SOLUTION BRIEF

AMD ROCM™ 7 SOFTWARE SOLUTION GUIDE FOR AMD INSTINCT™ GPUs

AMD ROCm

AMD ROCm™ 7 software is a next-generation open-source platform optimized to extract the best performance from modern GPUs for **AI and HPC workloads**, while maintaining seamless compatibility with industry-standard frameworks. It consists of a rich collection of drivers, libraries, and developer tools that enable GPU programming from low-level kernels up to end-user applications, and it's customizable to meet specific needs. With ROCm 7, teams can develop, collaborate, test, and deploy in a free, open ecosystem. Once a workload is optimized on ROCm 7 software, it is portable across different accelerators and interconnect architectures in a **vendor-agnostic, device-independent** manner. ROCm 7 is well-suited for GPU-accelerated high-performance computing (HPC), artificial intelligence (AI), scientific computing, and more. A wide range of pre-built GPU software containers and deployment guides (available via the AMD Infinity™ Hub) help speed up your deployments and time-to-insight.

OPTIMIZE NEXT GEN INNOVATION WITH ROCm SOFTWARE

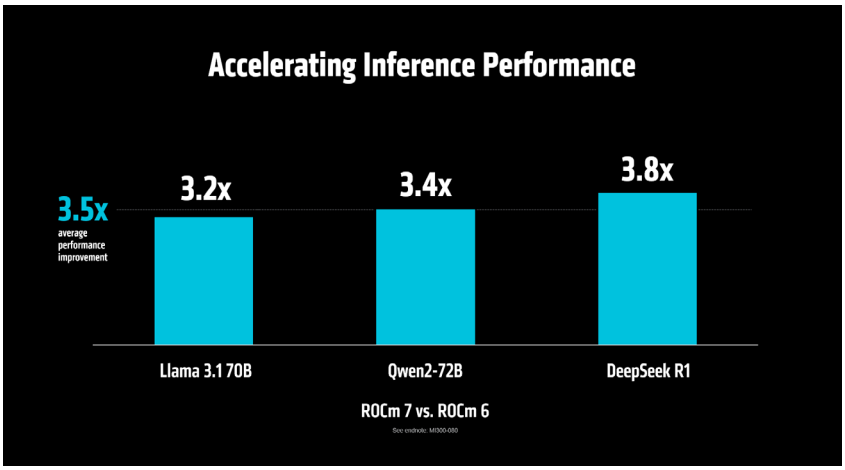
ROCm 7 delivers a highly optimized kernel library and unified compiler stack, powered by **Triton v3.3**, a kernel compiler generating efficient kernels from one codebase across NVIDIA (CUDA), AMD (HIP), and Intel GPUs. Fused implementations—such as flash attention and other critical Transformer operations in a preview version of ROCm 7—accelerate core AI operations higher than compared to traditional kernels[1]. ROCm 7 also introduces advanced low-precision support (**MXFP8, OCP-FP8, MXFP6, and MXFP4 datatypes**) for further AI throughput optimization. A preview version of ROCm 7 software showed an average **3.5× higher inference throughput performance on AMD InstinctTM MI300X 8x GPU platform**[2] and up to **3× faster training** versus AMD ROCm 6 software[3], driven primarily by software enhancements like improved GPU utilization and compute-communication overlap.

WHAT'S NEW IN ROCm 7.0?

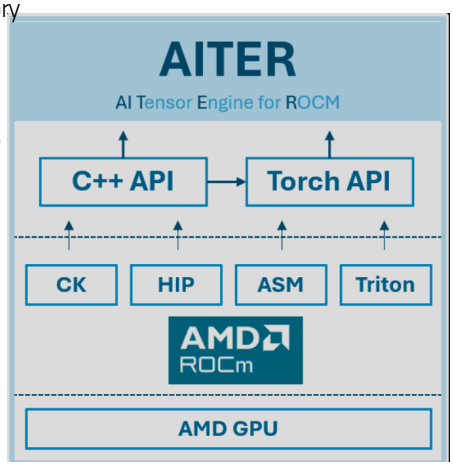
1. **Expanded Hardware & Platform Support:** ROCm 7 adds full support for the latest **AMD Instinct™ MI350 Series** GPUs (including new MXFP6/MXFP4 low-precision capabilities) and their AI Tensor Engines. It also expands beyond the data center – enabling development on select **AMD Radeon™ GPUs and Windows** environments for a consistent experience from cloud to edge. ROCm software continues to run across heterogeneous hardware without changes, preserving the project's open, multi-vendor approach.

