AI-READY ENTERPRISE PLATFORM

Unleashing AI for Every Enterprise with an End-to-End Platform from NVIDIA and VMware

Artificial intelligence (AI) is transforming every industry, whether it's by improving customer relationships in financial services, streamlining manufacturer supply chains, or helping doctors deliver better outcomes for patients. While most organizations know they need to invest in AI to secure their future, they struggle with finding the strategy and platform that can enable success.

Unlike traditional enterprise applications, AI apps are a relatively recent development for many IT departments. They're anchored in rapidly evolving, open-source, bleeding-edge code and lack proven approaches that meet the rigors of scaled production settings in enterprises. In fact, just 53 percent of projects make it from pilot to production, with the complexity of integrating AI solutions with existing infrastructure among the top three barriers to AI implementation.[1]

AI-Ready Platform from NVIDIA and VMware

VMware and NVIDIA have partnered together to unlock the power of AI for every business by delivering an end-to-end enterprise platform optimized for AI workloads. This integrated platform delivers best-in-class AI software, the NVIDIA AI Enterprise Suite, optimized and exclusively certified for the industry's leading virtualization platform, VMware vSphere®.

KEY CHALLENGES FOR ENTERPRISE AI WORKLOADS

> **Risk** – It's difficult to pull together an end-to-end AI solution from disparate products and integrate them with existing infrastructure.

> **Performance** – High performance is critical for AI, machine learning, and data analytics workloads. Fast time to deployment is also critical.

> **Scaling** – Going from proof of concept to enterprise deployments requires effective scaling—through efficient use of resources to ensure manageability and availability—and infrastructure cost management.
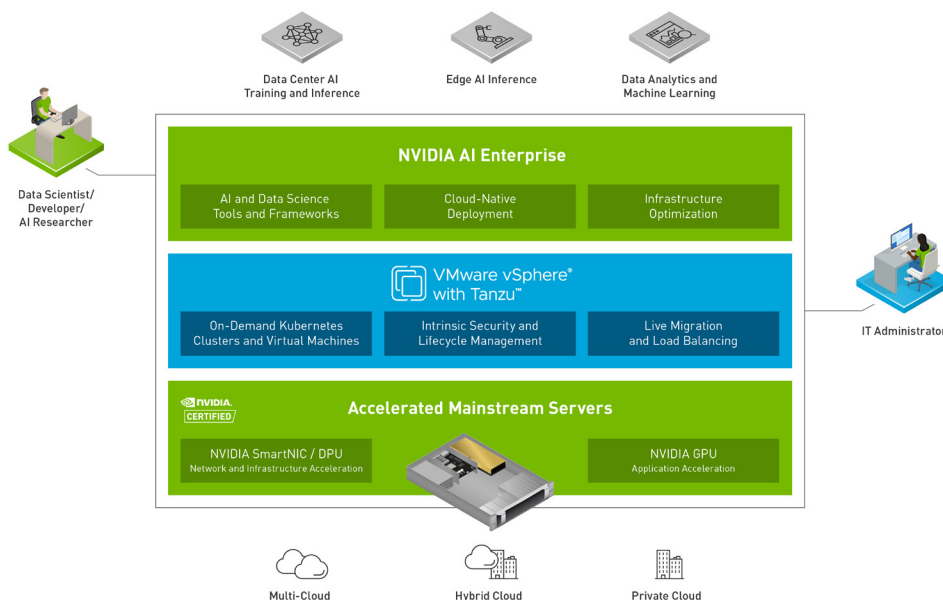

Figure 1. NVIDIA and VMware AI-Ready Platform: End users can access the software they need to build successful AI projects, and IT admins can support the projects using familiar tools.

Running on NVIDIA-Certified Systems™, industry-leading accelerated servers, this platform accelerates the speed at which developers can build AI and high-performance data analytics, enables organizations to scale modern workloads on the same VMware vSphere infrastructure they've already invested in, and delivers enterprise-class manageability, security and availability.

Key Enabling Technologies

The solution is comprised of the following building blocks.

NVIDIA AI Enterprise Suite

NVIDIA AI Enterprise is an end-to-end, cloud native suite of AI and data science applications and frameworks optimized and exclusively certified by NVIDIA to run on VMware vSphere with NVIDIA-Certified Systems. It includes key enabling technologies and software from NVIDIA for rapid deployment, management and scaling of AI workloads in the modern hybrid cloud. NVIDIA AI Enterprise is licensed and supported by NVIDIA.
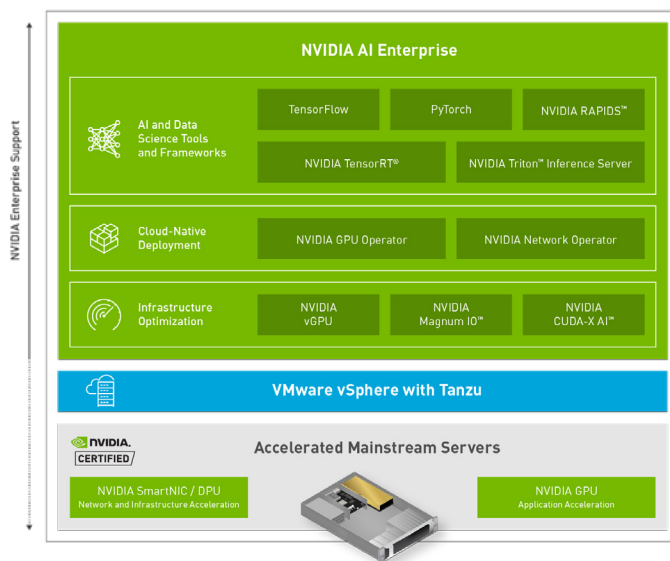

Figure 2. The NVIDIA AI Enterprise software suite includes the applications, frameworks, and tools used by AI researchers, data scientists, and developers, as well as tools for cloud-native deployments and infrastructure optimizations.

