Accelerate AI time to value with Red Hat and Cisco

AI opportunity is here but scaling to production remains a challenge

AI adoption across the enterprise has grown rapidly in recent years, bringing new capabilities and opportunities to build value for organizations of all sizes and industries.

Finance, healthcare, retail, and manufacturing organizations are seeing the benefits of using AI to automate key tasks and support better decision making, while the transportation, energy, agriculture, and entertainment industries are optimizing logistics and personalizing content, among many other emerging industry-specific use cases.

This has also resulted in a wave of valuable predictive AI and gen AI use cases being employed over industries of all types, including AI assistants, content generation (text, images, video, and code), visual computing, real-time multilingual translation, data insights and forecasting, domain-specific chatbots, and much more.

However, many organizations still face significant challenges as they look to scale their AI projects from pilots to production, resulting in rising project costs and delayed timelines.

AI models and frameworks, the applications being built with them, and the infrastructure supporting all of this, have created a new cyberattack surface and opened up new security vulnerabilities organizations must address. Model training and inferencing is increasing network traffic, which is slowing down efficiency and delaying time to value. And, most consequentially, a lack of high-performance infrastructure with integrated compute, network, and storage capabilities is increasing complexity and stalling AI projects.

These challenges are being further exacerbated by growing skill gaps and evolving data privacy and governance requirements. This has left organizations in need of an integrated AI platform that can help them optimize AI costs, performance, and resource use, while mitigating security and compliance risks, to accelerate time to value.

Why choose a prevalidated Cisco AI POD with Red Hat over a DIY approach?

Some organizations consider piecing together various hardware and software components for their AI infrastructure, but this often leads to unforeseen complexities and delays, hidden costs, and increased security risks. Effectively integrating disparate systems requires significant expertise, testing, and ongoing troubleshooting, which diverts valuable resources away from focusing on AI innovation.

Cisco AI PODs with Red Hat eliminate this burden by providing a prevalidated, full-stack AI infrastructure solution—delivered as a jointly engineered and rigorously tested reference architecture built on Cisco Validated Designs (CVDs).

This integrated approach helps optimize AI value with:

- **Accelerated deployment and reduced complexity.** Streamlined plug-and-play deployment reduces setup time by up to 50%[1] and minimizes misconfigurations, resulting in accelerated time to value.

[1] "Cisco AI PODs: Pre-validated, Flexible, and Modular AI Infrastructure Data Sheet." Cisco, 7 Aug. 2025."

- **Optimal performance and scalability.** Interoperability between Cisco's high-performance networking and compute, and Red Hat's orchestration and AI lifecycle management platforms, allows AI workloads to run efficiently and scale predictably, with streamlined training and inferencing.
- **Mitigated risk and enhanced reliability.** Extensive testing and validation for interoperability and performance helps mitigate deployment risks and delivers consistent, reliable operations for critical AI initiatives.
- **Focus on security and compliance by design.** Validated security and governance is embedded in every layer of the joint solution, with built-in capabilities offered by Cisco AI Defense and Cisco Hypershield, to help manage compliance requirements, safeguard data, and address evolving cyber threats.
- **Simplified operations and unified support.** A unified operational model and consolidated support experience (backed by the Cisco Technical Assistance Center) simplifies ongoing management and provides accelerated issue resolution, allowing IT teams to focus on strategic initiatives rather than addressing disparate operational issues.

Feature	DIY approach	Cisco AI PODs with Red Hat
Deployment speed	Months of manual integration and testing.	Streamlined setup via plug-and-play deployment.
Operational effort	Significant time spent on troubleshooting and firefighting.	Unified operational model with Cisco Technical Assistance Center support and Red Hat automation.
Reliability	Risk of misconfigurations and performance bottlenecks.	Prevalidated, rigourously tested reference architectures (CVDs).
Security and compliance	New cyberattack surfaces and fragmented security layers.	Security by design with Cisco Hypershield and Red Hat AI guardrails.
Scalability	Stalled projects due to lack of integrated compute/network/storage.	Modular scale-unit design that grows predictability with Kubernetes.
Training and interference	Network traffic congestion delaying time to value.	Optimized VLLM and Nexus networking for high-performance throughout.

Figure 1. Comparing a DIY approach to a prevalidated path from Red Hat and Cisco.

Overcome AI complexity with prevalidated architecture from Red Hat and Cisco

Cisco AI PODs with Red Hat® OpenShift® and Red Hat AI offer enterprises a unified, production-ready, security-focused AI foundation, rather than having to rely on disparate tools and platforms that can impede productivity and value.

