



Sovereign GPU Cloud as a Service: Build vs Buy Guide

Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-10-15 · sovereign gpu cloud as a service · gpu cloud build vs buy · red hat · rafay · openshift · sovereign ai infrastructure · gpuaas · gpu cloud economics · gpu cloud architecture

Original article: <https://gpusmith.com/articles/en/sovereign-gpu-cloud-build-vs-buy>



In this article

1. Executive Summary
 2. Introduction and Background
 3. Red Hat and Rafay: Adopt the Integrated Architecture
 4. Internal Assembly: Build the Service Factory
 5. Defer Commercialization: Preserve the Option
 6. Feature Comparison
 7. Tenant-Order Walkthrough (Hypothetical Example)
 8. Performance and Benchmarks
 9. Data Analysis and Evidence
 10. Implications and Future Directions
 11. Conclusion
-

Executive Summary

Sovereign GPU cloud as a service is a service-operations decision as much as an infrastructure decision. Red Hat dated its joint Rafay reference-architecture launch **September 9, 2026**, describing a tenant journey from service selection to a metered graphics processing unit (GPU) environment. (www.redhat.com) (www.redhat.com) The practical question for telecommunications operators and regional cloud builders is whether that blueprint supplies enough of the service factory to justify integration and subscription costs. The Architecture Center explicitly describes its architectures as generalizations of successful deployments; it does not make every implementation identical. (www.redhat.com)

Adopting Red Hat and Rafay offers a documented division between platform fulfillment and commercial orchestration. Red Hat supplies the platform-side functions; Rafay coordinates approvals, quota checks, reservations, and metering. (www.redhat.com) Procurement still needs an itemized entitlement map: **Red Hat AI Enterprise** and layered **OpenShift AI** have different subscription structures. (www.redhat.com) Rafay's service-provider frequently asked questions quote **Enterprise Support at an additional 20%** of the cluster or node subscription, a scoped published surcharge rather than a complete joint-stack price. (rafay.co) Internal assembly should therefore be compared with a written commercial proposal, using equivalent support coverage.

Building internally can preserve architectural choice, provided the operator accepts integration ownership. A dedicated tenant control plane does not establish worker-node isolation, and NVIDIA time-slicing does not provide memory or fault isolation between replicas. (kubernetes.io) (docs.nvidia.com) Likewise, monitoring is insufficient evidence of invoice accuracy: Prometheus explicitly cautions against accurate per-request billing use. (prometheus.io) Sovereignty claims also require bounded evidence. The National Institute of Standards and Technology (NIST) validates cryptographic modules under Federal Information Processing Standards (FIPS); module validation does not automatically certify a complete cloud service. (csrc.nist.gov)

The recommended decision is conditional. Adopt the integrated architecture when a pilot proves the complete tenant journey, deployment boundary, support model, and exit path. Build when existing engineering capabilities can operate those same controls at an acceptable fully loaded cost. Defer broad commercialization

when tenant isolation, disconnected continuity, invoice reconciliation, or demand evidence remains unresolved. The FinOps Foundation's guidance supports separating resource-efficiency metrics from business-unit economics and incorporating fully loaded costs. (www.finops.org) (www.finops.org) This report supplies an order walkthrough, ownership and capacity comparisons, symbolic cost sensitivities, and a weighted procurement rubric. It reports no measured operator margin, utilization improvement, or end-to-end performance advantage for either implementation.

Introduction and Background

An operator can own a capable GPU fleet while still lacking a product that procurement teams can buy repeatedly. The missing work includes defining an entitlement, allocating constrained capacity, proving isolation, measuring consumption, resolving billing differences, and removing access at contract end. Here, **GPU as a service (GPUaaS)** means a commercially governed capacity or inference offering. This report calls its connected operating process the **service factory**. Its output is an enforceable customer promise, backed by operational records and a reconciled commercial ledger. Tenant boundaries must cover the data plane as well as tenant administration. (kubernetes.io)

For this report, the operating unit is an order from a stock-keeping unit (**SKU**) to a metered environment. That framing is useful for operators already planning distributed infrastructure: evaluate the completeness of the commercial and operational chain, rather than treating cluster installation as the final milestone. A sovereign GPU cloud architecture must also explain who can administer each layer and which dependencies remain available inside the chosen boundary.

