SOLUTION BRIEF

AMD ROCM™ SOFTWARE STACK SOLUTIONS BRIEF FOR AMD INSTINCT™ GPUS

AMD ROCm

OPEN, SCALABLE AI SOFTWARE FOR PRODUCTION INFRASTRUCTURE

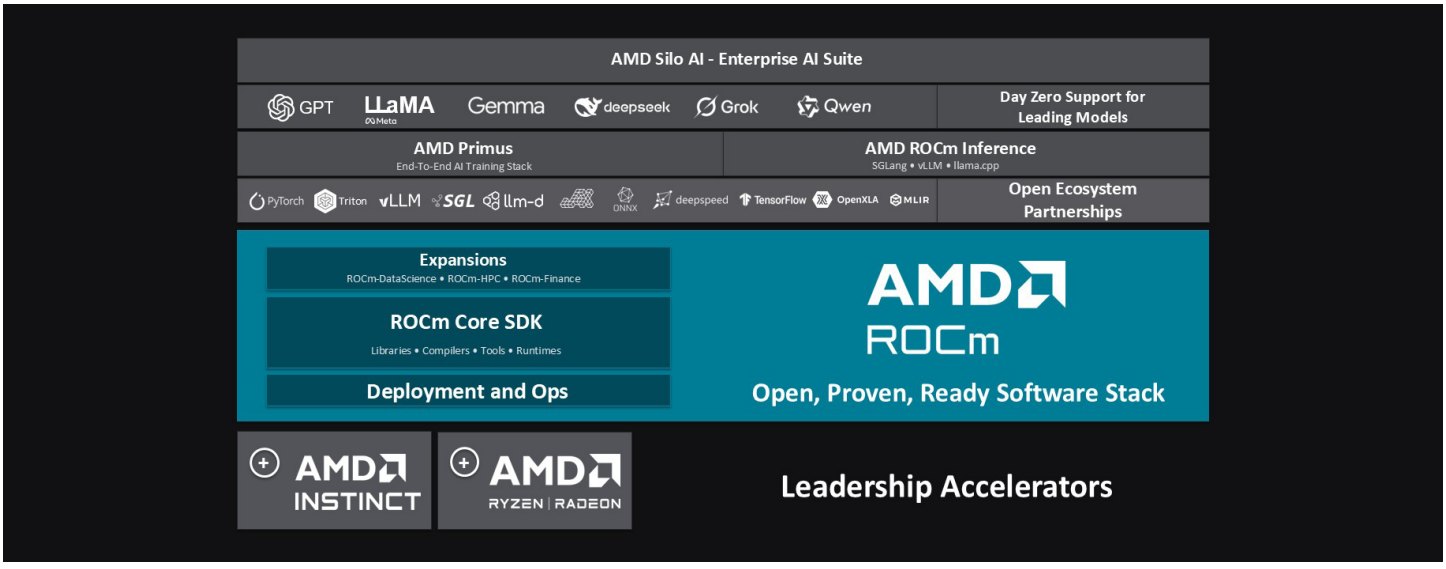
AMD ROCm™ software enables organizations to deploy high-performance AI and HPC workloads with confidence—without vendor lock-in. Built on an open, exascale-proven foundation, ROCm converts low-level kernel and compiler innovation into production-ready results that scale from single-GPU experimentation to large, multi-node clusters.

For AI and infrastructure leaders, ROCm delivers three critical advantages: architectural freedom through open standards, faster time-to-value via day-0 frameworks and validated platforms, and predictable operations at datacenter scale. Teams gain tuned math, graph, and communication libraries; validated training and inference engines; and Kubernetes-native lifecycle tooling—all designed to reduce integration risk while preserving long-term portability and maintainability.

Fast onboarding through validated servers and containers, reduced optimization effort via fused kernels and mixed-precision support, and confident operations enabled by automation, telemetry, and modular upgrades combine into an AI and HPC platform ready for speedy deployments. The same ROCm stack that powers leadership systems such as Frontier and El Capitan brings exascale rigor, reproducibility, and scale-out discipline directly into AI customer environments.