This jointly validated reference architecture brings together Cisco's high-performance networking, security, compute—including Graphics Processing Unit (GPU)-optimized servers—and observability with the orchestration and AI lifecycle management capabilities of Red Hat's application and AI platforms. It streamlines project deployments, mitigates risks, optimizes resource use, and simplifies operations on a consistent, pregoverned environment with built-in security capabilities and scalable AI tooling.

Operating within Cisco's integrated hardware and software stack, Red Hat OpenShift provides a flexible, consistent, and scalable application platform for containerized AI workloads across complex IT environments, including on-premise datacenters, hybrid or multicloud environments, and devices operating at the edge of a network.

Red Hat AI builds on Linux® and Kubernetes to provide a foundation for building, fine-tuning, training, deploying, and monitoring predictive and gen AI models, as well as AI agents, at scale and with a focus on security. It allows organizations to bring their own AI models, download popular open source models, and integrate with multiple AI/ML (machine learning) libraries, frameworks, and runtimes to connect models with private enterprise data.

Red Hat AI combines a sophisticated and powerful inference runtime (vLLM) and an innovative distributed inference framework (llm-d) to optimize model inference across the hybrid cloud—leading to more cost-effective and consistent deployments. It also provides consistent management of AI models across all IT environments, the ability to expand DevOps practices to the entire AI/ML lifecycle, and the scalability to meet the demands of modern AI workloads.

Cisco AI PODs with Red Hat are fully configurable to an organization's use cases and infrastructure strategy. It is built on a series of core layers, while also offering the option to add third-party solutions, such as storage and security, and a number of other customizable components. This includes:

- High-performance networking offered by Cisco Nexus.
- Computing power and GPU-powered servers delivered by Cisco Unified Computing System (UCS).
- Enterprise-grade Kubernetes deployment and management capabilities provided by Red Hat OpenShift.
- Popular open source AI tools for training, tuning, deploying, and running AI models, AI-powered applications, and AI agents provided by Red Hat AI.
- An operations software layer built on Cisco Intersight and Cisco Nexus Dashboard or Cisco Nexus Hyperfabric (depending on the operational model), and supported by enterprise-wide wide automation from Red Hat Ansible® Automation Platform.
- Optional integrations, including storage solutions offered by Cisco's partner ecosystem (such as VAST, NetApp, Nutanix, Pure Storage, Hitachi, and more), security solutions from Cisco, observability tools from Splunk, and workload management and operations tools from Isovalent, Slurm, and others.

