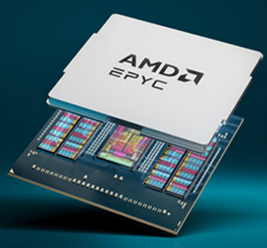
signal65

Improving AI Inference with AMD EPYC™ Host CPUs

Unlocking GPU Efficiency for Modern AI Workloads

AI inference is typically associated with GPUs, but host CPUs responsible for request handling, scheduling, and data movement play a critical role in enabling AI performance. 5th generation AMD EPYC High Frequency Processors deliver high performance per core, high frequency, and high memory bandwidth to avoid the CPU bottleneck associated with GPU-based AI and HPC workloads.



Why Host CPUs Matter for AI Inference

Choosing a capable host CPU helps maximize GPU utilization and overall AI performance.

Key CPU responsibilities include:

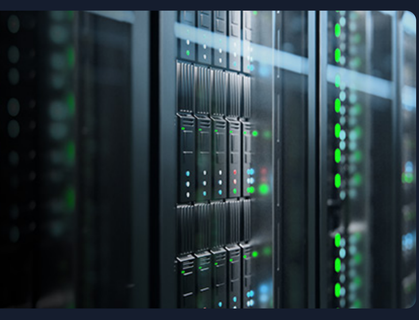
- Managing and routing inference requests
- Request batching and queue management
- Resource scheduling
- Data movement and serialization
- Returning inference results

Built for AI Inference Performance

5th Gen AMD EPYC processors are designed to meet the demands of GPU-accelerated AI environments.

Key specifications overview:

- Up to 64 cores
- Up to 5 GHz frequency
- 12-channel DDR5 memory
- Up to 160 PCIe® Gen5 lanes
- Up to 245 MB cache



Performance Testing

Signal65 evaluated the impact of host CPUs on AI inference by comparing Intel® Xeon® CPUs with AMD EPYC CPUs across 7 distinct AI models:

- GPT-OSS-120B
- Qwen2.5-VL-72B-Instruct
- Llama-3.3-70B-Instruct
- DeepSeek-R1-FP4
- Qwen2.5-Coder-Instruct
- Llama-4-Scout-17B-16E-FP4
- Llama-3.1-8B-Instruct

Key Performance Findings

Across all models tested, AMD EPYC host nodes consistently delivered higher AI inference performance.

Throughput
Up to
14%
higher throughput for both request processing and output tokens

Response Time
Up to
46.5%
faster time to first token for accelerated response start

Latency
Up to
11.4%
lower latency for better user experience

AMD EPYC™ High Frequency Processors help remove CPU bottlenecks and unlock the full potential of GPU-accelerated AI.

Explore the full Signal65 Lab Insights report >>