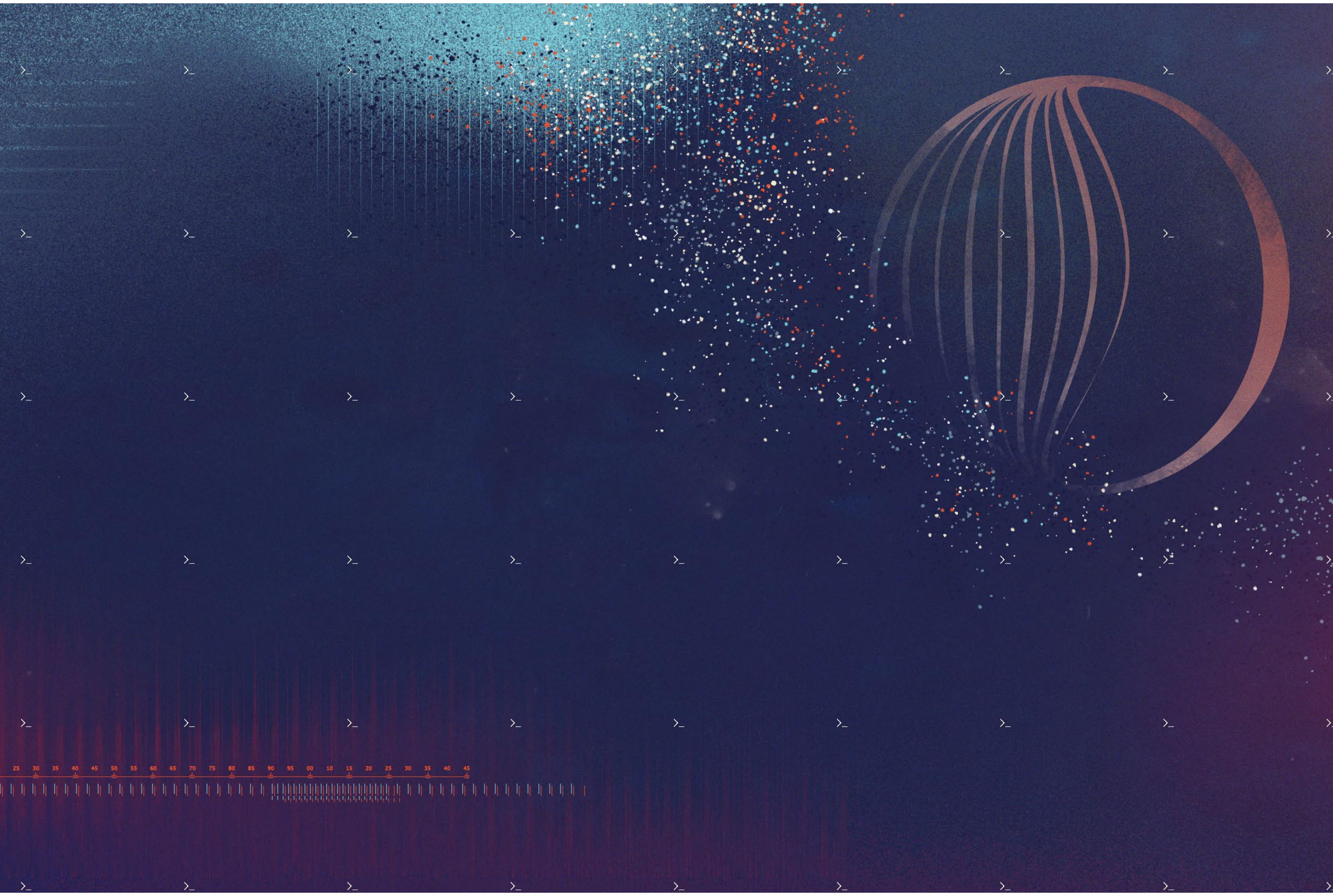


Meeting Enterprise AI Challenges and Building AI Factories – Evaluating Lenovo's Hybrid AI Server Platforms

AUTHOR

Russ Fellows
VP, Labs | Signal65

IN PARTNERSHIP WITH

MAY 2025

Overview

Technologists and business executives continue to learn more about the potential uses for AI every day. AI's ability to help augment company processes, people and productivity is becoming clear. In the latest Futurum Group CIO Insights survey, 89% of CIOs report leveraging AI for strategic improvements, with 71% re-evaluating their cloud workloads.[1]

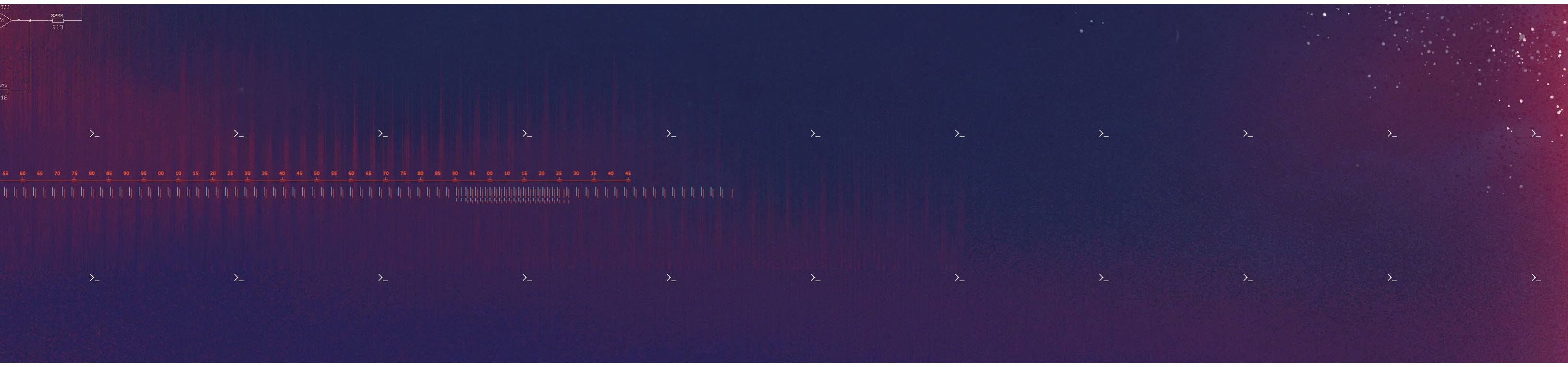
While the awareness and adoption rates of casual AI use among people generally has been astoundingly rapid compared to other new technologies, the integration rate for companies is still low. In market adoption terms, AI is in the early adopter phase, indicating that while companies have started AI projects, many have not yet implemented these solutions in a widespread or structured manner. Stated another way, we expect the use of AI within organizations to dramatically increase over the next 3 to 5 years.

