's own network perimeter. In contrast, cloud environments may pose higher privacy risks due to third-party data handling and shared infrastructure, making regulatory compliance and data sovereignty more complex. Understanding these factors is essential for making an informed decision that aligns with the organization's financial capacity and strategic goals.

Comparing On-prem vs Cloud costs

We will evaluate TCO across three core scenarios and a representative set of seven server configurations. This analysis focuses on three key GPU types commonly used in GenAI workloads: NVIDIA H100, H200, and L40S. For each on-premises Lenovo server configuration, we identify the equivalent cloud instance offered by cloud providers such as AWS and GCP.

Table 2. Configuration of servers compared for the study

Server	On-Prem Configuration	Cloud Equivalent
Server 1	ThinkSystem SR675 V3 - 8x NVIDIA H100 NVL 94GB PCIe Gen5	Amazon AWS EC2 p5.48xlarge - 8x NVIDIA Tesla H100 GPUs
Server 2	ThinkSystem SR675 V3 - 8x NVIDIA H200 NVL 141GB PCIe Gen5	Amazon AWS EC2 p5en.48xlarge - 8x NVIDIA Tesla H200 GPUs
Server 3	ThinkSystem SR675 V3 - 4x NVIDIA H100 NVL 94GB	Google GCP a3-highgpu-4g - 4x NVIDIA H100 80GB GPUs
Server 4	ThinkSystem SR675 V3 - 8x NVIDIA L40 48GB PCIe*	Google GCP a2-ultragpu-8g - 8x NVIDIA A100 80GB GPUs
Server 5	ThinkSystem SR650 V3 - 1x NVIDIA L40S 48GB, Intel Xeon Silver 4514Y (16 cores, 150W, 2.0GHz)	Amazon AWS EC2 g6e.8xlarge - 1x NVIDIA L40S Tensor Core GPU
Server 6	ThinkSystem SR650 V3 - 1x NVIDIA L40S 48GB PCIe Gen4, Intel Xeon Gold 6530 (32 cores, 270W, 2.1GHz)	Amazon AWS EC2 g6e.16xlarge - 1x NVIDIA L40S Tensor Core GPU
Server 7	ThinkSystem SR650a V4 - 4x NVIDIA L40S 48GB PCIe Gen4, Intel Xeon 6747P (48 cores, 330W, 2.7GHz)	Amazon AWS EC2 g6e.24xlarge - 4x NVIDIA L40S Tensor Core GPUs

* The A100 GPU is not used in this configuration because the A100 is now withdrawn from marketing

We provide three different comparison cases for the equivalent servers. For brevity, we have only posted calculations for Server 1 in this section. All calculations and figures for other servers are available in the Appendix.

- Case 1: Breakeven Point Analysis
- Case 2: Total Cost of Ownership and Savings Over Time
- Case 3: Hourly Utilization Threshold

Case 1: Breakeven Point Analysis

This scenario identifies the breakeven point—the point in time where the cumulative cost of cloud infrastructure matches the total investment in on-premises infrastructure. Prior to this point, cloud solutions may be more cost-effective. Beyond it, on-premises infrastructure delivers greater long-term savings.

Let's take Server 1 for instance, the ThinkSystem SR675 V3, equipped with 8x NVIDIA H100 NVL 94GB PCIe Gen5 GPUs. The cloud equivalent for this server is the Amazon EC2 P5 Instance (p5.48xlarge), which offers the same 8x NVIDIA H100 GPUs, along with 192 VCPUs and 2048 GiB memory. For this analysis, we are focusing on a single server configuration rather than a full rack, enabling a more focused and practical TCO comparison.

