5 REASONS TO CHOOSE THE AMD ROCm™ PLATFORM

AT A GLANCE

AMD ROCm™ software is an open stack including drivers, development tools and APIs for harnessing the parallel computing capabilities of GPUs. Following are five reasons to choose AMD ROCm software.

1

OPEN-SOURCE WITHOUT VENDOR LOCK-IN

AMD ROCm™ software has a vibrant and collaborative open-source community that contributes to the platform's development, leading to faster innovation and broader compatibility with a variety of software and hardware configurations. Because developers are not bound by proprietary software restrictions, they can customize, modify, and optimize the platform as needed for their specific environments and requirements.

2

DESKTOP TO DATACENTER GPU SUPPORT

Developers can adapt and optimize AMD ROCm software for a broader range of hardware, which may not be possible with proprietary solutions. Because AMD ROCm software supports AMD Radeon™ GPUs and Instinct™ server accelerators, developers can choose the hardware that best meets their needs as they scale, all while working with an open software platform. This offers more visibility, flexibility and compatibility than the competition.

3

VAST SOFTWARE DEVELOPMENT ECOSYSTEM

There is a vast AMD ROCm software community that's open 24/7 for collaboration. All are encouraged to participate in discussion forums, to share fixes, tap into community expertise, contribute to and explore documentation. With AMD in 162 of the TOP500 fastest supercomputers, thousands around the world are advancing AI and HPC every day.

4

EASY PORTING

Unlike CUDA, which uses the proprietary CUDA API, the AMD ROCm platform supports the open-source Heterogeneous-compute Interface for Portability (HIP) API. HIP is a C++ runtime API and kernel language for developers to port applications to AMD GPUs with minimal code changes, reducing the effort required to migrate existing framework-level codebases. This can result in lower development costs and faster time to market for GPU-accelerated applications.

5

COST-EFFECTIVENESS

As an open software, developers can access the AMD ROCm platform and its updates without software licensing fees. This can lead to substantial cost savings. And because the AMD ROCm platform is compatible with a range of GPUs, developers have more options, which can lead to lower hardware costs.

TECHNICAL DEEP DIVE

#1 OPEN-SOURCE WITHOUT VENDOR LOCK-IN

- AMD has technical collaborations with many AI industry and open software leaders, including AI Alliance, Hugging Face, Lamini and OpenAI. AMD participates in open-source cross-platform initiatives such as MLIR, OpenMP®, OpenXLA, PyTorch, TensorFlow and Triton.
- The AMD ROCm™ software stack provides open and easy-to-use tools built around industry standards that enable creating optimized portable software. AMD ROCm software source code is published on GitHub, including drivers, tools and libraries. Build environment scripts (CMake) are available to compile source for target devices.
- A broad set of models are supported out of the box, and there are 1M+ models available on Hugging Face.

#2 DESKTOP TO DATACENTER GPU SUPPORT

- AMD facilitates the adoption and use of multiple AMD GPUs from the desktop to the datacenter, and cross-platform development with the ROCm software ecosystem and programming toolset.
- The AMD ROCm software ecosystem includes drop-in support for major AI and HPC frameworks, leading software platforms and models, and networking infrastructures to help ease system adoption and deployment.
- The AMD ROCm platform supports 1M+ AI models out of the box – no porting needed, and the community is active 24x7.

#3 VAST SOFTWARE DEVELOPMENT ECOSYSTEM

- Some of the most popular frameworks are part of the AMD ROCm software ecosystem, including those designed to parallelize operations across multiple GPUs, handle memory hierarchies, and solve for linear systems. This frictionless software ecosystem includes drop-in support for major frameworks and programming models like PyTorch, JAX, TensorFlow and ONNX, as well as compilers like Triton.
- The AMD GPU Accelerated Applications Catalog lists a broad set of supporting libraries for AI and ML along with a comprehensive set of platform-compatible HPC applications for use cases such as astrophysics, climate and weather, computational chemistry, computational fluid dynamics, genomics, molecular dynamics and physics. Many of these are available through the AMD Infinity Hub, ready to download and run.

#4 EASY PORTING

- HIP provides hipify-clang, a Clang-based tool to quickly convert CUDA source code into HIP-compatible source code, minimizing the need for manual changes to the existing codebase. This automation in porting can save significant time and effort.
- HIP offers a mapping of most CUDA APIs, allowing developers to easily migrate their CUDA code to the AMD ROCm platform without extensive rewrites or modifications for a smoother transition.
- The HIP compiler allows compilation of CUDA code for AMD GPUs with minimal changes. It acts as a thin layer over CUDA, so the same code can run on both NVIDIA and AMD GPUs. This flexibility in development and deployment allows for more versatile hardware choices and easier integration into various environments.

