WHITE PAPER

SUPERMICRO AND INTEL® TEAM UP TO OFFER A HYBRID CPU+GPU ARCHITECTURE FOR SCALABLE, COST-EFFICIENT LLM INFERENCE USING INTEL® XEON® 6 AND INTEL® GAUDI® 3

Intel Xeon 6 6900P Processors Are Performance & Efficiency Focused

TABLE OF CONTENTS

Executive Summary

The generative AI (GenAI) landscape is undergoing a strategic paradigm shift, transitioning from monolithic, single-model interactions toward sophisticated, modular multi-agent systems. In this emerging architecture, complex user queries are no longer treated as atomic operations; instead, they are decomposed into orchestrated pipelines of specialized sub-tasks—ranging from data retrieval and intent classification to high-order reasoning—each demanding distinct Workload characteristics.

Our evaluation focuses on a hybrid CPU-GPU architecture that combines Intel Xeon 6 processors with Intel Gaudi 3 accelerators within the same system. We measured key performance metrics - including latency, throughput, Time to First Token (TTFT), and

Time Per Output Token (TPOT) - while simultaneously analyzing CPU optimization strategies to maximize GPU utilization by using Intel® VTune™ Profiler to identify and implement effective configurations. Building on this evaluation, we include a performance comparison with previous-generation configurations to quantify improvements in efficiency and scalability under concurrent user workloads, validated using MLPerf-aligned benchmarks.

Our key findings demonstrate that Intel Xeon 6 processors efficiently manage high-concurrency workloads for smaller LLMs, while Intel Gaudi 3 hybrid deployments significantly accelerate large-model inference. This architecture excels at processing extensive token sequences with high throughput and low latency. Furthermore, our analysis identifies specific configurations that optimize mixed-load stability, ultimately validating CPU-native inference as a high-efficiency tier for modern production environments.

This evaluation demonstrates that hybrid CPU–GPU architectures provide a versatile and efficient platform for modern LLM workloads. By intelligently partitioning tasks, offloading large-model inference to Intel Gaudi 3 accelerators while utilizing idle CPU cores for smaller models, the system maximizes overall utilization, reduces hardware requirements, and preserves GPU resources for high-value workloads. This approach enables high-throughput, low-latency inference across diverse model sizes, improves total cost of ownership (TCO) by reducing the need for additional GPUs or servers, and delivers higher efficiency per deployed GPU, ensuring better performance-per-dollar and more effective AI infrastructure investment.

Introduction

Agentic AI systems often orchestrate multiple smaller or domain-specific LLMs to accomplish task sequences or collaborate on subtasks, in contrast to the traditional approach of deploying a single large, monolithic LLM for everything. Smaller models can be more efficient for specialized workloads and adapt to diverse application requirements while consuming significantly less compute than large general-purpose models. Furthermore, running lightweight models efficiently often benefits from optimized serving stacks and architectures that strike a balance between throughput and cost. This shift underscores the importance of hybrid CPU–GPU inference architectures rather than relying solely on high-cost GPU resources.

Enterprises are facing significant challenges in scaling inference to support a high number of concurrent users while maintaining throughput and latency targets. While GPU-only architectures excel at massive parallel tensor operations for large-scale models, they are often underutilized during lighter inference workloads. This underutilization leads to inefficient capital expenditure (CAPEX) as concurrency scales.

Our empirical analysis indicates that serving large-scale models on the Supermicro Intel Gaudi 3 platform—despite its immense throughput—often leaves a majority of the Intel Xeon 6 CPU cores idle on the same server, as only a fraction of the processor's cycles are consumed by GPU orchestration and host-side management.

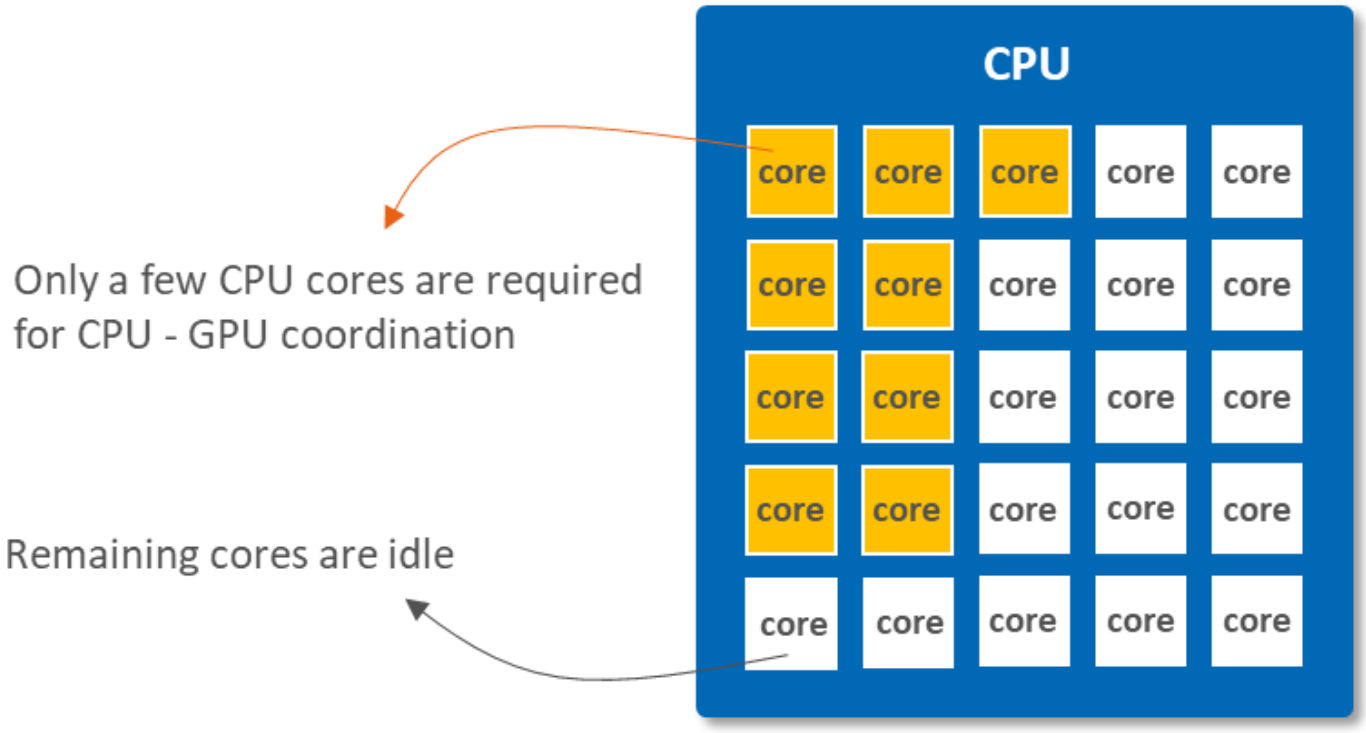

Figure 1: Asymmetric Workload Distribution: CPU-GPU Coordination

To optimize Total Cost of Ownership (TCO) and maximize infrastructure utilization, a more balanced architectural approach is required. Modern AI pipelines demand a system where the CPU can handle a significant portion of the inference chain - particularly for smaller models (e.g., Llama-3-8B or Phi-3) and pre/post-processing requests. Intel Xeon 6 processors with high-performance P-cores are specifically engineered for this paradigm. Featuring an increased core count (up to 128 cores per socket) and Intel® Advanced Matrix Extensions (Intel® AMX), these processors deliver the compute density needed to offload inference workloads directly to the CPU. This strategy not only preserves expensive GPU resources for high-parameter models but also delivers a more efficient, scalable, and cost-effective AI infrastructure for the enterprise.

Real World Scenario

To demonstrate the practical effectiveness of this hybrid architecture in real-world scenarios, we designed a "Research + Writing" multi-agent system. In this implementation, the multiple AI agents collaborate through a structured pipeline managed by the LlamaIndex AgentRunner:

1. Researcher Agent: Executes information retrieval using specialized tools (e.g., fetch background) to generate factual summaries.
2. Writer Agent: Performs reasoning and content generation to compose clear, engaging drafts.
3. Reviewer Agent: Handles review tasks, including linguistic polishing and ensuring the correctness of the final output.

