Virtualizing GPUs for AI with VMware and NVIDIA

Based on Dell Infrastructure

May 2023

H18903.3

White Paper

Abstract

This white paper describes the Dell Validated Design for Virtualizing GPUs for AI, designed in collaboration with VMware and NVIDIA, using VMware vSphere and Tanzu and NVIDIA AI Enterprise, based on Dell Infrastructure. It details how enterprises can run AI workloads along with existing applications in data centers without compromising performance.

Dell Technologies Solutions

Dell
Validated Design

NVIDIA

vmware

DELL Technologies

Copyright

Contents

Introduction

Executive summary

Dell Technologies, VMware, and NVIDIA are working collaboratively to develop solutions to improve the state-of-the-art practices for implementing AI in the enterprise. Flexibility and ease of use continue to be key tenets and desirable traits for artificial intelligence (AI) platforms that are used to train and host models for AI. This white paper describes the Dell Validated Design for Virtualizing GPUs for AI with VMware and NVIDIA, a solution that the three companies jointly engineered and validated. We provide background information and recommendations for how to implement a wide variety of NVIDIA graphics processing unit (GPU) acceleration products using several types of servers from Dell Technologies along with VMware vSphere and Tanzu technology. This solution can match the needs of almost any AI application requirements for enterprise customers who want to standardize on VMware for systems deployment and operations. The validated design also shows how to incorporate NVIDIA AI Enterprise, a comprehensive software suite of AI tools and frameworks, which enables organizations running VMware vSphere to virtualize and containerize AI workloads on NVIDIA-Certified Systems.

One lesson learned over the last decade is that AI platforms with custom architectures—whether on-premises or in a cloud service provider—introduce integration and operational challenges that in turn result in higher costs. With this validated design, VMware professionals can use their existing skill set for deployment and operations of this virtualized AI platform. The ability for an organization to maintain and evolve its existing operational centers of excellence allows for lower total cost of ownership and better overall business continuity.

Enterprises are embracing AI in every aspect of their business. Human resources are using AI for talent acquisition, marketing is using AI for pricing and demand forecasting, IT is using AI for cyber security, and customer services are using AI for chatbots. Such wide adoption of AI across an enterprise requires seamless integration of AI capabilities into its data center operations. Through NVIDIA AI Enterprise on VMware vSphere, this validated design enables enterprises to deploy, modernize, manage, and operate AI workloads along with their existing applications using the same tools with which they are familiar.

Document purpose

Using the information in this white paper, VMware professionals can quickly deploy and operate a full-featured platform to support advanced AI use cases that take advantage of NVIDIA-accelerated GPUs and curated AI software for AI researchers, data scientists, and developers on Dell Technologies infrastructures.

This white paper is a companion document to the Virtualizing GPUs for AI with VMware and NVIDIA Design Guide and Implementation Guide. See the design guide for more information about the reference architecture, configurations, and performance characterization. See the implementation guide for guidance about deploying the solution.

This updated white paper reflects the latest Dell PowerEdge servers, VMware vSphere 8, and NVIDIA AI Enterprise 3. For information about previous Dell PowerEdge servers, VMware vSphere 7, and NVIDIA AI Enterprise 1.x, download the following documents:

- White paper
- Design Guide
- Implementation Guide

Note: The contents of this document are valid for the described software and hardware versions. For information about updated configurations for newer software and hardware versions, contact your Dell Technologies sales representative.

Audience

This white paper is intended for solution architects, system administrators, and others who are interested in virtualized platforms for developing AI applications.

Revision Table

Table 1. Revision history

Date	Version	Change summary
May 2021	1.0	Initial release
October 2021	1.1	Updated to add support for NVIDIA AI Enterprise
March 2022	2.0	Updated to add support for VMware Tanzu
May 2023	3.0	Updated to add support for VMware vSphere 8, NVIDIA AI Enterprise 3.1, and the latest PowerEdge servers

Business challenges

The shadow AI challenge

Developing the skills to be successful at the integration of AI, software development, and IT operations has challenged everyone. The demands placed on IT to provide robust production systems for business-critical AI workloads while managing additional environments for development of new initiatives have severely taxed already limited resources. The pressure to advance from concept, through experimentation, and to production has frequently resulted in data scientists and developers abandoning collaboration with IT. Instead, they attempt to proceed faster alone. This situation, which surfaced during the years of rapid change in the business intelligence and microservice-oriented applications development eras, further strains the already challenged relationships between IT, the developer community, and the business management communities.

