PRODUCT BRIEF

SUPERMICRO X14 4-SOCKET SYSTEMS POWERED BY INTEL® XEON® 6 CPUS DEMONSTRATE 1.9X CONCURRENCY COMPARED TO 2-SOCKET SYSTEMS

Enterprise LLM Scaling and vLLM Performance on 4-Socket Systems Enables Higher Throughput, Greater User Concurrency, and Larger Models Beyond 2-Socket Architectures

Supermicro 4-Socket Server SYS-242H-NR

TABLE OF CONTENTS

Executive Summary

Enterprise adoption of generative AI is accelerating, with large language models (LLMs) deployed across mission-critical workloads, including chatbots, content generation, summarization, and knowledge assistance. Scaling these workloads requires higher concurrency, larger models, and strict service-level objectives (SLOs)—including Time-to-First-Token (TTFT) and Time-Per-Output-Token (TPOT)—while managing infrastructure footprint, power, and total cost of ownership (TCO). At the same time, LLM inference is shifting toward larger reasoning-capable models such as 32B-class

architectures, which significantly increase demands on compute density, memory capacity, and memory bandwidth. Traditional 2-socket (2S) servers are reaching their limits, often requiring multi-node scale-out, which adds networking overhead and leaves resources underutilized—particularly for CPU-based workloads like Llama 3.1 8B Instruct and Qwen/QwQ-32B.

4-socket (4S) Intel® Xeon® 6 systems address this gap by consolidating more cores, memory, and I/O bandwidth into a single chassis, enabling vertical scaling before horizontal expansion. Testing on Supermicro X14 systems with MLPerf Inference v6.0 and vLLM workloads demonstrates strong scaling and improved performance-per-watt and performance-per-rack-unit efficiency. The 4S configuration delivers up to ~1.9X higher concurrency and ~1.86X higher MLPerf v6.0 throughput than 2S systems for Llama 3.1 8B Instruct. It uniquely supports Qwen/QwQ-32B—enabling larger-model inference beyond 2S capability while providing better TCO and SLA-compliant performance across diverse generative AI workloads.

Concurrency and Throughput Performance on 4-Socket X14 Systems with Intel Xeon 6[1]

For large-scale generative AI deployments, the Supermicro X14 platform highlights the benefits of scaling from 2-socket to 4-socket configurations with Intel Xeon 6 processors. The 2S system with dual Intel Xeon 6787P processors delivers strong performance for mainstream inference workloads. In contrast, the 4S system with four Intel Xeon 6788P processors significantly increases compute capacity to up to 344 Performance cores, along with higher memory bandwidth and density needed for larger language models and low-latency performance. Using vLLM v0.19.0, benchmarking was conducted with BF16 Llama 3.1 8B Instruct and Qwen/QwQ-32B across varied token lengths to evaluate real-world inference scaling and validate SLA-compliant support for larger models, demonstrating the scalability advantages of 4S systems for modern generative AI workloads.

The Intel Xeon 6 4-socket platform demonstrated exceptional scaling efficiency for enterprise AI inference workloads. Running Llama 3.1 8B Instruct, the 4S configuration achieved up to ~1.9X higher concurrent user capacity compared to the 2S platform, enabling organizations to support significantly more simultaneous AI users with minimal scaling overhead. In MLPerf Inference v6.0 submissions, the 4S system also delivered approximately ~1.87X higher throughput for both Llama 3.1 8B and Whisper workloads compared to the 6787P-based 2S system v6.0 submission numbers, showcasing the platform's ability to accelerate demanding AI inference applications while maximizing infrastructure efficiency.

