GPU SMITH
COMPUTE ENGINEERING

GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

VMware Private AI Cloud in VCF 9.1.1: Now vs Next

> 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 · 2026-10-03 · vmware private ai cloud · vcf 9.1.1 · vmware ai factory · private ai services · model runtime · gpu infrastructure · ai gateway · private cloud · enterprise ai

Original article: https://gpusmith.com/articles/en/vmware-private-ai-cloud-vcf-9-1-1

In this article

1. Executive Summary
2. Introduction and Background
3. Product and Platform Architecture
4. VCF 9.1.1 Delivery Status and Dependencies
5. Hardware Path and Deployment Requirements
6. Use Cases and Functional Capabilities
7. Operations, Security, and Support Boundaries
8. Market Share, Adoption, and Competitor Context
9. Data Analysis and Evidence
10. Implications and Future Directions
11. Frequently Asked Questions (FAQs)
12. Conclusion

Executive Summary

VMware Cloud Foundation (VCF) 9.1.1 is available for a bounded Private AI Cloud pilot, with multi-tenant Model Runtime sharing as the clearest currently delivered private-AI feature. Broadcom introduced **VMware Private AI Cloud** and **VMware AI Factory** on August 31, 2026, and VMware announced VCF 9.1.1 general availability on September 3. The August launch groups new and forthcoming services; the September product article places Model Runtime sharing under generally available capabilities and AI Gateway, secure agent sandboxing, Agent Harness, and model autoscaling under future releases. A buyer should accept the former only after a workload test and keep the latter out of a current-release support commitment. (blogs.vmware.com) (blogs.vmware.com) (blogs.vmware.com)

The immediate deployment questions are concrete. Private AI Services runs on the vSphere Supervisor; the network choice affects how services and endpoints are deployed. The server, GPU, driver, guest operating system, GPU software, and Private AI Services package must match a supported configuration. NVIDIA's current certified-hypervisor listing identifies a vSphere 9.1 Hopper SXM scope, while AMD documents narrower ESXi 9.1 virtualization paths for selected Instinct parts. Those statements do not certify every proposed AI Factory bill of materials. The VCF 9.1.1 release notes document install and upgrade paths, but GPU mode and application recovery still require rehearsal. (blogs.vmware.com) (docs.nvidia.com) (rocm.docs.amd.com) (techdocs.broadcom.com)

There is a **licensing gate** before production purchase. VMware's September blog says Private AI Services are included in VCF, whereas its release notes describe availability under a Private AI Foundation with NVIDIA license. The customer needs a signed SKU, feature-entitlement, and support schedule for the exact build. Broadcom's claim that sharing removes redundant model deployments has no universal savings percentage; the report's capacity worksheet compares measured dedicated and pooled replicas, GPU memory, concurrency, and failover headroom at equal latency targets. An older MLCommons result on ESXi 9.0.0 is useful context, not a VCF 9.1.1 benchmark. (blogs.vmware.com) (techdocs.broadcom.com) (docs.mlcommons.org) (docs.nvidia.com)

The recommended decision is **pilot now** if the current model-sharing path, supported hardware, data isolation, and measured service level meet the use case. **Wait or narrow scope** if a future gateway or sandbox is mandatory, if the GPU matrix is unresolved, or if the entitlement cannot be documented. The proof of concept should include cross-tenant prompt and retrieval tests, noisy-neighbor load, GPU-host failure, offline update, and an upgrade rollback rehearsal. CISA's procurement guidance supports putting security requirements and patch expectations in the contract; NIST's AI risk framework supports testing before deployment and during operation. (kubernetes.io) (www.cisa.gov) (airc.nist.gov)

Introduction and Background

VMware Private AI Cloud is Broadcom's name for a private-cloud platform intended to place inference, agent applications, and conventional workloads under one operating model. **VMware AI Factory** is the software foundation of that proposition, while **VMware Cloud Foundation (VCF) 9.1.1** is the specific release an infrastructure team can install now. Broadcom introduced the Private AI Cloud and AI Factory names on August 31, 2026; VMware announced VCF 9.1.1 general availability on September 3. Those dates matter because the launch describes both existing and future functions. (investors.broadcom.com) (investors.broadcom.com)

The practical decision is narrower than the product name. A team can evaluate multi-tenant Model Runtime sharing on VCF 9.1.1, but it should not make the newer AI Gateway or secure agent sandbox a current acceptance criterion: VMware lists them under future release capabilities. It separately labels AI Assistant for VCF as a technology preview. A release announcement is evidence of a product boundary, not proof that every feature in a launch diagram is supported in the same software bill of materials. (blogs.vmware.com) (blogs.vmware.com)

This report is for enterprise architects, VCF administrators, security reviewers, and buyers deciding whether to pilot, upgrade, or defer. **GPU Smith** states that its private AI infrastructure work covers specification, procurement, integration, and validation against written acceptance criteria. That first-party perspective makes the relevant test a workload-specific bill of materials, isolation evidence, and a signed support path, rather than an abstract comparison of slogans. It does not make GPU Smith a competing software platform in the comparison below. (gpusmith.com)

