HITACHI

Reference Architecture Guide

Accelerate your AI Journey with Red Hat AI 3: Hitachi iQ with Hammerspace and Red Hat OpenShift AI

MK-SL-461-01
June 2026

Hitachi Vantara

© Hitachi Vantara LLC 2026. All Rights Reserved.

Hitachi iQ with Hammerspace and Red Hat OpenShift AI

Table of Contents

Notices and Disclaimer 3

Document conventions 4
Introduction 4
Intended Audience 4
Accessing product downloads 4
Getting Help 4
Comment 5
Document Revisions 5

Purpose 6

Solution Overview 7

Compute Platform Selection Guide 7
Key Benefits 8

Solution Components 9

Hitachi iQ MGX – Hitachi MGX Compute Platform 9
Hitachi iQ M Series – Hitachi AC520 Compute Platform 9
Hitachi HA810 G3 – Management Nodes 10
Hammerspace Storage Nodes 10
VSP One Block 28 10
Hammerspace Hyperscale NAS 11
Red Hat OpenShift AI – A Key Pillar of Red Hat AI 3 11
SLA-Aware Inference 11
Agentic AI with Llama Stack and MCP 12
Model Customization 12
GPU Resource Management 12
GPU Operator for Kubernetes 12

© Hitachi Vantara LLC 2026. All Rights Reserved.

Networking 13

Reference Configuration 13

Hardware Components — MGX Configuration 13
Hardware Components — AC520 Configuration 14
Software Components 14

Solution Design 16

Physical Architecture 16
OpenShift AI Deployment 16
Storage Architecture — Hammerspace Global File System 16

Solution Validation — Use Cases 18

Use Case 1: Generative AI Inference with Red Hat AI Inference Server 18
Use Case 2: Model Fine-Tuning with OpenShift AI 19
Use Case 3: Agentic AI with Llama Stack and MCP 20

Storage Performance 21

Conclusion 22

Next Steps 22

About the Vendors 23

Hitachi Vantara 23
Red Hat 23
Hammerspace 23

© Hitachi Vantara LLC 2026. All Rights Reserved.

Notices and Disclaimer

© 2026 Hitachi Vantara LLC. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or stored in a database or retrieval system for commercial purposes without the express written permission of Hitachi, Ltd., or Hitachi Vantara (collectively, "Hitachi"). Licensee may make copies of the Materials provided that any such copy is: (i) created as an essential step in utilization of the Software as licensed and is used in no other manner; or (ii) used for archival purposes. Licensee may not make any other copies of the Materials. "Materials" mean text, data, photographs, graphics, audio, video and documents.

Hitachi reserves the right to make changes to this Material at any time without notice and assumes no responsibility for its use. The Materials contain the most current information available at the time of publication.

Some of the features described in the Materials might not be currently available. Refer to the most recent product announcement for information about feature and product availability, or contact Hitachi Vantara at https://support.HitachiVantara.com/ en_us/contact-us.html.

Notice: Hitachi products and services can be ordered only under the terms and conditions of the applicable Hitachi agreements. The use of Hitachi products is governed by the terms of your agreements with Hitachi Vantara.

By using this software, you agree that you are responsible for:

1) Acquiring the relevant consents as may be required under local privacy laws or otherwise from authorized employees and other individuals to access relevant data; and
2) Verifying that data continues to be held, retrieved, deleted, or otherwise processed in accordance with relevant laws.

Notice on Export Controls. The technical data and technology inherent in this Document may be subject to U.S. export control laws, including the U.S. Export Administration Act and its associated regulations, and may be subject to export or import regulations in other countries. Reader agrees to comply strictly with all such regulations and acknowledges that Reader has the responsibility to obtain licenses to export, re-export, or import the Document and any Compliant Products.

EXPORT CONTROLS - Licensee will comply fully with all applicable export laws and regulations of the United States and other countries, and Licensee shall not export, or allow the export or re-export of, the Software, API, or Materials in violation of any such laws or regulations. By downloading or using the Software, API, or Materials, Licensee agrees to the foregoing and represents and warrants that Licensee is not located in, under the control of, or a national or resident of any embargoed or restricted country.

Hitachi is a registered trademark of Hitachi, Ltd., In the United States and other countries.

AIX, AS/400e, DB2, Domino, DS6000, DS8000, Enterprise Storage Server, eServer, FICON, Flash Copy, IBM, Lotus, MVS, OS/390, PowerPC, RS6000, S/390, System z9, System z10, Tivoli, z/OS, z9, z10, z13, z/VM, BCPii™ and z/VSE are registered trademarks or trademarks of International Business Machines Corporation.

Active Directory, ActiveX, Bing, Excel, Hyper-V, Internet Explorer, the Internet Explorer logo, Microsoft, the Microsoft Corporate Logo, MS-DOS, Outlook, PowerPoint, SharePoint, Silverlight, SmartScreen, SQL Server, Visual Basic, Visual C++, Visual Studio, Windows, the Windows logo, Windows Azure, Windows PowerShell, Windows Server, the Windows start button, and Windows Vista are registered trademarks or trademarks of Microsoft Corporation. Microsoft product screen shots are reprinted with permission from Microsoft Corporation.

All other trademarks, service marks, and company names in this document or web site are properties of their respective owners.

Preface

About This Document

This document presents a comprehensive performance evaluation of the Hitachi Virtual Storage Platform (VSP) 5600-4N storage system, conducted in a customer-aligned environment for Reliance Jio.

Document conventions

This document uses the following typographic convention:

Convention	Description
Bold	• Indicates text in a window, including window titles, menus, menu options, buttons, fields, and labels. Example: **Click OK**. • Indicates emphasized words in list items.
Italic	Indicates a document title or emphasized words in text.
`Monospace`	Indicates text that is displayed on screen or entered by the user. Example: `pairdisplay -g oradb`

Introduction

This guide outlines a validated reference architecture for building a scalable enterprise AI platform using Hitachi iQ, Hammerspace, and Red Hat OpenShift AI, supporting the full AI lifecycle.

Intended Audience

This document is intended for AI/ML professionals, data engineers, and IT architects with basic knowledge of GPU systems, Kubernetes, and AI/ML workflows.

Accessing product downloads

Product software, drivers, and firmware downloads are available on Hitachi Vantara Support Connect: https://support.hitachivantara.com/.

Log in and select Product Downloads to access the most current downloads, including updates that may have been made after the release of the product.