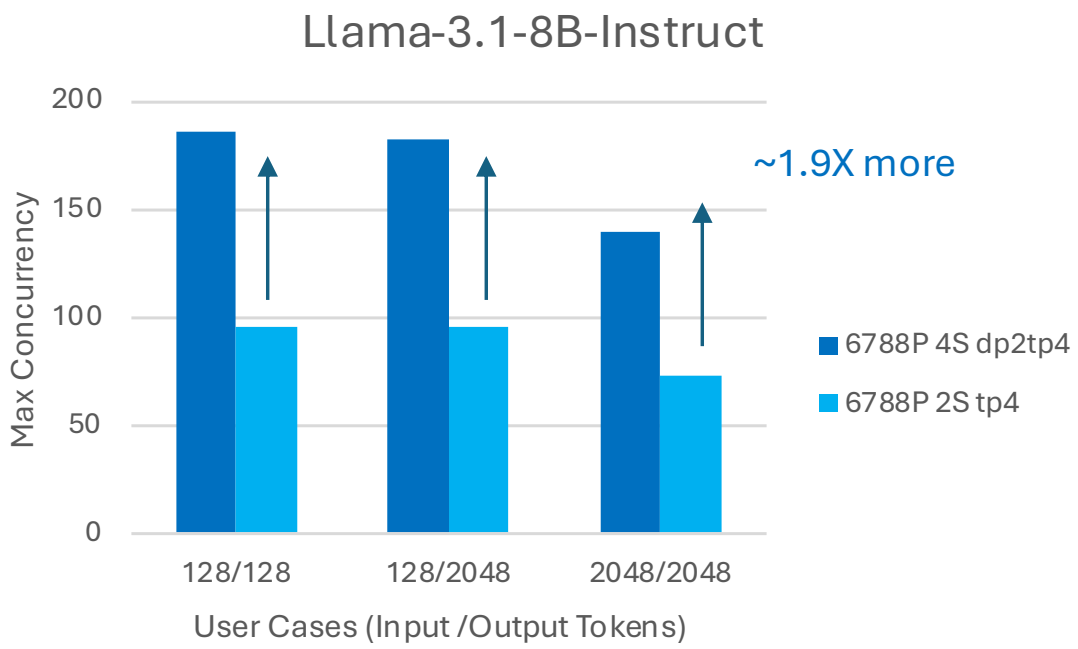


Figure 1 – Efficient 2S-to-4S Scaling for Maximum Llama 3.1 8B Concurrency in Inference

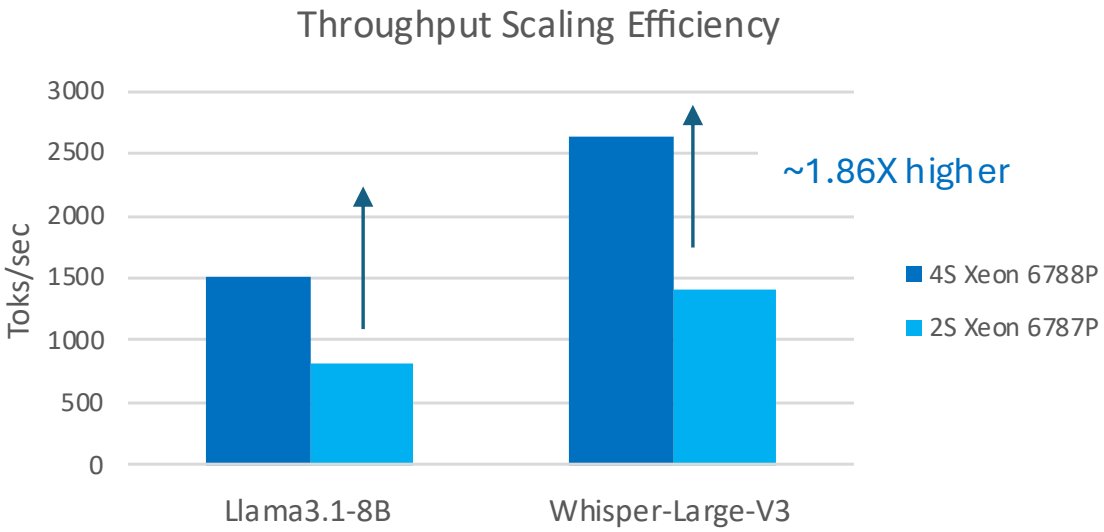
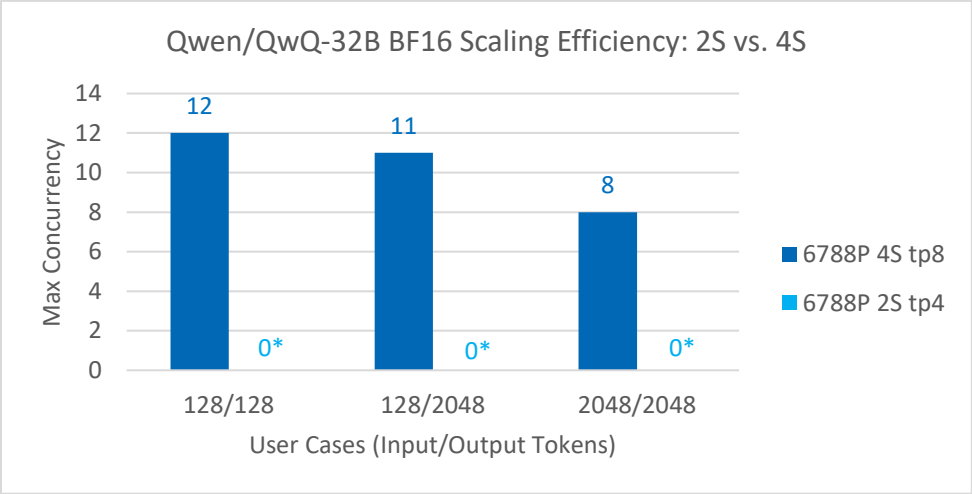


