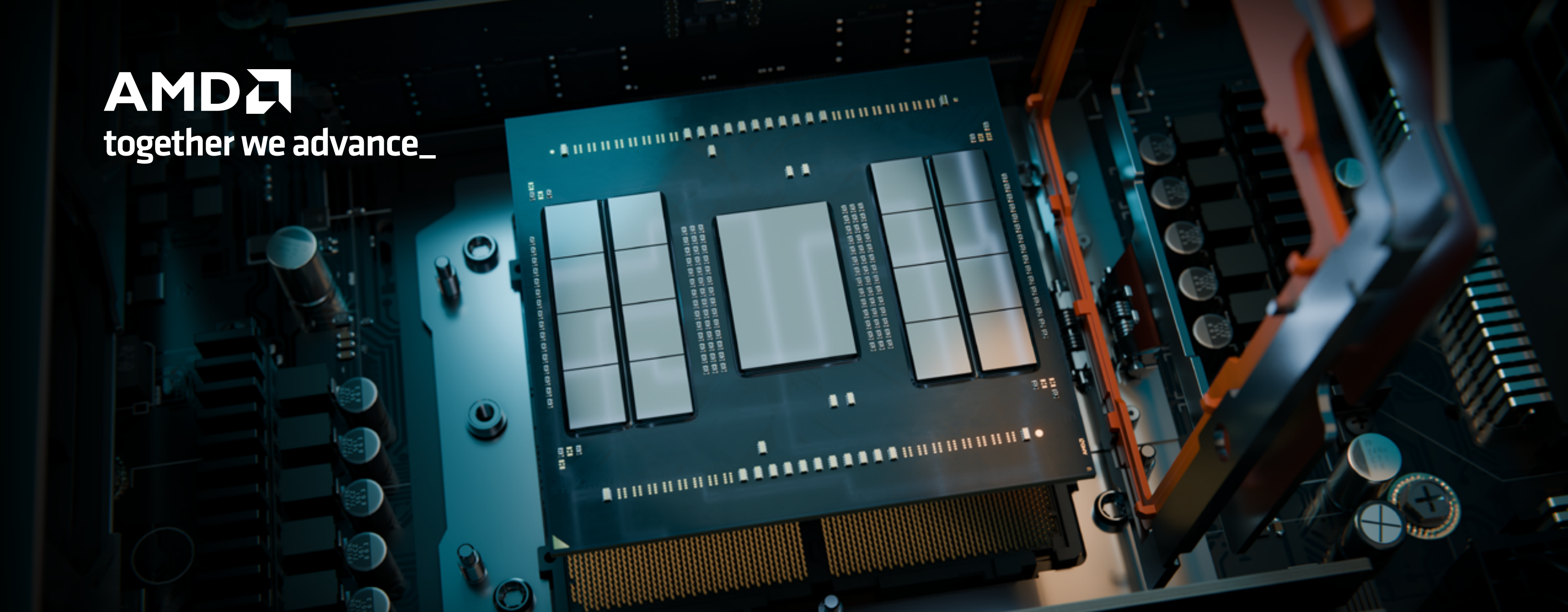
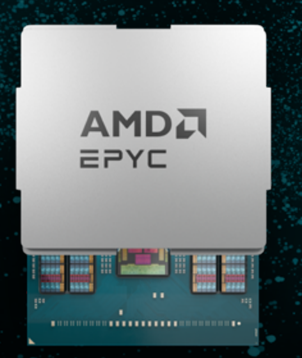
AMD
together we advance_



HOW TO GET UP TO *20% MORE AI PERFORMANCE* FROM HIGH-PERFORMANCE GPUs[1,2]

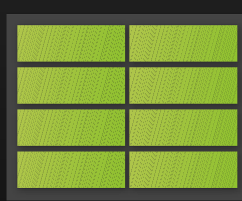


5th Generation AMD EPYC™ CPUs include higher-frequency models designed specifically for hosting accelerator platforms. These CPUs excel at orchestrating data movement and managing multiple virtual machines—critical capabilities that extract more performance from GPU platforms.[1,2]

