Democratizing Generative AI With VMware Private AI Foundation with NVIDIA

EXPT63094

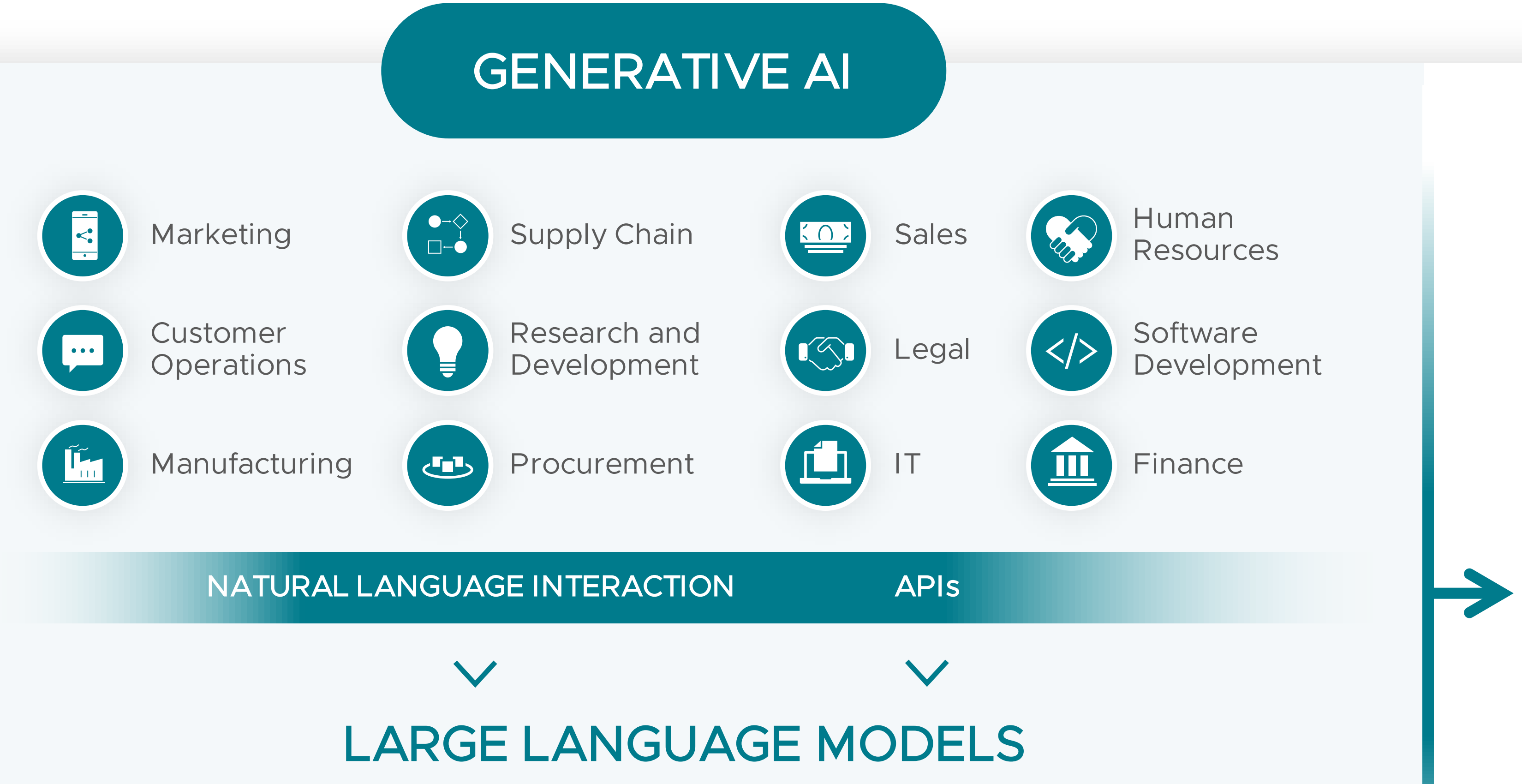
Shobhit Bhutani
VMware Private AI Product Marketing Lead

