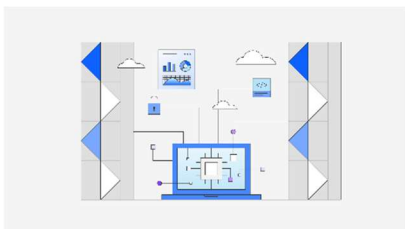


IBM AI Optimizer for Z



Optimize GenAI inferencing for IBM Z

Highlights

AI inferencing across Z with intelligent routing and multi-level caching deliver low-latency and high-throughput.

End-to-end observability with real-time insights for faster decision making.

IBM AI Optimizer for Z is designed to serve AI models and perform inference optimization directly on IBM Spyre™ accelerator (Spyre).

As **generative AI (GenAI)** continues to reshape business strategies, organizations reliant on IBM Z have a unique opportunity to accelerate innovation and deliver real business outcomes using AI. Without proper optimization, AI workloads are likely to become inefficient, slow, and prohibitively expensive.

AI Optimizer for Z addresses these challenges and introduces technical enhancements that improve both **performance and efficiency**.

Built on Red Hat OpenShift, it enables inferencing with intelligent tag-based routing, multi-level caching, and full-stack observability. It is particularly suited for inference workloads requiring **low latency, high throughput, and security**.

The solution is available in **two tiers**—Essentials and Advanced—providing flexibility for different business needs. AI Optimizer for Z 2.1 helps enterprises meet stringent SLAs, maintain data residency, and optimize resource utilization without compromising latency or compliance.

Core Capabilities

Real-Time Monitoring

Grafana dashboards provide visibility into usage, model performance, cache behavior, and system load. Using Grafana and Prometheus dashboards, AI Optimizer for Z 2.1 provides deep observability and near real-time insights into inferencing performance metrics, hardware and Spyre utilization, model usage patterns, and identifies bottlenecks and anomalies in model serving, e.g., Users can interpret complex data intuitively through one of the dashboards, avoid over-provisioning and plan future decisions on infrastructure and budget considerations using these metrics.

This visualization provides complete operational transparency.

Multi-level Caching

Intermediate inferencing results are reused to accelerate responses and reduce GPU/CPU computation. Eventually, with a staged delivery plan, two levels of caching that reuse previous computations for common token sequences across different inferencing requests, can be enabled:

1. Level 1 → Where Key Value (KV) caching can be done with one LLM deployed on several hardware units. Inferencing requests that have cached text will be accelerated and hardware utilization can be optimized.
2. Level 2 → Where caching can be shared across multiple LLM deployments, accelerating the inferencing, reducing time-to-first-token (TTFT) and increasing the throughput.

Inferencing Optimization

Execute AI workloads on IBM Z or external environments seamlessly, shifting dynamically based on latency or compliance. LLMs that run on Spyre can be automatically detected by AI Optimizer for Z and registered for inferencing optimization. Routing plans can be custom-built using the built-in intelligent router which considers availability, usage, and performance. LLMs serving similar applications or purposes can be grouped together by adding tags to them. The tags can also be configured by users following OpenAI APIs standards.

External model registration

External LLMs that are deployed on other infrastructures outside IBM Z and IBM LinuxONE can be registered with AI Optimizer for Z. These can be tagged and grouped with local LLMs running on Spyre to ensure use-case grouping and optimization. Depending on the LLM deployment, external LLMs monitoring can be integrated in the cross-platform monitoring dashboard to give a complete GenAI overview.

Depending on the business need for a GenAI use-case, Multiple models can be needed to achieve a certain goal. Therefore, AI Optimizer for Z allows registration of external models that are running outside IBM Z and IBM LinuxONE to unify the inferencing endpoints. External LLMs and Local LLMs can be grouped together through custom tags that can be used in inferencing requests to serve business needs.

