# Transform Cisco AI PODs into a Self-service GPU cloud

Unlock Your AI Potential with Cisco and Rafay to deliver sovereign and enterprise AI clouds

# Contents

# Operationalizing Cisco AI PODs with Rafay GPU Platform as a Service

To truly operationalize AI, organizations need to expose their infrastructure through self-service, multitenant consumption models. In this section, we explore how the Rafay GPU Platform as a Service overlays Cisco AI PODs to deliver a fully operational GPU cloud platform, one that enables SKU-based provisioning, GPU slicing, quota enforcement, and AI workload catalogs “out of the box.”

**From infrastructure to consumption:** A platform-led model

Enterprises need a platform operating model that exposes infrastructure as a governed, reusable service. Cisco provides the AI POD foundation; Rafay transforms that foundation into a production-grade GPU cloud, completing the vision of a simple-to-use, flexible, and modular AI infrastructure by extending these benefits from hardware deployment to seamless, self-service consumption.

# 1.0 The problems this solution addresses

Enterprises have made significant strides in deploying robust AI infrastructure, with solutions such as Cisco AI PODs providing pre-validated, high-performance foundations that simplify initial deployment and offer inherent security and scalability. These foundational investments embody simplicity, flexibility, and modularity for the underlying hardware. However, as these platforms mature, new challenges emerge in maximizing their value and enabling widespread, efficient consumption across diverse teams.

## 1.1 AI infrastructure is here, but it's not easy to consume for enterprises

In the last 18 months, enterprises have accelerated investments in AI infrastructure, deploying platforms such as Cisco AI PODs equipped with the latest NVIDIA hardware to enable generative AI, RAG, fine-tuning, and large-scale inferencing. These systems form the technical backbone for modern AI initiatives.

As adoption grows, organizations are encountering a new challenge: how to enable multiple teams to safely and efficiently consume the shared infrastructure. Teams are tasked with balancing access across different groups, each with different needs, priorities, and security requirements. Without standardized interfaces, fine-grained quotas, or policy-based controls, this leads to ad-hoc workflows, ticket queues, and duplicated setup efforts. What emerges is a need for a platform model where infrastructure is designed for multi-team AI consumption.

## 1.2 Maximizing GPU utilization in enterprises

While enterprises have invested considerably in high-performance GPU infrastructure, real-world usage often does not align with initial expectations. Various industry studies indicate that GPU utilization rates tend to be lower than anticipated, highlighting ongoing challenges in how these resources are allocated and consumed.

Several factors influence this dynamic:

- AI workloads such as inference, Retrieval-Augmented Generation (RAG), and fine-tuning are inherently variable, resulting in fluctuating GPU demand over time.
- Many enterprise environments currently lack the ability to finely partition, schedule, and share GPU resources across users, teams, and projects, which can lead to periods of both contention and inactivity.
- GPU infrastructure is frequently associated with static clusters or manually provisioned environments, creating fragmentation and limiting operational flexibility.

When GPU resources are not consistently utilized, organizations may miss opportunities to maximize the value of their infrastructure investments. By adopting advanced orchestration platforms and policy-driven consumption models, organizations can enhance GPU utilization, streamline operations, and extend the benefits of high-performance compute to more AI teams—helping to transform infrastructure into a foundation for innovation.

## 1.3 Streamlining GPU infrastructure provisioning

While Cisco AI PODs deliver a robust, production-ready foundation, organizations now have an opportunity to evolve how that infrastructure is consumed. Today, platform teams often rely on manual processes: configuring clusters, applying RBAC, and provisioning GPU quotas using YAML files or ad-hoc CI/CD pipelines. These workflows work, but they can slow down delivery and introduce inconsistency across environments.

By introducing orchestration and policy-driven automation, enterprises can transform this foundational hardware into a self-service platform enabling faster provisioning, consistent environments, and seamless access to GPU resources across teams. Instead of acting as a gate, infrastructure becomes an accelerator for AI innovation.

