LEADERSHIP NATURAL LANGUAGE AI PERFORMANCE
OUTPERFORMING 5TH GEN INTEL® XEON® WITH AMX

Powered by 4th Gen AMD EPYC™ Processors

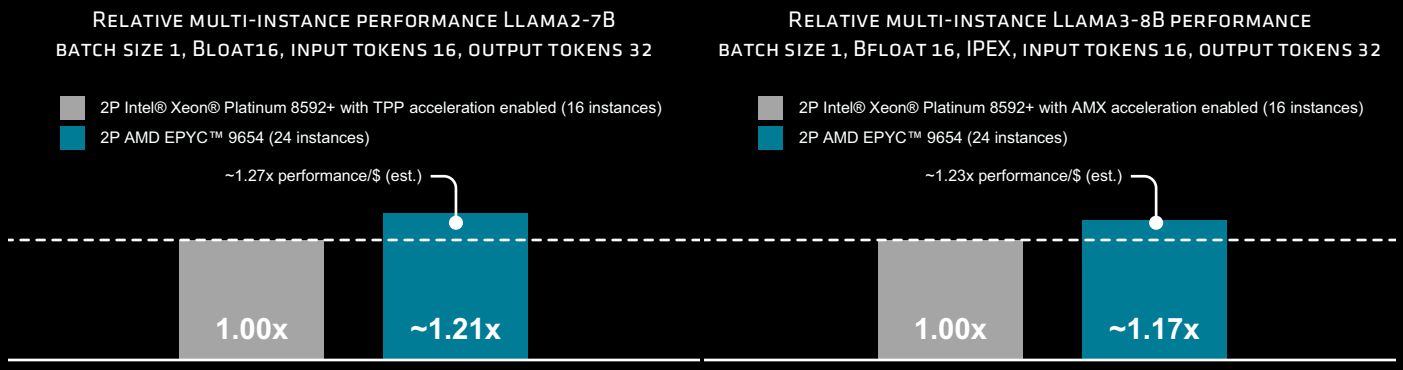
August 2024

AT A GLANCE

Systems powered by 4th Gen AMD EPYC™ processors deliver strong performance uplifts versus 5th Gen Intel® Xeon® Platinum 8592+ systems on a variety of AI workloads while running 50% more concurrent instances on multi-instance workloads.

PERFORMANCE HIGHLIGHTS

These charts highlight the relative single-instance Llama2-7B and Llama3-8B performance and price-performance. The 4th Gen AMD EPYC CPU-based system delivers performance advantages across both models with up to ~1.27x the performance per dollar.



KEY TAKEAWAYS

Llama2 is a popular Large Language Model (LLM) that represents a broad spectrum of AI use cases. Intel markets the AI capabilities of AMX in select Intel Xeon processors; however, the 2P AMD EPYC 9654 system delivers ~1.21x multi-instance Llama2-7B and ~1.17x multi-instance Llama3-8B uplifts vs. a 2P Intel Xeon Platinum 8592+ system with AMX acceleration.

- The 2P AMD EPYC 9654 system outperforms the 2P Intel Xeon Platinum 8592+ system for these use cases.
- The 2P AMD EPYC 9654 system also demonstrates superior cost effectiveness of ~1.27x the performance/est. $ in the Llama2-7B workload and ~1.23x the performance/est. $ in the Llama3-8B workload.
- These workloads represent smaller language model inference deployments and demonstrate that AMD EPYC processors are a superior and cost effective solution for small AI deployments.

IN THIS BRIEF

AMD EPYC 9004 SERIES PROCESSORS

General purpose AMD EPYC 9004 Series Processors continue to redefine the standards for modern data centers. 4th Gen AMD EPYC processors are built on the innovative x86 architecture and "Zen 4" core. 4th Gen AMD EPYC processors deliver efficient, optimized performance by combining high frequencies, the largest-available L3 cache, 128 lanes of PCIe® 5 I/O (1P), and synchronized fabric and memory clock speeds, plus support for up to 6 TB of DDR5-4800 memory. Built-in security features, such as Secure Memory Encryption (SME) and Secure Encrypted Virtualization (SEV-SNP), collectively known as AMD Infinity Guard, help protect data while it is in use.[1] Servers powered by 4th Gen AMD EPYC processors are ideal for inferencing and small AI model development, testing, and batch training.

