LEADERSHIP END-TO-END AI PERFORMANCE
OUTPERFORMING 5TH GEN INTEL® XEON®

Powered by 4th Gen AMD EPYC™ Processors

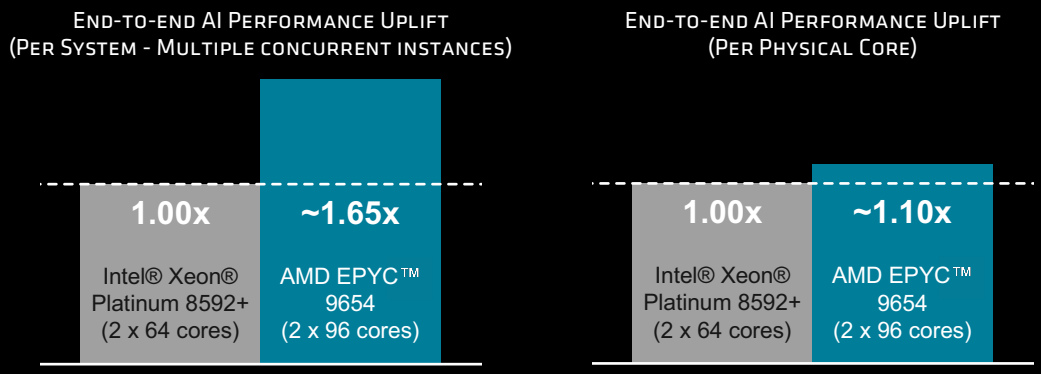
May 2024

AT A GLANCE

A 2P 4th Gen AMD EPYC™ 9654 system delivers a ~1.65x uplift on an end-to-end AI load derived from the TPC-Benchmark™ AI (TPCx-AI) benchmark vs. a 2P 5th Gen Intel® Xeon® Platinum 8592+ system while running 50% more concurrent instances.

PERFORMANCE HIGHLIGHTS

Testing compared the relative performance of a dual-socket 96-core AMD EPYC 9654 system and a dual-socket 64-core Intel Xeon Platinum 8592+ system running multiple instances of an end-to-end AI workload derived from the TPCx-AI benchmark. The AMD system with 192 total cores ran 6 concurrent 32-core instances versus the Intel system with 128 total cores, which ran 4 concurrent 32-core instances.



KEY TAKEAWAYS

AI is transforming daily life across an ever-broadening array of use cases from government to enterprises, academia, and defense. The use cases tested for this performance brief emulate real world AI and data science applications and cover loading, model training, and inference using a diverse 30 GB dataset (Scale Factor 30). Each instance used 32 processor cores based on experiments conducted to determine the optimal usage and scaling per instance.

The ~1.65x system-level and ~1.10x per-core uplifts delivered by the 2P 4th Gen AMD EPYC 9654 system showcases the ongoing performance leadership of AMD EPYC processors versus the latest competing processors.

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

- 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.
- AVX-512 instruction support for enhanced HPC and ML performance.
- AMD Infinity Guard technology to defend your data.[1]

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]

TEST METHODOLOGY

The use cases tested are open-source workloads derived from the TPC Benchmark -AI (TPCx-AI) standard and as such are not comparable to published TPCx-AI results, as the results do not comply with the TPCx-AI Benchmark Standard. The systems described in Tables 1 and 2, below, ran multiple concurrent instances of these workloads using 32 physical cores (with threading enabled for 64 total threads) per instance. Thus, the 4th Gen AMD EPYC 9654 system with dual 96-core processors (192 total cores) ran six concurrent instances (192/32=6). The 5th Gen Intel Xeon Platinum 8592+ system with dual 64-core processors (128 total cores) ran four concurrent instances (128/32=4).

The benchmark metric for each instance was obtained at AIUCpm@30.0 (AI use cases/minute @ SF30). The final results used for comparison are the aggregated scores (sums) from each concurrent instance (sum of 6 scores on the AMD EPYC system and of 4 scores on the Intel Xeon Platinum system), which provides the total end-to-end AI load measure per system. These tests extend TPCx-AI by maximizing CPU, memory, and other resources to maximize performance and demonstrate system capabilities.

USE CASES TESTED

These use cases emulate the following data science pipelines that cover end-to-end AI tasks, including loading data, training, and inference:

- **Customer segmentation:** This test uses retail orders and returns to define customer clusters and group those customers by spending behavior. A K-means clustering algorithm finds the optimum number of clusters and underlying customer segments based on these inputs.
- **Customer conversion:** This test emulates translating audio conversations into text, such as recorded customer service calls. Conversations are loaded, resampled, encoded, and finally translated to text.
- **Sales forecasting:** This test extrapolates a limited set of prior sales data to forecast sales for up to one year on a per-department basis. The input data consists of orders, products, line items, and store departments, along with markdowns. This testing assumes a chain of stores with multiple departments. A multiple regression model uses a combination of predictors to create the forecasts.
- **Spam detection:** This test uses a Naïve Bayes model to predict spam using comments, reviews, or descriptions as input.
- **Price prediction:** This test suggests or predicts the online retail price of an item based on factors such as brand, product name, description. A recurrent neural network model processes sequential data to derive results.
- **Hardware failure:** This test uses collected data and logs to predict hardware failure in advance using logged hardware events (event, component ID, date, and prior failure).
- **Product rating:** This test emulates a product recommendation system for an online marketplace. It uses a shopper's history with particular focus on how that shopper reviewed past purchases to predict items that the customer may be interested in.
- **Trip classification:** This test classifies customer trips to a retail location, such as normal weekly grocery shopping versus holiday shopping. This information can then be used to predict future trips, which can inform promotions, marketing, product placement, and other business decisions.
- **Face recognition:** This test emulates recognizing faces based on a set of previously recognized faces.
- **Fraud detection:** This test is designed to detect whether a financial transaction is fraudulent based on retail financial transaction data.

SYSTEM CONFIGURATION

AMD NODE 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=ON, 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, ulimit -l 2097152, ulimit -n 65535, ulimit -s unlimited, export OMP_NUM_THREADS=2, echo 3 > /proc/sys/vm/drop_caches, echo always > /sys/kernel/mm/transparent_hugepage/enabled

Table 1: AMD system configuration

INTEL NODE CONFIGURATION	
CPUs	2 x Intel Xeon Platinum 8592+
Frequency: Base | Boost[2]	1.9 GHz | 3.9 GHz
Cores	64 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=Enabled, Turbo boost=Enabled, SNC=Disabled
OS	Ubuntu® 22.04.3 LTS | 5.15.0-94-generic
OS Settings	tuned-adm profile throughput-performance, ulimit -l 2097152, ulimit -n 65535, ulimit -s unlimited, export OMP_NUM_THREADS=2, echo 3 > /proc/sys/vm/drop_caches, echo always > /sys/kernel/mm/transparent_hugepage/enabled

Table 2: Intel system configuration

CONCLUSION

Intel® released their 5th generation of Intel Xeon® Scalable processors (codenamed "Emerald Rapids") in December of 2023. The striking uplifts shown in this performance brief demonstrate that AMD's relentless focus 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

AUTHOR

Hari Surendran Nair

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 DATA ANALYTICS 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

DISCLAIMERS

COPYRIGHT NOTICE