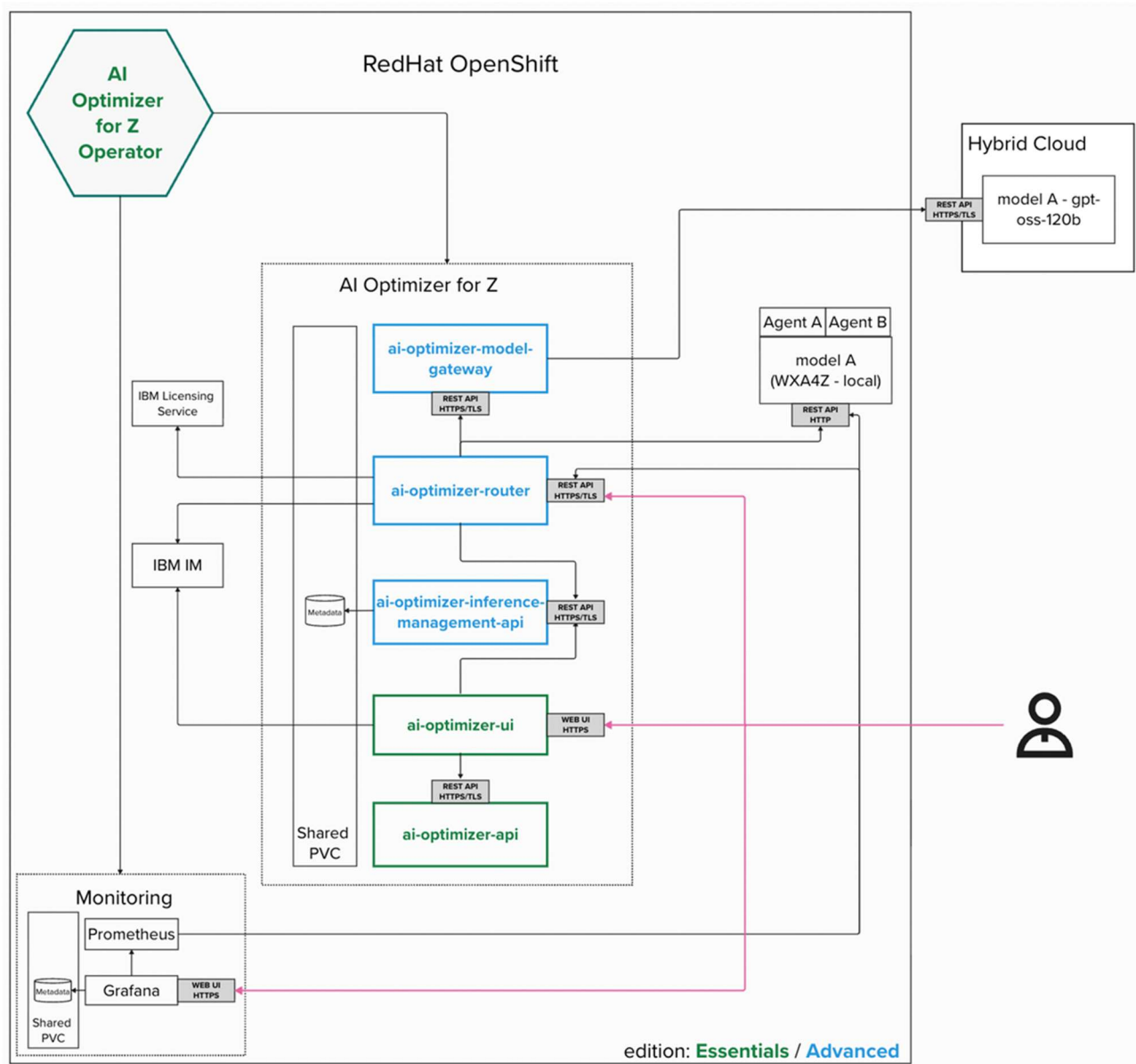


Figure 1: The product architecture.

Integration with watsonx

Assistant on Spyre

Combining AI Optimizer for Z with watsonx Assistant enables faster responses, lower latency, and seamless AI-powered experiences at scale.

[Learn more about the IBM AI Optimizer for Z](#) →

© Copyright IBM Corporation
2025

Produced in the
United States of America
December, 2025

IBM, the IBM logo, and IBM Spyre, are trademarks or registered trademarks of International Business Machines Corporation, in the United States and/or other countries. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on ibm.com/trademark.

Red Hat, OpenShift, and Ansible are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

© Copyright IBM Corporation 2025

Produced in the
United States of America
December, 2025