General purpose 4th Gen AMD EPYC 9004 processors are available in 1P and 2P configurations and feature:

- AVX-512 instruction support for enhanced HPC and ML performance.
- Up to 96 cores (192 threads) per processor.
- Up to 384 MB L3 cache.
- Up to 4 links of Gen 3 Infinity Fabric™ at up to 32 Gbps.
- 12 memory channels per socket that support up to 6 TB of DDR5-4800 memory.
- Support for 128 (1P) and up to 160 (2P) PCIe® Gen 5 with up to 32 Gbps bandwidth.
- AMD Infinity Guard technology to defend your data.

LLAMA2-7B SINGLE-INSTANCE (PER-SOCKET) THROUGHPUT

Figures 1 and 2 shows the relative throughput of a 2P AMD EPYC 9654 system versus a 2P Intel Xeon Platinum 8592+ system running one Llama2-7B instance per socket with a batch size of 1 and 16 input/32 output tokens. For this test:

- The 2P Intel Xeon Platinum system ran 1 instance per socket using 64 cores per instance.
- The 2P AMD EPYC 9654 system ran 1 instance per socket using 96 cores per instance and delivered ~1.13x the performance and ~1.19x the performance/est. $ of the Intel system.
- These tests used TPP v0.0.1.
- Results include Time To First Token (TTFT) times.

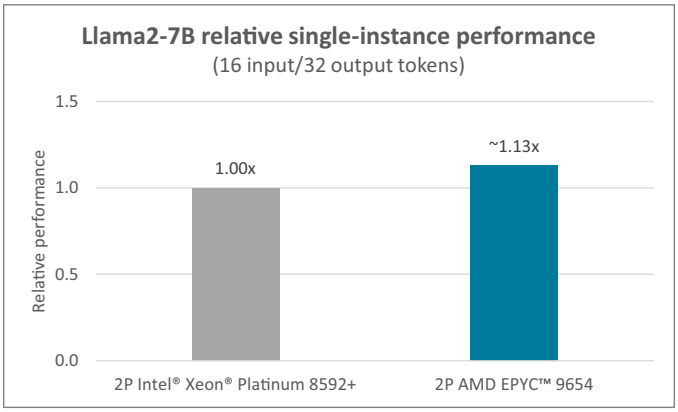


Figure 1: Relative single-instance Llama2-7B performance (upper right)

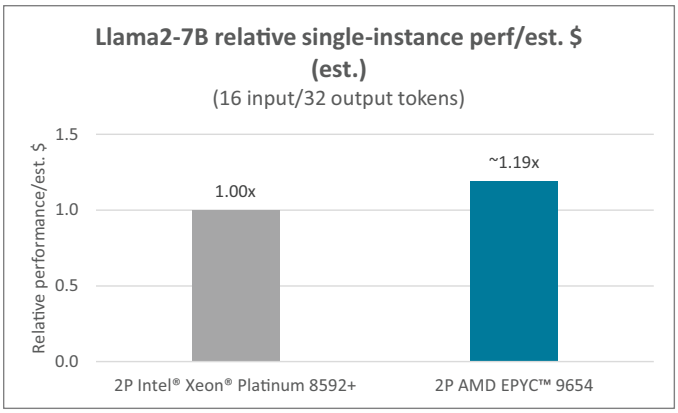


Figure 2: Relative single-instance Llama2-7B perf/est. $ (lower right)

LLAMA3-8B SINGLE-INSTANCE (PER-SOCKET) THROUGHPUT

Figures 3 and 4 show the relative throughput of a 2P AMD EPYC 9654 system versus a 2P Intel Xeon Platinum 8592+ system. The comparisons include a single Llama3-8B instance per socket with a batch size of 1 and 16 input/32 output tokens. For these tests:

- The 2P Intel Xeon Platinum system ran one instance per socket using 64 cores per socket.
- The 2P AMD EPYC 9654 system ran one instance per socket using 96 cores per socket and delivered ~1.11x the performance and ~1.16x the performance/est. $ of the Intel system.
- These tests used IPEX v2.3.0.
- Results include Time To First Token (TTFT) times.