## 1.4 AI clouds today require piecing together multiple vendors

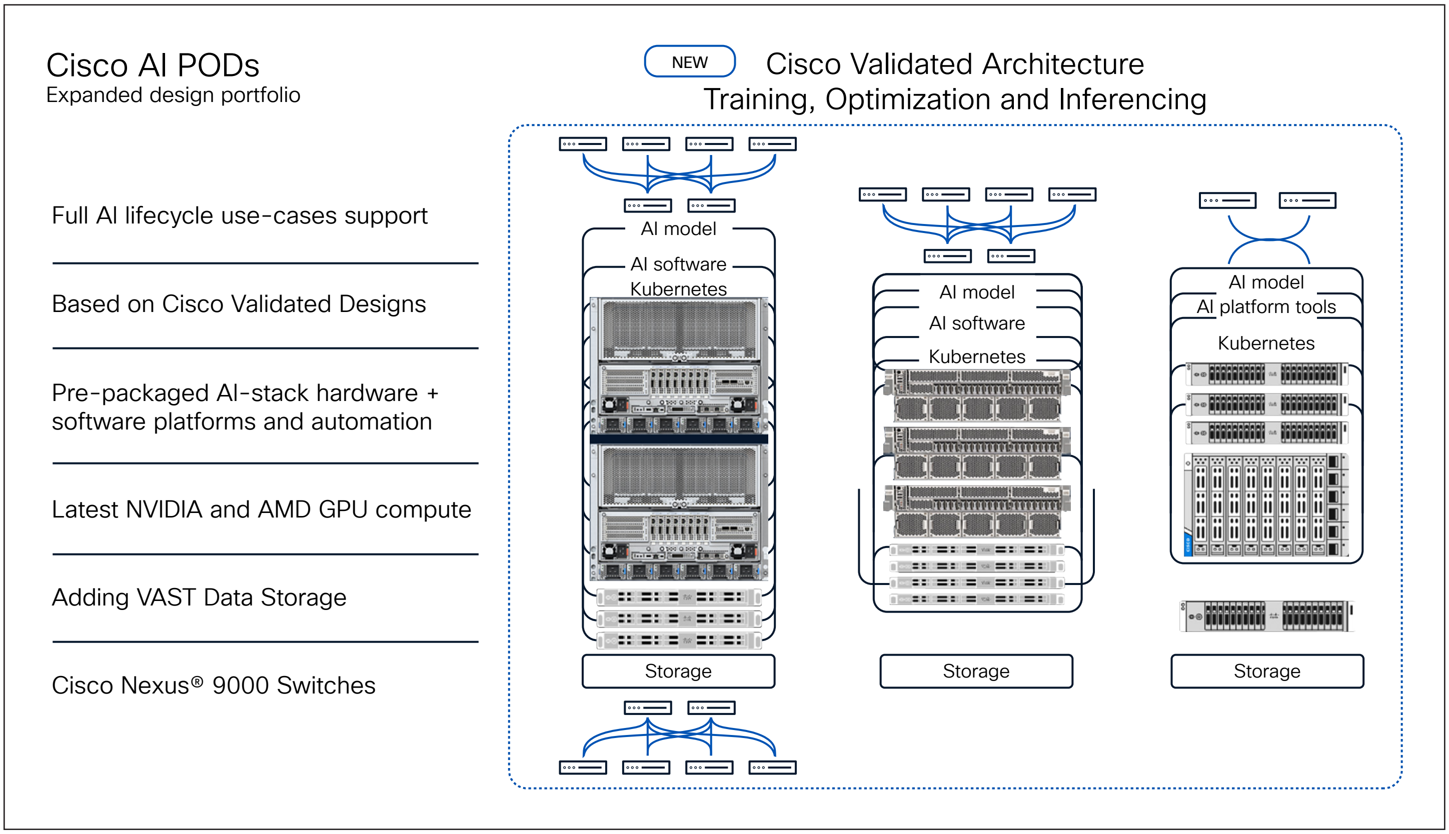
Unlike mature cloud environments, most enterprise AI infrastructure lacks a standardized interface for consumption. There is no SKU catalog to select from. No resource-based t-shirt sizing. No built-in governance over who is consuming what.

As a result:

- GPU infrastructure remains siloed per team, with limited sharing across departments.
- Platform teams cannot track utilization, enforce quotas, or implement cost controls.
- Developers and researchers are forced to navigate internal scripts, Slack threads, and service tickets just to run a job.

The absence of a well-defined consumption layer creates friction for developers and opacity for infrastructure owners, slowing time-to-value and undermining ROI.

## 2.0 Solution overview



Cisco AI PODs

Expanded design portfolio

- Full AI lifecycle use-cases support
- Based on Cisco Validated Designs
- Pre-packaged AI-stack hardware + software platforms and automation
- Latest NVIDIA and AMD GPU compute
- Adding VAST Data Storage
- Cisco Nexus® 9000 Switches

NEW Cisco Validated Architecture

Training, Optimization and Inferencing

Figure 1. Cisco AI PODs + Rafay GPU Platform as a Service = Sovereign and Enterprise GPU Cloud

Enterprises must evolve beyond the idea of AI infrastructure as a static, physical deployment. To truly unlock their investment in Cisco AI PODs, they need to operate GPU infrastructure as a **service**, one that supports:

- **Shared GPUs –** on-demand, self-service access to GPU environments
- **Multitenant isolation** for secure, compliant infrastructure sharing
- **Standardized, version-controlled** SKUs that map to workload needs
- **Automated policy enforcement**, quota management, and telemetry

To address this, Cisco and Rafay have collaborated to deliver a modular, fully validated GPU cloud architecture that combines Cisco's AI POD infrastructure with Rafay's GPU Platform as a Service (PaaS).