2. **Key AI Features & Optimizations:** New optimizations target large-scale **AI** and **LLM** deployments. ROCm 7 incorporates pre-optimized transformer engine kernels (via the AI Tensor Engine) to leverage OCP-FP8/MXFP8/MXFP6/MXFP4 precisions, greatly increasing throughput with minimal accuracy loss.
   It introduces robust distributed inference support through close integration with frameworks like **vLLM v1**, **llm-d**, and **SGLang**, co-developed to enable efficient multi-GPU and multi-node inference serving. Additionally, highly optimized "flash" attention algorithms and improved collective communication libraries (NCCL/RCCL, abstracted by the new NIXL layer) ensure maximal GPU utilization for large models by reducing overhead and eliminating bottlenecks.

3. **Optimized Performance:** A preview version of ROCm 7 demonstrated a dramatic performance leap for AI workloads, with **up to 3.5× higher AI inferencing throughput**[2] **and up to 3× faster training** compared to ROCm 63. By exploiting lower-precision data types and advanced kernel fusion, ROCm 7 preview version squeezes more useful work out of each GPU cycle while lowering memory and I/O strain.



The **AI Tensor Engine** for **ROCm software (AITER)** is designed to significantly accelerate key AI operations—including GEMM, attention mechanisms, and Mixture-of-Experts—by providing drop-in, pre-optimized Triton kernels for the new AMD Instinct MI350 Series GPUs. AITER eliminates the complexity and overhead of manual kernel tuning, enabling immediate performance gains such as **faster decoder execution1, higher Multi-Headed Attention performance**[1], and throughput improvement for large language model inference1. Enhanced precision modes like **FP16/BF16, OCO-FP8, MXFP8,** and **INT8**, along with fused operations (e.g., Fused-MoE), are designed to reduce memory footprints and training time while maintaining accuracy. AITER also supports PyTorch with new **MXFP8/OCP-FP8** and **MXFP4** datatypes, simplifying quantized model training and inference without custom kernels. These capabilities enable maximum throughput, minimal latency, and efficient memory use, especially critical for generative AI and transformer-based workloads on the new AMD Instinct MI350 Series GPUs.