Getting Help

Hitachi Vantara Support Connect is the destination for technical support of products and solutions sold by Hitachi Vantara. To contact technical support, log on to Hitachi Vantara Support Connect for contact information: https://support.hitachivantara.com/en_us/contact-us.html.

Hitachi Vantara Community is a global online community for customers, partners, independent software vendors, employees, and prospects. It is the destination to get answers, discover insights, and make connections. **Join the conversation today!** Go to community.hitachivantara.com, register, and complete your profile.

© Hitachi Vantara LLC 2026. All Rights Reserved.

Comment

Send any comments on this document to Docs-Feedback@hitachivantara.com. Include the document title, including the revision level, and refer to specific sections and paragraphs whenever possible. All comments become the property of Hitachi Vantara LLC.

Thank you.

Document Revisions

Revision Number	Date	Details
v1.0	June 2026	Initial release

© Hitachi Vantara LLC 2026. All Rights Reserved.

Purpose

In this reference architecture guide, the design for a production-ready enterprise AI platform is described by combining the Hitachi iQ AC520 and/or MGX compute and storage infrastructure with Hammerspace Hyperscale NAS and Red Hat OpenShift AI. This guide is intended for organizations that require a scalable, high-performance, and operationally simplified foundation for artificial intelligence (AI) and machine learning (ML) workloads, ranging from generative AI inference and model fine-tuning to agentic AI workflows.

Red Hat AI 3 is an enterprise AI platform that delivers a unified solution for building and running AI models. Its three core components are optimized for different scales and environments.

- **Red Hat OpenShift AI**: A comprehensive MLOps platform for large-scale AI workloads. It runs on Red Hat OpenShift clusters and provides the tools required to build, train, deploy and manage models in a secure, scalable, container-based environment.
- **Red Hat Enterprise Linux (RHEL) AI**: An AI-optimized operating system designed for model development and tuning. It runs on bare-metal servers or virtual machines (VMs), allowing you to build and test, and optimize models directly on your hardware, close to your data.
- **AI Inference Server**: A flexible, portable inference engine for serving production-ready models. As a containerized service, it can be deployed across environments – on OpenShift, RHEL, or other standard Kubernetes and Linux platforms – to expose model's predictions as a scalable API.

Red Hat OpenShift AI, powered by Red Hat AI 3, delivers an open-source, enterprise-grade AI platform for building, training, fine-tuning, deploying, and monitoring AI models across hybrid cloud environments. The Hitachi iQ hardware platform provides two GPU compute node options – the Hitachi MGX and the Hitachi AC520 – allowing organizations to select the platform best aligned with workload scale, performance requirements, and budget. Hammerspace delivers a global parallel file system that unifies storage across multiple systems and sites, ensuring AI workloads consistently have fast, scalable, parallel access to datasets and model checkpoints.

This reference architecture builds on Hitachi Integrated Systems with Red Hat AI. For more information, see the following documents available on the Hitachi Vantara Product Documentation Portal:

- MK-SL-373-01 Accelerate your AI Journey with Red Hat AI and NVIDIA with the Hitachi Integrated Systems Platform and Virtual Storage Platform One
- MK-SL-332-00 Hitachi iQ M Series Solution with Hammerspace and Hitachi Virtual Storage Platform One Block

The design introduces Hammerspace Hyperscale NAS as the unified storage layer and adopts Red Hat AI 3 as the AI software platform. Compute support is expanded to include the Hitachi MGX for entry-level to midrange workloads and the Hitachi AC520 platform for midrange to high-end workloads, reflecting the full breadth of the Hitachi iQ hardware portfolio.

The intended audience of this document is:

- AI and ML architects and engineers
- Data scientists and data engineers
- Storage and infrastructure administrators
- IT professionals evaluating AI platform options
- Technology sales engineers and solution architects

The audience for this guide is expected to have familiarity with GPU-based server platforms, NFS and SMB storage concepts, Kubernetes orchestration, and general AI and ML workflows.

© Hitachi Vantara LLC 2026. All Rights Reserved.

Solution Overview

This reference architecture integrates three complementary technology layers to deliver a complete and validated enterprise AI platform:

- **Hitachi iQ hardware platform with Hammerspace**: Purpose-built GPU compute, available in two server form factors – the Hitachi MGX and Hitachi AC520 – is combined with Hammerspace Hyperscale NAS backed by Hitachi VSP One Block NVMe storage. Organizations can select the GPU compute platform that best aligns with workload requirements and budget, when using a common Hammerspace storage layer and Red Hat OpenShift AI software stack across both options.

- **Red Hat OpenShift AI (as part of Red Hat AI 3)**: A Kubernetes-native AI platform that supports the full AI lifecycle, ranging from JupyterLab workbenches and distributed model training to enterprise-grade inference using the Red Hat AI Inference Server (vLLM), agentic AI workflows enabled by Llama Stack and Model Context Protocol (MCP), and integrated production observability. Built on open standards, Red Hat AI 3 supports any model on any supported hardware accelerator.

- **Hitachi VSP One Block**: All-NVMe enterprise block storage connected directly to Hammerspace DSX nodes using NVMe over TCP, providing the low-latency, high-bandwidth persistent storage layer that Hammerspace presents to all clients as a unified global file system.

Flexibility at the compute layer is a core design principle of the architecture. Organizations in the early stages of AI adoption, such as those running smaller models, evaluating use cases, or operating under tighter budget constraints – can deploy the Hitachi MGX platform with NVIDIA RTX PRO Blackwell GPUs. As requirements evolve, workloads can scale within the MGX platform or transition to the higher-capacity Hitachi AC520 platform with NVIDIA H200 NVL or L40S GPUs. Across both options, the Hammerspace storage infrastructure, Red Hat OpenShift AI platform, and networking architecture remain consistent.

Red Hat AI 3 enables enterprises to move, AI initiative out of experimentation and into production. Built on open-source technologies, such as vLLM, llm-d, Llama Stack, InstructLab, and Kubernetes, the platform provides long-term flexibility, allowing organizations to select any model from any provider and deploy on any supported hardware accelerator as the AI landscape continues to evolve.

Compute Platform Selection Guide

The following table summarizes the key differences between the two Hitachi iQ hardware compute platforms to help organizations in selecting the option best aligned with

their workload requirements:

Attribute	Hitachi MGX	Hitachi AC520
Positioning	Entry-level to midrange AI/ML and generative AI workloads	Midrange to high-end AI/ML and generative AI workloads
Form Factor	2U (SYS-221GE-FNR2)	5U (SYS-522GA-NRT)
Max GPUs per Node	4 double-width PCIe	8 double-width PCIe

© Hitachi Vantara LLC 2026. All Rights Reserved.

