

# Artificial Intelligence (AI)

FPGA community forums and blogs on [community.intel.com](https://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 6® Processors: Delivering High Throughput and Low Latency with Data Center LLMs

Subscribe

Article Options



Intel AI Employee

09-22-2025

2 0 2,097

### Authors:

Chris Liebert, System and Software Optimization Engineer, Intel  
Padma Apparao, Senior Principal Engineer, AI Solutions, Intel

Large language models (LLMs) are in high demand in data centers, powering everything from chatbots to content creation to language translation tools. However, artificial intelligence (AI) workloads are computational and memory bandwidth-intensive, requiring significant processing power to handle complex calculations and large amounts of data. Low-latency hardware solutions are essential today to ensure fast inferences, real-time interactions, and scalability. Graphics processing units (GPUs) are just one option for running LLMs. The industry is exploring new central processing unit (CPU) approaches, from custom AI accelerators to advanced memory technologies designed to reduce latency. [Intel® Xeon® 6 processors](#) offer a compelling solution, leveraging integrated technologies such as Intel® Advanced Matrix Extensions ([Intel® AMX](#)) and high-bandwidth memory bandwidth multiplexed rank DIMMs ([MRDIMMs](#)).

Hardware is only one piece of the equation. Intel brings decades of experience working with software developers, contributing to frameworks like PyTorch and building a strong ecosystem full of opportunities to optimize software for Xeon processors. This close relationship is deepened by Intel's proven commitment to providing detailed documentation and optimization guides through joint efforts with major [contributions to PyTorch](#). AMD has also made contributions with the ZenDNN Zentorch library. Let's look closer to see how these optimized solutions for data centers stack up.

### Comparative Analysis of Intel Xeon 6980P Versus AMD EPYC 9755

A thorough and standardized benchmarking methodology is crucial to assess the competitive performance of LLMs. Throughput, a key performance metric, measures the rate at which a model can generate responses to requests within a given timeframe. Concurrency measures the number of simultaneous requests being processed. To create better user experiences, using a service level agreement (SLA) such as time per output token (TPOT) of less than 100 milliseconds (ms) provides a clear, quantifiable, and comparable measurement. When it comes to benchmarking LLM throughput, TPOT is critical as it measures how fast the user receives a response. Throughput can be improved by increasing the batch size; however, doing so also increases request latency. By increasing the batch size until TPOT has exceeded 100ms, we can tune each experiment to measure optimal throughput.

The following comparisons show latency-constrained throughput held to an SLA of 100ms (TPOT P90) on the 128-core Intel Xeon 6980P versus the 128-core AMD EPYC 9755.<sup>(1)</sup> We use Meta AI's Llama family and EleutherAI's GPT-J LLMs. These open-source models are quickly improving performance compared to proprietary models, making them especially valuable for researchers and developers who want flexibility and transparency.

For the first test, we simulated LLM abstract title generation. This scenario involves using the LLM to generate titles or grab headlines for articles, documents, and other content. Figure 1 shows that Intel Xeon 6980P delivered higher performance up to 2.8x when compared with AMD EPYC 9755 and supported a greater number of concurrent requests.