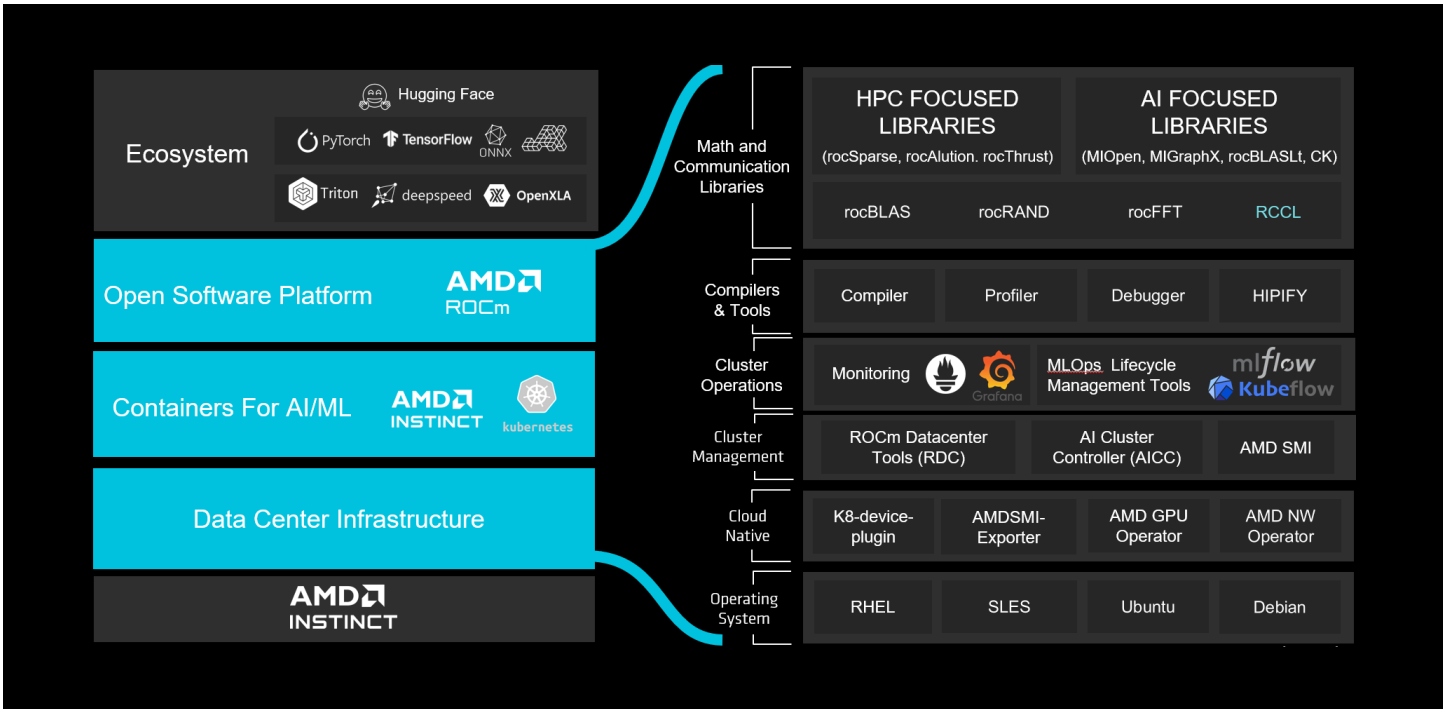


Built for elastic scalability, ROCm 7 software incorporates robust orchestration frameworks—including **vLLM**, **llm-d**, and **SGLang**—developed with open-source community collaboration to seamlessly scale large AI models across multiple GPUs and nodes. Dynamic batching and automated pipeline/tensor parallelism keep GPU utilization high while maintaining strict latency targets. The ROCm 7 software scheduler dynamically balances workloads, cleverly overlaps compute and data transfers (**DeepEP engine**), and supports elastic cluster resizing. The platform introduces optimized **PD KV-Cache** transfer to handle large language models, efficiently batching and streaming token cache data between GPUs via PCIe®, or RDMA to achieve lower latency and prevent decoding stalls. As a result, ROCm 7 provides scalable, low-latency inference for even multi-hundred-billion parameter models distributed across large GPU clusters.

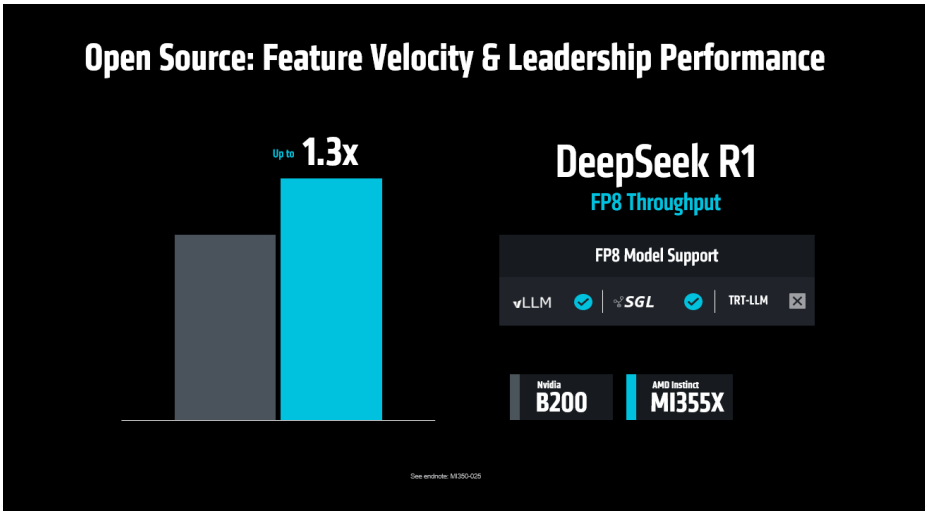
4. **Enabling Developer Success:** User experience is a major focus in ROCm 7. A new **ROCm Enterprise AI** suite debut as an MLOps platform for enterprise users, making it easier to fine-tune models on domain-specific data and deploy AI services in production. Installation and setup have been streamlined – for instance, developers can now get started with a simple pip install rocm flow, going from zero to a running GPU workload in no time. ROCm 7 also continues to support advanced optimization features like model **quantization** libraries, enabling developers to further compress models or improve throughput without extensive manual effort. From new profiling tools to integrated workflow scripts, ROCm 7 is built to help engineers and data scientists be productive quickly.