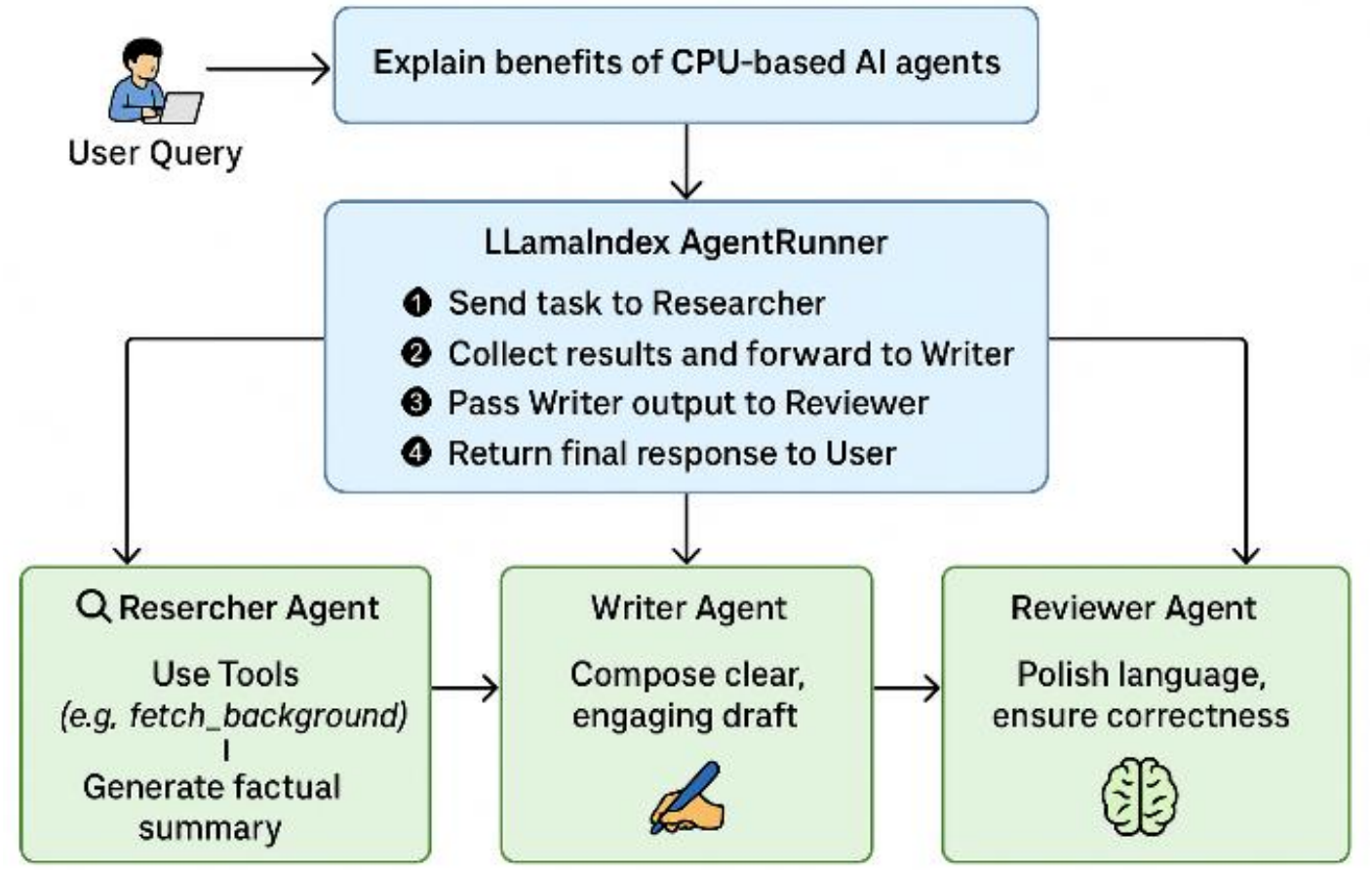

Figure 2 - Automated Content Pipeline: Researcher, Writer, and Reviewer Agents

The table below illustrates how specific AI agent roles are mapped to the hybrid CPU–GPU architecture to maximize operational efficiency.

The following section provides a deep dive into our methodology, accompanied by empirical evidence demonstrating how this hybrid architecture maximizes hardware efficiency. We show that leveraging high-performance CPU cores for orchestration directly enhances accelerator-based inference, optimizing overall throughput and ensuring adherence to strict responsiveness SLAs across diverse model scales.

Stage	Role	Core Operation	Relative Complexity	Common input/output length	Model Type Fit	Arch Fit
Researcher	Retrieve and summarize relevant info	Moderate reasoning, retrieval, synthesis	Medium	1500/400	Smaller reasoning model (e.g., Llama 3 8B, Mistral 7B, Qwen3 8B)	**CPU**
Writer	Generate structured, coherent output	High creative + compositional load	High	1000/1200	Larger fluent model (e.g., Llama 3 70B/405B, Mixtral 8×7B, or Claude 3 Sonnet)	**Gaudi3**
Reviewer	Analyze, evaluate, refine text	Logical, critical reasoning	Medium-High	1200/1200	Models with strong evaluation skills (e.g., GPT-4-Turbo, Claude 3.5 Sonnet, Llama 3 70B/405B)	**CPU Gaudi3**

Table 1 - Resource Allocation Strategy for Researcher, Writer, and Reviewer Agents

Methodology

We adopted a hybrid CPU - GPU architecture in our testing methodology to comprehensively evaluate the combined performance of the Intel Gaudi 3 accelerator and Intel® Xeon® 6 processors using large-scale and smaller language models. This section describes the experimental setup, the models evaluated, benchmark scenarios, and the metrics for assessing performance.

Experimental Setup

The experiments were conducted on a high-performance Supermicro server featuring a hybrid CPU–GPU architecture. The system configuration is designed to maximize performance for both CPU- and accelerator-based inference workloads:

- CPU: 2 × Intel Xeon 6960P
  - Total cores: 144
  - NUMA nodes: 6 across 2 sockets
  - Features: Intel AMX for efficient tensor processing and CPU-assisted inference, pre/post-processing, and hybrid workload handling

- Accelerators: 8 × Intel Gaudi 3 OAM mezzanine cards
  - High memory bandwidth and scalable interconnects support large model inference.
  - Optimized for low-latency, high-throughput transformer workloads.

Models Evaluated

To evaluate system performance across different workload scales, the following large language models (LLMs) were selected and used for inference testing:

- **Llama 3-405B:** A large-scale model designed to stress accelerator compute capacity. It is used to evaluate performance under high-throughput, latency-sensitive scenarios typical of content generation and complex inference workloads.
- **Llama 3.1-8B:** a compact language model with 8 billion parameters, representative of CPU-friendly inference workloads and suitable for evaluating CPU-only execution efficiency.

Note:

- The full list of models supported for CPU-based inference can be found in the vLLM documentation: https://docs.vllm.ai/en/latest/models/hardware_supported_models/cpu/

- The full list of models validated and supported on Intel® Gaudi® 3 accelerators is available in the vLLM documentation: https://docs.vllm.ai/projects/gaudi/en/latest/getting_started/validated_models.html

Performance Metrics

The following metrics were measured during the experiments to gauge inference performance:

- Time to First Token (TTFT) measures the latency from the initiation of a request to the delivery of the first generated token. It is critical for assessing the initial system overhead and user-perceived responsiveness.
- Time Per Output Token (TPOT) represents the average duration required to compute each subsequent token. This provides a measure of sustained inference throughput and the model's efficiency during continuous generation tasks.

These metrics together provide insight into both initial responsiveness and sustained compute efficiency, which are essential for real-world deployment scenarios such as conversational AI and content generation.

Benchmark Scenarios

The hybrid CPU–GPU architecture was evaluated under two primary scenarios to characterize its performance across diverse workloads:

- Idle CPU resources serving lightweight chatbot workloads:
    - The Llama 3.1-8B, configured with 128-token input and output sequence lengths, was deployed on Intel Xeon 6900P series processors.
    - This scenario assessed CPU inference efficiency, system responsiveness, and the ability to handle multiple lightweight tasks concurrently without degrading performance.

- Accelerators serving content generation workloads:
    - Llama 3-405B, configured with 2K-token input and output sequence lengths, was executed on 8 × Intelsig accelerators.
    - The goal was to measure throughput and latency when processing large, transformer-based models that demand high compute and memory bandwidth.
    - This scenario also evaluated CPU and GPU utilization, demonstrating how resources are efficiently allocated across the hybrid architecture to maximize overall system performance.

Results and Analysis

Testing demonstrates that idle CPU cores can run inference for the Llama 3.1-8B model while meeting SLA targets **(TTFT < 3 s, TPOT < 100 ms)**, without constraining Intel® Gaudi® 3 accelerators dedicated to the Llama 3-405B workload.

With an optimal allocation of CPU cores, Llama 3-405B running on Intel® Gaudi® 3 also achieves SLA-compliant performance **(TTFT < 10 s, TPOT < 100 ms)**. This hybrid CPU–GPU configuration enables concurrent execution of small and large models, maximizing overall system throughput while maintaining SLA compliance.

Throughput Analysis

Observations: As the number of CPU cores dedicated to the accelerators increases from 6 to 18, throughput rises from 302 tokens/sec to a peak of 812 tokens/sec. This demonstrates that CPU support significantly improves accelerator pipeline efficiency by enabling parallel data preparation, task scheduling, and execution of parallelizable workloads on the CPU. Beyond 18 cores, throughput declines slightly to 736 tokens/sec at 24 cores, indicating that excessive CPU allocation can lead to diminishing returns due to scheduling overhead and resource contention.

