VMware Baseline Reference Architecture for Generative AI

KEY COMPONENTS

- Hugging Face Transformers Library, PyTorch
- Ray, Kubeflow
- VMware Cloud Foundation 5.0

SUMMARY

The VMware GenAI Baseline Reference Architecture provides organizations with the fastest path from ML project to production. Leveraging industry-leading virtualization and cloud technologies, the platform is both efficient and secure, accelerating AI transformation by delivering simple and efficient management of GenAI workloads at scale, with state-of-the-art training and inferencing tools and GPU support. A validated GenAI technology drastically simplifies operationalization of AI/ML workloads by delivering an integrated solution leveraging open source components that is easy to operate and built-in automated lifecycle management of the VMware platform.

KEY BENEFITS

- **Accelerate time to results**: Speed up time to market by eliminating complexities in system design, testing, bring-up, configuration, and provisioning processes.
- **Trusted deployment**: Enable quick, repeatable, secure deployments based on a standardized, VMware Validated Architecture.
- **Improve ROI**: Maximize resource utilization and reduce time to support different organizations on a common base platform.
- **Flexible infrastructure**: Run all AI workloads, both predictive and GenAI on a common on-prem cloud infrastructure.

VMware's GenAI Reference Architecture offers a comprehensive solution for integrating AI and Enterprise workloads, improving operational efficiency, and enabling rapid infrastructure changes. The architecture includes recommended hardware, software, integration, and support, streamlining implementation and deployment without piecing together disparate elements. By leveraging VMware virtualization, sensitive data and applications can be secured within dedicated virtual environments, ensuring data privacy and protection against unauthorized access. Administrators gain comprehensive control over resource allocation and network configurations, enhancing management and customization of computing resources to meet specific business needs.

A VMware validated solution for GenAI will reduce the time, effort, and cost for building and management of the LLM infrastructure and LLM software stack, accelerating the time to value and enabling enterprises to focus on delivering value from their GenAI-based applications, securely and while maintaining data privacy. IT teams can leverage VMware management tools for rapid deployment, management, and scaling of AI workloads on GPU accelerated software-defined cloud.

VMware GenAI Platform for LLM Customization and Application Deployment

The solution can be leveraged by customers for two primary LLM workflows – customization (fine-tuning, prompt-tuning, and others etc.) and inference at scale. When it comes to LLMs, both these workflows demand more compute capacity than traditional ML or Deep Learning workloads.

- Customization of open-source LLMs (ranging from a few billion to 15 billion parameters or more) will require continuous distributed training on multiple GPUs spread across multiple servers.
- Inference will also require GPU resources and can be significant depending on application needs and number of concurrent users.

Underutilized GPU resources in the environment can be easily leveraged using vSphere capabilities for distributed training of models or inference depending on compute and SLA requirements – thus improving infrastructure utilization and improving overall productivity for ML workflows.

By leveraging an integrated and validated solution setup from VMware will enable organizations to start realizing the benefits of the solution almost immediately, accelerating the time it takes to move GenAI models to production and realize value for their investment.

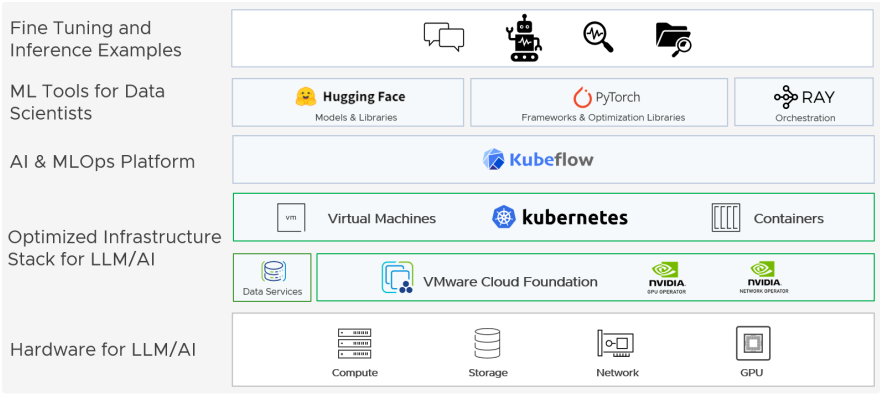
LLM Stack overview

Key vSphere Functionalities

- NVIDIA NVSwitch Support
- Device Groups
- Simplified Hardware Consumption with Device Groups
- Heterogeneous vGPU Profile
- Elevated Security

