



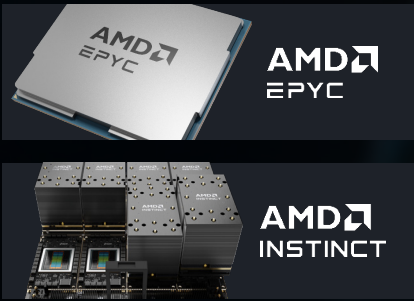
AMD EPYC™ 9005 PROCESSORS ADVANCE ENTERPRISE AI INFERENCE

AMD LEADERSHIP ENTERPRISE AI PORTFOLIO HELPS CONSOLIDATE TRADITIONAL & AI WORKLOADS

Key features enable AMD EPYC to consolidate infrastructure, optimize costs, and adapt to evolving needs of both traditional and AI workloads.

- Advanced Memory Management
- Robust Software Ecosystem
- Hardware Acceleration
- High Core Count & Multithreading
- Large Cache Sizes

AMD
Offers Leadership
Performance & Efficiency
for both CPUs & GPUs



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

- AI Training & Dedicated Deployments
- Medium to Large Models
- Large-Scale Inference

LEADERSHIP ENTERPRISE AI CPU INFERENCE & ENERGY EFFICIENCY SOLUTIONS

AMD EPYC processors deliver incredible performance for AI inference

~39% Faster LLM
Inference Throughput

Llama 3.1-8B BF16 (tokens/sec)
[9xx5-009](#)

~86% Faster
Similarity Search

FAISS (Requests/Hour)
(16 core instances at FP32)
[9xx5-011](#)

~128% Faster
End-to-End AI

TCPxAI @ SF30 (AI use cases/min)
[9xx5-012](#)

Comparing 2P servers with 5th Gen 192-core AMD EPYC 9965 CPUs vs 4th Gen 96-core AMD EPYC 9654 CPUs

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.



“We have found that AMD has the most powerful processor on the market and that helps us build systems that can increase the throughput for each server while reducing the hardware investment costs.”

Professor Minh Hoai Nguyen
Principal Research Scientist & Head of Smart Edge, VinAI

[Case Study: https://www.amd.com/en/resources/case-studies/vin-ai.html](https://www.amd.com/en/resources/case-studies/vin-ai.html)

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](https://www.amd.com/en/legal/claims/epyc).

©2024 Advanced Micro Devices, Inc. all rights reserved. AMD, the AMD arrow, EPYC, AMD Instinct and combinations thereof, are trademarks of Advanced Micro Devices, Inc. TPC is a trademark of the Transaction Processing Performance Council. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies. Certain AMD technologies may require third-party enablement or activation. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.