GPU Smith is an adjacent engineering advisor in this comparison. Its published scope includes independent technical due diligence for AI datacenter and GPU-cloud investments. (gpusmith.com) That perspective favors written requirements, reproducible tests, and cost models reconciled with actual facilities. The consultancy is not treated as a competing cloud platform or an alternative subscription product.

The report date is **October 13, 2026**; live source verification occurred **October 15, 2026**. Dated releases establish their stated publication dates. Undated documentation describes what was accessible during verification, without proving that every current detail was available on the report date. Product documentation, proposed acceptance tests, and this report's derived equations are distinguished throughout.

Regulatory context is similarly bounded. The European Commission identifies cloud-computing and data-center providers among the entities addressed by the implementing regulation associated with the **Directive on measures for a high common level of cybersecurity across the Union (NIS2 Directive)**. (digital-strategy.ec.europa.eu) The **European Union Agency for Cybersecurity (ENISA)** supplies implementation advice and examples of evidence. (www.enisa.europa.eu) Those sources support collecting an applicability and control-evidence dossier; they do not establish certification of the architecture considered here. Jurisdiction-specific applicability and contractual obligations require qualified legal assessment.

Red Hat and Rafay: Adopt the Integrated Architecture

Capabilities

The published architecture requires **OpenShift and CoreOS** to be operational before Rafay capabilities function. (www.redhat.com) Within that dependency, the reference assigns provisioning, workload scheduling, and security enforcement to Red Hat layers. (www.redhat.com) **Red Hat Advanced Cluster Management** drives fleet lifecycle; Rafay's **SKU Studio** defines offers, while commercial workflows coordinate approvals, quota checks, reservations, and metering. (www.redhat.com) These are complementary ownership boundaries that the operator must turn into an operational responsibility matrix.

Rafay's usage documentation distinguishes a fixed hourly price for an entire SKU from dimension-based pricing that aggregates resource costs. (docs.rafay.co) (docs.rafay.co) Its billing announcement describes on-demand and **Monthly Recurring Charge (MRC)** models, alongside reservations. (docs.rafay.co) Documented reservation protection prevents other demand from encroaching on reserved capacity. (docs.rafay.co) This makes reservation admission a commercial capacity decision, rather than simply another scheduler setting.

The documented reservation charging trigger is especially consequential: billing begins for accepted quantities even when instances are not running. (docs.rafay.co) The operator should reproduce that event in a pilot and align it with the customer's contract. Acceptance, allocation, application activity, and invoice recognition must remain separate timestamps.

Adoption

The architecture's division of functions is a starting hypothesis for implementation. Adoption should begin with a versioned bill of materials, customer-order schema, site topology, entitlement schedule, and escalation map. The procurement package should identify every component assumed to be included, separately licensed, supplied by an infrastructure partner, or operated by the customer.

Red Hat's subscription guide says layered OpenShift AI requires a base OpenShift subscription. (www.redhat.com) The same guide restricts included OpenShift entitlements in AI Enterprise to AI use cases and discusses separately purchased management add-ons. (www.redhat.com) (www.redhat.com) Consequently, quoting an AI product name alone does not establish coverage for the whole commercial service. Obtain written rights for the intended service-provider and tenant uses.

Self-hosted sizing also needs release-specific confirmation. Rafay's air-gapped prerequisites specify **at least three high-availability controller nodes** and a **2 TB data disk**, while a separate self-hosted matrix lists **1 TB**. (docs.rafay.co) (docs.rafay.co) (docs.rafay.co) These pages describe differing published requirements; a signed sizing schedule should resolve the applicable configuration before procurement.

Strengths and Limitations

The integrated approach supplies explicit commercial and platform interfaces to evaluate. It can reduce the number of interfaces an operator designs from scratch, but that is an implementation opportunity, not a measured delivery-time saving. A pilot must establish which workflows are packaged, configured, customized, or still owned by the operator.

Rafay's release notes document **Token Factory on OpenShift**, supporting an inference-service path in addition to infrastructure rentals. (docs.rafay.co) That capability should be purchased and tested against the intended model-serving service, rather than assumed to follow automatically from a GPU-instance subscription.

The reviewed sources do not establish a complete joint-stack price, an independently measured operator margin, or comparative implementation effort. Their absence makes a reproducible pilot and itemized quote necessary; it does not determine which option will be cheaper.