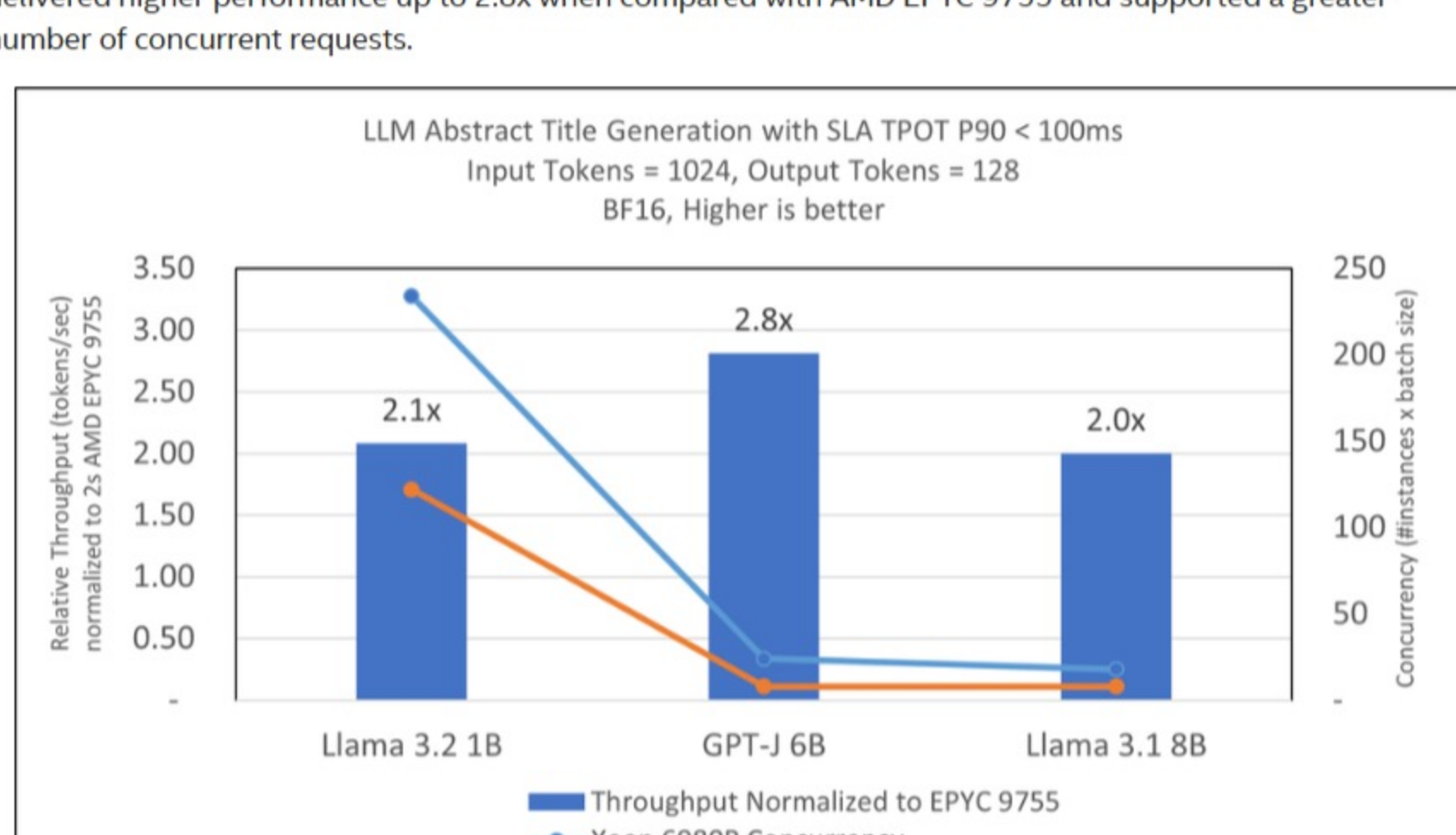


Figure 1: Simulation of LLM generating abstract titles.

For the second test, we simulated a chatbot. An example of this scenario is simulating human conversation through LLMs, understanding context, and engaging with users in real-time. Chatbots are commonly used in customer service, virtual assistance, and other business applications to enhance user experience and streamline operations. Figure 2 shows Intel Xeon 6980P delivered higher performance up to 2.6x when compared with AMD EPYC 9755, supporting a greater number of concurrent requests.