Together, the platform transforms traditional GPU infrastructure into a secure, self-service, multitenant cloud platform, extending the inherent simplicity, flexibility, and modularity of Cisco AI PODs from infrastructure deployment to a fully operational, consumption-driven AI cloud.

- **Cisco AI PODs** provide the compute, fabric, storage, and pre-validated design.
- **Rafay GPU PaaS** delivers orchestration, policy enforcement, and developer access abstraction.

This white paper outlines how the combined solution enables enterprises to rapidly launch and operate GPU clouds with:

- Full-stack orchestration and workload lifecycle management
- Declarative SKU provisioning with real-time quota enforcement
- Namespace isolation and cost attribution for internal and external users

## 2.1 Who is this for

This paper is designed for **IT operations**, **platform engineering**, and **cloud infrastructure** leaders responsible for deploying and scaling AI-ready infrastructure within private or sovereign and enterprise environments.

It provides a blueprint for how to move from bare-metal GPUs to a **production-grade AI consumption platform**, validated by joint field deployment results.

## 3.0 Cisco AI PODs: pre-validated infrastructure for AI at scale

To power production-grade AI, enterprises need infrastructure that delivers both raw performance and operational readiness. Cisco AI PODs provide this foundation by offering a pre-validated, full-stack architecture engineered to support diverse AI workloads including RAG pipelines, optimization, and inferencing. Cisco AI PODs deliver a pre-validated, full-stack architecture, embodying simplicity, flexibility, and modularity to provide the foundational infrastructure for production-grade AI. Engineered to support diverse AI workloads including RAG pipelines, optimization, and inferencing, AI PODs simplify procurement, deployment, and scaling, thereby ensuring both raw performance and operational readiness.

Built on Cisco's AI-ready data-center strategy and integrated with NVIDIA-optimized compute, AI PODs simplify the procurement, deployment, and scaling of modern AI infrastructure.

### 3.1 Powered by next-generation Cisco UCS platforms

Cisco AI PODs are available in several form factors, each designed for specific performance and density needs:

- **Cisco UCS C225 M8 Rack Server:** Single-socket 4th Gen. and 5th Gen. AMD EPYC CPUs and up to 3 TB of memory capacity. Read Spec ↗
- **Cisco UCS C220 M8 Rack Server:** Two-socket 4th Gen. and 5th Gen. Intel Xeon Scalable Processors with up to 60 cores per processor. Up to 4TB with 32 x 128GB DDR5-5600 DIMMs with 5th Gen. Intel Xeon Scalable Processors. Read Spec ↗
- **Cisco UCS® C885A M8 Server**: optimized for high-density, GPU-intensive workloads. Supports up to 4 double-wide GPUs (for example, NVIDIA H100), dual 5th Gen Intel® Xeon® processors, and up to 8 TB of DDR5 memory. Ideal for RAG, and foundation model fine-tuning. Read Spec ↗
- **Cisco UCS C845A M8 Rack Server**: a versatile 2U rack server designed for accelerated inferencing and smaller-scale AI workloads. Offers balanced performance for edge AI, analytics, and real-time inference. Read Spec ↗
- **Cisco UCS X-Series Modular System**: designed for composability and scalability, the Cisco UCS X210c M7 Compute Node supports NVIDIA GPUs through PCIe expansion. Ideal for GPU-cloud deployments requiring flexible, software-defined resource pools. Explore UCS X ↗

Each AI POD configuration integrates with Cisco Nexus switches, high-performance storage (for example, VAST, Pure Storage, Netapp), and Kubernetes platforms such as Red Hat OpenShift, providing full-stack readiness.

### 3.2 Cisco Validated Designs (CVDs): built for the entire AI lifecycle

AI PODs are delivered with full **Cisco Validated Designs**, ensuring that the end-to-end stack is tested and production-ready. CVDs are continuously updated to reflect the latest NVIDIA GPUs (HGX, MGX, L40S, H100, H200), storage integrations, and orchestration choices.

This means faster time-to-value, less integration-complexity, and a repeatable pattern for deploying AI across environments.

## 3.3 AI-ready, secure by design

Cisco AI PODs are built with zero-trust principles and enterprise-grade operational visibility:

- Workload isolation and segmentation at the network and cluster level
- RBAC and governance integration with Cisco Intersight® and OpenShift
- Fabric-based scalability to hundreds of nodes per cluster
- Out-of-band management and compliance observability through Intersight

This secure-by-default architecture ensures that the infrastructure is ready for enterprise AI and sovereign deployments alike.

| Cisco® system | GPU support | Ideal workloads |
|---|---|---|
| **Cisco UCS C885A M8 Server** | 8x H200 | Optimization (RAG, fine-tuning), large-scale Inference |
| **Cisco UCS C845A M8 Rack Server** | 2 to 8xRTX PRO 6000/L40S/L4 | Inference, optimization (RAG) |
| **Cisco UCS X210c M7 Compute Node** | Modular (PCIe) | Multitenant orchestration |
| **Cisco UCS C220 M8 Rack Server** | 3xL4 | AI inferencing, data analytics, and graphics |
| **Cisco UCS C225 M8 Rack Server** | 3xL4 | AI inferencing, data analytics, and graphics |

Cisco AI PODs provide compute, storage, and fabric in an infrastructure that is needed to run AI workloads at scale. But to truly operationalize AI, organizations need to expose this infrastructure through self-service, multi-tenant consumption models.