Internal Assembly: Build the Service Factory

Capabilities

An internal assembly must reproduce the business behavior of the integrated option. It needs a catalog, identity federation, entitlement checks, reservation authority, provisioning workflow, usage-event pipeline, invoice export, and teardown orchestration. These functions can have independent implementations, but their handoffs must share durable tenant, order, resource, and rate-version identifiers.

OpenID Connect defines an identity layer over OAuth, providing a standards-based authentication interface. (openid.net) **Keycloak** documents a separate backchannel URL for internal communications. (www.keycloak.org) These offer implementation choices for identity networking; neither establishes that the joint architecture uses that arrangement or that an external identity dependency can survive disconnection.

Open Policy Agent (OPA) supports signed policy bundles and optional persistence of activated bundles for recovery. (www.openpolicyagent.org) This is a possible internal policy-distribution mechanism. Local recovery still needs a test of stale policy, lost connectivity, signing-key rotation, and behavior when the last valid policy is unavailable.

Adoption

Engineering teams should define commercial event contracts before selecting interchangeable components. **OpenMeter** documents event-based metering, including compute-runtime use cases. (openmeter.io) Its documentation makes key meter-definition fields immutable because changes would redefine already collected usage. (openmeter.io) An internal service therefore needs meter-version migration and historical reproducibility, alongside live aggregation.

Identity also extends beyond tenant administrators. **SPIFFE** identifies workloads, providing a basis for service-to-service identity. (spiffe.io) The operator must still decide which trust domains, issuing authorities, and revocation behavior belong inside each sovereign boundary. Identity issuance and workload isolation remain distinct engineering responsibilities.

Network encryption requires equally precise scoping. **Cilium** documents that its default WireGuard encryption covers traffic between Cilium-managed pods. (docs.cilium.io) An internal architecture should map management, storage, host, ingress, and accelerator-network paths separately, then verify the controls selected for each path.

Strengths and Limitations

Internal assembly gives the operator direct responsibility for interface and data-model design. It also creates a continuing integration obligation: authentication changes must preserve entitlements, provisioning retries must preserve reservations, and meter evolution must preserve invoice reproducibility. Source availability alone does not determine the staffing or operational cost of those obligations.

FOCUS, the FinOps Open Cost and Usage Specification, requires pricing-unit semantics to agree with published pricing and relevant invoice units. (focus.finops.org) That is a useful contract for an internal ledger, particularly when reserved capacity, runtime, storage, and tokens appear together.

The build option is strongest when the operator already owns these business capabilities and can demonstrate their operation. A catalog of available projects is insufficient adoption evidence. Require an existing service owner, maintained integration interfaces, on-call coverage, and a funded release process before giving internal assembly a favorable procurement score.

Defer Commercialization: Preserve the Option

Capabilities

Deferral means continuing infrastructure validation or bounded tenant pilots while withholding a broad commercial promise. It can preserve time to establish demand, operational controls, and contract terms. It should have explicit exit criteria, a spending limit, and a review date, rather than becoming an indefinite holding pattern.

A useful pilot can permit tenant onboarding, allocation, serving, and teardown while producing a shadow invoice. The operator can compare that invoice with the intended contract before accepting production billing obligations. This is a proposed operating method, not a claim about an implemented vendor workflow.

Adoption

Deferral is appropriate when an essential requirement cannot yet be demonstrated. For example, **Federal Risk and Authorization Management Program (FedRAMP)** guidance directs agencies to examine inherited controls, provider responsibilities, and secure-configuration guidance. (www.fedramp.gov) A collection of component descriptions is insufficient evidence for a customer's complete service assessment.

Similarly, NIST advises checking the validated cryptographic module's specific version or release. (csrc.nist.gov) Where a customer requires such evidence, the operator should resolve the actual deployed module and environment before promising compliance outcomes. The choice to defer follows from an incomplete evidence package, not from a general judgment about either platform.

A bounded pilot should include these sovereign-infrastructure acceptance checks:

- **Identity:** Test existing and new sessions during loss of external identity access.
- **Keys:** Document custody, rotation, recovery, and administrative access within the permitted jurisdiction. Verify encryption-at-rest configuration explicitly. (kubernetes.io)
- **Artifacts:** Verify local registries, signed updates, model acquisition, and rollback.

