CLOUD-NATIVE WORKLOADS ON AMD EPYC™ 9754 PROCESSORS
COMPREHENSIVE PERFORMANCE LEADERSHIP

Powered by 4th Gen AMD EPYC™ 9754 Processors

June 2023

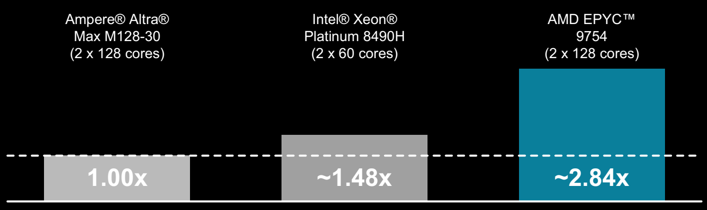
AT A GLANCE

2P systems powered by top-of-stack 128-core 4th Gen AMD EPYC™ 9754 CPUs show strong performance uplifts on key cloud-native workloads compared to top-of-stack Intel® Xeon® Platinum 8490H and Ampere® Altra® Max CPUs.

PERFORMANCE HIGHLIGHTS

A 2P 128-core 4th Gen AMD EPYC 9754 system demonstrates a composite average ~2.84x performance uplift on key cloud-native workloads vs. a top of stack 2P Ampere® Altra® Max systems. It also significantly outperforms a top of stack Intel® Xeon® 8490H system by ~1.92x. This Performance Brief showcases several key cloud workloads.

Cloud-Native Workloads average performance uplift



KEY TAKEAWAYS

The demand for cloud services and infrastructure continues to grow, challenging service providers to keep pace with the evolving requirements to scale—particularly in terms of managing costly space and power requirements. 4th Gen AMD EPYC 97x4 processors address this need by providing robust, energy-efficient processor cores, multithreading support, the highest thread density, and a full ecosystem of services to run the most demanding and scalable cloud native services and enterprise applications. This Performance Brief showcases AMD EPYC 9754 leadership on several industry-leading workloads, including Server-Side Java®, NGINX®, Redis®, Memcached, Cassandra®, FFmpeg™, MySQL™, TPROC-C, and MySQL TPC-H benchmarks.

4th Gen AMD EPYC 97x4 processors are available in 1P and 2P configurations and feature:

- Up to 128 cores (256 threads) per processor.
- Up to 256MB L3 cache.
- Up to 4 links of Gen 3 Infinity Fabric™ at up to 32 Gbps.
- 12 memory channels that support up to 6TB of DDR5-4800 memory.
- Support for PCIe® Gen 5 at up to 32 Gbps.
- AVX-512 instruction support for enhanced HPC and ML performance.
- AMD Infinity Guard technology to defend your data.[2]

IN THIS BRIEF

DETAILED PERFORMANCE RESULTS

Figure 1 shows the performance benchmark results of AMD EPYC 9754 processors versus top of stack Intel Xeon Platinum 8490H and Ampere Altra Max systems for the following key industry standard cloud native application workloads: Server-Side Java, Database performance on MySQL TPROC-C, NGINX, Redis, Memcached, Casandra, and FFmpeg. The AMD EPYC 9754 demonstrates significant cloud native application performance uplifts and leadership on each individual workload.