- On-Demand Cost for cloud instance: $98.32 per hour (at the time of writing)
- 1-Year Reserved Instance Cost: $77.43 per hour
- Estimated On-Prem Power & Cooling Cost: ~$0.87 per hour (at $0.15/kWh for server + HVAC)
- On-Prem Total System Cost: ~$833,806 (considering no sales discounts)

To calculate the breakeven point—the number of hours at which the total cost of using cloud services equals the total cost of on-premises infrastructure—we compare the two cost models:

- cloud_cost = 98.32 * x
- onprem_cost = 0.87 * x + 833806

To find the breakeven point, we set the two equations equal:

98.32 * x = 0.87 * x + 833,806.00

Solving for x:

x ≈ 8556 hours

So, the breakeven point is reached at approximately 8,556 hours or **11.9 months** of usage. Beyond this point, operating on-prem infrastructure becomes more cost-effective than continuing with cloud services. Plotting these two equations we can visualize the breakeven point and the savings region.

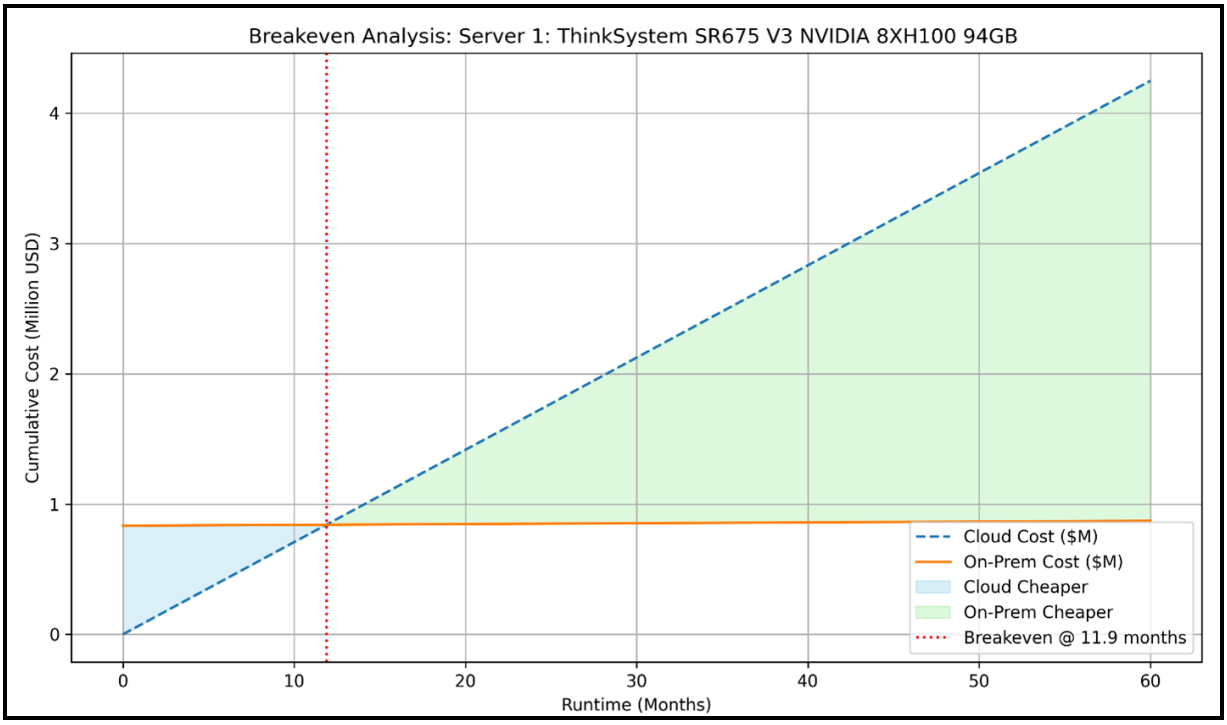

Figure 2. Breakeven analysis of Server 1 considering hourly costs

For the same server, instead of the hourly costs we can repeat the breakeven calculation using discounted hourly rates from AWS savings plans:

- **1-Year Reserved Instance Cost:**
  Breakeven: x = 833,806.00 / (77.427 - 0.87) ≈ 10,890 hours ~ 15.13 months
- **3-Year Reserved Instance Cost:**
  Breakeven: x = 833,806.00 / (53.945 - 0.87) ≈ 15710 hours ~ 21.82 months (a little less than 2 years)

These longer breakeven points indicate that while reserved pricing offers lower hourly cloud rates, **on-prem infrastructure still becomes more cost-effective in the long run for sustained usage**.

