

Artificial Intelligence (AI)

FPGA community forums and blogs on community.intel.com are migrating to the new Altera Community and are read-only. For urgent support needs during this transition, please visit the FPGA Design Resources page or contact an Altera Authorized Distributor.

Intel Community / Blogs / Tech Innovation / Artificial Intelligence (AI)

798 Discussions

Intel® Xeon® Processors: The Most Preferred CPU for AI Host Nodes

Subscribe

Article Options



Divakar Employee
08-26-2025

16 0 6,516

Author: Divakar Mariyanna, Cloud Systems and Solutions Engineer, Intel

In the rapidly evolving world of artificial intelligence (AI), the parallel processing capabilities of graphics processing units (GPUs) are essential in training and deploying large language models (LLMs). However, today's AI workloads are not purely offloaded to GPU accelerators. Host central processing units (CPUs) such as the Intel® Xeon® 6 processor with Performance Cores (P-cores) play a significant role in maximizing the performance of AI-accelerated systems.

Host CPUs are essential for overall system management. As the system's control center, host CPUs orchestrate tasks or agents across GPUs by preprocessing data for training models, transmitting data to the GPU for parallel processing, managing checkpointing to system memory, and processing mixed workloads running on the same infrastructure. As the most deployed host processors in the market for AI accelerator platforms, Intel Xeon processors are the host CPUs of choice for the world's most powerful AI systems.⁽¹⁾

Comparative Analysis with AMD EPYC Host CPUs

In a comparative analysis of Intel Xeon 8592V with AMD EPYC 9575F host CPUs, the Intel Xeon host processor delivered on-par performance in latency-constrained throughput compared to the AMD EPYC processor-hosted system. As agentic AI workloads call for stricter performance requirements on serving LLMs, host CPUs can balance latency-constrained throughput (or goodput) and latency performance. The higher the goodput for time-to-first-token (TTFT), the better the performance and responsiveness of the AI system.

We tested latency-constrained inference serving performance on an 8x GPU system with a 2-socket CPU, using the virtual LLM (vLLM) inference runtime to serve Llama-3.3-70B in a TP8 configuration. With the Intel Xeon 8592V as host node using the v0.9.1 June 2025 release of vLLM V1 engine, the request-rate was set to 512, and throughput was plotted against TTFT constraints (50ms, 100ms, 200ms, 300ms, 400ms, 500ms, and 600ms). In Figure 1 below, we graphed our results against [AMD's analysis](#) of the AMD EPYC 9575F host CPU.