The public evidence is unusually precise in some places and incomplete in others. Broadcom's VCF 9.1.1 notes call the release a maintenance update with an updated bill of materials, while its September product blog identifies model sharing as a new generally available capability. The Private AI Services notes still describe version 2.1 for VCF 9.1 and use different license language from the later blog. The buyer therefore needs a dated entitlement and compatibility statement for the proposed VCF 9.1.1 deployment. (techdocs.broadcom.com) (techdocs.broadcom.com)

Product and Platform Architecture

What the names cover

Private AI Cloud is the umbrella proposition. **AI Factory** denotes the software-defined foundation. **VCF** supplies the private-cloud platform, and **VCF Private AI Services** supplies services such as Model Runtime and model-facing applications. Broadcom describes the AI Factory as the foundation of Private AI Cloud, and its earlier Private AI Services description names Model Runtime, an existing machine-learning API gateway, data indexing, and Agent Builder. The newly announced **AI Gateway** has a broader future-release description, including routing across on-premises and cloud models; the older gateway name should not be treated as evidence that this new scope already ships. (blogs.vmware.com)

The model-serving layer sits on the **vSphere Supervisor**, which hosts Kubernetes services inside the VCF environment. VMware's deployment guidance says the architecture can use either a Virtual Distributed Switch (VDS) networking path or a Virtual Private Cloud (VPC) path. VCF Automation requires a Supervisor configured for VPC networking; with VDS, administrators perform some Private AI Services activation and endpoint deployment manually through a command-line interface and YAML manifests. Changing between those Supervisor networking modes requires planned redeployment. These are design choices for a pilot, not details to postpone until rollout. (blogs.vmware.com) (blogs.vmware.com) (blogs.vmware.com)

The service boundary includes infrastructure that may sit outside the model runtime. VMware's retrieval-augmented generation example provisions its PostgreSQL vector database externally as a prerequisite. A buyer should list the database, model registry, identity provider, certificate authority, image mirror, observability destination, and backup target in the deployment diagram, with a named owner for each. The deployment guide supports the external-database point; the remaining items are explicit design questions for the proposed implementation. (blogs.vmware.com)

Multi-tenant sharing, with an isolation test

VMware says VCF 9.1.1 enhances Model Runtime so one service can run and scale models for the organization while teams use separate namespaces for private data. It says model sharing avoids redundant deployments. That is a plausible resource-efficiency mechanism, but the number of GPU replicas and the memory saved depend on workload, model footprint, concurrency, and the failure policy. The capacity worksheet later in this report deliberately treats those as measured variables. (investors.broadcom.com)

A **Kubernetes namespace** scopes namespaced resources within a cluster; it does not scope cluster-wide objects such as nodes or persistent volumes. Kubernetes also allows ingress and egress by default in a namespace without policies, and NetworkPolicy requires a network plugin that enforces it. Consequently, a shared Model Runtime pilot should verify the entire prompt, retrieval, response, log, trace, and credential path across two test tenants. Namespace labels alone cannot demonstrate data-path isolation. (kubernetes.io) (kubernetes.io) (kubernetes.io)

For a controlled trial, create tenants A and B with distinct model endpoint permissions, data stores, keys, audit identities, and network policies. Send recognizable synthetic prompts and retrieval records from each. Prove that a tenant cannot query the other's private material, discover its secrets, or access its logs. Then induce a

burst on A and measure B's time to first token, token throughput, queue depth, and error rate. The experiment tests both isolation and noisy-neighbor behavior; it makes no assumption that namespaces guarantee performance isolation. (kubernetes.io) (mlcommons.org)

VCF 9.1.1 Delivery Status and Dependencies

The decisive distinction is between **general availability**, **technology preview**, **future release**, and **unverified entitlement**. Broadcom's August 31 AI Factory release says its list contains “new and forthcoming” services. VMware's September 3 article is more useful for an acceptance plan because it explicitly divides generally available Model Runtime sharing from future capabilities. The release notes provide a separate check on packaged software, upgrade path, and support language. (investors.broadcom.com) (techdocs.broadcom.com)

Table 1 maps claims to evidence and to the proof a buyer should require before signing off a VCF 9.1.1 pilot. “Future” means no delivery date or contract entitlement is inferred from the launch announcement.

Capability	Status on VCF 9.1.1	Dependency or evidence	Acceptance test
Multi-tenant Model Runtime sharing	**GA**, per VMware's September 3 feature statement (blogs.vmware.com)	Private AI Services deployment, compatible VCF and GPU stack	Run two isolated namespaces against one service; prove separate data paths and measure contention.
VCF Automation GitOps service	**Tech preview**; the underlying Argo CD Supervisor service has a distinct earlier support history (blogs.vmware.com)	VCF Automation Org integration	Obtain written support terms before using it for a production delivery dependency.
AI Assistant for VCF	**Tech preview** in the VCF 9.1.1 announcement (blogs.vmware.com)	Locally configured model or another supported private endpoint (blogs.vmware.com)	Test on non-production operational data and record retention behavior.
New AI Gateway, agent sandbox, Agent Harness, model autoscaling	**Future release** in VMware's September 3 classification (blogs.vmware.com)	Version, entitlement, and support date unannounced in cited material	Make each a separate roadmap option with an agreed substitute or deferral criterion.
Heterogeneous physical-server automation	**Announced integration objective**, not established VCF 9.1.1 GA (investors.broadcom.com)	MetalSoft integration, validated OEM firmware and network stack	Require a shipping build number and supported hardware list, then rehearse provisioning and repaving.
Private AI Services entitlement	**Contract clarification required** (techdocs.broadcom.com)	September blog says included in VCF; release notes name a Private AI Foundation with NVIDIA license	Map each service, GPU software, support entitlement, and license key to a signed SKU schedule.