Attribute	Hitachi MGX	Hitachi AC520
Supported GPUs	NVIDIA RTX PRO 6000 Server Edition (up to 4); NVIDIA RTX PRO 4500 Blackwell Server Edition (up to 6)	NVIDIA L40S 48 GB (up to 8); NVIDIA H200 141 GB NVL (up to 8)
Processors	Dual Intel Xeon Gold 6548Y (32C each)	Dual Intel Xeon 6952P (96C each)
Memory	Up to 1,024 GB DDR5	Up to 1,536 GB DDR5
Storage Fabric NIC	200G ConnectX-7 NDR (dual-port)	200G ConnectX-6 VPI (dual-port)
Scale-out	2 to 16 nodes	2 to 16 nodes
Hammerspace Storage	Same: Hammerspace + VSP One Block	Same: Hammerspace + VSP One Block
AI Platform	Same: Red Hat OpenShift AI (part of Red Hat AI 3)	Same: Red Hat OpenShift AI (part of Red Hat AI 3)

Table 1 Key differences between the two Hitachi iQ hardware compute platforms

Key Benefits

- **Flexible Compute for Every Stage of AI Adoption**: The Hitachi MGX platform provides a cost-effective entry point for organizations starting their AI initiatives, whereas the Hitachi AC520 platform delivers higher GPU density and memory capacity to support larger models and higher concurrency. Both platforms share a common storage and software architecture, protecting infrastructure investments as AI workloads scale.
- **Scalable Parallel Storage Optimized for AI**: Hammerspace Hyperscale NAS delivers linear performance scaling as DSX nodes are added, providing the sequential throughput and parallel data access required to keep GPU clusters efficiently used during training and inference.
- **Open, Enterprise-Grade AI Platform**: Red Hat AI 3 is built entirely on open-source technologies, enabling flexibility and avoiding vendor lock-in while delivering enterprise-grade security, lifecycle management, and support.
- **Accelerated Time to Value**: Red Hat OpenShift AI reduces the operational complexity of assembling and maintaining AI tooling, enabling teams to transition from infrastructure deployment to AI model development and prototyping significantly faster than traditional, manual approaches.
- **End-to-End AI Lifecycle on a Single Platform**: The platform supports dataset preparation, model training, fine-tuning, deployment, serving, monitoring, and agentic AI workflows all managed through a unified Kubernetes-based platform.
- **Agentic AI Readiness**: Red Hat AI 3 introduces support for Llama Stack and MCP, providing a foundation for building autonomous AI agents that interact with enterprise data sources and tools.
- **Unified Enterprise Support**: Hitachi Vantara delivers SLA-backed support across compute, storage, and networking paired with Red Hat's enterprise support for the AI software platform, simplifying vendor relationship for enterprise AI deployments.

© Hitachi Vantara LLC 2026. All Rights Reserved.

Solution Components

Hitachi iQ MGX – Hitachi MGX Compute Platform

The Hitachi MGX is a 2U PCIe GPU compute platform designed for entry-level to midrange AI, ML, and generative AI workloads. Its compact 2U form factor and support for NVIDIA RTX PRO Blackwell GPUs make it well suited for organizations that require cost-effective GPU acceleration for inference workloads, smaller model training runs, or initial AI proof-of-concept deployments, with a clear path to scale toward higher-density platforms as requirements grow.

Hitachi MGX GPU Compute Node (SYS-221GE-FNR2)

- Dual Intel Xeon Gold 6548Y processors (32 cores each, up to 250W TDP)
- Up to 1,024 GB DDR5 RAM (16 x 64 GB DIMMs)
- Support for up to double-width PCIe GPU accelerator cards, including:
  - 2 to 4 x NVIDIA RTX PRO 6000 Server Edition, or
  - 2 to 6 x NVIDIA RTX PRO 4500 Blackwell Server Edition
- Up to 8 front hot-swap E1.S NVMe drive bays; 2 x 960 GB M.2 NVMe boot drives
- Dual-port 200G ConnectX-7 NDR (QSFP112, VPI) for storage fabric connectivity
- Dual-port 10 GbE (AOC-STGS-I2T-P) for in-band operating system management
- Four redundant 2000W Titanium-class power supplies
- Scale-out support ranging from 2 to 16 nodes

Hitachi iQ M Series – Hitachi AC520 Compute Platform

The Hitachi AC520 is a 5U dual-root PCIe GPU compute platform designed for midrange to high-end AI, ML, and generative AI workloads. With support for up to 8 GPU accelerators per node, including the NVIDIA H200 NVL with 141 GB HBM3e memory – the platform is well-suited for large language model training, fine-tuning, and high-concurrency inference workloads that demand maximum GPU memory, bandwidth, and compute density.

Hitachi AC520 GPU Compute Node (SYS-522GA-NRT)

- Dual Intel Xeon 6952P processors (96 cores each, up to 500W TDP)
- Up to 1,536 GB DDR5 RAM (24 x 64 GB DIMMs)
- Up to 8 double-width PCIe GPU accelerator cards:
  - 2 to 8 x NVIDIA L40S 48 GB, or
  - 2 to 8 x NVIDIA H200 141 GB NVL
- 4 x 3.84 TB NVMe local drives + 2 x 960 GB M.2 NVMe boot drives
- Dual-port 200 GbE NVIDIA ConnectX-6 VPI (QSFP56) for storage fabric connectivity
- Dual-port 10 GbE OCP 3.0 adapter for in-band operating system management
- 6 redundant N+N 2700W Titanium-class power supplies
- Scale-out support ranging from 2 to 16 nodes

© Hitachi Vantara LLC 2026. All Rights Reserved.

Hitachi HA810 G3 – Management Nodes

HA810 G3 servers function as Red Hat OpenShift Container Platform control plane nodes in both the MGX and AC520 configurations. These nodes run the Kubernetes API server, etcd, and the scheduler, but do not run AI workloads. Cluster lifecycle management, including installation, upgrades, and ongoing management, are handled natively by Red Hat OpenShift.

- Dual Intel Xeon Gold 5418Y processors (24 cores each, 2.0 GHz) with 512 GB RAM
- Dual 480 GB NVMe M.2 boot devices configured in RAID-1 using the NS204i-u module
- Dual-port 100 GbE QSFP28 NIC for storage fabric connectivity; plus, a 10 GbE OCP 3.0 adapter for management traffic

Hammerspace Storage Nodes

The Hammerspace storage infrastructure remains consistent across both the MGX and AC520 compute configurations. A common set of Anvil metadata nodes, DSX data nodes, and VSP One Block storage systems supports GPU compute platforms or both simultaneously in mixed deployments.