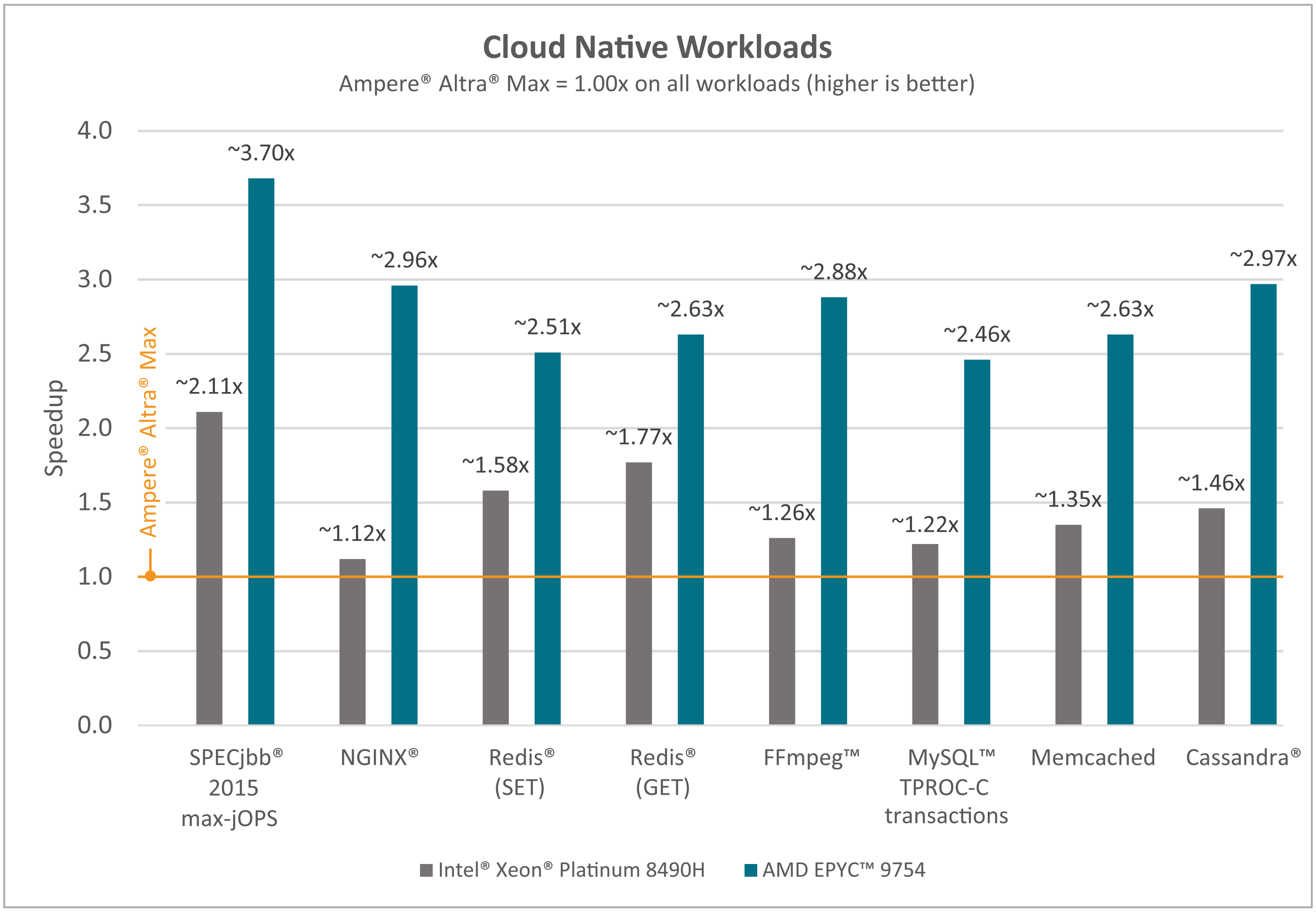


Figure 1: Performance benchmark results of AMD EPYC 9754 processors vs. top of stack Intel Xeon and Ampere Altra Max processors

TEST METHODOLOGY

The AMD EPYC 9754, Intel Xeon Platinum 8490H, and Ampere Altra Max systems were each tuned to exuberate maximum performance. System performance was maximized by creating the appropriate number of instances/VMs depending on the available resources.

