VMware Private AI Foundation with NVIDIA

Unlock AI and unleash productivity with lower TCO

At a glance

VMware Private AI Foundation with NVIDIA is a joint AI platform that will enable enterprises to run RAG workflows, fine-tune and customize LLM models, and run inference workloads in their data centers, addressing privacy, choice, cost, performance and compliance concerns.

Get private AI deployments

Broadcom and NVIDIA have collaborated to develop the joint AI platform called VMware Private AI Foundation with NVIDIA. This platform enables enterprises to fine-tune LLM models, deploy RAG workflows, and run inference workloads in their data centers, addressing privacy, choice, cost, performance, and compliance concerns. Built and run on the industry-leading private cloud platform, VMware Cloud Foundation, VMware Private AI Foundation with NVIDIA includes VCF Private AI Services, NVIDIA AI Enterprise, NVIDIA NIM™ inference microservices for the latest AI models -including NVIDIA Nemotron models and leading community models - and NVIDIA Blueprints. VMware Cloud Foundation (VCF), VMware's full-stack private cloud infrastructure solution, offers a secure, comprehensive, and scalable platform for building and operating AI workloads, providing organizations with agility, flexibility, and scalability to meet their evolving business needs. Note that NVIDIA AI Enterprise licenses need to be purchased separately from NVIDIA.

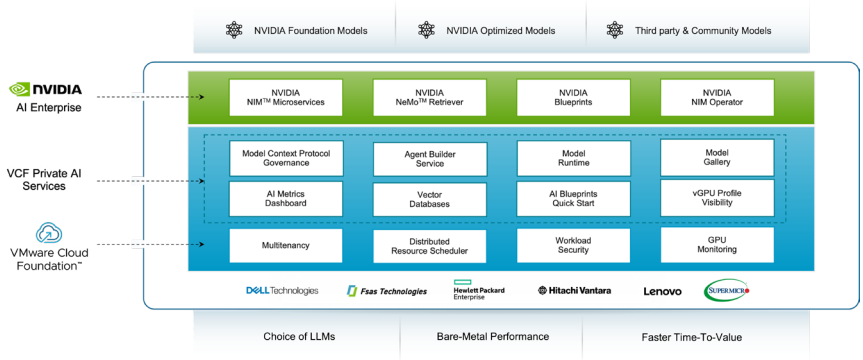

Figure 1: VMware Private AI Foundation with NVIDIA platform architecture.

Overcoming AI challenges- Privacy is the most important

Artificial Intelligence (AI) has become a cornerstone of digital transformation across industries. By enabling machines to learn from data and make decisions, AI helps organizations streamline operations, automate repetitive tasks, and enhance overall efficiency. Whether it's routing customer inquiries through chatbots or detecting fraud in financial transactions, AI delivers measurable business value by augmenting human capabilities. With such massive potential, it's no surprise that companies are eager to leverage this technology to boost productivity across every aspect of their organizations.

Benefits of VMware Private AI Foundation with NVIDIA

- Enable Privacy & Security of AI Models
- Simplify Infrastructure Management
- Streamline Model Deployment

However, AI has several challenges that enterprises must address for successful deployments

- **Privacy:** Privacy is the biggest consideration enterprises face in launching AI. Training models on open or cloud-based projects may unintentionally incorporate proprietary data into the language model's training data that may not comply with applicable privacy and/or intellectual privacy legal considerations. path to private cloud infrastructure. property, proprietary data and access control are very important.
- **Choice:** Enterprises want to choose LLMs that fit their use cases, industry vertical requirements, and retain their ability to shift to other LLMs as their needs evolve.
- **Cost:** AI models are complex and costly to architect since they rapidly evolve with new vendors, SaaS components, and bleeding edge AI software continuously launched and deployed. This can lead to rapidly escalating costs.
- **Performance:** Fine-tuning, customizing, deploying and querying LLMs can be intensive, and scaling up can cause performance issues without access to adequate resources.
- **Compliance:** Organizations in different industries and countries have different compliance and legal needs that enterprise solutions, including AI, must meet.