The matrix does not imply that a technology preview is unusable in a lab. It means a production dependency needs support and rollback terms matching its status. The notable discrepancy is licensing: one September VMware blog says Private AI Services are included in VCF, while the release notes say their releases are available under a VMware Private AI Foundation with NVIDIA license. An OEM product guide also describes

Private AI Foundation as part of the VCF license. The public statements do not settle an individual customer's entitlement, so procurement must get the governing order form and support schedule. (lenovopress.lenovo.com)

The same discipline applies to version history. Broadcom's Private AI Services release notes dated May 12, 2026 describe version 2.1 for VCF 9.1, with a default of up to 15 model endpoint replicas per namespace. That is useful evidence about a documented package, not a guarantee that a later Model Runtime sharing implementation inherits every limit unchanged. Ask for the specific Private AI Services package version and its limits on VCF 9.1.1. (techdocs.broadcom.com) (techdocs.broadcom.com)

Hardware Path and Deployment Requirements

VCF AI ReadyNodes combine VCF, NVIDIA AI Enterprise software, and certified hardware in Broadcom's ecosystem description. The August AI Factory release names Cisco, Dell Technologies, Lenovo, and Supermicro among providers. A name in a partner list is a starting point, not a certified bill of materials: Broadcom directs buyers to its compatibility guide for exact server and GPU status. (investors.broadcom.com) (blogs.vmware.com)

NVIDIA's certification list names vSphere 9.1 and identifies Hopper SXM in its current listed scope. (docs.nvidia.com) Broadcom's August 27 post describes Blackwell and Hopper coverage; a Blackwell procurement therefore needs written configuration-level confirmation. (blogs.vmware.com) Its certification explanation says a result applies to the tested partner stack, GPU architecture, hypervisor configuration, and driver baseline. NVIDIA's vGPU support documents add that server hardware, hypervisor version, and guest operating system must align. The purchaser should therefore demand exact server model, BIOS and firmware, GPU, host driver, guest driver, GPU Operator, VCF build, and guest operating system versions in one compatibility record. (docs.nvidia.com) (docs.nvidia.com)

There are at least two materially different GPU paths. **Passthrough** gives a virtual machine a physical GPU; **vGPU for Compute** has separate software and driver requirements. NVIDIA says GPU passthrough and vGPU for Compute differ in NVIDIA license requirements, while hypervisor licensing is separate. Broadcom also documents that DirectPath passthrough can limit some virtual-machine operations. The migration and recovery plan must be tested for the chosen mode rather than inferred from generic VCF high-availability features. (docs.nvidia.com) (knowledge.broadcom.com)

AMD support requires the same precision. AMD's current ROCm release notes list VMware ESXi 9.1 passthrough with an Ubuntu 24.04 guest for Instinct MI350P, and a distinct ESXi 9.1 SR-IOV path for MI355X and MI350X. Broadcom's August AI Factory release frames the wider AMD configuration and zero-touch deployment as collaboration and future delivery. An ESXi virtualization entry is not by itself proof that a specific VCF 9.1.1 AI Factory service stack, GPU operator, and server are jointly supported. (rocm.docs.amd.com) (rocm.docs.amd.com)

The infrastructure qualification package should contain these records:

- **Compatibility:** the precise VCF, ESXi, Private AI Services, GPU software, firmware, and guest versions, with a dated vendor support statement. (docs.nvidia.com)

- **Mode:** passthrough, vGPU, or SR-IOV configuration and its license, VM sizing, peer-to-peer, migration, and recovery restrictions. (docs.nvidia.com) (docs.nvidia.com)
- **Server:** OEM-supported chassis, power and thermal envelope, accelerator topology, network adapters, and BIOS profile. (blogs.vmware.com)
- **Network:** separation of management and GPU traffic where the reference design requires it; Broadcom's Supermicro design uses a separate RoCEv2 GPU fabric. (blogs.vmware.com)
- **Provisioning:** a shipping MetalSoft integration version and repeatable firmware, driver, and bare-metal reprovisioning procedure if that automation is in scope.
- **Supplier ownership:** Broadcom for VCF software, the OEM for server hardware, the GPU vendor for accelerator software, and an identified model provider for weights and serving behavior. Broadcom explicitly describes the VCF/OEM support split. (blogs.vmware.com)