The experimental nature of data science work makes collaboration and planning challenging. Allocation of IT resources in an environment in which resources are required for “development labs” is difficult to predict. The uncertainty of time to value makes budgeting and workload management nearly impossible. The IT department can feel that it lacks sufficient information to allocate resources effectively, and developers often feel that there is a lack of priority in response to changes in a previously agreed-to plan.

These factors have produced incentives for groups that are involved in the rush to implement AI workloads to behave in ways that are not cost effective for their organizations. The most common ways that groups attempt to “go faster” is to use unmonitored public cloud usage, reuse of equipment for an unintended effort, or acquisition of business unit funding for siloed development initiatives that are outside the official IT capital budgeting process. These types of information systems, which exist largely hidden from managers and official IT units, create the ‘‘shadow IT’’ problem with which all larger organizations must deal. The circumstances that motivate groups to choose the shadow IT route are particularly acute in the AI application and machine intelligence research areas.

This focus on AI, once the interest of only a small community of researchers and computer scientists, is now so intense that many organizations feel that despite considerable investment, they are falling further behind their competitors. The rate of new data management and AI being brought to market increases the cost of system planning and increases the risk of having to change course more frequently.

Solution overview

Dell Technologies, NVIDIA, and VMware are offering enterprises a way forward with the launch of an integrated solution to democratize and unlock AI across the enterprise. This validated design is jointly engineered and validated to help organizations capitalize on the benefits of virtualization for AI workloads. The design includes the latest version of VMware vSphere and Tanzu combined with the NVIDIA AI Enterprise suite on Dell PowerEdge servers. The design also includes Dell PowerScale, which provides the necessary analytics performance and concurrency at scale to consistently feed the most data hungry AI algorithms.

The following figure shows the solution components:

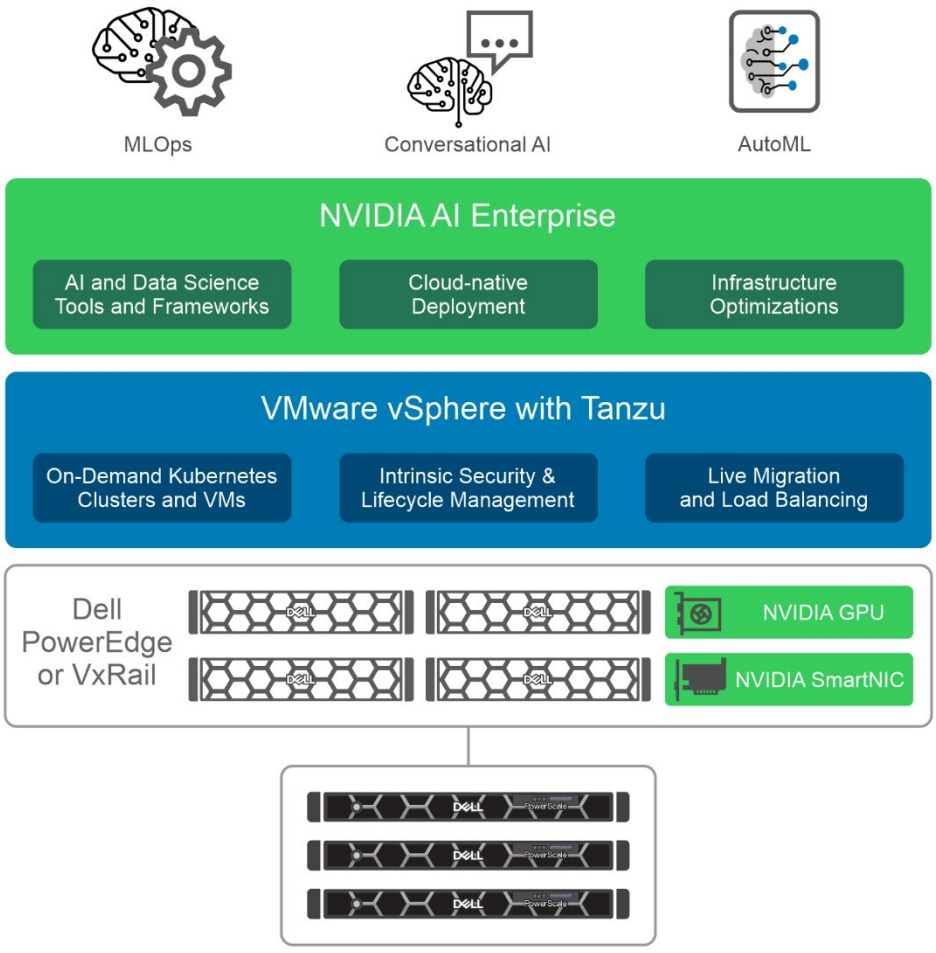

Figure 1. Overview of solution components

Benefits of Validated Design for AI

This combination of leading-edge technologies makes it possible to adopt the latest NVIDIA Ampere GPUs using the predictability and security of vSphere for virtualization with VMware-optimized infrastructure. This validated design provides the following key benefits:

- **No siloed infrastructure for AI**—Customers can use the same data center tools and processes with which they are familiar for building and operating AI

infrastructure. With integration to the VMware ecosystem, customers can avoid silos of AI-specific systems that are difficult to manage and secure. They can also mitigate the risks of shadow AI deployments, where data scientists and machine learning engineers procure resources outside of the IT ecosystem.

- **Consistent tools for management and operations**—GPU resources can now be virtualized similarly to CPU, memory, network, and storage resources. This virtualization and container orchestration allows IT administrators to use the same tools for management and operations for both their AI workloads and other data center workloads.
- **AI workload orchestration**—Through integration with NVIDIA AI Enterprise and VMware Tanzu, this validated design enables automation of the AI workload's life cycle, including provisioning, deployment, scaling, networking, and load balancing. Administrators can now simplify their complex AI deployment through production-grade Kubernetes container orchestration.
- **Curated end-to-end AI software with Enterprise grade support**—The NVIDIA AI Enterprise software suite includes AI and data science tools and frameworks that are packaged as containers for easy and rapid deployment. These containers support end-to-end AI development and are validated on VMware vSphere. NVIDIA Support Services for the NVIDIA AI Enterprise software suite provides access to comprehensive software patches, updates, upgrades, and technical support. These services help customers with an easy and reliable way to improve productivity and reduce downtime for their AI infrastructure.
- **Near bare-metal performance and scalability**—AI workloads can run at near bare-metal performance on virtualized GPUs. These workloads can scale across multiple GPUs and multiple nodes, allowing training of even the largest deep learning models.

Key features

Some of the key features of this validated design include:

- **GPU virtualization and allocation**—VMware vSphere 7 and later supports virtualization for NVIDIA Ampere GPUs. The virtualized GPUs can be assigned to virtual machines (VMs) and containers through Single-Root Input/Output Virtualization (SR-IOV). Also, vSphere supports:
  - **Partitioning of GPUs** using NVIDIA Multi-Instance GPU (MIG) technology, which increases GPU use. MIG-partitioned virtual GPU (vGPU) instances are fully isolated with an exclusive allocation of high-bandwidth memory, cache, and compute. A common use case is for administrators to partition available GPUs into multiple units for allocation to individual data scientists through VMs or containers. Each data scientist can be assured of predictable performance due to the isolation and Quality of Service guarantees of the vGPU partitioning technology.
  - **GPU aggregation** allowing multiple virtual GPUs to be assigned to VMs and containers to allow deep learning jobs that are compute intensive. GPUDirect RDMA from NVIDIA provides more efficient data exchange between GPUs that perform multinode training at scale. It enables a direct peer-to-peer data path between the memory resources of two or more GPUs using ConnectX network adapter ports on the host.

- **Support for GPU virtualization with Tanzu container orchestration**—Virtualized GPUs can now be made available to enterprise-grade Kubernetes container orchestration through Tanzu. Administrators can provision AI workloads as Kubernetes pods or through Helm deployments, which use virtualized GPUs.
- **Availability and continuous maintenance using VMware vSphere vMotion**—vSphere enables live migration (using vSphere vMotion) for NVIDIA vGPU-powered VMs, simplifying infrastructure maintenance such as consolidation, expansion, or upgrades, and enabling nondisruptive operations.