#5 COST-EFFECTIVENESS

- AMD continues to advance its ROCm software stack, bringing the latest features to support leading training and inference on generative AI workloads. The open AMD ROCm platform now includes support for critical AI features like FP8 datatype, Flash Attention 3, Kernel Fusion and more. With these new additions, AMD ROCm 6.2 software, compared to AMD ROCm 6.0 software, provides up to a 2.4X performance improvement on inference[MI300-62] and 1.8X on training for a variety of LLMs.[MI300-63]
- By avoiding vendor lock-in, developers can choose the most cost-effective hardware and software solutions for their specific needs. They can also protect existing investments in GPU-accelerated applications and infrastructure, as they are not dependent on a single vendor's support or roadmap.

END NOTES

For details on the claims used in this document, visit amd.com/en/legal/claims/instinct.

MI300-62
Testing conducted by internal AMD Performance Labs as of September 29, 2024 inference performance comparison between ROCm 6.2 software and ROCm 6.0 software on the systems with 8 AMD Instinct™ MI300X GPUs coupled with Llama 3.1-8B, Llama 3.1-70B, Mixtral-8x7B, Mixtral-8x22B, and Qwen 72B models.
ROCm 6.2 with vLLM 0.5.5 performance was measured against the performance with ROCm 6.0 with vLLM 0.3.3, and tests were performed across batch sizes of 1 to 256 and sequence lengths of 128 to 2048.

Configurations:
1P AMD EPYC™ 9534 CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, Supermicro AS-8125GS-TNMR2, NPS1 (1 NUMA per socket), 1.5 TiB (24 DIMMs, 4800 mts memory, 64 GiB/DIMM), 4x 3.49TB Micron 7450 storage, BIOS version: 1.8, , ROCm 6.2.0-00, vLLM 0.5.5, PyTorch 2.4.0, Ubuntu® 22.04 LTS with Linux kernel 5.15.0-119-generic. vs.
1P AMD EPYC 9534 CPU server with 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, Supermicro AS-8125GS-TNMR2, NPS1 (1 NUMA per socket), 1.5TiB 24 DIMMs, 4800 mts memory, 64 GiB/DIMM), 4x 3.49TB Micron 7450 storage, BIOS version: 1.8, ROCm 6.0.0-00, vLLM 0.3.3, PyTorch 2.1.1, Ubuntu 22.04 LTS with Linux kernel 5.15.0-119-generic.

Server manufacturers may vary configurations, yielding different results. Performance may vary based on factors including but not limited to different versions of configurations, vLLM, and drivers.

MI300-631
AMD Instinct MI300X platform (8x GPUs) with ROCm 6.2 running Llama2 7B, Llama2-70B, Qwen1.5-14B using Megatron-LM delivers a combined average 1.83x training performance uplift (83% higher) over AMD Instinct MI300X platform with ROCm 6.0 all with FP16 datatype. "Testing conducted by internal AMD Performance Labs as of September 29, 2024 training performance comparison between ROCm 6.2 software with compared to ROCm 6.0 software both with Megatron-LM on systems with 8 AMD Instinct™ MI300X GPUs running Llama 2-7B, Llama 2-70B (4K), Qwen1.5-14B models using custom docker container for each system.

ROCm 6.2 with megatron-LM TFLOPs was measured against the TFLOPs with ROCm 6.0 with megatron-LM.

Configurations:
CPU: 1P AMD EPYC 9454 48-core processor, ost memory: 2x3.5 T GB, GPU: AMD Instinct MI300X, 1P AMD EPYC™ 9454 CPU, 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, American Megatrends International LLC BIOS version: 1.8, ROCm 6.2 internal release, Megatron-LM code branches hanl/disable_te_llama2 for Llama 2-7B, guihong_dev for LLama 2-70B, renwuli/disable_te_qwen1.5 for Qwen1.5-14B, PyTorch 2.4, Ubuntu 22.04 LTS with Linux kernel 5.15.0-117-generic.
vs.
1P AMD EPYC 9454 CPU 48-core processor, 8x AMD Instinct™ MI300X (192GB, 750W) GPUs, American Megatrends International LLC BIOS version: 1.8, ROCm 6.0.0, Megatron-LM code branches hanl/disable_te_llama2 for Llama 2-7B, guihong_dev for LLama 2-70B, renwuli/disable_te_qwen1.5 for Qwen1.5-14B, PyTorch 2.2, Ubuntu 22.04 LTS with Linux kernel 5.15.0-72-generic Server manufacturers may vary configurations, yielding different results. Performance may vary based on factors including but not limited to different versions of configurations, megatron-LM, and drivers.
Results: MI300X with ROCm 6.2 delivers average 1.83X the (83% higher) training throughput than ROCm 6.0."

AMD INSTINCT™ ACCELERATORS
TOGETHER WE ADVANCE_

AMD.COM/EN/ROCM

DISCLAIMER: The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes. GD-18.