5. **Expanded Ecosystem & Community Collaboration:** ROCm 7 strengthens its ties with the broader AI and HPC ecosystems. It offers improved support for many of the latest models and frameworks – ensuring day-0 compatibility with new releases of PyTorch, TensorFlow, JAX, ONNX and more. Over **2 million pre-trained models** from community hubs like Hugging Face can be directly loaded and accelerated on ROCm 7, giving organizations unparalleled flexibility in model selection. The AMD open-source commitment means that many ROCm innovations (kernel optimizations, compiler improvements, etc.) are upstreamed to projects like **ONNX**, **MLIR**, **DeepSpeed**, and others, benefiting the community at large. This broad ecosystem support and collaborative development ensure that ROCm 7 remains **stable, forward compatible, and ready for cutting-edge workloads** as they emerge.

Additionally, ROCm 7 supports all leading AI frameworks—including **PyTorch**, **TensorFlow**, **JAX**, and **ONNX Runtime**—and a broad range of tools such as **DeepSpeed and CuPy**, with comprehensive model compatibility through the Hugging Face Hub. Emerging models like Llama 4 and Gemma 3 run out-of-the-box on day one, removing the typical porting or manual tuning burdens. **AMD Infinity™ Hub** further streamlines deployment, providing containerized, ready-to-run frameworks and models for rapid prototyping and production deployment. ROCm 7 software helps ensure that organizations can effortlessly maximize AI performance and minimize friction, from single-server workloads to enterprise-scale deployments.