The right host makes a difference

SELECT AMD EPYC HOST NODE CPUs INCREASE INFERENCE AND TRAINING THROUGHPUT

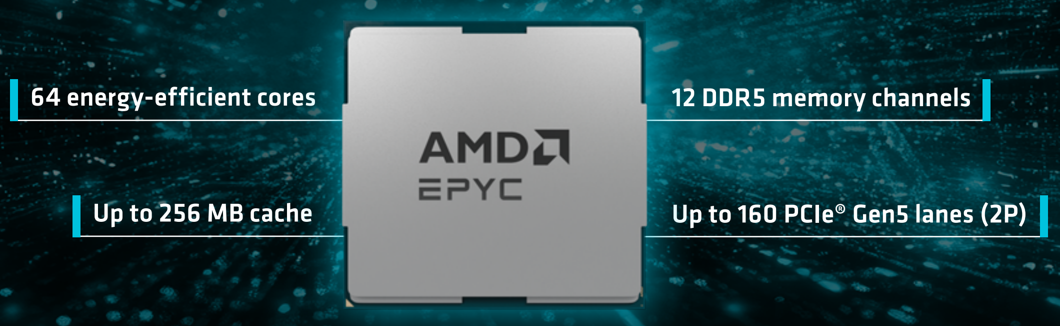
HOST CPU	GPU PLATFORM	INFERENCE	TRAINING
 2P AMD EPYC™ 9575F CPUs (128 total cores)	 8x AMD Instinct™ MI300X GPUs	Llama 3.1-70B @ FP8 Up to **8%** more tokens/sec[3] 700K more tokens/second inference on a 1K node cluster of AMD Instinct GPUs	DeepSpeed 0.14.0 @ FP8 Stable Diffusion XL v2 training set Up to **20%** more samples/sec[4]

HOST CPU	GPU PLATFORM	INFERENCE	TRAINING
 2P AMD EPYC™ 9575F CPUs (128 total cores)	 8x NVIDIA H100 GPUs	Llama 3.1-70B @ FP8 Up to **20%** more tokens/sec[5]	Llama 3.1-8B @ BF16 Max sequence length 1024 Up to **15%** more samples/sec[6]

Performance compared to 2P Intel® Xeon® Platinum 8592+ (128 total cores) hosting the same GPUs running identical workloads.

Built to boost AI accelerator performance

5 GHz EPYC 9575F max boost is 28% higher than Intel® Xeon® Platinum 8592+.[7]



64 energy-efficient cores

Up to 256 MB cache

12 DDR5 memory channels

Up to 160 PCIe® Gen5 lanes (2P)

A range of options for hosting GPUs

With frequencies up to 5 GHz and support for up to 6 TB of memory, 5th Generation AMD EPYC CPUs provide multiple models specifically designed for GPU clusters.

PROCESSOR	CORE COUNT	MAX BOOST FREQUENCY
9575F	64	5 GHz
9475F	48	4.8 GHz
9375F	32	4.8 GHz
9275F	24	4.8 GHz
9175F	16	5 GHz

5TH GENERATION AMD EPYC™ CPUs: *THE BEST CPU FOR ENTERPRISE AI*[8]

Gain industry-leading server performance for AI, enterprise, and cloud workloads with 5th Generation AMD EPYC processors.

Explore AMD EPYC

[1] Stable Diffusion XL v2 training results based on AMD internal testing as of 10/10/2024. SDXL configurations: DeepSpeed 0.14.0, TP8 Parallel, FP8, batch size 24, results in seconds 2P AMD EPYC 9575F (128 Total Cores) with 8x AMD Instinct MI300X-NPS1-SPX-192GB-750W, GPU Interconnectivity XGMI, ROCm™ 6.2.0-66, 2304GB 24x96GB DDR5-6000, BIOS 1.0 (power determinism = off), Ubuntu® 22.04.4 LTS, kernel 5.15.0-72-generic, 334.80 seconds 2P Intel Xeon Platinum 8592+ (128 Total Cores) with 8x AMD Instinct MI300X-NPS1-SPX-192GB-750, GPU Interconnectivity XGMI, ROCm 6.2.0-66, 2048GB 32x64GB DDR5-4400, BIOS 2.0.4, (power determinism = off), Ubuntu 22.04.4 LTS, kernel 5.15.0-72-generic, 400.43 seconds for 19.600% training performance increase. Results may vary due to factors including system configurations, software versions, and BIOS settings. (9xx5-059A)