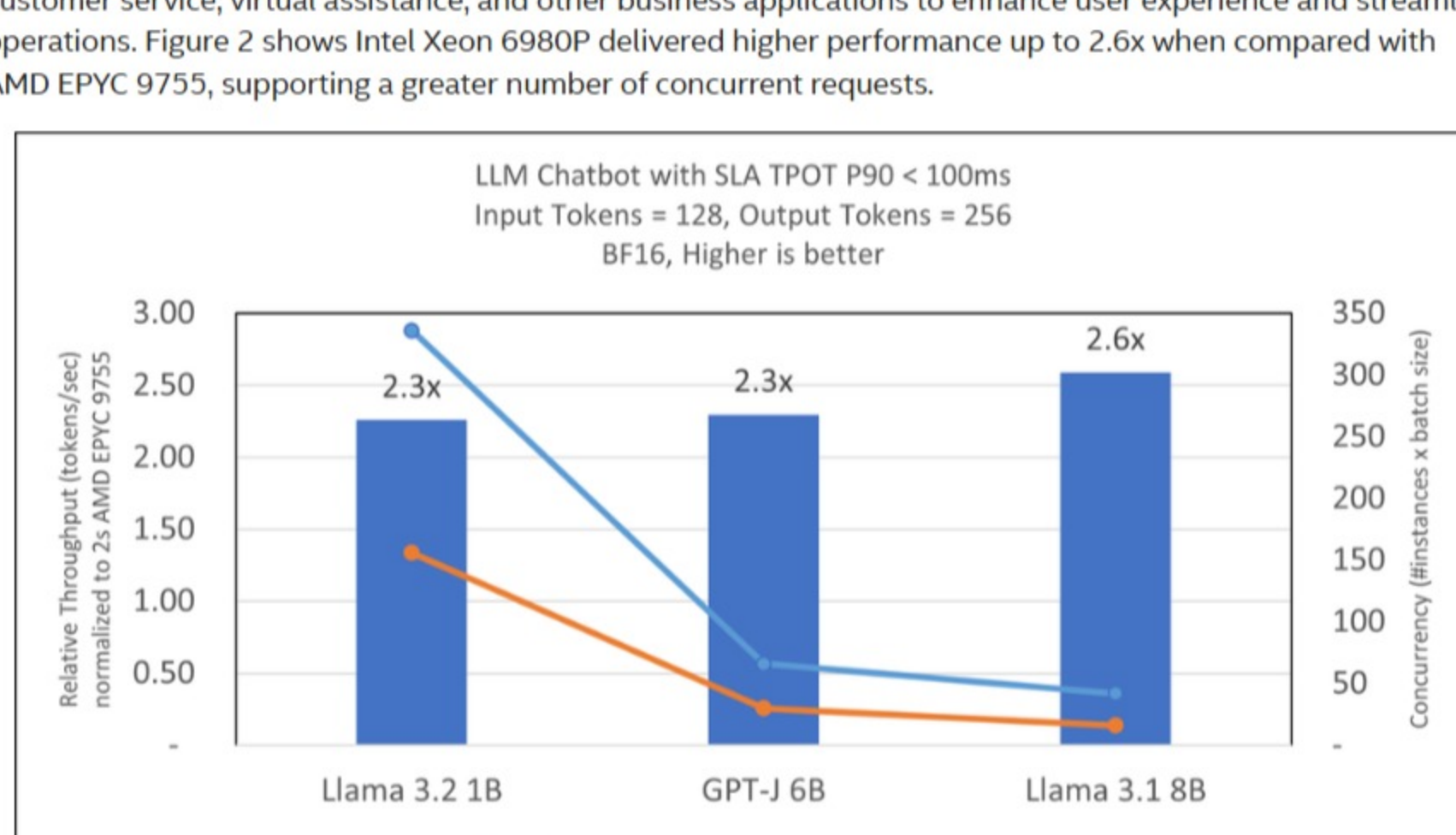


Figure 2: Simulation of LLM chatbot functions

For the last test, we simulated LLM translation. An example of this scenario is using LLMs to convert text from one language to another while preserving context and idiomatic expressions across the translation. Figure 3 shows Intel Xeon 6980P delivered up to 2.4x higher performance compared with AMD EPYC 9755, and supporting a greater number of concurrent requests.