In the next section, we explore how the Rafay GPU Platform as a Service overlays Cisco AI PODs to deliver a fully operational **GPU cloud platform**, one that enables SKU-based provisioning, GPU slicing, quota enforcement, and AI workload catalogs “out of the box.”

# 4.0 Operationalizing Cisco AI PODs with Rafay GPU Platform as a Service

## 4.1 From infrastructure to consumption: a platform-led model

Enterprises need a platform operating model that exposes infrastructure as a governed, reusable service. Cisco provides the AI POD foundation. Rafay transforms this foundation into a production-grade GPU cloud, completing the vision of a truly simple, flexible, and modular AI infrastructure by extending these benefits from hardware deployment to seamless, self-service consumption.

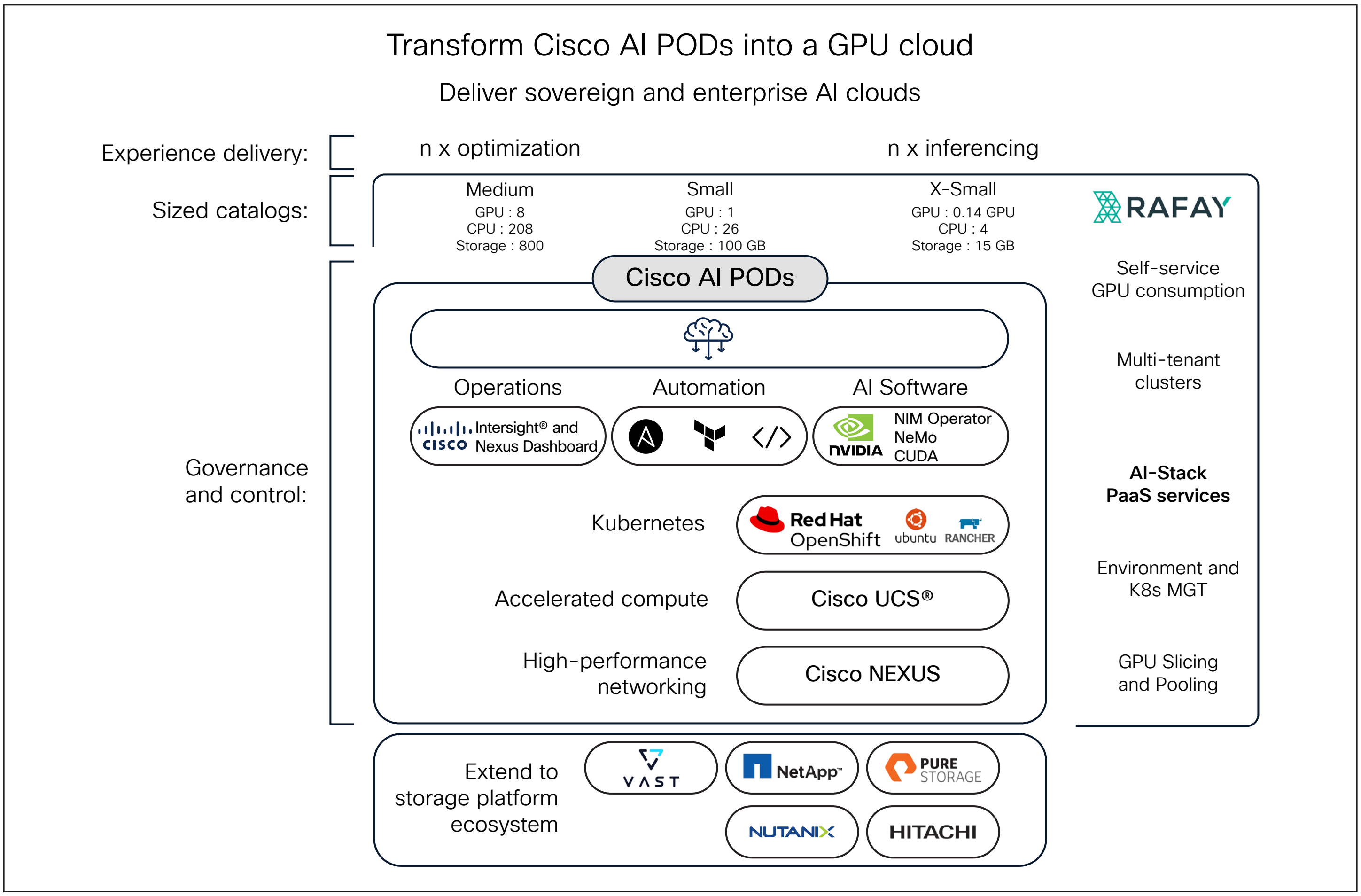


Figure 2. Cisco delivers compute, fabric, Kubernetes platforms and lifecycle control. Rafay overlays a GPU PaaS experience that includes SKU management, self-service portals, multi-tenancy controls, environment + Kubernetes management and GPU slicing and pooling