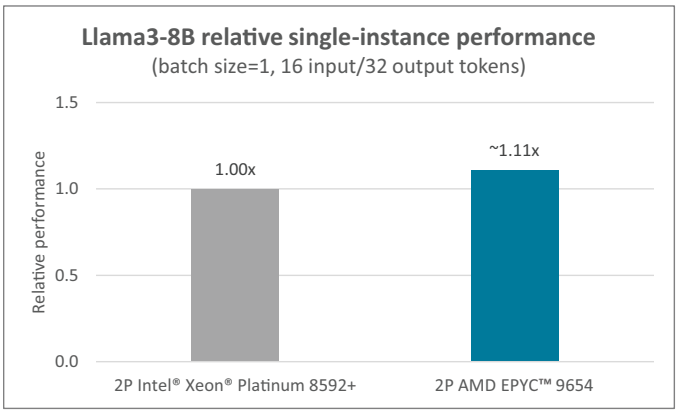


Figure 3: Relative single-instance Llama3-8B performance (upper right)

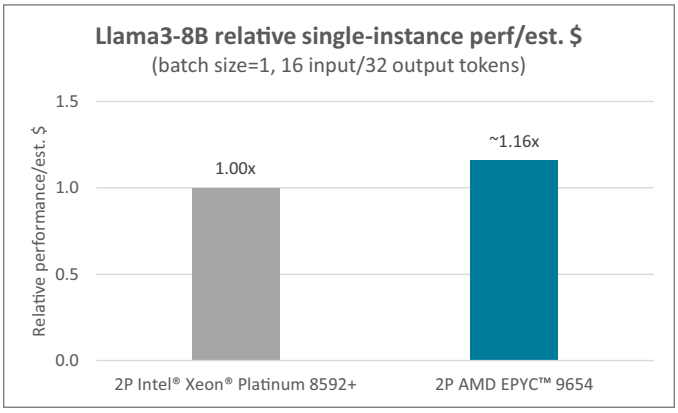


Figure 4: Relative single-instance Llama3-8B perf/est. $ (lower right)

LLAMA2-7B MULTI-INSTANCE THROUGHPUT

Figures 5 and 6 show the relative throughput of the 2P AMD EPYC 9654 system versus the 2P Intel Xeon Platinum 8592+ system running multiple concurrent Llama2-7B instances with a batch size of 1 and 16 input/32 output tokens. For these tests:

- The 2P Intel Xeon Platinum system ran 16 instances using 8 cores per instance.
- The 2P AMD EPYC 9654 system ran 24 instances using 8 cores per instance and delivered ~1.21x the performance and ~1.27x the performance/est. $ of the Intel system while running 50% more concurrent instances than the Intel Xeon Platinum 8592+ system.
- These tests used TTP 0.0.1.
- Results include Time To First Token (TTFT) times.

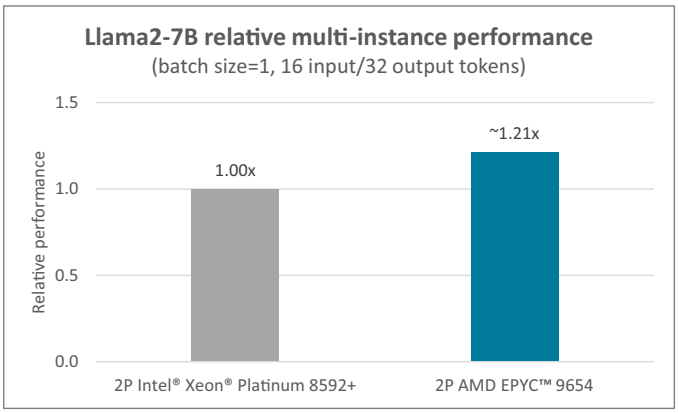


Figure 5: Relative multi-instance Llama3-8B performance (right top)

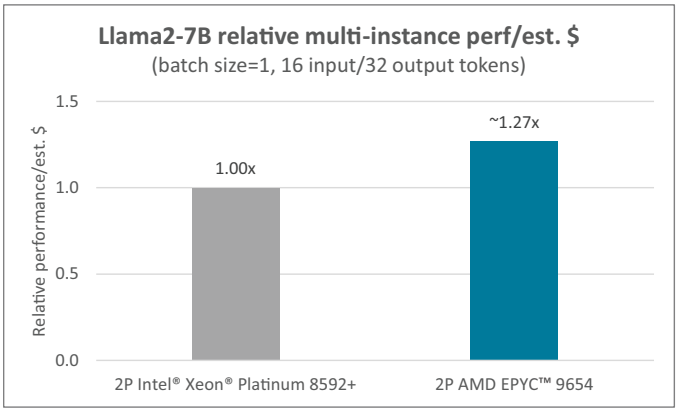


Figure 6: Relative multi-instance Llama3-8B perf/est. $ (right bottom)

