PRODUCT BRIEF

SUPERMICRO X14 NVIDIA HGX B300 SYSTEM WITH INTEL® XEON® 6 PRIORITY CORE TURBO (PCT) TECHNOLOGY DELIVERS 1.8X PERFORMANCE GAIN FOR LARGE-SCALE AI INFERENCE

Accelerating AI Workloads with Intelligent Core Prioritization and Turbo Boosting

SYS-GPU SuperServer SYS-822GS-NB3RT

TABLE OF CONTENTS

Executive Summary 1
Why Priority Core Turbo? 2
How Does Priority Core Turbo Work? 2
How To Enable PCT Technology on Intel Xeon 6? 3
PCT Performance on Intel Xeon 6 PCT 3
Supermicro System for Testing 6
Summary 6
For More Information 6

Executive Summary

Supermicro X14 GPU Servers powered by Intel Xeon 6 Priority Core Turbo (PCT) SKUs deliver industry-leading performance for demanding AI inference workloads by combining ultra-high CPU clock speeds with dense, high-bandwidth GPU configurations, enabling industries across sectors to gain significant competitive advantages from this platform.

Supermicro's system integration expertise enables this solution to leverage both the massive GPU compute capacity fully and the Intel PCT capabilities of the underlying hardware—delivering industry-leading performance for demanding large-

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

scale AI inference workloads by combining elevated CPU clock speeds with dense, high-bandwidth GPU configurations. Enterprises across industries gain significant competitive advantages from this tightly integrated, production-ready platform.

In typical GPU-accelerated systems, where one or two Xeon host CPUs manage four to eight GPUs per node, this architecture benefits from efficient CPU-to-GPU coordination, minimizing context switching and page migration. Host CPUs in intensive AI environments face ongoing challenges in balancing resource allocation and maintaining low-latency responsiveness. When the processor operates under full load, Priority Core Turbo (PCT) strategically prioritizes a subset of cores, allowing high-priority cores to run at boosted turbo frequencies. In contrast, non-priority cores operate at lower speeds to remain within system power and thermal limits. This targeted allocation allows latency-sensitive or serial tasks—such as GPU orchestration and critical pipeline operations—to run on the fastest cores. At the same time, background workloads execute on the remaining cores.

Why Priority Core Turbo?

PCT addresses these challenges by dynamically allocating power and frequency to the most critical cores. This capability ensures that the accelerated AI system can effectively manage the computational demands of complex AI and mixed workloads, providing a stable and efficient environment for critical tasks. Through PCT, Intel Xeon 6 processors with P-cores can achieve a competitive per-watt frequency. By optimizing resource allocation with specialized Intel Xeon 6 processor SKUs, PCT enables organizations to achieve optimal performance from their AI-accelerated systems and improve GPU utilization—all while offloading certain tasks to the CPU.

The introduction of Priority Core Turbo (PCT) in select Intel Xeon 6 processor SKUs delivers a significant boost to AI system performance. PCT enables designated high-priority cores to reach elevated peak turbo frequencies, accelerating demanding AI workloads while Intel continues to evolve CPU capabilities to meet escalating AI demands and maximize GPU utilization efficiency. By enabling priority cores to exceed standard all-core turbo limits and approach higher turbo frequency levels, PCT accelerates essential operations, reduces processing delays, and improves overall system responsiveness. As AI workloads become increasingly complex and data-intensive, this capability provides a more stable and efficient platform for maintaining low-latency performance and sustaining peak GPU utilization across modern AI infrastructure.

How Does Priority Core Turbo Work?

The default PCT setting divides cores into four partitions and designates the first core in each partition as the PCT priority core[1]. The operating system (OS) improves performance by assigning priority work to the priority cores and non-priority work (or no work) to the remaining cores. PCT allows priority cores to achieve frequencies higher than the standard all-core turbo (P0n) and up to the half-core turbo (P0half), as shown in Table 1. By enabling these cores to operate at higher performance levels, PCT can significantly reduce latency and enhance overall system responsiveness. Systems powered by CPUs that support PCT will boot with PCT capability enabled.

