AMD
together we advance_

THE AI CONTINUUM: WHAT INFRASTRUCTURE WORKS BEST FOR INFERENCE?

Planning for AI is essential to any data center refresh. GPUs are critical for large AI workloads, but the latest generations of CPUs can handle a wide range of AI tasks alongside general-purpose workloads. Keep these considerations in mind as you assess your growing AI inference needs.

A lot of AI doesn't need real-time results

Modern CPUs can run small to mid-sized AI inference workloads with sub-second latency. As AI inference workloads grow or response times shrink, you may need to add discrete accelerators.

BATCH PROCESSING	MID LATENCY	LOW LATENCY	NEAR-REAL TIME	REAL TIME
Minutes to days	Seconds to minutes	~500 ms to seconds	~100 ms to ~500 ms	<10 ms to ~50 ms
USE CASES				
• Document processing and classification • Data mining and analytics • Scientific simulations	• Translation • Indexing • Content moderation • Predictive maintenance	• Virtual assistants • Chatbots • Expert agents • Video captioning	• Fraud detection • Decision-making • Dynamic pricing • Audio and video filtering	• Financial trading • Telecommunications and networking • Autonomous systems
CPUs		**CPUs + GPUs**		**Multiple GPU clusters**

As AI workloads rise, GPUs become increasingly cost-effective

CPUs alone can support mixed enterprise workloads and AI. As model size, complexity, and volumes increase, GPU clusters can deliver more performance per dollar.

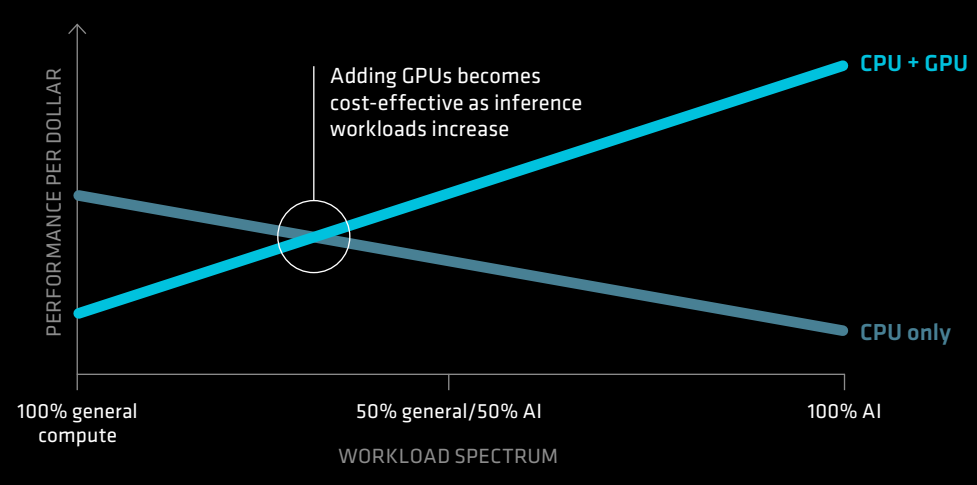
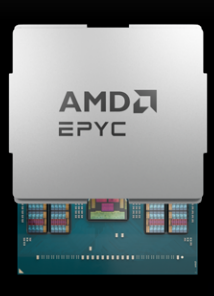


Chart for illustrative purposes only. Crossover point varies based on specific workloads and processor models.

Different models have unique processing needs

Machine learning, graph processing, and statistical methods run exceptionally well on CPUs. Small to mid-sized large language models (LLMs) perform well on the latest CPUs. Larger models can realize significant benefit from AI accelerators.

	MODEL SIZE	PROCESSOR	AMD SOLUTION
Classical machine learning	~1 MB to ~200 MB	CPUs, embedded to data center	AMD Ryzen™ CPUs AMD EPYC™ CPUs
Deep learning	~60 million parameters to ~20 billion	CPUs, data center	AMD EPYC CPUs (high core count)
Transformers/LLMs	~20 billion parameters to ~450 billion	CPUs + GPUs	AMD EPYC CPUs (high frequency) + AMD Instinct™ GPUs **or** NVIDIA GPUs
	~450 billion parameters and greater	Multiple GPU clusters	



AMD EPYC CPUs excel with enterprise-class AI

5th Generation AMD EPYC CPUs deliver major performance improvements for AI workloads:

Up to 3.8X the throughput for end-to-end AI compared to competitor CPUs[1]

Up to 90% faster throughput on Llama 3.1 8B at BF16 compared to competitor CPUs[2]

Up to 86% faster Facebook AI Similarity Search (FAISS) compared to previous-generation EPYC CPUs[3]

5th GENERATION AMD EPYC™ CPUs: THE BEST CPU FOR ENTERPRISE AI[4]

See why 5th Generation AMD EPYC CPUs excel with AI inference workloads.