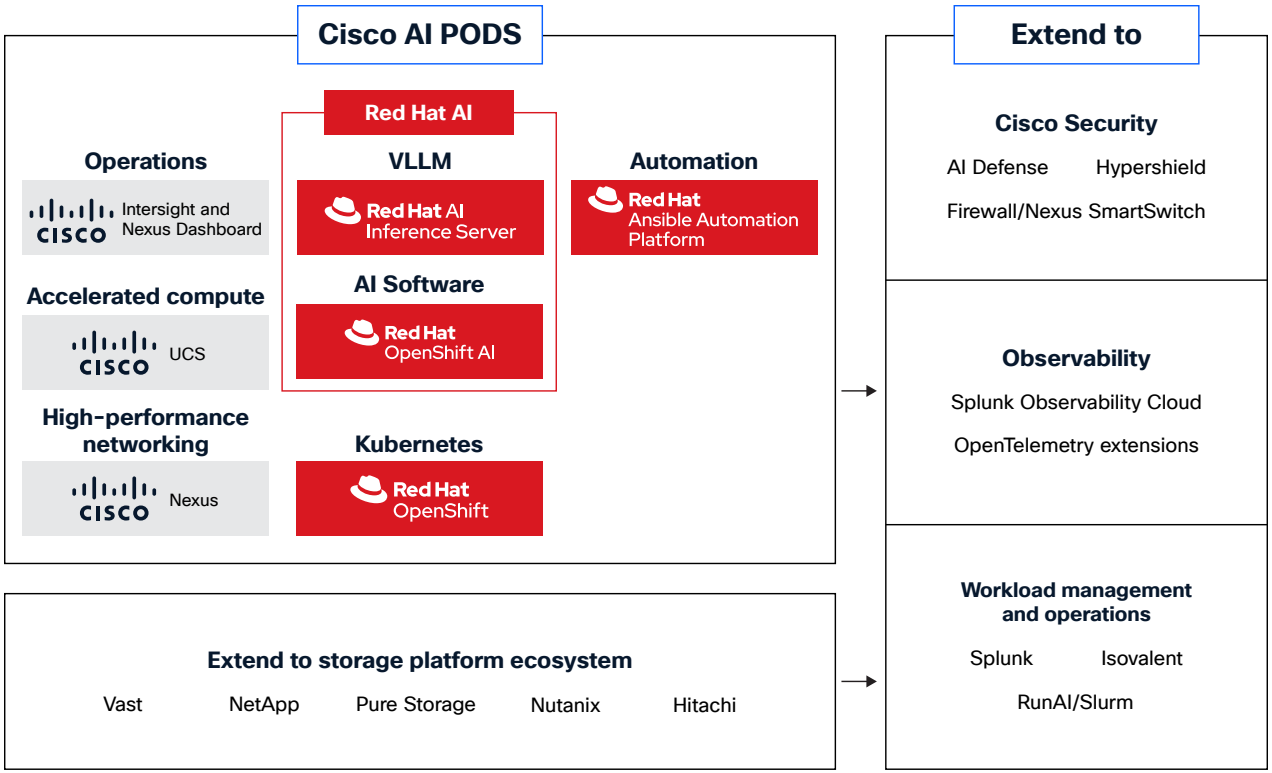

Figure 2. Cisco AI PODs with Red Hat AI full-stack infrastructure for AI training, optimization and inferencing

Accelerate AI innovation on a unified, production-ready platform

Red Hat and Cisco have jointly engineered this solution to deliver a preintegrated and ready-to-use platform that combines high-performance hardware with a flexible, cloud-native application and AI platform. It provides organizations with everything they need to accelerate time to value as they move from pilot to production, including unified infrastructure, streamlined deployment, AI-optimized security features, and more.

Unified, scalable architecture for any AI workload

This comprehensive AI platform from Red Hat and Cisco brings together high-performance compute, networking, and management capabilities and a consistent, cloud-native foundation for building, training, deploying, monitoring, and scaling AI models across complex environments.

It provides lossless, low-latency networking and sustained throughput for AI model training, fine-tuning, and inferencing, and supports the evolving needs of modern AI workloads through a modular scale-unit design and the proven scalability of Kubernetes-powered Red Hat OpenShift.

Streamlined deployment through validated integration

As part of Cisco's collection of CVDs, this reference architecture offers fully integrated, tested, and validated AI infrastructure, with 100+ pages of documentation, including Ansible playbooks for automating deployment and Day 2 operations.

In tandem with its plug-and-play deployment, this helps organizations reduce setup time and complexity, including minimizing misconfigurations and mitigating related risks.

Flexible and efficient operations

Cisco AI PODs with Red Hat offer the flexibility to run AI-powered applications in containers or Virtual Machines (VMs), with centralized and unified management for traditional and emerging workloads. It reinforces this flexibility with Red Hat OpenShift delivering consistency across hybrid or multicloud environments, enterprise-wide automation capabilities offered by Ansible Automation Platform, and streamlined lifecycle management through Cisco Intersight.

Cisco Intersight provides a unified, cloud-native platform for managing the entire compute infrastructure, including Cisco UCS servers and GPU-optimized systems, with policy-based provisioning, deployment, and monitoring. This allows IT teams to centrally manage fleets at scale, automate complex tasks, and significantly reduce setup time for AI workloads across datacenters, colocation facilities, and distributed edge environments.

Reliable, security-focused infrastructure

This solution is jointly engineered and validated for interoperability and performance, and designed to provide consistent policy enforcement, governance, and observability across complex IT environments.

It provides an embedded and validated focus on security and governance at every layer, and offers the option to further reinforce this with proactive, AI-optimized security capabilities from Red Hat AI, Cisco AI Defense, Cisco Hypershield, Red Hat OpenShift, and a range of tested and integrated third-party security solutions.

Simplified choice and flexibility

Offered as a fully configurable solution within a range of Cisco AI POD bundle options, this flexible offering gives customers the ability to tailor their AI infrastructure to their specific use cases and infrastructure strategy.

It offers the flexibility to choose between a number of operational models. This includes a choice of operational models for hardware (e.g. Cisco UCS, AI accelerators) and software (e.g. Red Hat OpenShift, Red Hat AI), backed up by a comprehensive ecosystem of integrated partner hardware and software solutions, and 2 key operational models for network management, depending on an organization's needs.