PCT depends on two Intel Speed Select Technologies (SST):

- SST-TF (Turbo Frequency): Determines how many CPU cores can operate at elevated turbo frequencies (high-performance cores).
- SST-CP (Core Power / CLOS): Maps CPUs to specific Classes of Service (CLOS). Only CPUs assigned to CLOS0 are recognized by PCT as high-priority cores.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

The Intel Xeon 6776P (64 cores) boosts from a 3.9 GHz max turbo to 4.6 GHz on up to 8 specific cores using Priority Core Turbo (PCT) and Speed Select Technology – Turbo Frequency (SST-TF).

Learn more about Intel Xeon 6 processors SKUs with PCT on Intel's website: Intel Xeon 6776P processor.

Table 1 - Example core frequency ranges on CPUs with PCT

Description	Name	Example frequency
All-core turbo (P0n)	F1	3.6 GHz
Low-priority frequency	F2	2.3 GHz
PCT high-priority frequency	F3	4.6 GHz
Half-core turbo (P0half)	F4	3.9 GHz
Maximum turbo frequency (P0max)	F5	4.6 GHz

PCT optimizes power and thermal management, helping systems maintain high performance under demanding AI workloads without compromising reliability. Furthermore, PCT offers flexibility to configure and optimize core performance based on workload requirements.

How To Enable PCT Technology on Intel Xeon 6?

PCT Technology divides CPU threads into N partitions (default 4), with each partition associated with a GPU. Within each partition, one or more consecutive cores are designated as high-priority (HP) cores, which run at maximum turbo frequency (P0max), while the remaining low-priority (LP) cores run at a lower frequency (~P1). The BIOS enables PCT by default (if supported), automatically partitions cores, assigns HP cores, and provides ACPI/MADT information to the OS. No extra configuration is needed if CPU-GPU mapping follows the default assumptions.

The OS/software is responsible for scheduling important tasks onto HP cores. Alternatively, users can manage PCT dynamically at runtime using the Intel Speed Select technologies tool, which allows enabling/disabling PCT and reconfiguring HP cores without rebooting. The solution supports various SKUs and core counts; if cores or PCT allocations cannot be evenly divided across partitions, the algorithm rounds down and treats surplus cores as non-PCT cores.

Please refer to the Priority Core Turbo Technology (PCT Technology) Technical Article for more details.

PCT Performance on Intel Xeon 6 PCT with GPU

Long Context Inference on Intel® Xeon® 6 PCT with HGX B300[2]

Performance results demonstrate the advantages of Intel Xeon 6 processors with PCT in AI host-node scenarios. An NVIDIA HGX platform configured with a PCT-capable Intel Xeon 6776P processor and eight NVIDIA HGX™ B300 GPUs was evaluated in a long-context inference test using the QWEN3-235B model with a 100K token context and with FP16 precision. In this test, the system with PCT enabled delivered a significant performance boost, achieving a goodput (i.e., the number of tokens successfully delivered within application service-level objectives [SLOs]) of 218 tokens/sec. This is a 1.8X performance improvement over the 121 tokens/sec achieved without PCT. The system sustained this performance while handling a request rate of 6 and meeting SLOs with a 400 ms goodput target. Intel ran these tests in early 2026.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Goodput, or the amount of compute time spent on meaningful work, is an essential metric for assessing AI cluster performance.

Overall performance gains come from the combined contributions of the CPU and GPU: the CPU prepares and orchestrates workload tasks, such as tokenization, kernel launches, data movement, and vLLM management, enabling the GPU to produce the first token faster. As the measurement window expands, goodput naturally decreases. Under tight SLOs, even small CPU-side gains can translate into disproportionately large improvements in measured goodput, especially when many requests sit just above the service-level agreement (SLA) threshold. The system is further optimized by binding GPUs to PCT cores, fully using PCT frequency, and ensuring peak performance. The results were validated across multiple runs, although additional testing was limited due to restricted system access.