[2] Llama3.1-70B inference throughput results based on AMD internal testing as of 09/01/2024. Llama3.1-70B configurations: TensorRT-LLM 0.9.0, nvidia/cuda 12.5.0-devel-ubuntu22.04, FP8, Input/Output token configurations (use cases): [BS=1024 I/O=128/128, BS=1024 I/O=128/2048, BS=96 I/O=2048/128, BS=64 I/O=2048/2048]. Results in tokens/second. 2P AMD EPYC 9575F (128 Total Cores) with 8x NVIDIA H100 80GB HBM3, 1.5TB 24x64GB DDR5-6000, 1.0 Gbps 3TB Micron_9300_MTFDHAL3T8TDP NVMe®, BIOS T20240805173113 (Determinism=Power,SR-IOV=On), Ubuntu 22.04.3 LTS, kernel=5.15.0-117-generic (mitigations=off, cpupower frequency-set -g performance, cpupower idle-set -d 2, echo 3> /proc/syss/vm/drop_caches), 2P Intel Xeon Platinum 8592+ (128 Total Cores) with 8x NVIDIA H100 80GB HBM3, 1TB 16x64GB DDR5-5600, 3.2TB Dell Ent NVMe® PM1735a MU, Ubuntu 22.04.3 LTS, kernel-5.15.0-118-generic, (processor.max_cstate=1, intel_idle.max_cstate=0 mitigations=off, cpupower frequency-set -g performance), BIOS 2.1, (Maximum performance, SR-IOV=On), I/O Tokens Batch Size EMR Turin Relative 128/128 1024 814.678 1101.966 1.353 128/2048 1024 2120.664 2331.776 1.1 2048/128 96 114.954 146.187 1.272 2048/2048 64 333.325 354.208 1.063 for average throughput increase of 1.197x. Results may vary due to factors including system configurations, software versions, and BIOS settings. (9xx5-014)

[3] As of 10/10/2024; this scenario contains several assumptions and estimates and, while based on AMD internal research and best approximations, should be considered an example for information purposes only, and not used as a basis for decision making over actual testing. Referencing 9xx5-056A: "2P AMD EPYC 9575F powered server and 8x AMD Instinct MI300X GPUs running Llama3.1-70B select inference workloads at FP8 precision vs 2P Intel Xeon Platinum 8592+ powered server and 8x AMD Instinct MI300X GPUs has ~8% overall throughput increase across select inference use cases" and 8763.52 tokens/s (9575F) versus 8,048.48 tokens/s (8592+) at 128 input / 2048 output tokens, 500 prompts for 1.089x the tokens/s or 715.04 more tokens/s. 1 Node = 2 CPUs and 8 GPUs. Assuming a 1000 node cluster, 1000 * 715.04 = 715,040 tokens/s for ~700,000 more tokens/s. Results may vary due to factors including system configurations, software versions, and BIOS settings. (9xx5-087)

[4] See note 1 above.

[5] See note 2 above.

[6] Llama3.1-8B (BF16, max sequence length 1024) training testing results based on AMD internal testing as of 09/05/2024. Llama3.1-8B configurations: Max Sequence length 1024, BF16, Docker: huggingface/transformers-pytorch-gpu:latest 2P AMD EPYC 9575F (128 Total Cores) with 8x NVIDIA H100 80GB HBM3, 1.5TB 24x64GB DDR5-6000, 1.0 Gbps 3TB Micron_9300_MTFDHAL3T8TDP NVMe®, BIOS T20240805173113 (Determinism=Power,SR-IOV=On), Ubuntu 22.04.3 LTS, kernel=5.15.0-117-generic (mitigations=off, cpupower frequency-set -g performance, cpupower idle-set -d 2, echo 3> /proc/syss/vm/drop_caches), For 31.79 Train Samples/Second. 2P Intel Xeon Platinum 8592+ (128 Total Cores) with 8x NVIDIA H100 80GB HBM3, 1TB 16x64GB DDR5-5600, 3.2TB Dell Ent NVMe® PM1735a MU, Ubuntu 22.04.3 LTS, kernel-5.15.0-118-generic, (processor.max_cstate=1, intel_idle.max_cstate=0 mitigations=off, cpupower frequency-set -g performance), BIOS 2.1, (Maximum performance, SR-IOV=On), For 27.74 Train Samples/Second for average throughput increase of 1.146. Results may vary due to factors including system configurations, software versions, and BIOS settings. (9xx5-015)

[7] Comparison of the highest-frequency 5th Generation AMD EPYC 9575F CPU (up to 5 GHz) and the highest-frequency Intel Xeon Platinum 8592+ CPU (up to 3.9 GHz), based on published specifications.

[8] 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_