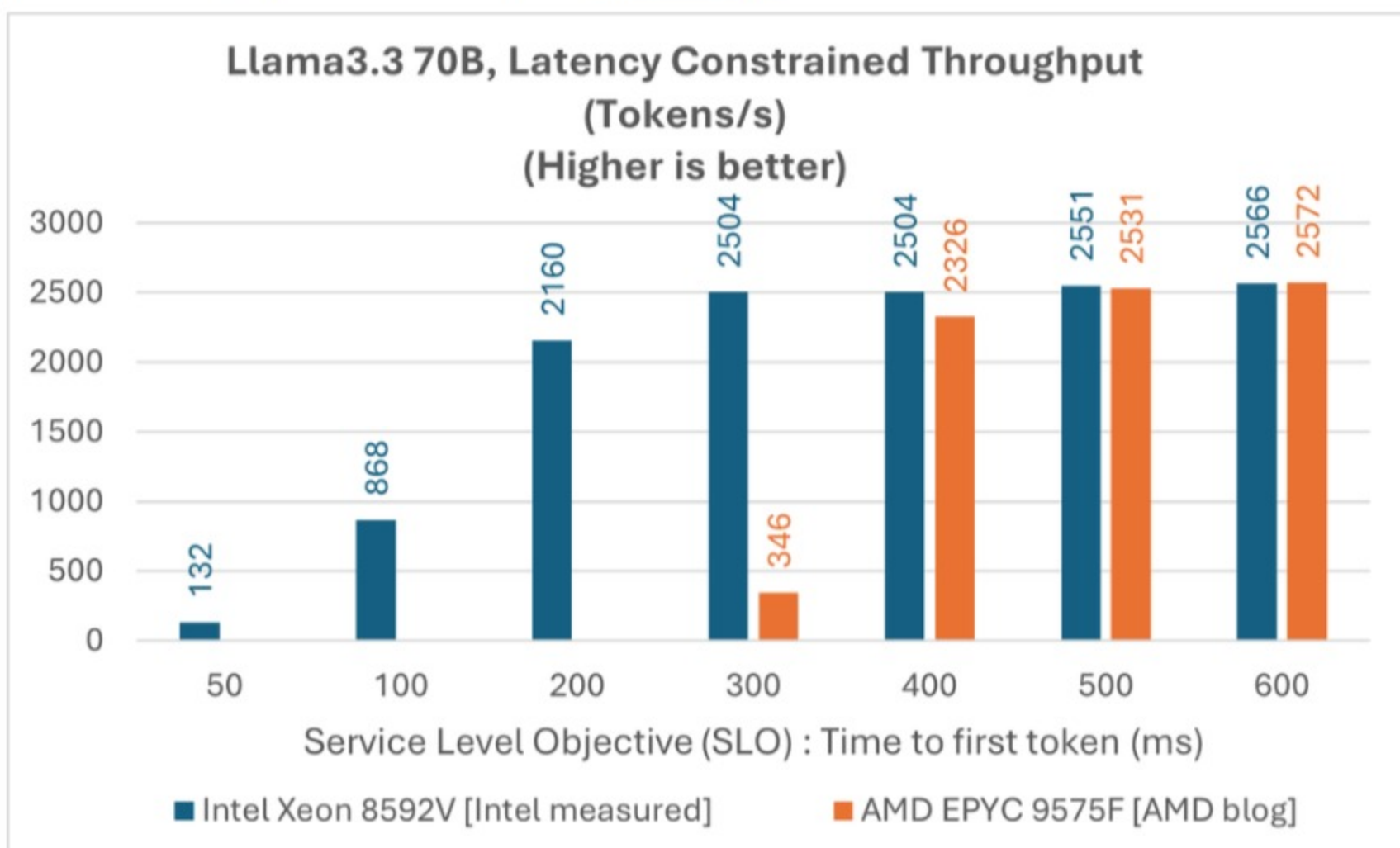


Figure 1: Latency-constrained throughput versus TTFT latency constraint. AMD EPYC 9575F figures are based on AMD Analysis.

The Intel Xeon 8592V host node met all service level objective (SLO) TTFT limitations, starting as low as a 50ms constraint. The Intel Xeon 8592V host node can enable GPUs to perform at greater stringent constraints. In addition, the Intel Xeon 8592V based host node scales up to 2000+ tokens/second at the 200ms constraint and stabilizes at 2500+ tokens/second at constraint of 300ms and beyond. The Intel Xeon 8592V platform shows solid performance benefits when used as a host CPU for a GPU system.

In reviewing a similar comparison analysis conducted by AMD using the Intel Xeon 8592+, we were unable to replicate numbers published on vLLM goodput performance. The vLLM V1 engine has multiple GitHub releases (0.8.0 to 0.9.2) and it is not clear which vLLM V1 release was used. AMD blog also has a reference to MAX_NUM_REQS parameter, as the vLLM benchmarking script doesn't recognize this parameter, any reference to rate control is unclear. In addition, AMD did not publish goodput numbers under 300ms.

Choosing the Right Host CPU Based on Performance

Choosing the right host CPU can help alleviate bottlenecks and increase work time for training and inference workloads. Host CPUs for GPU-accelerated systems need to perform efficiently with high throughput and low latency to ensure that the entire GPU-accelerated system operates in a balanced and optimal way, unlocking greater GPU performance and driving better cost-effectiveness.

Look for the following performance features when selecting a host CPU:

- High memory capacity and bandwidth:** A 2 DIMMs per channel (2DPC) configuration can support CPU memory capacities beyond 4.6 TB, up to 8 TB per system, perfect for training large models. 2DPC on Intel Xeon 6 processors delivers up to 18% higher memory speeds and bandwidth compared to the latest AMD EPYC processor.⁽²⁾
- Superior I/O support with the latest-generation PCIe:** Intel Xeon 6 processors with P-cores can deliver up to 192 PCIe 5.0 lanes per 25 server, compared to only 160 lanes in a 25 configuration with the latest AMD EPYC processor. Higher I/O bandwidth accelerates data offloads and elevates operational efficiency.
- Improved single-thread performance:** The Intel Xeon 6 processor's single-threaded core performance drives fast data transfers to GPU accelerators, increasing the time available for GPU processing and shortening model time to train. With up to 128 P-cores per CPU, Intel Xeon 6 processors deliver 2x more cores per socket than the previous generation.
- Dynamic prioritization of high-priority cores:** Intel Xeon 6 processors with P-cores feature select SKUs with Intel® Priority Core Turbo (Intel® PCT), allowing eight priority cores to operate dynamically at higher turbo frequencies for peak GPU efficiency. In parallel, lower priority cores operate at base frequency, ensuring optimal distribution of CPU resources.
- Improved speed with vector database processing:** 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 2.6x compared to AMD EPYC 9575F processors.⁽³⁾
- Dedicated RAS support:** Intel's industry-leading reliability, availability, and serviceability (RAS) support has monitoring and control capabilities to keep systems running at optimal performance and reduce costly downtime.

A Trusted Partner in AI Accelerated Platforms

Customers have choices when evaluating CPUs, and the vast majority are choosing Intel Xeon processors as the host CPU for their accelerated systems. Most notably, Nvidia, the AI infrastructure leader, chose the Intel Xeon 6776P processor as the host for their latest [DGX B300 system](#). Other cloud solution providers (CSP) and original equipment manufacturer (OEM) partners have adopted this Intel Xeon host CPU and Nvidia GPU system configuration, including Google Cloud, AWS, Microsoft, CoreWeave, Dell, HPE, and Lenovo.

As enterprises modernize their infrastructure to handle the increasing demands of AI, Intel Xeon 6 processors provide the ideal combination of performance and energy efficiency. These processors support a wide range of data center and network applications, solidifying Intel's position as the leader in AI-optimized CPU solutions.

Learn how [Intel Xeon 6 processors](#) can meet your diverse power, performance, and efficiency requirements.

Product and Performance Information

Software Configuration

- Model:** Llama3.3 70B Instruct⁽⁴⁾
- Data Set:** Sonnet3.5-SlimOrcaDedupCleaned⁽⁵⁾
- Server Engine:** vLLM V1⁽⁶⁾ PKIs are an average of 10 runs
- Server Command:** vllm serve \${model_path} --dtype half --kv-cache-dtype auto --pp 1 -tp 8
- Client Command:** python /workdir/vllm/benchmarks/benchmark_serving.py --model \${model_path} --tokenizer \${model_path} --dataset-name sharegpt --dataset-path \${dataset_path} --trust-remote-code --save-result --num-prompts 512 --request-rate 512 --goodput tftf:\${slo_tftf} --percentile-metrics tftf,pot,ttl,e2el

Hardware Configuration

1-node, NVIDIA DGXH100 with 8x H100 SXM 80GB HBM3, 2x INTEL® XEON® PLATINUM 8592V, 64 cores, 330W TDP, HT Off, Turbo On, Total Memory 1024GB (16x64GB DDR5 5600MT/s [4800MT/s]), BIOS 1.5.5, microcode 0x210002a9, 1x Ethernet Controller 10G X550T, 12x MT2910 Family [ConnectX-7], 2x 1.7T SAMSUNG M21L21T9HCLS-00A07, 8x 3.5T KCM6DRUL3T84, Ubuntu 22.04.5 LTS, 5.15.0-1081-nvidia.

Test by Intel as of July 2025. Your results may vary. Intel technologies may require enabled hardware, software, or service activation.⁽⁷⁾

Endnotes

- IDC Server Tracker report, based on Q1'24 system volume.
- 8-channel 2DPC for a 25 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.
- See [7D220] [intel.com/processorclaims](#): Intel Xeon 6. Results may vary.
- <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>
- <https://huggingface.co/datasets/Gryphe/Sonnet3.5-SlimOrcaDedupCleaned>
- <https://github.com/vllm-project/vllm/tree/v0.9.1>
- Performance varies by use, configuration, and other factors. Learn more at [intel.com/performanceindex](#).

Notices and Disclaimers

Performance varies by use, configuration, and other factors. Learn more on the [Performance Index site](#). Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration 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. © Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

16 Kudos

You must be a registered user to add a comment. If you've already registered, sign in. Otherwise, register and sign in.

Comment

Community support is provided Monday to Friday. Other contact methods are available [here](#).

Intel does not verify all solutions, including but not limited to any file transfers that may appear in this community. Accordingly, Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

For more complete information about compiler optimizations, see our [Optimization Notice](#).