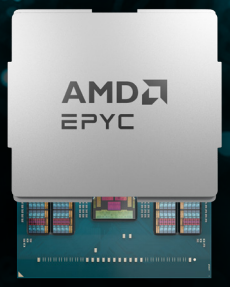
AMD



5 REASONS AMD EPYC™ CPUs GET UP TO *20% MORE AI PERFORMANCE* FROM GPUs[1,2]

High-performance GPU nodes are essential for heavy AI workloads, but high-compute, high-performance accelerators depend on fast CPU hosts to maximize their performance. 5th Generation AMD EPYC™ CPUs include high-frequency models specifically designed to get more throughput from GPU clusters. Here's how they do it.

1 HIGH-FREQUENCY PROCESSING

Higher-frequency CPUs improve data movement and virtual machine performance

Higher-frequency CPUs can move data, orchestrate tasks, and serve multiple virtual machines, enabling higher AI accelerator performance. AMD EPYC 9575F CPUs run at up to 28% higher frequency than Intel® Xeon® Platinum 8592+, making them invaluable as GPU host nodes.[3]

2 50% AVERAGE HIGHER PERFORMANCE ACROSS FUNDAMENTAL GPU HOSTING TASKS[4]

Faster kernel launches, memory swaps, and data transfers boost performance

In tests, 5th Generation AMD EPYC CPUs increased the general performance of an 8x NVIDIA H100 platform by an average 50%, raising Grok-1 inference ~52%, Kernel Launch ~30% and MemCopy tasks ~138% compared to Intel® Xeon® Platinum 8592+ CPUs.[5]

3 AMPLE RAM FOR HOLDING ENTIRE MODELS AND DATASETS IN MEMORY

12 DDR5 channels support up to 6 TB of RAM per socket

5th Gen AMD EPYC CPUs can support enough RAM to store large datasets and entire models in memory, helping reduce read/write cycles and transfers to and from storage. Keeping data in memory allows AMD EPYC CPUs to process and feed more data to GPU clusters faster.

AMD

4

WIDE, FAST DATA MOVEMENT FOR MASSIVE PARALLEL PROCESSING

Up to 160 PCIe® Gen5 lanes speed data transfers

PCIe lanes are the communication channels between the CPU, GPUs, and storage. CPUs with fewer lanes can rapidly become bottlenecks that throttle GPU performance. With up to 128 PCIe lanes (single socket) and up to 160 PCIe Gen5 lanes (dual socket), 5th Generation AMD EPYC CPUs can move huge data volumes to and across GPUs to maximize their capacity.

5

UP TO 20% HIGHER GPU THROUGHPUT[6,7]

GPU performance on heavy AI workloads jumps with AMD EPYC hosts

When hosted by 5th Generation AMD EPYC CPUs, 8x GPU platforms gain ~20% on the Stable Diffusion XL v2 (FP8) training benchmark[8] and up to 20% on the Llama 3.1-70B (FP8) inference benchmark compared to equivalent nodes hosted by Intel® Xeon® Platinum 8592+ CPUs.[9]

GET MORE FROM GPU ACCELERATORS WITH *AMD EPYC CPUs*

High-frequency 5th Generation AMD EPYC CPUs are uniquely designed to optimize performance for GPU clusters.

Learn more about the benefits of using AMD EPYC CPUs for hosting.

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. Llama 3.1-70B 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. Comparison of the highest-frequency 5th Generation AMD EPYC CPU (5 GHz) and the highest-frequency Intel Xeon Platinum 8592+ CPU (3.9 GHz), based on published specifications.
4. Comparisons based on AMD internal testing as of 11/05/2024. Workloads: MemCopy v1.0 (8 threads / 8 GPUs, nvhpc 24.3 KernelLaunch v2.0 (8 threads / 8 GPUs, nvhpc 24.3) Grok1-324B (FP16, JAX 0.4.25, nvhpc 24.3, sentencepiece 0.2.0, numpy 1.26.4, dm_haiky 0.0.12, 2 / 8 experts, 11 token input prompt with 105 token output prompt). 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), average over 3 runs 77.13 seconds (MemCopy), average over 3 runs 3.21 seconds (Kernel Launch), average over 3 runs 99.00 seconds (Grok) 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), average over 3 runs 183.58 seconds (MemCopy), average over 3 runs 4.18 seconds (Kernel Launch), average over 3 runs 163.98 seconds (Grok), for a 138.01% performance gain in MemCopy, a 30.22% performance gain in KernelLaunch, and a 51.77% performance gain in Grok1-324B, or 49.67% the overall performance gain (geometric mean). Results may vary due to factors including system configurations, software versions and BIOS settings. (9xx5-084A)
5. Ibid.
6. See footnote 1 above.
7. See footnote 2 above.
8. See footnote 1 above.
9. See footnote 2 above.

AMD
together we advance_