One of the most important considerations for many organizations is not if they will use AI, but how they can do so while protecting their valuable corporate data, maintaining security, and adhering to regulations regarding privacy and sovereignty. This challenge explains why firms may have begun with cloud-based AI, and are beginning to implement AI solutions at scale with hybrid and private AI factories. Hybrid and private AI are options that can help companies maintain control of their private data.

Lenovo asked Signal65 to evaluate their latest hybrid AI platform option in the context of running typical enterprise AI workloads and building AI factories. We utilized the Signal65 AI inferencing test suite, developed in conjunction with Kamiwaza.ai, to run a variety of workloads. Our focus was on scenarios for enterprises who are looking to begin deploying AI tools on-premises in order to boost their productivity, while also maintaining control of corporate data.

Through our testing, we found that the Lenovo ThinkSystem SR680a V3 with NVIDIA H200 GPUs provides an excellent foundation for AI deployments, including support for the largest language models—such as Llama-405B and DeepSeek-R1—while also supporting use for RAG and fine-tuning models to improve accuracy and relevance. In short, this system is a flexible and powerful entry point for private AI, for companies of any size, and across industry verticals.

Signal65 Comments: *We found that the Lenovo ThinkSystem SR680a GPU server using NVIDIA GPUs delivered excellent performance on a wide range of AI use cases, including inferencing, RAG and fine-tuning workloads. The flexibility and performance of this system shows that this system would serve as an excellent starting point for companies looking to build and grow their in-house AI capabilities.*



[1] Futurum Research: Futurm-Research-2025-Key-Issues-20250310.pdf

Analysis Overview & Approach

The objective of this analysis was to answer the questions many people have with respect to a particular hardware configuration. The type of questions we sought to answer included:

- How many users can a particular system support?
- How do different sizes of LLMs perform?
- Are we able to run fine-tuning and RAG workloads on the system?
- Can we run the very largest LLM's, like Llama-405B and DeepSeek-R1?

In order to answer these questions, we ran a wide range of AI workloads, utilizing multiple GPU configurations, a variety of LLM models, and tweaking their parameters. Additionally, we tested a fine-tuning workload, using a base model, that was enhanced by additional training with private data. For this we utilized a publicly available legal training data set, and the Axolotl fine-tuning tool.

Signal65 Comments: *Generative AI has quickly become adept at assisting in the creation of content based upon specific content, which can then be leveraged by companies to help augment their workforce across a variety of disciplines. This includes content generation for marketing or programming code, along with helping to provide content for employees including internal or external customer service, finance, HR and other functional roles.*

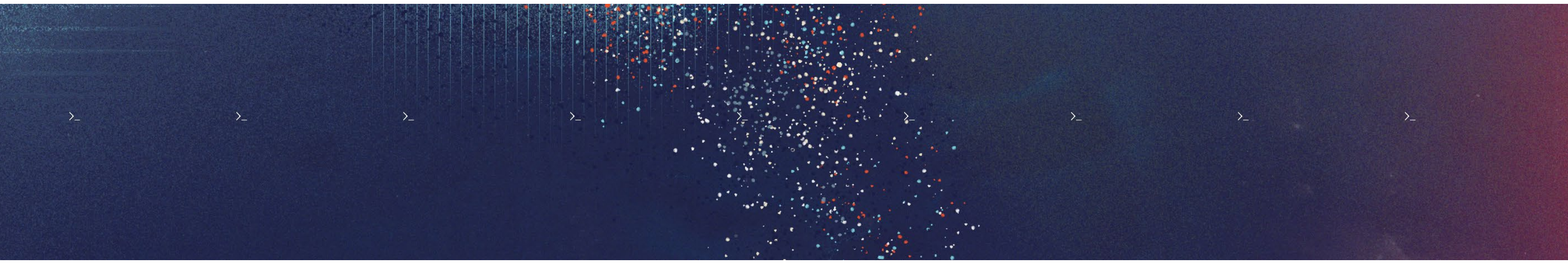
We also ran a simulated Retrieval Augmented Generation (RAG) workload, by reproducing large input sizes as would occur with RAG. The focus of our testing was on the inferencing performance of various LLM workloads, including with RAG. Typically, RAG relies upon specialized databases to retain and retrieve vectors, which can add significant variability to the process.

The RAG process works to augment an LLM with additional information contained in vector databases. This information is queried, and relevant context is then added to the original prompt. In our usage, we did not use a vector database, but instead dramatically increased the input prompt sizes up to 32k tokens. This process helped remove the variability of different vector databases from our testing and allowed us to focus on the impact RAG has on the inferencing portion of the workload. As a result, we were able to measure how particular LLMs and the Lenovo SR680a GPU server performed the inferencing portion of the workload, after retrieving data from a vector database.

