5 REASONS 4TH GEN AMD EPYC™ 9004 PROCESSORS ADVANCE ENTERPRISE AI INFERENCE

AT A GLANCE

The insatiable demand for AI has become a crucial factor for enterprises as they strive to understand its value. In today's fast-paced business environment, enterprises that fail to adopt AI innovation risk being left behind the competition.

4th generation AMD EPYC 9004 processors accelerate your AI journey, making it easier by providing a platform for data center consolidation, a host processor for GPU-accelerated machine learning, and an efficient processor for AI inference.

1

CONSOLIDATE TRADITIONAL & ENTERPRISE AI WORKLOADS

AMD EPYC is a powerful solution for organizations looking to unlock the potential of AI while simultaneously powering their established business applications. Its robust architecture, advanced features, and growing software ecosystem make it a key driver for innovation and achieving business goals in the AI-driven era

2

LEADERSHIP PORTFOLIO TO ADDRESS ENTERPRISE AI CHALLENGES

AMD EPYC processors & AMD Instinct™ accelerators provide a full solution portfolio for your entire AI needs. AMD EPYC consolidation advantages and inference get enterprises ready for AI, and can be used for smaller scale AI inference deployments. AMD Instinct offers leadership generative AI performance supporting larger AI models, such as LLMs.

3

LEADERSHIP ENTERPRISE AI CPU INFERENCE PERFORMANCE

AMD EPYC processors deliver incredible performance for enterprise AI CPU inference, and the optimized AMD ZenDNN software library helps deliver even greater performance gains. With AMD ZenDNN performance enhancements optimized for the most popular open-source frameworks, choose your model and take advantage of the best that AMD has to offer.

4

CREATE ENERGY-EFFICIENT SOLUTIONS

AMD EPYC processors power the world's most energy-efficient servers, delivering exceptional performance and helping reduce energy costs.[EPYC-028D] Discover new ways to optimize core usage, impact your TCO and advance your sustainability goals.

5

DEPLOY WITH THE CONFIDENCE OF ADVANCED SECURITY FEATURES & OPEN STANDARDS

Compute with confidence, knowing that your business is addressing today's security challenges with the advanced security features of AMD Infinity Guard [1]. Plus long and consistent AMD commitments to supporting open standards is critical to the development of a healthy and competitive computing ecosystem

TECHNICAL DEEP DIVE

#1 CONSOLIDATE TRADITIONAL & ENTERPRISE AI WORKLOADS

- Key features enable AMD EPYC to consolidate infrastructure, optimize costs, and adapt to evolving needs of both traditional and AI workloads
  - **High Core Count & Multithreading**: Enables efficient handling of demanding workloads and parallel processing.
  - **Large Cache Sizes**: Empowers high performance by providing fast access to frequently used data.
  - **Hardware Acceleration**: Offloads computationally intensive AI tasks to improve performance.
  - **Advanced Memory Management**: Delivers high bandwidth and low latency for demanding datasets and AI communication.
  - **Robust Software Ecosystem**: Supports leading AI frameworks and tools for optimized performance.

#2 LEADERSHIP PORTFOLIO TO ADDRESS YOUR ENTERPRISE AI CHALLENGES

- AMD offers impressive performance and efficiency for both CPU & GPUs. AMD can work with you to help identify the best path to solve your AI computing challenges.

AMD EPYC
- Mixed workload Inference
- Small to Medium Models
- Batch/Small Scale Inference

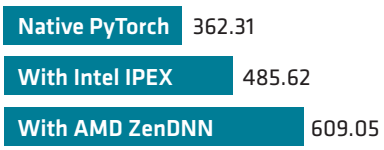
AMD INSTINCT
- Dedicated AI Deployments, Training
- Medium to Large Models
- Large-Scale Inference

- Achieve outstanding end-to-end AI throughput performance on a wide variety of use cases. Using the TPCx-AI SF30 benchmark, 2P servers with 96C AMD EPYC 9654 processors deliver up to an aggregate of ~65% more AI test cases per minute vs. 2P servers with 64C Intel Xeon Platinum 8592+.[SP5-051A]