A MODULAR, OPEN ARCHITECTURE BUILT TO SCALE

AMD ROCm™ spans layered capabilities that can be adopted end-to-end or integrated selectively: frameworks and serving engines for training and inference; a compiler and graph layer including Triton-based kernel generation; core math and communication libraries for dense, sparse, FFT, and collective operations; and deployment and operations controls engineered for enterprise environments. This architecture enables organizations to scale performance without locking into brittle, per-model tuning or proprietary software paths. With AMD ROCm™ 7, the stack strengthens this end-to-end foundation through day-0 framework readiness with more than 2M+ Models to choose from, and architecture-specific packaging that enables lean, reproducible deployments, while expanding optimized building blocks such as drop-in transformer kernels and improved collective reliability to support production-scale training and distributed inference. Together, these ROCm™ 7 advancements reduce integration risk and accelerate the path from evaluation to sustained cluster-scale operations.



OPENNESS: FREEDOM, PORTABILITY, AND UPSTREAM COLLABORATION

AMD ROCm™ is built upstream-first to deliver performance without lock-in—meeting developers where they already work while giving organizations the freedom to evolve their infrastructure over time. Day-0 support for PyTorch, TensorFlow, and JAX, along with validated inference engines such as vLLM and SGLang, reduces friction for training and serving, while architecture-specific wheels and containers enable lean, reproducible deployments aligned with enterprise CI/CD practices.

At the kernel and runtime level, ROCm converts openness into measurable performance. The AI Tensor Engine for ROCm (AITER) provides drop-in Triton kernels for core transformer operations—including GEMM, attention, and Mixture-of-Experts—delivering immediate throughput gains without bespoke tuning. Stream-K improves GEMM utilization through automatic load balancing, while HIP 7.0 strengthens portability and toolchain stability so teams can maintain a single codebase across AMD Instinct™ GPUs. Mixed-precision execution, fused transformer primitives, and optimized math, vision, media, and HPC libraries support high-throughput, predictable-latency AI and compute workloads.

To accelerate adoption, AMD Infinity™ Hub offers ready-to-run containers for LLMs, scientific codes, and domain applications, while AMD Quark provides production-ready FP8 and MXFP4 model packages—shortening the path from evaluation to deployment across AI and HPC environments.

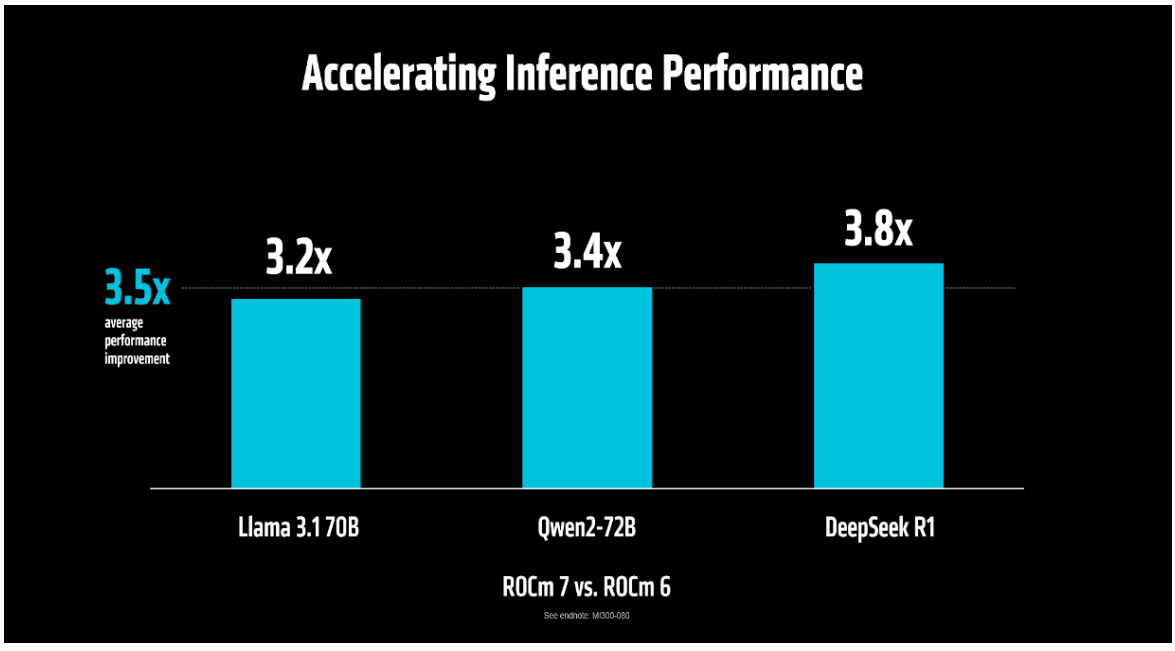
LEADERSHIP PERFORMANCE: FROM KERNELS TO CLUSTERS