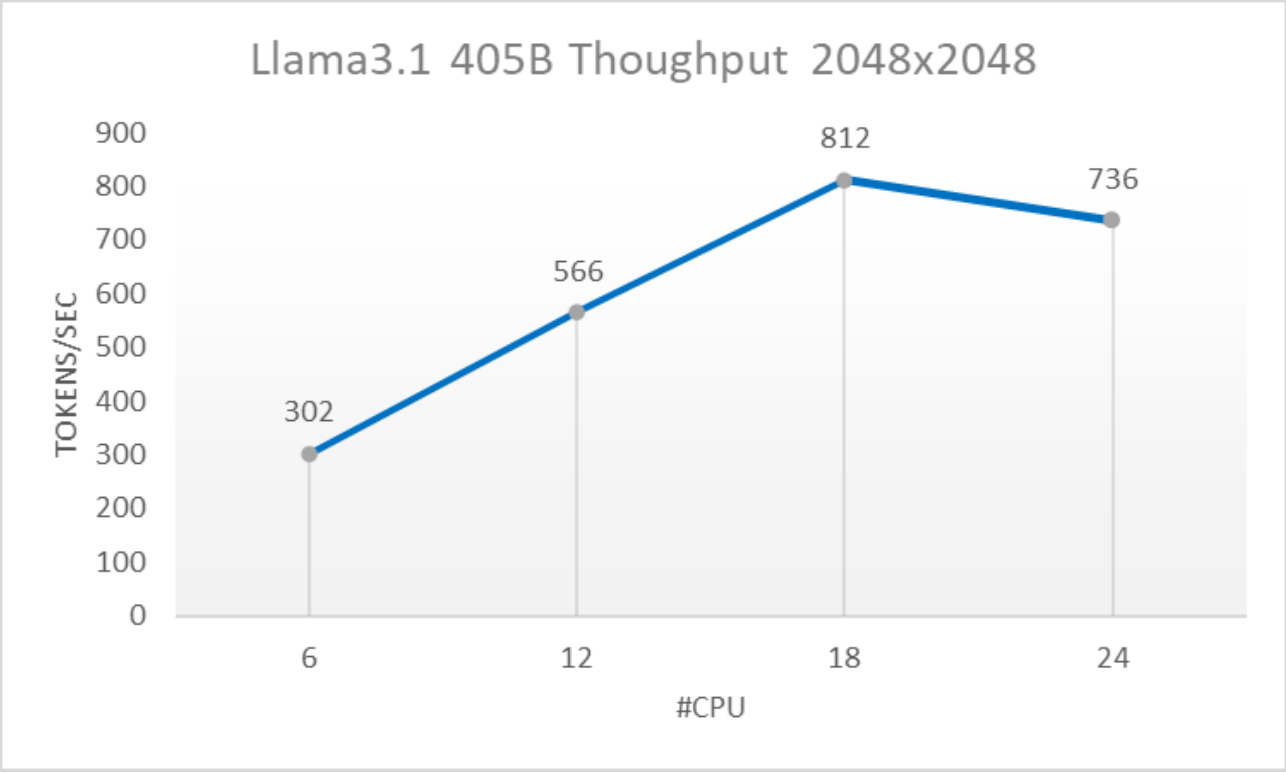

Figure 3 - Throughput Scaling of Llama 3.1 405B Across CPU Cores

TPOT (Time Per Output Token) Analysis

Observations Correspondingly, TPOT decreases as CPU allocation increases, from 209 ms at 6 cores to a minimum of 76 ms at 18 cores. Increasing the allocation beyond 18 cores results in a slight increase in latency to 84.67 ms at 24 cores, consistent with the observed drop in throughput. This behavior highlights the impact of scheduling overhead and CPU frequency reduction when over-allocating cores.

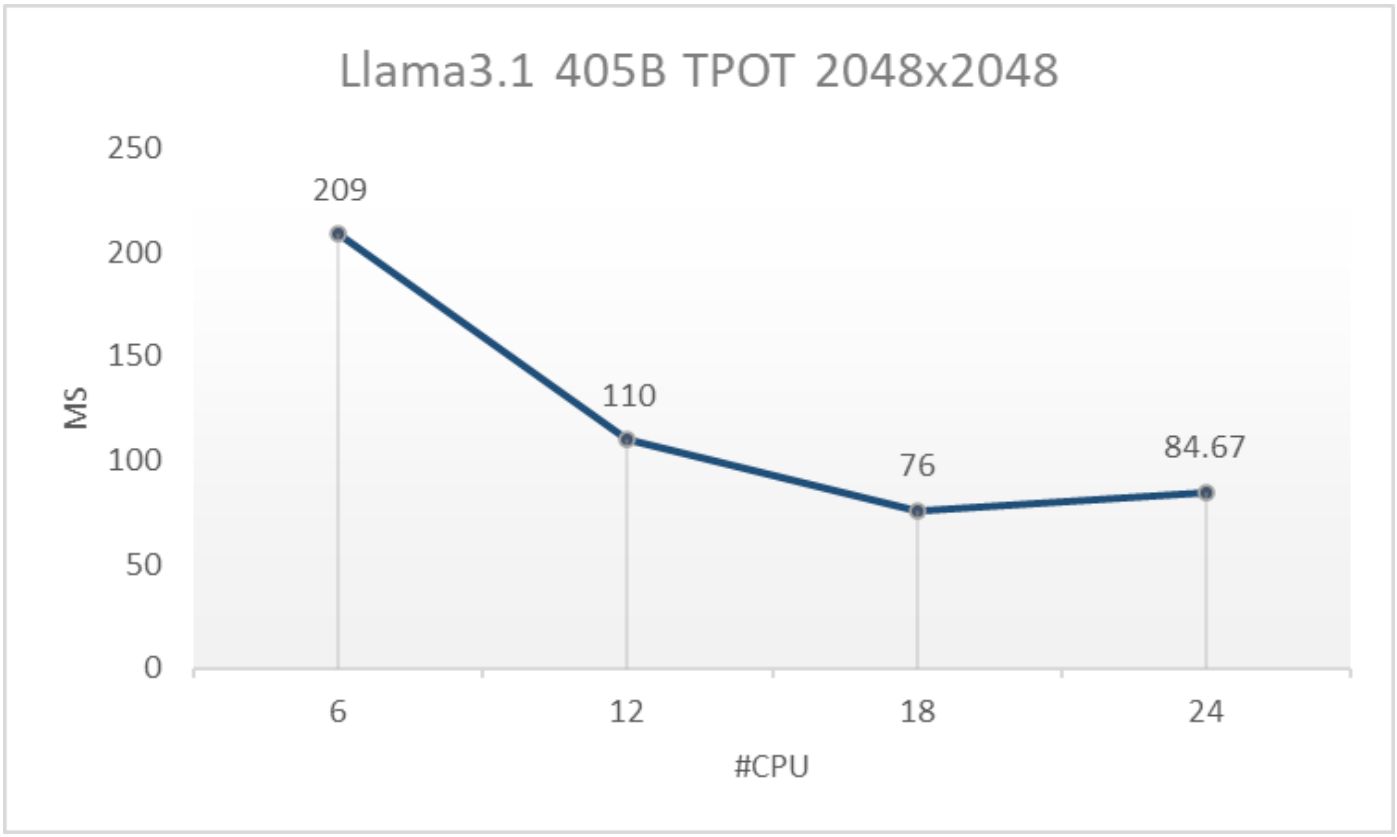

Figure 4 - Llama 3.1 405B TPOT Scaling: Latency vs. CPU Core Count

VTune Performance Analysis: Impact of CPU Core Binding

We conducted a detailed performance analysis using Intel VTune Profiler to quantify how CPU core binding fundamentally improves hardware efficiency when executing lama-3 405B inference on an Intel Gaudi 3 accelerator configuration. The results demonstrate that CPU core binding plays a critical role in stabilizing CPU behavior, improving microarchitectural efficiency, and ensuring balanced execution in large-scale, multi-accelerator inference workloads.

The observed performance differences are systematically evaluated by comparing configurations with CPU core binding enabled against those using default OS scheduling (unbound) across several key performance metrics.

- Binding Improves Efficiency: Binding the threads to specific, high-performance CPU cores allows the CPU to run within a high-frequency range for a longer duration.
- Eliminates Low-Frequency Core Use: When unbound, code can migrate between many different frequency cores, including low-frequency ones. Binding almost completely eliminates the use of low-frequency cores, removing this performance bottleneck.
- Better CPI and CPU Time Distribution: This binding strategy results in a lower CPI (meaning higher execution efficiency) and a more uniform and predictable distribution of CPU time, reducing unnecessary thread migration between cores.

