AMD EPYC™ 9754 AI/ML PERFORMANCE
STABLE DIFFUSION AND DLRM

Powered by 4th Gen AMD EPYC™ 9754 Processors

August 2023

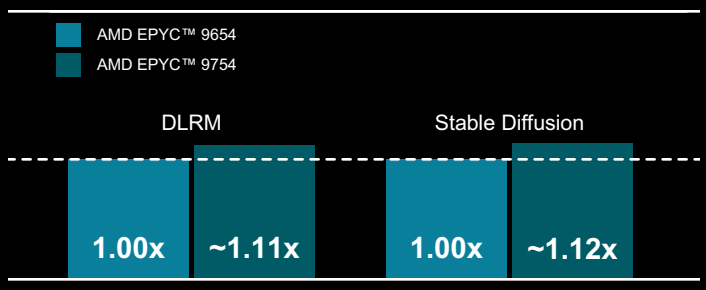
AT A GLANCE

A 1P system powered by a 128-core 4th Gen AMD EPYC™ 9754 processor shows significant performance uplift vs. a 1P system powered by a 96-core 4th Gen AMD EPYC 9654 processor running DLRM and Stable Diffusion AI/ML workloads.

PERFORMANCE HIGHLIGHTS

A single 1P 128-core 4th Gen AMD EPYC 9754 system demonstrates the following uplifts vs. a single 1P 96-core AMD EPYC 9654 system on DLRM and Stable Diffusion. The same data format (BF16) and framework (default PyTorch) were used across all tests:

DLRM queries/s (QPS) & Stable Diffusion frames/s (FPS) , 1P 128c AMD EPYC™ 9754 vs. 1P 96c AMD EPYC 9654 (normalized to AMD EPYC 9654)



KEY TAKEAWAYS

Stable Diffusion and DLRM are popular models that serve as proxies for today's generative and recommendation workloads. 4th Gen AMD EPYC 9754 processors with 128 cores demonstrate significant performance uplifts on DLRM (up to ~1.11x) and Stable Diffusion (up to ~1.12x) throughput using the bfloat16 data format versus 4th Gen AMD EPYC 9654 processors with 96 cores. Default PyTorch was used across all of the above tests. The AMD EPYC 9754 delivers no-compromise performance for cloud-native environments.

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 384 MB L3 cache.
- Up to 4 links of Gen 3 Infinity Fabric™ at up to 32 Gbps.[3]
- 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.[3]

IN THIS BRIEF

SYSTEM CONFIGURATION

AMD SYSTEM CONFIGURATIONS		
CPUs	2 x AMD EPYC 9654	2 x AMD EPYC 9754
Frequency: Base | Boost[4]	2.40 GHz | 3.55 GHz (up to)	2.25 GHz | 3.10 GHz (up to)
Cores	96 cores/socket (192 threads)	128 cores/socket (256 threads)
L3 Cache	384 MB per CPU	256 MB per CPU
Memory	1.5 TB (24x) Dual-Rank DDR5 4800 64 GB DIMMs 1 DPC	
NIC	25Gb (MT4119 - MCX512A-ACAT) CX512A -ConnectX-5 (fw 16.35.2000)	
InfiniBand	200Gb HDR (MT4123 - MCX653106A-HDAT) ConnectX-6 VPI fw 20.35.2000	
Storage: OS | Data	SAMSUNG MZQL21T9HCJR-00A07	
BIOS Version	1007D	
BIOS Settings	SMT=OFF; Determinism=Power	
OS	RHEL 8.7, Kernel 4.18.0-425.3.1.el8.x86_64	
OS Settings	amd_iommu=on iommu=pt mitigations=off; clear caches; NUMA balancing=0; randomize_va_space=0; THP=ON; CPU Governor=performance	
Software Versions	Python 3.8; PyTorch_v1.12	

Table 1: AMD system configurations

TEST METHODOLOGY

Both the DLRM and Stable Diffusion implementations were retrieved from this page*, with AMD modifications to the DLRM model for improved BF16 execution (please see "DLRM," below). All DLRM and Stable Diffusion benchmark throughputs were calculated as the aggregate of all the throughputs across all simultaneously running instances.