With the Distributed Resource Scheduler (DRS), vSphere provides automatic initial workload placement for AI infrastructure at scale for optimal resource consumption and to avoid performance bottlenecks.

- **Support for VM suspend and resume operations with virtual GPUs multinode training**—GPUDirect RDMA from NVIDIA enables a direct peer-to-peer data path between the GPU memory and ConnectX network adapters. This path provides a significant decrease in GPU-to-GPU communication latency and completely offloads the CPU, removing it from all GPU-to-GPU communications across the network. GPUDirect RDMA from NVIDIA enables near bare-metal performance on multinode training.

Engaging the Dell Technologies Customer Solutions Center

The Dell Technologies Customer Solution Center helps you plan and achieve your business goals to accelerate your digital future:

- **Proof of Concept**—Validate that your preferred solution meets your needs with a custom Proof of Concept. Dell Technologies solution architects enable practical, hands-on implementation based on your test cases.
- **Design Session**—Collaborate with Dell Technologies experts to design a solution framework. Brainstorm with our experts to explore your current IT environment, your future objectives, and business solutions.
- **Technical Deep Dive**—Dive into the technical solution details that you are considering for your business. Learn from live product demonstrations and solution-focused discussions with Dell Technologies subject matter experts.

Contact your Dell Technologies Sales Representative today to schedule a customized briefing or solutions engagement for this or any other Dell Validated Design for AI.

Technology overview

VMware

VMware vSphere 8

VMware vSphere 8 includes the following features to support AI and machine learning workloads:

- Support for the latest generation of GPUs from NVIDIA, including support for the spatial partitioning-based NVIDIA MIGs.
- Enhanced performance of device-to-device communication, building on the existing NVIDIA GPUDirect functionality by enabling Address Translation Services (ATS) and Access Control Services (ACS) at the PCIe bus layer in the ESXi kernel.

- Support for device groups for multi-GPU and multinode training. Device groups enable virtual machines to consume complementary hardware devices more easily. NVIDIA Smart NICs and GPU devices are supported in vSphere 8. Device groups are added to virtual machines by using the existing *Add New PCI Device* workflows. Device groups aid in automatic configuration of virtual machine and creation of worker nodes in Tanzu. vSphere DRS and vSphere HA are aware of device groups and places VMs appropriately to satisfy the device group.
- VMware vSphere licensing per CPU socket. Licensing is available for the following editions:
  - vSphere Standard
  - vSphere Enterprise Plus
  - vSphere Essentials
  - vSphere Essentials Plus

  This validated design requires the vSphere Enterprise Plus editions. NVIDIA vGPU and distributed virtual switches (required for load balancing in Tanzu) require the Enterprise Plus edition.

VMware vSphere with Tanzu

vSphere with Tanzu enables administrators to transform vSphere into a platform for running Kubernetes workloads natively on the hypervisor layer. When enabled on a vSphere cluster, vSphere with Tanzu provides the capability to run Kubernetes workloads directly on ESXi hosts and to create upstream Kubernetes clusters in dedicated resource pools.

vSphere administrators can enable existing vSphere clusters for Workload Management, to create a Tanzu Kubernetes cluster in the ESXi hosts that are part of the cluster. The Tanzu Kubernetes cluster is a full distribution of the open-source Kubernetes container orchestration platform that is built, signed, and supported by VMware. Tanzu Kubernetes Grid (TKG) Service provisions and operates Tanzu Kubernetes cluster on vSphere.

Tanzu Kubernetes Grid (TKG), available with VMware vSphere 8, supports virtualizing NVIDIA GPUs through NVIDIA AI Enterprise. With TKG, virtual GPUs are automatically provisioned and configured on the Tanzu Kubernetes Cluster worker nodes and made available to AI workload containers.

VMware vSphere with Tanzu can be licensed through vSphere+ or Tanzu Kubernetes Operations. For more information, see the VMware vSphere Product Line Comparison and VMware Tanzu for Kubernetes Operations Documentation.

