Accelerating Innovation: 5 Reasons for Choosing Intel® Xeon® 6 Processors as Host CPUs to Drive AI Success

Design an advanced AI-accelerated system capable of running demanding AI workloads by making use of Intel Xeon 6 processors as the host CPU of choice.

Why choose Intel Xeon 6 processors as host CPUs?

Intel Xeon processors are the host CPUs of choice for the world's most powerful AI accelerator platforms, being the most benchmarked host processors for these systems.[1]

The performance and utilization of AI systems rely on the CPU host to coordinate computation, manage input/output (I/O) traffic, and keep throughput running efficiently. Choosing the right CPU for your AI system can help alleviate bottlenecks and increase work time for training and inference workloads. The NVIDIA DGX Rubin NVL8 platform leverages Intel Xeon 6 processors with Intel® Priority Core Turbo (Intel® PCT) as the host CPU.

Here are five reasons why Intel Xeon 6 processors are the best choice for host CPUs in AI-accelerated systems.

1 Higher memory capacity and bandwidth

Training large generative AI (GenAI) models requires large memory capacity to store model parameters and weights before the models are sent to the GPU. As a general guideline, these accelerated AI systems should support 2x the CPU memory capacity relative to GPU memory per system, meaning that eight 288 GB GPUs would require 4.6 TB CPU memory capacity per system (8 x 288 GB). The Intel Xeon 6776P processor with 128 GB DIMMs in a 2 DIMMs per channel (2DPC) configuration can support CPU memory capacities beyond 4.6 TB, up to 8 TB per system. The 2DPC configuration provides the system memory capacity needed to train large models, and it delivers the best memory performance and total cost of ownership (TCO).[2]

In addition, Intel Xeon 6 processors with Performance-cores (P-cores) use Multiplexed Rank DIMMs (MRDIMMs) to deliver higher memory bandwidth. This innovative memory technology boosts bandwidth and performance while reducing latency for memory-bound AI and high-performance computing (HPC) workloads. High system memory bandwidth is important for large-scale inference on GPUs, especially with key-value (KV) cache sizes increasing due to larger context sizes during inference. Higher memory bandwidth is also important to support the emergence of new agentic AI inference applications, where the host CPU serves as the orchestrator of tasks.

2DPC on Intel Xeon 6 processors delivers **up to 18% higher memory speeds** compared to the latest AMD EPYC processor.[2]

MRDIMMs deliver **up to 2.3x higher memory bandwidth** compared to 5th Gen Intel Xeon processors.[3]

2 Improved single-thread performance

Accelerated AI systems require a minimum ratio of 8–12 CPU cores to 1 GPU, and Intel Xeon 6 processors deliver 2x more cores per socket than the previous generation. The processors' single-threaded core performance drives faster data transfers to GPU accelerators, increasing the time available for GPU processing and shortening model time to train, which also aids in data preprocessing—a critical function of the host CPU.

Intel Xeon 6 processors with P-cores feature select SKUs with Intel PCT technology, allowing eight priority cores to operate dynamically at a higher frequency. This enables faster data transfer of model parameters and weights to/from memory and the orchestration of tasks originating on the GPU or accelerator-based system. The remaining cores operate at a base frequency, ensuring optimal distribution of CPU resources. Intel PCT technology allows GPUs to operate at peak efficiency levels, increasing utilization for the system—a capability critical for workloads that demand sequential or serial processing—with the right volume of high-frequency cores per GPU. The right number of cores running at high frequency enables optimal TDP power.

Up to **128 P-cores per CPU** delivers 2x more cores per socket than the previous generation.

The 64-core Intel Xeon 6776P processor with Intel PCT cores and 4.6 GHz frequency delivers **up to a 17% frequency gain** compared to the Intel Xeon 6767P processor.[4]

3 Superior I/O support with the latest-generation PCIe

PCIe high performance and lane numbers determine the I/O performance of an AI system, which needs to be designed for maximum PCIe lane availability. Intel Xeon processors provide a range of 32 GB/s PCIe Gen 5 lanes for accelerator, networking, and storage workloads—from 192 lanes in a dual-socket configuration for the Intel Xeon 6900P processor to 136 lanes in an Intel Xeon 6700P one-socket configuration.

