Red Hat AI Inference Server 3.2

Supported product and hardware configurations

Supported hardware and software configurations for deploying Red Hat AI Inference Server

Last Updated: 2025-07-22

Red Hat AI Inference Server 3.2 Supported product and hardware configurations

Supported hardware and software configurations for deploying Red Hat AI Inference Server

Legal Notice

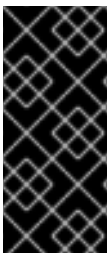
Abstract

Learn about supported hardware and software configurations for Red Hat AI Inference Server.

Table of Contents

PREFACE

This document describes the supported hardware, software, and delivery platforms that you can use to run Red Hat AI Inference Server in production environments.



IMPORTANT

Technology Preview and Developer Preview features are provided for early access to potential new features.

Technology Preview or Developer Preview features are not supported or recommended for production workloads.

Additional resources

- Red Hat AI Inference Server documentation
- Red Hat AI on Hugging Face
- LLM Compressor techniques

CHAPTER 1. PRODUCT AND VERSION COMPATIBILITY

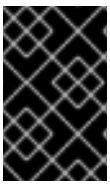
The following table lists the supported product versions for Red Hat AI Inference Server 3.2.

Table 1.1. Product and version compatibility

Product	Supported version
Red Hat AI Inference Server	3.2
vLLM core	v0.9.2
LLM Compressor	v0.6.0

CHAPTER 2. SUPPORTED AI ACCELERATORS

The following tables list the supported AI data center grade accelerators for Red Hat AI Inference Server 3.2.



IMPORTANT

Red Hat AI Inference Server only supports data center grade accelerators.

Red Hat AI Inference Server 3.2 is not compatible with CUDA versions lower than 12.8.

Table 2.1. Supported NVIDIA AI accelerators

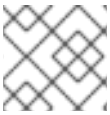
Container image	vLLM release	AI accelerators	Requirements	vLLM architecture support	LLM Compressor support
rhaiis/vllm-cuda-rhel9	vLLM v0.9.2	NVIDIA data center GPUs: • Turing: T4 • Ampere: A2, A10, A16, A30, A40, A100 • Ada: L4, L40, L40S • Hopper: H100, H200, GH200 • Blackwell: B200, RTX PRO 6000 Blackwell Server Edition	• CUDA Toolkit 12.8 • NVIDIA Container Toolkit 1.14 • NVIDIA GPU Operator 24.3 • Python 3.12 • PyTorch 2.7.0	• x86 • Aarch64 Developer Preview	x86 Technology Preview

NOTE

NVIDIA T4 and A100 accelerators do not support FP8 (W8A8) quantization.

Table 2.2. Supported AMD AI accelerators

Container image	vLLM release	AI accelerators	Requirements	vLLM architecture support	LLM Compressor support
rhaiis/vllm-rocm-rhel9	vLLM v0.9.2	• AMD Instinct MI210 • AMD Instinct MI300X	• ROCm 6.2 • AMD GPU Operator 6.2 • Python 3.12 • PyTorch 2.7.0	x86	x86 Technology Preview



NOTE

AMD GPUs support FP8 (W8A8) and GGUF quantization schemes only.

Table 2.3. Google TPU AI accelerators (Developer Preview)

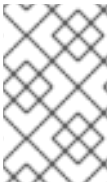
Container image	vLLM release	AI accelerators	Requirements	vLLM architecture support	LLM Compressor support
rhaiis/vllm-xla-rhel9	vLLM v0.8.5	Google TPU v6e	• Python 3.12 • PyTorch 2.7.0	x86 Developer Preview	Not supported

CHAPTER 3. SUPPORTED DEPLOYMENT ENVIRONMENTS

The following deployment environments for Red Hat AI Inference Server are supported.

Table 3.1. Red Hat AI Inference Server supported deployment environments

Environment	Supported versions	Deployment notes
OpenShift Container Platform (self-managed)	4.14 – 4.19	Deploy on bare-metal hosts or virtual machines.
Red Hat OpenShift Service on AWS (ROSA)	4.14 – 4.19	Requires ROSA STS cluster with GPU-enabled P5 or G5 node types.
Red Hat Enterprise Linux (RHEL)	9.2 – 10.0	Deploy on bare-metal hosts or virtual machines.
Linux (not RHEL)	-	Supported under third-party policy deployed on bare-metal hosts or virtual machines. OpenShift Container Platform Operators are not required.
Kubernetes (not OpenShift Container Platform)	-	Supported under third-party policy deployed on bare-metal hosts or virtual machines.



NOTE

Red Hat AI Inference Server is available only as a container image. The host operating system and kernel must support the required accelerator drivers. For more information, see Supported AI accelerators.

CHAPTER 4. OPENSHIFT CONTAINER PLATFORM SOFTWARE PREREQUISITES FOR GPU DEPLOYMENTS

The following table lists the OpenShift Container Platform software prerequisites for GPU deployments.

Table 4.1. Software prerequisites for GPU deployments

Component	Minimum version	Operator
NVIDIA GPU Operator	24.3	NVIDIA GPU Operator OLM Operator
AMD GPU Operator	6.2	AMD GPU Operator OLM Operator
Node Feature Discovery [1]	4.14	Node Feature Discovery Operator

[1] Included by default with OpenShift Container Platform. Node Feature Discovery is required for scheduling NUMA-aware workloads.

CHAPTER 5. LIFECYCLE AND UPDATE POLICY

Security and critical bug fixes are delivered as container images available from the **registry.access.redhat.com/rhaiis** container registry and are announced through RHSA advisories. See RHAIIS container images on catalog.redhat.com for more details.