Solution Brief: **February 2022**

AI Ready Infrastructure with NVIDIA and VMware Cloud Foundation™

An Overview of VMware Cloud Foundation integration with the NVIDIA AI Enterprise Software Suite

vmware®

Table of contents

Flexible, scalable hybrid cloud platform for AI/ML workloads

Artificial Intelligence (AI) applications are transforming every business

Artificial Intelligence (AI) and Machine Learning (ML) initiatives are leading the path to digital transformation across all industries enabling newer AI powered workloads and applications. These AI/ML workloads have special requirements across the AI lifecycle, including unique performance characteristics for training, inferencing and performing data analytics. For example, deep learning training requires high performance to execute massively parallel processing using private cloud infrastructure with GPUs or other special purpose processors and hardware components. Near real time processing requirements of inferencing require the ability to process data coming through AI-trained models for the desired outcomes. These inferencing workloads may need to be within or adjacent to the enterprise, and the distributed nature of these workloads imposes specific security requirements to ensure compliance. As a result, enterprises and their IT organizations are looking to provide superior performance and acceleration to these AI workloads in a secure fashion.

To further complicate matters, these AI workloads are typically deployed with containerized and stateful workloads to take full advantage of the modern infrastructure required to support the AI/ML lifecycle. IT organizations are realizing that their staff does not have the necessary skillset to setup and manage different silos of container, VM and AI workloads. This increased complexity can significantly increase costs and delay results of the strategic initiatives within organizations. To address these challenges, what's needed is a hybrid infrastructure that is flexible enough to support the needs of the business while still ensuring that it is performant and scalable. Organizations must address the skillset gaps that are amplified by the complexity of this AI infrastructure. As a result, IT teams are looking for the ability to leverage hyperconverged infrastructure (HCI) in software-defined environments with accelerated GPUs in an automated fashion utilizing existing management tools to minimize complexity and simplify deployments.

Deliver AI-ready infrastructure

VMware Cloud Foundation™ with Tanzu® provides a full-stack hybrid cloud platform that delivers AI-ready infrastructure to enable customers to accelerate AI/ML workloads as well as traditional enterprise apps. Based on a proven and comprehensive software-defined stack including VMware vSphere®, VMware vSAN™, VMware NSX-T™ Data Center, and VMware vRealize® Suite, VMware Cloud Foundation™ with Tanzu® provides a complete set of secure software-defined services for compute, storage, network security, Kubernetes management, and cloud management. The result is agile, reliable, efficient cloud infrastructure that offers consistent operations across private and public clouds. In addition, VMware Cloud Foundation contains built-in automated lifecycle management to simplify the administration of the software stack, from initial deployment, to patching and upgrading. VMware and NVIDIA have partnered to unlock the power of AI by delivering an end-to-end platform optimized for AI workloads. This integrated platform delivers best in class AI software, the NVIDIA AI Enterprise Suite which is optimized to work with VMware Cloud Foundation with Tanzu

Solution architecture

Figure 1: VMware Cloud Foundation Solution Architecture

VMware Cloud Foundation capabilities

Application-focused management

Most AI/ML workloads are cloud native in nature, deployed in containers and optimized to work with GPUs. They need to co-exist with other workloads on a shared infrastructure which results in a mix of virtual machines (VMs) and containers.

With VMware Cloud Foundation, customers get unified visibility of VMs, containers, and Kubernetes clusters in vCenter. Containers and Kubernetes clusters are treated as first-class citizens like VMs, from a vCenter perspective.

The Kubernetes concept of a namespace is integrated into vSphere and becomes the unit of management. By grouping resource objects such as VMs and containers into logical applications via namespaces, Virtual Infrastructure

(VI) admins who previously managed thousands of VMs can now manage just dozens of application namespaces, resulting in a massive increase in scale and reduction in cognitive load. This agility and flexibility is very effective in driving up efficiency and utilization in AI pipelines.

Performance, security and compliance for AI workloads

AI workloads consume massive amounts of data using infrastructure that must meet stringent requirements for security, high availability and resiliency. Secure, low-latency access to this data during training, inferencing and analytics requires performance acceleration using GPUs to deliver faster business outcomes and meet the desired SLAs. AI workloads must be able to dynamically consume GPU resources to meet increasing performance requirements across larger data sets, while maintaining compliance with privacy and security mandates.

With VMware Cloud Foundation, policies for security, performance and availability are centrally managed to ensure consistency across the enterprise landscape. Admins can define QoS, security policies, firewall rules, encryption settings, availability and access control rules at namespace level, reducing the time it takes to manage and troubleshoot large scale AI applications.

NSX-T has been designed as the pod networking for VMware Cloud Foundation with Tanzu. NSX-T provides the full stack networking and security to VMware Cloud Foundation with Tanzu, including distributed switching and routing, distributed firewalling, load balancing, ingress control and more. Integrations with Kubernetes enables context-aware security policies with namespace isolation.