LLAMA3-8B MULTI-INSTANCE THROUGHPUT

Figures 7 and 8 show the relative throughput of the 2P AMD EPYC 9654 system versus the 2P Intel Xeon Platinum 8592+ system running multiple concurrent Llama3-8B instances with a batch size of 1 and 16 input/32 output tokens. For these tests:

- The 2P Intel Xeon Platinum system ran 16 instances using 8 cores per instance.
- The 2P AMD EPYC 9654 system ran 24 instances using 8 cores per instance and delivered ~1.17x the performance and ~1.23x the performance/est. $ of the Intel system while running 50% more concurrent instances than the Intel Xeon Platinum 8592+ system.
- These tests used IPEX v2.3.0.
- Results include Time To First Token (TTFT) times.

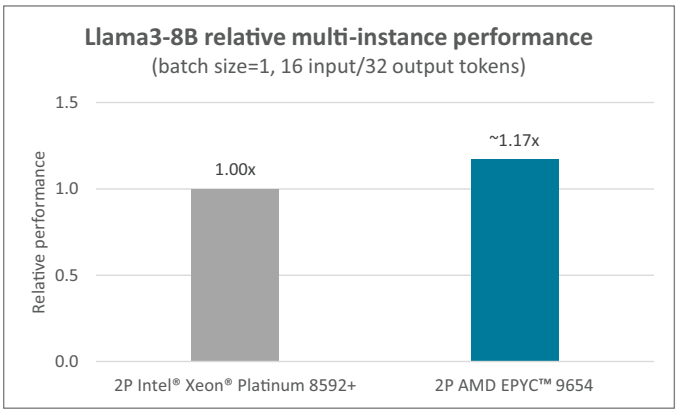


Figure 7: Relative multi-instance Llama3-8B performance (right top)

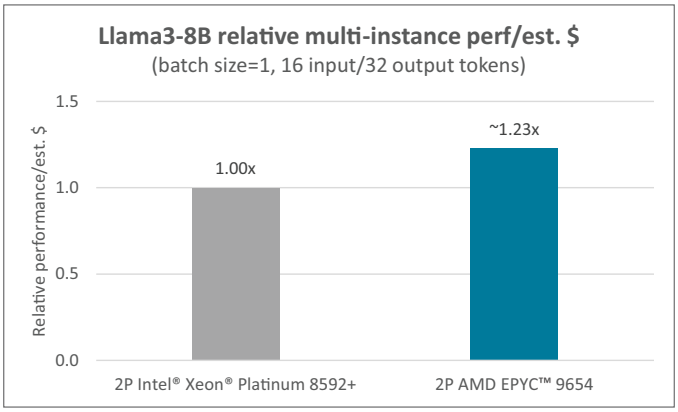


Figure 8: Relative multi-instance Llama3-8B perf/est. $ (right bottom)

TEST METHODOLOGY

Llama 2 and Llama3 are a family of Large Language Models (LLMs) developed and released to the public by Meta. Llama2 consists of pre-trained or tuned generative text models containing either 7 billion (7B), 13 billion (13B), or 70 billion (70B) parameters. Of these, the 7B model is optimal for use cases involving dialogue. Llama3 uses models containing either 8 billion (8B) or 70 billion (70B) parameters. Llama 2 is designed for English commerce and research use cases in English. Users can adapt pre-trained models for various natural language tests, while tuned models offer assistant-like chat natural language generation tasks. Llama3 is designed to increase performance and deliver better results compared to Llama2.

Testing used the Llama2-7B and Llama3-8B models. Each test consisted of three sets that consisted of five runs per set. Two runs of each five-run set were discarded as warm-up runs, and the mean was taken of the remaining three runs. The multi-instance tests used `numactl` to pin instances to individual processor cores.

