signal65

# The Lenovo ThinkSystem SR680a V3 with NVIDIA H200 GPUs

Organizations are turning to hybrid or private AI deployments that provide greater control and performance. Signal65 went hands-on with the Lenovo ThinkSystem SR680a V3 to prove its value proposition for a wide range of AI use cases and model sizes.

> **The Lenovo ThinkSystem SR680a V3 with NVIDIA H200 GPUs represents a compelling solution for deploying AI workloads securely on-premises.**
>
> signal65

## Tech Specs

- 5U GPU server with Neptune liquid cooling
- Dual 5th Gen Intel Xeon Scalable processors
- Up to 8x NVIDIA H200 GPUs with 141GB HMB per GPU
- NVLink and Infinity Fabric support
- Modular 3-2-1 design for future scalability

## Performance Overview

The Lenovo ThinkSystem SR680a V3 was able to scale enterprise AI workload performance with a variety of model sizes, using additional GPUs in **Figure 1**, and through increased batch sizes seen in **Figure 2**.

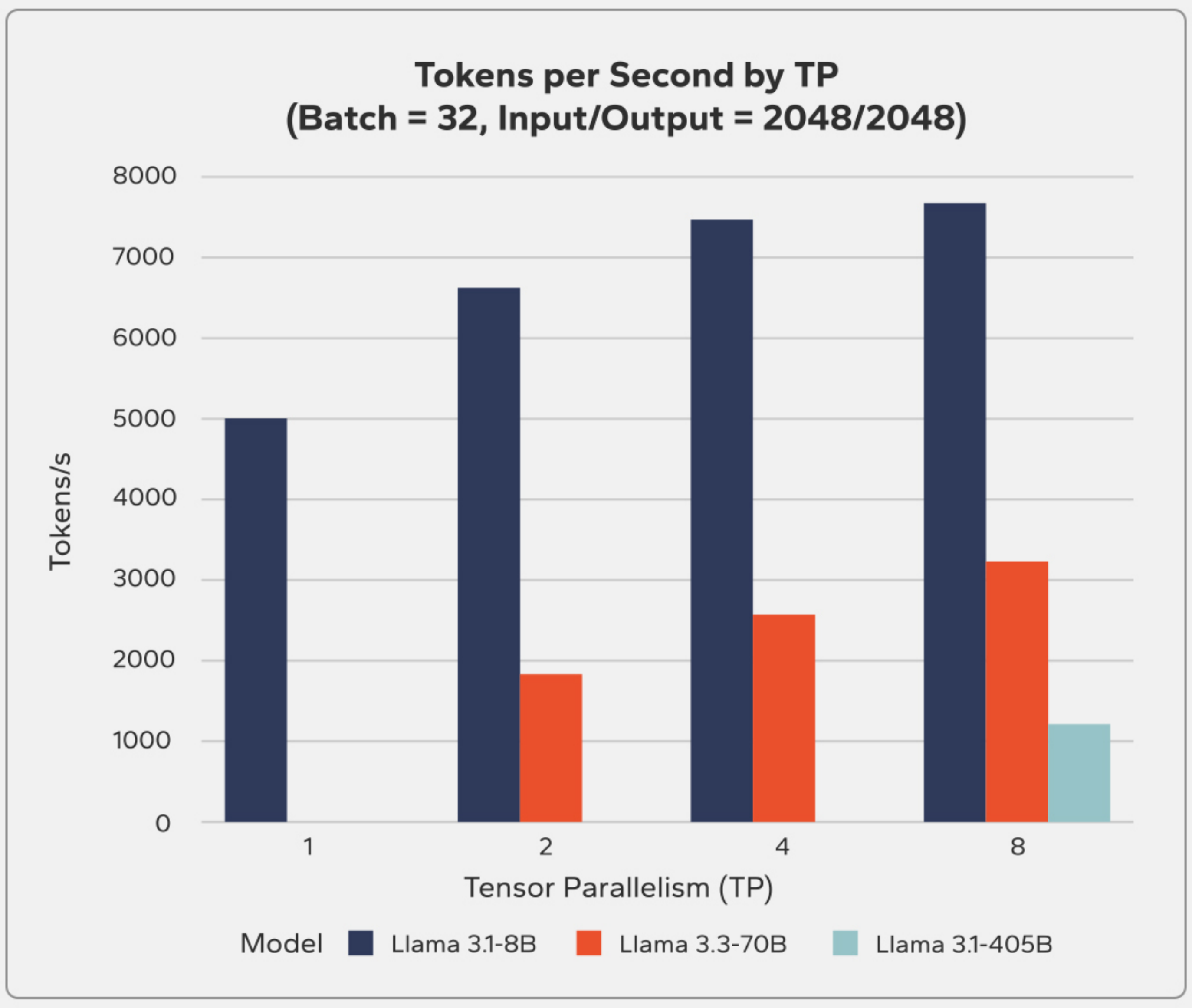

*Figure 1: Scaling Tokens per Second across GPUs with Various LLMs (Source: Signal65)*