Figure 2 – Throughput Scaling Efficiency of 4S versus 2S Systems in MLPerf Inference v6.0

As enterprise LLM deployments evolve from basic chatbot and summarization workloads to more complex reasoning and long-context use cases, customers increasingly need larger models such as 32B-class LLMs to improve response quality, instruction following, and task accuracy; however, these larger models significantly increase demands on compute density, memory capacity, and memory bandwidth, making it more challenging to maintain strict low-latency SLOs. The Supermicro X14 4-socket

system addresses these requirements, with Qwen/QwQ-32B meeting enterprise SLA targets only on the 4S configuration, achieving TTFT below 3 seconds and TPOT under 100 milliseconds in chatbot use cases. These results highlight how the 4-socket Intel Xeon 6 architecture enables deployment of larger models while sustaining strict latency objectives through higher compute density, expanded memory capacity, and improved system efficiency.



** 2-socket system currently shows 0, indicating that the system did not meet the required SLA targets. Only a 4-socket system meets the SLA targets.*

Figure 3 - Qwen/QwQ-32B BF16 Scaling Efficiency: 2S vs. 4S

TCO Analysis of AI Infrastructure Scaling: Performance-per-Watt Efficiency

Assuming equivalent platform overhead across configurations, the 4S system consumes a total of 3402.3 W, including CPUs, motherboard, memory, fans, and storage, with four 350 W CPUs contributing 1400 W of the overall power envelope. By scaling the CPU contribution linearly, the 2S system is estimated to consume 2702.3 W, comprising 700 W from two CPUs (2 × 350 W) and 2002.3 W of shared platform power attributable to non-CPU components such as memory, storage, I/O, and cooling. From a TCO perspective, this corresponds to a ~1.26X increase in total system power when moving from 2S to 4S. At the same time, the 4S platform delivers up to ~1.9X higher concurrency and ~1.87X higher throughput, resulting in a net improvement in performance-per-watt efficiency of approximately ~1.5X. Although the 4S configuration increases absolute power consumption, it more than offsets this with disproportionate gains in compute capacity, yielding lower effective power per unit of AI inference work and improved infrastructure efficiency—particularly in high-density deployment environments where power, rack space, and scaling costs are key constraints.

Supermicro System for Testing

The Supermicro SYS-242H-NR integrates four Intel® Xeon® 6 processors into a 2U system, a factory-validated platform designed for enterprise and data center workloads. Below are the details of the system tested:

System	SYS-242H-NR
CPUs	Four Intel Xeon 6788P processors (86 cores, 350W TDP)
Memory	64 DIMM slots Up to 8TB DDR5-6400MT/s (1DPC) Up to 16TB DDR5-5200MT/s (2DPC)

Summary

Overall, the test highlights that Supermicro 4-socket systems with Intel Xeon 6788P processors provide a compelling TCO advantage for enterprise AI deployments. By combining higher sustained performance, better hardware consolidation, and the ability to support larger models under strict SLAs, the 4S platform enables organizations to reduce infrastructure footprint while maximizing both performance-per-watt efficiency in production AI environments.

For More Information:

Supermicro SYS-242H-NR: https://www.supermicro.com/en/products/system/mp/2u/sys-242h-nr

Intel Xeon product information: https://www.intel.com/content/www/us/en/products/sku/241837/intel-xeon-6788p-processor-336m-cache-2-00-ghz/specifications.html

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements. Visit www.supermicro.com

INTEL

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, visit

Visit www.intel.com

[1] Results may vary based on system configuration, workload characteristics, hardware environment, service level objectives, and other operational factors. Performance improvements are dependent on proper setup and optimization. Configuration: 1-node, Supermicro Super Server, 4x Intel(R) Xeon(R) 6788P, 86 cores, 350W TDP, HT On, Turbo On, Total Memory 2048GB (32x64GB DDR5 6400MT/s [6400MT/s]), BIOS 1.3, microcode 0x1000405, 2x Ethernet Controller X550, 1x 894.3G SAMSUNG MZ1L2960HCJR-00A07, 1x 14T SAMSUNG MZWLO15THBLA-00A07, 1x 2.9T KIOXIA KCD81PUG3T20, Ubuntu 24.04.3 LTS, 6.8.0-111-generic. Test by Intel as of Mon May 11 11:30:02 PM UTC 2026.