VMware and NVIDIA Unlock Generative AI for Enterprises

New VMware Private AI Foundation With NVIDIA Enables Enterprises to Ready Their Businesses for Generative AI; Platform to Further Support Data Privacy, Security and Control

VMware Explore—VMware Inc. (NYSE: VMW) and NVIDIA (NASDAQ: NVDA) today announced the expansion of their strategic partnership to ready the hundreds of thousands of enterprises that run on VMware's cloud infrastructure for the era of generative AI.

VMware Private AI Foundation with NVIDIA will enable enterprises to customize models and run generative AI applications, including intelligent chatbots, assistants, search and summarization. The platform will be a fully integrated solution featuring generative AI software and accelerated computing from NVIDIA, built on VMware Cloud Foundation and optimized for AI.

"Generative AI and multi-cloud are the perfect match," said Raghu Raghuram, CEO, VMware. "Customer data is everywhere — in their data centers, at the edge, and in their clouds. Together with NVIDIA, we'll empower enterprises to run their generative AI workloads adjacent to their data with confidence while addressing their corporate data privacy, security and control concerns."

"Enterprises everywhere are racing to integrate generative AI into their businesses," said Jensen Huang, founder and CEO, NVIDIA. "Our expanded collaboration with VMware will offer hundreds of thousands of customers — across financial services, healthcare, manufacturing and more — the full-stack software and computing they need to unlock the potential of generative AI using custom applications built with their own data."

Full-Stack Computing to Supercharge Generative AI
To achieve business benefits faster, enterprises are seeking to streamline development, testing and deployment of generative AI applications. McKinsey estimates that generative AI could add up to $4.4 trillion annually to the global economy.(1)

VMware Private AI Foundation with NVIDIA will enable enterprises to harness this capability, customizing large language models; producing more secure and private models for their internal usage; and offering generative AI as a service to their users; and, more securely running inference workloads at scale.

The platform is expected to include integrated AI tools to empower enterprises to run proven models trained on their private data in a cost-efficient manner. To be built on VMware Cloud Foundation and NVIDIA AI Enterprise software, the platform's expected benefits will include:

- Privacy — Will enable customers to easily run AI services adjacent to wherever they have data with an architecture that preserves data privacy and enable secure access.
- Choice — Enterprises will have a wide choice in where to build and run their models — from NVIDIA NeMo™ to Llama 2 and beyond — including leading OEM hardware configurations and, in the future, on public cloud and service provider offerings.
- Performance — Running on NVIDIA accelerated infrastructure will deliver performance equal to and even exceeding bare metal in some use cases, as proven in recent industry benchmarks.
- Data-Center Scale — GPU scaling optimizations in virtualized environments will enable AI workloads to scale across up to 16 vGPUs/GPUs in a single virtual machine and across multiple nodes to speed generative AI model fine-tuning and deployment.
- Lower Cost — Will maximize usage of all compute resources across, GPUs, DPUs and CPUs to lower overall costs, and create a pooled resource environment that can be shared efficiently across teams.
- Accelerated Storage — VMware vSAN Express Storage Architecture will provide performance-optimized NVMe storage and supports GPUDirect® storage over RDMA, allowing for direct I/O transfer from storage to GPUs without CPU involvement.
- Accelerated Networking — Deep integration between vSphere and NVIDIA NVSwitch™ technology will further enable multi-GPU models to execute without inter-GPU bottlenecks.
- Rapid Deployment and Time to Value — vSphere Deep Learning VM images and image repository will enable fast prototyping capabilities by offering a stable turnkey solution image that includes frameworks and performance-optimized libraries pre-installed.

The platform will feature NVIDIA NeMo, an end-to-end, cloud-native framework included in NVIDIA AI Enterprise — the operating system of the NVIDIA AI platform — that allows enterprises to build, customize and deploy generative AI models virtually anywhere. NeMo combines customization frameworks, guardrail toolkits, data curation tools and pretrained models to offer enterprises an easy, cost-effective and fast way to adopt generative AI.