Higher I/O bandwidth accelerates data offloads and elevates operational efficiency. A higher PCIe lane count helps accommodate high-throughput GPUs, network interface controllers (NICs), and storage devices.

Boost I/O bandwidth with up to **20 percent more PCIe lanes** than the previous generation.

Intel Xeon 6 processors with P-cores can deliver **up to 192 PCIe 5.0 lanes per 2S server**, compared to only 160 lanes in a 2S configuration with the latest AMD EPYC processor.

4 Speed vector database processing with Intel® AMX instructions

Loading data to AI accelerators can bottleneck GPU utilization, as the GPUs sit idle until the data is loaded. Utilizing a vector database can improve the data-loading speed and increase overall GPU efficiency. Intel Xeon 6 processors feature Intel® Advanced Matrix Extensions (Intel AMX), an instruction set that improves vector processing.

With moderate AI compute capabilities through Intel AMX, Intel Xeon 6 processors are designed to support a wide variety of tasks as host CPUs, delivering system performance and efficiency.

Intel Xeon 6 processors with P-cores and Intel® Scalable Vector Search (Intel® SVS) optimizations enabled can **improve vector indexing and search by up to 2.75x** compared to AMD EPYC 9575F processors.[5]

Intel AMX includes **newly added support for FP16** precision arithmetic to support data preprocessing and other host CPU responsibilities in AI-accelerated systems.

5 Dedicated RAS support and confidential AI

Uptime is the key to system optimization. Intel's industry-leading reliability, availability, and serviceability (RAS) support has monitoring and control capabilities to keep systems running with optimal performance and to reduce costly downtime for accelerated AI systems. Advanced management capabilities include telemetry, platform monitoring, control over shared resources, and real-time firmware updates.

Intel and NVIDIA are actively collaborating on AI data security. Intel Xeon CPUs and NVIDIA GPUs can operate their own Trusted Execution Environments (TEEs) for confidential computing. An encrypted bounce buffer and Intel® Trust Domain Extensions (Intel® TDX) Connect provide hardware-protected connectivity between the TEEs on both the CPU and GPU, enabling end-to-end AI confidentiality. This provides a variety of benefits for AI applications, like helping prevent unauthorized data exposure, helping protect inference or training data from exposure, and strengthening cybersecurity against known or new threats.

Reduce business disruptions, improve uptime, and **deliver end-to-end AI confidentiality** with Intel Xeon 6 processors.

Learn about additional benefits that Intel Xeon 6 processors can deliver as the host CPU of choice for AI-accelerated systems:
intel.com/content/www/us/en/products/details/processors/xeon.html

See how Intel Xeon 6 processors enhance AI/HPC workloads. Examine the latest workload performance metrics:
https://edc.intel.com/content/www/us/en/products/performance/benchmarks/intel-xeon-6/

Review product specifications and find the best processor for your unique computing needs:
https://ark.intel.com/content/www/us/en/ark/products/series/595/intel-xeon-processors.html