Metric	Unbound (OS Default)	Binding (Pinned)
Average CPU Frequency	1.9–3.9 GHz (wide)	3.2–4.15 GHz (tight)
CPI (Cycles Per Instruction)	0.17–0.19	0.13–0.17
CPU Time	Highly uneven	Uniform
Instruction Retired	Uneven	Balanced
Low-Frequency Cores	Many	Almost none

Table 2 - Comparative Analysis: Unbound OS Scheduling vs. CPU Binding

Note: Detailed instructions for enabling CPU core binding in vLLM are available in the official documentation: https://docs.vllm.ai/projects/gaudi/en/latest/getting_started/quickstart/quickstart_configuration.html#pinning-cpu-cores-for-memory-access-coherence

CPU Frequency and Accelerators Utilization Analysis

It shows that allocating more CPU cores does not necessarily improve accelerator utilization. In this case, 18 pinned CPU cores provide the optimal balance between CPU operating frequency and Intel Gaudi 3 accelerator efficiency.

Observations:

With 18 CPU cores allocated, the CPU frequency averages 2.3 GHz, while the Intel Gaudi 3 accelerator utilization achieves 40%, as illustrated:

#CPU=18

CPU Frequency is around 2300 Hz.

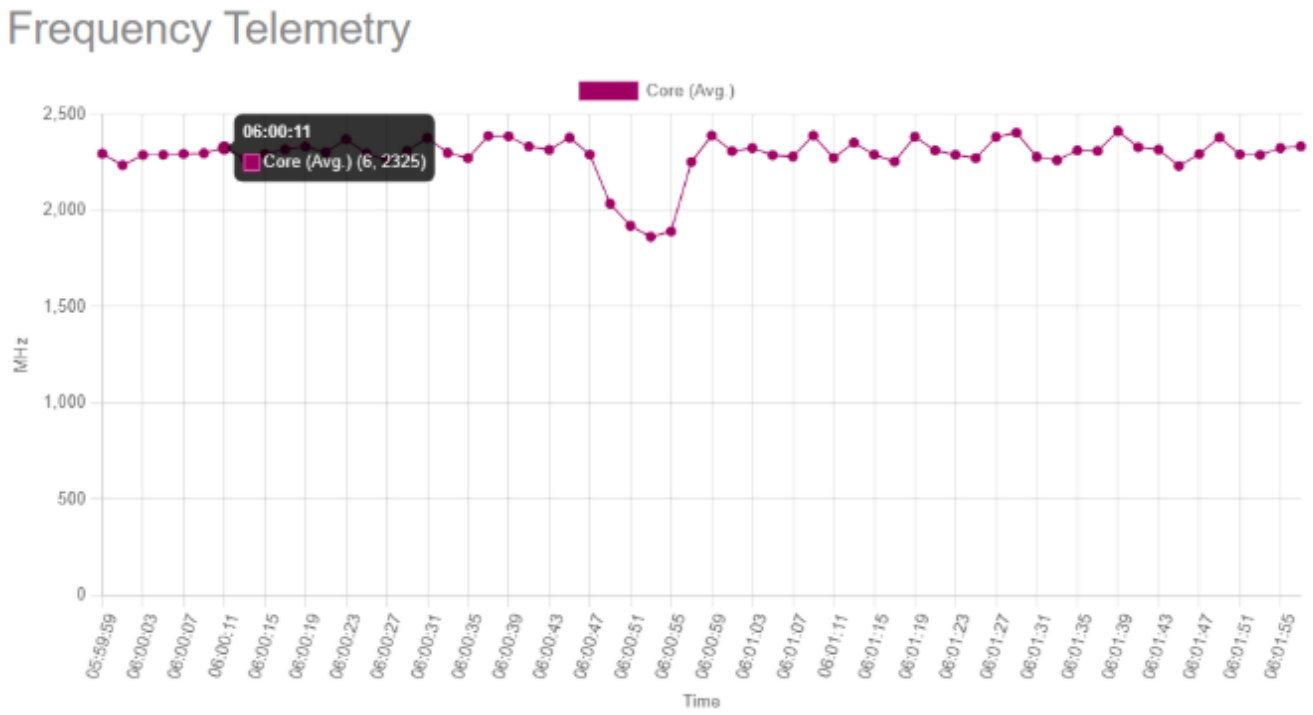

Figure 5 - CPU Frequency Stability at 18 Cores

Gaudi utilization is around 40%.

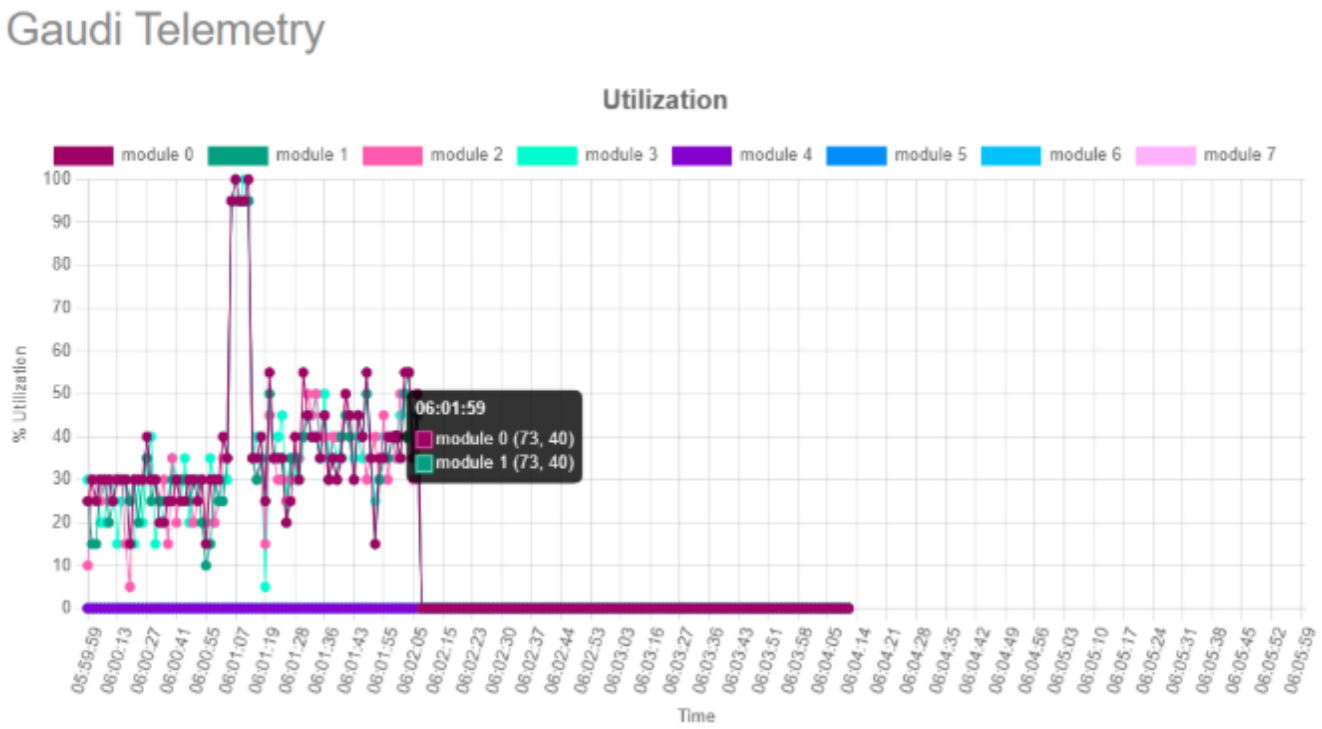

Figure 6 – Intel Gaudi Accelerator Utilization (40% Load)

With 24 CPU cores allocated, the CPU frequency drops to 1.8 GHz, resulting in 30% Intel Gaudi 3 utilization as illustrated:

#CPU = 24
CPU frequency dropped to ~1800 Hz

Figure 7 - CPU Frequency Stability at 24 Cores

Gaudi utilization dropped to 30%.