- **Control:** Prove order admission, provisioning, usage retention, and license continuity while disconnected.
- **Recovery:** Restore policy, controller state, identities and meters without unapproved dependencies.
- **Evidence:** Map each applicable obligation to a configuration, record, test, owner, and review date, using guidance such as ENISA's evidence examples. (www.enisa.europa.eu)

NIST module validation and ENISA's evidence-oriented guidance help define separate assurance questions; neither substitutes for the service-specific dossier. (csrc.nist.gov) (www.enisa.europa.eu) SPIFFE's documented short-lived, automatically rotated credentials also make identity continuity an explicit test concern. (spiffe.io)

Strengths and Limitations

Deferral can prevent an unproven service-level agreement (**SLA**) from becoming the operating plan. Its cost is real: infrastructure, engineering, financing, and opportunity costs continue during the pilot. The operator should track those costs explicitly and compare them with a narrower initial offering, such as one geography and one isolation class.

The U.S. Department of Energy cautions that **power usage effectiveness (PUE)** does not define the overall efficiency of a data center. (www.energy.gov) Facility efficiency alone therefore cannot close a commercialization gate. Demand evidence and useful work per unit of fully loaded cost remain necessary. A written decision date should determine whether to adopt, build, narrow the product, or continue only justified validation work.

Feature Comparison

Table 1 maps proposed operating accountability. **A** means accountable, **R** responsible for implementation, and **C** consulted. These assignments are a procurement starting point, not contractual vendor commitments.

Layer or decision	Red Hat and Rafay approach	Internal assembly	Deferral gate
Catalog and commercial admission	Operator A; Rafay R for commercial orchestration; Token Factory meters Red Hat AI's vLLM/llm-d serving through an application programming interface (API) and billing layer. (www.redhat.com)	Operator A/R for catalog, entitlement and admission services.	Approved product definition and reservation contract.
Cluster fulfillment and policy	Operator A; Red Hat R for the reference architecture's provisioning and enforcement layer. (www.redhat.com)	Operator A/R; distribution and component suppliers C.	Reproducible fulfillment and denied-access tests.
Tenant control-plane boundary	Select the documented container, virtual-machine, or physical isolation level. (docs.redhat.com)	Document control-plane and worker isolation separately. (kubernetes.io)	Accepted boundary for the specific tenant workload.

Layer or decision	Red Hat and Rafay approach	Internal assembly	Deferral gate
Meter and invoice handoff	Operator A; vendor chargeback export connects to the existing billing system. (rafay.co)	Operator A/R for event lineage, aggregation, reconciliation and invoice export.	Shadow invoice reconciles to orders and service events.
Recovery and service ownership	Operator A; assign platform, controller, hardware and billing responders in writing.	Operator A/R across the integrated service.	Restored service and usage ledger meet agreed objectives.

The table makes the operator's continuing accountability visible. Buying orchestration does not assign ownership of customer terms, tax treatment, facility capacity, or end-to-end recovery. Building internally changes the implementation owner while preserving those service obligations.

Define the SKU before allocating hardware

A usable rate card should bind the following worksheet fields into a versioned offer. The chargeable unit must be explicit before testing the journey.

- **Accelerator:** GPU model, count, partition profile, and exclusivity boundary.
- **Reservation:** Start and end, acceptance trigger, renewal, cancellation, and unused-capacity treatment.
- **Location:** Permitted primary, backup, support, and disaster-recovery geographies.
- **Network:** Connectivity class, bandwidth commitment, egress allowance, and encryption boundary.
- **Storage:** Capacity, performance class, snapshot retention, and key ownership.
- **Software:** Runtime, model-serving entitlement, permitted tenant use, and supported version set.
- **Service:** Support hours, availability objective, credit rules, metering unit, and invoice currency.

Each field is an operator input or contractual requirement. Add request-rate limits, spending thresholds, administrative audit records, and a tested suspension process to the operating policy. A supplier demonstration should show that changing one field produces the intended capacity, policy, and pricing consequences without silently changing the others.

Tenant-Order Walkthrough (Hypothetical Example)

This hypothetical example is a proposed acceptance scenario. An event-based metering layer is one documented implementation path, as OpenMeter illustrates. (openmeter.io)

- **Order and identity:** Create a tenant request carrying a stable order identifier and rate-card version. Authenticate its administrator and verify that an unauthorized principal cannot approve it.
- **Admission and reservation:** Apply the tenant quota and reserve the promised capacity before announcing acceptance. Test simultaneous orders against the same capacity, and require explicit terms before enabling overbooking. Verify the charging trigger against Rafay's documented accepted-reservation behavior where that model is selected. (docs.rafay.co)