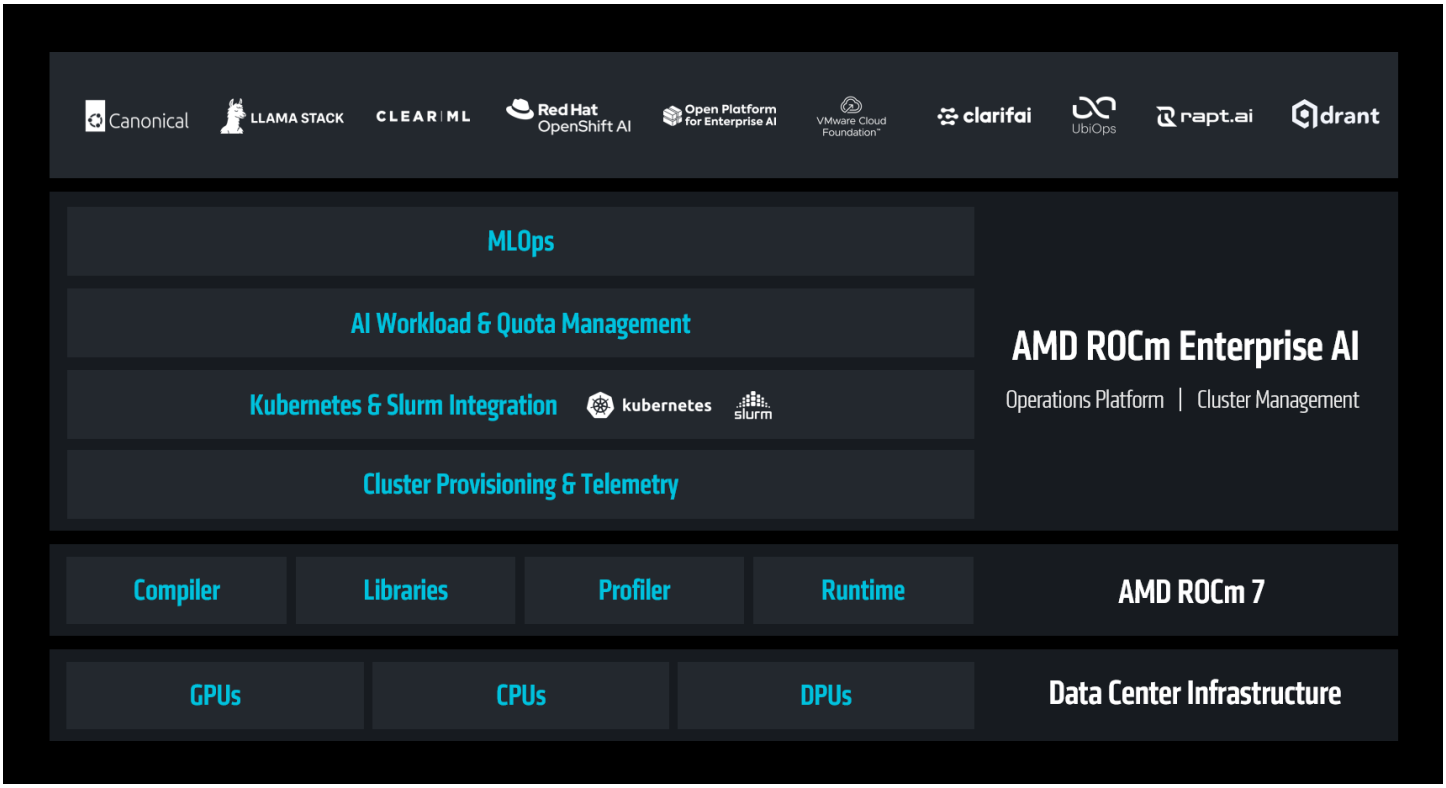
ACCELERATE YOUR HIGH-PERFORMANCE COMPUTING WORKLOADS



Many of the most popular HPC frameworks are integrated into the ROCm platform, including tools to parallelize workloads across multiple GPUs and nodes, handle complex memory hierarchies, and solve large-scale linear algebra problems. The **ROCm GPU Application Catalog** encompasses a vast set of HPC applications across domains – astrophysics, climate and weather modeling, computational chemistry, fluid dynamics, earth science, genomics, geophysics, molecular dynamics, physics, and more. Researchers and engineers can readily deploy these optimized applications on AMD Instinct GPUs, with many containerized solutions available on **AMD Infinity Hub** for convenience. Using DeepSeekR1, it is possible to achieve up to a 1.3x increase in FP8 throughput with a preview version of ROCm 7.0, comparing the AMD Instinct MI355X 8x GPU platform to Nvidia B200 8x GPU platform4, with support for vLLM and SGLang. HPC developers and scientific computing teams can benefit from the flexibility of an open, portable software stack – avoiding vendor lock-in and ensuring their codes run efficiently on diverse hardware.

REAL-WORLD IMPACT AND USE CASES

ROCm 7 software's blend of AI-focused optimizations and HPC-grade scalability unlocks a wide range of applications. For example, enterprises can deploy real-time conversational AI (chatbots) serving thousands of users with faster response times, thanks to ROCm 7 software's efficient batching and GPU utilization. Streaming recommendation systems can score content for millions of users with millisecond latency by leveraging micro-batch inferencing. Research labs and universities can accelerate scientific simulations or train open-source LLMs on large public datasets – using tools that run on national supercomputers, such as El Capitan at Lawrence Livermore National Laboratory and Frontier at Oak Ridge National Laboratory, using AMD ROCm™ 6.0 software. Critically, ROCm software's continuous integration and testing enables reliability in production. This means organizations can adopt ROCm 7 with confidence that it will remain stable, up-to-date with the latest models, and optimized for any new hardware that comes along.



