BROAD SPECTRUM AI WORKLOAD PERFORMANCE LEADERSHIP
OUTPERFORMING 5TH GEN INTEL® XEON®

Powered by 4th Gen AMD EPYC™ Processors

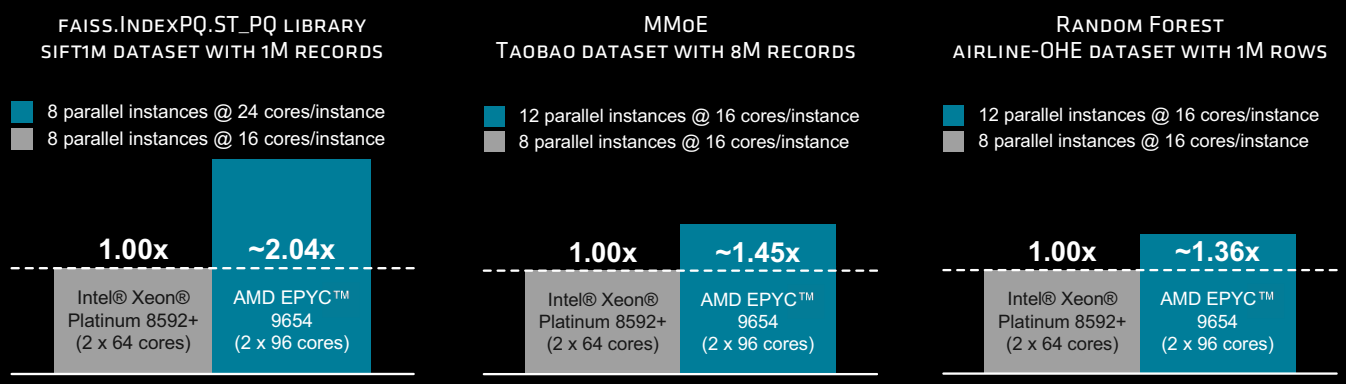
May 2024

AT A GLANCE

A dual-socket 4th Gen AMD EPYC™ 9654 system delivers impressive performance leadership versus a dual-socket 5th Gen Intel® Xeon® Platinum 8592+ system on a variety of AI workloads.

PERFORMANCE HIGHLIGHTS

A dual-socket 96-core AMD EPYC 9654 system shows throughput performance uplifts of ~1.36x to ~2.04x on Facebook AI Similarity Search (FAISS), Multi-Gate Mixture of Experts (MMoE), and Random Forest (RF) versus a dual-socket Intel Xeon Platinum 8592+ system utilizing all system cores.



KEY TAKEAWAYS

The Facebook AI Similarity Search (FAISS) library, Multi-gate Mixture of Experts (MMoE), and Random Forest represent popular AI training and inference methods used for diverse purposes that include finding similarities between multimedia files, individuated advertising recommendations, and predicting airline flight delays, to name just the few examples included in this performance brief. A 2P 4th Gen AMD EPYC 9654 system delivers stellar performance uplifts of ~2.04x (FAISS multimedia search), ~1.45x (MMoE advertising recommendation), and ~1.36x (random forest airline flight delay prediction) versus a 2P Intel Xeon Platinum 8592+ system. Every classical AI/ML problem behaves differently, and only AMD EPYC processors deliver the top core counts needed to easily scale server throughput of a server when not running day-to-day business applications.

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 up to 160 lanes (2P), 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

This section describes the AI models and datasets used for this performance briefs, how testing was performed, and how results were calculated and reported. All testing occurred on the systems described in Tables 1 and 2, below.

FAISS

The Facebook AI Similarity Search (FAISS) library enables fast, scalable searches for similar multimedia files. It transcends legacy databases by allowing nearest-neighbor searches across large datasets with optimal memory-speed-accuracy balancing. AMD tested Faiss.Index PQ.ST_PQ (Product Quantization with Subspace Tree) using the sift1m* Scale-Invariant Feature Transform (SIFT) image descriptor dataset. Faiss.IndexPQ.ST is a library for efficiently searching for similar images and clustering vectors, where PQ quantizes vectors and ST is a hierarchical clustering method of organizing those quantized vectors. Results were obtained in queries per millisecond from which throughput is calculated. Uplifts are calculated using the median of three runs on both the AMD and Intel systems.

- The AMD EPYC system ran 8 parallel instances of 24 cores/instance, where each instance was pinned to one processor quadrant each, as follows: Instance 1: 0-23, Instance 2: 24-47, Instance 3: 48-71, Instance 4: 72-95, Instance 5: 96-119, Instance 6: 120-143, Instance 7: 144-167, and Instance 8: 168-191.[2]
- The Intel Xeon system ran 8 parallel instances of 16 cores/instance, where each instance was pinned to a sequence of 16 cores, as follows: Instance 1: 0-15, Instance 2: 16-31, Instance 3: 32-47, Instance 4: 48-63, Instance 5: 64-79, Instance 6: 80-95, Instance 7: 96-111, and Instance 8: 112-127.[2]

The AMD EPYC system achieved an uplift of ~2.04x over the Intel Xeon system while only using 50% more cores per instance because of the highly performant "Zen 4" core architecture.

MMOE

A Multi-gate Mixture of Experts (MMoE) sends inputs to multiple AI "experts." For each gate, the outputs from these experts receive different weightings to obtain the desired output for that gate. AMD tested a Taobao* dataset containing 8M records that uses historical advertising data to predict customer click and buy behavior, which can enable more effective, individuated ad recommendations. Throughput results were obtained in steps per second in the same 1001 iteration across all tests, and uplifts were calculated using the median of three runs on both the AMD EPYC and Intel Xeon Platinum systems.

