Scalable AI Inferencing Platform: Red Hat OpenShift AI powered by Rebellions NPUs

Unlock scalable, cost-effective AI inference for enterprise workloads

As organisations adopt generative AI across more use cases, they face rising infrastructure costs, deployment complexity, and the need for flexible, secure environments.

Traditional GPU-based setups often struggle to meet performance and efficiency demands at scale especially in regulated sectors where data sovereignty and compliance are critical.

Rebellions and Red Hat deliver a smarter way to scale AI

This joint solution combines Red Hat OpenShift AI with Rebellions' energy-efficient NPUs to deliver a validated full-stack AI inference platform. It simplifies deployment, reduces operational costs, and enables consistent performance across environments.

- Enterprise-ready AI at scale: Run large language models and inference workloads with high throughput, low latency and superior power efficiency, leveraging vLLM integrated with Rebellions' rack-scale NPU solutions for distributed processing.
- Secure and compliant: Keep data on-premises and meet regulatory requirements with Red Hat's trusted platform and Rebellions' secure hardware.
- Simplified operations: Manage NPUs like GPUs, using Red Hat's unified platform, reducing complexity and accelerating adoption.
- Flexible and scalable: Deploy where your data lives, from core to edge, with linear scale-out.

Technology Used

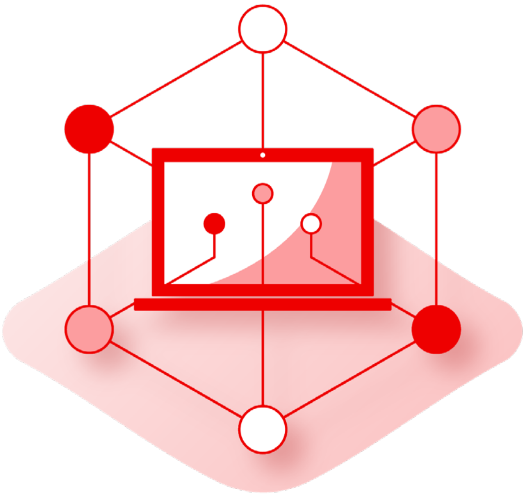
- Red Hat AI Inference Server
- Red Hat OpenShift AI
- Red Hat OpenShift
- Rebellions SDK
- Rebellions NPU Operator
- Rebellions NPUs

Key benefits

- **Enterprise-grade** AI inference at scale
- **Lower** infrastructure costs and energy usage
- **Seamless integration** with existing AI workflows
- **Secure, on-premises deployment** for compliance
- **Simplified operations** with GPU-like NPU experience
- **Validated full-stack architecture** with rapid deployment

Traditional AI infrastructure versus Rebellions + Red Hat OpenShift AI

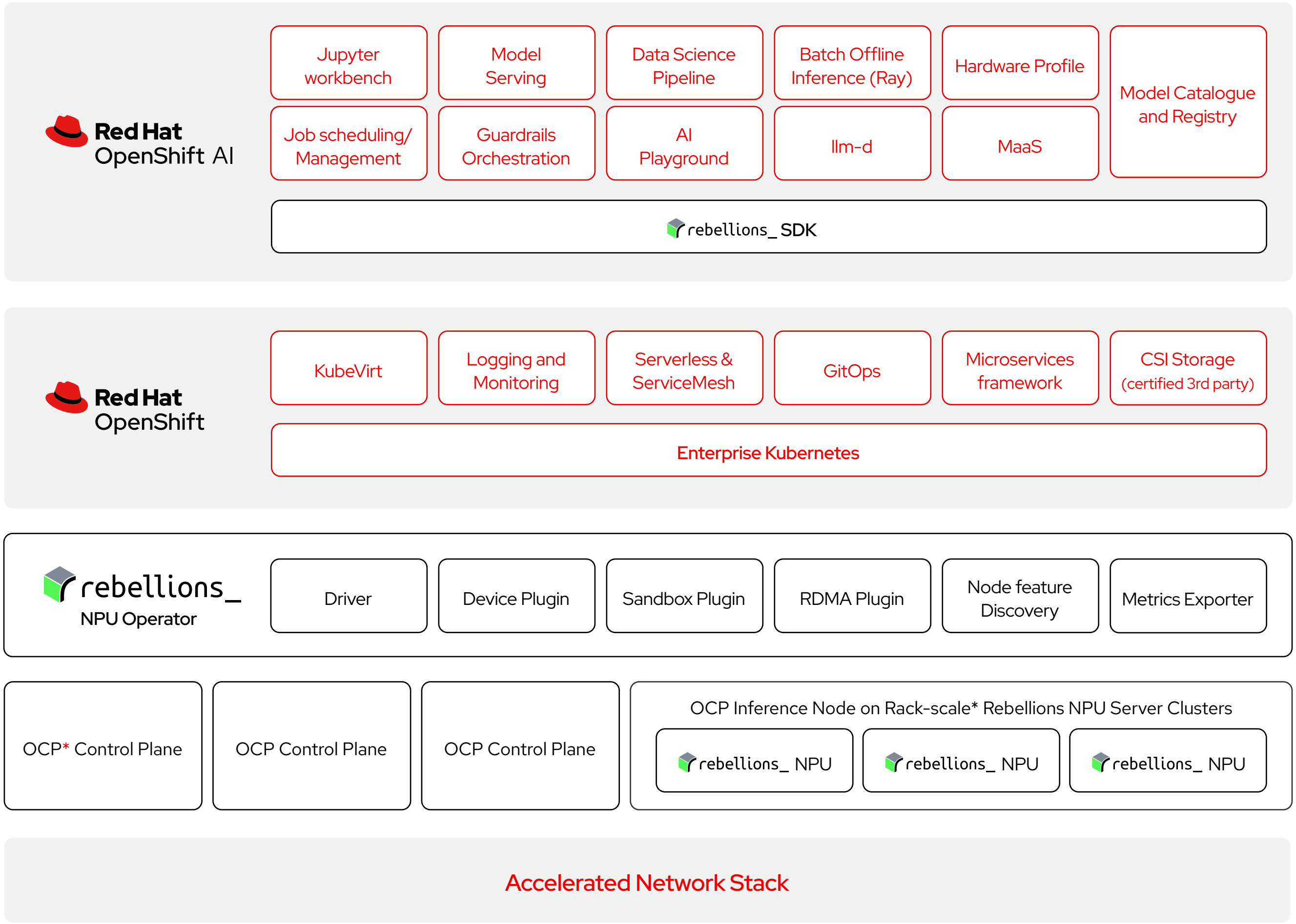
Traditional AI infrastructure	Rebellions + Red Hat OpenShift AI
High power consumption and cost	Energy-efficient NPUs with lower total cost of ownership
Complex multi-cloud deployment	Validated full-stack with seamless integration
Limited flexibility and vendor lock-in	Open-source platform with flexible deployment
Cumbersome deployment and optimization	Streamlined integration with OpenShift AI
Security and compliance challenges	On-premises, secure and compliant by design



