signal65

EXECUTIVE SUMMARY

Evaluating Lenovo's ThinkSystem SR680a V3 for Enterprise AI

AUTHOR

Russ Fellows
VP, Labs | Signal65

IN PARTNERSHIP WITH

Lenovo

JUNE 2025

Introduction: The Emerging Demand for Enterprise AI

Enterprise leaders are rapidly shifting from exploring artificial intelligence (AI) to implementing real solutions. According to The Futurum Group CIO Insights survey[1], 89% of CIOs are leveraging AI for strategic advantage, and 71% are reassessing cloud workloads. However, most companies remain in the early adoption phase, constrained by concerns over data security, privacy, and regulatory compliance.

To address these challenges, organizations are turning to hybrid or private AI deployments that provide greater control and performance. The Lenovo ThinkSystem SR680a V3 with NVIDIA H200 GPUs represents a compelling solution for deploying AI workloads securely on-premises.

Platform Overview: Lenovo SR680a V3 with NVIDIA H200 GPUs

The ThinkSystem SR680a V3 is a 5U, water-cooled GPU server supporting up to 8x NVIDIA H200 GPUs with 141GB HBM per GPU. Key platform features include:

- Dual 5th Gen Intel Xeon processors
- Full N+N power redundancy
- NVLink and Infinity Fabric support
- The modular '3-2-1' design allows future scalability, while Neptune™ liquid cooling enhances energy efficiency and system reliability.

Test Methodology and Performance Overview

Signal65 evaluated the platform using the Signal65 AI test suite, co-developed with Kamiwaza.ai. We tested inferencing, Retrieval Augmented Generation (RAG), and fine-tuning using various LLM sizes:

- **Small models:** 2–8B parameters
- **Large models:** ~70B parameters
- **Very large models:** Up to 405B parameters

The evaluation focused on balancing interactive low-latency scenarios with high-throughput batch processing workloads.

[1] Futurum Group Enterprise AI and CIOs 2025

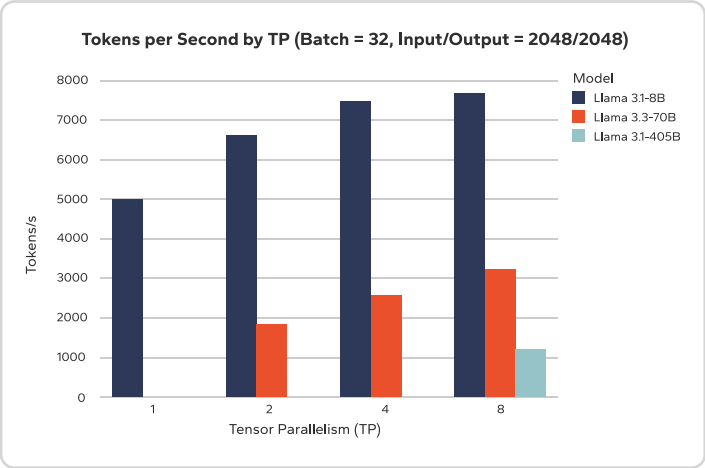

Figure 1: *Scaling Tokens per Second across GPUs with Various LLMs (Source: Signal65)*

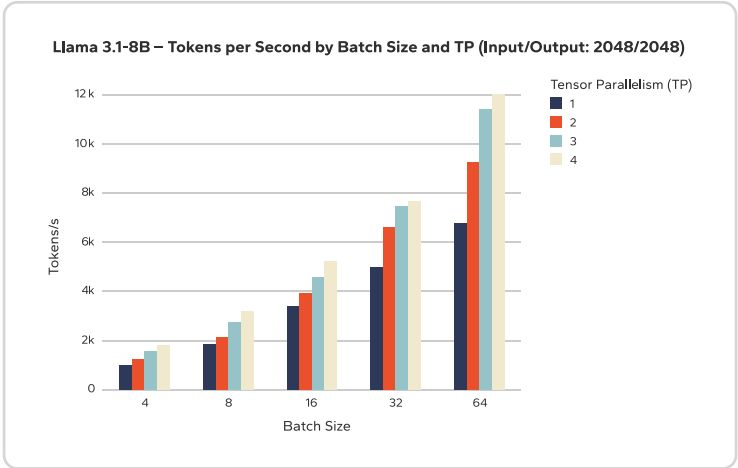

Figure 2: *Scaling Tokens per Second by Batch-Size across 3 LLMs (Source: Signal65)*

The Lenovo ThinkSystem SR680a V3 was able to scale performance using a variety of model sizes, including LLama 8B, 70B and 405B as shown in Figures 1 and 2. Scaling occurred via both using additional GPUs (aka tensor parallelism) in Figure 1, and through increased batch sizes seen in Figure 2. The system also delivered scalable fine-tuning capabilities for two different model sizes as displayed in Figure 3.

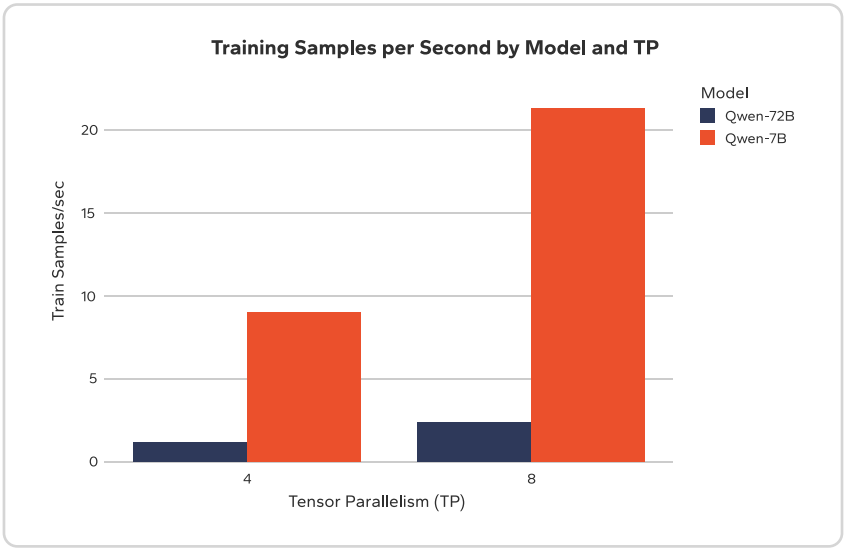

Figure 3: *Fine-tuning Performance using 4 or 8 GPUs (Source: Signal65)*

Summary & Real-World Scenarios

Lenovo's objective 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
- **Engage:** Customer service engagement support including chatbots, website content, language translation and customer service agents
- **Assist:** Knowledge Assistants for Legal, HR, Finance and other workplace assistants

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

The Lenovo ThinkSystem SR680a V3 is a future-ready, enterprise-grade AI platform. It supports a wide range of AI workloads including large language model inferencing, fine-tuning, and Retrieval Augmented Generation. Its hardware design—featuring high-memory GPUs and modular scalability—enables companies to retain data control while benefiting from rapid AI adoption.

The system tested demonstrated a strong performance for multiple workloads, making the system suitable for both real-time interactions and high-volume background processing. Additionally, AI use cases such as fine-tuning and RAG, the SR680a V3 also showed strong capabilities by scalability different model sizes for fine-tuning workloads. 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.

Signal65 Comments: *In summary, the Lenovo ThinkSystem SR680a V3 system with NVIDIA H200 GPUs provides a versatile and powerful solution for organizations pursuing on-premises AI deployments. 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.*

Whether deployed as a standalone solution or as part of a clustered infrastructure, the SR680a V3 provides an ideal foundation for secure, high-performance enterprise AI.

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

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