- **Isolation:** Provision the chosen boundary. Kubernetes recommends default-deny communication for strict tenant isolation, and Secrets require configured encryption at rest rather than reliance on their default encoding. (kubernetes.io) (kubernetes.io)
- **Allocation and endpoint:** Bind the intended accelerator resource and model-serving endpoint to the order. Record the serving configuration, exposed interfaces, credential scope, and activation event.
- **Metering:** Emit durable, attributable commercial events. Prometheus' caution about accurate per-request billing supports keeping invoice evidence separate from sampled operational telemetry. (prometheus.io) Time-sliced deployments also need specific attribution tests. (docs.nvidia.com)
- **Invoice and correction:** Reconcile the accepted order, delivered entitlement, meter events, and invoice. Stripe documents asynchronous event processing and warns that late-event acceptance does not guarantee invoice inclusion. (docs.stripe.com) (docs.stripe.com)
- **Termination:** Revoke access, stop entitlements and charges according to contract, and dispose of retained data under an approved process. Kubernetes volume-reuse controls and NIST sanitization guidance address different parts of that process. (kubernetes.io) (csrc.nist.gov)

For an internal billing integration, require duplicate-event tests, correction tests, boundary timestamps, partial intervals, and rounding checks. Stripe documents idempotency guidance and a single-event cancellation mechanism. (docs.stripe.com) (docs.stripe.com) Those are examples of behaviors to specify, not proof of a supplied Rafay integration. Stripe's current documentation recommends Metronome for most new usage-based use cases; it should not be assumed that legacy billing examples prescribe the preferred new implementation. (docs.stripe.com)

Performance and Benchmarks

The reviewed evidence establishes no end-to-end performance winner. Test tenant isolation, network/storage tiers, serving configuration, billing, order fulfillment, and recovery alongside accelerator throughput.

Table 2 separates capacity products that can otherwise look interchangeable on a rate card. **Multi-Instance GPU (MIG)** and virtual machine (**VM**) describe different boundaries.

Capacity product	Documented boundary or behavior	Fragmentation exposure	Required operator test
Dedicated whole GPU or cluster	Define GPU, host and fabric exclusivity contractually; a dedicated API alone does not establish worker isolation. (kubernetes.io)	Unsold whole devices, topology groups, or sites.	Reallocate across tenant sizes while preserving exclusivity.
MIG-backed partition	NVIDIA documents hardware-level memory and fault isolation. (docs.nvidia.com)	Free profiles may not satisfy the next requested shape.	Measure profile mix, placement, reclamation and reconfiguration downtime.
Time-sliced resource	NVIDIA documents no memory or fault isolation between replicas and no proportional-compute guarantee from multiple requests. (docs.nvidia.com) (docs.nvidia.com)	Share count can diverge from deliverable performance.	Test noisy-neighbor behavior and the advertised entitlement.

Capacity product	Documented boundary or behavior	Fragmentation exposure	Required operator test
VM or dedicated worker product	Operator requirements to verify: workload and guest isolation, host allocation, and GPU device-assignment boundaries.	Stranded host resources and reserved topology.	Verify guest access, host allocation, device assignment and recovery.

Red Hat's VM-based isolation example concerns hosted control plane pods running inside VMs.

(docs.redhat.com) Worker and GPU device boundaries require separate operator verification.

Maintain separate inventory and evidence for each promised boundary. Partition shares and whole devices are different sellable units; partitioning benefits depend on demand and are not quantified here.

NVIDIA's MIG Manager requires user workloads to stop before reconfiguration, and reconfiguration can involve a node reboot. (docs.nvidia.com) (docs.nvidia.com) Capacity plans should therefore distinguish immediately allocatable profiles from capacity that becomes available only after a maintenance action.

Metering constraints also affect product eligibility. NVIDIA documents that **DCGM-Exporter**, its [Data Center GPU Manager metrics exporter](#), cannot associate metrics with containers when device-plugin time-slicing is enabled. (docs.nvidia.com) A service using that mode needs another proven attribution method before promising tenant-specific usage accounting.

Reproduce the service, not a headline throughput figure

MLPerf Inference specifies datasets and quality targets and uses a standard load generator for defined request patterns. (mlcommons.org) (mlcommons.org) These methodological controls are useful references for a local comparison, but they do not constitute measurements of this joint architecture.