HA815 G6 – Hammerspace Anvil (Metadata) Nodes

The MGX and AC520 configurations use HA815 G6 servers as Hammerspace Anvil metadata nodes.

- Dual AMD EPYC 9224 24-core processors (2.5 GHz, 48 total cores)
- 1 TB RAM (16 x 64 GB DDR5-4800)
- 480 GB NVMe SSD boot device; 2 x 7.68 TB SSD for metadata
- Dual-port 100 GbE QSFP28 (Intel E810) for data network
- Dedicated 25 GbE direct HA heartbeat link between the Anvil node pair

HA815 G3 – Hammerspace DSX (Data) Nodes

HA815 G3 servers function as Hammerspace DSX data nodes are the same in both the MGX and AC520 configurations.

- Dual AMD EPYC 9354 32-core processors (3.25 GHz, 64 total cores)
- 512 GB RAM (16 x 32 GB DDR5-4800)
- 480 GB NVMe SSD boot device; 2 x 7.68 TB SSD for caching
- Dual-port 100 GbE QSFP28 (Intel E810) for client-facing data network
- Dual-port 100 GbE QSFP56 (ConnectX-6) for NVMe/TCP backend to VSP One Block
- Support for 2 to 4 DSX nodes per VSP One Block 28 storage systems

VSP One Block 28

Virtual Storage Platform One (VSP One) Block 28 is an all-NVMe enterprise block storage system that provides primary persistent storage for the Hammerspace DSX nodes through NVMe over TCP. This storage platform is common to both the MGX and AC520 configurations, ensuring a consistent storage foundation regardless of the selected GPU compute platform.

© Hitachi Vantara LLC 2026. All Rights Reserved.

- 24 x 3.8 TB NVMe drives, providing approximately 91 TB of raw storage capacity)
- 1 TiB cache with 4 x Compression Accelerator Modules (CAM)
- 8 x NVMe TCP 100 G host ports for connectivity to Hammerspace DSX nodes
- Always-on compression and deduplication enabled through Advanced Data Reduction (ADR)
- Dynamic Drive Protection (DDP), RAID-6 equivalent resilience with distributed spare capacity
- Hitachi Ops Center Clear Sight embedded with management, including a built-in sustainability and energy-efficiency dashboard

Hammerspace Hyperscale NAS

Hammerspace is a software-defined data orchestration solution that creates a unified Global Data Environment spanning multiple storage systems, sites, and cloud environments. It presents storage to clients via standard NFS, SMB, and S3 protocols, making it completely transparent to AI frameworks such as PyTorch, TensorFlow, and Hugging Face.

Internally, Hammerspace uses pNFS with the Flexible Files layout type (pNFS Flex Files), allowing multiple clients to read and write in parallel directly to DSX nodes — delivering parallel file system performance without requiring specialized clients or drivers. This is what enables the GPU compute nodes (whether MGX or AC520) to feed large datasets to their GPUs without storage becoming a bottleneck.

- **Anvil Nodes:** The metadata control plane. Anvil nodes manage namespace, access control, file system metadata, snapshots, and replication. They are deployed as an HA pair and are never in the data path — clients negotiate file locations via Anvil, then transfer data directly to DSX nodes.
- **DSX Nodes:** The data plane. DSX nodes handle all I/O and scale out linearly — adding DSX nodes increase both throughput and capacity independently proportionally, regardless of the GPU compute platform in use.
- **Storage (VSP One Block):** Hammerspace consumes block storage from VSP One Block through NVMe/TCP, surfacing it as internal capacity within the global file system. Up to 4 DSX nodes connect per VSP One Block 28.
- **Clients:** Standard NFS, SMB, and S3 clients — including Linux OpenShift worker nodes running (MGX or AC520), Windows data science workstations, and S3-compatible applications.

Red Hat OpenShift AI — A Key Pillar of Red Hat AI 3

Red Hat OpenShift AI is the AI platform layer of this reference architecture. It runs on Red Hat OpenShift Container Platform and provides a Kubernetes-native environment for the full AI lifecycle. OpenShift manages GPU resources, schedules AI workloads, provisions persistent storage from Hammerspace's global file system, and delivers the developer-facing AI tools that data scientists and AI engineers need to be productive — across both the Hitachi MGX-based systems and Hitachi AC520 compute platforms.

Red Hat AI 3 delivered enterprise AI capabilities across three strategic pillars:

SLA-Aware Inference

The llm-d provides Kubernetes-native distributed inference for large language models, generally available in OpenShift AI 3.0. By intelligently distributing inference across nodes, llm-d delivers consistent resource allocation and predictable response times essential for production generative AI applications — whether running on MGX or AC520 GPU nodes.

The Red Hat AI Inference Server (version 3.2 and later) serves as the core model serving engine — an enterprise-grade version of vLLM with model optimization capabilities and extended hardware accelerator support covering NVIDIA and AMD GPUs. Red Hat AI 3 also provides a validated and optimized third-party model catalog featuring frontier open-source models (including DeepSeek-R1, OpenAI's gpt-oss, Whisper, and more) delivered as scanned and traceable containers.

For organizations managing model access across teams, Models as a Service (MaaS) — introduced as a developer preview in OpenShift AI 3.0 — provides an API gateway, role-based access control, and cost-tracking metrics that allow IT teams to centralize and govern model consumption.

© Hitachi Vantara LLC 2026. All Rights Reserved.

Agentic AI with Llama Stack and MCP

Red Hat AI 3 introduces Llama Stack — an open-source framework providing a standardized interface for retrieval-augmented generation (RAG), safety guardrails, agent evaluation, and tool calling — as a technology preview in OpenShift AI 3.0. Llama Stack enables teams to build AI agents that maintain context across multi-turn conversations, call external tools, and integrate with knowledge bases, all through a consistent OpenAI-compatible API.

Model Context Protocol (MCP) support connects AI agents to external tools, data sources, and enterprise applications through a standardized interface. Red Hat has emerged as an early adopter of MCP and OpenShift AI 3.0 and later includes MCP server support and a curated collection of MCP connectors — allowing enterprise teams to integrate their existing software with AI agents without custom integrations.

