Omdia
by informa techtarget

OCTOBER 2026

Architecting the Modern Edge with Cisco and Red Hat

Scott Sinclair, Senior Research Director; and Monya Keane, Senior Analyst

Abstract: Cisco and Red Hat have partnered to transform IT infrastructure at the edge. They are delivering multiple advances in high-performance compute, automated lifecycle management, data storage, and data security to run real-time AI inference workloads at scale while supporting existing business-critical legacy applications. Cisco and Red Hat support industry leaders in storage such as NetApp, Everpure Portworx, and Qumulo, as well as native OpenShift Data Foundation (ODF). Combining Cisco Unified Edge with Red Hat OpenShift and Red Hat AI Inference delivers a flexible, enterprise-grade blueprint designed to serve large language models (LLMs) and more targeted small language models locally, directly at network edge sites without introducing the complexity that comes with fragmented architectures.

Introduction

Businesses in nearly every industry are modernizing edge infrastructures to stay competitive in the AI era. The effort must, however, ensure continued support for existing virtual machine (VM) application environments and balance next-generation capabilities with simplicity to support existing operations. For example:

- In **industrial manufacturing**, organizations are seeking to deploy emerging AI apps (such as using computer vision quality control for high-speed defect detection), but also to deliver the optimal user experience for existing applications, such as traditional inventory management.
- In **financial services**, new investments in AI are helping to deliver real-time fraud detection via sub-millisecond ATM inference and local transaction scoring, while basic transaction routing continues to run on existing VMs.
- In **healthcare**, medical imaging diagnoses are benefiting from local AI-powered scan analysis, while the existing VM environment is still required to run the underlying imaging software.
- In **retail**, businesses are leveraging local AI to enhance loss prevention and personalize the shopper experience through real-time shelf analytics, all while maintaining the high availability of critical, long-standing VM-based point-of-sale (POS) and legacy store management systems.

As businesses increase their AI investments, competitive success will depend upon high-value data generated at the edge. Because of its relationship to other mission-critical business operations, edge data presents a treasure trove of value—value that is unfortunately often trapped in traditional applications and siloed infrastructures. Therefore, success in the AI era requires infrastructure and application platform modernization at the edge that can provide significant enhancements to scalability and flexibility while maintaining support for existing VM-based apps. Cisco and Red Hat are two leaders in the industry that have partnered to provide a solution based on an integrated, validated design to simplify environments while providing scalability and flexibility to serve current and future needs.

This Showcase from Omdia was commissioned by Cisco and Red Hat and is distributed under license from TechTarget, Inc.

© 2026 TechTarget, Inc. All Rights Reserved

Architecting the edge for an AI future

Rampant AI investment has been leading IT decision makers to recognize the need to transform their edge environments. According to the research, 80% of organizations agreed that AI initiatives are significantly changing their edge infrastructure plans and designs.[1] For example:

- Infrastructure and application platforms must be able to **scale** to support application modernization efforts for current or eventual deployment of agentic AI workloads, and they must free up data to better tune private models, all while reducing the burden on administrators.
- Similarly, the uncertainty that accompanies the rapid pace of innovation around AI means new investments must provide **flexibility** to facilitate quick integration of new technologies, tune the infrastructure to the specific needs of the site, and provide management consistency to simplify collaboration and smooth operations across teams.
- A modern infrastructure must be able to **run traditional VM applications** but also allow for a seamless transition to modern container workloads. Many organizations are exploring alternative hypervisors and want to be ready for AI container workloads. They need a solution that can do this in a very simple way.