The range of parameters tested also included various input and output token counts while using different batch sizes. Across the variety of tests, we generated thousands of data points. In this paper we are highlighting some of the most important, including the following:

1. Testing different LLM sizes to showcase potential throughput and performance:
    - Small models (2–8B parameters) for high total throughput
    - Larger models (~70B parameters) for higher-quality results
    - Very large models (up to 405B) to confirm the system's ability to handle massive inferencing workloads with the highest quality LLMs

2. Testing system scalability by using different GPU counts:
    - 1, 4, and 8 GPUs were used to examine how performance scales with GPU resources on the Lenovo ThinkSystem
3. Common use case scenarios:
    - Single-node workloads (8 GPUs)
    - Inferencing, including RAG
    - Fine-tuning tasks, such as training an LLM on domain-specific data
4. Highlighting the unique capabilities of NVIDIA H200 GPUs:
    - Significantly higher memory (141 GB per GPU vs. 80 GB typical on H100)
    - Running 70B models at FP16 on fewer GPUs (e.g., 2 GPUs for 70B on H200, whereas 4 are usually required on H100)
    - Enhanced performance with lower-precision inference (FP8, INT8, etc.) to improve throughput
5. Examining performance trade-offs:
    - Small vs. larger batch sizes for lower latency vs. higher throughput
    - Impact of tensor parallelism (scaling from 1 to 8 GPUs)
    - The performance of small (8B), large (70B) and very large (405B) LLMs



Lenovo AI Infrastructure Lineup

Lenovo provides a range of server, storage, and networking solutions optimized for AI workloads through its ThinkSystem product line. These servers include both air-cooled and water-cooled options suitable for various data center environments.

Lenovo offers over 80+ AI-ready systems in its portfolio. To highlight a few:

- **Lenovo ThinkSystem SR680a V3 and SR780a V3:** Air- and water-cooled servers in a 5U chassis, supporting up to eight GPUs. For the water-cooled ST780a, cooling is facilitated by braided stainless-steel lines with copper connectors.
- **Lenovo ThinkSystem SR685a V3:** An 8U air-cooled server compatible with NVIDIA and AMD GPUs.
- **Lenovo ThinkSystem SR675a V3:** A 3U rack-mounted server supporting up to 4 GPUs. It can also be configured with and supports up to eight PCIe H200 GPUs. It features Lenovo's Neptune™ hybrid cooling module, which dissipates heat using direct liquid cooling (DLC) to a liquid-to-air heat exchanger, minimizing the need for extensive plumbing infrastructure.
- **Lenovo ThinkSystem SR650a V4:** A 2U-rack mounted server supporting up to eight GPUs. It also features Lenovo's Neptune hybrid cooling module.

These hybrid AI platforms scale efficiently from single-server setups with four GPUs up to large-scale, rack-level deployments. The modular architecture, exemplified by Lenovo's flexible "3-2-1" design in SR680a and SR685a servers, allows customizable combinations of CPUs and GPUs to address diverse AI tasks such as training and inference.

Additionally, Lenovo's PCIe-based AI nodes offer configurable GPU risers and support NVLink technology to enhance GPU-to-GPU communication. Cooling demands are managed through Lenovo's Neptune liquid cooling, technology initially developed for major supercomputing projects.

Lenovo complements its hardware solutions with integrated networking, storage, software, and professional services, enabling organizations to efficiently deploy and scale AI infrastructure according to their specific needs.

The Lenovo ThinkSystem SR680a V3 with NVIDIA H200

Signal65 performed all our testing and benchmarking using the Lenovo ThinkSystem Sr680a V3, with eight NVIDIA H200 GPUs. Some of the key features and capabilities of the system include:

- 8x GPUs (NVIDIA H100 640 GB HBM, H200 with 1,128 GB HBM, or AMD MI300X with 1,536 GB HBM)
- Support for NVIDIA NVLink or AMD Infinity Fabric up to 900 GB/s
- Two 5th Gen Intel® Xeon® Scalable processors with 32 DDR5 DIMM sockets
- Up to 10x PCIe Gen5 x16 slots with 8x directly connected to the GPU complex
- Up to 16x 2.5" NVMe drives for maximum storage capacity and performance
- Designed with 8x power supplies for full N+N redundancy without throttling
- ThinkSystem SR680a includes N+1 hot-swap fans, with headroom for future GPUs

Figure 1: *Lenovo ThinkSystem SR680a V3 GPU Server*

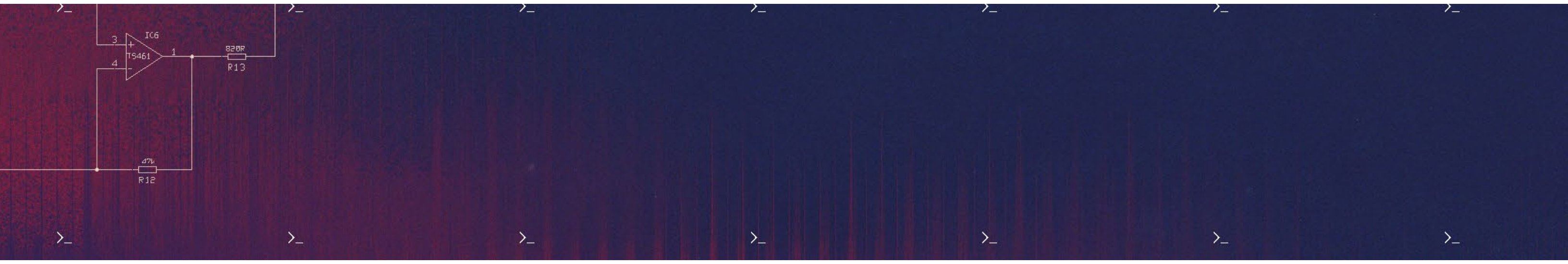
Lenovo has published a summary of recent MLPerf benchmark results, highlighting that the SR680a V3 with NVIDIA H200 GPUs placed in 1st, 2nd or 3rd across 16 different categories. These results include LLM training, image generation and Medical Image training and others.[2]