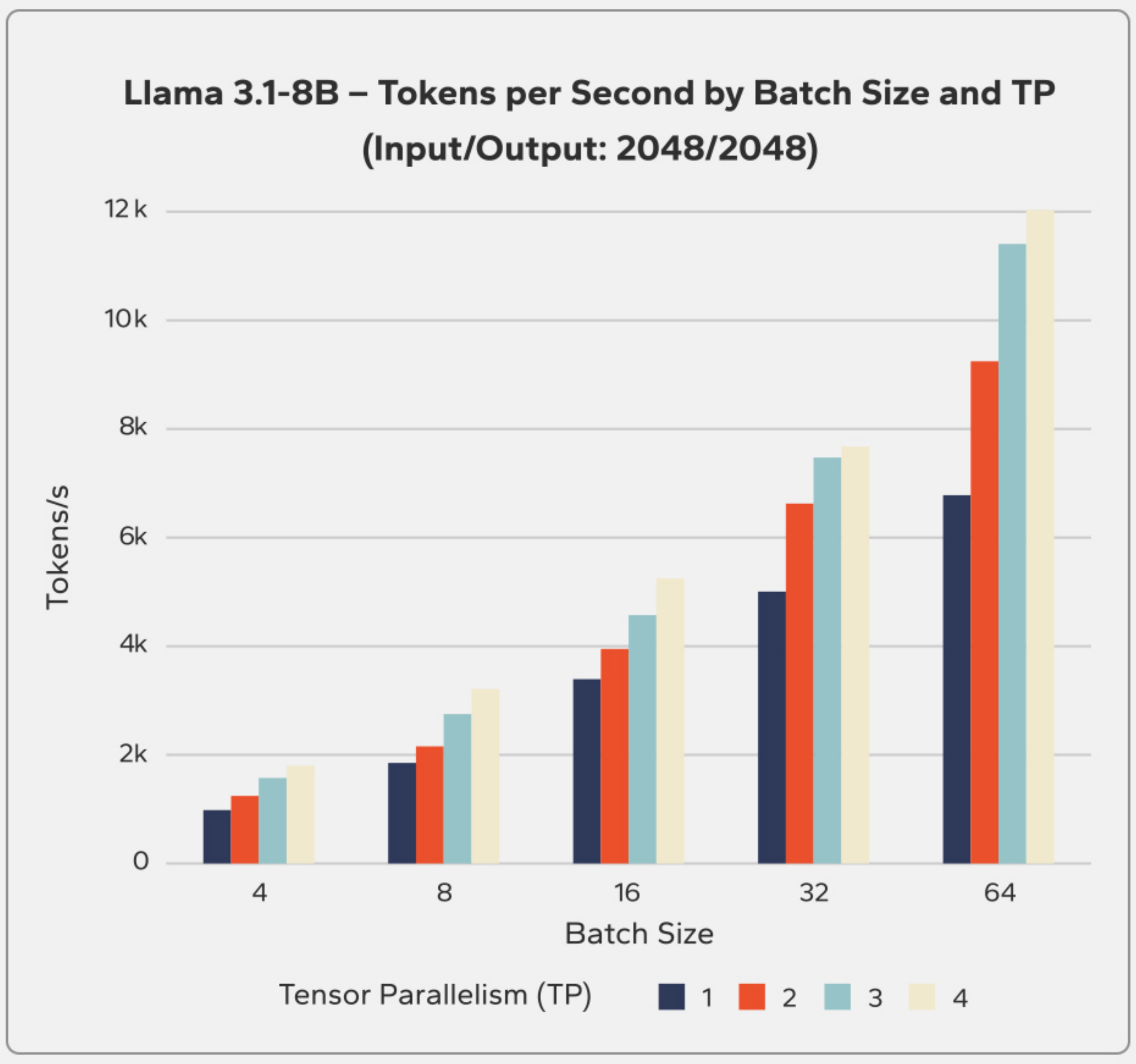

*Figure 2: Scaling Tokens per Second by Batch-Size across 3 LLMs (Source: Signal65)*

## Based on hands-on testing by Signal65, the Lenovo ThinkSystem SR680a V3 is well-equipped to handle these key AI workload areas:

### 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.

**Visit signal65.com for a more detailed breakdown and full report.**

**In partnership with:** Lenovo