- **Workload:** Fix model, precision, input/output distribution, context length, and quality target.
- **Topology:** Record GPU profile, host, network path, storage tier, and tenant boundary.
- **Latency:** Measure time to first token, subsequent-token latency, and tail behavior under the promised load.
- **Contention:** Repeat with competing tenants and background maintenance.
- **Lifecycle:** Drain, upgrade, reboot, restore and fail over while recording interrupted orders and endpoints.
- **Commercial accuracy:** Inject missing, duplicate, delayed and corrected events, then reconcile the invoice.
- **Reproducibility:** Preserve configuration, timestamps, event identifiers, and acceptance results.

Do not average independently calculated latency quantiles across instances: Prometheus documents that this produces statistically nonsensical values. (prometheus.io) Select a distribution-based aggregation method appropriate to the actual measurement system.

Lifecycle support also needs a joint test plan. Red Hat documents separate node-pool updates; Rafay warns of possible controller downtime during air-gapped upgrades. (docs.redhat.com) (docs.rafaq.co) Sequence platform, controller, firmware, driver, accelerator configuration and application changes through a jointly

accepted compatibility matrix. Recovery planning should include a restore exercise: Rafay's air-gapped backup guide uses a separate private-network server for object storage. (docs.rafaqay.co) That example is not evidence of independent-site disaster recovery.

Data Analysis and Evidence

Separate reported facts from operator economics

The public evidence supplies technical requirements and some commercial rules, rather than a complete business case. Examples include Rafay's scoped **20% Enterprise Support surcharge** and its reservation-acceptance charging behavior. (rafay.co) (docs.rafaqay.co) Neither establishes the operator's base subscription, realized selling rate, financing cost, billable demand, or contribution margin.

FinOps recommends fully loaded unit costs and distinguishes resource-efficiency measures from business measures. (www.finops.org) (www.finops.org) Accordingly, the model below uses operator-supplied variables. It contains no assumed hardware price, power tariff, staffing level, utilization result, or vendor discount. The equations are analytical definitions, not published financial results.

Let **H** be recognized billable GPU-hours, **r** the realized revenue per recognized hour, **F** period fixed cost, and **v** variable cost per recognized hour. Recognized means accepted under the operator's contract and finance policy, not merely emitted by a utilization counter. The derived modeled service surplus is $H \times r - F - H \times v$. A positive result is not automatically accounting profit.

For **GPU cloud total cost of ownership (TCO)**, use a cost schedule that includes:

- **Infrastructure:** Allocated acquisition or lease cost, residual-value assumptions, and the chosen ownership period.
- **Facility:** Space, power, cooling, interconnects, and committed capacity charges.
- **Platform:** Software subscriptions, support, management infrastructure, and billing integrations.
- **Operations:** Staffing, service management, security operations, and audit evidence.
- **Resilience:** Spares, repairs, recovery capacity, and planned maintenance.
- **Capital:** Financing and explicitly modeled idle-capacity allocation.

Run the worksheet separately for each SKU and geography before aggregating results. These are proposed model categories. Finance must reconcile their treatment to avoid charging acquisition and depreciation, or financing and capital costs, twice. FinOps explicitly assigns validation of metric assumptions and cross-source reconciliation to finance. (www.finops.org)

Calculate utilization and reservation sensitivity

Let **G** be installed GPUs, **T** period hours, **a** the fraction of physical capacity offered after maintenance and resilience holds, and **u** the fraction of offered capacity recognized commercially. Then $H = G \times T \times a \times u$, provided every counted hour uses the same whole-GPU-equivalent definition.

The derived fixed cost per recognized hour is $F \div H$. Holding the other inputs constant, halving u doubles that fixed-cost allocation. This arithmetic is a sensitivity, not an asserted utilization forecast. Break-even utilization is $F \div [G \times T \times a \times (r - v)]$ where the denominator is positive. A result above physically achievable capacity indicates that the entered assumptions cannot cover the modeled cost.