Shobhit Bhutani@Linkedin

vmware by Broadcom

NVIDIA

Impact of Generative AI in the Enterprise



GENERATIVE AI

Marketing

Supply Chain

Sales

Human Resources

Customer Operations

Research and Development

Legal

Software Development

Manufacturing

Procurement

IT

Finance

NATURAL LANGUAGE INTERACTION

APIs

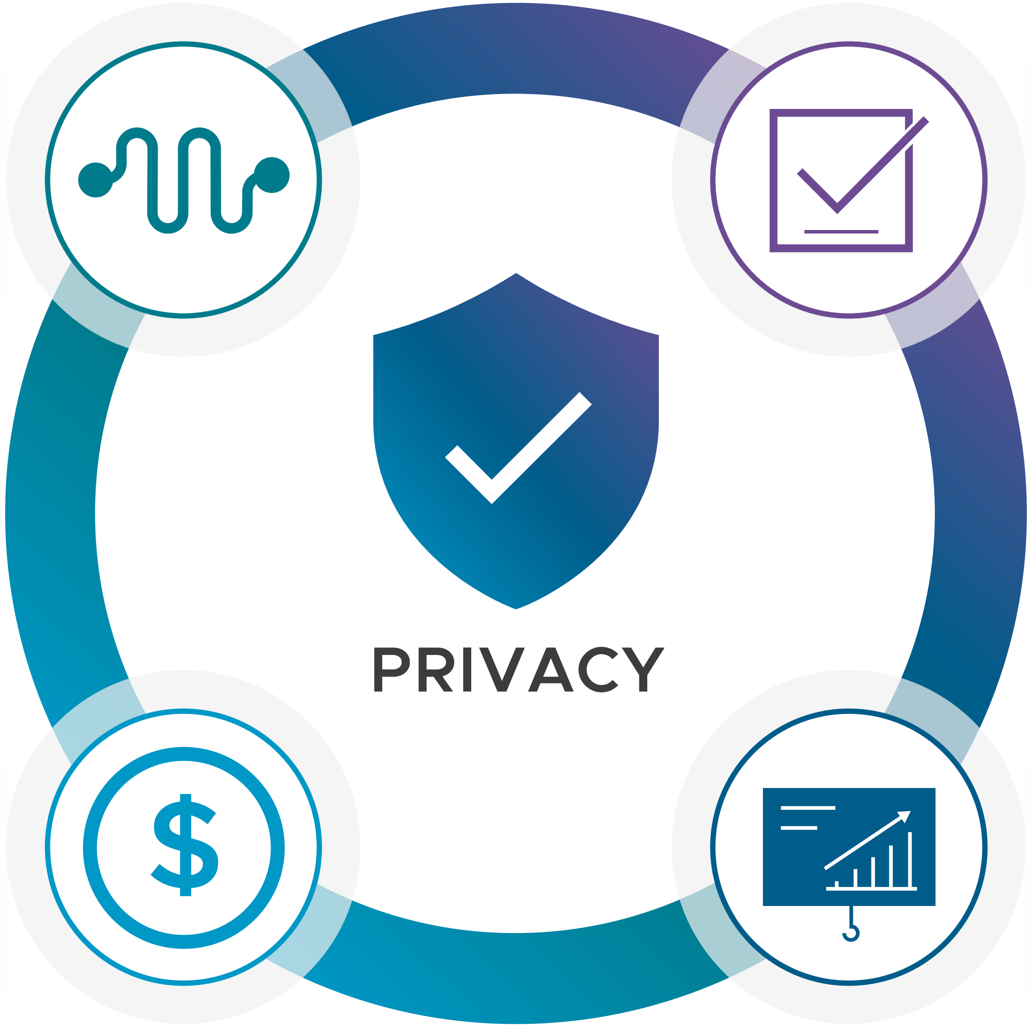
LARGE LANGUAGE MODELS

$4.4T

annual economic value

Source: McKinsey, The economic potential of generative AI: The next productivity frontier, June 2023

