

# AMD ROCm™ 7 SOFTWARE: 5 REASONS WHY FOR AMD INSTINCT™ GPUs



## AT A GLANCE

AMD ROCm™ 7.0 is the open GPU software stack purpose-built for the generative AI and HPC era. With enablement for AMD Instinct™ MI350 series accelerators and their AI Tensor Engine, ROCm™ 7.0 adds FP4/FP6/FP8 precision (via AMD Quark) and optimized kernels (AITER, Stream-K) to drive higher throughput and lower latency. It brings distributed inference to life through native integrations with vLLM and SGLang (including prefill-decode disaggregation) while delivering day-0 support for leading frameworks like PyTorch, TensorFlow, and JAX. For enterprise rollout, AMD Resource Manager and AMD AI Workbench streamline orchestration, quotas, telemetry, and reproducible pipelines from development to production. Combined with HIP 7.0 portability and an open-source philosophy, ROCm™ 7.0 lets engineering and IT teams train and serve large models at scale—without lock-in and with the confidence to move fast.

### 1 EXPANDED HARDWARE & PLATFORM SUPPORT:

Full enablement of AMD Instinct™ MI350 GPUs (FP4/FP6/FP8, AI Tensor Engine), plus support for Radeon GPUs + Windows; consistent, open, multi-vendor platform.

ROCm 7 adds full support for the latest **AMD Instinct™ MI350 Series GPUs** (including new FP6/FP4 low-precision capabilities) and their AI Tensor Engines. It also expands beyond the data center – enabling development on select **AMD Radeon™ GPU** and **Windows** environments for a consistent experience from cloud to edge. AMD ROCm™ software continues to run across heterogeneous hardware without changes, preserving the project's open, multi-vendor approach.

### 2 NEW AI FEATURES & OPTIMIZATIONS:

Pre-optimized transformer kernels, distributed inference via vLLM/SGLang/llm-d, and enhanced GPU utilization through Flash Attention and improved RCCL implementations.

New optimizations target large-scale **AI** and **LLM** deployments. ROCm 7 incorporates pre-optimized transformer engine kernels (via the AI Tensor Engine) to leverage FP8/FP6/FP4 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 ensure maximal GPU utilization for large models by reducing overhead and eliminating bottlenecks.

## AMD ROCm™ 7.0 TL;DR SUMMARY

### 1. EXPANDED HARDWARE & PLATFORM SUPPORT –

Full enablement of AMD Instinct™ MI350 GPUs (FP4/FP6/FP8, AI Tensor Engine), plus support for Radeon GPUs + Windows; consistent, open, multi-vendor platform.

### 2. NEW AI FEATURES & OPTIMIZATIONS –

Pre-optimized transformer kernels, distributed inference via vLLM/SGLang/llm-d, and upgrades like flash attention + improved RCCL to maximize GPU utilization.

### 3. GENERATIONAL PERFORMANCE UPLIFT –

ROCm 7 preview testing shows up to **3.5× faster inference<sup>1</sup>** and **3.0× faster training<sup>2</sup>** vs. ROCm 6, powered by lower-precision formats and advanced kernel fusion.

### 4. DEVELOPER & ENTERPRISE READY –

New Enterprise AI suite with Resource Manager + AI Workbench, streamlined installs (pip flows, smaller Dockers), quantization support, and better profiling tools.

### 5. OPEN ECOSYSTEM & COMMUNITY COLLABORATION –

Day-0 support for PyTorch, TensorFlow, JAX, ONNX, plus access to 2M+ Hugging Face models; open-source contributions upstreamed to ONNX, MLIR, DeepSpeed, and more.

### 3 OPTIMIZED GENERATIONAL PERFORMANCE:

ROCm 7 preview testing shows up to **3.5× faster inference**<sup>1</sup> and **3.0× faster training**<sup>2</sup> vs. ROCm 6, powered by lower-precision formats and advanced kernel fusion.

A preview version of ROCm 7 demonstrated a dramatic performance leap for AI workloads, with **up to 3.5× higher AI inferencing throughput**<sup>1</sup> and **up to 3.0× faster training** compared to ROCm 6<sup>2</sup>. 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.

### 4 ENABLING DEVELOPER & ENTERPRISE:

New **Enterprise AI suite** with Resource Manager + AI Workbench, streamlined installs (pip flows, smaller Dockers), quantization support, and better profiling tools.

User experience is a major focus in ROCm 7. A new **ROCm Enterprise AI** suite debuts 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 OPEN ECOSYSTEM & COMMUNITY COLLABORATION:

Day-0 support for PyTorch, TensorFlow, JAX, ONNX, plus access to 2M+ Hugging Face models; open-source contributions upstreamed to ONNX, MLIR, DeepSpeed, and more.

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.

## 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), 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 an average TPS over the (3) LLMs tested. Server manufacturers may vary configurations, yielding different results. Performance may vary based on configuration, software, vLLM version, and the use of the latest drivers and optimizations.

2. MI300-081: Testing by AMD 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. Server manufacturers may vary configurations, yielding different results. Performance may vary based on configuration, software, vLLM version, and the use of the latest drivers and optimizations.

## COPYRIGHT NOTICE

© 2025 Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur laoreet odio non tortor maximus ultricies. Ut tellus ligula, interdum et ligula eget, egestas varius eros. Ut libero ante, consectetur id neque sit amet, porta rhoncus lorem. Cras eu fringilla elit. Nam lobortis in velit vel sagittis. Sed tempus dolor at urna ultrices fermentum. Aenean lacinia tellus in ullamcorper dictum. Nam sit amet scelerisque eros. (Footer Text)