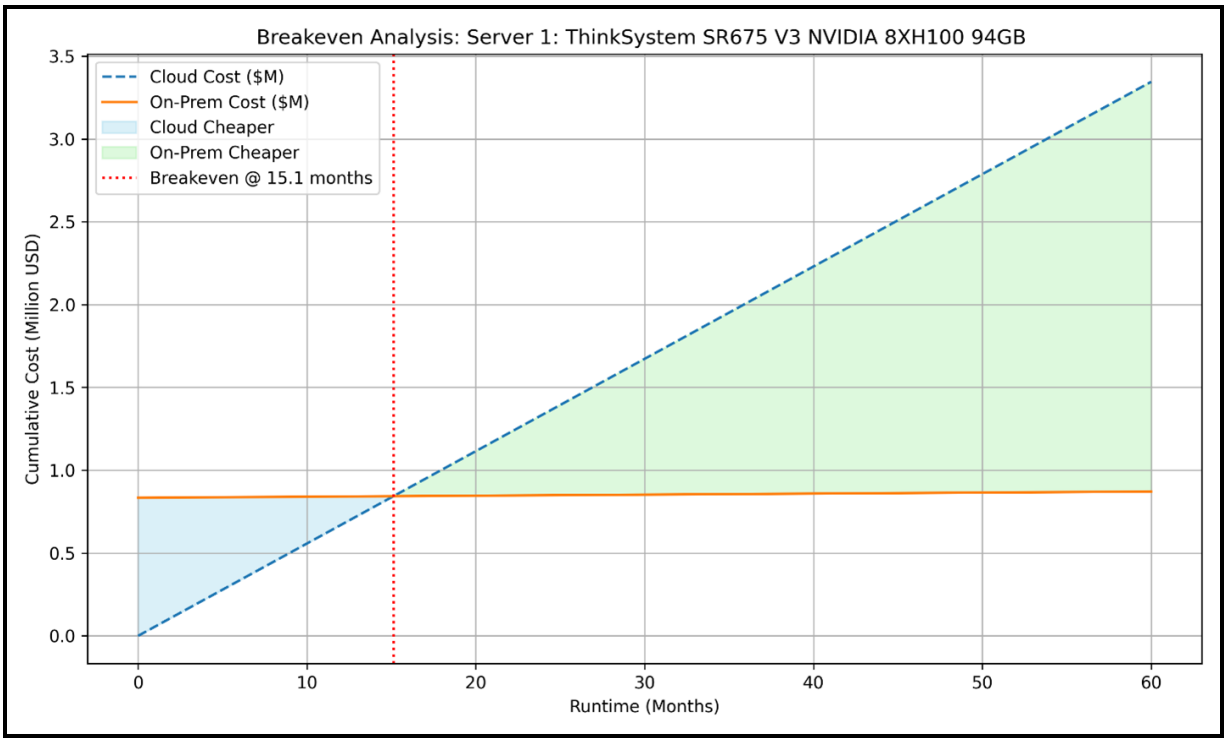

Figure 3. Breakeven analysis of Server 1 considering yearly discounted cost

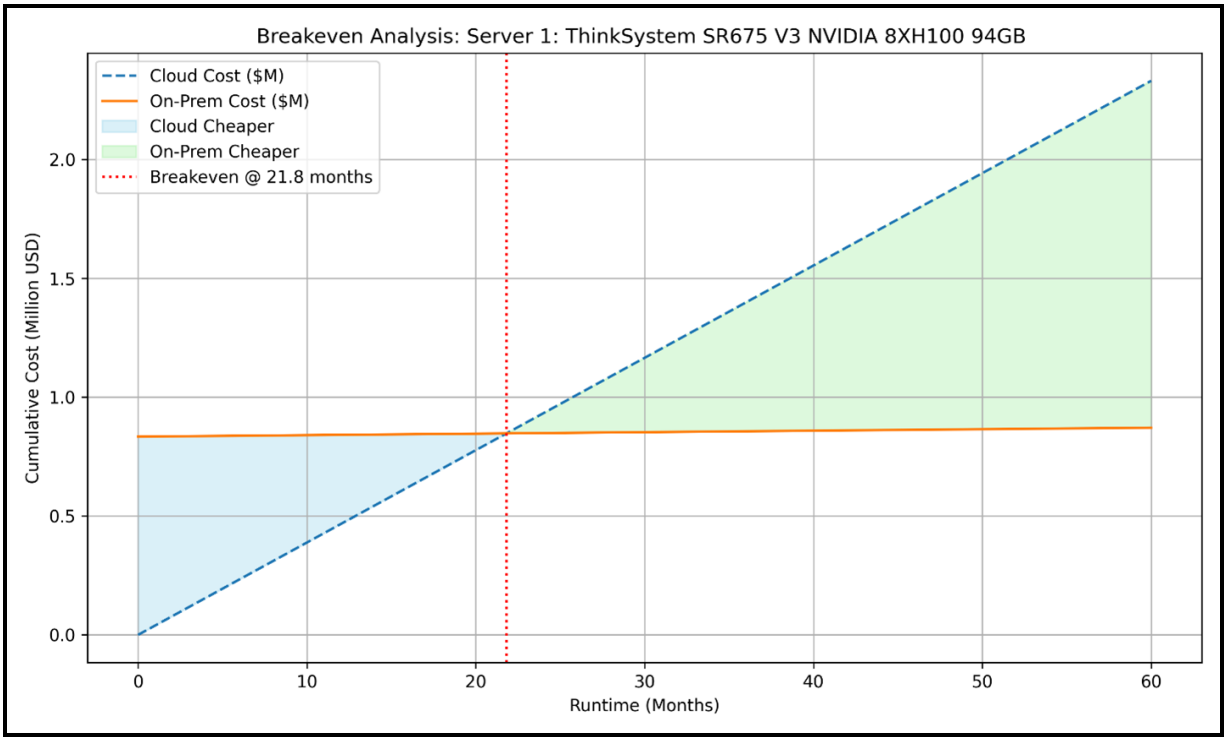

Figure 4. Breakeven analysis of Server 1 with 3-year savings plan

Case 2: Total Cost of Ownership and Savings Over Time

Assuming a 5-year operational lifespan for on-premises servers, this scenario compares total costs over time. It quantifies both annual and cumulative savings achieved by using on-prem infrastructure instead of cloud services. A 5-year lifespan means that we let the server fully depreciate with no recovery value. This means that when you purchase let's say an NVIDIA H100 GPU, you spread the purchase cost over its useful life. To understand long-term cost implications, we calculate the total 5-year cost of running the system continuously (24 hours per day) on both cloud and on-premises infrastructure. This helps quantify cumulative savings over a typical server lifecycle.

Assumptions:

- Continuous operation: 24 hours/day for 5 years
- Total hours over 5 years: 24 × 365 × 5 = 43,800 hours

Cost Calculations:

- Cloud 5-Year Cost: cloud_hourly × 43,800
- On-Prem 5-Year Cost: onprem_base_cost + (onprem_hourly × 43,800)
- Savings: cloud_cost_5yr - onprem_cost_5yr

Taking Server 1 ThinkSystem SR675 V3 — 8x NVIDIA H100 NVL 94GB PCIe Gen5 and cloud equivalent Amazon EC2 P5 Instances (p5.48xlarge) with 8× NVIDIA H100 GPUs as an example.

- **On-Prem Cost:** $833,806 + (0.87 x 43800) = $871,912
- **Cloud Cost:** $98.32 * 43800 = $4,306,416.00
- **Total Savings Over 5 Years:** $ 3,434,504

The following figure illustrates both the **annual savings** and the **cumulative cost difference** between on-premises and cloud deployments over a 5-year period, highlighting the growing financial advantage of on-prem infrastructure with sustained usage.