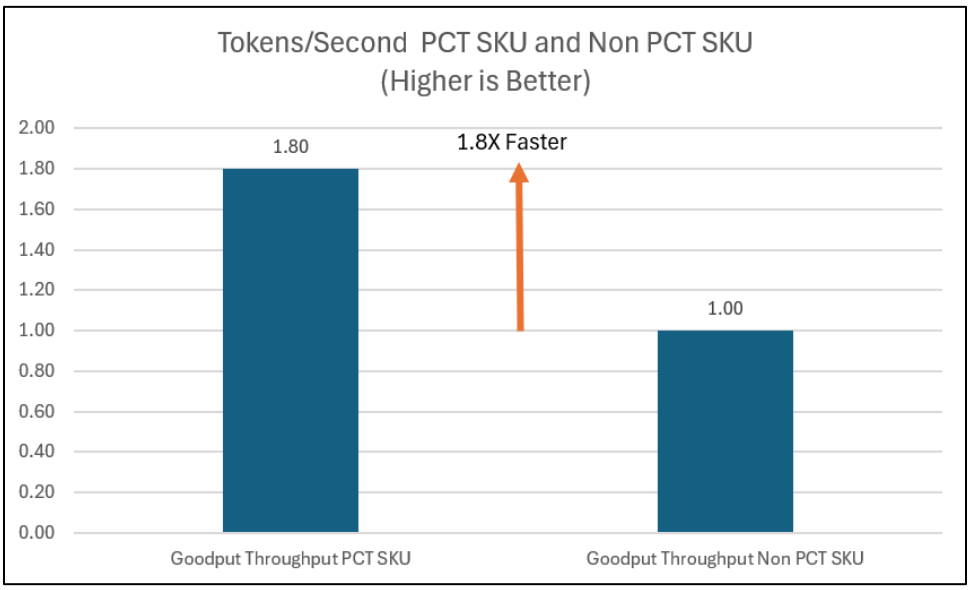


Figure 1 - Comparison of Performance (Long-context Inference) of Non-PCT vs. PCT SKUs

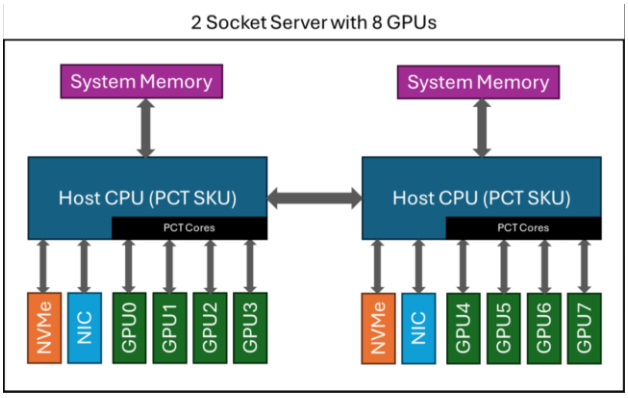


Figure 2 - Optimized configuration: GPUs are bound to PCT cores to fully leverage PCT frequency and ensure peak performance.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Checkpointing Performance Gains with PCT on Intel® Xeon® 6 and HGX B300[3]

Performance results highlight the value of Intel Xeon 6 processors with PCT for AI host-node workloads. An NVIDIA HGX platform equipped with an Intel Xeon 6776P processor and eight NVIDIA HGX B300 GPUs was evaluated using large-scale models based on the Llama 3.3 architecture. The checkpointing workflow includes state extraction, GPU-to-CPU transfer, serialization, and SSD writeback. Among these stages, serialization is the most performance-sensitive because it relies on single-threaded Python pickle processing and memory operations. At the same time, the remaining phases are primarily constrained by PCIe or storage bandwidth. We trained a synthetic Llama 3.3-140B-parameter causal language model by multiplying the hidden layers of the Llama 3.3-70B model in bfloat16 precision.

Results show that the PCT-enabled configuration consistently reduced checkpoint completion time across both standard and distributed checkpointing methods. With PCT, the Llama 3.3-70B model achieved a 5.4% reduction, while the Llama 3.3-140B[4] achieved the largest improvement at 7.2% under Distributed Checkpointing (DCP). These results demonstrate that Intel Xeon 6 with PCT technology can significantly reduce checkpoint overhead for large-scale AI training environments, improving overall system training efficiency and GPU utilization.