#3 LEADERSHIP ENTERPRISE AI CPU INFERENCE PERFORMANCE

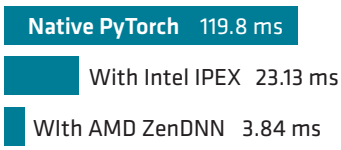
- We assembled several PyTorch models as-is, with the Intel® Extension for PyTorch (IPEX), and with the AMD ZenDNN plug-in. We tested AI inference workloads using servers with two 96-core EPYC 9654 processors and observed increases in throughput and reductions in latency with the AMD ZenDNN plug-in:

~68% Higher Image Classification Throughput [ZD-052]



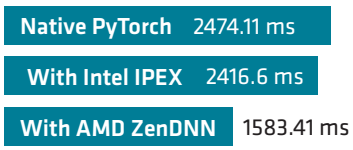
YOLO v5 images/sec at BF16 precision
more is better, batch size=960

~96.8% Faster Image Recognition [ZD-053]



ResNet50 latency at BF16 precision
lower is better, batch size=1

~36% Faster Natural Language Processing [ZD-053]



Llama2 13B latency at FP32 precision
lower is better, batch size=1

- Comparing AI inference workload performance between servers with two 96-core EPYC 9654 processors to those with two 64-core Xeon 8592+CPUs we find:
  - Classification on random decision forests ~36% faster (SciKit-Learning RandomForest airline_ohe) [SP5-184A]
  - Multi-Gate Mixture-of-Experts (MMoE) recommendations help predict customer behavior ~45% quicker (MMoE r1.15.5-deeprec230) [SP5-183A]
  - Extreme gradient boosting with the Higgs particle explosion boson data set runs ~70% faster (XGBoost 2.0.3) [SP5-251]
  - Clustering dense vectors runs ~100% faster (FAISS v1.7.4 1000 throughput) [SP5-185A]

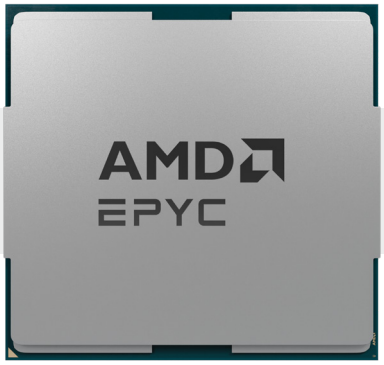
#4 CREATE ENERGY-EFFICIENT SOLUTIONS

- AMD delivers enterprise-ready IT and AI solutions that offer energy efficiency, technology innovation, and low TCO.
- 2P 4th Gen EPYC 9754 (128C)-powered servers have up to 2.4x (avg. 72% higher) the inference FPS per watt vs. 2P 5th Gen Xeon 8592+ (64C)-powered servers when comparing select published Phoronix Test Suite OpenVINO™ workloads. SP5-252
- A 2P 128-core EPYC 9754 powered server has 2.25x the overall ssj_ops/W of a 2P 64-core Xeon Platinum 8592+ base server running SPECpower_ssj®2008.SP5-011F

#5 DEPLOY WITH THE CONFIDENCE OF ADVANCED SECURITY FEATURES & OPEN STANDARDS

- The AMD "Security by Design" approach includes state-of-the-art security features and a silicon embedded security subsystem. It starts with a foundation of data and cyberthreat management from AMD Infinity Guard[1] features that have been consistently added to AMD EPYC processors across its longstanding security roadmap.
- 4th Gen AMD EPYC processors add more layers for both physical and virtual security, addressing special security concerns about migrating sensitive applications and data.
- Through our long and consistent commitment to open standards, AMD EPYC processors enable businesses to deploy confidently with unparalleled flexibility and choice, effortless scalability, and exceptional cost-efficiency.

4TH GEN AMD EPYC™ PROCESSORS ADVANCE DATA CENTER AI



TOGETHER WE ADVANCE_AI

LEARN MORE AT AMD.COM/AI

1 AMD Infinity Guard features vary by EPYC™ Processor generations and/or series. 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-183A

For details on the claims used in this document, visit amd.com/en/legal/claims/epyc.

PID 242805453-A August 2024