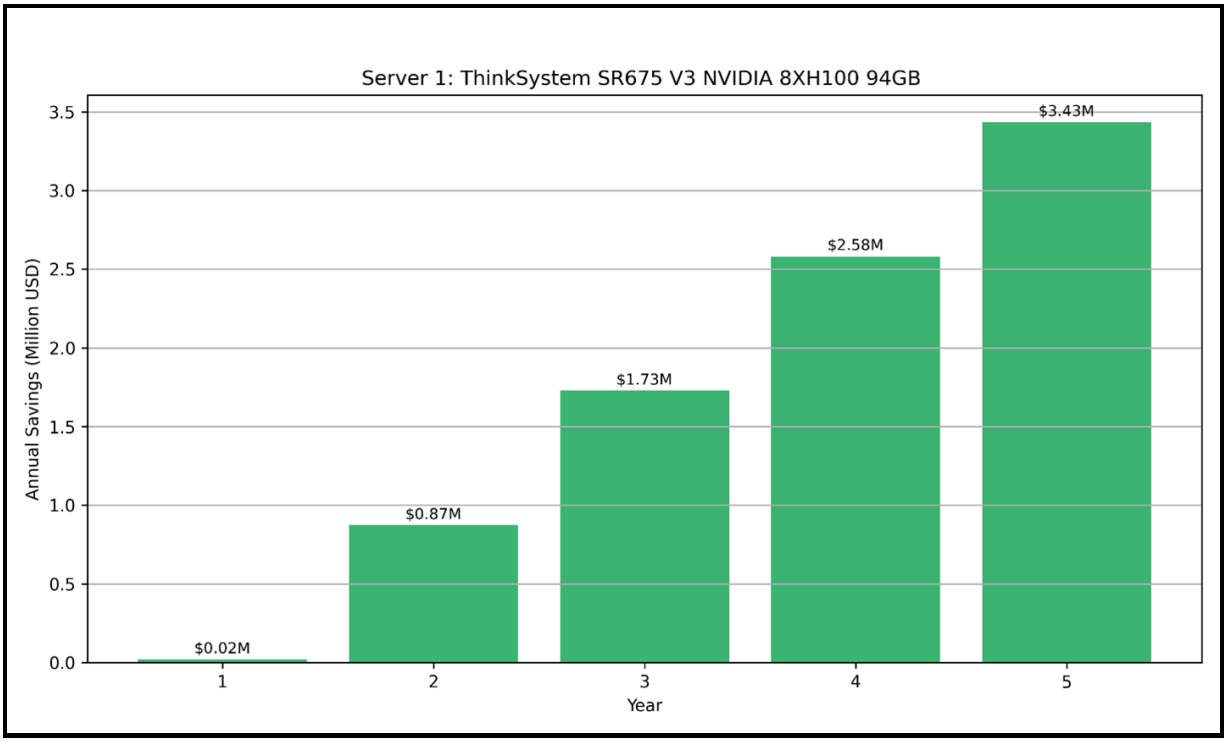

Figure 5. Savings for Server 1 when using on-premise compared to cloud hourly costs

Considering 1 Year Savings plan cost:

- Yearly savings plan cloud cost for 5 years: $77.427 * 43800 = $3,391,302.6
- Total savings over 5 years: $2,519,390.6