DEPLOYMENT & DOCUMENTATION RESOURCES

To explore ROCm 7.0 or deploy it in your environment, the following resources will be helpful:

- **AMD ROCm Official Documentation** – Comprehensive guides, release notes, and tutorials for installing and using ROCm 7.
- **AMD ROCm Technical Blog** – Comprehensive review of all libraries and frameworks in ROCm 7.0
- **AMD Infinity Hub** – A catalog of ready-to-run containerized applications and frameworks (for both HPC and AI) optimized for ROCm, along with deployment guides.
- **ROCm Application Catalog** – A detailed list of libraries and applications supported on the ROCm platform across various domains (AI, scientific computing, etc.).
- **ROCm Developer Hub** – Portal for developers featuring training materials, webinars, community forums, and best practices for ROCm.

For more information, visit **amd.com/ROCm** and join the open AI/HPC community driving ROCm forward.

ENDNOTES

1. **AITER Kernel Level Blog**
2. (MI300-080): Testing by AMD as of May 15, 2025, measuring the inference performance in tokens per second (TPS) of AMD ROCm 6.x software, vLLM 0.3.3 vs. AMD ROCm 7.0 preview version SW, vLLM 0.8.5 on a system with (8) AMD Instinct MI300X GPUs running Llama 3.1-70B (TP2), Qwen 72B (TP2), and Deepseek-R1 (FP16) models with batch sizes of 1-256 and sequence lengths of 128-204. Stated performance uplift is expressed as the average TPS over the (3) LLMs tested. Results may vary.
3. (MI300-081): Testing conducted by AMD Performance Labs as of May 15, 2025, measuring the training performance (TFLOPS) of ROCm 7.0 preview version software, Megatron-LM, on (8) AMD Instinct MI300X GPUs running Llama 2-70B (4K), Qwen1.5-14B, and Llama3.1-8B models, and a custom docker container vs. a similarly configured system with AMD ROCm 6.0 software.
4. (MI350-025) Testing by AMD Performance Labs as of May 25, 2025, measuring the inference performance in tokens per second (TPS) of the AMD Instinct MI355X platform with ROCm 7.0 pre-release build 16047, running DeepSeek R1 LLM on SGLang versus NVIDIA Blackwell B200 platform with CUDA version 12.8. Server manufacturers may vary configurations, yielding different results. Performance may vary based on hardware configuration, software version, and the use of the latest drivers and optimizations.

   Additional Hardware Configuration(s) 2P AMD EPYC™ 9575F CPU server with 8x AMD Instinct™ MI355X (288GB, 1400W) GPUs, Supermicro AS-4126GS-NMR0LCC, 3 TiB (24 D IMMs, 6400 mts memory, 128 GiB/DIMM), 2x 3.49TB Micron 7450 storage, BIOS version: 1.4a. 2P Intel Xeon 6972P CPU server with 8x NVIDIA B200 (180GB, 1000W) GPUs, Supermicro SYS-A22GA-NBRT, 2.95 TiB (24 DIMMs, 4800 mts memory, 128 GiB/DIMM), 2x 3.5 TB Micron 7450 storage, BIOS version: 1.8.

   Additional Software Configuration(s) Ubuntu 22.04 LTS with Linux kernel 6.8.0-59-generic, ROCm 7.0.0 (pre-release build 16047) + amdgpu 6.14.5 (build 2168543) Pre-release Docker: rocm/aigmodels private:experimental_950_5_26 (cache off, --chunked prefill size 131072, torch compile), TP8+DP8 vs. Ubuntu 22.04.5 LTS with Linux kernel 5.15.0-72-generic, Driver Version: 570.133.20 CUDA Version: 12.8 Public Docker: lmsysorg/sglang:blackwell.

DISCLAIMERS

COPYRIGHT NOTICE