Key Enterprise Challenges



INTRODUCING

VMware Private AI Foundation with NVIDIA

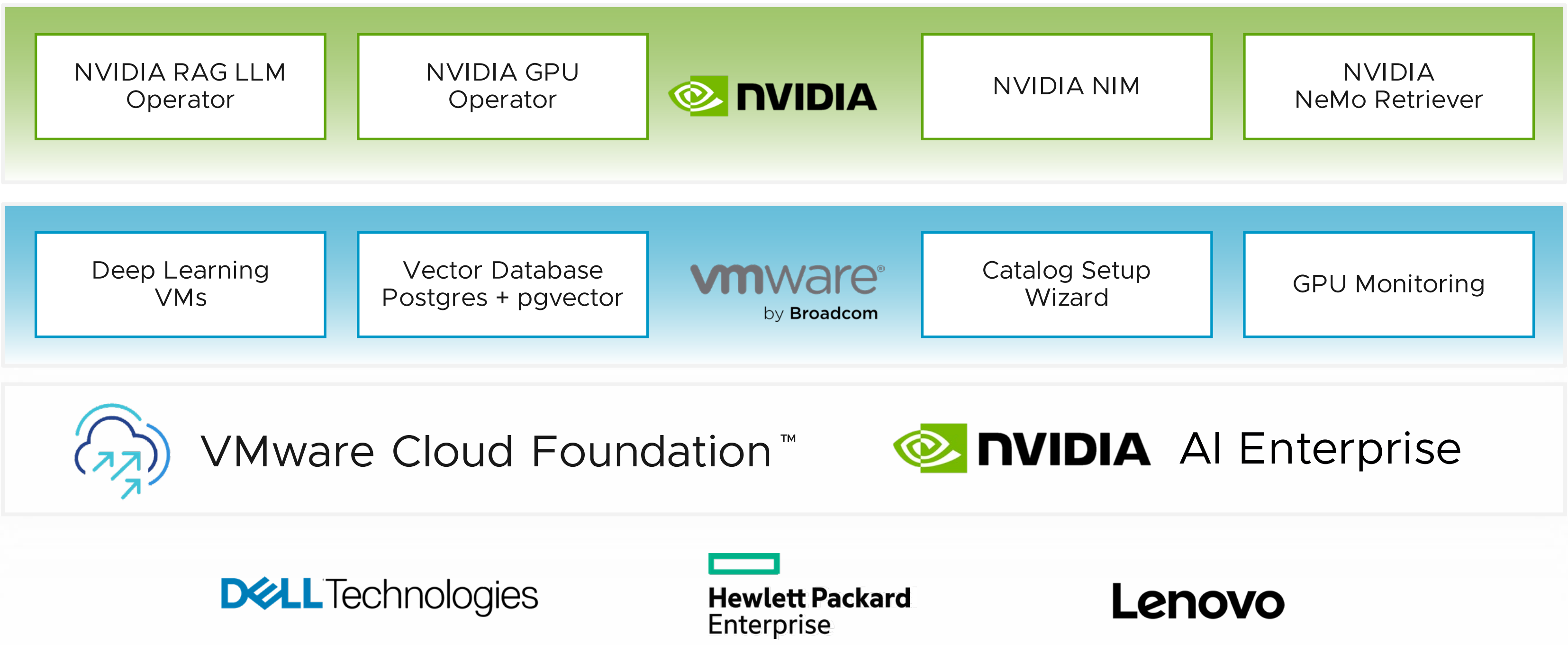
Joint Gen AI Platform:

Initially Available!

NVIDIA Foundation Models

NVIDIA Fine-Tuned Models

Third party & Community Models



NVIDIA RAG LLM Operator

NVIDIA GPU Operator

NVIDIA

NVIDIA NIM

NVIDIA NeMo Retriever

Deep Learning VMs

Vector Database Postgres + pgvector

vmware by Broadcom

Catalog Setup Wizard

GPU Monitoring

VMware Cloud Foundation™

NVIDIA AI Enterprise

DELL Technologies

Hewlett Packard Enterprise

Lenovo

Choice of LLMs

Bare-Metal performance

Faster Time-To-Value

Leveraging VCF as a Platform for GenAI Workloads

Solving for Common GenAI Workloads with VMware Cloud Foundation

Scale Quickly

Scale compute and storage in tandem with your GenAI needs, and provide low-latency, optimized networking

Management and Operations

Quick actions, quick troubleshooting and ability to bring-up and tear-down environments with a few clicks

Cost and Capacity Planning

GenAI workloads can be expensive to host and resource demand can bloat quickly. Stay in control and predict growth

Custom Dashboards

Track what's important whether it's capacity management, bottlenecks or other advanced KPIs