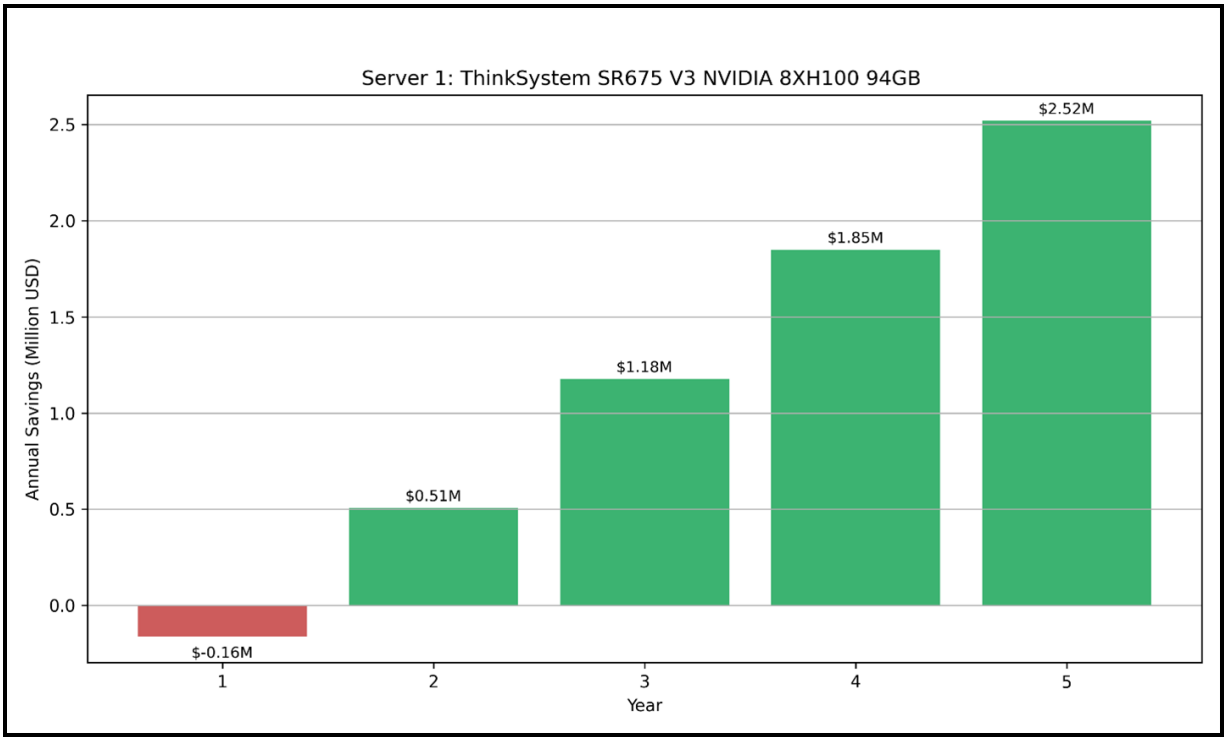

Figure 6. Savings for Server 1 when using on-premise compared to cloud yearly plan

Considering 3 year Savings plan cost:

- 5 year Cloud cost= $ 53.94547 x 43800 = $2,362,811.59
- Total savings over 5 years: $1,490,899.59