Key features of the solution

- **Full-stack AI platform:** Integrated from hardware to model serving, validated by Red Hat and Rebellions for enterprise-grade compatibility.
- **Red Hat Verified Operator:** The Rebellions NPU Operator is officially certified, ensuring seamless integration and trusted support.
- **Seamless SDK integration:** Rebellions software stack runs natively on OpenShift AI, eliminating overhead and accelerating deployment.
- **Performance and efficiency:** Achieve high throughput with low latency and superior power efficiency, ideal for large-scale inference.
- **Flexible deployment:** Supports on-premises and multi-cloud environments, enabling data sovereignty and regulatory compliance.

Red Hat OpenShift AI powered by Rebellions NPUs



Red Hat OpenShift AI

Jupyter workbench

Model Serving

Data Science Pipeline

Batch Offline Inference (Ray)

Hardware Profile

Model Catalogue and Registry

Job scheduling/ Management

Guardrails Orchestration

AI Playground

llm-d

MaaS

rebellions_ SDK

Red Hat OpenShift

KubeVirt

Logging and Monitoring

Serverless & ServiceMesh

GitOps

Microservices framework

CSI Storage (certified 3rd party)

Enterprise Kubernetes

rebellions_ NPU Operator

Driver

Device Plugin

Sandbox Plugin

RDMA Plugin

Node feature Discovery

Metrics Exporter

OCP* Control Plane

OCP Control Plane

OCP Control Plane

OCP Inference Node on Rack-scale* Rebellions NPU Server Clusters

rebellions_ NPU

rebellions_ NPU

rebellions_ NPU

Accelerated Network Stack

*OCP: OpenShift Container Platform

This diagram shows how Red Hat OpenShift AI and Rebellions NPUs work together to deliver a complete AI inference platform. OpenShift AI provides enterprise AI services like model serving and batch inference, while the Rebellions SDK accelerates these workloads on NPUs.

The Red Hat–verified NPU Operator ensures seamless integration, and the underlying infrastructure uses an optimised network and resilient control plane for scalable, low-latency performance.

Getting started is simple

Deployment can be completed in just a few weeks using validated reference architectures and preconfigured solutions.

Deployment steps include:

- Assessment & planning (1–2 weeks): Review workloads and define deployment architecture.
- Stack deployment (1–2 weeks): Install OpenShift AI and integrate Rebellions NPUs.
- Model optimisation & validation (1–2 weeks): Tune workloads and validate performance.

rebellions_

About Rebellions

Rebellions develops AI accelerators optimised for Large Language Models and large-scale inference, delivering industry-leading energy efficiency.

Its flagship REBEL-Quad chip uses chiplet architecture and 144GB HBM3E memory for massive-scale AI inference.

The platform includes software for seamless datacenter integration. Building on ATOM's proven mass-production experience since 2024, Rebellions is backed by SK Telecom, SK Hynix, Aramco's Wa'ed Ventures, and KT.

Strengthened by its merger with SK SAPEON, Rebellions is now taking the lead in APAC and setting the stage for global AI semiconductor leadership.

About Red Hat

Red Hat is the world's leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, and Kubernetes technologies. Red Hat helps customers develop cloud-native applications, integrate existing and new IT applications, and automate and manage complex environments. A trusted adviser to the Fortune 500, Red Hat provides award-winning support, training, and consulting services that bring the benefits of open innovation to any industry. Red Hat is a connective hub in a global network of enterprises, partners, and communities, helping organizations grow, transform, and prepare for the digital future.

facebook.com/redhatinc
@RedHat
linkedin.com/company/red-hat

North America
1-888-REDHAT1
www.redhat.com

Europe, Middle East, and Africa
00800 7334 2835
europe@redhat.com

Asia Pacific
+65 6490 4200
apac@redhat.com

Latin America
+54 11 4329 7300
info-latam@redhat.com