[2] Lenovo Press Release: https://lenovopress.lenovo.com/lp2036-breaking-barriers-in-ai-inference-mlperf-41

Signal65 Comments: *Lenovo's participation with the MLCommons organization is an important part of helping to create and promote standardized AI performance results. This provides clarity and much-needed information to help IT and business executives make decisions.*

Here we note some of Lenovo's recent notable results in various MLCommons benchmarks, highlighting the capabilities of their AI-optimized server solutions:

1. **Elevating AI Performance:** (Published August 20, 2024): Lenovo's MLPerf Training 4.0 benchmarks, performance of the Lenovo ThinkSystem SR685a V3 equipped with 8x NVIDIA H100 GPUs. Lenovo Press
2. **Unleashing the Power of AI:** (Published June 6, 2024): Highlights the performance of Lenovo's ThinkEdge SE455 V3 and ThinkEdge SE450 servers in MLPerf benchmarks. Lenovo Press
3. **Breaking Barriers in AI Inference:** (Published October 11, 2024): Showcases the Lenovo ThinkSystem servers in the MLPerf v4.1 benchmarks, particularly the SR680a V3 and SR685a V3 models equipped with NVIDIA GPUs. Lenovo Press



AI Testing Background

In general, larger models produce higher quality results than smaller models but require more resources. Large models have more parameters, with sizes ranging from approximately 3 billion parameters up to 400 billion and beyond.

In addition to the size of the model, the size of the data used when processing requests matter, with larger data sizes like 16-bit data types (FP16 and BF16) typically producing better quality results than smaller data sizes, such as 8-bit data types like FP8 or INT8. More recently, inferencing software and models are being created that take the quantization (reduced data sizes) to new levels, with new support for sizes as small as 2 bits.

There are other trade-offs that can be made, including how many requests are submitted at the same time, known as the batch size. Reserving an entire GPU for a single user provides excellent performance with few delays but can be an inefficient use of a valuable resource. In contrast, assigning many requests together in one batch increases total token throughput, but does so at the cost of latency delays for individual users waiting for token responses.

There is no consensus on what the best set of trade-offs is for any one problem. Thus, Signal65 chose to test a range of possibilities, to demonstrate the range of workloads, and the performance levels that should be expected when using a Lenovo ThinkSystem SR680a V3 with eight NVIDIA H200 GPUs for common enterprise AI workloads.

There are two primary metrics for examining the speed or performance of AI inferencing. One is to evaluate how many tokens are processed over some time period, with the focus on total throughput rather than how long it takes to obtain individual tokens. This measurement is appropriate for offline or "batch processing" where multiple requests may be submitted, and the results are examined later. Using large batches will maximize GPU utilization, although at the expense of time to first token and the time between tokens. Latency isn't critical since the process is running asynchronously.

The other common use case for LLM inferencing is for live, interactive usage where low latency is an important factor in the perceived response. In general, less time for a response is better. There are limits in terms of how fast is useful, and for text responses producing results faster than people can read does not improve the experience. Typically, responses that take less than 2 or 3 seconds to begin, and rates close to 200 words per minute for textual responses are perceived to be very fast. However, for code generation the response rate can go even higher, as the intention is to run or compile the code, often without reading the response.

Scaling Token Rates with Batch Sizes

As discussed previously, there are many choices when inferencing, including the model size (such as 8B, 70B or 405B) parameters, along with the batch size used when submitting queries.

In Figure 1 below, we examine how increasing batch sizes can yield a greater number of total tokens per second for three different sized LLMs. In this example we utilized all eight GPUs, with Tensor Parallelism (TP) set to 8, and a constant input / output size.

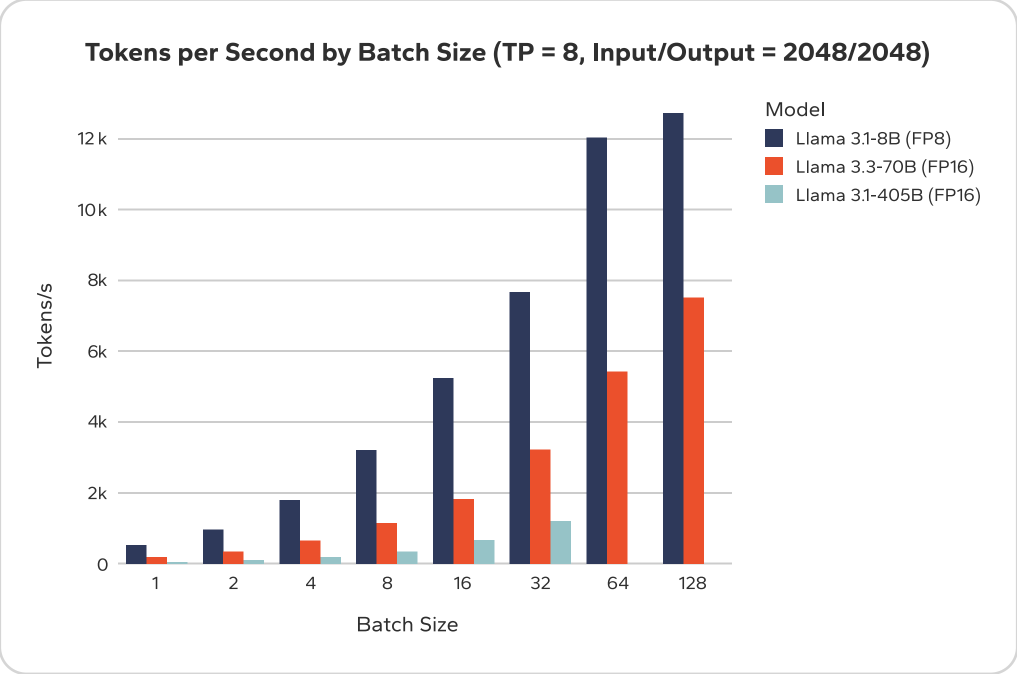


Figure 2: *Tokens per Second by Batch-Size across 3 LLMs*

Next, in Figure 3, we explore how using a different number of GPUs can scale performance, by holding our batch size and our input / output size constant. Again, we examine the impact across the same three different models. Note that Llama-70B required a minimum of 2 GPUs to run, and Llama-405B requires all eight GPUs to run. While results do scale, they are not linear. This is expected behavior and occurs in part due to the complexity of scheduling workloads across multiple resources.

