VMware Private AI Foundation with NVIDIA

Unlock GenAI and unleash productivity

At a glance

VMware Private AI Foundation with NVIDIA is a joint GenAI 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.

Initially available in early 2024, the platform comprises NVIDIA AI Enterprise, which includes NVIDIA NIM microservices, NVIDIA LLMs, and access to other community models (such as Hugging face models) running on VMware Cloud Foundation™. This platform is an add-on SKU on top of VMware Cloud Foundation. NVIDIA AI Enterprise licenses will need to be purchased separately.

Generative AI will transform business in much the same way that the PC, the internet, and the smartphone did. With a potential $4.4 trillion annual economic value for enterprises,[1] 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 bringing generative AI into organizations.

Privacy is the key challenge of generative 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 generative 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** – Generative AI can be costly to architect because it is a rapidly evolving technology. New vendors enter the market constantly, and new software is 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 latency.

1. McKinsey & Company. "The economic potential of generative AI: The next productivity frontier." Michael Chui, Eric Hazan, Roger Roberts, Alex Singla, Kate Smaje, Alex Sukharevsky, Lareina Yee, and Rodney Zemmel, June 2023.

Benefits of VMware Private AI Foundation with NVIDIA

- Enable privacy, security and compliance of AI models
- Get accelerated performance for generative AI models regardless of the chosen LLMs
- Simplify generative AI deployment and optimize costs

- **Compliance** – Organizations in different industries and countries have different compliance and legal needs that enterprise solutions, including generative AI, must meet. Generative AI access control, workload placement, and audit readiness are vital when deploying generative 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 this, Broadcom and NVIDIA have collaborated to develop a joint generative AI platform called VMware Private AI Foundation with NVIDIA. This joint GenAI 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.

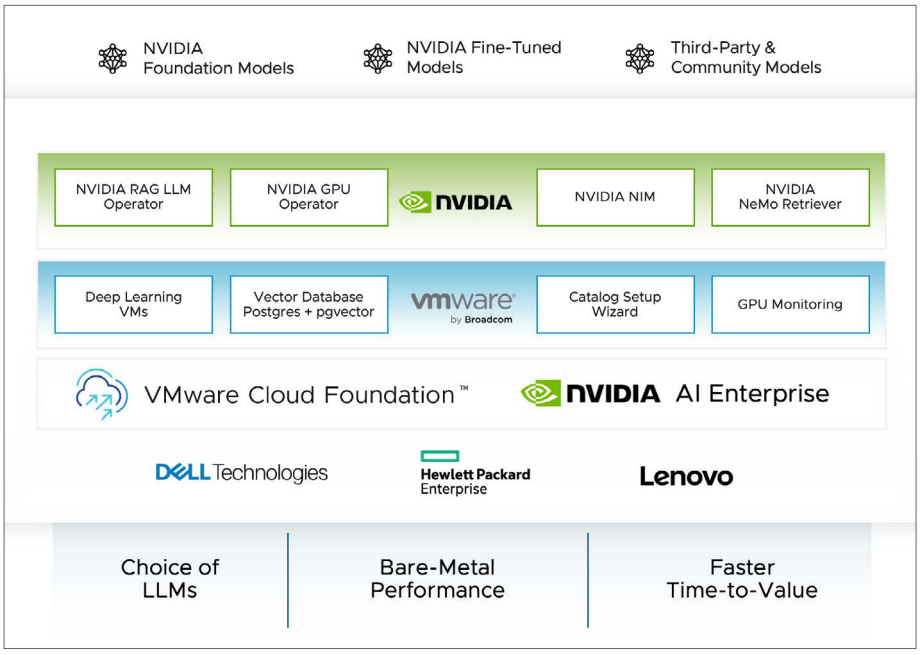

Figure 1: The VMware Private AI Foundation with NVIDIA platform architecture.

95%

of tech companies are integrating AI features into new apps.

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

Components of this platform

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

- **VMware Cloud Foundation** – VMware Cloud Foundation offers a full-stack scalable, software-defined architecture designed to deliver a self-service unified platform and leverage an automated IT environment that simplifies the deployment and management of all workloads utilizing VMs, containers and AI technologies. 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.
- **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 and HPE support this platform.

This platform is an add-on SKU on top of VMware Cloud Foundation. NVIDIA AI Enterprise licenses will also need to be purchased separately.

Unlock the power of generative 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 request access!

To learn more, visit vmware.com/AIML.

---

* Submitting this form does not guarantee access to software. Availability is subject to approval and may be limited.

vmware by Broadcom | NVIDIA

Item No: FY24-7496-BCM-PRIVATE-AI-NVIDIA-SO-USLET-WEB-20240320 3/24