The OpenShift AI Gen AI Playground, available as a technology preview in OpenShift AI 3.3 provides an interactive environment where AI engineers can experiment with deployed models, test prompts, configure MCP tools, and prototype Retrieval-Augmented Generation (RAG) workflows directly from the OpenShift AI dashboard. For production workflows, Llama Stack capabilities are accessed through the OpenAI-compatible API or from JupyterLab notebooks using the Llama Stack Python SDK.

Model Customization

- **InstructLab**: Enables fine-tuning of foundation models using synthetic data generation, reducing reliance on large, labeled datasets or deep ML expertise. Red Hat AI 3 modularizes InstructLab into specialized Python libraries for data ingestion, synthetic data generation (SDG), tuning, and evaluation, enabling integration into automated pipelines.
- **RAG Pipelines**: An enhanced RAG workflow integrates with docling for document ingestion, supports multiple vector databases (pgvector, FAISS, Qdrant), and provides OpenAI-compatible APIs for embeddings and completions.
- **Distributed Training**: Kubeflow Trainer v2 (technology preview in OpenShift AI 3.2+) provides Kubernetes-native distributed training for PyTorch workloads, featuring a unified TrainJob API and intelligent checkpoint management compute platform is deployed.
- **Model Registry**: Centralized discovery, versioning, and lifecycle management for AI models and artifacts, enabling governed reuse and promotion across teams and environments.

GPU Resource Management

- **GPU-as-a-Service**: Accelerator slicing for MIG-enabled GPUs partitions a single GPU across multiple users or workloads.
- **Kueue Integration**: Queue-based scheduling supports Ray training jobs, training operator jobs, and inference services across shared GPU resources.
- **Hardware Profiles**: Administrators define profiles that target specific worker nodes, allowing workbenches and model-serving workloads to be pinned to the appropriate GPU compute platforms based on performance and capacity requirements (for example, MGX nodes for lighter workloads and AC520 nodes for heavier ones.
- **Built-in Observability**: Zero-configuration GPU monitoring through OpenTelemetry provides built-in dashboards for time-to-first-token, throughput, latency, and GPU utilization.

GPU Operator for Kubernetes

The NVIDIA GPU Operator is an open-source Kubernetes operator that automates provisioning of GPU resources within an OpenShift cluster. It handles installation and lifecycle management of the GPU driver stack — CUDA runtime, device plug-in, container runtime, and node labeling — on GPU worker nodes, whether those nodes are MGX or AC520 servers. This ensures that AI workloads scheduled by OpenShift AI can access the full GPU resources of each node without manual driver management.

© Hitachi Vantara LLC 2026. All Rights Reserved.

Networking

- **Storage Frontend Network (100 GbE):** Connects GPU compute nodes (MGX or AC520) and Hammerspace Anvil and DSX nodes using dual Cisco Nexus 93600CD-GX switches. This network carries all AI client access to the Hammerspace global file system traverses this network.
- **Storage Backend Network (100 GbE direct):** Provides dedicated connections between Hammerspace DSX nodes and VSP One Block 28 through NVMe/TCP. Each DSX node is equipped with 2 x 100 GbE links supporting up to 4 DSX nodes per VSP One Block 28.
- **Management Network (1/10 GbE):** Supports out-of-band (iLO) and in-band OS management using dual Cisco Nexus 93108TC-FX3 switches and also carries Kubernetes control plane traffic.
- **Anvil HA Network (25 GbE direct):** A dedicated private Layer 2 heartbeat network between the two Anvil nodes for HA coordination.
- **Inter-Switch Links (400 GbE):** High-bandwidth 400 GbE ISL between the paired Nexus 93600CD-GX storage switches for east-west fabric redundancy.

Reference Configuration

Hardware Components – MGX Configuration

The following table outlines the hardware components for the Hitachi iQ MGX reference configuration. The minimum configuration begins with 2 MGX compute nodes and scales up to 16 nodes.

Vendor	Hardware	Description	Version	Qty
Hitachi Vantara	HA810 G3	OpenShift control plane nodes. Dual Intel Xeon Gold 5418Y, 512 GB RAM, 10 GbE OCP3, RHEL 9.4	SPV 7.42	3
Hitachi MGX	SYS-221GE-FNR2	GPU compute / OpenShift worker nodes. Dual Intel Xeon Gold 6548Y (32C), 1024 GB DDR5, up to 4x GPU (RTX PRO 6000 or RTX PRO 4500 Blackwell), 200G ConnectX-7, RHEL 9.4	BMC 01.05.08 / BIOS 2.7	2–16
Hitachi Vantara	HA815 G6 (Anvil)	Hammerspace metadata nodes (HA pair). Dual AMD EPYC 9224 (48C), 1 TB DDR5, 2x 7.68 TB SSD, 100 GbE dual port	OS v5.1.24-318	2
Hitachi Vantara	HA815 G3 (DSX)	Hammerspace data nodes. Dual AMD EPYC 9354 (64C), 512 GB DDR5, 2x 7.68 TB SSD, 100 GbE client + 100 GbE NVMe/TCP backend	OS v5.1.24-318	2–4
Hitachi Vantara	VSP One Block 28	All-NVMe block storage. 24x 3.8 TB NVMe, 1 TiB cache, 4x CAM, 8x NVMe TCP 100G, always-on ADR	SVOS 10.3.1	1+
Cisco	Nexus 93600CD-GX	100 GbE storage fabric switch. 28x 100G QSFP28 + 8x 400G QSFP-DD, 12 Tbps	NXOS 10.4(3)	2

© Hitachi Vantara LLC 2026. All Rights Reserved.

Vendor	Hardware	Description	Version	Qty
Cisco	Nexus 93108TC-FX3	1/10 GbE management switch. 48x 10G BASE-T + 6x 100G QSFP28	NXOS 10.4(3)	2

Table 2 Hardware components for the Hitachi iQ MGX reference configuration

Hardware Components – AC520 Configuration

The following table describes the hardware components for the Hitachi iQ AC520 reference configuration. The minimum configuration begins with two AC520 compute nodes and scales to 16 nodes. Hammerspace storage nodes, VSP One Block storage systems, and networking components are identical to those used in the MGX configuration, with the exception of the Anvil server generation.

