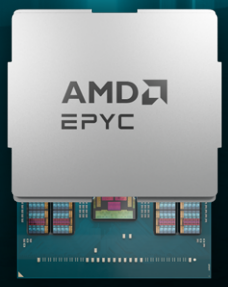
AMD

5 AI WORKLOADS
THAT CAN RUN ON A CPU



A modern data center must support multiple AI workloads, few of which are equal. An AI service may use a spectrum of small to large models, where quality of service levels will vary based on application. Plus, many enterprise applications now include AI capabilities. With careful provisioning, data centers can support many of these AI services at enterprise scale on less-expensive CPUs and reserve GPUs for heavier lifts.

1

CLASSIC MACHINE LEARNING

Traditional machine learning algorithms don't benefit from parallel computing GPUs

Machine learning tasks that use decision trees, random forests, and linear statistical models benefit from high core count CPUs and don't generally take advantage of the parallel computing GPUs offer. If tasks like sentiment analysis, text and image classification, fraud detection, or time-series forecasting make up a large portion of your workloads, CPUs with the highest available core counts are a smart investment.

2

COMPUTER VISION

Pattern recognition and deep learning vision models perform well on CPUs

Facial recognition, object detection, image classification, heat mapping—even difficult defect and anomaly detection—can run blazingly fast on GPUs but may not require that level of responsiveness. At enterprise-level and edge use-case scales, CPUs can generally handle vision tasks quite efficiently.

3

MEMORY-INTENSIVE GRAPH ANALYSIS

For graph analysis on large datasets, CPUs often outperform GPUs

Complex networks—social networks, IT systems, logistics, and supply chains—have complex nodes, interactions, and patterns best analyzed with graph algorithms. They can also create immense datasets. CPUs have direct, low-latency access to system RAM, which allows them to run large datasets in-memory, eliminating read/write cycles to storage. Choose CPUs with the highest available memory speed and capacity for maximum performance.

AMD

4

SMALL TO MID-SIZED RECOMMENDATION SYSTEMS

CPUs are a good fit for real-time recommendation engines

CPUs with higher frequencies and higher core counts provide sufficient parallelization and processing speeds for recommendation systems. For real-time recommendation systems, opt for CPUs with large caches that support high-speed RAM and max out system memory.

5

FINE-TUNED, EXPERT AI AGENTS

Fine-tuning models for specific tasks can significantly reduce their footprint

Techniques like Parameter Efficient Fine Tuning (PEFT) and Low-Rank Adaptation (LoRA) can transform large, general-purpose models into smaller, more efficient models that deliver highly accurate results. Models fine-tuned on specific knowledge bases like product catalogs, technical documentation, or escrow documents can support expert agents, chat services, and decision-making applications that run efficiently on CPUs.

RUN ENTERPRISE-CLASS AI *ON AMD EPYC™ CPUs*

Up to 192 cores, 5 GHz max frequency
Processing power for classic machine learning, recommendation systems, and AI inference

12 channels of DDR5 memory running at up to 6400 Mbps
Hold large databases, AI models, and training data in memory

Up to 160 PCIe® Gen5 lanes (2P)
Move large datasets faster for more responsive AI

5th Generation AMD EPYC™ CPUs deliver 1.9X the LLM performance of 5th Gen Intel® Xeon® Scalable processors,[1] a performance edge that makes AI—including small and mid-sized LLMs—a practical CPU workload.

Get the details on using AMD EPYC CPUs for on-chip AI inference.

1. 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)

AMD
together we advance_