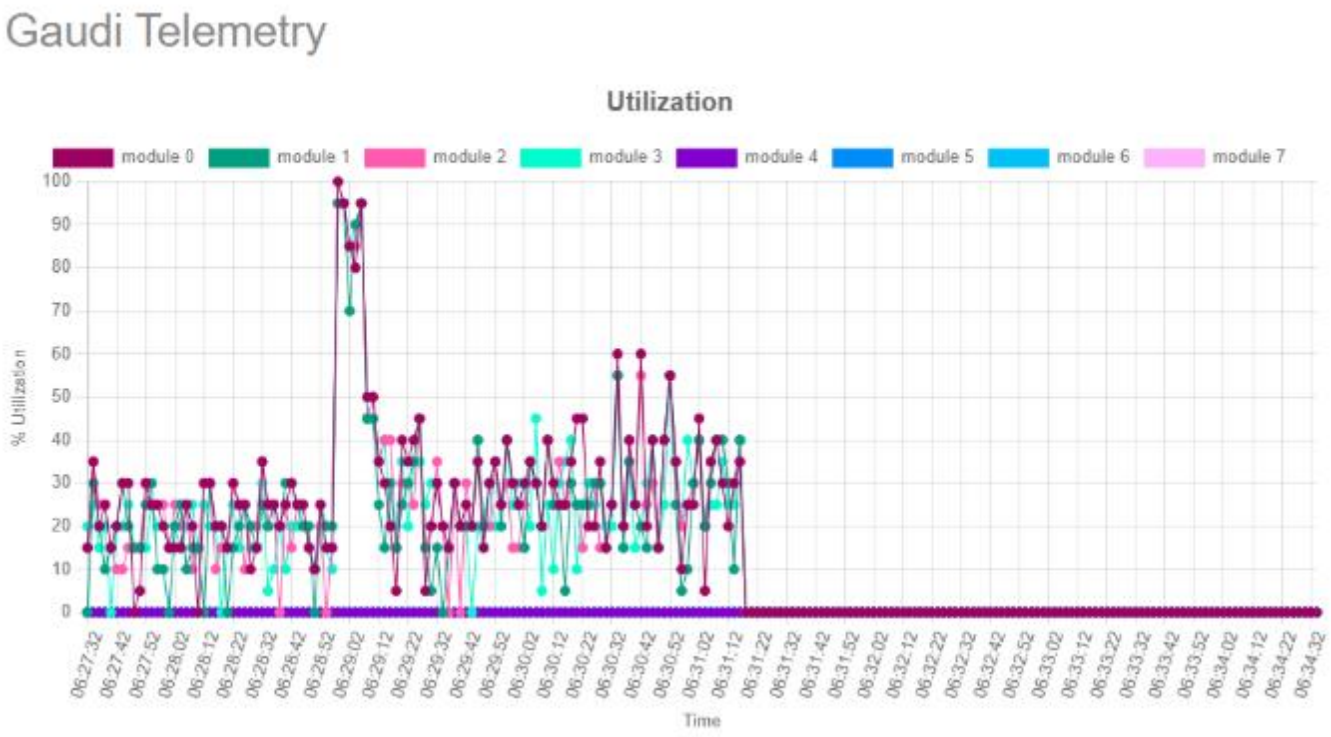

Figure 8 – Intel Gaudi 3 Accelerator Underutilization (30% Load)

Significant Inference Uplift from the 5th Gen Intel Xeon Processors to the Intel Xeon 6 6900P Processors

The Intel Xeon 6 6900P processors deliver substantial gains over the 5th Gen Intel Xeon processors in both CPU-only and hybrid Intel Gaudi 3 inference. CPU-only concurrency improves by **2.66X for Llama-3 8B**, while hybrid CPU+GPU deployments scale to **128 users** with proper CPU core pinning. This section describes our experiments on CPU-only, hybrid CPU+GPU, and GPU-only inference scenarios.

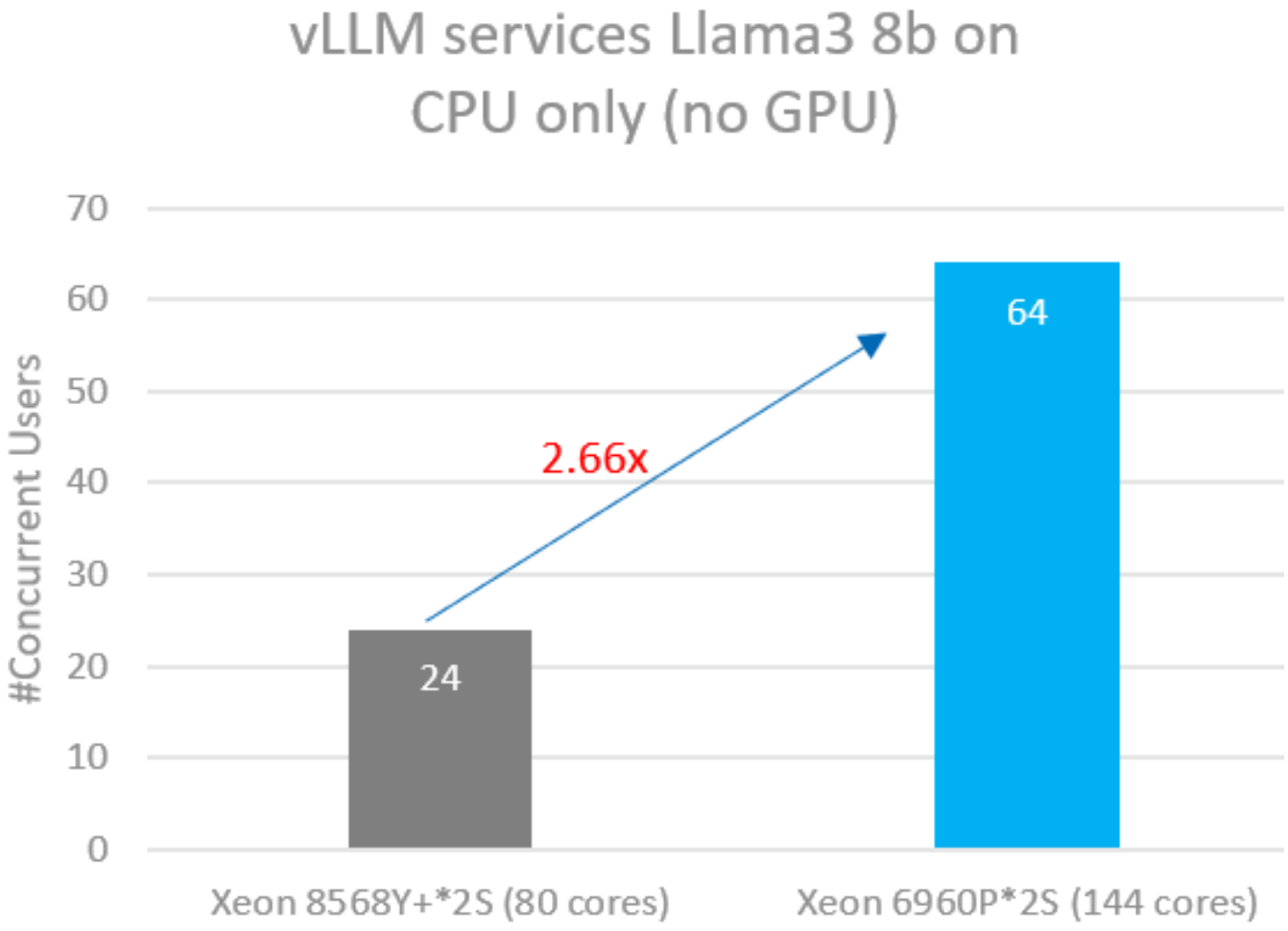

Figure 9 - Scaling Llama 3 8b Performance on CPU: Intel Xeon 8568Y vs. Intel Xeon 6960P processors