The median of 3-5 runs per instance/VM was calculated and then aggregated across all VMs/instances per server to obtain overall workload performance score. This section also describes the tests and test configurations included in this Performance Brief. Please see Tables 1-11 for detailed system configuration information. The workloads shown in Figure 1 include:

- **SPECjbb® 2015:** SPECjbb® 2015[1] is the industry-standardized, CPU, and memory intensive benchmark for measuring and comparing compute intensive Java performance, stressing a system's processor, memory subsystem and Java® VM. SPECjbb® 2015 measures both a pure throughput metric and a metric that measures critical throughput under service level agreements (SLAs). This Performance Brief targets max JOPs throughput metric. The total amount of variability between individual runs was within run-to-run variations allowed as per SPECjbb® 2015 run and reporting rules at https://www.spec.org/jbb2015/docs/runrules.pdf*. The 2P Ampere results are estimates.

- **NGINX:** The NGINX web server benchmark test profile uses WRK to serve HTTP requests over a fixed period of time using a configurable number of concurrent clients/connections. Testing focuses on NGINX server throughput as measured by requests per second served from the NGINX server to a client such as WRK. Testing used multiple concurrent NGINX instances to fully load the systems being tested.

SYSTEM	INSTANCES
2P AMD EPYC 9754 (128 cores, 256 threads/CPU)	32 NGINX server, 32 WRK instances; 16 threads per instance; server & client sharing threads
2P Intel Xeon Platinum 8490H (60 cores, 120 threads/CPU)	40 NGINX server, corresponding 40 WRK client; 6 threads per instance; server & client sharing threads
2P Ampere Altra Max (128 cores/CPU)	32 NGINX server, 32 WRK; instances; 8 threads per instance; server & client sharing threads

Table 1: NGINX instance configurations

- **Redis:** Redis is an open source (BSD licensed), in-memory data structure store used as a database, cache, message broker, and streaming engine. Redis testing focuses on server throughput as measured by the number of SET and GET requests per second served from the redis-server to a client redis-benchmark from the OS distro (e.g., Ubuntu 22.04). Testing used multiple concurrent redis-server instances on the systems being tested to scale to max resource utilization.

SYSTEM	INSTANCES
2P AMD EPYC 9754 (128 cores, 256 threads/CPU)	64 Redis server; 64 Redis client; 8 threads per instance
2P Intel Xeon Platinum 8490H (60 cores, 120 threads/CPU)	40 Redis Server; 40 Redis client; 6 threads each
2P Ampere Altra Max (128 cores/CPU)	32 Redis server; 32 Redis client; 8 threads per instance

Table 2: Redis instance configurations

- **FFmpeg:** FFmpeg is a complete, cross-platform solution to record, convert and stream audio and video. FFmpeg testing calculates the Frames per Hour metric for a given encoder using the FFmpeg package provided by the operating system distribution, such as Ubuntu® 22.04. Testing used multiple concurrent FFmpeg instances to fully load the systems being tested. `numactrl` was used to set 16, 8, or 4 CPU threads per instance, and each server also includes a WRK client instance per FFmpeg instance.

SYSTEM	INSTANCES
2P AMD EPYC 9754 (128 cores, 256 threads/CPU)	64 FFMpeg jobs; 8 threads per job
2P Intel Xeon Platinum 8490H (60 cores, 120 threads/CPU)	60 FFMpeg jobs; 4 threads per job
2P Ampere Altra Max (128 cores/CPU)	64 FFMpeg jobs; 4 threads per job

Table 3: FFmpeg instance configurations

- **MySQL TPROC-C:** TPC Benchmark-C™ is an on-line transaction processing (OLTP) benchmark that defines a set of functional requirements that can be run on any transaction processing system, regardless of hardware or operating system. Testing used Ubuntu 22.04 OS and MySQL 8.0.33. The HammerDB benchmark tool was used to build and generate the TPROC-C workload. The HammerDB TPROC-C workload is an open-source workload derived from the TPC-C Benchmark Standard and as such is not comparable to published TPC-C results, as the results do not comply with the TPC-C Benchmark Standard. VMs were created with ~32 CPU threads, 128GB memory, and one NVMe disk running one MySQL database per VM and TPROC-C schema created for 1000 warehouses. The TPROC-C workload was run on all VMs simultaneously, and the median of five separate runs on each VM was added to calculate the total New Orders per Minute (NOPM) aggregate.