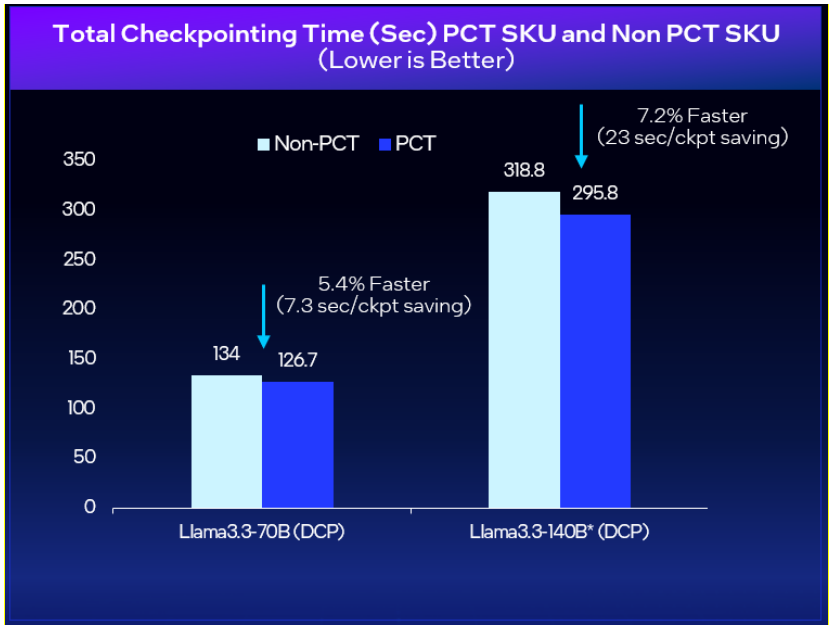
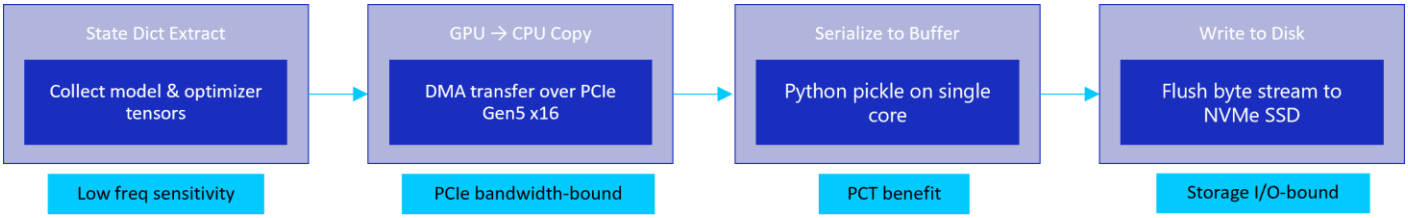


Figure 3 - Comparison of Checkpointing Performance of Non-PCT vs. PCT SKUs

Phases of Model Checkpointing (Saving model state)

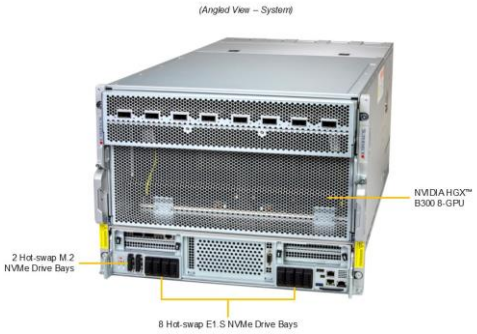


© 2026 Copyright Super Micro Computer, Inc. All rights reserved

Supermicro System for Testing:

The Supermicro SYS-822GS-NB3RT is an 8U GPU System purpose-built to integrate eight NVIDIA HGX B300 GPUs with dual Intel Xeon 6 processors into a single, factory-validated platform. Below are details regarding the system tested.