Higher productivity with less friction

Data scientists and developers need the ability to develop and test their AI models in test environments and quickly iterate to get to their desired outcome. As a result, they need to easily create and test their AI/ML workloads before deploying them in production. However, many IT organizations rely on slow ticketing systems to provide infrastructure services to developers because it is the only way to provide governance over developer applications and processes. VMware Cloud Foundation provides the ability to manage at the namespace level so that admins can set policies, quota and role-based access to a namespace once. Developers can then self-service into the namespace within the predefined boundaries.

With Kubernetes embedded into the control plane of vSphere, developers can create and consume cloud resources such as Kubernetes clusters, volumes (including persistent volumes for stateful applications), and networks using Kubernetes and RESTful APIs with a declarative model. This reduces the time and effort it takes for infrastructure provisioning and scaling so that developers can focus on building apps. Meanwhile, IT operators maintain visibility into those cloud resources created by developers through the VMware interfaces with which they are familiar.

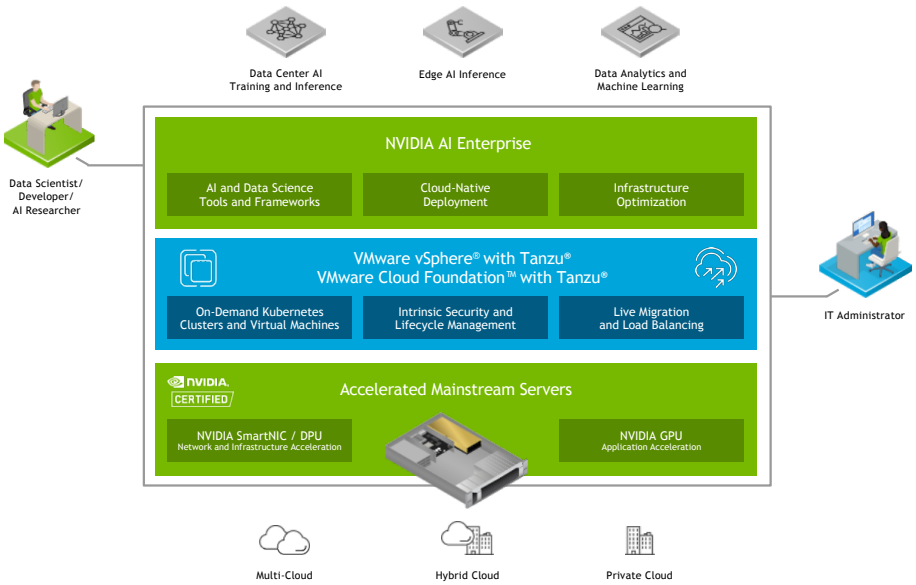

Figure 2: VMware Cloud Foundation with NVIDIA AI Ready Enterprise Suite

Unlock the raw performance of NVIDIA with full stack agility

For best efficiency and scale, AI workloads should be deployed alongside traditional workloads using a combination of container and VM based infrastructure. However, many IT organizations lack the necessary skill sets for deploying advanced AI and Kubernetes workloads without creating silos. The additional complexity comes from the tools and services required to maintain these complex application environments running at peak performance.

With the integration of the NVIDIA AI Enterprise Suite and VMware Cloud Foundation, AI and data science tools and frameworks such as TensorFlow and Pytorch can be used in conjunction with cloud native deployment tools such as NVIDIA GPU operator and NVIDIA network operator to enable IT admins to deploy these workloads easily.