SYSTEM CONFIGURATION

AMD SYSTEM CONFIGURATION	
CPUs	2 x AMD EPYC 9654
Frequency: Base | Boost[2]	2.4 GHz |3.7 GHz (up to)
Cores	96 cores/socket (192 threads) - 1 NUMA domain/socket
L3 Cache	384 MB per CPU
Memory	1.5TB (24 x Dual-Rank DDR5-5600 64GB DIMMs, 1DPC [platform supports up to 4800MHz])
NIC	2 x 100 GbE Mellanox CX-5 (MT28800)
Storage	3.2 TB Samsung MO003200KYDNC U.3 NVMe®
BIOS Version	1.56
BIOS Settings	SMT=OFF, Determinism=Power, NPS=1, PPL=400W, Turbo Boost = Enabled
OS	Ubuntu® 22.04.3 LTS | 5.15.0-94-generic
OS Settings	tuned-adm profile throughput-performance
Estimated system costs	$37,265.01 (base system) Cost info from https://dcsc.lenovo.com/* as of July 23, 2024.

Table 1: AMD system configuration

INTEL SYSTEM CONFIGURATION	
CPUs	2 x Intel Xeon Platinum 8592+
Frequency: Base | Boost[2]	1.9 GHz | 3.9 GHz
Cores	62 cores/socket (128 threads) - 1 NUMA domain/socket
L3 Cache	320 MB per CPU
Memory	1024 GB (16 x Dual-Rank DDR5-5600 64GB DIMMs, 1DPC)
NIC	4 x 1GbE Broadcom NetXtreme BCM5719 Gigabit Ethernet PCIe
Storage	3.84TB KIOXIA KCMYXRUG3T84 NVMe
BIOS Version	ESE124B-3.11
BIOS Settings	Hyperthreading=Off, Turbo boost=Enabled, SNC=Disabled
OS	Ubuntu® 22.04.3 LTS | 5.15.0-94-generic
OS Settings	tuned-adm profile throughput-performance
Estimated system costs	$39,173.01 (base system) Cost info from https://us-dcsc.lenovo.com/* as of June 19, 2024)

Table 2: Intel system configuration

SOFTWARE AND VERSIONS	
Python	3.10
Transformers	4.35.0
PyTorch Extensions	Intel Tensor Processing Primitives, IPEX 2.3.0, TPP 0.0.1

Table 3: Software used

CONCLUSION

Intel released their 5th generation of Intel Xeon Scalable processors (codenamed "Emerald Rapids") in December of 2023. Meanwhile, AMD has released four generations of AMD EPYC™ processors to date. The striking uplifts shown in this performance brief demonstrate that the relentless focus by AMD on data center performance continues to surpass the competition. The thriving AMD EPYC ecosystem includes over 250 distinct server designs and more than 800 unique cloud instances. AMD EPYC processors hold over 300 world records for performance across a wide spectrum of workloads, and Intel's latest release does not change this leadership position.

FOR ADDITIONAL INFORMATION

Please see the following additional resources for more about 4th Gen AMD EPYC features, architecture, and available models:

- AMD EPYC™ 9004 Series Processors
- AMD Documentation Hub

REFERENCES

1. AMD Infinity Guard features vary by EPYC™ Processor generations. Infinity Guard security features must be enabled by server OEMs and/or Cloud Service Providers to operate. Check with your OEM or provider to confirm support of these features. Learn more about Infinity Guard at https://www.amd.com/en/technologies/infinity-guard. GD-183
2. Maximum boost for AMD EPYC processors is the maximum frequency achievable by any single core on the processor under normal operating conditions for server systems. EPYC-18

AUTHORS

Dinesh Chitlangia, Rema Hariharan, Gitu Jain, Ziwei Peng

RELATED LINKS

- AMD EPYC Processors
- AMD Documentation Hub

**Links to third party sites are provided for convenience and unless explicitly stated, AMD is not responsible for the contents of such linked sites and no endorsement is implied.*

SUPERB AI PERFORMANCE

Enterprises of all sizes rely on ever-growing datasets to query and analyze data to derive mission-critical business insights that support key decisions. Systems powered 4th Gen AMD EPYC™ 9xx4 processors deliver superb data analytics performance across even the most demanding workloads and dataset.

"ZEN 4" CORE & SECURITY FEATURES

General-purpose support for up to:

- 96 physical cores, 192 threads
- 384 MB of L3 cache per CPU
- 32 MB of L3 cache per CCD
- 6 TB of DDR5-4800 memory
- Up to 128 1P, up to 160 2P PCIe® Gen 5 lanes

Infinity Guard security features[1]

- Secure Boot

Encrypted memory with SME