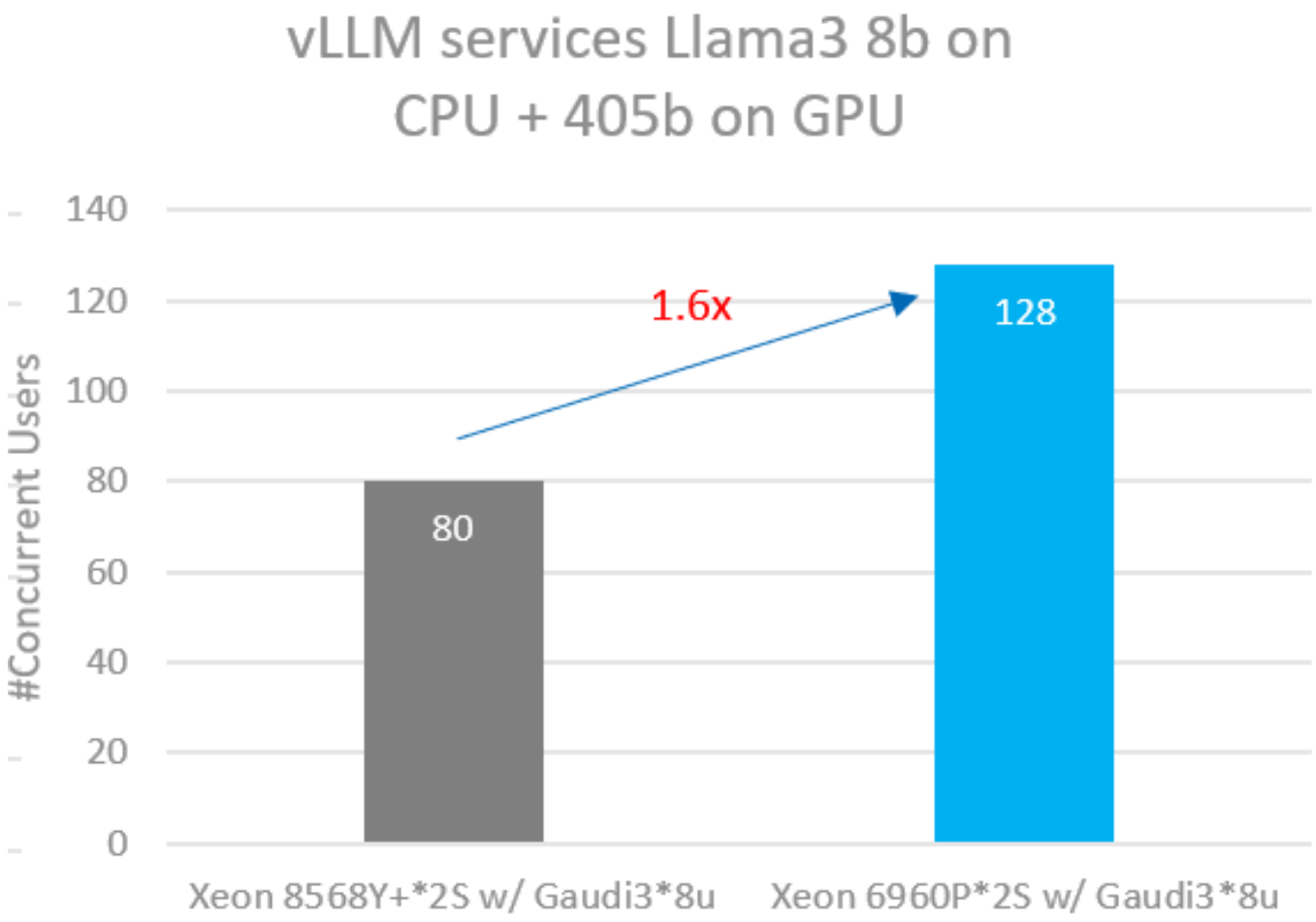

Figure 10 - Heterogeneous Scaling: Combining Llama 3 8b (CPU) and 405b (GPU)

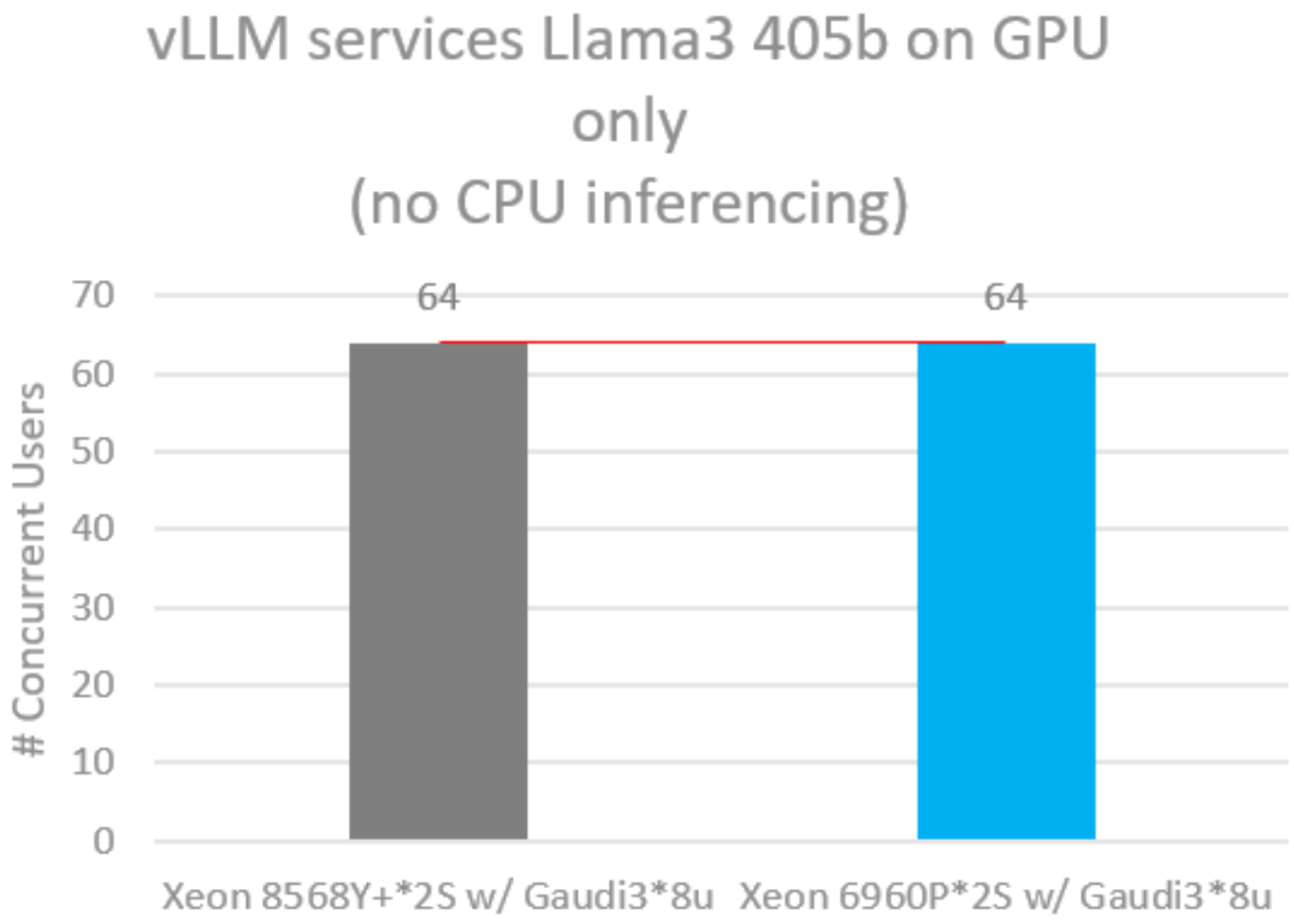

Figure 11 - GPU-Bound Performance: Constant Capacity Across Intel Xeon Generations

CPU-only inference scalability:

When utilizing the full CPU resources of an Intel Gaudi 3 server without GPU involvement, Intel Xeon 6 6972P delivers 64 concurrent users for Llama-3 8B inference, compared to 24 users on the 5th Gen Intel Xeon 8568Y+ with 80 CPU cores and 2 NUMA nodes under identical test conditions and SLA (TTFT <= 3 sec , TPOT <=100ms), representing a 2.66× generation-to-generation performance uplift driven by higher core counts and enhanced AI capabilities.

CPU + GPU co-processing benefits:

In hybrid deployments running Llama-3 405B on Intel Gaudi 3 accelerators with proper CPU core pinning, pairing Intel Gaudi 3 with Intel Xeon 6 processors scales concurrency to 128 users, versus 80 users on the 5th Gen Intel Xeon processors-based

systems, demonstrating that improved CPU-side AI performance materially enhances end-to-end throughput in accelerator-attached inference scenarios.

GPU-bound baseline comparison:

When inference is executed exclusively on Intel Gaudi 3 accelerators without CPU-side inference, both Intel Xeon 6 6900P- and the 5th Gen Intel Xeon processors-based Intel Gaudi 3 servers achieve the same 64 concurrent users with 2K input length and 2K output length under SLA (TTFT <= 15 sec, TPOT<= 100 ms), confirming that the observed gains in hybrid configurations are attributable to GNR-AP's superior CPU AI processing capability, not GPU limitations.

Intel Xeon 6 6900P Throughput Performance

The Intel Xeon 6 6900P processors boost CPU inference throughput over **3X** versus prior generation, supports concurrent CPU+GPU workloads efficiently, and Intel Gaudi 3 throughput remains stable across CPU generations