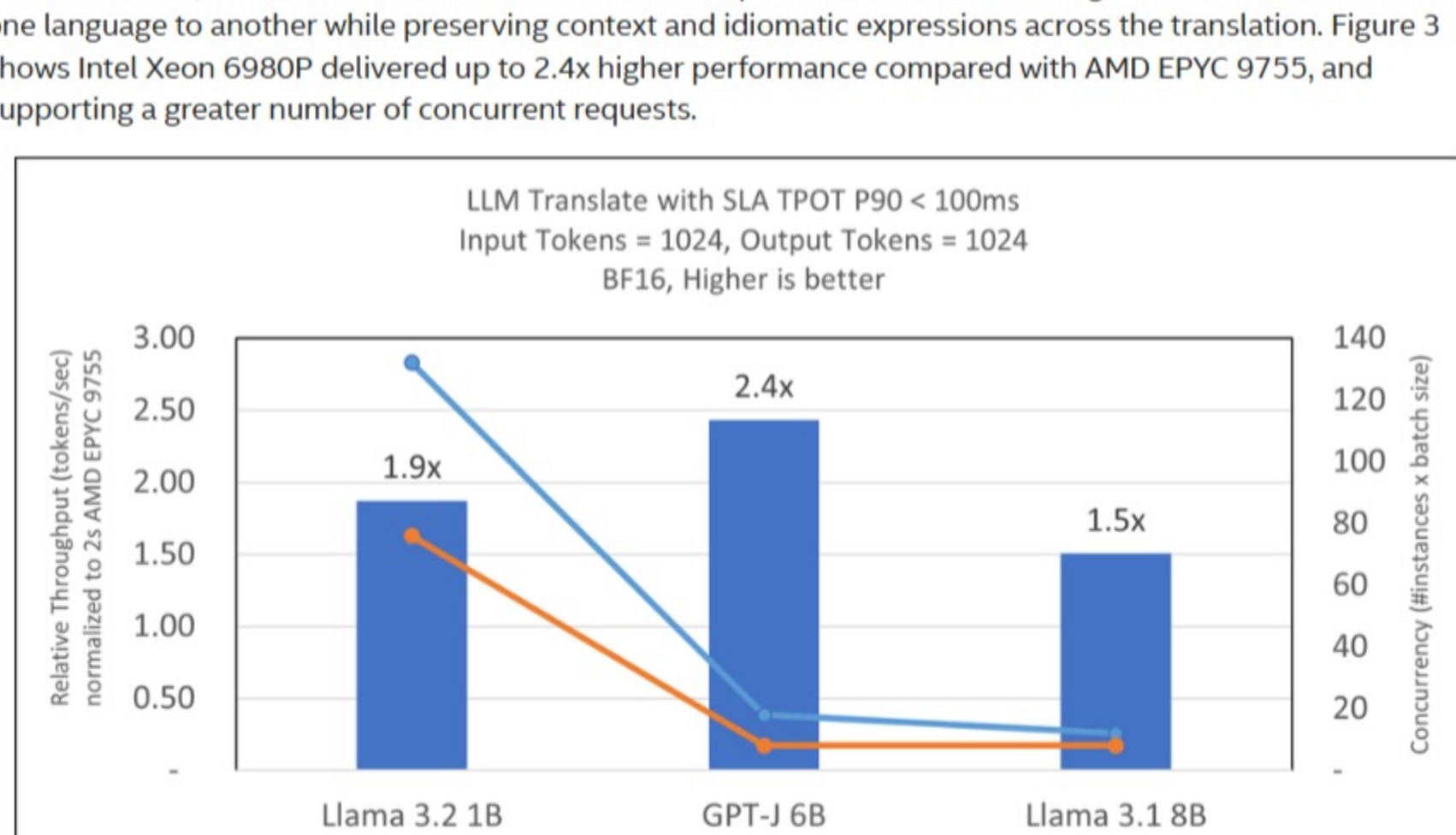


Figure 3: Simulation of LLM translation

### Performance of the 192-core AMD EPYC 9965 "Turin Dense"

The higher-core-count 192-core AMD EPYC 9965 did not produce comparable throughput because the lowest measured batch sizes exceeded the 100ms TPOT SLA, making it less suitable for scenarios requiring low latency and high scalability.