Vendor	Hardware	Description	Version	Qty
Hitachi Vantara	HA810 G3	OpenShift control plane nodes. Dual Intel Xeon Gold 5418Y, 512 GB RAM, 10 GbE OCP3, RHEL 9.4	SPV 7.42	3
Hitachi Vantara	AC520 (SYS-522GA-NRT)	GPU compute / OpenShift worker nodes. Dual Intel Xeon 6952P (96C), 1536 GB DDR5, up to 8x GPU (L40S 48 GB or H200 141 GB NVL), 200G ConnectX-6 VPI, RHEL 9.4	BMC 01.00.12	2–16
Hitachi Vantara	HA815 G6 (Anvil)	Hammerspace metadata nodes (HA pair). Dual AMD EPYC 9224 (48C), 1 TB DDR5, 2x 7.68 TB SSD, 100 GbE dual-port	OS v5.1.24-318	2
Hitachi Vantara	HA815 G3 (DSX)	Hammerspace data nodes. Dual AMD EPYC 9354 (64C), 512 GB DDR5, 2x 7.68 TB SSD, 100 GbE client + 100 GbE NVMe/TCP backend	OS v5.1.24-318	2–4
Hitachi Vantara	VSP One Block 28	All-NVMe block storage. 24x 3.8 TB NVMe, 1 TiB cache, 4x CAM, 8x NVMe TCP 100G, always-on ADR	SVOS 10.3.1	1+
Cisco	Nexus 93600CD-GX	100 GbE storage fabric switch. 28x 100G QSFP28 + 8x 400G QSFP-DD, 12 Tbps	NXOS 10.4(3)	2
Cisco	Nexus 93108TC-FX3	1/10 GbE management switch. 48x 10G BASE-T + 6x 100G QSFP28	NXOS 10.4(3)	2

Table 3 Hardware components for the Hitachi iQ AC520 reference configuration

Software Components

The following software components apply to the MGX and AC520 configurations.

© Hitachi Vantara LLC 2026. All Rights Reserved.

Component	Version / Details
Red Hat OpenShift Container Platform	4.16 and later
Red Hat OpenShift AI	3.0 and later (part of Red Hat AI 3)
Red Hat AI Inference Server	3.2 and later (enterprise vLLM, part of Red Hat AI 3)
llm-d (distributed inference)	GA in OpenShift AI 3.0
Llama Stack	Technology Preview – OpenShift AI 3.0 and later
Hammerspace Software	5.1.24-318
VSP One Block SVOS	10.3.1
NVIDIA GPU Operator	24.9.2 and later (open source Kubernetes operator)
OS – Management / GPU Nodes	RHEL 9.4
OS – Hammerspace Nodes	CentOS 8.5.2111
Cisco Nexus Switches	NXOS 10.4(3)

Table 4 Software components applicable to MGX and AC520 configurations

© Hitachi Vantara LLC 2026. All Rights Reserved.

Solution Design

Physical Architecture

Both the MGX and AC520 configurations follow the same logical physical architecture and modular rack design. The architecture is scalable, starting with a minimum of two GPU compute nodes and expanding to a maximum of 16 nodes across one or multiple racks. The key physical groupings include:

- **Management Nodes (HA810 G3 – 3 nodes)**: Serve as Red Hat OpenShift Container Platform control plane nodes, hosting the Kubernetes API server, etcd, and scheduler. These nodes are dedicated to cluster management and do not run AI workloads. OpenShift natively manages the full cluster lifecycle.
- **GPU Compute Nodes – MGX or AC520 (2 to 16 nodes)**: Function as OpenShift worker nodes. Each server connects to the 100 GbE storage fabric for access to the Hammerspace global file system. The NVIDIA GPU Operator automatically provisions GPU resources as nodes join the cluster. Customers select MGX nodes for entry-level to midrange workloads, or AC520 nodes for larger-scale AI training and inference.
- **Hammerspace Storage Nodes**: Two HA815 G6 Anvil nodes form the HA metadata cluster across the MGX and AC520 configurations. Two to four HA815 G3 DSX nodes provide the scalable data plane. Networking connectivity and integration with VSP One Block remain consistent across configurations.
- **VSP One Block 28**: Connects directly to the DSX nodes through NVMe/TCP over a dedicated 100 GbE backend network. Additional VSP One Block systems can be added to scale storage capacity and throughput linearly.

OpenShift AI Deployment

Red Hat OpenShift Container Platform is deployed on the Hitachi iQ hardware platform in a bare-metal configuration using the standard OpenShift installer. GPU worker nodes (MGX or AC520 systems) are enrolled as OpenShift worker nodes.

After the cluster is established, the following operators are deployed from OperatorHub:

- Node Feature Discovery (NFD) Operator: Labels nodes with hardware attributes, including GPU type, kernel version, and PCIe card inventory
- NVIDIA GPU Operator: Provisions the GPU driver stack, device plug-in, container runtime, and DCGM monitoring on GPU worker nodes
- Red Hat OpenShift Serverless: Required for KServe single-model inference
- Red Hat OpenShift Service Mesh: required for KServe traffic management and routing
- Red Hat OpenShift AI Operator: Installs the complete OpenShift AI platform, including workbenches, model serving, AI pipelines, Llama Stack, and observability

Persistent storage for OpenShift AI workbenches, model serving, and pipeline artifacts is provisioned from the Hammerspace global file system using the Hammerspace CSI driver or the Kubernetes NFS CSI driver, giving AI engineers access to the same high-performance parallel file system used by training jobs directly from their interactive notebook environments.

Note: Hardware Profiles in OpenShift AI allow administrators to define node selectors for specific GPU hardware. In a mixed deployment where MGX and AC520 nodes are present, administrators can configure separate hardware profiles to route lighter inference workloads to MGX nodes and larger training or model-serving workloads to AC520 nodes – all managed from the unified OpenShift AI dashboard.

Storage Architecture – Hammerspace Global File System

The storage architecture is identical across both compute configurations. All GPU compute nodes – whether MGX or AC520 – access a single Hammerspace global file system, through standard protocols.

© Hitachi Vantara LLC 2026. All Rights Reserved.

- **AI Training (NFS / pNFS Flex Files):** GPU worker nodes mount the Hammerspace file system using NFSv4.1 with pNFS Flex Files. Training jobs read dataset shards and write checkpoints in parallel directly across multiple DSX nodes, with throughput scaling linearly with the number of DSX nodes.
- **Interactive Workbenches (NFS):** OpenShift AI workbench persistent volumes are provisioned from Hammerspace through NFS PersistentVolumes, giving data scientists direct, shared access to training datasets from their JupyterLab environments.
- **Windows / SMB Access:** Data science workstations and data labeling tools running Windows mount shares from the same global file system over SMB3.
- **S3 Access:** OpenShift AI data connections, model storage, and pipeline artifacts access Hammerspace through its S3 gateway.
- **VSP One Block Backend (NVMe/TCP):** DSX nodes connect directly to VSP One Block 28 over a dedicated 100 GbE NVMe/TCP network.