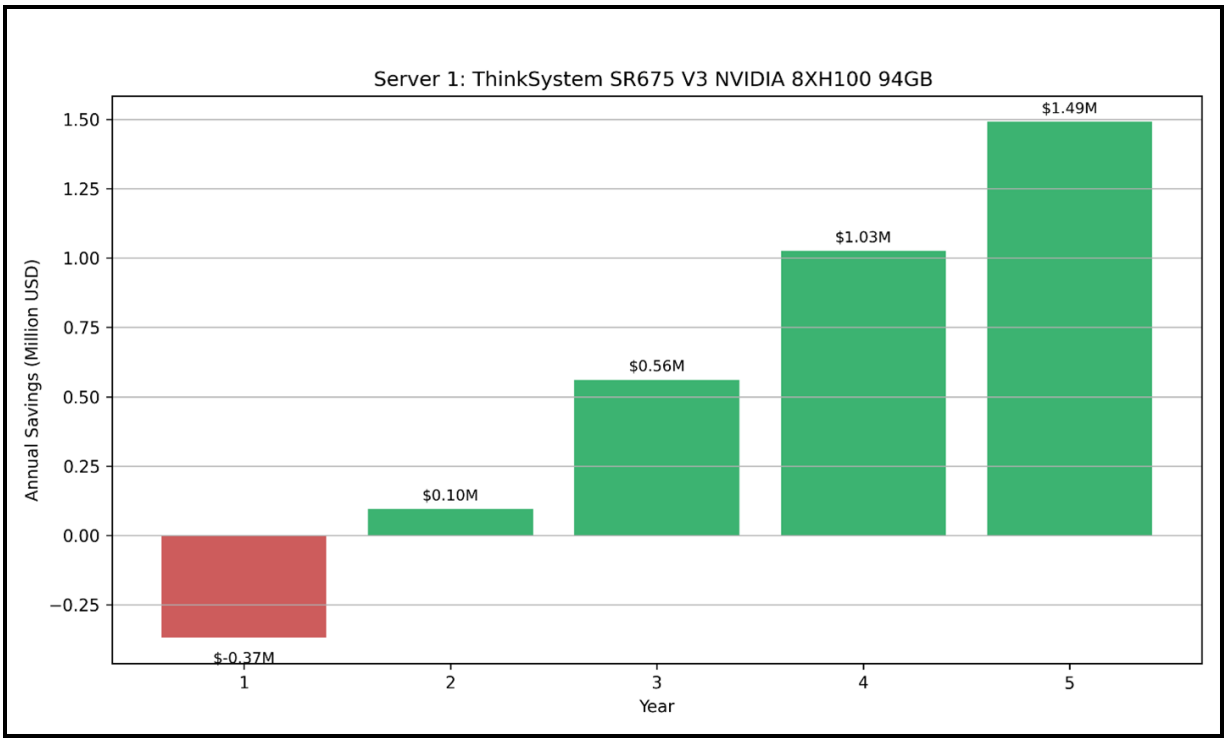

Figure 7. Savings for Server 1 when using on-premise compared to cloud 3-year plan

Case 3: Hourly Utilization Threshold

This scenario determines the **minimum number of hours per day** that a system must be in use for on-premises infrastructure to become more cost-effective than cloud services over a 5-year period. This will help the organization evaluate whether their usage levels justify the capital investment in on-prem infrastructure. For simplicity, this analysis assumes **100% system and GPU utilization** during active hours, representing a high-demand scenario typical of **inference workloads**, where GPUs are consistently engaged. While actual usage may vary, this assumption helps define a clear breakeven threshold and serves as a reference point for scaling cost comparisons under lower utilization scenarios.

We first calculate the 5-year total cost of running both cloud and on-prem setups continuously (24 hours/day for 5 years, totaling 43,800 hours). Then, we compute the usage ratio:

- Usage Ratio = On-Prem 5-Year Cost / Cloud 5-Year Cost

This ratio reflects the proportion of time at which on-prem and cloud costs are equal. To find the daily threshold:

- Daily Threshold (in hours) = Usage Ratio × 24

Example:

- On-Prem 5-Year Cost = $1,000,000
- Cloud 5-Year Cost = $4,000,000
- Usage Ratio = 1,000,000 / 4,000,000 = 0.25
- Daily Threshold = 0.25 × 24 = 6 hours/day

Interpretation: If your system runs more than **6 hours per day on the cloud**, it becomes more expensive than running the same workload on a purchased on-prem server. This analysis is particularly useful for organizations with regular, moderate workloads that don't run 24/7 but still require predictable and economic performance.

From the calculations above for Server 1 for hourly cloud costs, we can calculate the Usage Ratio as:

- $871, 912 / $4,306,416 = 0.2025
- Daily threshold = usage ratio x 24 = 4.93 ~ 5 hours a day.

Similarly for 1 year and 3-year savings plan:

- 1 year cloud costs: ($871, 912 / $ 3,391,302.6) x 24 = 6.17 hours a day
- 3-year costs: ($871, 912/ $ 2,362,811.59) x 24 = 8.86 ~ 9 hours a day

From the above calculations we can infer that as AWS Savings Plans (such as 1-year and 3-year reserved pricing) offer lower hourly rates compared to on-demand pricing, they **raise the hourly utilization threshold** at which on-prem infrastructure becomes more cost-effective. In other words, under discounted cloud pricing, you must run the system for **more hours per day** to justify the cost of purchasing on-prem hardware.