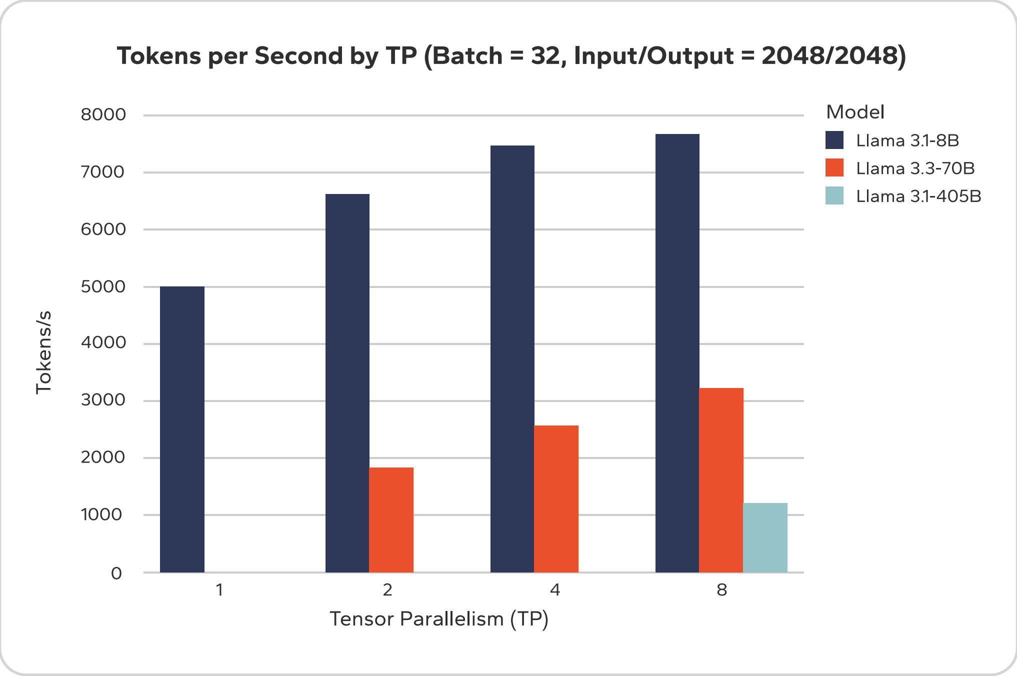


Figure 3: *TPS by Tensor Parallelism, with Constant Batch and Size*

In our next chart, we examine the impact of scaling GPUs, with TP of 1, 2, 4 and 8, while highlighting different batch sizes with a single LLM, Llama-3.1-8B. Again, we see good scalability, with results increasing, albeit at a less than linear rate. These results are also expected and in-line with other test results for these GPUs and model parameters, shown below in Figure 4.

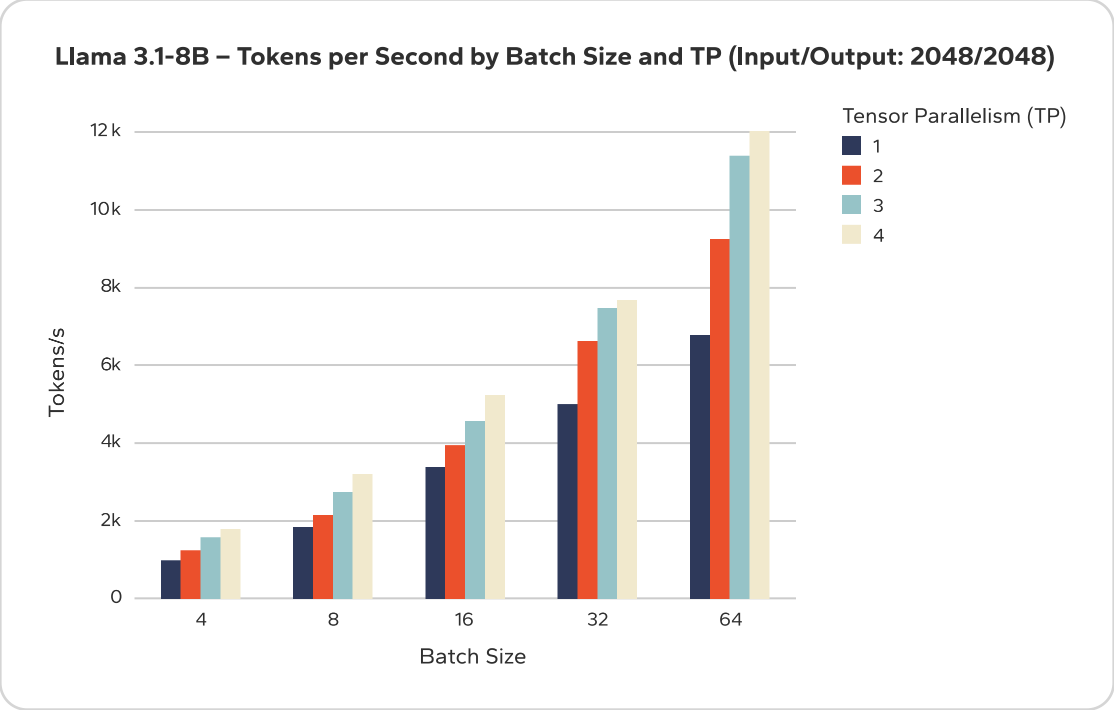


Figure 4: *Llama-8B TPS by Batch Size, with Constant Size*

Simulated RAG Inferencing

In the last of our inferencing comparisons, we show how RAG results can scale, by again examining the total tokens per second that can be processed, for three different model sizes. In these scenarios we significantly increased the number of input tokens above 2k, with input values of 8,000, 16,000 and 32,000 which represent significantly long input context sizes, that are similar to workloads where RAG would query a Vector DB, find relevant information, and then add that content to the input provided to the LLM to inference.