SYSTEM	INSTANCES
2P AMD EPYC 9754 (128 cores, 256 threads/CPU)	16 vCPU(16 cores | 16c/32 threads, 128G)
2P Intel Xeon Platinum 8490H (60 cores, 120 threads/CPU)	8 vCPU(14 cores / 28 threads, 128G)* *Configured to prevent tile and socket crossover
2P Ampere Altra Max (128 cores, 128 threads/CPU)	8 vCPU (32 cores | 32 threads, 128 G)

Table 4: MySQL TPROC-C instance configurations

- **Memcached:** Memcached is a distributed memory object caching system, an in-memory key-value store for small chunks of arbitrary data from the results of database calls, API calls, or page rendering. Each server was configured to run multiple concurrent memcached instances for maximum utilization. Every memtier client starts with 10 connections, 8 pipelines, 8 threads, a SET:GET ratio of 1:10, default data size of 32, and uses memcached text protocol to the corresponding memcached server instance based on the port number differentiation.

SYSTEM	INSTANCES
2P AMD EPYC 9754 (128 cores, 256 threads/CPU)	64 memcached server; 64 memtier clients; 8 threads; server & client sharing threads.
2P Intel Xeon Platinum 8490H (60 cores, 120 threads/CPU)	28 memcached server, 28 memtier clients; 8 threads; server & client sharing threads
2P Ampere Altra Max (128 cores/CPU)	32 memcached server, 32 memtier clients; 8 threads; server & client sharing threads

Table 5: Memcached instance configurations

- **Apache® Cassandra®:** Cassandra is an open-source NoSQL distributed database that offers scalability and high availability without compromising performance. The Cassandra test environment was configured on 2-socket servers running Ubuntu 22.04, OpenJDK11 and Cassandra 4.1.2. To simulate a cloud environment, multiple VMs were created with vCPUs configured to fit within each socket, for overall optimal performance. The Cassandra-stress tool is a Java-based stress testing utility for basic benchmarking and load testing Cassandra database, which is started by adding 1 million values to the database. An operation rate of mixed load test with 25% write/75% read ratio was performed and recorded for each VM and median across 5 runs accounted per VM. The total aggregate across all VMs was collected as the overall score per server.

SYSTEM	INSTANCES
2P AMD EPYC 9754 (128 cores, 256 threads/CPU)	16 vCPU (16 cores/32 threads, 64 GB memory)
2P Intel Xeon Platinum 8490H (60 cores, 120 threads/CPU)	8 vCPU (15 cores/30 threads, 128 GB memory)
2P Ampere Altra Max (128 cores/CPU)	8 vCPU(32 cores/32 thread, 128 GB memory)

Table 6: Cassandra instance configurations

SYSTEM CONFIGURATION

AMD SYSTEM CONFIGURATION	
CPUs	2 x AMD EPYC 9754
Frequency: Base | Boost[3]	2.25 GHz | 3.10 GHz (up to)
Cores	128 cores/socket (256 threads)
L3 Cache	256 MB per CPU
Memory	Non-database workloads: 1.5 TB (24x) Dual-Rank DDR5 4800 64 GB DIMMs 1 DPC Database workloads: 3TB (24) Dual-Rank DDR5-4800 128GB DIMMs, 1DPC
NIC	NetXtreme BCM5720 Gigabit Ethernet PCIe
Storage: OS | Data	2 x Samsung MZQL21T9HCJR-00A07 NVMe 1.7 TB
BIOS Version	1006C
BIOS Settings	SMT=ON; NPS=4