Broadcom's VCF 9.1.1 release notes allow new environments to deploy directly on the release and document direct upgrades from VCF 5.2.x or 9.0.x. For a 9.1.0.x patch, the fleet lifecycle component is updated first. Those paths establish installability; they do not substitute for the application and GPU-driver rollback rehearsal the buyer should require. (techdocs.broadcom.com) (techdocs.broadcom.com)

Use Cases and Functional Capabilities

Where a pilot has a defensible scope

The strongest current use case is an organization with an existing VCF operating model that wants several teams to consume a common model runtime while keeping each team's private retrieval data and policy in its own namespace. VMware specifically describes one service scaling models for the organization and separate namespaces for teams. The trial should measure whether fewer active replicas actually meet the same service-level objective under simultaneous load; the vendor's asserted infrastructure saving is not a measured result for an individual buyer.

A second use case is a controlled **retrieval-augmented generation (RAG)** service. Private AI Services already describes data indexing and retrieval, and VMware's networking guide shows an externally provisioned PostgreSQL vector database. The acceptance boundary should include document ingestion, index update, tenant-specific retrieval filtering, and an end-to-end deletion test. A shared model endpoint can still leak context if the retrieval and logging layers are not separately governed. (blogs.vmware.com) (kubernetes.io)

A third use case is a disconnected deployment, provided it is treated as a full supply-chain exercise. Broadcom documents a VCF offline depot with local metadata as well as installation binaries, and Private AI Services 2.1 introduced an artifact mirroring tool. That supports an offline design path, but the buyer still needs a tested import procedure for every container, model weight, driver, license, certificate, and update. The cited materials do not establish one universal offline update sequence for the complete AI Factory stack. (knowledge.broadcom.com) (techdocs.broadcom.com)

The pilot should be written as a set of pass/fail observations:

- **Isolation:** tenant A's prompt, retrieval result, response, trace, and secret must not appear in tenant B's accessible surfaces. (kubernetes.io) (kubernetes.io)

- **Authorization:** record which identities can create endpoints, change a model, rotate keys, and read usage data. (kubernetes.io)
- **Concurrency:** replay measured arrival rates and peak simultaneous sessions; retain per-tenant latency distributions. (mlcommons.org)
- **Noisy neighbor:** saturate one namespace and document queueing, token rate, and tail latency in the other. (mlcommons.org)
- **Failover:** remove a GPU host and an inference replica; measure recovery, degraded throughput, and any prompt loss. (docs.nvidia.com)
- **Change:** upgrade a driver, Private AI Services package, and VCF component in a staging copy; demonstrate a documented recovery path. (docs.nvidia.com)
- **Cost:** compute reserved GPU memory, hardware power, license and support costs, and cost per successful request from the measured workload. (www.finops.org)

The future-facing uses need different treatment. The September product article puts AI Gateway routing, secure agent sandboxes, an Agent Harness, and model autoscaling under future release capabilities. A buyer whose production architecture requires central cross-cloud routing or isolated execution of agent-generated code should defer that dependency until Broadcom publishes an eligible build and support statement. An older ML API Gateway entry in Private AI Services should be evaluated on its documented functions, not renamed into the new roadmap item.

Operations, Security, and Support Boundaries

Day-two evidence

Broadcom describes token throughput, latency, and compute and memory utilization as observability targets for AI Factory. Its Private AI Services 2.1 release notes describe observability across inference engines, GPU utilization, knowledge-base indexing, and agents, with tracing among users, models, agents, and knowledge bases. The VCF Operations blog additionally describes two-second metric streaming for vSphere Kubernetes Service clusters. These are different metric surfaces; a pilot should demonstrate that each required metric exists at the intended tenant, model, replica, GPU, and request granularity. (techdocs.broadcom.com) (blogs.vmware.com)

An exporter configuration is not proof that the hardware emits a value. NVIDIA's Data Center GPU Manager documentation says GPU, driver, permissions, and deployment configuration determine metric availability. A dashboard review should deliberately compare counter samples with raw driver telemetry and request logs. Use a synthetic load with known token counts to validate units, missing data, clock alignment, and whether a tenant administrator can see another tenant's usage. OpenTelemetry's semantic conventions can provide shared names for traces and metrics, but they do not create missing instrumentation. (docs.nvidia.com) (opentelemetry.io)

Security gates should follow the complete data path. Broadcom describes vDefend microsegmentation at the VPC layer; Kubernetes says NetworkPolicy requires a supporting plugin and that namespaces without policies allow traffic by default. NIST's AI Risk Management Framework calls for testing before deployment and

regularly during operation. The right acceptance evidence is an exported policy set plus observed deny and allow tests, including model endpoint, vector database, metadata service, registry, and administrative interfaces. (blogs.vmware.com) (kubernetes.io) (airc.nist.gov)

For agent applications, do not imply that the newly announced sandbox is a current VCF 9.1.1 control. The September article presents it as future, while the August release uses future tense for new vDefend agentic enhancements. Until a shipping build is documented, constrain execution through existing tested workload controls and a clearly owned tool-access policy. If agents have credentials, require evidence of where those credentials reside, who can rotate them, and what the audit trail records. (investors.broadcom.com)