Although the rate increases, it is less than linear, indicating that the total time required to process the request and return results increases as the input grows. This is expected behavior and aligns with results achieved by other test cases of RAG inferencing shown in Figure 5.

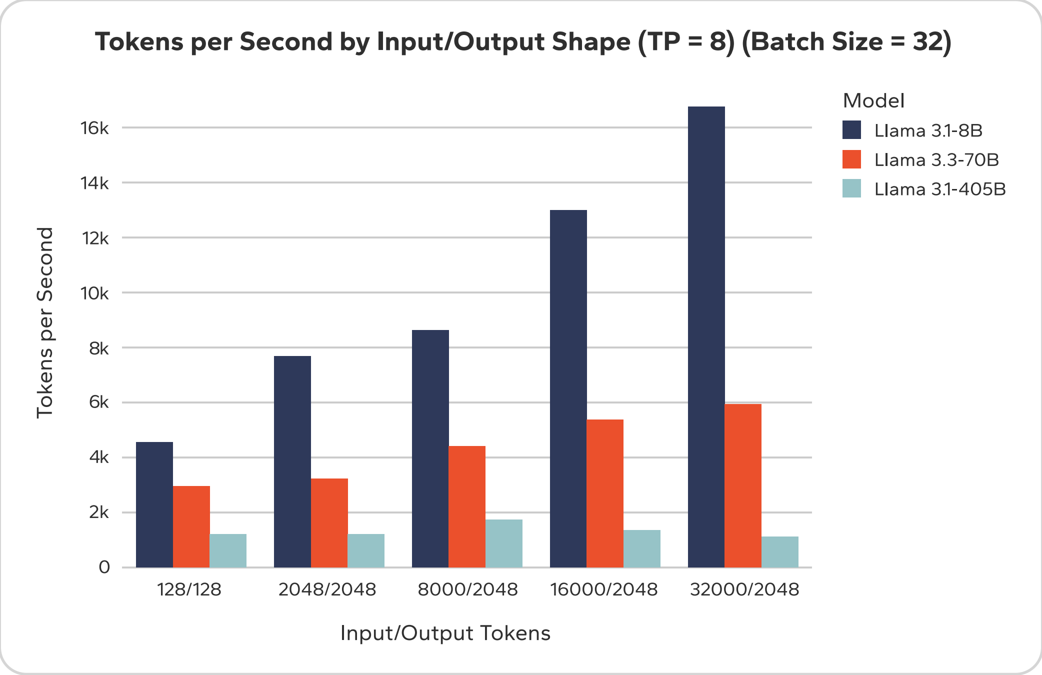


***Figure 5:** TPS by input / output shape, Batch 32, TP=8*

Fine Tuning Workload

Fine-tuning can be an important part of the process of building efficient and effective AI tools that are tailored for particular use cases. By adding additional examples based upon corporate data, companies can create customized AI models that can provide more accurate results, while also utilizing smaller base models as their foundation. While larger models have proven to provide more accurate responses, this is in large part because they contain more information, that may apply to a particular data set or use case. Fine-tuning is different, in that instead of hoping that a large model will contain some relevant data, a company can instead add additional training with private corporate data to instill relevant information into a model.

In many cases, companies have reported obtaining excellent results by fine-tuning models as small as 3 billion parameters. In our testing, we used a well-known fine-tuning framework known as Axolotl to train a small, 7B model and a large 72B parameter model with a legal dataset.

Note: *While our fine-tuning example used a legal dataset, this can be applied to nearly any industry or vertical segment. Adding in specific examples from corporate data is what moves an LLM from providing interesting suggestions, to actionable insights based on specific, relevant information. This is why combining LLM's, with corporate data is the objective companies are attempting to implement. Our example showcases doing so in several hours.*

