Implement AI on VCF with 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.

Built and run on the industry-leading private cloud platform, VMware Cloud Foundation (VCF), VMware Private AI Foundation with NVIDIA includes the 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. NVIDIA AI Enterprise licenses need to be purchased separately from NVIDIA.

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. Now, the evolution of AI has taken a leap forward with Generative AI (Gen AI). Generative AI will transform businesses in much the same way that the PC, the internet, and the smartphone did.

AI momentum is accelerating at an unprecedented scale. AI solutions and services will generate global impact of $22.3 Trillion by 2030[1]. 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.

However, there are several challenges that must be confronted before widespread deployment of AI and Gen AI in organizations.

Privacy is the key challenge of AI

The latest wave of AI innovation is being driven by Gen AI. While the potential of Gen AI is virtually limitless, their open design presents inherent privacy risks, making privacy the biggest challenge. Enterprise data and intellectual property (IP) needs to be protected to prevent leakage outside the organizational boundary. Infrastructure and data access must be tightly controlled.

Further challenges presented by AI

In addition to privacy, there are other challenges organizations need to consider.

- **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.

Performance – Fine-tuning, customizing, deploying and querying LLMs can be intensive, and scaling up can be challenging without access to adequate resources. Efficient allocation of GPU resources is critical to ensure low

1. IDC Blog, "IDC Predicts AI Solutions & Services will Generate Global Impact of $22.3 Trillion by 2030" (01 April 2025).

Benefits of VMware Private AI Foundation with NVIDIA

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

- **Performance** – Fine-tuning, customizing, deploying and querying LLMs can be intensive, and scaling up can be challenging without access to adequate resources. Efficient allocation of GPU resources is critical to ensure low latency.
- **Compliance** – Organizations in different industries and countries have different compliance and legal needs that enterprise solutions, including AI, must meet. Access control, workload placement, and audit readiness are vital when deploying AI and Gen AI models.
- **Infrastructure** – The deployment and scaling of AI infrastructure encounter several critical infrastructure-specific issues that can hinder adoption of large language models based on their specific AI use cases. Without addressing these challenges, IT architects will find it very difficult to deploy, configure and reconfigure compute, storage and networking infrastructure to support the needs of AI workloads as dictated by the business.

The solution: VMware Private AI Foundation with NVIDIA

To address the challenges, Broadcom and NVIDIA have collaborated to develop a joint AI platform called VMware Private AI Foundation with NVIDIA. This joint AI platform enables enterprises to fine-tune LLM models, deploy retrieval augmented generation (RAG) workflows, and run inference workloads in their data centers, addressing privacy, choice, cost, performance and compliance concerns. VMware Private AI Foundation with NVIDIA simplifies AI deployments for enterprises by offering an intuitive automation tool, deep learning VM images, vector database, and GPU monitoring capabilities.

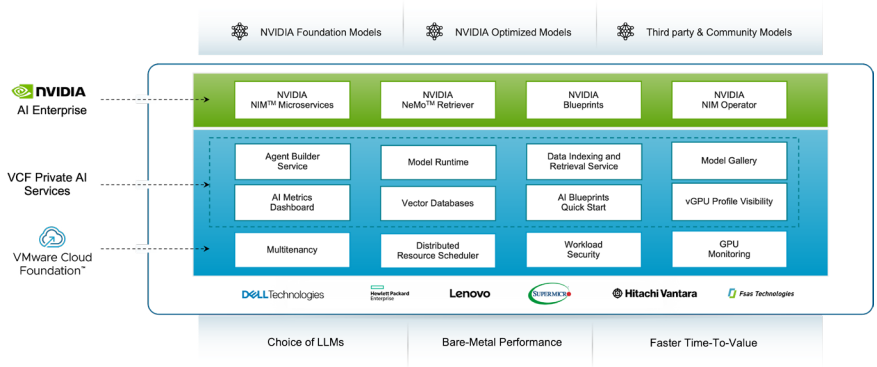

Figure 1: The VMware Private AI Foundation with NVIDIA platform architecture.

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.

Key Benefits of VCF

- Compute, Storage and GPU Virtualization
- Networking
- Management and Operations

- **VCF Private AI Services** – VCF Private AI Services provide 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 need to be purchased separately.

VCF

Let's get into the details of the value that VCF provides:

Compute and Storage Virtualization – The integration between VCF and NVIDIA AI Enterprise and the ability to virtualize GPUs (vGPU technology) enables maximize performance and utilization of physical GPU resources across multiple users. The platform also enables highly scalable and performance-optimized storage, including the ability to disaggregate storage and compute resources so that they can be allocated independently to optimize capacity, utilization and performance.

Networking – One of the major strengths of leveraging VMware Cloud Foundation is that networking is a fundamental construct of the design and implementation, offering connectivity to the closest storage and compute node, especially when routing at the edge. VCF Networking, delivered through NSX, provides a software-defined networking solution that supports the high-bandwidth and low-latency networking requirements of AI workloads.

Management and Operations – Beyond the individual capabilities at each layer of the full stack architecture, one of the key considerations in infrastructure operations is to optimize the operational aspects of deploying AI solutions at an enterprise scale. AI workloads demand quick actions, quick troubleshooting, and the ability to bring up and tear down environments with a few clicks. The integrated VCF Operations offers management tools for automating the deployment, monitoring, and operation of AI workloads within the VCF environment. This includes capabilities for performance monitoring, and operation of AI workloads within the VCF environment. This includes capabilities for performance monitoring, capacity management, and compliance, which help streamline the management of AI infrastructure and reduce operational overhead.

Capabilities enabled through VCF

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 VMware.com/AIML-NVIDIA

vmware by Broadcom | NVIDIA

Item No: O-11347 -9.1 VCF heavy VPAIF Solns brief 4/26