### Conclusion

Overall, the Intel Xeon 6980P 128-core CPU delivers up to 2.8x higher performance and up to 2.2x higher concurrency across various tested use cases compared to the AMD EPYC 9755 (128 core). For real-time deployments, the latest generation Intel Xeon 6 CPUs can provide significantly higher LLM performance and support a larger number of concurrent requests.

### Accelerate AI Workloads with Built-In Intel AMX

To help make your deep learning workloads more efficient, cost-effective, and easier to train and deploy, Intel AMX on Intel Xeon 6 processors delivers acceleration for inferencing and training while minimizing the need for specialized hardware. Make the most of your CPU to power AI training and inferencing workloads at scale with the following benefits:

- Improved performance:** CPU-based acceleration can improve power and resource utilization efficiencies, giving you better performance for the same price.
- Reduced total cost of ownership (TCO):** As an integrated accelerator on Intel Xeon 6 processors that you may already own, Intel AMX enables you to maximize your investments and get more from your CPU.
- Reduced development time:** Intel works closely with the open-source community, including TensorFlow and PyTorch projects, to optimize frameworks for Intel hardware. This enables you to take advantage of the performance benefits of Intel AMX by adding a few lines of code, reducing overall development time.

In addition, Intel Xeon 6 processors use MRDIMMs to deliver higher memory bandwidth. Not supported on 5th Gen AMD EPYC processors, this innovative memory technology boosts bandwidth and performance while reducing latency for memory-bound AI and high-performance computing (HPC) workloads.

Learn how [Intel Xeon 6 processors](#) with integrated Intel AMX accelerators and MRDIMMs memory can help you to meet the demands of today's data centers.

### Product and Performance Information

#### Configurations

The software configuration for this environment as of 8/29/2025 is as follows: meta-llama/Llama-3.2-1B, EleutherAI/gpt-j-6b, meta-llama/Llama-3.1-8B-Instruct. AMD ZenDNN 5.1 Python 3.11: PyTorch 2.7.0, Intel IPEX 2.5.0 Python 3.10: PyTorch 2.5.0.

**AMD EPYC 9755:** 1-node, 2x AMD EPYC 9755 128-Core Processor, 128 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400MT/s [6000MT/s]), microcode 0xb002116, Ubuntu 24.04.1 LTS, 6.8.0-47-generic. Using physical cores only.

**AMD EPYC 9965:** 1-node, 2x AMD EPYC 9965 192-Core Processor, 192 cores, 500W TDP, SMT On, Boost On, Total Memory 1536GB (24x64GB DDR5 6400 MT/s [6000 MT/s]), microcode 0xb101021, Ubuntu 24.04 LTS, 6.8.0-47-generic. Using physical cores only. 192-core AMD EPYC 9965 did not produce comparable throughput because the lowest measured batch sizes exceeded the 100ms TPOT SLA, making it less suitable for scenarios requiring low latency and high scalability.

**Intel Xeon 6980P:** 1-node, 2x Intel(R) Xeon(R) 6980P, 128 cores, 500W TDP, HT On, Turbo On, Total Memory 1536GB (24x64GB DDR5 8800MT/s [8800MT/s]), microcode 0x10003a5, Ubuntu 24.04 LTS, 6.8.0-47-generic. Using physical cores only.

Testing conducted by Intel on 8/29/2025. Your results may vary. Intel technologies may require enabled hardware, software, or service activation.<sup>(2)</sup>

#### Endnotes

- Compared to the 128-core Intel Xeon 6980P, the higher 192-core AMD EPYC 9965 did not produce comparable throughput because the lowest measured batch sizes exceeded the 100ms TPOT SLA, making it less suitable for scenarios that require both low latency and high scalability.
- Performance varies by use, configuration, and other factors. Learn more at [intel.com/performanceindex](https://www.intel.com/performanceindex).

#### Notices and Disclaimers

Performance results 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.

#### Translate

2 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](#).