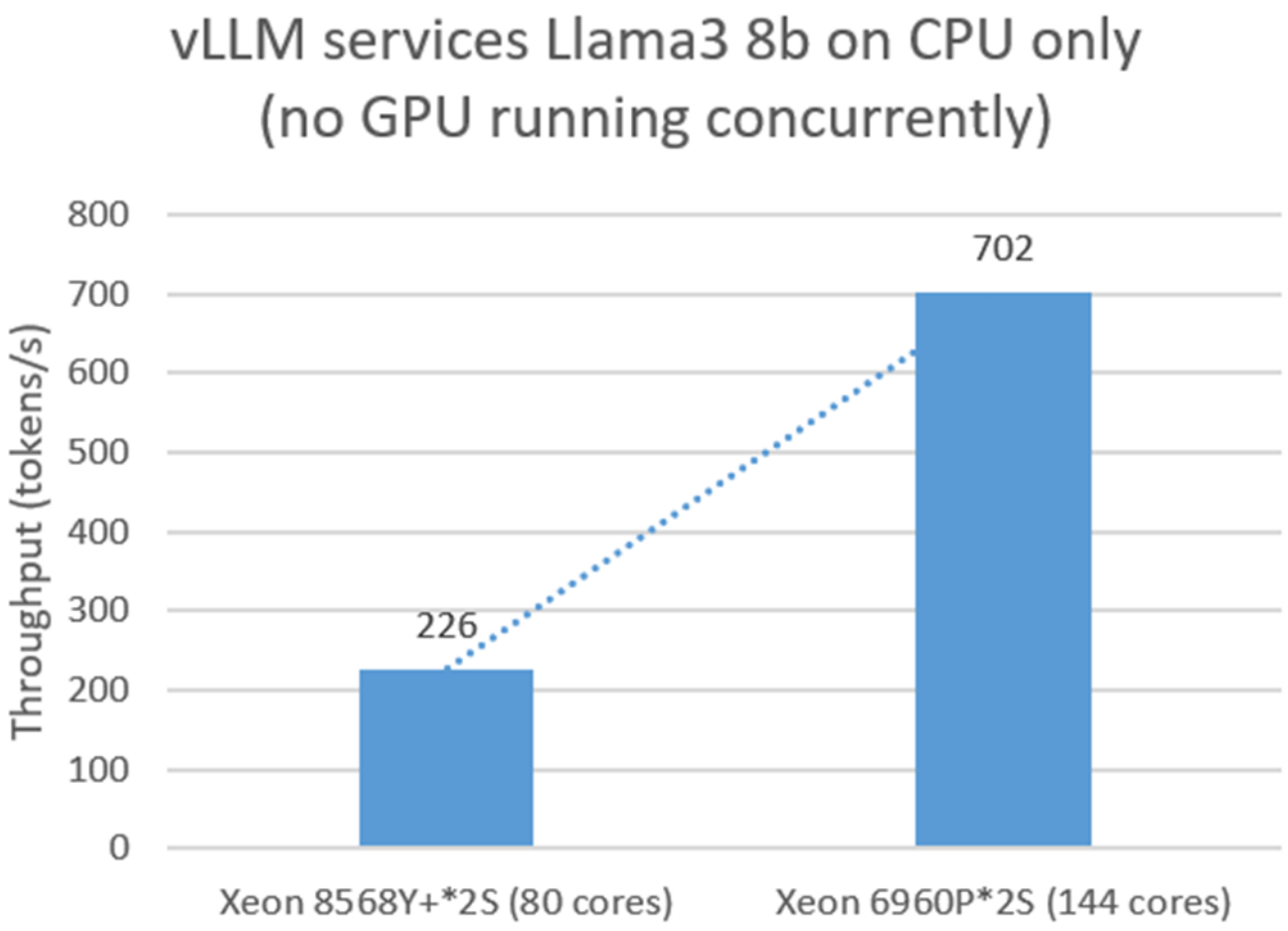

Figure 12 - CPU-Only Throughput Scaling: Llama 3 8b Benchmarks

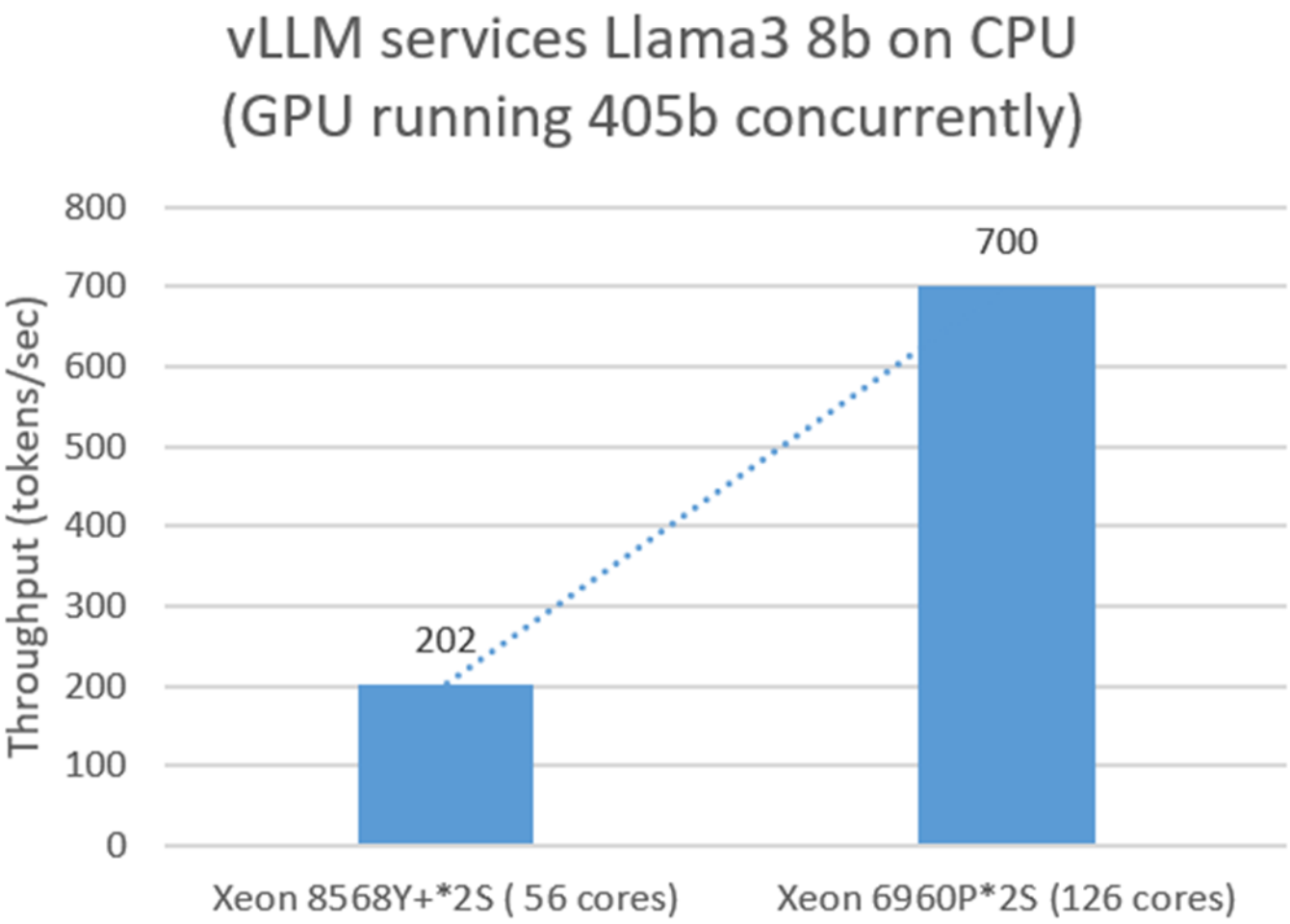

Figure 13 - CPU Throughput with Concurrent 405b GPU Workload

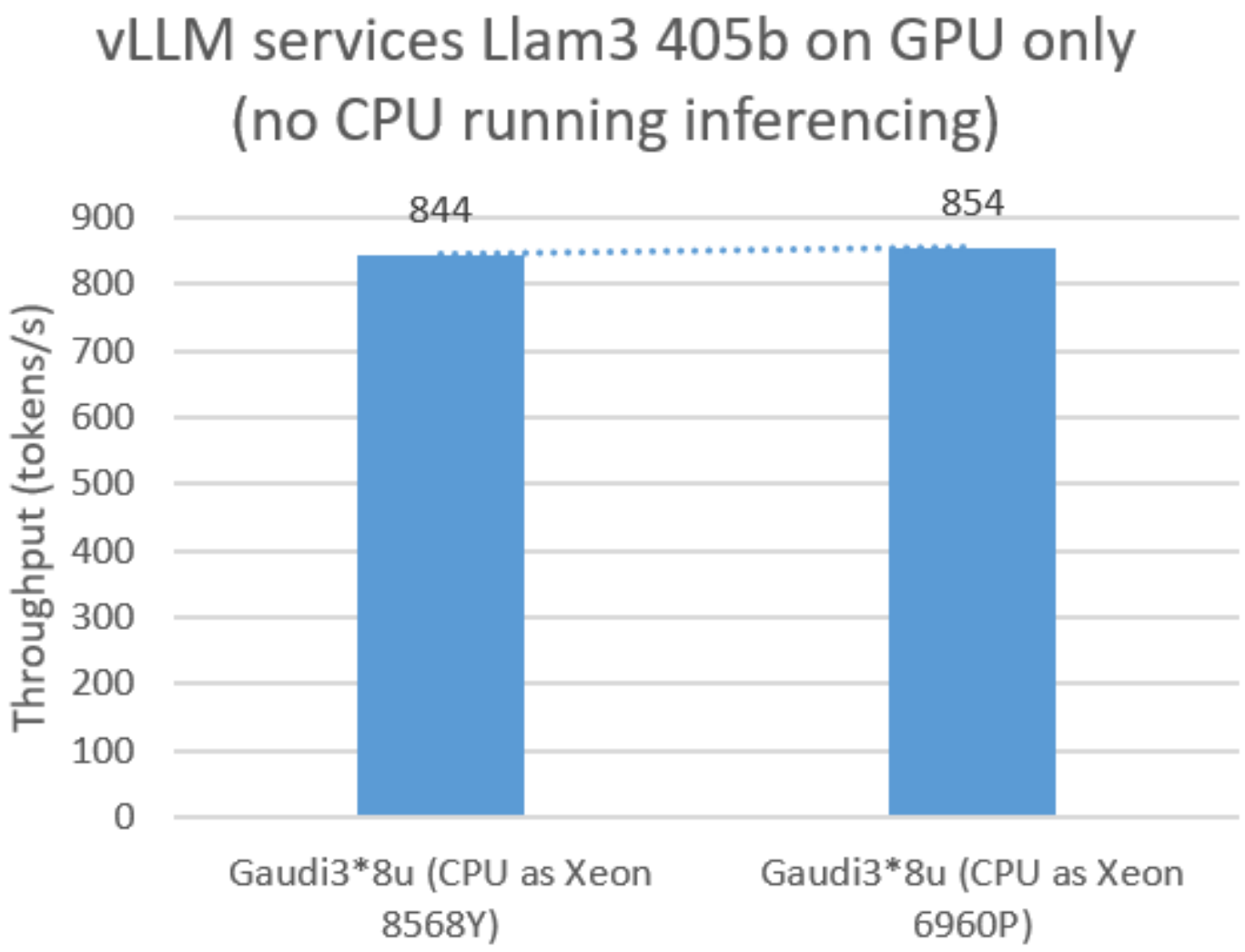

Figure 14 - GPU Throughput (Intel Gaudi 3) for Llama 3 405b

Significant CPU inference throughput uplift with Intel Xeon 6 processors:

The Intel Xeon 6 6900P processors deliver >3× higher Llama3-8B vLLM throughput versus prior-generation Xeon®, demonstrating strong scalability and positioning CPU inference as an effective option for production deployments.

Enables efficient CPU and GPU concurrency:

The Intel Xeon 6 6900P processors sustain CPU inference performance while GPUs concurrently execute large-model workloads, supporting mixed-workload AI environments without compromising system efficiency.

Intel Gaudi 3 throughput remains consistent across CPU generations:

The Intel Gaudi 3 demonstrates similar throughput when paired with either the 5th Gen Intel Xeon or the Intel Xeon 6 6900P processors, indicating stable, predictable GPU behavior independent of the host CPU generation.

MLPerf v5.1 Validates Intel Xeon 6 as a High-Performance AI Inference CPU

The Intel Xeon 6 6900P processors boost CPU inference throughput over **3X** versus prior generation, supports concurrent CPU+GPU workloads efficiently, and Intel Gaudi 3 throughput remains stable across CPU generations.

MLPerf validates generation-to-generation Intel CPU AI performance improvement across diverse workloads:

MLPerf v5.1 demonstrates ~1.9×–2.0× inference performance gains for the Intel Xeon 6 6900P processors versus the 5th Gen Intel Xeon processors, based on v4.1 submissions on non-LLM AI workloads such as DLRM v2 and RetinaNet, showing that the Intel Xeon 6 6900P processors deliver strong AI performance beyond Llama-based models.

Continuously improved AMX for higher AI efficiency:

Generation-to-generation enhancements to Intel® AMX significantly improve tensor execution efficiency, enabling Intel Xeon 6 6900P processors to deliver increasingly efficient, scalable CPU-based AI inference across a broad range of AI workloads.

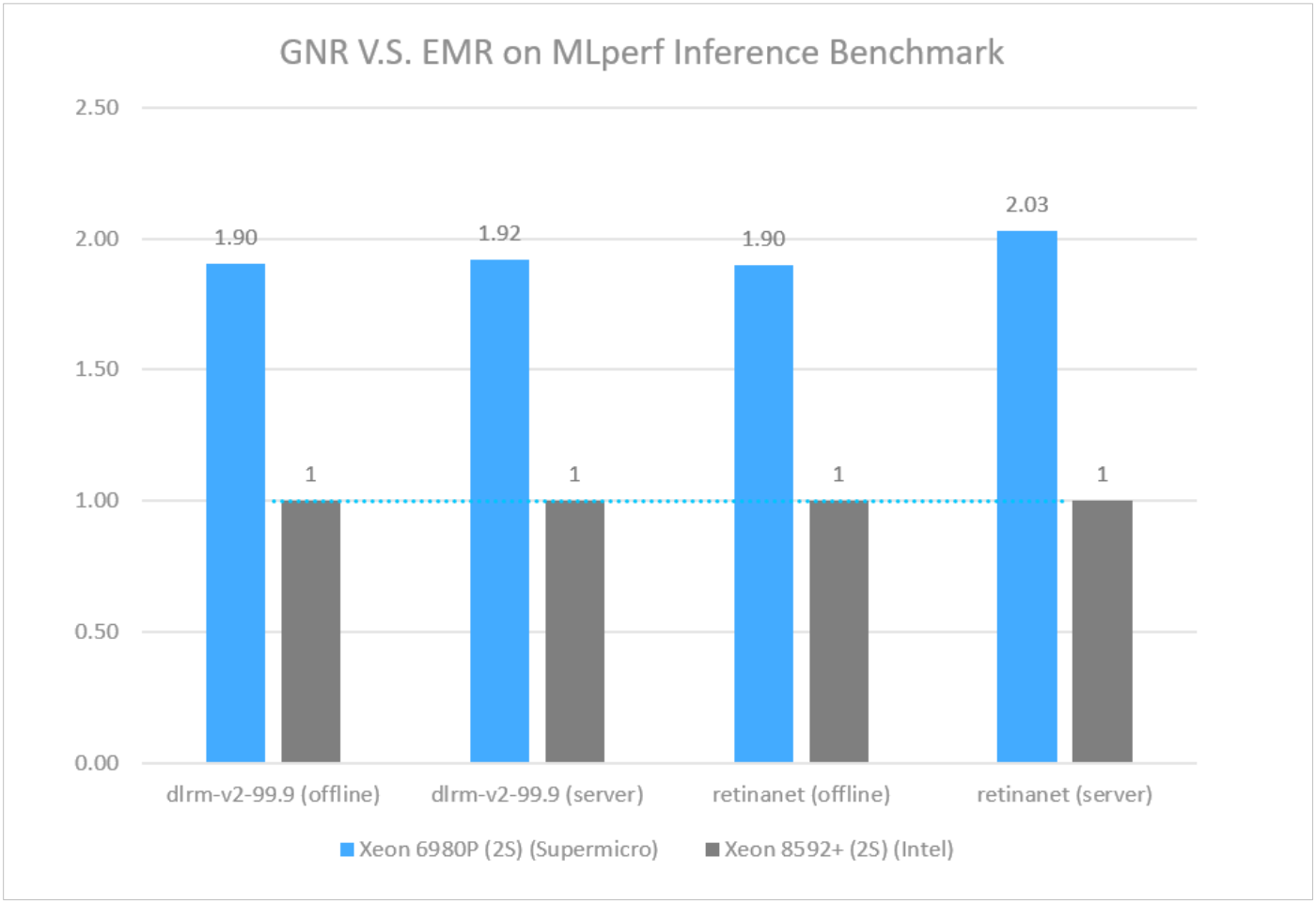

Figure 15 - MLperf Inference Performance: Intel Xeon 6 vs. the 5th Gen Intel Xeon

Conclusion

This technical evaluation validates that a hybrid architecture combining Intel Xeon 6 and Intel Gaudi 3 provides a high-efficiency platform for modern, multi-agent GenAI systems. By strategically partitioning workloads—offloading large-scale models like Llama 3-405B to Intel Gaudi 3 accelerators while leveraging high-performance CPU cores for smaller models like Llama 3.1-8B—the platform maximizes total system utilization and avoids the common pitfall of GPU underutilization.

Our analysis demonstrates that optimized CPU core binding is critical for hardware efficiency; using 18 pinned cores achieves peak throughput of 812 tokens/sec and reduces TPOT to 76 ms, while stabilizing execution across multi-accelerator workloads by eliminating thread migration to low-frequency cores. Furthermore, the transition to Intel Xeon 6 processors delivers a significant 2.66× generational performance uplift in CPU-only inference and scales hybrid concurrency to 128 users while strictly meeting TTFT and TPOT SLAs.

Ultimately, this intelligent role separation improves the Total Cost of Ownership (TCO) by reducing hardware overprovisioning, potentially saving approximately 1/8 of a GPU server's cost—thereby ensuring a superior performance-per-dollar return on investment in AI infrastructure.

For More Information:

Supermicro SSG-222B-NE3X24R: https://www.supermicro.com/en/products/system/storage/2u/ssg-222b-ne3x24r

Intel Gaudi 3: https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi.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