The **air-gap** gate has two layers. Broadcom's offline-depot guidance instructs operators to transfer metadata and binaries into a local depot; its disconnected-licensing instructions describe a separate registration workflow. A local Harbor repository appears in Broadcom's earlier VCF AI deployment guidance. The customer should run a full installation, patch, model update, certificate rotation, and support-bundle export without live external network access, then document any staging workstation and transfer control. (knowledge.broadcom.com) (knowledge.broadcom.com) (blogs.vmware.com)

The support plan needs explicit handoffs:

- **Broadcom:** VCF build, Private AI Services package, feature status, upgrade path, and support scope.
- **OEM:** server, GPU carrier, firmware, power and cooling, and replacement-part support.
- **GPU vendor:** driver matrix, vGPU or passthrough mode, license service, and diagnostics. (docs.nvidia.com) (docs.nvidia.com)
- **Model provider:** weight license, model version, supported serving engine, and model-level behavior.
- **Customer:** data policy, tenant identities, prompt and log retention, workload tests, and change control. (airc.nist.gov)

CISA recommends putting product-security requirements into procurement language and asking about machine-readable software bills of materials, baseline logging, single sign-on, and patch installation. For this stack, the contract appendix should name supported versions, a critical-vulnerability response contact and timetable, update cadence, rollback responsibilities, and remedies if a roadmap feature is required for the intended service. Public product pages are not a substitute for that signed appendix. (www.cisa.gov) (www.cisa.gov) (www.cisa.gov)

Market Share, Adoption, and Competitor Context

No public adoption figure surfaced in the researched primary material that isolates **VMware Private AI Cloud on VCF 9.1.1**. Broadcom's claim that VCF customers can run more than 150 models is a capability statement, not a count of installed Private AI Cloud deployments. Likewise, MLCommons reported 30 submitters and 120 systems in its September 2026 inference round, but those are benchmark participation counts for an industry test, not VMware market share. This report therefore does not assign a market-share percentage to the new product name. (investors.broadcom.com) (mlcommons.org)

The same separation applies to performance evidence. MLCommons has an earlier Dell and Broadcom submission using eight H200 GPUs on VMware ESXi 9.0.0. It demonstrates that a virtualized configuration was submitted to that benchmark, but it is neither a VCF 9.1.1 test nor evidence of the new Model Runtime sharing implementation. MLCommons distinguishes available, preview, and experimental submission categories and separates Offline, Server, and Interactive test scenarios; any vendor result used in procurement should identify the exact configuration, scenario, and release. (docs.mlcommons.org) (docs.mlcommons.org)

Several alternative architectures offer private model serving, but their labels and delivery boundaries differ. **Red Hat OpenShift AI** identifies KServe RawDeployment as the model-deployment migration path in its 3.5 support-removals notes; its current release notes describe model-server dashboards. **Nutanix Enterprise AI** announced a generally available 2.8 release in August 2026 and a generally available Model Context Protocol gateway. Cisco documents AI POD designs that combine its infrastructure with Red Hat OpenShift or Nutanix Enterprise AI. These are architecture comparisons, not claims that the products have identical control sets. (docs.redhat.com) (docs.redhat.com) (www.nutanix.com) (www.nutanix.com) (www.cisco.com)

Table 2 compares the component a buyer would actually evaluate. Each row names the source of the platform claim; exact license and support terms still need a quote for the selected hardware and workload.

Platform path	Documented serving or control boundary	Procurement test
VMware Private AI Cloud on VCF 9.1.1	GA multi-tenant Model Runtime sharing through separate namespaces; new AI Gateway remains future release (blogs.vmware.com)	Prove namespace data isolation, GPU and token telemetry, license entitlement, and upgrade support.
Red Hat OpenShift AI (3.5)	KServe RawDeployment is the documented model-deployment migration path; release notes describe serving metrics (docs.redhat.com) (docs.redhat.com)	Confirm current serving mode, support status, OpenShift dependencies, and the same workload latency target.
Nutanix Enterprise AI	Nutanix announced Enterprise AI 2.8 GA and a Model Context Protocol gateway; its inference API documents API-key access (www.nutanix.com) (www.nutanix.com) (www.nutanix.dev)	Verify endpoint policy, exact NVIDIA and Kubernetes dependencies, and production eligibility of chosen capabilities.
Cisco AI POD with partner software	Cisco describes combinations of UCS, Nexus, NVIDIA software, and Red Hat or Nutanix serving layers (www.cisco.com) (www.cisco.com)	Identify who supports each layer and compare a complete bill of materials, not a hardware-only price.