System	SYS-822GS-NB3RT
CPUs	Dual Intel Xeon 6776P processors (64 cores, 350W TDP)
Memory	3TB Total (32 x 96GB DDR5 6400 MT/s)
GPUs	NVIDIA HGX B300 8-GPU with 5th Generation NVLink® 1.8TB/s, 2.3TB of HBM3e GPU memory per system



Summary

In this work, Supermicro worked closely with the Intel engineering team to bring the latest Intel Xeon 6 CPU into a critical role in orchestration, handling large-scale tokenization, chunk scheduling, memory coordination, and data pipeline management, ensuring stable and efficient workload execution across the GPU cluster. Combined with robust accelerator support, these capabilities make Supermicro servers with Intel Xeon 6 processors with Priority Core Turbo Technology a versatile platform for modern data centers, delivering exceptional efficiency and performance for large-scale, long-context AI inference and checkpointing workloads.

For More Information:

Supermicro SYS-822GS-NB3RT: https://www.supermicro.com/en/products/system/gpu/8u/sys-822gs-nb3rt

Intel PCT Details: https://www.intel.com/content/www/us/en/content-details/846906/priority-core-turbo-technology-pct-technology-technical-article.html?DocID=846906

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

© 2026 Copyright Super Micro Computer, Inc. All rights reserved

[1] The number of core partitions is configurable in the BIOS

[2] 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, 2 x Intel Xeon 6776P processor, 64 cores, 350 W TDP, Intel HT Technology on, Intel Turbo Boost Technology on, 3,072 GB total memory (32 x 96 GB DDR5 6,400 MT/s [5,200 MT/s]), BIOS 1.4, microcode 0x10003d0, 3 x unknown NIC, 2 x Intel Ethernet Controller X710 for 10GBASE-T, 8 x ConnectX-8 family, 2 x MT43244 BlueField-3 integrated ConnectX-7 network controller, 4 x ConnextX-7 MT2910 family, 1 x 3.5 TB SAS 3808N, 8 x 7 TB HFS7T6GFMWX192N, Ubuntu 24.04.3 LTS, 6.8.0-88-generic. Software configuration: Model: Qwen3-235B-A22B-Instruct-2507, dataset: generated from a book, context length: 100K, precision: FP16, serving engine: vLLM V0.15.0, KPIs are an average of 10 runs, server command: VLLM_ALLOW_LONG_MAX_MODEL_LEN=1 vllm serve Qwen/Qwen3-235B-A22B-Instruct-2507 --host 0.0.0.0 --port 8000 --trust-remote-code --dtype half --kv-cache-dtype auto --pipeline-parallel-size 1 --tensor-parallel-size 8 --max-model-len 1010000 --enable-chunked-prefill; client command: vllm bench serve --model Qwen/Qwen3-235B-A22B-Instruct-2507 --dataset-name custom --dataset-path custom_100k.jsonl --num-prompts 32 --request-rate 6 --trust-remote-code --save-result.

[3] 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, 2 x Intel Xeon 6776P processor, 64 cores, 350 W TDP, Intel HT Technology on, Intel Turbo Boost Technology on, 3,072 GB total memory (32 x 96 GB DDR5 6,400 MT/s [5,200 MT/s]), BIOS 1.4, microcode 0x10003d0, 3 x unknown NIC, 2 x Intel Ethernet Controller X710 for 10GBASE-T, 8 x ConnectX-8 family, 2 x MT43244 BlueField-3 integrated ConnectX-7 network controller, 4 x ConnextX-7 MT2910 family, 1 x 3.5 TB SAS 3808N, 8 x 7 TB HFS7T6GFMWX192N, Ubuntu 24.04.3 LTS, 6.8.0-88-generic. Software configuration: Model: Llama 3.3 70B and Synthetic Llama 3.3-140 B parameter causal language models created by multiplying the hidden layers of the Llama 3.3-70B model. Precision: Bfloat16. MBS=1, Total Steps=100, Checkpoint frequency=every 10 steps. Numbers are the average of 10 checkpoints. Pore-condition commands: sync; sudo sysctl -w vm.drop_caches=3; sudo fstrim -v ${ckpt_dir}. The benchmark is software mapped to PCT cores using numactl. Checkpoints are saved using the DCP (distributed checkpointing) module from torch. Synthetic training samples.

[4]Synthetic Llama 3.3-140B parameter causal language models created by multiplying the hidden layers of Llama 3.3-70B mode.

© 2026 Copyright Super Micro Computer, Inc. All rights reserved