VMware vSphere with Tanzu

VMware vSphere® with VMware Tanzu™ is the industry's leading virtualization platform that runs over 70 million workloads across hundreds of thousands of organizations worldwide. VMware vSphere transforms bare-metal servers (including CPU- and GPU-based resources) into centrally managed AI and machine learning infrastructure pools that can quickly provision virtual machines and containers on demand.

VMware vSphere optimizes performance, increases availability, tightens security, streamlines maintenance, and reduces costs to create an agile, efficient, resilient and intrinsically secure infrastructure platform that supports existing workloads and next-gen applications such as AI.

VMware vSphere has been enhanced to optimally run the NVIDIA AI Enterprise software suite leveraging NVIDIA's latest NVIDIA Ampere architecture-based GPUs.

BENEFITS FOR ENTERPRISES

> **Deploy with confidence** – Leverage the jointly certified NVIDIA and VMware solution to deploy and manage AI workloads with optimized software.

> **Boost AI performance** – Achieve near bare-metal performance for AI training and inference.

> **Scale without compromise** – Enable large, more complex AI training and data analytics with new scale-out capabilities.
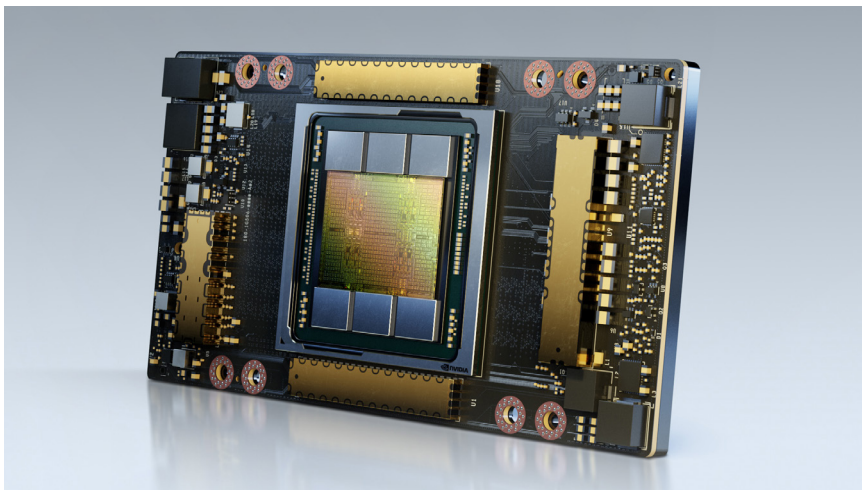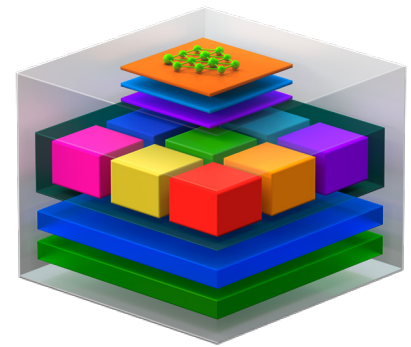
BENEFITS FOR ENTERPRISES

> Avoid AI silos and simplify management by leveraging virtualization to fold AI deployments into existing enterprise infrastructure.

> Get near bare-metal performance for AI and machine learning training and inference, with the added benefits of enterprise-grade security, manageability, and efficiency.

> Enable high availability for AI workloads while simplifying infrastructure maintenance such as consolidation, expansion and upgrades.

> Simplify management at scale across AI infrastructure with automatic initial placement of workloads for optimal resource consumption and deterministic costs.

NVIDIA-Certified Systems

The NVIDIA AI Enterprise Suite is certified to run on NVIDIA-Certified Systems, which include the following:

> NVIDIA Ampere architecture-based GPUs such as the NVIDIA A100 Tensor Core GPU. The Tensor Core technology included in the Ampere architecture has brought dramatic speedups to AI operations, bringing down training times from weeks to hours and providing massive acceleration to inference.

> NVIDIA® ConnectX® SmartNICs and the NVIDIA BlueField® data processing unit (DPU) provide a host of software-defined hardware engines for accelerating networking and security. These enable the best of both worlds: best-in-class AI training and inference performance, with all the necessary levels of enterprise data privacy, integrity and reliability.

The NVIDIA A100 Tensor Core GPU delivers unprecedented acceleration at every scale.

Contact an Account Manager for more information.

Business Solutions	Enterprise Solutions	Public Sector Solutions
1.800.800.0014	**1.800.369.1047**	**1.800.800.0019**

www.connection.com/nvidia

C1619321 0122