The table's central implication is that a feature name is not a substitute for a supported stack. VMware's current strength for a VCF customer is an integrated path to shared model runtime under an existing private-cloud operating model. An alternative may offer a different gateway or serving interface today, but it may also move responsibility among platform, GPU, and model vendors. Evaluate the same prompt mix, latency target, security boundary, power budget, and support contract across rows. (www.cisco.com) (mlcommons.org)

Data Analysis and Evidence

A capacity worksheet for shared model runtime

The economic claim most worth testing is avoided redundant model deployment. Broadcom says sharing lets multiple tenants use one Model Runtime and reduces duplicated model copies, while its announcement also speaks of GPU pooling. Neither statement gives a transferable savings percentage. A buyer should measure active replicas, model weight memory, key-value cache memory, peak token rates, and failure reserve in the proposed configuration. NVIDIA's sizing guidance likewise starts from measured response time and expected peak load. (docs.nvidia.com)

Table 3 is a symbolic worksheet, not a benchmark. All inputs are measured for the same model, precision, serving engine, prompt distribution, and latency objective. Let **T** be the number of tenants, $\mathbf{r_i}$ each tenant's dedicated active replicas, **R** the pooled active replicas measured under simultaneous load, $\mathbf{F_D}$ and $\mathbf{F_p}$ the dedicated and pooled failover replicas, **W** the measured GPU memory per loaded model replica, and $\mathbf{K_D}$ and $\mathbf{K_p}$ the measured peak key-value cache and runtime overhead. (docs.nvidia.com) (docs.nvidia.com)

Measure	Dedicated copies	Shared runtime	Evidence required
Active replicas	Sum of r_i across T tenants	R under coincident demand	Replica and request traces at target latency.
Failover reserve	F_D additional replicas	F_p additional replicas	Single-host and single-replica failure test.
Reserved GPU memory	(Sum of r_i + F_D) × W + K_D	(R + F_p) × W + K_p	GPU memory counters and serving-engine allocation logs.
Peak service quality	Per-tenant token rate and tail latency	Same per-tenant targets under pooled load	Concurrent load test using identical prompts and output lengths. (mlcommons.org)
Net capacity difference	Baseline dedicated reservation	Dedicated reservation minus shared reservation	Report only with equal service quality, failover headroom, and compatible license costs. (www.finops.org)

The calculation can be positive, zero, or negative. Pooling is attractive when tenants' peaks are not perfectly aligned and the shared service can run fewer model replicas. It can lose the expected advantage when simultaneous concurrency requires the same replica count, extra failover headroom, or larger key-value caches. Even a memory reduction is not automatically a cost reduction if licensed GPU capacity, network fabric, storage, and operations remain unchanged. Those are deductions from the worksheet, not claimed VCF 9.1.1 benchmark results. (www.finops.org)

For reproducibility, report the model identifier and checksum, quantization, input and output token distributions, arrival process, warm-up period, concurrency, batching, GPU profile, driver, serving engine, and all observed tail latencies. MLCommons distinguishes datacenter Offline, Server, and Interactive scenarios; its Endpoints guidance includes throughput, interactivity, 95th-percentile time to first token, and concurrency. This is a useful disclosure standard even when the pilot does not submit to MLPerf. FinOps Foundation guidance defines utilization efficiency as actual utilization divided by provisioned capacity and says unit economics should extend beyond cost per token. (docs.mlcommons.org) (mlcommons.org) (www.finops.org) (www.finops.org)

Other public quantities deserve careful labels. VMware says VCF Operations can stream Kubernetes metrics every two seconds, but that is not an independently measured token-latency sampling rate. Private AI Services 2.1 release notes specify up to 15 model endpoint replicas per namespace and a default /24 network allocation per replica; those are dated package limits to check against the proposed build. Broadcom's offline-depot guide recommends at least 1 TB of storage for its depot, a staging requirement rather than GPU capacity. None of these figures establishes Private AI Cloud market share or a universal total-cost-of-ownership benefit. (blogs.vmware.com) (knowledge.broadcom.com)

Implications and Future Directions

An enterprise already operating VCF can make a bounded **pilot decision now**: follow the documented upgrade path, obtain a supported GPU configuration (docs.nvidia.com), confirm service entitlement, and test multi-tenant Model Runtime sharing with real workload traces. A team whose business case depends on the new AI Gateway, agent sandbox, automated heterogeneous server provisioning, or a promised autoscaling function should make the delivery date and support terms a separate purchase condition (www.cisa.gov). VMware classifies the gateway, sandbox, and autoscaling as future capabilities; Broadcom's MetalSoft description uses future tense.

The request for proposal should ask for the following evidence, each with an owner and acceptance date:

- **Software bill:** exact VCF, Private AI Services, VKS, GPU Operator, driver, and inference-engine versions. (docs.nvidia.com)
- **Entitlement:** signed SKUs, Private AI Services inclusion, NVIDIA software rights, support level, and renewal terms. (docs.nvidia.com)
- **Hardware matrix:** server, accelerator, network adapter, firmware, BIOS, guest operating system, and GPU mode. (docs.nvidia.com)
- **Isolation record:** namespace, VPC, firewall, network policy, secrets, storage, logging, and audit boundaries. (kubernetes.io)
- **Scale envelope:** maximum endpoints and replicas, namespace limits, model sizes, and supported serving engines for the offered build.
- **SLO run:** prompt mix, peak concurrency, token throughput, tail latency, and failover headroom. (mlcommons.org)
- **Upgrade cadence:** supported order for VCF, Private AI Services, GPU drivers, and Kubernetes components. (docs.nvidia.com)
- **Rollback:** supported snapshot, backup, restore, and version-reversal procedure, demonstrated in staging.
- **Security response:** vulnerability intake and patch timetable, software bill of materials, signatures, and audit-log retention. (www.cisa.gov) (contribute.cncf.io)
- **Offline behavior:** depot, model artifacts, activation, telemetry destinations, and support-bundle export in a disconnected test.
- **Roadmap remedy:** delivery milestone and contractual alternative for every future control in the proposed architecture.
- **Responsibility:** escalation path across Broadcom, OEM, GPU vendor, model provider, and customer.