VMware Kubernetes ecosystem

VMware offers several products under the Tanzu portfolio to enhance the capabilities of vSphere on Tanzu. These products enable administrators to build, run, and manage the AI workload along with modern applications and continuously deliver value to customers. Depending on the Tanzu edition, these software products are bundled with VMware vSphere with Tanzu and are fully supported by VMware. Some key products that are applicable to this validated design include:

- **Harbor**—An open-source, trusted, cloud native container registry that stores, signs, and scans content. Harbor extends the open-source Docker distribution by adding functionalities such as security, identity control, and management.
- **Tanzu Kubernetes Grid**—Includes signed binaries for Harbor that you can deploy on a shared services cluster to provide container registry services for other Tanzu Kubernetes clusters.
- **Prometheus**—An open-source systems monitoring and alerting toolkit. Prometheus collects and stores metrics as time series data, that is, metrics information is stored with the timestamp at which it was recorded, along with optional key-value pairs. Tanzu Kubernetes Grid includes signed binaries for Prometheus that you can deploy on Tanzu Kubernetes clusters to monitor cluster health and services.
- **Grafana**—Open-source software that allows you to visualize and analyze metrics data collected by Prometheus on Tanzu Kubernetes clusters. Tanzu Kubernetes Grid includes a Grafana package that you can deploy on the clusters.
- **VMware NSX Advanced Load Balancer**—NSX Advanced Load Balancer (formerly known as Avi Networks) with Cloud Services has multicloud load balancing, web application firewall, and container ingress services. The software-defined, scale-out architecture of NSX Advanced Load Balancer provides on-demand autoscaling of elastic load balancers. The distributed software load balancers and the backend applications can scale up or down in response to real-time traffic monitoring.

  NSX Advanced Load Balancer provides network access and load balancing for Tanzu Kubernetes clusters. You can use it to load balance AI use cases such as Machine Learning Operation applications or inference workloads.
- **Tanzu Mission Control**—A centralized hub for simplified, multicloud, multicluster Kubernetes management. Tanzu Mission Control provides centralized policy management that enables administrators to apply consistent policies, such as for access and security, to a fleet of clusters and namespaces at scale. It provides life cycle management for Kubernetes clusters enabling administrators to provision, scale, upgrade, and delete Tanzu Kubernetes Grid clusters.

The following additional software is available from VMware to manage and orchestrate container workloads. These software tools address general-purpose application development and are not validated as part of this validated design.

- **VMware Tanzu Application Platform** is a modular, application-aware platform that provides a rich set of developer tools and a path to production to build and

deploy software quickly and securely on any compliant public cloud or on-premises Kubernetes cluster.

- **Tanzu Observability** enables Kubernetes monitoring with full-stack visibility of nodes, pods, and containers. It provides instant insight into Tanzu Application Service platform health across foundations and the impact of code in production.
- **Tanzu Service Mesh** provides advanced, end-to-end connectivity, security, and insights for modern applications—across application end-users, microservices, APIs, and data—enabling compliance with Service Level Objectives and data protection and privacy regulations.
- **VMware application catalog** is a customizable selection of trusted, prepackaged open-source application components that are continuously maintained and verifiably tested for use in production environments.
- **Tanzu Build Service** automates container creation, management, and governance at enterprise scale while boosting security and reducing risk from Common Vulnerability Exposure.

  Tanzu Data Services is a portfolio of on-demand caching, messaging, and database software on VMware Tanzu for development teams building modern applications.

VMware vSAN 8

vSAN is a software-defined storage solution from VMware, built from the ground up for vSphere VMs. It abstracts and aggregates locally attached disks in a vSphere cluster to create a storage solution that you can provision and manage from vCenter and the vSphere client. vSAN is embedded in the hypervisor, therefore, storage and compute for VMs are delivered from the same x86 server platform running the hypervisor.

vSAN is the market leader in HCI infrastructure. Traditional applications such as Microsoft SQL Server and SAP HANA, and next-generation applications such as AI workloads can run on vSAN. Paradigms associated with traditional infrastructure deployment, operations, and maintenance include various disaggregated tools and often specialized skill sets. The hyperconverged approach of vSphere and vSAN simplifies these tasks using familiar tools to deploy, operate, and manage private-cloud infrastructure.

vSAN 8 Express Storage Architecture (ESA) is the latest major enhancement available for vSphere 8 clusters. vSAN 8 ESA uses a file system that is optimized to take full advantage of certified NVMe storage devices and 25 Gbps+ networking to greatly improve performance and capacity over previous versions. vSAN 7 is now referred to as Original Storage Architecture (OSA).