For reservation mix, let q be the share of offered capacity sold as paid reservations, b the on-demand billable fraction of the remainder, and rR and rO their respective realized rates. Derived revenue is $G \times T \times a \times [q \times rR + (\text{one minus } q) \times b \times rO]$. This separates reservation payment from workload activity. Pricing units should remain consistent with the rate card and invoice, following FOCUS semantics. (focus.finops.org)

Power should use metered infrastructure inputs. The Green Grid defines PUE as total data-center energy divided by equipment energy. (www.thegreengrid.org) If the electricity model starts with measured information-technology energy, multiply that energy by the applicable PUE and tariff once. (www.thegreengrid.org) Avoid adding cooling energy again if already included. PUE measures supporting infrastructure efficiency rather than useful computational output. (www.energy.gov)

Implications and Future Directions

Make sovereignty a tested operating boundary

An outbound-only connection still has a destination. The reference describes that connection to the Rafay control plane; hosted-control-plane guidance distinguishes self-hosted deployments. ([www.redhat.com/docs.rafay.co](https://www.redhat.com/docs/rafay.co)) Require a flow inventory covering commands, identities, telemetry, usage records, support bundles, artifacts, keys, and backups, with destination, owner, retention, and disconnection behavior for each.

Rafay documents air-gapped operation without Internet software downloads, while NVIDIA requires locally hosted operator images for its air-gapped installation path. (docs.rafay.co) (docs.nvidia.com) These are component mechanisms. They do not independently prove continuity of the complete commercial service.

Use weights without inventing scores

Table 3 is a weighted procurement rubric. The operator supplies w for each criterion and evidence-based option scores s ; the weighted result is **sum of $w \times s$ divided by sum of w** . No vendor scores or implementation estimates are invented.

Criterion	Weight input	Integrated option evidence	Internal option evidence
Implementation effort	w-effort	Accepted work breakdown, customization and staffing quote.	Funded backlog, interface ownership and delivery estimate.
Portability	w-exit	Export and replacement-controller exercise; exit terms.	Restore and migrate using documented internal interfaces.
Auditability	w-audit	Tenant-to-invoice lineage and control-evidence package.	Equivalent lineage, policy records and retained evidence.
Support boundary	w-support	Named platform, controller, hardware and billing responders.	Named internal service owner and supplier escalation paths.

Criterion	Weight input	Integrated option evidence	Internal option evidence
Margin risk	w-margin	Operator-cost model using written entitlements and fees.	Equivalent model including integration and ongoing operations.

The scorecard is useful only after mandatory requirements pass. A high aggregate score must not compensate for an unmet tenant boundary, legal requirement, or unworkable cost model. Deferral remains appropriate when neither implementation clears those gates.

The request for proposal (**RFP**) should demand a versioned subscription schedule, hardware and accelerator compatibility matrix, measurement granularity and rounding rules, reservation/overbooking policy, and jointly owned recovery targets. Rafay's matrix says GPU resources in imported Kubernetes clusters cannot automatically scale up or down, illustrating why deployment mode belongs in the capacity plan. (docs.rafay.co)

Before accepting portability claims, perform an exit exercise:

- **Export:** Recover tenant identities, entitlements, rate cards, meter definitions and usage history.
- **Recreate:** Rebuild a tenant and serving endpoint with documented replacement procedures.
- **Revoke:** Remove users, workload credentials, endpoint keys and delegated access.
- **Dispose:** Apply retention and sanitization rules to volumes, snapshots, replicas and backups, with sanitization assessed separately from access removal. (csrc.nist.gov)
- **Reconcile:** Produce the final invoice, credits, reservation release and retained audit record.

GPU Smith's published method emphasizes written acceptance criteria and as-built documentation. (gpusmith.com) In this decision, that provides an appropriate advisory perspective: compare reviewable operating evidence, and retain the records needed to run or replace the resulting service.

Conclusion

A sovereign GPU cloud becomes commercially credible when a tenant order can be admitted, isolated, fulfilled, measured, billed, corrected, and terminated under a defined operating boundary. Owning accelerators establishes capacity; operating this connected process establishes the service.

The Red Hat and Rafay architecture gives operators a concrete integration model to evaluate. Adoption remains conditional on the selected product editions, deployment mode, support responsibilities, and successful acceptance tests. Internal assembly offers a different allocation of integration work and control. It deserves the same scrutiny of staffing, evidence, recovery, and lifecycle costs. Deferral is a valid decision when it has bounded expenditure, measurable exit criteria, and a scheduled review.

The strongest procurement sequence begins with a specific SKU and a representative tenant journey. It then tests isolation, disconnected operation, recovery, invoice reconciliation, and migration. The financial model should use the same entitlement units as the service contract, distinguishing paid reservations from active workload consumption and treating idle capacity explicitly.