CISA's secure-demand guide recommends using procurement language for product-security expectations; Cloud Native Computing Foundation guidance recommends checking artifact hashes and signatures. Those general practices become concrete here as a signed configuration matrix, artifact manifest, and an offline update rehearsal. GPU Smith's published method says its assessment produces a reference architecture and itemized bill of materials; an independent engineering review can apply that discipline to this product without becoming a software-vendor row in the platform table. (www.cisa.gov) (contribute.cncf.io) (gpusmith.com)

Frequently Asked Questions (FAQs)

Is VMware Private AI Cloud itself a new VCF 9.1.1 SKU?

The public material distinguishes the Private AI Cloud umbrella, AI Factory foundation, VCF release, and Private AI Services, but does not settle a customer's purchasable SKU and entitlement from those labels alone. The September blog says Private AI Services are included in VCF, whereas the release notes refer to a Private AI Foundation with NVIDIA license. Ask Broadcom and the reseller for a signed line-item mapping.

Can AMD GPUs be used today?

AMD documents specific ESXi 9.1 passthrough and SR-IOV paths for named Instinct parts and Ubuntu guests. That supports evaluation of a hardware path. Validation of AI Factory services, server combinations, and ROCm versions together on VCF 9.1.1 still requires a combined compatibility record. Require the combined compatibility record and an application-level pilot. (rocm.docs.amd.com) (rocm.docs.amd.com) (docs.nvidia.com)

Does one shared model runtime guarantee tenant privacy or lower GPU cost?

No universal guarantee follows from the launch statement. VMware says separate namespaces maintain tenant privacy and eliminate redundant model deployments, while Kubernetes documents that namespaces do not isolate cluster-wide resources and that network policies need an enforcing plugin. Privacy requires a data-path test; savings require the worksheet's measured replica count, memory, concurrency, and failover reserve. (kubernetes.io) (kubernetes.io)

What should trigger a wait decision?

Wait or narrow the project if a mandatory control is only listed as a future release, the required GPU and server combination lacks a support statement, licensing is unresolved, or the pilot misses its agreed isolation and latency targets. AI Gateway, secure agent sandbox, and autoscaling are in VMware's future-release group, while the VCF Automation GitOps service is a technology preview. Those are different statuses and should have different contract treatment.

Conclusion

VCF 9.1.1 is a defensible base for a bounded private-AI pilot, especially when an enterprise already has VCF operations and can validate a supported GPU stack. The available-now centerpiece is multi-tenant Model Runtime sharing. Its value depends on whether a common service meets tenant isolation, latency, failover, and

cost targets on the buyer's actual prompts and concurrency. The public launch does not supply a transferable benchmark or a universal saving. (docs.nvidia.com)

The procurement boundary is equally clear. New AI Gateway routing, secure agent sandboxing, Agent Harness, and model autoscaling belong to a later release path in VMware's September classification. GitOps integration and the VCF operations assistant are labeled technology previews. Broadcom's physical-server automation partnership is an announced objective that needs a shipping-version check. A production design should accept each component only at its documented status and with an explicit entitlement.

The remaining work is concrete: secure a signed software and hardware matrix, reconcile the Private AI Services license language, test the full tenant data path and GPU failure domain, rehearse upgrades and offline updates, and attach measured pass/fail results to the purchase decision. That turns the question “now or next?” into an evidence gate that can be revisited when a later release and its support documents actually arrive. (airc.nist.gov)