DLRM

- AMD improved upon the base code from the following DLRM repo linked below:
  - Base DLRM model: PyTorch implementation based on official facebookresearch/dlrm (https://github.com/facebookresearch/dlrm*)
  - Publicly available representative DLRM model: https://github.com/amd/UIF/blob/main/docs/2_model_setup/model-list/pt_dlrm_terabytes_13_26_20K_1.1_Z4.0/model.yaml*

  Scripts from this representative DLRM model directory were used in this performance brief to load the 10M model.
- Model improvements for bfloat16 DLRM execution:
  - All the MLP sequential layers in the model, including linear/dense and reLU layers, were converted to bfloat16 through data type conversion
  - Remaining layers were run with default FP32
- Model hyperparameters:
  - arch_sparse_feature_size = 64
  - arch_mlp_bot = "13-512-512-64"
  - arch_mlp_top = "1024-1024-1024-1"
  - mini_batch_size = 4032

  - num_batches = 1
  - num_indicies_per_lookup = 100
- Total number of parameters: ~514M parameters
- Threads/instances were optimized for maximum throughput, as shown in the following table:

AMD THREAD/INSTANCE CONFIGURATION (DLRM)		
CPU	AMD EPYC 9754	AMD EPYC 9654
DLRM threads/instance|# of instances	4 | 32	4 | 24

STABLE DIFFUSION

- https://huggingface.co/stabilityai/stable-diffusion-2-1-base*
- Graph torch JIT scripted
- Checkpoint: Version 2.1, 512-base-ema.ckpt
- Base model in FP32 precision, ~1.3 billion parameters
- txt 2 image mode
- 50 sampling steps, batch size of 3, generated image resolution 512x512
- Threads/instances were optimized for maximum throughput, as shown in the following table:

AMD THREAD/INSTANCE CONFIGURATION (STABLE DIFFUSION)		
CPU	AMD EPYC 9754	AMD EPYC 9654
Default PyTorch threads/ instance | # of instances	8 | 16	8 | 12

The uplift is calculated as the ratio of the systems under test (*sut*) to the reference systems (*ref*). In this Performance Brief, the AMD EPYC 9654 benchmarking results are the *ref* systems, and 4th Gen AMD EPYC 9754 processors are the *sut*. The total amount of variability between individual runs was <1%.

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™ Products
- AMD EPYC™ Tuning Guides

REFERENCES

1. The DLRM throughput comparisons are:
   -AMD EPYC 9654, bfloat16, default PyTorch, 24 instances; measured by AMD.
   -AMD EPYC 9754, bfloat16, default PyTorch, 32 instances; measured by AMD.
2. The Stable Diffusion throughput comparisons are:
   -AMD EPYC 9654, bfloat16, default PyTorch, 12 instances; measured by AMD.
   -AMD EPYC 9754, bfloat16, default PyTorch, 16 instances; measured by AMD.
3. 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
4. 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

Sarina Sit, Michael Senizaiz, Miro Hodak, Arun Ramachandran, Arun Muraleedharan, Ratan Prasad, Prakash Raghavendra, and Anthony Hernandez 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 4C" CORE & SECURITY FEATURES

Support for up to:

- 128 physical cores, 256 threads
- 384 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[3]

- Secure Boot
- Encrypted memory with SME

AMD EPYC FOR AI WORKLOADS

AI algorithms are pervasive and becoming an integral part of our daily lives, from pruning junk email from inboxes or suggesting what movies one may be interested in. AI is at an inflection point in the semiconductor industry, and AMD has accelerated its focus on innovative AI solution. AMD EPYHC 9754 processors deliver solid performance across multiple AI/ML workloads.