VMware Private AI Foundation with NVIDIA solves these challenges. Let's get into the details of this platform. Components of this platform

Components of this platform

Here are the key components that enable organizations to securely harness the power of AI.

- **VMware Cloud Foundation (VCF)** - VMware Cloud Foundation is the industry's first private cloud platform that delivers public cloud scale and agility, private cloud security, resilience and performance, and low overall total cost of ownership for your AI workloads. The versatility offered through this architecture enables cloud admins to utilize different workload domains, which can each be customized to support specific workload types, optimizing for workload performance and resource utilization, specifically GPUs.
- **VCF Private AI Services** - VCF Private AI Services provides powerful capabilities like Model Gallery, Model Runtime, Vector Databases, Deep Learning VMs, Data Indexing and Retrieval service, AI Agent Builder service and more to enable privacy and security, simplify infrastructure management and streamline model deployment.
- **NVIDIA AI Enterprise** - NVIDIA AI Enterprise is a secure, end-to-end, cloud native software platform that accelerates the data science pipeline and streamlines development and deployment of production-grade AI applications, including generative AI, computer vision, speech AI, and more. NVIDIA NIM allows enterprises to run inference on a range from LLMs from NVIDIA models to community models.

- **Major server OEM support** - Major server OEMs such as Dell, Lenovo, HPE, Supermicro, Hitachi Vantara and Fsas Technologies support this platform

NVIDIA AI Enterprise licenses will also need to be purchased separately.

Key Benefits and Capabilities

- **Enable Privacy & Security of AI Models:** VMware Private AI Foundation with NVIDIA's architectural approach for AI services enables privacy and control of corporate data and integrated security and management. Broadcom and NVIDIA's partnership can help enterprises build and deploy private and secure AI models with VCF Private AI Services and NVIDIA AI Enterprise.
  - Model Context Protocol (MCP) support with governance: Get a secure and standardized method to integrate AI assistants with internal content repositories and external MCP tools from Oracle, Microsoft SQL Server, ServiceNow, GitHub, Slack, PostgreSQL, and more—without building and maintaining custom connectors.
  - Model Gallery: With Model Gallery capability, ML Ops teams and data scientists can now curate and provide secure LLMs with integrated access control (RBAC). This can ensure governance and security for the environment and the privacy of enterprise data and IP.
  - Air-Gap Support: Through VCF Automation, VMware Private AI Foundation with NVIDIA can be deployed in air-gapped environments, supporting the business needs of customers, with data confidentiality and isolation for their critical workloads.
  - AI Software Security Patching through NVIDIA AI Enterprise: For NVIDIA AI Enterprise customers, NVIDIA commits to patch critical and high CVEs (common vulnerabilities and exposures) monthly for production branches and quarterly for long-term support branches while maintaining API compatibility up and down the stack.
- **Simplify Infrastructure Management:** AI models are complex and costly to architect since they are rapidly evolving with new vendors, SaaS components, and bleeding-edge AI software continuously launched and deployed. In this complex environment, VMware Private AI Foundation with NVIDIA comes with specially architected capabilities that help simplify infrastructure management of AI environments and optimize costs.
  - DirectPath Enablement for GPUs: This capability enables high-performance, exclusive GPU access to a single VM and the VM will be able to fully utilize GPU capabilities. With this new capability, enterprises can deploy AI projects on VMware Private AI Foundation with NVIDIA in DirectPath mode.
  - vGPU Profile Visibility: View all vGPUs across their GPU footprint via DirectPath Profiles through an easy-to-use UI screen in vCenter and eliminate the manual tracking of vGPUs, reducing admin time.
  - Vector Databases for Enabling RAG Workflows: Broadcom has enabled Vector databases by leveraging pgvector on PostgreSQL. This capability is managed through Data Services Manager and enables quick deployment of vector databases to support retrieval-augmented generation AI applications.
  - AI Blueprints Quick Start: With this capability, LOB Admins on Day 0 to

quickly design, curate, and offer infrastructure catalog objects through VCF's self-service portal (formerly VCF Automation), greatly simplifying the deployment of AI workloads.