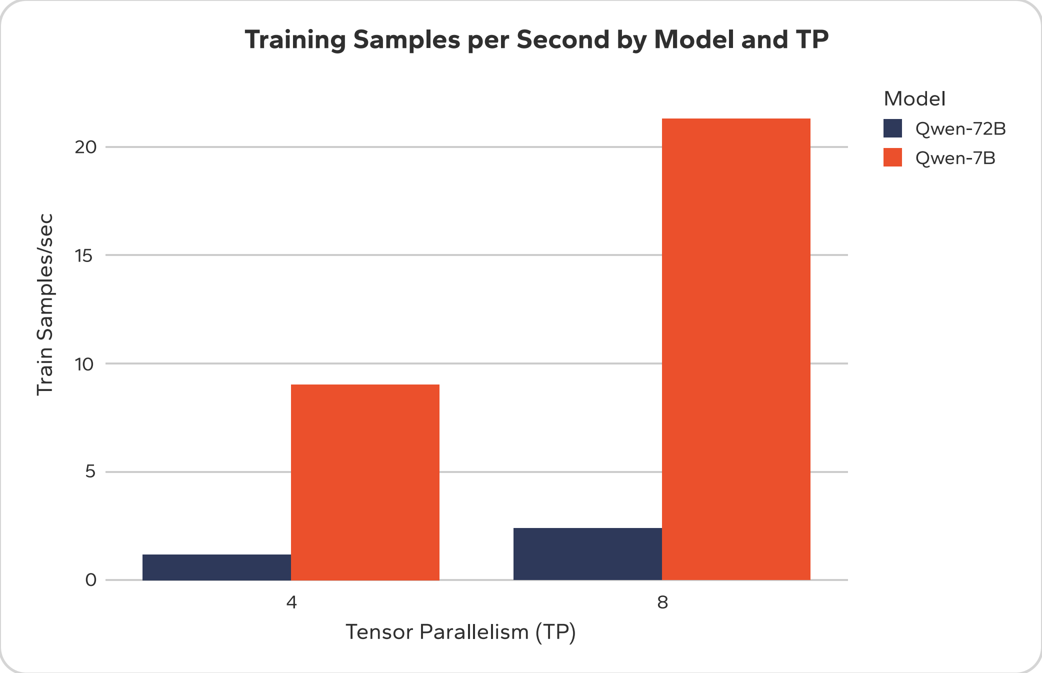


Figure 6: *Training rate, by Model Size, and Number of GPUs (TP)*

The graph shown in Figure 6 shows the rate of training, in terms of samples per second for these two models when using four GPUs, and eight GPUs. As seen, the results scale extremely well with increased Tensor Parallelism.

Signal65 Comments: *The ability for companies to create and continually update LLM models with corporate data in under a day is a powerful capability that can provide a significant boost to the usefulness of AI for nearly any company in any vertical segment. A retail firm may update their model with the latest sales data on a weekly data, which in turn can then help further improve their sales, order tracking and inventory processes. Also, it is important to note that fine-tuning is different from RAG, and using both together can have even greater benefits.*

Additional Observations

In addition to testing inferencing with the three Llama models, we also examined running DeepSeek-R1, examined inferencing for two Qwen models, 7B and 72B models, both of which demonstrated excellent performance scaling with higher batch sizes and tensor parallelism (TP).

In Table 1, we provide more data points for the fine-tuning process, which include the learning rates, the number of training cycles, or epochs, along with the total training time. The time to train, or more accurately fine-tune a particular model with a data set is an important factor. With the power and capabilities of the Lenovo system and the H200 GPUs, the entire process can be completed in several hours, even for the larger Qwen-72B model.

Model Name	Tensor Parallelism	Training Samples / Sec.	Training Iterations	Train Time (H:MM:SS)
Qwen-72B	8	2.394	2	9:20:33
Qwen-7B	8	21.298	4	2:06:01
Qwen-72B	4	1.186	2	18:51:11
Qwen-7B	4	9.027	4	4:57:19

Table 1: *Fine-Tuning process on 4 or 8 GPUs for 7B and 70B models*

Signal65 Comments: *We found that the Lenovo ThinkSystem SR680a GPU server using NVIDIA GPUs delivered excellent performance on a wide range of AI use cases, including inferencing, RAG and fine-tuning workloads. The flexibility and performance of this system shows that this system would serve as an excellent starting point for companies looking to build and grow their in-house AI capabilities.*

These findings underscore how well AI workloads scale when the hardware and software stacks are designed for parallelization. Large-batch, long-sequence inference—such as summarizing lengthy documents—benefits significantly from higher GPU counts, while smaller-batch or single-request scenarios still gain from parallelization (just less dramatically).

Performance Testing Summary

As discussed, there are many ways to configure and use large language models, with choices leading to trade-offs between, speed, result quality, and responsiveness. As shown in this testing, companies implementing the SR680a system with eight NVIDIA H200 GPUs have a lot of possible options. With prior generations of GPUs, it was not possible to run DeepSeek-R1 or Llama-405B on a single, 8x GPU system, due to the large GPU memory requirements.

The SR680a system serves as a scalable building block for enterprises that desire high token processing rates combined with large GPU memory to enable running the largest LLMs. By clustering Lenovo GPU servers together with Lenovo networking and storage infrastructure, companies can scale their AI infrastructure to meet their needs, regardless of the scale required.

For companies who may not require eight GPUs to run their inferencing workloads, other servers such as the Lenovo ThinkSystem SR675 V3 system, with four GPUs in a 3U enclosure can provide a more flexible and cost-effective entry point into delivering AI inferencing services to their internal and external clients.

Based upon Signal65's testing, we found that the Lenovo ThinkSystem SR680 V3 can provide inferencing performance rates for various LLM sizes as follows:

Very Large Inferencing (e.g., Llama 3.1–405B)

- **Performance:** Rates up to 1,000 tokens per second, but a focus on quality over speed.
- **Findings:** On the SR680a V3 can support 2–8 simultaneous users, with performance adequate for interactive sessions. For offline scenarios, the total tokens per second achieved was approximately 1,700, with a batch size of 32 and large input / output sizes.
- **Use Cases:** Deep research, or in-depth queries that require extensive knowledge and high-quality results as is typical during the design stages of product development.

Large Model Inferencing (e.g., Llama 3.3–70B)

- **Performance:** Roughly 6× faster than 405B, scaling nearly linearly with parameter count. At larger input sizes, rates were approximately 6,000 tokens per second.
- **Findings:** Up to 32 users in interactive sessions was achievable with under two minutes of total turnaround time for large prompts. For offline use, this model achieved more than 8,000 tokens per second with large batch sizes and large input and output combinations.
- **Use Cases:** Real-time Q&A, chatbots, or summarization tasks for departmental or company-wide usage—while maintaining on-premises data security. When combined with RAG, or fine-tuning, can provide extremely high-quality and accurate results with good performance.

Small Model Inferencing (e.g., Llama 3.1–8B)

- **Performance:** Up to 8× faster than a 70B model, supporting a high volume of interactive sessions. We saw rates exceeding 16k tokens per second on 8 GPU configurations.
- **Findings:** These models can run effectively on fewer GPUs, or a single GPU, thus freeing others for separate tasks or parallel instances. For offline use, this model achieved more than 28,000 tps with large batch sizes and large input / output sizes.
- **Use Cases:** Ideal for large-scale internal usage (e.g., thousands of employees) needing moderate LLM capabilities. Enhancing with RAG is desirable. Also appropriate for fine-tuned small models that add domain specific expertise to a smaller model.

