intel xeon

Accelerating Innovation: 5 Reasons for Choosing Intel® Xeon® 6 Processors 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 do you need an AI-accelerated system?

As predictive AI, generative AI (GenAI), and high-performance computing (HPC) workloads grow in complexity, their performance and energy-efficiency requirements likewise grow. One approach for achieving an optimal balance of performance and total cost of ownership (TCO) for these workloads is to design an AI-accelerated system using a **host CPU** and **discrete AI accelerators**.

In an AI-accelerated system, the host CPU optimizes processing performance and resource utilization by delivering efficient task management and high-performance preprocessing—two factors critical for ensuring that model training pipelines stay well fed and tha discrete AI processors are kept running at optimal utilization levels.

Intel Xeon 6 processors with Performance-cores (P-cores) are ideal host CPUs. Serving as the brain of an AI-accelerated system, the hos CPU performs a wide variety of management, optimization, preprocessing, processing, and offloading tasks to facilitate system performance and efficiency.

GPUs and Intel® Gaudi® AI accelerators provide a system's high-powered muscles. These discrete AI accelerators dedicate their parallel-processing capabilities to large language model (LLM) training for GenAI and to model training for predictive AI.

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]

Here are five more reasons to choose Intel Xeon 6 processors as your host CPUs for AI-accelerated systems.

1 Superior I/O performance

Higher input/output (I/O) bandwidth accelerates data offloads and elevates operational efficiency.

Boost I/O bandwidth with up to **20 percent more PCIe lanes** than the previous generation (up to 192 PCIe 5.0 lanes per 2S system).

2 Higher core counts and single-threaded performance

Higher CPU core counts and single-threaded performance translate into faster data feeds for GPUs/accelerators, which helps shorten models' time-to-train. High max turbo processor frequencies boost single-threaded CPU performance.

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

3 Higher memory bandwidth and capacity

High memory capacities and performance are critical requirements for AI systems. Intel Xeon 6 processors with P-cores provide higher memory speeds with 2 DIMMs per channel (2DPC) to deliver the best memory performance and TCO compared to the competition.[2] Additionally, Intel Xeon 6 processors with P-cores can deliver even higher memory bandwidth with Multiplexed Rank DIMMs (MRDIMMs). This innovative memory technology boosts bandwidth and performance while reducing latency for memory-bound AI and HPC workloads, and it is not currently supported on AMD EPYC processors.

Intel Xeon 6 processors feature up to 504 MB L3 cache, combined with support from Compute Express Link (CXL). CXL maintains memory coherency between the CPU memory space and memory on attached devices, enabling high-performance resource sharing, reduced software stack complexity, and lower overall system cost.

2DPC on Intel Xeon 6 processors delivers **up to 30% higher memory speeds** compared to the latest AMD EPYC processor.[2]

MRDIMMs deliver **up to 2.3x higher memory bandwidth** compared to the previous generation.[3]

4 Dedicated RAS support

Intel's industry-leading reliability, availability, and serviceability (RAS) support reduces costly downtime for large AI/HPC systems. Advanced management capabilities include telemetry, platform monitoring, control over shared resources, and real-time firmware updates. RAS benefits from the collective expertise of platform partners, ISVs, and solution integrators.

Minimize business disruptions with Intel Xeon 6 processors, **built to maximize uptime** and operational efficiency.

5 Flexibility for mixed workloads

Intel Xeon 6 processors are designed to support a wide variety of workloads as host CPUs, delivering both performance and efficiency. In some cases, host CPUs in AI systems might need to support limited AI functionality during the data preprocessing phase.

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

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.

intel xeon

[1] Based on MLPerf benchmark testing as of 2024. For details, visit https://mlcommons.org/.
[2] 2DPC on an Intel Xeon 6700P processor = 5,200 MT/s RDIMM speed; 2DPC on a the latest AMD EPYC processor = 4,000 MT/s RDIMM speed.
[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 megatransfers per second [MT/s]), BIOS version 3B07.TEL2P1, microcode 0x21000200, 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 0x81000200, 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 "-i4 -r8 -O3 -xCORE-AVX2 -fno-alias -fp-model fast=2 -align array64byte -fimf-use-svml=true."

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for additional details.
No product or component can be absolutely secure.
Your costs and results may vary.
Intel technologies may require enabled hardware, software or service activation.

Printed in USA 0425/DR/PRW/PDF Please Recycle 363270-002US