However, that need for edge-related transformation comes at a time when software and hardware complexity already hinder operations and limit collaboration between IT and operational technology (OT) teams. Consider that:

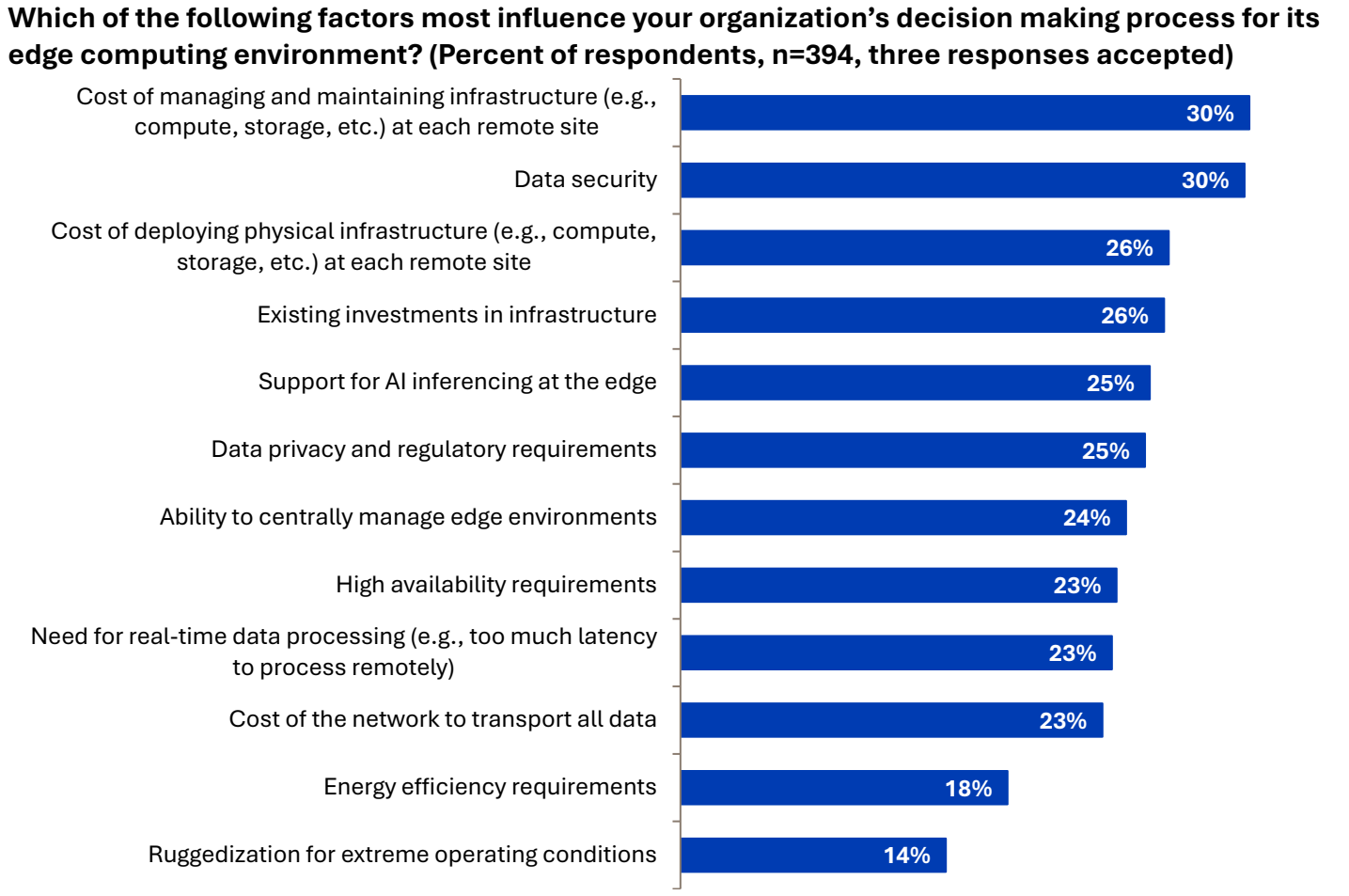
- **66%** of organizations agreed that maintaining software or hardware at the edge is complex and hinders operations.
- **69%** of organizations agreed that collaboration between IT and OT teams is often complex at edge environments.

When Omdia asked decision makers to identify their top three decision factors related to edge infrastructure (see **Figure 1**), they identified management costs (cited by 30% of respondents) and security (30%) as top-tier concerns. Those concerns were accompanied by multiple other factors, including the need to support AI (25%), centrally manage edge environments (24%), and manage the cost of the network (23%).

[1] Source: Omdia Research Report, *IT Modernization Report: Cloud, Applications, and Infrastructure Amid the AI Revolution*, August 2026. All Omdia research references and charts in this Showcase are from this report unless otherwise noted.