Table 7: AMD system configuration

INTEL SYSTEM CONFIGURATION	
CPUs	2x Intel Xeon Platinum 8490H
Frequency: Base | Boost	1.90 GHz | 3.85 GHz (up to)
Cores	60 cores per socket (120 threads)
L3 Cache	112.5 MB per CPU
Memory	1.5 TB (24x) Dual-Rank DDR5 4800 64 GB DIMMs 1 DPC
NIC	NetXtreme BCM5720 Gigabit Ethernet PCIe
Storage: OS | Data	2 x Intel SSDPF2KE032T1O NVMe 2.91 TB
BIOS Version	ESE110Q-1.10
BIOS Settings	Hyperthreading=ON, PSNC=4

Table 8: Intel system configuration

AMPERE SYSTEM CONFIGURATION	
CPUs	2x Ampere Altra Max Neoverse N-1
Frequency: Base | Boost	3.00 GHz | 3.00 GHz (up to)
Cores	128 cores per socket (256 threads)
L3 Cache	32 MB
Memory	1.0 TB (16x) Dual-Rank DDR4 3200 64 GB DIMMs 1 DPC
NIC	Intel I350 Gigabit Network Connection (igb: 5.15.0-71-generic) MT28800 Family [ConnectX-5 Ex] (5.15.0-69-generic)
Storage: OS | Data	2 x Samsung MZWLJ3T8HBLS-00007 1.7 TB
BIOS Version	0ACOD009 (SCP: 2.10.20221028)
BIOS Settings	default

Table 9: Ampere system configuration

FOR ADDITIONAL INFORMATION

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

- AMD EPYC™ 9004 Series Processors
- AMD EPYC™ Tuning Guides
- AMD EPYC™ Products

REFERENCES

Ampere Altra Max results measured in compliance with SPECjbb 2015 run and reporting rules.

1. As of 6/13/2023, 2P Ampere Altra Max M128-30 128, 238,494 SPECjbb 2015 MultiJVM max-jOPS (165,757 SPECjbb 2015 MultiJVM critical-jOPS) using 4 groups, Ubuntu 22.04 (kernel 5.15.0-72-generic), Oracle Java SE 17.0.7, 16 x 64GB M393A8G40AB2-CWE DDR4-3200 (Dual-Rank) 2P Intel® Xeon® Platinum 8490H (60-core) 505,379 SPECjbb 2015 MultiJVM max-jOPS (253,829 SPECjbb 2015 MultiJVM critical-jOPS) https://spec.org/jbb2015/results/res2023q1/jbb2015-20230119-01006.html; 2P AMD EPYC™9754 (128-core, SMT on) 883,097 SPECjbb 2015 MultiJVM max-jOPS (383,660 SPECjbb 2015 MultiJVM critical-jOPS) Link https://spec.org/jbb2015/results/res2023q2/jbb2015-20230517-01044.html;
2. 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
3. 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

Gnankumar Rajaram, Jaganmohanarao Porana, Bhuvaneshwara Veerabhadrachary, and Deepak Kenchappanavara contributed to this Performance Brief

RELATED LINKS

- AMD EPYC Processors
- AMD EPYC Technical Briefs

**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.*

NO-COMPROMISE CLOUD NATIVE COMPUTING

Cloud native development practices are emerging as an optimized approach for developers to rapidly deliver more efficient and scalable services. The demand for cloud services and infrastructure continues to grow. Systems featuring 4th Gen AMD EPYC™ 97x4 processors can provide a robust, efficient environment to run the most demanding and scalable cloud native services and enterprise applications.

"ZEN 4" CORE & SECURITY FEATURES

Support for up to:

- 128 physical cores, 256 threads
- 256 MB of L3 cache per CPU
- 96 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[2]

- Secure Boot
- Encrypted memory with SME

DESIGNED FOR CLOUD NATIVE COMPUTING

Platforms powered by 4th Gen AMD EPYC™ 97x4 Series processors provide core compute densities (up to 128 cores/socket), memory capacities, bandwidth, and up to either 128 (1P) or 160 (2P) lanes of high-speed PCIe® Gen5 lanes I/O to meet design challenges encountered by Cloud Service Providers (CSPs) fulfilling the growing demands of public cloud infrastructures.