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 Private AI Package, NVIDIA AI Enterprise, NVIDIA NIM inference microservices for the latest AI models — including NVIDIA Nemotron models and leading community models— and NVIDIA Blueprints. This platform is an advanced service on top of VMware Cloud Foundation. 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.

Gen AI momentum is growing fast. Generative AI is expected to drive up to $4.4 Trillion[2] in annual economic value for enterprises. " By 2028, 95% of organizations will have integrated generative AI into daily operations—up from just 15% in 2025." According to Gartner®, Enterprise application software revenue is also projected to reach $536 billion by 2033*. 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 large language models (LLMs) that process massive data sets. While the potential of LLMs is virtually limitless, their open design presents inherent privacy risks, making privacy the biggest challenge. Enterprise data and intellectual property (IP) is private to the enterprise and critically valuable when training LLMs to serve organizations' specific needs. This data 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.

1. *Danielle Casey's presentation, Where Generative AI Works, Will Soon Work and Might Never Work, Gartner Tech Growth & Innovation Conference 2025. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.
2. McKinsey Report: The economic potential of Generative AI: The next productivity frontier

vmware by Broadcom | NVIDIA

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 GenAI 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 GenAI 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 GenAI deployments for enterprises by offering an intuitive automation tool, deep learning VM images, vector database, and GPU monitoring capabilities

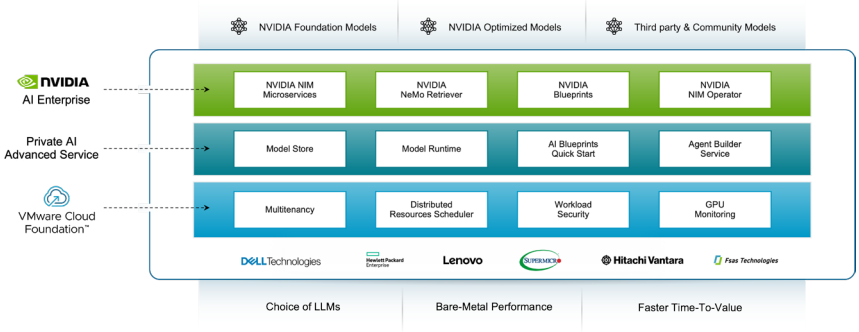

Figure 1: The VMware Private AI Foundation with NVIDIA platform architecture.

Components of this platform

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

95%

of tech companies are integrating AI features into new apps.

(Source: VMware FY24 Q2 Executive Pulse, N=450 Enterprise Technology Executives)

- Private AI Package – The Private AI Package provides powerful capabilities like 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.

This platform is an advanced service on top of VMware Cloud Foundation. NVIDIA AI Enterprise licenses will need to be purchased separately.

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: 9.0 UPDATE VPAIF Solutions brief 6/25