## 4.2 Developer self-service portal

Developers access the GPU platform through a web console (GUI), API, or CLI by leveraging a self-service portal that eliminates the need for support tickets or manual intervention. They can launch instances from a curated catalog of pre-approved SKUs, streamlining workflows and simplifying setup.

Rafay simplifies provisioning through a declarative, repeatable launch process. For example:

- A data scientist selects the H100-Inference-vLLM SKU.
- This SKU automatically provisions a MIG slice, deploys a secure vLLM container, and applies a 48-hour TTL.
- RBAC ensures exclusive access to the instance for the requesting user.

The result: faster onboarding, minimal operational overhead, and consistent application of security best practices.

## 4.3 Secure multi-tenancy

Rafay offers native, secure multi-tenancy through a layered architectural model that provides strict isolation, governance, and operational efficiency across teams and business units.

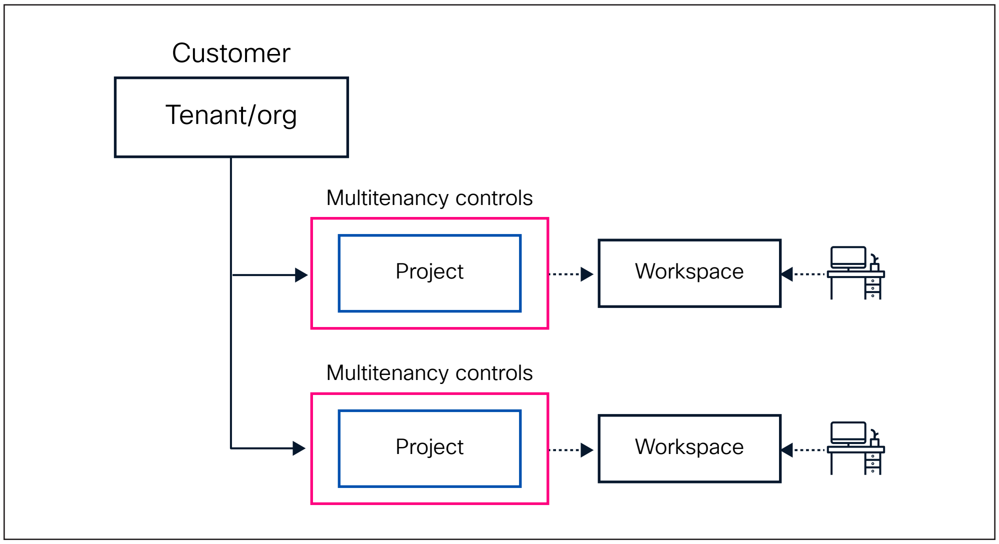


Figure 3. Role-based access control with organizational structure

Rafay employs a hierarchical Role-Based Access Control (RBAC) model where users are assigned one or more roles that define their permissions and access scope.

Each customer (tenant) can define multiple projects. Projects serve as isolated units, often aligned with teams or business units within which users such as developers, ML engineers, or data scientists can create and manage workspaces. These users can also delegate access by assigning workspace-level or collaborator roles, enabling fine-grained control over who can access and manage specific resources. User lifecycles can be managed externally by the customer's IDP (identity provider).

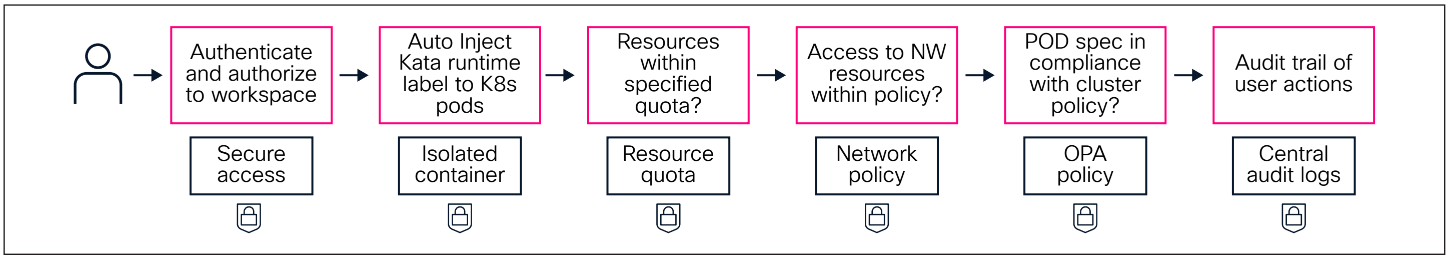

Authenticate and authorize to workspace
Auto Inject Kata runtime label to K8s pods
Resources within specified quota?
Access to NW resources within policy?
POD spec in compliance with cluster policy?
Audit trail of user actions
Secure access
Isolated container
Resource quota
Network policy
OPA policy
Central audit logs