Fine Tuning (Using Qwen 7B and 70B)

- **Performance:** Good training rates, with linear performance scaling, 2X higher for both sizes moving from 4 GPU to 8 GPUs.
- **Findings:** Linear scaling, Qwen 7B training time fell from 4h:57m, to 2h:06m by scaling from 4 to 8 GPUs, with Qwen 70B going from 18h:51m to 09h:20m
- **Use Cases:** Training time of less than 1 day for a large language model of 70B parameters is considered very good. This enables companies to train models monthly or even weekly with new data sets for enhanced operational accuracy on their private data.

Mapping Results to Real-World Scenarios

Lenovo's objective with their AI portfolio is to help their clients deliver business value through horizontal capabilities along with targeted vertical solutions. Currently, the three areas of horizontal focus for AI workloads are:

- **Create:** Content Creation, including audio, text, video and computer code
  - Using the largest LLMs, or diffusion-based AI models require multiple GPUs with significant GPU memory in order to perform well for multiple interactive users. As seen with the inferencing rates of 1,000 tokens per second with Llama 405B, this would support interactive coding with the very largest LLMs to enhance developer's speed. Additionally, diffusion models for audio, images and video require substantial GPU processing.
- **Engage:** Customer service engagement support including chatbots, website content, language translation and customer service agents
  - This objective can be met using a Large model such as Llama 70B, which when run on the system tested delivered approximately 6,000 tokens per second. These token rates are high enough to support more than 100 interactive users with excellent performance.
- **Assist:** Knowledge Assistants for Legal, HR, Finance and other workplace assistants
  - These types of workloads are often best accomplished through a combination of fine-tuning smaller models such as Qwen or Llama 7B and then utilizing these models in batch processing mode with corporate data sets. With processing rates of 16,000 tokens per second, this translates to processing a 15 page document per second.

Each of these horizontal enablers can be significantly enhanced by utilizing a Lenovo SR680a V3 system with NVIDIA GPUs as tested.

Summary

The Lenovo ThinkSystem SR680a V3 system equipped with NVIDIA H200 GPUs presents a robust and scalable hybrid AI platform well-suited for enterprise and private AI deployments. Our evaluation showed the system is capable of strong performance across a range of AI workloads, from small-scale inferencing, RAG inferencing, and even fine-tuning large, 70B parameter models.

The system showed the ability to support a wide number of model sizes, handling everything from smaller 2–8 billion models to very large-scale 405 billion parameter models. The expanded GPU memory available in the H200 GPUs proved especially beneficial, enabling larger models to run efficiently using fewer GPUs than typically required by H100-based configurations. This not only improves performance but also optimizes hardware utilization and reduces operational complexity.

According to Lenovo, the use of their Neptune liquid cooling solution can reduce power consumption by up to 40% compared to similar air-cooled systems in data centers. Additionally, liquid cooling can reduce equipment rack usage, by moving some portion of the cooling external to the rack or datacenter. As a result, customers can achieve a reduction in power and space, while possibly increasing equipment utilization and longevity through enhanced thermal stability.

The system also demonstrated a strong balance between batch and interactive workloads. In batch scenarios, throughput remained consistently high, while in interactive inferencing tasks, latency remained low. Notably, time-to-first-token metrics remained within a few seconds across a variety of configurations—even when working with relatively large prompts. This performance makes the system suitable for both real-time user interactions and high-volume background processing.

For specialized AI use cases such as fine-tuning and RAG, the SR680a V3 also showed strong capabilities. It showed very good scalability of different model sizes for fine-tuning workloads small and large language models. For RAG enhanced inferencing tasks, the system processed large prompt sizes effectively, supporting sequences of 30,000 tokens or more. This demonstrates the platform's ability to support advanced AI techniques that rely on extensive context or domain-specific adaptation.

Lenovo Hybrid AI Advantage™

Lenovo Hybrid AI Advantage™ help organizations improve productivity, increase agility and innovate with trust through standardized and accelerated development and deployment of AI use case solutions. Lenovo Hybrid AI Advantage™ bring the power of Lenovo AI library and validated, tested hybrid AI factories (hybrid AI platforms, workstations, servers, storage, network, software, models, services, partner ecosystem) to the enterprises.

The hybrid AI factory is designed to support hybrid deployments at the Edge, data centers, colos and business locations with cloud integration. It offers flexibility of model, infrastructure choice, enables a wide range of AI applications, agentic and machine learning workflows, and real-time data analysis. Lenovo's hybrid AI platforms power the hybrid AI factory, and they can scale from a single server with just four GPUs as starter environment to a rack scalable unit (SU) as a turnkey infrastructure solution with partner technology choice.

Signal65 Comment: *In summary, the Lenovo ThinkSystem SR680a V3 system with NVIDIA H200 GPUs provides a versatile and powerful solution for organizations pursuing on-premises AI. It is well-suited for real-time inferencing, high batch processing throughput, along with AI uses cases requiring RAG and fine-tuning. For teams seeking a secure, high-performance, and future-ready AI infrastructure, the SR680a V3 offers both a strong starting point and a reliable starting point to begin building a scalable AI platform.*

All of these results are important within the context of supporting enterprises efforts towards standing up internal AI operations, to provide the benefits of AI while ensuring data security and privacy. For smaller companies, a single SR680a system with eight GPUs was shown to support hundreds of simultaneous users. As firms grow, or for larger enterprises, the Lenovo SR680a systems may be clustered to provide even more capabilities.

Important Information About this Report

CONTRIBUTORS

Russ Fellows
VP, Labs | Signal65

PUBLISHER

Ryan Shrout
President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

This document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

IN PARTNERSHIP WITH

Lenovo

ABOUT SIGNAL65

Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

signal65

signal65

CONTACT INFORMATION

Signal65 I signal65.com