VMware vSAN is licensed per CPU socket. It is available in the following editions: Standard, Advanced, Enterprise, and Enterprise Plus. For this validated design, we recommend vSAN Enterprise license. vSphere Enterprise Plus, and VMware Tanzu Standard are required to use the Data Persistence platform. The Data Persistence platform is available in vSAN Enterprise and Enterprise Plus only.

NVIDIA

NVIDIA AI Enterprise

NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data analytics software. It is optimized, certified, and supported by NVIDIA to run exclusively on VMware vSphere with NVIDIA-Certified Systems, as shown in the following figure:

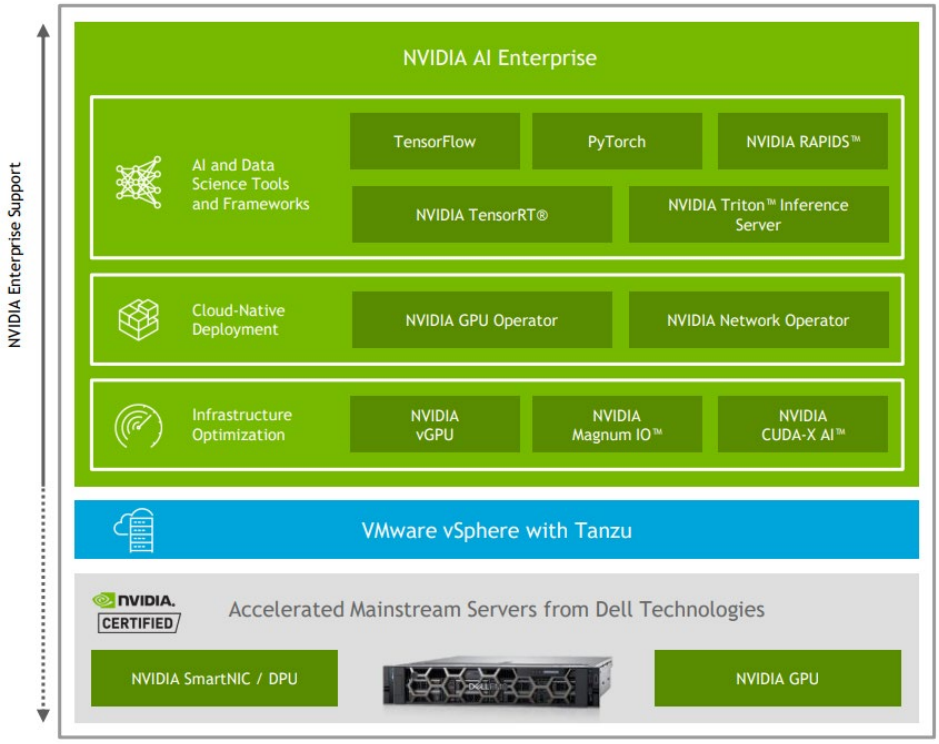

Figure 2. NVIDIA AI Enterprise—a comprehensive AI suite

NVIDIA AI Enterprise includes key enabling technologies and software from NVIDIA for rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud. NVIDIA licenses and supports NVIDIA AI Enterprise.

The software in the NVIDIA AI Enterprise suite is organized into the following layers:

- Infrastructure optimization software:
  - **NVIDIA vGPU**—NVIDIA vGPU software creates virtual GPUs that can be shared across multiple VMs enabling IT to use the management and security benefits of virtualization and the performance of NVIDIA GPUs.
  - **NVIDIA CUDA Toolkit**—CUDA Toolkit includes GPU-accelerated libraries, debugging and optimization tools, a C/C++ compiler, and a runtime library to build and deploy your AI application.
  - **NVIDIA Magnum IO**—Magnum IO stack contains the libraries that developers need to create and optimize applications IO across the entire stack, including:
    - Networking across NVIDIA NVLink
    - Ethernet

  - InfiniBand
  - Storage APIs
  - In-networking compute to accelerate multinode operations
  - IO management of networking hardware