Figure 4. Built-in security controls

Rafay automatically enforces a comprehensive set of security controls to mitigate risks such as lateral movement and privilege escalation. These controls ensure that teams can safely share infrastructure without interference, while platform teams maintain oversight and control. Key protections include network segmentation, resource quotas, secure remote access, support for isolated Kata containers, and customizable security policies, all designed to enable secure, multitenant operations by default.

## 4.4 Rafay Kubernetes Operations Platform

The Rafay Kubernetes Operations Platform provides a production-grade control layer for managing multi-cluster environments at scale. It streamlines day-2 operations, enforces policy-driven governance, and ensures environment consistency across any Kubernetes distribution.

Key capabilities include:

- **Blueprinting**: standardizes deployment of critical add-ons (for example, GPU operators and security tools) across clusters with automated drift-detection to block unauthorized changes
- **Fleet operations:** enables coordinated updates at scale, including add-on rollouts, Kubernetes upgrades, and OS patches, reducing operational burden and aiding compliance
- **Centralized visibility:** real-time dashboards and audit-ready reports that offer clear operational insights and simplify compliance verification

## 4.5 Comprehensive GPU slicing, monitoring, and cost attribution at scale

Rafay delivers an end-to-end solution for managing GPU infrastructure with slicing, visibility, metering, and cost attribution, all tailored for multitenant environments.

Platform teams can centrally deploy and standardize GPU operator configurations, including GPU slicing strategies such as MIG, using Rafay’s declarative blueprints and add-ons. These configurations are applied consistently across clusters, ensuring alignment with organizational policies.

Rafay’s multitenant GPU dashboards offer detailed visibility into:

- GPU inventory and allocation (for example, servers and MIG slices)
- SKU-usage patterns and frequency
- Instance-level activity and user attribution
- Health status, uptime, and trends over time
- Most-active users and failed-instance reports

Real-time metrics are collected at both node and instance levels, stored in a time-series database, and made accessible to platform operators and end users.

The billing metrics API enhances financial accountability by:

- Aggregating usage by SKU and instance
- Calculating billable compute and storage
- Generating detailed, auditable usage reports
- Enabling chargebacks through finance system integrations

For advanced insights, Cisco Intersight telemetry can be integrated to correlate GPU health and power metrics with Rafay's usage data, enabling teams to monitor resource efficiency, detect under utilization, and dynamically price services.

## 4.5 SKU as a service: productizing GPU infrastructure into on-demand AI environments

**At the core of Rafay's platform is the** SKU Studio, a robust, purpose-built catalog system that empowers platform teams to create, customize, and deliver AI-ready infrastructure and applications as reusable SKUs for end users**.**