VMware Cloud Foundation™

vmware® by Broadcom

NVIDIA

Designed for Enterprises that Run their Business on AI

NVIDIA AI Enterprise: Production-Grade Software for AI

Accelerated Computing
increases productivity while lowering TCO

Generative AI

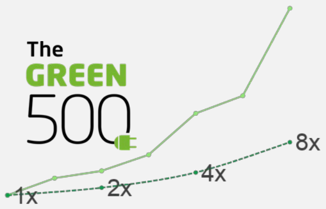
ETL/Spark

Inference

#1

MLPerf

The GREEN 500



Enterprise-Grade
security, stability, manageability & support

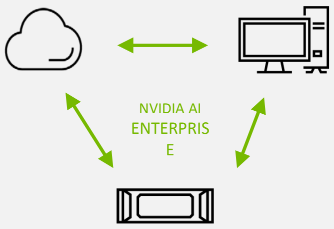
CVE Patching

API Stability

End-to-End Manageability

SLAs with NVIDIA Support

Cloud Native & Certified
to run everywhere



aws

Microsoft Azure

Google Cloud

ORACLE CLOUD

DELL

Hewlett Packard Enterprise

Lenovo

...

VMware Private AI Foundation with **NVIDIA**

Unlock Gen AI
& Unleash Productivity

Enable Privacy, Security & Compliance for AI Models

Simplify GenAI Deployment and Optimize Costs

Accelerate Performance regardless of the selected LLM

vmware by Broadcom

NVIDIA

Enable Privacy, Security & Compliance for Your AI Models

Build and deploy with an integrated GenAI platform with multi-layered built-in security and management

VCF Delivers Enhanced Security

Secure Boot, Virtual TPM, vSphere Trust Authority, advanced threat protection, network micro segmentation and more.

Enterprise-Grade Security with NVIDIA AI Enterprise

Continuous monitoring and regular releases of security patches for critical and common vulnerabilities and exposures (CVEs).

VCF Enables Modern Identity and Access Management for infrastructure

Integrations include: VMware Identity Manager
Third-party identity providers–Okta and Microsoft Entra ID.

Simplify GenAI Deployments & Optimize Costs

Get architected capabilities to simplify & expedite deployments

NVIDIA RAG

Leverage built-in RAG workflow with NVIDIA RAG LLM Operator

Vector Databases for RAG workflows

Enable fast querying of data and real-time updates using Vector Databases built using pgvector on PostgreSQL

Deep Learning VM Templates

Get Deep learning VMs with pre-configured software, NVIDIA NGC and software libraries

Catalog Setup Wizard & AI Deployment Guide

Expedite deployment with comprehensive deployment guide and enable the rapid creation, customization, and availability of intricate GenAI catalog items.

vmware by Broadcom

NVIDIA

Accelerate Performance regardless of the selected LLM

Integrated monitoring & flexibility

GPU Monitoring

Get GPU related resource optimization across clusters and hosts

Elevate Performance with NVIDIA NIM

NVIDIA NIM optimizes models for the highest performance

Choice of LLMs

Get the choice of LLMs- NVIDIA Foundation Models, Community, Third party and NVIDIA fine-tuned models,

VMware Private AI Foundation with NVIDIA

Unlock GenAI and unleash productivity

VALUE PROPOSITION

VMware Private AI Foundation with NVIDIA is a joint GenAI platform by Broadcom and NVIDIA, which helps enterprises unlock GenAI & unleash productivity. With this platform, enterprises can deploy AI workloads faster, fine-tune and customize LLM models, deploy retrieval augmented generation (RAG) workflows, and run inference workloads in their data centers, addressing privacy, choice, cost, performance, and compliance concerns.

VALUE

- **Enable** Privacy, Security & Compliance for AI Models
- **Simplify** GenAI Deployment and Optimize Costs
- **Accelerate** performance regardless of the selected LLM

ARCHITECTURE

VMware Cloud Foundation™

NVIDIA AI Enterprise

vmware by Broadcom

NVIDIA

Learn More

VMware Private AI Foundation with NVIDIA Resources

- VMware AI and ML page
- VMware Private AI Foundation with NVIDIA Launch Blog
- Initial Availability Access Request Form

Thank You