AMD ROCm™ software delivers leadership-class AI and HPC performance by translating low-level GPU innovation into reproducible, production-scale results. Designed from kernels to clusters, ROCm unifies optimized execution, low-precision math, and scale-out communication so training and inference sustain high utilization as workloads scale from early experiments to large, multi-node deployments.

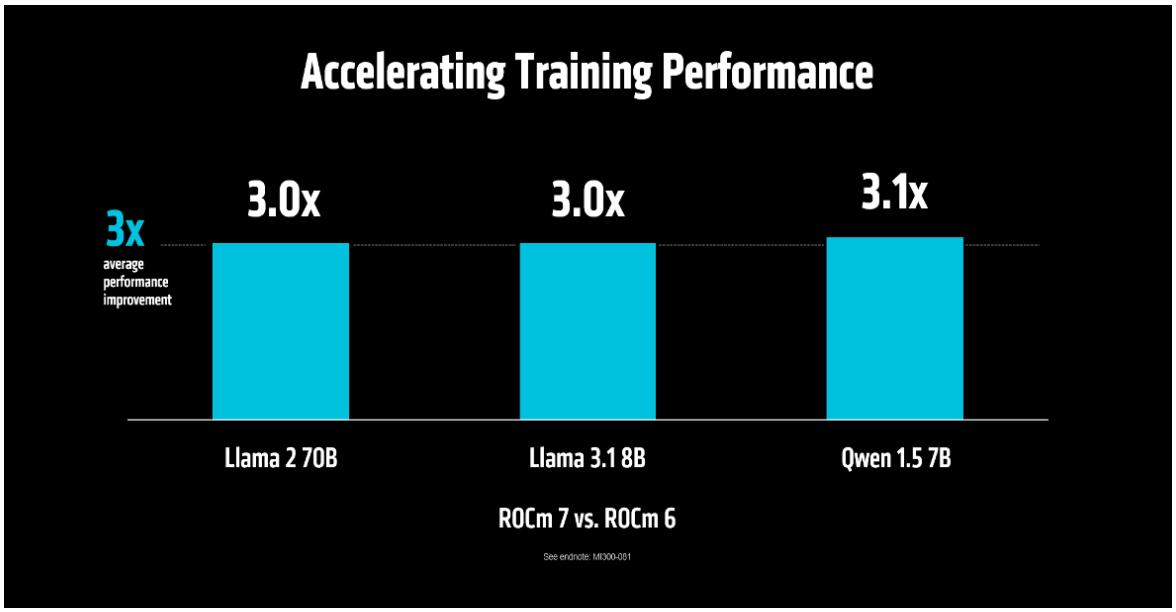
Built upstream-first and developed in the open, ROCm pairs day-0 integration for mainstream frameworks and inference engines with optimized kernels and collective communication—enabling high throughput without brittle, per-model tuning or proprietary lock-in. This performance discipline is proven at extreme scale: ROCm underpins leadership systems such as Frontier and El Capitan, demonstrating that the same open software foundation delivers consistent, high-performance execution from national supercomputers to large customer AI clusters.

PERFORMANCE MOMENTUM

In internal testing of a preview version of ROCm 7.0 on an 8× AMD Instinct™ MI300X platform running Llama 3.1-70B, Qwen 72B, and DeepSeek-R1 with vLLM, these software advancements contributed to an average ~3.5×[1] increase in inference throughput.