- **Streamline Model Deployment:** Broadcom and NVIDIA have also enabled software and capabilities to easily deploy and manage AI workloads for data scientists and IT teams.
  - Model Runtime: The Model Runtime service enables data scientists to create and manage Model endpoints for their applications. This simplifies model usage and the scalability of LLMs.
  - Agent Builder Service: The Agent Builder Service allows for GenAI application developers to build AI Agents by using resources from the Model Store, Model Runtime, and Data Indexing and Retrieval Service.
  - AI Metrics Observability Dashboard: This dashboard provides enhanced visibility into model and GPU metrics, enabling data scientists and MLOps teams to identify bottlenecks, optimize resource allocation, and significantly improve throughput and performance.
  - CPU-only Workload Deployment: Reduce TCO through CPU-only AI workloads, through the integration of Model Runtime with Llama.cpp inferencing engine. This enables the deployment of less resource-intensive environments for testing, proof-of-concept initiatives, or AI applications with minimal or no GPU requirements.
  - NVIDIA NIM: NVIDIA NIM, part of NVIDIA AI Enterprise, is a set of easy-to-use microservices designed for secure, reliable deployment of high-performance AI model inferencing across the cloud, data center, and workstations. These prebuilt containers support a broad spectrum of AI models—from open-source community models to NVIDIA AI Foundation models, as well as custom AI models.

Capabilities Enabled through VCF: VMware Cloud Foundation (VCF) is the industry's first private cloud platform to deliver public cloud scale and agility with private cloud security, resilience, and performance. VCF offers a full-stack software-defined architecture designed to deliver a self-service unified platform and leverage an automated IT environment for the deployment and management of AI, non-AI workloads, and containers. This unified approach provides overall lower TCO and excellent performance. **The recently released benchmark** study compared against bare metal using MLPerf Inference v5.1 standards, show performance onpar with bare metal. This allows customers to benefit from the increased Agility, Availability, and Flexibility that VCF provides while leveraging excellent performance. Let's look at some of the key capabilities in the platform through VCF.

Workload security: VCF has several built-in capabilities to improve workload security. These include Secure Boot, Virtual TPM, vSphere Trust Authority, VM Encryption, and more

- Identity and Access Management: VCF integrates with various identity and access management solutions, including VMware Identity Manager and third-party identity providers. This ensures that only authorized users and applications can access AI models and data sets.
- Network security: VCF helps protect applications with micro-segmentation,

full-stack networking & security, and advanced threat prevention at the network level via software-dedicated firewalls for applications and their associated AI models and data sets.

- Distributed Resource Scheduler (DRS): This industry-leading capability helps achieve excellent cost optimization and workload performance by optimally placing workloads on hosts. ESXi hosts are grouped into resource clusters to segregate the computing needs of different business units.
- Multi-tenancy: With VCF 9.0's multi-tenancy capability, cloud service providers and enterprise administrators can enable secure and private environments for tenants on the same infrastructure and achieve high efficiency, scalability, and lower TCO.
- GPU and vGPU Monitoring Improvements: VCF offers powerful GPU monitoring capabilities at the host, cluster, and VM levels, giving administrators deep visibility into GPU utilization. These capabilities help identify GPU over-provisioning or under-utilization, optimize total cost of ownership (TCO), accelerate issue resolution, and enhance overall performance.

Unlock the power of AI

VMware Private AI Foundation with NVIDIA can help bring new levels of productivity to every department of organizations while maintaining the privacy and control of corporate data and IP.

Ready to go on your AI/ML journey? Complete this form to contact us!

To learn more, visit www.vmware.com/aiml-nvidia

vmware by Broadcom

Item No: VPAIF Datasheet 6/25