©2026 TechTarget, Inc. All Rights Reserved.

Figure 1. Top decision factors for edge infrastructure



Source: Omdia

Consideration factors most important for the modern edge

Functionality, availability, and resilience will always be priorities for edge infrastructures. But the importance of ensuring that the organization can optimize current and future edge AI initiatives makes certain capabilities even more valuable. Those capabilities are described as:

- **Future-proof solutions** that offer application modernization and infrastructure scalability and support/stability for existing workloads (which are still required at the sites) while minimizing operational complexity. Ensuring scalability across the three dimensions of compute (performance, storage, and networking) is essential to supporting future application demands, including agentic workloads, while minimizing budget impact.

  On the application side, solutions must provide an integrated path to support existing, traditional VM-based applications, as well as cloud native, container-based, AI, and agentic workloads. 77% of organizations expect to have at least some of their deployments on VMs and containers residing on the same compute clusters in two years.

©2026 TechTarget, Inc. All Rights Reserved.

Improved scalability for both infrastructure capabilities and application support also increases the consolidation potential of the infrastructure, simplifying the environment by reducing the number of systems the IT and OT teams must manage.

- **Modular solutions for cost optimization and flexibility.** The pace of AI development and the diversity of use cases mean infrastructure design is often an uncertain effort. As organizations scale AI operations at the edge, it will be imperative to embrace solutions that offer flexibility and modularity to best tune the infrastructure to the specific needs of the site. Having the flexibility to deploy multiple processing options, increase networking bandwidth, or adjust storage density is especially important to ensure optimization for emerging, as-yet-undefined AI workloads. Modular, single-chassis solutions can provide the ability to tune a deployment to the specific needs of a site, adjust it when demands change, and attain the management consistency necessary to simplify overall operations.
- **Solutions with advanced data and physical security.** The rise of next-generation threats, including AI-driven attacks and emerging quantum computing risks, increases the complexity of edge infrastructure design and management. Investments must help ensure long-term data integrity, operational resilience, and the protection of key assets without slowing down business operations. Solutions that can simplify remote update management, especially with automation, help teams to deliver fixes more consistently, which is essential to ensuring cyber resilience across complex edge environments.

Cisco and Red Hat partner to offer a unified edge

Cisco and Red Hat have partnered to offer a joint solution combining Cisco Unified Edge, Red Hat OpenShift, and Red Hat AI Inference. The solution is deployed on Cisco Unified Edge infrastructure as a single, three-rack unit (3U) chassis that integrates compute, networking, and storage. Multiple processing/accelerator options are available with support for Intel Xeon and NVIDIA GPUs.

The solution provides support for up to five individual nodes, with the ability to mix and match node types. For example, a single chassis could be configured to support Kubernetes clusters, custom bare-metal applications, and networking infrastructure, all tailored to the needs of an edge use case. Offering a consistent platform to deploy at the edge to accommodate whatever the organization's workload needs are at any time is a major differentiator.

The solution is also designed to run Red Hat OpenShift, the Ansible Automation Platform, Red Hat Enterprise Linux, and Red Hat AI Inference, pre-validated to suit the infrastructure. Red Hat offers a consistent platform from the core to the edge, and its OpenShift solution has been shown to provide an up to 636% ROI over five years.[2] This solution features differentiators that can deliver significant value to organizations modernizing edge environments, including the following:

- **A highly customizable, modular infrastructure.** Cisco's Unified Edge modular design enables organizations to scale performance, capacity, and networking up or down to align with the specific needs of the edge location. The scalable modular architecture allows from one to five nodes, supporting a wide variety of needs while providing an option to start small and grow with demand. OpenShift supports diverse edge topologies, including compact edge clusters running three-node high availability with

[2] Source: Red Hat OpenShift Overview, "Gain business value with Red Hat OpenShift Virtualization," October 2024.

©2026 TechTarget, Inc. All Rights Reserved.

combined control and worker nodes. Cisco's chassis is also designed to offer a long lifespan to ensure organizations can maximize the value of their investments.