Visit EPYC for AI inference

1. TPCxAI @SF30 Multi-Instance 32C Instance Size throughput results based on AMD internal testing as of 09/05/2024 running multiple VM instances. The aggregate end-to-end AI throughput test is derived from the TPCx-AI benchmark and as such is not comparable to published TPCx-AI results, as the end-to-end AI throughput test results do not comply with the TPCx-AI Specification. 2P AMD EPYC 9965 (384 Total Cores), 12 32C instances, NPS1, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled) 2P AMD EPYC 9755 (256 Total Cores), 8 32C instances, NPS1, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT0090F (SMT=off, Determinism=Power, Turbo Boost=Enabled) 2P AMD EPYC 9654 (192 Total cores) 6 32C instances, NPS1, 1.5TB 24x64GB DDR5-4800, 1DPC, 2 x 1.92 TB Samsung MZQL21T9HCJR-00A07 NVMe, Ubuntu 22.04.3 LTS, BIOS 1006C (SMT=off, Determinism=Power) Versus 2P Xeon Platinum 8592+ (128 Total Cores), 4 32C instances, AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe, Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled) Results: CPU Median Relative Generational Turin 192C, 12 Inst 6067.531 3.775 2.278 Turin 128C, 8 Inst 4091.85 2.546 1.536 Genoa 96C, 6 Inst 2663.14 1.657 1 EMR 64C, 4 Inst 1607.417 1 NA. Results may vary due to factors including system configurations, software versions, and BIOS settings. TPC, TPC Benchmark and TPC-C are trademarks of the Transaction Processing Performance Council. (9xx5-012)
2. Llama3.1-8B throughput results based on AMD internal testing as of 09/05/2024. Llama3-8B configurations: IPEX.LLM 2.4.0, NPS=2, BF16, batch size 4, Use Case Input/Output token configurations: [Summary = 1024/128, Chatbot = 128/128, Translate = 1024/1024, Essay = 128/1024, Caption = 16/16]. 2P AMD EPYC 9965 (384 Total Cores), 6 64C instances 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1 DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.3 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C, (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=22P AMD EPYC 9755 (256 Total Cores), 4 64C instances, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu 22.04.3 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=22P AMD EPYC 9654 (192 Total Cores) 4 48C instances, 1.5TB 24x64GB DDR5-4800, 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 5.15.85-051585-generic (tuned-adm profile throughput-performance, ulimit -l 1198117616, ulimit -n 500000, ulimit -s 8192), BIOS RVI1008C (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=2Versus 2P Xeon Platinum 8592+ (128 Total Cores), 2 64C instances , AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe®, Ubuntu 22.04.4 LTS 6.5.0-35-generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled). Results: CPU 2P EMR 64c 2P Turin 192c 2P Turin 128c 2P Genoa 96c Average Aggregate Median Total Throughput 99.474 193.267 182.595 138.978 Competitive 1 1.943 1.836 1.397 Generational NA 1.391 1.314 1. Results may vary due to factors including system configurations, software versions, and BIOS settings. (9xx5-009)
3. FAISS (Requests/Hour) throughput results based on AMD internal testing as of 09/05/2024. FAISS Configurations: sift1m Data Set, 16 Core Instances, FP32, MKL 2024.2.1 2P AMD EPYC 9965 (384 Total Cores), 24 16C instances, 1.5TB 24x64GB DDR5-6400 (at 6000 MT/s), 1DPC, 1.0 Gbps NetXtreme BCM5720 Gigabit Ethernet PCIe, 3.5 TB Samsung MZWLO3T8HCLS-00A07 NVMe®, Ubuntu® 22.04.4 LTS, 6.8.0-40-generic (tuned-adm profile throughput-performance, ulimit -l 198096812, ulimit -n 1024, ulimit -s 8192), BIOS RVOT1000C (SMT=off, Determinism=Power, Turbo Boost=Enabled), NPS=42P AMD EPYC 9654 (192 Total cores) 12 16C instances, 1.5TB 24x64GB DDR5-4800, 1DPC, 2 x 1.92 TB Samsung MZQL21T9HCJR-00A07 NVMe, Ubuntu 22.04.3 LTS, BIOS 1006C (SMT=off, Determinism=Power), NPS=4 Versus 2P Xeon Platinum 8592+ (128 Total Cores), 8 16C instances, AMX On, 1TB 16x64GB DDR5-5600, 1DPC, 1.0 Gbps NetXtreme BCM5719 Gigabit Ethernet PCIe, 3.84 TB KIOXIA KCMYXRUG3T84 NVMe, Ubuntu 22.04.4 LTS, 6.5.0-35 generic (tuned-adm profile throughput-performance, ulimit -l 132065548, ulimit -n 1024, ulimit -s 8192), BIOS ESE122V (SMT=off, Determinism=Power, Turbo Boost = Enabled) Results: CPU Median Relative Throughput Generational 2P Turin 192C 64.2 3.776 1.861 2P Genoa 96C 34.5 2.029 1 2P EMR 64C 17 1 NA. Results may vary due to factors including system configurations, software versions, and BIOS settings. (9xx5-011)
4. Comparison based on thread density, performance, features, process technology and built-in security features of currently shipping servers as of 10/10/2024. EPYC 9005 series CPUs offer the highest thread density, leads the industry with 500+ performance world records including world record enterprise leadership Java® ops/sec performance, top HPC leadership with floating-point throughput performance, AI end-to-end performance with TPCx-AI performance and highest energy efficiency scores. Compared to 5th Gen Xeon, the 5th Gen EPYC series also has more DDR5 memory channels with more memory bandwidth and supports more PCIe® Gen5 lanes for I/O throughput, and has up to 5x the L3 cache/core for faster data access. The EPYC 9005 series uses advanced 3-4nm technology, and offers Secure Memory Encryption + Secure Encrypted Virtualization (SEV) + SEV Encrypted State + SEV-Secure Nested Paging security features. (EPYC-029D)

AMD
together we advance_