Conclusion

This analysis highlights a clear distinction between the strategic use cases for cloud and on-premises infrastructure in the era of Generative AI. Cloud platforms offer unmatched flexibility and scalability, making them ideal for short-term needs such as model experimentation, fine-tuning, or dynamic workloads. However, as usage becomes sustained and predictable, cloud costs can grow substantially due to recurring compute charges, data transfer fees, and storage costs.

In contrast, on-premises deployments require a larger upfront investment but deliver **significant long-term cost savings**, especially once capital expenditures are amortized. The concept of a breakeven point is essential—beyond this threshold, on-prem infrastructure consistently outperforms cloud options in terms of total cost of ownership.

For organizations considering deployment options, the duration of use is a decisive factor. Short-term requirements, such as retraining or fine-tuning models, often make cloud services the most practical choice due to their scalability and agility. Conversely, long-term workloads, such as serving models for inference, frequently reach a breakeven point where on-premises infrastructure becomes the more cost-effective option.

Lenovo's ThinkSystem servers further amplify these benefits with industry-leading performance, reliability, and energy efficiency. They are purpose-built for enterprise-grade AI workloads and provide a solid foundation for organizations committed to long-term GenAI initiatives, including large-scale inference and model serving.

Ultimately, the optimal deployment strategy depends on workload duration, usage intensity, and financial goals. For organizations anticipating continuous AI operations over multiple years, **on-premises infrastructure emerges as the most cost-effective and sustainable solution**.

Appendix

Use the links below to download the supporting documents and spreadsheets for plots, operational and server cost breakdowns, and a summary of all calculations.

- All Plots for Server 2-7: Appendix Plots.docx
- Summary for costs: Onprem_calculations.xlsx
- Operational (Power + Cooling) calculations for the server: Power_Calculations_Cost_v2.xlsx

References

For more information, see these web pages:

- Evaluating the Total Cost of Ownership for an On-Premise Application System (Kenny & Company) https://michaelskenny.com/wp-content/uploads/2017/08/POV-Evaluating-the-Total-Cost-of-Ownership-for-an-On-Premise-Application-System.pdf
- Amazon EC2 On-Demand Pricing https://aws.amazon.com/ec2/pricing/on-demand/
- Microsoft Azure Pricing calculator https://azure.microsoft.com/en-us/pricing/calculator/
- Google Cloud VM instance pricing https://cloud.google.com/compute/vm-instance-pricing?hl=en
- What Is Cloud TCO? (Total Cost of Ownership) (CloudZero) https://www.cloudzero.com/blog/cloud-tco/
- Unlocking Savings: Why Buying NVIDIA H100 GPUs Beat AWS Rental Costs (TRG Datacenters) https://www.trgdatacenters.com/resource/unlocking-savings-why-nvidia-h100-gpus-beat-aws-rental-costs/
- $ Cost of LLM continued pre-training (Medium) https://medium.com/@gilinachum/cost-of-llm-continued-pre-training-0c1998cb44ec

Authors

Sachin Gopal Wani is an AI Data Scientist at Lenovo, working on end-to-end Machine Learning (ML) applications for varying customers, and developing the NewTalk AI framework. He graduated from Rutgers University as a gold medalist specializing in Machine Learning and has secured the J.N. Tata Scholarship.

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.

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.

Matthew Ziegler is the Director of Lenovo Neptune and Sustainability at Lenovo's Infrastructure Solutions Group. He leads efforts in liquid-cooling and sustainability for data centers. Matthew began his career in life science research, spending a decade in the field before shifting his focus to the design and architecture of x86-based supercomputers (HPC) for various industries, including life sciences, energy, digital media, and atmospheric sciences. He joined IBM in 2003, where he broadened his HPC expertise before transitioning to Lenovo in 2014 following IBM's acquisition. At Lenovo, he continued to drive HPC innovations and now dedicates his work to liquid-cooling solutions. Matthew holds a BA in Molecular, Cellular, and Developmental Biology from the University of Colorado, Boulder.

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:

- AI Servers
- 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, LP2225, was created or updated on May 23, 2025.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2225
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2225.

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® 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.