Note: Key architectural principle: Hammerspace separates the metadata path (Anvil nodes) from the data path (DSX nodes). Clients negotiate file locations with Anvil, then transfer data directly to DSX nodes — Anvil is never in the data path. This enables linear performance scaling as DSX nodes are added and delivers parallel file system throughput through standard NFS/SMB/S3 protocols.

© Hitachi Vantara LLC 2026. All Rights Reserved.

Solution Validation – Use Cases

The following use cases apply to both the MGX and AC520 compute configurations. The Red Hat OpenShift AI platform, Hammerspace storage architecture, and operational workflows are identical across both – with the primary difference being the GPU capacity available per node.

Use Case 1: Generative AI Inference with Red Hat AI Inference Server

This use case demonstrates deploying a production-ready generative AI inference endpoint using the Red Hat AI Inference Server (enterprise vLLM) and llm-d on the Hitachi iQ hardware platform.

Prerequisites

- Red Hat OpenShift cluster deployed on Hitachi iQ hardware platform (with the control plane running on HA810 G3 and GPU worker nodes on MGX or AC520)
- Node Feature Discovery and GPU Operator deployed and running on GPU worker nodes
- Red Hat OpenShift Serverless and Red Hat OpenShift Service Mesh deployed
- Red Hat OpenShift AI Operator deployed with DataScienceCluster created
- Hammerspace file system accessible from worker nodes through NFS for model artifacts and data storage

Deployment Steps

Step 1. Create a Data Science Project

From the OpenShift AI dashboard, create a Data Science Project (a dedicated Kubernetes namespace) to contain the model server, workbenches, and associated storage.

Step 2. Deploy a Serving Runtime

Within the project, configure the Red Hat AI Inference Server as the model serving runtime. Select a model from the Red Hat AI model catalog or provide a custom model stored on the Hammerspace file system. Configure the GPU count, replica count, and persistent volume size (backed by Hammerspace NFS). Use a Hardware Profile to target MGX nodes for smaller models or AC520 nodes for larger or more resource-intensive models.

Step 3. Enable llm-d for Distributed Inference (Optional)

For large language models that exceed the memory of a single GPU node, enable llm-d to distribute inference across multiple nodes. llm-d integrates with the Kubernetes Gateway API Inference Extension to provide intelligent request routing and consistent response times.

Step 4. Create a Workbench and Test Inference

Create a JupyterLab workbench within the same project. The workbench persistent volume is provisioned from Hammerspace. Use the model server's internal endpoint URL to issue inference requests from notebook code, validating the deployment before exposing it to downstream applications.

© Hitachi Vantara LLC 2026. All Rights Reserved.

Result

Teams can deploy a production-ready inference endpoint in minutes. The GPU Operator handles driver provisioning automatically, model artifacts are served from the Hammerspace global file system without requiring data movement, and usage and performance are monitored through OpenShift AI's built-in observability dashboards, including time-to-first-token latency and throughput metrics.

Use Case 2: Model Fine-Tuning with OpenShift AI

This use case demonstrates fine-tuning a pretrained foundation model using OpenShift AI workbenches and the GPU resources of the Hitachi iQ. The Hammerspace file system provides the parallel file system throughput required to keep GPUs fully used during training.

Workflow

Step 1. Import Training Dataset

Create a JupyterLab workbench with a PyTorch notebook image and a data connection to an S3-compatible bucket on Hitachi VSP One Object. Download and extract the training dataset into the Hammerspace file system. Because Hammerspace presents a single global namespace, the dataset is immediately accessible to all GPU worker nodes without requiring data duplication.

Step 2. Fine-Tune the Model

Load a pretrained foundation model and run the training loop. The GPU Operator ensures GPU resources are available to the PyTorch training process. Hammerspace delivers dataset shards in parallel to all training workers through pNFS Flex Files, eliminating the storage I/O bottleneck. Model checkpoints are written to Hammerspace, providing durability and enabling training to resume seamlessly if interrupted.

Step 3. Evaluate, Export, and Register

Evaluate the model against a validation set, export it in a portable format (such as ONNX or Hugging Face SafeTensors), and register it in the OpenShift AI Model Registry. The model is then discoverable, versioned, and reusable by other teams and projects with no additional data movement.

Step 4. Deploy the Retrained Model

Deploy the fine-tuned model through the Red Hat AI Inference Server directly from the Model Registry. Because the model is already stored on Hammerspace, no data transfer is required between training and serving.

Result

The Hammerspace file system enables seamless data flow from training through serving — the same global namespace supports dataset access, checkpoint storage, model registration, and inference serving, all without requiring data copies or transfers between separate storage systems.

© Hitachi Vantara LLC 2026. All Rights Reserved.

Use Case 3: Agentic AI with Llama Stack and MCP

This use case demonstrates how to use Red Hat OpenShift AI's Llama Stack capabilities to build an AI agent grounded in enterprise knowledge — the foundational pattern for enterprise RAG and agentic AI applications.

Workflow

Step 1. Deploy a Llama Stack Server

Enable the Llama Stack Operator in the OpenShift AI DataScienceCluster custom resource. The operator deploys a Llama Stack distribution that includes the Red Hat AI Inference Server (vLLM), a vector store (pgvector, FAISS, or qdrant), and MCP server support within the OpenShift cluster.

Step 2. Ingest a Knowledge Base

Store enterprise documents on the Hammerspace file system. Use docling (integrated with OpenShift AI's RAG pipeline tooling) to ingest, chunk, and embed the documents, storing the resulting vectors in the configured vector database.

Step 3. Build and Test the Agent through Notebook and the OpenShift AI Playground

Using a JupyterLab workbench, interact with the deployed Llama Stack server through its OpenAI-compatible API or the Llama Stack Python SDK. Issue chat completions against the running model, enable the vector store tool to retrieve grounded responses from the knowledge base, and enable safety shields (Llama Guard for content safety; Prompt Guard for injection detection) using the API. Compare responses with and without RAG enabled to validate proper knowledge grounding.

Optionally, the Gen AI Playground, available as a technology preview in OpenShift AI 3.3, provides a dashboard-based interface for the same experimentation without writing code; however, it should not be used for production workflows at this time.