- The AMD EPYC system ran 12 parallel instances of 16 cores/instance.[3]
- The Intel Xeon system ran 8 parallel instances of 16 cores/instance.[3]

The AMD EPYC system achieved a ~1.45x throughput uplift.

RANDOM FOREST

This common Machine Learning (ML) algorithm predicts results using multiple concurrent and uncorrelated decision trees that independently train multiple models and then aggregate the results to generate the final prediction. AMD performed random forest testing using the airline-OHE* dataset with 1M rows of data to predict the likelihood of flight delays. Totat time taken to run the datasets were obtained from which Throughput is calculated [runs per hour] Uplifts are calculated using the median of three runs on both the AMD and Intel systems.

- The AMD EPYC system ran 12 parallel instances of 16 cores/instance.[4]
- The Intel Xeon system ran 8 parallel instances of 16 cores/instance.[4]

The AMD EPYC system achieved a ~1.36x throughput uplift despite while running 50% more instances because of the higher core count and high-performance "Zen 4" cores.

SYSTEM CONFIGURATION

AMD NODE CONFIGURATION	
CPUs	2 x AMD EPYC 9654
Frequency: Base | Boost[5]	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 DDR5-4800 64GB DIMMs, 1DPC)
Storage	2 x 1.92 TB Samsung MZQL21T9HCJR-00A07 NVMe®
BIOS Version	1006C
BIOS Settings	SMT=OFF, Determinism=Power, NPS=1
OS	Ubuntu® 22.04.3 LTS

Table 1: AMD system configuration

INTEL NODE CONFIGURATION	
CPUs	2 x Intel Xeon Platinum 8592+
Frequency: Base | Boost[5]	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 DDR5-4800 64GB DIMMs, 1DPC)
Storage	3.2 TB Intel SSDPF2KE032T10 NVMe
BIOS Version	1.4.4
BIOS Settings	Hyperthreading=Disabled, Profile=Maximum Performance
OS	Ubuntu® 22.04.3 LTS

Table 2: Intel system configuration

CONCLUSION

Intel® released their 5th generation of Intel Xeon® Scalable processors (codenamed "Emerald Rapids") in December of 2023. AMD has released four generations of AMD EPYC™ processors since 2017. The uplifts shown in this performance brief demonstrate that the relentless AMD 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. FAISS v1.7.4 1000 throughput workload claim based on AMD internal testing as of 4/19/2024. 2P server configurations: 2P EPYC 9654 (96C/96T), BIOS 1006C (SMT=off, NPS=1, Power Determinism), 1.5TB (24x 64GB DDR5-4800), 2 x Samsung MZQL21T9HCJR-00A07 1.92 TB, Ubuntu® 22.04.3 LTS running 8 instances/24 cores/instance scoring 39.6 median throughput is 2.04x the performance of 2P Xeon Platinum 8592+ (64C/64T), BIOS 1.4.4 (HT=off, Profile=Maximum Performance), 1TB (16x 64GB DDR5-4800), Intel SSDPF2KE032T10 3.2TB NVMe, Ubuntu 22.04.3 LTS running 8 instances/16 cores/instance scoring 19.4 median throughput. Results may vary due to factors including system configurations, software versions and BIOS settings. SP5-185A
3. MMoE r1.15.5-deeprec2306 global_steps/sec workload claim based on AMD internal testing as of 4/19/2024. 2P server configurations: 2P EPYC 9654 (96C/192T), BIOS 1006C (NPS=1, Power Determinism), 1.5TB (24x 64GB DDR5-4800), 2x Samsung MZQL21T9HCJR-00A07 1.92 TB, Ubuntu® 22.04.3 LTS running 12 instances/16 cores/instance scoring 28.86 median global_steps/sec is 1.45x the performance of 2P Xeon Platinum 8592+ (64C/128T), BIOS 1.4.4 (HT=OFF, Profile=Maximum Performance), 1TB (16x 64GB DDR5-4800), Intel SSDPF2KE032T10 3.2TB NVMe, Ubuntu 22.04.3 LTS running 8 instances/16 cores/instance scoring 19.97 avg. global_steps/sec. Results may vary due to factors including system configurations, software versions and BIOS settings. SP5-183A
4. SciKit-Learning Random Forest v2023.2 airline_ohe data set throughput workload claim based on AMD internal testing as of 4/19/2024. 2P server configurations: 2P EPYC 9654 (96C/96T), BIOS 1006C (SMT=off, NPS=1, Power Determinism), 1.5TB (24x 64GB DDR5-4800), 2x Samsung MZQL21T9HCJR-00A07 1.92 TB, Ubuntu® 22.04.3 LTS running 12 instances/16 cores/instance scoring 166.8 median throughput is 1.36x the performance of 2P Xeon Platinum 8592+ (64C/64T), BIOS 1.4.4 (HT=off, Profile=Maximum Performance), 1TB (16x 64GB DDR5-4800), Intel SSDPF2KE032T10 3.2TB NVMe, Ubuntu 22.04.3 LTS running 8 instances/16 cores/instance scoring 123.1 median throughput. Results may vary due to factors including system configurations, software versions and BIOS settings. SP5-184A
5. 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

Rema Hariharan

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