Operators should select the approach that clears mandatory requirements and produces the most defensible weighted result using their own costs and demand evidence. The reference architecture can organize that evaluation, but it cannot supply missing operator economics or acceptance results. A decision supported by reproducible records is more useful than a product comparison that assumes the difficult commercial and operational work is already complete.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. <https://www.redhat.com/en/blog/rafay-and-red-hat-publish-joint-reference-architecture-sovereign-ai-cloud-service>
2. <https://www.redhat.com/architect/portfolio/detail/276-sovereign-ai-cloud-as-a-service>
3. <https://www.redhat.com/en/resources/ai-subscription-guide-detail>
4. <https://rafay.co/solutions/gpu-paas-for-service-providers>
5. <https://kubernetes.io/docs/concepts/security/multi-tenancy/>
6. <https://docs.nvidia.com/datacenter/cloud-native/gpu-operator/latest/gpu-sharing.html>
7. <https://prometheus.io/docs/introduction/overview/>
8. <https://csrc.nist.gov/projects/cryptographic-module-validation-program/validated-modules>
9. <https://www.finops.org/framework/capabilities/unit-economics/>
10. <https://gpusmith.com/>
11. <https://digital-strategy.ec.europa.eu/en/library/nis2-commission-implementing-regulation-critical-entities-and-networks>
12. <https://www.enisa.europa.eu/publications/nis2-technical-implementation-guidance>
13. <https://docs.rafay.co/learn/aiml/gpu-paas/usage/overview/>
14. <https://docs.rafay.co/blog/2026/03/23/flexible-gpu-billing-models-for-modern-cloud-providers--powering-the-ai-factory-with-rafay/>
15. <https://docs.rafay.co/releasenotes/2026/gpupaas/v3.1-39/>
16. <https://docs.rafay.co/selfhosted/airgapped/pre-requisites/>
17. https://docs.rafay.co/selfhosted/support_matrix/
18. https://openid.net/specs/openid-connect-core-1_0.html
19. <https://www.keycloak.org/server/hostname>
20. <https://www.openpolicyagent.org/docs/management-bundles>
21. <https://openmeter.io/docs/metering/overview/>
22. <https://openmeter.io/docs/metering/guides/creating-meters>
23. <https://spiffe.io/docs/latest/spiffe/concepts/>
24. <https://docs.cilium.io/en/stable/security/network/encryption-wireguard/>
25. <https://focus.finops.org/docs/specification/v1-0/columns/cost-and-usage/pricing-unit/>
26. <https://www.fedramp.gov/2026/agencies/use/>
27. <https://kubernetes.io/docs/concepts/security/secrets-good-practices/>
28. <https://www.energy.gov/sites/default/files/2024-07/best-practice-guide-data-center-design.pdf>

29. https://docs.redhat.com/en/documentation/openshift_container_platform/4.21/html/hosted_control_planes/hcp-networking
30. <https://docs.stripe.com/billing/subscriptions/usage-based/recording-usage>
31. <https://docs.stripe.com/billing/subscriptions/usage-based/recording-usage-api>
32. <https://csrc.nist.gov/pubs/sp/800/88/r2/final>
33. <https://docs.stripe.com/api/billing/meter-event-adjustment/create>
34. <https://docs.nvidia.com/datacenter/cloud-native/gpu-operator/latest/gpu-operator-mig.html>
35. <https://gpusmith.com/articles/en/dcgm-exporter-gpu-monitoring-metrics-alerts-triage>
36. <https://gpusmith.com/articles/en/mlperf-inference-6-1-gpu-procurement>
37. <https://mlcommons.org/benchmarks/inference-datacenter/>
38. <https://prometheus.io/docs/practices/histograms/>
39. https://docs.redhat.com/en/documentation/openshift_container_platform/4.22/html/hosted_control_planes/hcp-overview
40. <https://docs.rafaq.co/selfhosted/airgapped/upgrade/>
41. <https://docs.rafaq.co/selfhosted/airgapped/backup-restore-s3-compatible/>
42. <https://www.thegreengrid.org/node/372>
43. https://docs.rafaq.co/security/whitelist_paas/
44. <https://docs.rafaq.co/selfhosted/airgapped/overview/>
45. <https://docs.nvidia.com/datacenter/cloud-native/gpu-operator/latest/install-gpu-operator-air-gapped.html>
46. https://docs.rafaq.co/supportmatrix/gpu_paas/

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the [hardware library](#), [GPU references](#), [networking references](#) and [vendor research](#). [Contact GPU Smith](#) with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: <https://gpusmith.com>

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.