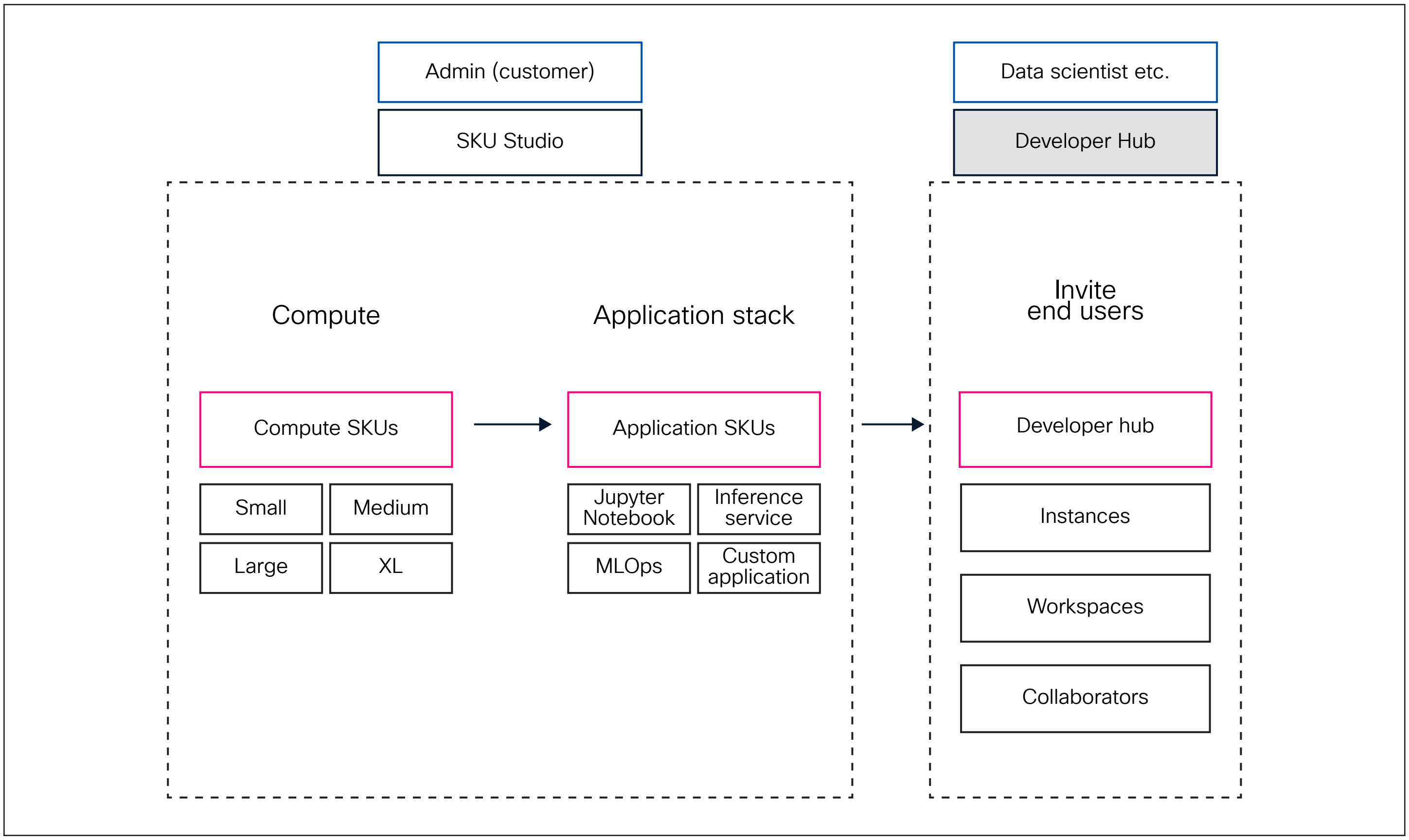


Figure 5. Rafay SKU Studio and developer hub

Each **SKU** is a declarative, modular abstraction of GPU infrastructure that bundles together:

- **Compute configuration**: GPU/MIG profiles, CPU, memory, and storage
- **Application stack**: pre-integrated tools such as vLLM, Triton, Jupyter Notebook, and more
- **Policy controls**: TTLs, RBAC, multitenancy isolation, quotas, and scheduling rules
- **Billing metadata**: usage units, cost attribution, and visibility into billable components

Table 1. **GPU instances tiers**

| SKU tier | Workload | GPU | CPU | Storage |
|---|---|---|---|---|
| **Medium** | Optimization/RAG | 8 | 208 | 800 GB |
| **Small** | Inference API | 1 | 26 | 100 GB |
| **X-Small** | Fractional model serving | 0.14 | 4 | 15 GB |

Users with the appropriate RBAC permissions can use SKU Studio to design and configure a SKU by:

- Defining the provisioning logic with embedded orchestration workflows
- Customizing the end-user experience with icons, documentation, and dynamic input forms
- Locking or exposing specific input fields to balance control and flexibility
- Displaying transparent billing estimates at launch time
- Configuring output visibility and runtime metadata
- Enforcing TTLs or schedules to avoid sprawl and control resource consumption
- Adapting SKU definitions to fit different team needs, customer environments, or SLAs, thereby ensuring a consistent, secure, and cost-effective self-service experience across the organization

## 4.6 Why this layer matters (DIY vs. Rafay)

**Table 2. Comparison table**

| Without Rafay | With Rafay GPU PaaS |
|---|---|
| Static, hard-coded clusters | Dynamic SKU provisioning per team |
| Manual RBAC, YAML, scripts | Self-service + policy-driven SKUs |
| Idle GPU fragmentation | Sliced GPU pooling w/ MIG + quotas |
| No real metering | Token-based usage tracking + billing |
| Inconsistent environments | Versioned, reproducible SKUs |

Together, Cisco and Rafay offer the infrastructure and orchestration stack to turn AI infrastructure into an internal GPU cloud: secure, scalable, and operational within weeks.

## 5.0 Who benefits from a unified GPU cloud platform

This jointly validated solution is purpose-built for a diverse range of customers unified by a common challenge: the need to operationalize GPU infrastructure for modern AI workloads, with security, speed, and scale.

Tables 3 through 8 provide a breakdown of Ideal Customer Profiles (ICP) that benefit most from this solution from Cisco and Rafay.

