IBM Newsroom

IBM Announces Red Hat AI Inference and Red Hat OpenShift Virtualization Service on IBM Cloud

- **IBM delivers Red Hat AI Inference, Red Hat OpenShift Virtualization Service as managed services**

- **New offerings designed to enable enterprises to operationalize AI and securely run virtualized workloads at scale**

ARMONK, N.Y., May 12, 2026 — IBM (NYSE: IBM) today announced two new managed services – Red Hat AI Inference on IBM Cloud and Red Hat OpenShift Virtualization Service on IBM Cloud – to help enterprises accelerate AI adoption and run security-forward, scalable and predictable virtualization environments. Red Hat AI Inference on IBM Cloud, with built-in governance controls, is designed to help clients reliably integrate real-time AI inferencing directly into their production workflows across hybrid cloud environments. Red Hat OpenShift Virtualization Service on IBM Cloud provides a managed path to help clients migrate and run virtual machines (VMs) securely and at scale. With these new offerings, IBM continues to provide clients with the full spectrum of Red Hat managed platform offerings to help accelerate hybrid cloud adoption.

As enterprises move beyond AI experimentation and into production, IBM is delivering a cloud foundation built on Red Hat technology to help clients innovate with speed and predictability to manage the compounding inference demand. Powered by Red Hat AI and delivered on IBM Cloud's enterprise grade infrastructure, Red Hat AI Inference on IBM Cloud is delivered as a managed service designed to span developer teams and agents. It is built to enable organizations to standardize the orchestration, performance and governance of AI models across the enterprise while freeing developers and platform teams to focus on delivering the value-added applications and services their clients need. Additionally, organizations are managing the need to migrate to purpose-built virtualization environments that are optimized for operational stability, security, compliance and predictable costs. Red Hat OpenShift Virtualization Service on IBM Cloud is a managed virtualization service that can help enterprises migrate and operate VM-based workloads on Red Hat OpenShift with Kubernetes-based infrastructure, automated lifecycle management and a consistent foundation toward containerization and application modernization. These new services build on IBM's existing managed offerings across Red Hat Enterprise Linux, Red Hat OpenShift, Red Hat Ansible Automation Platform and Red Hat AI.

"Enterprises are eager to operationalize AI, but the gap between pilot and productionmay hold them back. With Red Hat AI Inference on IBM Cloud, we're giving clients a managed platform that isbuilt for real workloads, not just experiments. At the same time, our new virtualization offering on IBM Cloud is enabling enterprises to migrate to a resilient and security-focused

virtualization environment while giving them the flexibility to adopt Red Hat OpenShift at their own pace for future AI workloads and containerization," said ***Jason McGee, CTO, IBM Cloud****.*

"These new managed services are the next step in our work with IBM to help enterprises drive innovation in the era of AI with an open, consistent hybrid cloud platform. By bringing Red Hat AI Inference and Red Hat OpenShift Virtualization Service to IBM Cloud, we are empowering clients to modernize at their own pace while preparing for an AI-driven future," said ***Ashesh Badani, senior vice president and chief product officer, Red Hat****.*

Helping Enterprises Scale Real-Time Inference Workloads Built for Consistent Performance and Predictable Cost

Red Hat AI Inference on IBM Cloud is an enterprise-ready, fully managed inference service designed to empower clients to run production-grade AI models without the complexity of managing GPUs, infrastructure or AI platforms. It brings together Red Hat AI's high-performance inference engine with IBM Cloud's enterprise-grade capabilities to help enterprises deploy AI models built for consistent performance and predictable cost. This approach helps organizations move from pilot deployments to steady-state production usage across a wide range of models and hardware without the cost volatility that real-world scaling requirements can bring. Operated and maintained by IBM Cloud, the service demonstrates how Red Hat AI can operate at enterprise scale with the security capabilities, reliability, and performance required for production workloads. IBM Cloud is the only cloud that provides a fully managed Red Hat AI add-on providing access to the full capabilities of Red Hat AI.

Key features include:

- *Production grade performance at enterprise scale:* The service is powered by vLLM and Red Hat AI's inference engine, optimized for high throughput and low latency, and designed to enable agents and applications to deliver consistent real-time performance. The model catalog includes Granite 4.0 H Small (IBM), Mistral-Small-3.2-24B-Instruct, Llama 3.3 70B Instruct, GPT-OSS-120B, and Nemotron-3-Nano-30B-FP8 with more open models and custom models planned starting in May 2026.
- *Accelerated time to production*: Designed to allow developers to integrate quickly using familiar OpenAI compatible APIs and without the need to manage GPUs or tuning runtimes, helping accelerate time-to-value.
- *Built-in security capabilities and governance*: Integration with IBM Cloud IAM, audit logging, privacy controls, and SLA backed reliability is designed to give enterprises full visibility and governance over model use and to support mission critical applications.
- *Models-as-a-Service*: Red Hat AI Inference enables organizations to set up AI models as API-accessible, shared resources, promoting rapid AI adoption while reducing infrastructure burden.

Enabling Enterprises to Modernize Virtualized Workloads with Ease

With many enterprises reassessing their virtualization strategies and seeking more predictable economics and a clearer path to upgrading VMs, Red Hat OpenShift Virtualization Service on IBM Cloud delivers a managed, enterprise-ready virtualization platform purpose-built for hybrid cloud environments. Backed by IBM Technology Expert Labs, IBM Consulting, Red Hat Services, and global system integrator partners, Red Hat OpenShift Virtualization Service on IBM Cloud helps enterprises migrate at scale and mitigate risk to accelerate time to value. Running on IBM Cloud VPC Bare Metal, the service is built to deliver predictable performance and total cost of ownership (TCO) while helping clients migrate virtual machines onto a Kubernetes-based virtualization environment.

IBM manages the platform lifecycle – including upgrades, patching, automated recovery, and worker-node remediation – so

organizations can offload the operational burden of running their own virtualization infrastructure and free IT teams to focus on the virtual machines, applications, and workloads themselves rather than the platform underneath them. With integrated migration tooling, including the Migration Toolkit for Virtualization, engineered so that clients can move from legacy environments quickly and with minimal disruption, the service is designed to create a clear path for future modern workload adoption on Red Hat OpenShift on IBM Cloud.

Availability

Red Hat AI Inference on IBM Cloud will be generally available on May 22, 2026Learn more here.

Red Hat OpenShift Virtualization Service on IBM Cloud is in limited availability and is expected to be generally available in June 2026. Learn more here.

Product features and timelines are subject to change at IBM's sole discretion and may not be available in all countries. Nothing in this release creates any warranties or alters applicable license terms.

Statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Learn more: https://www.ibm.com/products/cloud/redhat

About IBM

IBM is a leading provider of global hybrid cloud and AI, and consulting expertise. We help clients in more than 175 countries capitalize on insights from their data, streamline business processes, reduce costs and gain the competitive edge in their industries. Thousands of governments and corporate entities in critical infrastructure areas such as financial services, telecommunications and healthcare rely on IBM's hybrid cloud platform and Red Hat OpenShift to affect their digital transformations quickly, efficiently and securely. IBM's breakthrough innovations in AI, quantum computing, industry-specific cloud solutions and consulting deliver open and flexible options to our clients. All of this is backed by IBM's long-standing commitment to trust, transparency, responsibility, inclusivity and service.

Visit www.ibm.com for more information.

Media contact:

Kate Gazzillo
IBM
kate.gazzillo@ibm.com

https://stage.mediaroom.com/ibmnewsroom/2026-05-12-ibm-announces-red-hat-ai-inference-and-red-hat-openShift-virtualization-service-on-ibm-cloud