Also these advancements helped achieve up to ~3×[2] higher training performance versus an internal version of ROCm 6 using Megatron-LM on comparable models. These results illustrate how ROCm software improvements translate directly into end-to-end performance gains, with actual performance dependent on system configuration and workload.



OPERATE AT FLEET SCALE: KUBERNETES-READY, OBSERVABLE, GOVERNABLE

AMD ROCm™ is engineered for organizations that operate AI as critical infrastructure—where uptime, scale, and cost efficiency matter as much as raw performance. ROCm integrates the operator-grade controls required to deploy and manage large GPU fleets reliably, while supporting distributed inference and training across nodes. Kubernetes alignment through the AMD GPU Operator automates driver and runtime rollout, supports Red Hat OpenShift and Ubuntu, and lets teams do safe rolling upgrades—even in restricted networks. A decoupled AMD GPU driver model separates kernel and user space, enabling predictable lifecycle management, compatibility control, and safer upgrades at fleet scale.

For platform and Site reliability engineering teams, ROCm delivers full-stack observability and repeatable operations. Device Metrics Exporter surfaces Prometheus-ready telemetry for ECC events, utilization, power, and health, while AMD Resource Manager and AMD AI Workbench add orchestration, quota management, and standardized development pipelines that connect notebooks to production fleets. At the same time, ROCm validates cluster-scale inference engines so multi-GPU serving becomes routine rather than bespoke: SGLang enables prefill/decode disaggregation and optimized KV-cache movement, vLLM delivers a v1 engine tuned for lower time-to-first-token and higher throughput, and RCCL provides the collective communication backbone that scales serving efficiently across GPUs and nodes.

For organizations seeking a supported, production-ready path, AMD Enterprise AI Suite builds on ROCm with a curated, validated set of frameworks, inference engines, and management tools designed for enterprise deployment. Enterprise AI Suite reduces integration effort and operational risk by delivering tested software combinations, predictable update paths, and commercial support—allowing teams to move from pilot to production more quickly while retaining the openness and performance advantages of the ROCm platform.

EXASCALE PROVEN: TRUSTED FOR SOVEREIGN AI AND HPC

ROCm provides a continuous path from science-grade computation to leadership-class AI—without switching software stacks. Organizations can begin with FP64 simulation and established HPC workflows, then integrate AI-assisted discovery using the same libraries, communication layers, and operational patterns.

This approach is proven where reliability and scale are non-negotiable. Leadership systems such as Frontier and El Capitan run on AMD ROCm™, and the AMD collaboration with the U.S. Department of Energy—including next-generation systems at Oak Ridge National Laboratory—extends ROCm's exascale discipline into future sovereign AI and HPC infrastructure.

For enterprises, this matters because the same software rigor, validation processes, and scale-out capabilities used in national laboratories are available in commercial ROCm deployments. Teams gain confidence that their AI platforms can scale, evolve, and remain maintainable over multi-year infrastructure lifecycles.

START NOW WITH INFINITY HUB

Teams can begin with AMD ROCm™ documentation for installation and release notes, then pull containers from AMD Infinity™ Hub or the ROCm application catalog to validate workflows quickly on AMD Instinct™ GPUs. For enterprise rollouts, the AMD GPU Operator streamlines cluster enablement while AMD Resource Manager and AMD AI Workbench introduce orchestration and developer workflows that connect pilots to production.

- **AMD ROCm Official Documentation** – Comprehensive guides, release notes, and tutorials for installing and using ROCm.
- **AMD ROCm Technical Blog** – Comprehensive review of all libraries and frameworks in ROCm 7.0 (Release Specific)
- **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.

ENDNOTES

1. (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), Qwen2 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.
2. (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), Qwen1.5-7B, and Llama3.1-8B models, and a custom docker container vs. a similarly configured system with AMD ROCm 6.0 software.

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 (MI350-025)

DISCLAIMERS

COPYRIGHT NOTICE