Source links

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

1. https://blogs.vmware.com/cloud-foundation/2026/09/03/announcing-general-availability-of-vmware-cloud-foundation-9-1-1/
2. https://blogs.vmware.com/cloud-foundation/2026/09/03/explore-2026-vmware-ai-factory-and-other-new-ai-innovations-in-vcf/
3. https://blogs.vmware.com/cloud-foundation/2026/06/11/deploying-vmware-cloud-foundation-private-ai-services-navigating-supervisor-networking-stack/
4. https://docs.nvidia.com/certification-programs/certified-hypervisors/latest/hypervisors-list.html
5. https://rocm.docs.amd.com/en/latest/about/release-notes.html
6. https://techdocs.broadcom.com/us/en/vmware-cis/vcf/vcf-9-0-and-later/9-1/release-notes/vmware-cloud-foundation-9-1-1-0-release-notes.html
7. https://techdocs.broadcom.com/us/en/vmware-cis/private-ai/foundation-with-nvidia/9-0/private-ai-release-notes/vmware-private-ai-services-release-notes.html
8. https://docs.mlcommons.org/inference_results_v5.1/
9. https://docs.nvidia.com/enterprise-reference-architectures/nim-llm-runai-vanilla-kubernetes/latest/sizing-guidelines.html
10. https://gpusmith.com/articles/en/gpu-cluster-reliability-budget
11. https://kubernetes.io/docs/concepts/overview/working-with-objects/namespaces/
12. https://www.cisa.gov/resources-tools/resources/secure-demand-guide
13. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/
14. https://investors.broadcom.com/news-releases/news-release-details/broadcom-introduces-vmware-private-ai-cloud-enabling-enterprises
15. https://investors.broadcom.com/news-releases/news-release-details/broadcom-announces-vmware-ai-factory-enabling-faster-time
16. https://gpusmith.com/
17. https://blogs.vmware.com/cloud-foundation/2026/01/15/activate-vcf-private-ai-services/
18. https://kubernetes.io/docs/concepts/services-networking/network-policies/

19. https://kubernetes.io/docs/concepts/policy/
20. https://mlcommons.org/benchmarks/endpoints/
21. https://blogs.vmware.com/cloud-foundation/2026/09/03/from-bottleneck-to-breakthrough-centralizing-gitops-at-enterprise-scale-with-vcf-9-1-1/
22. https://blogs.vmware.com/cloud-foundation/2026/09/03/new-ai-and-kubernetes-private-cloud-operations-capabilities-in-vmware-cloud-foundation-9-1-1/
23. https://lenovopress.lenovo.com/lp2338-lenovo-thinkagile-fx650-v4-hyperconverged-system
24. https://blogs.vmware.com/cloud-foundation/2024/05/02/vmware-private-ai-foundation-with-nvidia-server-guidance/
25. https://blogs.vmware.com/cloud-foundation/2026/08/27/vcf-achieves-nvidia-hypervisor_certification/
26. https://docs.nvidia.com/certification-programs/certified-hypervisors/latest/nvidia-certified-hypervisors.html
27. https://docs.nvidia.com/vgpu/latest/grid-vgpu-release-notes-vmware-vsphere/validated-platforms.html
28. https://knowledge.broadcom.com/external/article?legacyId=2142307
29. https://docs.nvidia.com/ai-enterprise/release-7/latest/infra-software/vgpu/features/multi-vgpu-p2p.html
30. https://blogs.vmware.com/cloud-foundation/2026/08/31/deploy-vcf-on-supermicro-hgx-servers-with-nvidia-gpus-for-ai-workloads/
31. https://blogs.vmware.com/cloud-foundation/2026/02/25/vcf-9-0-server-certification-preserving-your-hardware-investment-and-giving-the-best-roi/
32. https://gpusmith.com/articles/en/air-gapped-llm-deployment-guide
33. https://knowledge.broadcom.com/external/article/451834/how-to-configure-metadata-file-for-vmwar.html
34. https://docs.nvidia.com/vgpu/latest/grid-vgpu-release-notes-vmware-vsphere/release-notes.html
35. https://www.finops.org/framework/capabilities/unit-economics/
36. https://docs.nvidia.com/datacenter/dcgm/latest/reference/dcgm-exporter-metrics.html
37. https://opentelemetry.io/docs/specs/semconv/
38. https://blogs.vmware.com/security/2025/06/announcing-vdefend-for-vcf-9
39. https://knowledge.broadcom.com/external/article/404155/licensing-vcenter-9x-and-esxi-9x-in-air-gapped-environments
40. https://blogs.vmware.com/cloud-foundation/2025/06/19/use-cases-for-vmware-private-ai-foundation-with-nvidia-in-vcf-9-0/
41. https://mlcommons.org/2026/09/chairs-mlperf-inference-v6-1/
42. https://docs.mlcommons.org/inference/submission/
43. https://docs.redhat.com/en/documentation/red_hat_openshift_ai_self-managed/3.5/html/release_notes/support-removals_relnotes
44. https://docs.redhat.com/en/documentation/red_hat_openshift_ai_self-managed/3.5/html/release_notes/new-features-and-enhancements_relnotes
45. https://www.nutanix.com/press-releases/2026/nutanix-gives-enterprises-the-freedom-to-run-production-agentic-ai-their-way
46. https://www.cisco.com/site/us/en/solutions/artificial-intelligence/infrastructure/ai-pods.html?ccid=cc006075
47. https://www.nutanix.dev/api_reference/apis/nai2.4.html
48. https://www.cisco.com/c/en/us/products/collateral/hyperconverged-infrastructure/compute-hyperconverged/ai-pods-with-nutanix-aag.html
49. https://docs.nvidia.com/enterprise-reference-architectures/nim-llm-runai-vanilla-kubernetes/latest/performance-and-scale-methodology.html

50. https://www.finops.org/framework/technology-categories/ai/
51. https://knowledge.broadcom.com/external/article/312168/vcf-offline-depot-deployment.html
52. https://contribute.cncf.io/projects/best-practices/security/supply-chain/supply-chain-best-practices/

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.