Step 4. Connect External Tools Using MCP

Register MCP server connectors through the Llama Stack Operator. AI engineers can consume pre-registered connectors from the Playground or notebook code, enabling agents to call external enterprise tools without managing complex integration configurations.

Result

Red Hat OpenShift AI's Llama Stack and MCP capabilities enable teams to build production-ready, enterprise-grade AI agents grounded in enterprise knowledge and connected to enterprise tools — while maintaining safety guardrails and full operational visibility. The Hammerspace file system provides a high-performance, scalable storage layer for both the document ingestion pipeline and the inference server.

© Hitachi Vantara LLC 2026. All Rights Reserved.

Storage Performance

Hammerspace with VSP One Block 28 scales linearly with the number of DSX nodes. These performance numbers apply equally to both the MGX and AC520 compute configurations — as the storage architecture is identical. Data was measured in a validated lab environment with small (1 TB) and large (16 TB) datasets, with 2x file replication per DSX node pair.

Configuration	Rnd Read (K/IO)	Rnd Write (K/IO)	Sequential Read (GB/s)	Sequential Write (GB/s)
2x DSX Nodes	502.58	98.62	14.69	1.92
4x DSX Nodes	1274.69	183.12	58.13	3.57

Table 5 Performance scaling from 2 to 4 DSX nodes across random and sequential workloads.

- Random read performance scales approximately 2.5x from 2 to 4 DSX nodes, driven by the parallelized pNFS Flex Files architecture.
- Sequential read throughput scales linearly from 2 to 4 DSX nodes. This is the critical metric for AI training workloads, where data loaders read sequentially through sharded training datasets.
- The 4-DSX configuration delivers approximately 58 GB/s of sequential read throughput — sufficient to continuously feed large GPU clusters across MGX and AC520 node types.
- Adding additional VSP One Block systems and DSX nodes extends both capacity and performance, following the same linear scaling model.

Note: Sequential read throughput is the critical storage metric for AI model training. The 4-DSX configuration's 58 GB/s throughput is sufficient to keep even large GPU clusters continuously fed — whether those GPUs are NVIDIA RTX PRO Blackwell units in MGX nodes or H200 NVL units in AC520 nodes.

© Hitachi Vantara LLC 2026. All Rights Reserved.

Conclusion

The Hitachi iQ M Series and Hitachi iQ MGX with Hammerspace and Red Hat OpenShift AI delivers a complete, validated enterprise AI platform covering the full AI lifecycle — from data ingestion and distributed model training through fine-tuning, production inference, and agentic AI.

By offering two GPU compute platforms — the Hitachi MGX for entry-level to midrange workloads and the Hitachi AC520 for midrange to high-end workloads — the Hitachi iQ M Series and Hitachi MGX allows organizations to match compute investment to their actual AI requirements at each stage of adoption. Critically, both platforms share the same Hammerspace storage infrastructure and Red Hat OpenShift AI software stack, protecting infrastructure investments and simplifying operations as workloads grow or compute platforms evolve.

Hammerspace Hyperscale NAS eliminates the data bottlenecks that commonly limit GPU utilization during training, while providing unified access to the same datasets and model artifacts from every component of the AI platform through a single, global namespace. Red Hat AI 3 provides an open, enterprise-grade AI platform that avoids lock-in to proprietary AI software, giving organizations the flexibility to choose any model from any provider and deploy on any hardware accelerator.

Key benefits of this reference architecture:

- Two GPU compute platforms serving entry-level through high-end AI workloads — MGX and AC520 — on a shared, unified storage and software stack
- Validated, modular architecture that scales from 2 to 16+ GPU compute nodes on either platform
- Linear storage performance scaling with additional Hammerspace DSX nodes — applicable to MGX and AC520 deployments
- Full AI lifecycle support: data preparation, training, fine-tuning, serving, and monitoring on a single platform
- Agentic AI readiness with Llama Stack, MCP, and the OpenAI-compatible Llama Stack API
- Open-source AI platform (Red Hat AI 3) — any model, any accelerator, any cloud
- Unified enterprise support across compute, storage, and AI platforms from Hitachi Vantara and Red Hat

Next Steps

- **Hitachi Vantara:** Contact your Hitachi Vantara account team to discuss MGX and AC520 configurations, Hammerspace storage sizing, and proof-of-concept engagement options.
- **Red Hat OpenShift AI:** Visit red.ht/openshift-ai for documentation, trial access, and guided tutorials for deploying Red Hat OpenShift AI.
- **Red Hat AI 3:** Visit redhat.com/en/products/ai to explore the full Red Hat AI portfolio, including Red Hat AI Inference Server, RHEL AI, and OpenShift AI.
- **Hammerspace:** Visit hammerspace.com for technical documentation on the Global Data Environment and Hyperscale NAS capabilities.

© Hitachi Vantara LLC 2026. All Rights Reserved.

About the Vendors

Hitachi Vantara

Hitachi Vantara, a wholly owned subsidiary of Hitachi, Ltd., guides our customers from what's now to what's next by solving their digital challenges. Working alongside each customer, we apply our unmatched industrial and digital capabilities to their data and applications to benefit both business and society. More than 80% of the Fortune 100 trust Hitachi Vantara to help them develop new revenue streams, unlock competitive advantages, lower costs, enhance customer experiences, and deliver social and environmental value.

The Hitachi iQ portfolio provides flexible, high-performance AI infrastructure specifically engineered to simplify the enterprise AI journey — delivering improved time to value, better economics at scale, and the flexibility to meet the demands of any enterprise AI workload. The Hitachi iQ hardware platform, spanning the MGX and AC520 compute platforms, is backed by Hitachi Vantara's award-winning global support and professional services organizations.

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 AI is Red Hat's portfolio of solutions for building, deploying, and managing AI workloads across the hybrid cloud — including Red Hat AI Inference Server, Red Hat Enterprise Linux AI (RHEL AI), and Red Hat OpenShift AI.

Red Hat AI 3 brings together the latest innovations across the Red Hat AI portfolio to simplify the complexities of high-performance AI inference at scale, enable agentic AI workflows, and empower organizations to move AI workloads confidently from proof of concept to production — on any hardware, in any environment.

Hammerspace

Hammerspace is a software-defined data management company that provides a global data environment spanning on-premises, multi-cloud, and edge environments through its Hyperscale NAS solution. Hammerspace enables organizations to orchestrate, manage, and access data globally through a unified parallel file system – delivering the data agility and performance required for modern AI and high-performance computing workloads.

© Hitachi Vantara LLC 2026. All Rights Reserved.