Table 3. General enterprises adopting private AI infrastructure

| Feature | Value to enterprise AI teams |
|---|---|
| **Challenge** | Siloed on-premises compute, high platform-engineering overhead |
| **Need** | Internal GPU cloud with self-service access, quota enforcement, and usage governance |
| **Rafay+Cisco fit** | Enables federated self-service for AI workloads with centralized visibility |
| **Infra admin benefit** | Centralized quota and RBAC enforcement, shared infrastructure visibility, and improved GPU utilization |
| **Outcome** | Reduced infra duplication, enhanced developer agility, and governance baked-in |

Table 4. Government agencies and public sector institutions

| Feature | Value to regulated AI environments |
|---|---|
| **Challenge** | Need for air-gapped, compliant environments with strict access control |
| **Need** | Secure multitenancy, audit logging, and offline deployment capability |
| **Rafay+Cisco fit** | Validated stack for sovereign infrastructure with RBAC, policy enforcement, and out-of-band observability |
| **Infra admin benefit** | Policy enforcement for compliance, air-gapped observability, and centralized audit + access controls. |
| **Outcome** | Rapid compliance onboarding, usable by both IT and research divisions without drift or rework |

Table 5. CSPs offering Rafay GPU Platform as a Service and AI services

| Feature | Value to GPU monetization strategies |
|---|---|
| **Challenge** | Manual tenant onboarding and limited cost metering capabilities |
| **Need** | Multitenant GPU-as-a-service platform with branded developer portals and chargeback enforcement |
| **Rafay+Cisco fit** | White-labeled, multitenant PaaS with chargeback reports, isolated environments, and quota control |
| **Infra admin benefit** | Tenant lifecycle automation, built-in chargeback metering, and simplified operational management. |
| **Outcome** | Faster time-to-market for service monetization with operational simplicity and partner branding support |

Table 6. Existing Cisco UCS customers (brownfield)

| Feature | Value to current Cisco infrastructure operators |
|---|---|
| **Challenge** | Extending existing infrastructure to support AI without replatforming |
| **Need** | Seamless GPU orchestration atop existing Cisco UCS deployments |
| **Rafay+Cisco fit** | Turnkey overlay with agent-based GPU orchestration on existing clusters |
| **Infra admin benefit** | Deploy GPU orchestration on existing Cisco UCS clusters with no infra re-architecture or downtime. |
| **Outcome** | Extends ROI of Cisco UCS deployments, introduces Rafay GPU Platform as a Service without infrastructure rework |

Table 7. Greenfield builders of new AI platforms

| Feature | Value to net-new GPU infrastructure projects |
|---|---|
| **Challenge** | Starting from scratch with a high-performing, production-ready AI stack |
| **Need** | Fully integrated solution that spans from racking to model deployment |
| **Rafay+Cisco fit** | Pre-validated, AI-ready PODs + PaaS orchestration reduce time from procurement to productivity |
| **Infra admin benefit** | Out-of-the-box orchestration stack with minimal platform overhead; GitOps-native workflows included. |
| **Outcome** | AI platforms operationalized in weeks instead of months, with support for GenAI, RAG, and inference use cases |

Table 8. Telecommunications operators

| Feature | Value to network AI and SLA-sensitive use cases |
|---|---|
| **Challenge** | Operational complexity and regulatory sensitivity in network AI use cases |
| **Need** | On-premises, sovereign AI cloud with secure slicing for inference and optimization workloads |
| **Rafay+Cisco fit** | Centralized GPU resource pool with tenant-specific provisioning, monitoring, and control |
| **Infra admin benefit** | SLA-compliant quota policies, secure resource segmentation, and internal usage reporting per tenant. |
| **Outcome** | Enables telcos to offer internal AI-as-a-service, with performance SLAs and compliance baked in |

## 6.0 Conclusion: Operationalizing AI with confidence

By pairing Cisco's validated AI infrastructure with Rafay's GPU PaaS control plane, organizations can transform GPU systems into fully governed internal platforms. What emerges is a modular, consumption-driven architecture where GPU resources are accessible through standardized SKUs, workloads are isolated and policy-bound, and observability is embedded into every layer, from cluster telemetry to per-tenant usage dashboards. Developers gain self-service access without compromising platform security or consistency. Operators enforce quotas, track consumption, and ensure that AI infrastructure is being fully utilized.

As organizations continue to scale AI programs across private, hybrid, and sovereign environments, this architecture offers a clear path forward: deliver GPU infrastructure as a service, enable secure and compliant multitenancy, and make consumption predictable, observable, and cost-aligned from day 1.

## Learn more

To learn more or explore a tailored deployment path, contact your Cisco representative or visit rafay.com to see the platform in action.

Discover Cisco AI PODs at https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ai-pods-aag.html.

### Legal notice

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. A listing of Cisco's trademarks can be found at www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company.

Specifications and product availability subject to change without notice. Cisco reserves the right to make changes to this document and the products described herein at any time, without notice. The information in this document is provided "as is" without warranty of any kind, either expressed or implied.