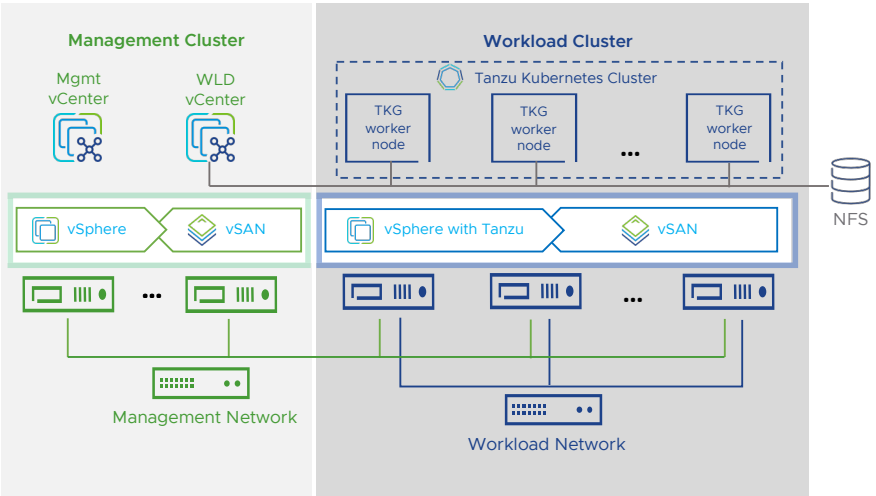
A high-level overview of the solution is depicted below starting with the infrastructure components up to the LLM application layer. The application layer is composed of the following Open-Source ML components:

Ray - a cutting-edge open-source distributed computing framework tailored to meet the demands of modern data-driven and AI-powered applications. Hugging Face Transformers library to work with pre-trained transformer models leveraging PyTorch deep learning framework which provides a rich ecosystem of libraries and tools specifically designed for GenAI workflows. Organizations can leverage Kubeflow to streamline the building and deployment of AI models at scale in a cloud native environment.



LLM Reference Architecture on VMware Cloud

VMware Cloud is a next-generation, multi-cloud IaaS software that empowers IT operations to deliver the right infrastructure for every application. This solution provides a full stack of enterprise-class Compute, Networking, Storage, Management, and security services with a well architected framework. The validated optimized infrastructure stack leverages Tanzu Kubernetes Grid Service (TKGS) for providing Kubernetes infrastructure for the cloud native Ai & MLOps platform components. TKG offers a consistent and scalable Kubernetes experience, simplifying the complexities of container orchestration, and enabling enterprises to focus on delivering value through their AI/ML workloads.



vSphere AI Capabilities

vmware

With the new Update 1 release of VMware vSphere® 8, the enterprise workload platform, customers benefit from enhanced operational efficiency for admins, supercharged performance for higher-end AI/ML workloads, and elevated security across the environment.

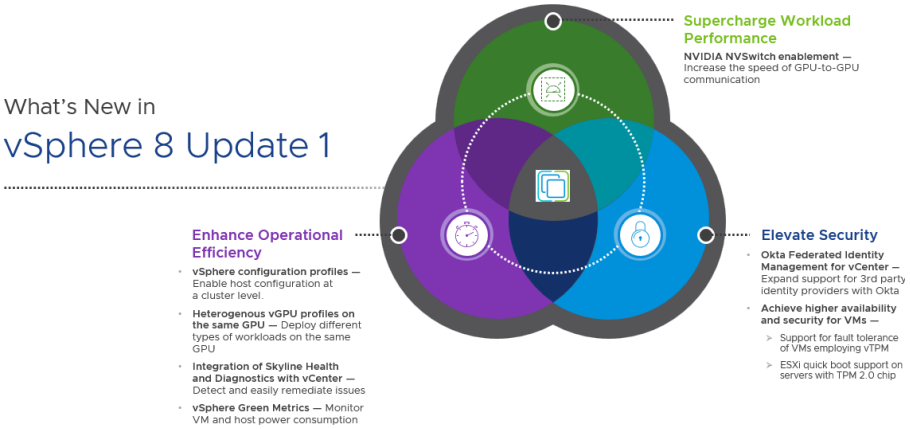
For up to 8 GPUs per host, vSphere now supports the deployment of NVIDIA NVSwitch technology, greatly improving large sized AI/ML workload performance. All 8 GPUs or a subset of them, can be allocated to a single VM.

Device Groups makes Virtual Machines consuming complementary hardware devices simpler in vSphere 8. NIC and GPU devices are supported in vSphere 8 GA. Compatible vendor device drivers are required and subject to vendor release. NVIDIA® will be the first partner supporting Device Groups with upcoming compatible drivers.

Device Groups are added to virtual machines using the existing Add New PCI Device workflows. vSphere DRS and vSphere HA are aware of device groups and will place VMs appropriately to satisfy the device group.

vSphere reduces cost by improving GPU utilization and reducing workload fragmentation in GPUs by the addition of support for heterogenous vGPU profiles on the same GPU. This capability allows for different types of workloads to be deployed on the GPUs, such as VDI applications, compute applications, graphics applications.

Elevates security by adding support for Okta Identity federation, Quick Boot support for servers with TPM 2.0 chip, and Fault Tolerance of VMs that use virtual TPM.



vmware®