- Cloud-Native Deployment software, which is required for VMware Tanzu support:
  - **NVIDIA GPU Operator** uses the operator framework in Kubernetes to automate the management of all NVIDIA software components needed to provision the GPU. These components include the NVIDIA drivers (to enable CUDA), the Kubernetes device plug-in for GPUs, the NVIDIA Container Runtime, automatic node labeling, DCGM-based monitoring, and others.
  - **NVIDIA Network Operator** uses the operator framework in Kubernetes to manage networking-related components to enable fast networking, RDMA, and GPUDirect for workloads in a Kubernetes cluster. Network Operator works with GPU Operator to enable GPU-Direct RDMA on compatible systems.
- AI and data science frameworks that include the following validated containers on VMware vSphere:
  - **NVIDIA RAPIDS** is an open-source machine learning framework. RAPIDS brings GPU optimization to problems traditionally solved by using tools such as Hadoop or Scikit-learn and pandas. RAPIDS is a useful tool for working with tabular and other data formats; it is also an essential tool for data preparation, data formatting, and data labeling. RAPIDS is a critical component to start any AI pipeline that requires data preprocessing.
  - **TensorFlow** is an open-source framework for machine learning implemented in a combination of C++ and NVIDIA CUDA tools. First developed by Google, it has been a mainstream tool for deep learning since its debut in 2015. The provided TensorFlow container has full support for GPUs, as well as multi-GPU and multinode capabilities along with NVIDIA-tested GPU optimizations.
  - **PyTorch** is an open-source Python Deep Learning Framework. Facebook created PyTorch, which like Tensorflow, is a leading AI framework. The PyTorch containers published through NVIDIA AI Enterprise include the software needed to run single GPU, multi-GPU, or multinode workloads.
  - **NVIDIA TensorRT** converts models developed in frameworks such as TensorFlow and PyTorch by compiling them into a format optimized for inferencing on a specific runtime platform. When compiling a model with TensorRT, features including bit-precision optimizations, neural network graph optimizations, and automatic tuning result in a more performant model for inference. The performance benefits can be significant depending on the type of model being developed. Generally, models compiled with TensorRT take up less memory and perform inference tasks faster than the original format.

- **NVIDIA Triton Inference Server** is an open-source model serving software that simplifies the deployment of production AI models at scale. It lets teams deploy trained AI models from any framework, including optimized TensorRT models on any single GPU, multi-GPU, or CPU-based infrastructure. When models are retrained, IT staff can easily deploy the updates without restarting the inference server or disrupting the calling application. Triton supports multiple inferencing types including real-time, batch, and streaming. It also supports efficient model ensembles if your pipeline has multiple models that share inputs and outputs, such as in conversational AI.

Licensing and enterprise support

NVIDIA AI Enterprise is licensed per CPU socket and can be purchased through Dell Software & Peripherals. You can purchase NVIDIA AI Enterprise products either as a perpetual license with support services or as an annual or multiyear subscription. The perpetual license provides the right to use the NVIDIA AI Enterprise software indefinitely, with no expiration. You must purchase NVIDIA AI Enterprise with perpetual licenses with one-year, three-year, or five-year support services. A one-year support service is also available for renewals. For more information, see the NVIDIA AI Enterprise Packaging, Pricing, and Licensing Guide.

NVIDIA Support Services for the NVIDIA AI Enterprise software suite provides seamless access to comprehensive software patches, updates, upgrades, and technical support.

NVIDIA Ampere GPU

The Tensor Core technology in the Ampere architecture has brought dramatic performance gains to AI workloads. Large-scale testing and customer case studies prove that Ampere GPUs that are based on Tensor Core can decrease training times significantly. Two types of Ampere GPUs are available for compute workloads:

- **NVIDIA A100 GPU**—This Tensor Core GPU can achieve massive acceleration for training workloads. IT professionals benefit from reduced operational complexity by using a single technology that is easy to onboard and manage for these use cases. The A100 GPU is a dual-slot 10.5-inch PCI Express (PCIe) Gen4 card that is based on the NVIDIA Ampere A100 GPU. It uses a passive heat sink for cooling. The A100 PCIe supports double precision (FP64), single precision (FP32), and half precision (FP16) compute tasks. It also supports unified virtual memory and a page migration engine.
- **NVIDIA A30 GPU**—This Tensor Core GPU is the most versatile mainstream compute GPU for AI inference and mainstream enterprise workloads. It supports a broad range of math precisions, providing a single accelerator to expedite every workload. Built for AI inference at scale, the same compute resource can rapidly retrain AI models with TF32, as well as accelerate high-performance computing (HPC) applications using FP64 Tensor Cores. MIG and FP64 Tensor Cores combine with fast 933 GB/s of memory bandwidth in a low 165 W power envelope, all running on a PCIe card that is optimal for mainstream servers.