- **An application platform designed for the needs of today and the upcoming requirements of the AI era.** OpenShift enables administrators to manage VMs and containers side by side. Ansible provides automation and agentic operational tools to simplify and automate infrastructure management, reducing the burden on IT and OT personnel. For production AI applications, the integration of Red Hat AI Distributed Inference with llm-d assists in the selection and management of models, including open source and local models. The combination provides a software platform that can run a tailored mix of existing legacy apps while providing a foundation to support future AI use cases. By integrating Red Hat AI Inference serving frameworks with continuous request batching or distributed inference with llm-d, the combined solution enables organizations to scale AI initiatives more efficiently, improving user experience. Separating the prefill and decode phases helps further improve the AI user experience by helping to maximize throughput and optimize latency.
- **Unified fleet-based management.** Cisco offers centralized management via Cisco Intersight, which can scale up to thousands of sites at a global scale. With integrated automation and observability functionality, Cisco's approach minimizes the time and costs associated with maintaining and updating infrastructure. It accelerates troubleshooting and facilitates collaboration between IT and OT teams via a consistent set of tools.
- **Data and infrastructure security**. Cisco's multilayered security capabilities are designed to protect data, models, and infrastructure at the edge. From a management perspective, SaaS-based Cisco Intersight management safeguards against unverified access. The hardware offers physical security (including intrusion detection and locking bezel), along with digital protection (such as signed firmware to ensure verified updates).

  Automated policy enforcement helps eliminate configuration drift. Network security capabilities provide threat defense and network segmentation features to further secure edge infrastructures. Red Hat provides additional security layers to better protect the business against AI-driven attacks or quantum-based risks.

Conclusion

Taken together, edge computing and AI may herald one of the most significant infrastructure confluence points of the decade. The Cisco-Red Hat partnership exemplifies the kind of strategic collaboration required to address a transformation like this one. What distinguishes the Cisco Unified Edge with Red Hat solution is not only its technical specifications but also its acknowledgment of a competitive marketplace reality: Too many organizations can't afford to rip and replace existing infrastructure, but can't afford to sit still either.

IT leaders today face an important choice. While 80% of them acknowledged that AI is reshaping their edge strategies, two-thirds simultaneously reported that maintaining software or hardware at the edge is complex and hinders operations. The infrastructure modernization needed to capture AI's value compounds operational challenges, and the Cisco-Red Hat approach addresses that dilemma architecturally, supporting legacy VM workloads alongside modern containerized applications within a fluid, unified management framework, enabling modernization without disruption.

©2026 TechTarget, Inc. All Rights Reserved.

The modular chassis design of this solution improves edge infrastructure economics, enabling right-sizing for specific site requirements while maintaining upgrade pathways as workloads evolve. Red Hat AI Inference capabilities with enterprise-grade orchestration through OpenShift position organizations to operationalize AI at scale. The security of OpenShift is built on defense-in-depth, including Federal Information Processing Standards (FIPS) compliance and post-quantum cryptography readiness.

This is a big deal because success in edge AI may help to create market leaders in multiple industries going forward. Manufacturing, financial services, healthcare, and numerous other sectors will find competitive and other vital advantages in processing and acting on intelligence where that intelligence originates—stored and waiting at the edge. The Cisco Unified Edge with Red Hat solution is a production-ready answer to the challenge of capitalizing on it.

Copyright notice and disclaimer

The Omdia research, data, and information referenced herein (the "Omdia Materials") are the copyrighted property of TechTarget, Inc. and its subsidiaries or affiliates (together "Informa TechTarget") or its third-party data providers and represent data, research, opinions, or viewpoints published by Informa TechTarget and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice, and Informa TechTarget does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an "as-is" and "as-available" basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness, or correctness of the information, opinions, and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa TechTarget and its affiliates, officers, directors, employees, agents, and third-party data providers disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa TechTarget will not, under any circumstance whatsoever, be liable for any trading, investment, commercial, or other decisions based on or made in reliance of the Omdia Materials.

Get in touch: www.omdia.com askananalyst@omdia.com

© 2026 TechTarget, Inc. All rights reserved. Unauthorized reproduction prohibited.