For deploying generative AI in production, NeMo uses TensorRT for Large Language Models (TRT-LLM), which accelerates

and optimizes inference performance on the latest LLMs on NVIDIA GPUs. With NeMo, VMware Private AI Foundation with NVIDIA will enable enterprises to pull in their own data to build and run custom generative AI models on VMware's hybrid cloud infrastructure.

At VMware Explore 2023, NVIDIA and VMware will highlight how developers within enterprises can use the new NVIDIA AI Workbench to pull community models, like Llama 2, available on Hugging Face, customize them remotely and deploy production-grade generative AI in VMware environments.

Broad Ecosystem Support for VMware Private AI Foundation With NVIDIA
VMware Private AI Foundation with NVIDIA will be supported by Dell Technologies, Hewlett Packard Enterprise and Lenovo — which will be among the first to offer systems that supercharge enterprise LLM customization and inference workloads with NVIDIA L40S GPUs, NVIDIA BlueField®-3 DPUs and NVIDIA ConnectX®-7 SmartNICs.

The NVIDIA L40S GPU enables up to 1.2x more generative AI inference performance and up to 1.7x more training performance compared with the NVIDIA A100 Tensor Core GPU.

NVIDIA BlueField-3 DPUs accelerate, offload and isolate the tremendous compute load of virtualization, networking, storage, security and other cloud-native AI services from the GPU or CPU.

NVIDIA ConnectX-7 SmartNICs deliver smart, accelerated networking for data center infrastructure to boost some of the world's most demanding AI workloads.

VMware Private AI Foundation with NVIDIA builds on the companies' decade-long partnership. Their co-engineering work optimized VMware's cloud infrastructure to run NVIDIA AI Enterprise with performance comparable to bare metal. Mutual customers further benefit from the resource and infrastructure management and flexibility enabled by VMware Cloud Foundation.

Availability
VMware intends to release VMware Private AI Foundation with NVIDIA in early 2024.

About NVIDIA
Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com/.

About VMware
VMware is a leading provider of multi-cloud services for all apps, enabling digital innovation with enterprise control. As a trusted foundation to accelerate innovation, VMware software gives businesses the flexibility and choice they need to build the future. Headquartered in Palo Alto, California, VMware is committed to building a better future through the company's 2030 Agenda. For more information, please visit www.vmware.com/company.

1. "The economic potential of generative AI: The next productivity frontier," McKinsey, 2023

Certain statements in this press release including, but not limited to, statements as to: the benefits, impact, performance, features and availability of our products and technologies, including NVIDIA AI Enterprise, NVIDIA NeMo, TensorRT, NVIDIA L40S GPUs, NVIDIA BlueField-3 DPUs, and NVIDIA ConnectX-7 SmartNICs; NVIDIA's partnership with VMware, including the benefits, impact, features, and availability of the VMware Private AI Foundation with NVIDIA platform; enterprises everywhere racing to integrate generative AI into their businesses; NVIDIA's expanded collaboration with VMware offering hundreds of thousands of customers — across financial services, healthcare, manufacturing and more — the full-stack software and computing they need to unlock the potential of generative AI using custom applications built with their own data; estimates that generative AI could add up to $4.4 trillion annually to the global economy; and broad ecosystem support for VMware Private AI Foundation with NVIDIA and third parties supporting the platform are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the most recent reports NVIDIA files with the Securities and Exchange Commission, or SEC, including, but not limited to, its annual report on Form 10-K and quarterly reports on Form 10-Q. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

Many of the products and features described herein remain in various stages and will be offered on a when-and-if-available

basis. The statements above are not intended to be, and should not be interpreted as a commitment, promise, or legal obligation, and the development, release, and timing of any features or functionalities described for our products is subject to change and remains at the sole discretion of NVIDIA. NVIDIA will have no liability for failure to deliver or delay in the delivery of any of the products, features or functions set forth herein.

Shannon McPhee
NVIDIA Corporation
+1-310-920-9642
smcphee@nvidia.com
Eloy Ontiveros
VMware
+1-650-427-6145
eontiveros@vmware.com