Endnotes

[1] Based on ML Perf benchmark testing as of 2024. For details, visit https://mlcommons.org/.
[2] 8-channel 2DPC for a 2S system on Intel Xeon 6700P processor. 16 DIMMs per socket totaling 32 DIMMs. Comparing 5,200 megatransfers per second (MT/s) RDIMM speed vs. 4,400 MT/s RDIMM speed on a 5th Gen AMD EPYC processor.
[3] Based on Intel analysis as of May 2024. **Baseline:** 1-node, 2 x Intel Xeon Platinum 8592+ processors, 64 cores, Intel® Hyper-Threading Technology (Intel® HT Technology) on, Intel® Turbo Boost Technology on, NUMA configuration SNC2, 1,024 GB total memory (16 x 64 GB DDR5 5,600 MT/), BIOS version 3B07.TEL2P1, microcode 0x3000290, Ubuntu 24.04, Linux version 6.8.0-31-generic, tested by Intel as of May 2024. **New:** 1-node, pre-production platform, 2 x Intel Xeon 6 processors with P-cores, Intel HT Technology on, Intel Turbo Boost Technology on, NUMA configuration SNC3, 3,072 GB total memory (24 x 128 GB MCR 8,800 MT/s), BIOS version BHSDCRB1.IPC.0031.D97.2404192148, microcode 0x10001200, Ubuntu 23.10, kernel version 6.5.0-28-generic. **Software:** NEMO v4.2.2. ORCA025 dataset from CMCC. Intel® Fortran Compiler Classic and Intel® MPI from 2024.1; Intel® oneAPI HPC Toolkit. Compiler flags "-o4 -i8 -O3 -xCORE-AVX2 -fno-alias -fp-model fast=2 -align array64byte -fimf-use-svml=true."
[4] Tested by Intel analysis as of April 2025. **Baseline:** 1 node, 2 x Intel Xeon 6767P, 64 cores, 350 W TDP, Intel HT Technology on, Intel Turbo Boost Technology on, 1,024 GB total memory (16 x 64 GB DDR5, 8,800 MT/s [8,000 MT/s]), BIOS FIT, microcode 0x1000380, 2 x Intel® Ethernet Controller X710 for 10GBASE-T, 2 x Intel® Ethernet Controller E810-C for QSFP, 1 x 3.5 TB Intel SSDPF2KX038TZ, 2 x NVIDIA H100 NVL PCIe GPU, Ubuntu 24.04.2 LTS, 6.8.0-54-generic, Tested by Intel as of April 2025. **New:** 1 node, 2 x Intel Xeon 6776P (pre-production), 64 cores, 350 W TDP, Intel HT Technology on, Intel Turbo Boost Technology on, 1,024 GB total memory (16 x 64 GB DDR5, 8,800 MT/s [8,000 MT/s]), BIOS FIT, microcode 0x1000380, 2 x Intel Ethernet Controller X710 for 10GBASE-T, 2 x Intel Ethernet Controller E810-C for QSFP, 1 x 3.5 TB Intel SSDPF2KX038TZ, 2 x NVIDIA H100 NVL PCIe GPU, Ubuntu 24.04.2 LTS, 6.8.0-54-generic, Tested by Intel as of April 2025. **Software:** CUDA runtime version: 12080. CUDA driver version: 12080. NVIDIA driver: 570.133.20. PyPerformance: Python benchmark suite 1.11.0, https://github.com/python/pyperformance. NVbandwidth: nvbandwidth version 0.7. https://github.com/NVIDIA/nvbandwidth. TensorRT-LLM: TensorRT-LLM version 0.17.0. https://github.com/NVIDIA/TensorRT-LLM/, container: tensorrt_llm/release: 0a8915a87949. Python v.3.12. BioGPT: BioGPT ver e83601f8. Python v3.12. SVS: Similarity Search Graph-based: similarity_search on DataSets Wiki41M (45 million vectors, 1,536 dimensionality) SVS-FP16 (AMD EPYC 9575F) SVS-VectorCompression with Intel optimizations (Intel), Library: robin-map/fmtlib/ee/tsimplu/plus/pybind11/MKL/spdlog/catch2/GSL, narrow. SVS Inverted Index Construction IVF-based index construction for vector search. SVS IVF BF16, AMD AOCL 4.2.0 for AMD, OneMKL for Intel, Build times are an average of five runs, number of centroids used: 70K for Wiki-45M.
[5] See [2Q2025] intel.com/processorclaims: Intel Xeon 6. Results may vary.

Printed in USA 0326/DR/PRW/PDF Please Recycle 363270-002US