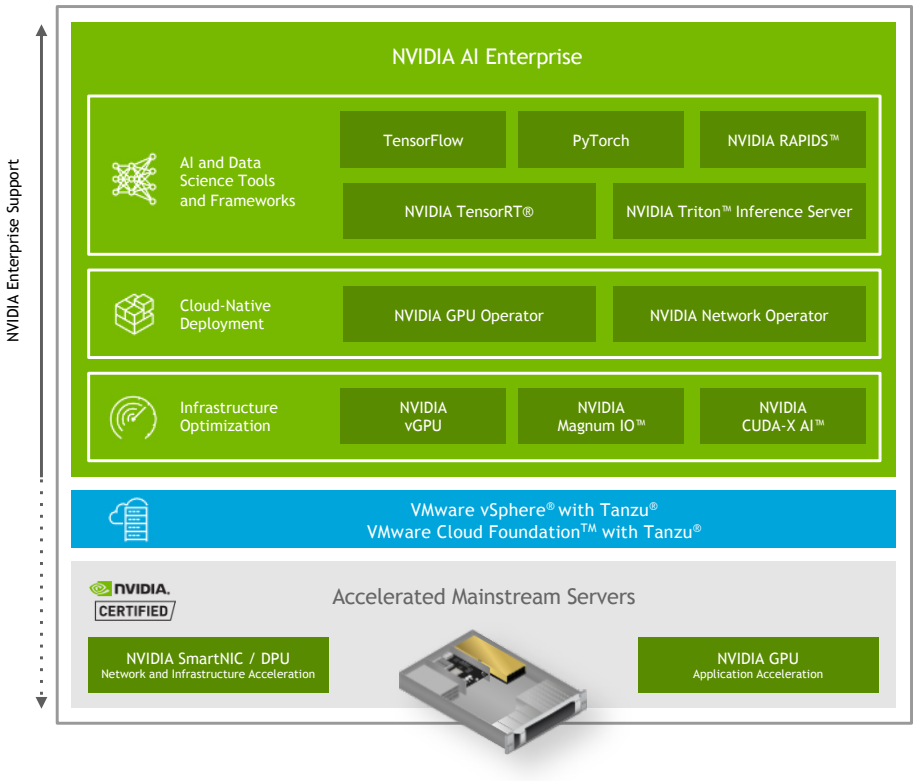

Figure 3: NVIDIA AI Enterprise Suite

Given the distributed nature of AI/ML workloads, there is a need to deploy workloads easily and rapidly on-premises, at the edge or in the cloud. VMware Cloud Foundation addresses this need with distributed workload domains. A workload domain is a policy-based resource construct with specific availability and performance attributes. It combines compute (vSphere), storage (vSAN), networking (NSX-T) and cloud management (vRealize) into a single consumable entity. Workload domains greatly speed up the instantiation of Kubernetes, deploying both the underlying infrastructure and Kubernetes components in an automated fashion. Workload domains also allow IT operators and developers to securely sandbox and allocate the right infrastructure for containers alongside VMs.

Enhanced raw performance for AI workloads while maximizing utilization

VMware Cloud Foundation with Tanzu now supports the NVIDIA Ampere A30 and A100 GPUs through integration with the NVIDIA AI Enterprise Suite. This allows GPUs to be shared across multiple instances to deliver increased utilization and optimization of resources with the benefit of reduced costs.

AI Training and Inferencing workloads require a highly performant infrastructure and benefit from GPUs and network accelerators. NVIDIA AI Enterprise integrates with VMware Cloud Foundation for provisioning and management of a fleet of GPUs in a simplified full stack HCI environment. It includes optimizations for near bare metal performance of accelerated VMs and containers. NVIDIA GPUs can be configured for time slicing or multi-instance GPU (MIG) to provide predictable performance while increasing utilization of these critical resources.

Efficient use and easy management of GPUs for AI/ML workloads

IT admins can quickly and easily provision self-serve capabilities to Data scientist and DevOps teams when building AI/ML data pipelines using these vGPUs with VMware Cloud Foundation and the NVIDIA AI Enterprise Suite. Developers can consume GPU resources that can be scaled up or down independently when needed by data scientist independently of the IT admin teams by selecting pre-configured vGPU profiles.

Infrastructure life cycle management

VMware Cloud Foundation offers automated lifecycle management to simplify management of the infrastructure stack on a per-workload domain basis. Available updates for all components are tested for interoperability and bundled with the necessary logic for proper installation order. The update bundles are then scheduled for automatic installation on a per-workload domain basis. This allows the admin to target specific workloads or environments (development vs. production, for example) for updates independent from the rest of the environment.

Resources

- Learn more about VMware Cloud Foundation
- Check out Cloud Foundation Blog, Twitter, and YouTube for the latest updates on Cloud Foundation
- Want to try Cloud Foundation for yourself? Visit the VMware Cloud Foundation Hands-On Lab

Intrinsic security at every layer of the stack

At the container image layer, Tanzu Kubernetes Grid includes a best-in-class container registry with built-in vulnerability scanning, image signing and auditing.

At the compute layer, vSphere provides comprehensive built-in security for protecting data, infrastructure and access that is operationally simple. Policy-driven security provides VM- or pod-level encryption to protect unauthorized data access both at rest and in motion.

At the network layer, NSX-T delivers micro-segmentation and granular security to the individual VM or pod workload, enabling a fundamentally more secure data center. Security policies travel with the workloads, independent of where workloads are in the network topology.

At the storage layer, vSAN offers data at rest and data in transit encryption at the cluster level. Storage Encryption is built for compliance requirements and offers simple key management with support for all Key Management Interoperability Protocol (KMIP) compliant key managers.

At the management layer, vRealize solutions automate manual tasks to eliminate human error, provide monitoring and auditing the full stack, and provide self-driving operations to quickly remediate issues as they are identified.

Cloud operating model extending across private and hybrid cloud

The same core software-defined infrastructure stack leveraged in private cloud deployments of VMware Cloud Foundation is also the underpinning technology of VMware-based public clouds like VMware Cloud on AWS and other VMware Cloud Provider™ Program partners. With VMware Cloud Foundation powered clouds offering consistent infrastructure and operations, customers can transform to a different way of operating IT, where service delivery is better aligned to the service consumption needs of the business. Adopting a cloud operating model represents a move toward application modernization and new application architectures that enable digital initiatives.

VMware Cloud Foundation – AI-ready infrastructure

- **Easy to deploy** and run integrated AI-ready infrastructure, including compute, storage, networking, security and cloud management services for modern applications on the same platform as for traditional applications.
- **Boosts developer productivity**, allowing app and data science teams to access cloud resources that they are already familiar with through industry standard APIs.
- **Simple to operate and future proof hybrid cloud strategy** that is consistent and compatible across on- and off-premises environments with the ability to deploy VMs, containers, and any next-generation application needs.

Item No: AI Ready Infrastructure with NVIDIA and VMware Cloud Foundation 2/22