1. **Air-gapped, on-premise networking management** with Cisco Nexus Switches and Cisco Nexus Dashboard, and a broad range of configurability options. This operational model, which is backed by Red Hat AI's ability to support air-gapped, on-premise deployments, meets the needs of organizations with high compliance requirements, such as those in the financial, government, or healthcare sectors.
2. **A cloud-managed operational model** that offers Cisco Nexus Hyperfabric as the network management platform, while keeping all data on premise. This provides a cloud-like experience for customers repatriating workloads, with a centralized dashboard for building design templates, autogenerating blueprints and Bills of Materials (BOMs), and accessing mobile-friendly deployment guides, as well as real-time connection validation and lifetime monitoring.

Build for the future of AI on a foundation that supports all use cases

Cisco AI PODs with Red Hat are built to meet the diverse AI needs of the present and future, with customizable solution designs delivered on a standardized, security-focused, and scalable foundation that supports all AI use cases.

The full range of Cisco AI POD solution bundles offers a number of different configurations, based on use cases and scale, with a comprehensive array of compute platforms to support all use cases. Each of these architecture options is fully tested and validated by Cisco, and include a full design and deployment guide.

Each solution bundle provides specifically curated designs that are fully optimized for model training, model optimization, and small-, medium-, and large-scale inferencing, as well as options to include additional security, observability, and data service solutions.

Model training. Red Hat and Cisco help organizations accelerate model training with high-performance compute and networking solutions that reduce training times and accelerate iteration, supported by an MLOps platform that provides distributed training frameworks and dynamic resource allocation and scheduling capabilities.

This joint solution offers the flexibility for organizations to bring their own proprietary AI models, while also providing access to optimized and validated third-party open models that run efficiently on vLLM across the platform, including meeting key performance benchmarks and accuracy evaluations.

Model optimization. Cisco AI PODs with Red Hat help organizations improve the accuracy and domain-specific relevance of their AI outputs by refining pretrained models with scalable and cost-effective model alignment tools and the ability to connect models to organizational data through Retrieval-Augmented Generation (RAG) for context-aware responses.

It also offers a Large Language Model (LLM) compressor toolkit, which helps organizations optimize for cost by lowering the size and computational requirements of a model, and a performance evaluation framework. The performance evaluation tooling allows organizations to benchmark artificial workloads on their AI models to test for performance with real-world traffic and concurrent requests, and receive actionable insights to help fine-tune their AI models.

AI inferencing. Cisco AI PODs with optimized inferencing from Red Hat AI help organizations improve the speed and accuracy of predictions and insights, streamline resource usage, and lower inference costs for AI-powered applications in diverse environments and at scale with optimized model inference across on-premise datacenters, hybrid or multicloud environments, and edge locations.

Inferencing capabilities are provided through high-performance inferencing frameworks, such as vLLM, that accelerate output of gen AI-powered applications by optimizing use of GPU memory to boost throughput reduce latency, and distributed inferencing frameworks, such as llm-d, that support cost-effective, predictable performance at scale for gen AI models and delivers improved performance as volume increases.

Security, observability, and data services. Red Hat and Cisco help organizations reinforce their security focus, extend their observability across complex environments, and bolster their data management with:

- Red Hat AI, which includes guardrails for inputs and outputs during inference, data bias and data drift detection, and more AI-specific security capabilities.
- Cisco AI Defense, which supports full application security, application and model validation, and runtime application protection.
- Splunk Observability Cloud Dashboard, which provides real-time monitoring and troubleshooting, actionable insights, efficient data ingestion, and telemetry data processing.
- VAST, which offers a universal data platform that simplifies full-stack orderability.

Discover what it takes to be AI-ready

Red Hat and Cisco both offer assessment tools to help organizations understand how prepared they are for the AI journey, what they need to do to get ready, and how Red Hat and Cisco can help.

Try the Red Hat AI Readiness Assessment.

Try the Cisco AI Readiness Index Assessment Tool.

Start accelerating AI time to value with Red Hat and Cisco

Access these resources to learn more about Cisco AI PODs, Red Hat OpenShift, Red Hat AI, Cisco Intersight, Nexus Dashboard, or how to operationalize AI more quickly.

Explore this CVD to learn more about the full-stack solution reference architecture based on Cisco AI PODs with Red Hat AI.

Contact your local Cisco, Red Hat, or partner sales consultant to discuss how this joint solution from Red Hat and Cisco can help your organization accelerate AI time to value.