A100 and A30 GPUs support the MIG feature, which allows administrators to partition a single GPU into multiple instances, each fully isolated with its own high-bandwidth memory, cache, and compute. The A100 PCIe card supports MIG configurations with up

to seven GPU instances per A100 GPU, while the A30 GPU supports up to four GPU instances. For more information, see the section about virtual GPUs in the Virtualizing GPUs for AI with VMware and NVIDIA design guide.

ConnectX SmartNICs

The ConnectX-6 Dx SmartNIC is a secure and advanced cloud network interface card that accelerates mission-critical, data center applications, such as virtualization, SDN/NFV, big data, machine learning, network security, and storage. ConnectX-6 supports Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE), the network protocol required for multinode training with GPUDirect RDMA. In this validated design, we use ConnectX-6 Lx for 25 Gb/s Ethernet connectivity and optionally ConnectX-6 Dx for 100 Gb/s Ethernet connectivity.

NVIDIA-Certified Systems

To be successful with machine learning and AI initiatives, enterprises need a modern coherent computing infrastructure that provides functionality, performance, security, and scalability. Organizations also benefit when they can run both development and production workloads with common technology. With NVIDIA-Certified Systems from Dell Technologies, enterprises can confidently choose performance-optimized hardware that runs VMware and NVIDIA software solutions—all backed by enterprise-grade support.

Dell Technologies produces a range of PowerEdge servers that are qualified as NVIDIA-Certified Systems. NVIDIA-Certified Systems are shipped with NVIDIA Ampere architecture A100 and A30 Tensor Core GPUs and the latest NVIDIA Mellanox ConnectX-6 network adapters.

A subset of NVIDIA-Certified Systems goes through additional certification, including VMware GPU certification, to ensure compatibility with NVIDIA AI Enterprise. An NVIDIA-Certified System that is compatible with NVIDIA AI Enterprise conforms to NVIDIA design best practices and has passed certification tests that address a range of use cases on VMware vSphere infrastructure. These use cases include deep learning training, AI inference, data science algorithms, intelligent video analytics, security, and network and storage offload for both single-node and multinode clusters.

Dell Technologies

Dell PowerEdge servers

The latest Dell PowerEdge servers are certified for VMware vSphere 8 and vSAN 8. PowerEdge Intel-based servers use the latest 4th Generation Intel Xeon Scalable processors. Other key features include:

- Intel Advanced Vector Extensions 512 (Intel AVX-512), which can accelerate classical machine learning and other workloads in the end-to-end AI workflow, such as data preparation. It can accelerate in-memory workloads with up to 32 DDR5 RDIMMS at up to 4800 MT/sec, using eight memory channels per CPU
- Support for two double-wide or six single-wide GPUs for workloads requiring acceleration
- Storage options that include SAS3/SAS4/SATA and NVMe Gen4/NVME Gen5 options
- Multiple Gen4 and Gen5 riser configurations

PowerEdge AMD-based servers use the latest EPYC 4th generation processors 4th Gen AMD EPYC processors. Other key features include:

- Strong generational performance uplifts across multiple workloads and demonstrate leadership TPCx-AI benchmark performance.
- Accelerate in-memory workloads with up to 24 DDR5 RDIMMS at up to 4800 MT/sec, using up to 12 memory channels per CPU
- Support for GPUs including 2 x double-wide or 6 x single-wide for workloads requiring acceleration
- Storage options include SAS3/SAS4/SATA, NVMe Gen4/NVME Gen5
- Multiple Gen4 and Gen5 riser configurations

The following table lists the PowerEdge servers that are supported with NVIDIA AI Enterprise and the number of Ampere GPUs that are supported with each server model:

Table 2. Supported PowerEdge servers

Server	Maximum A100 GPUs and A30 GPUs
PowerEdge R760	2
PowerEdge R7625	2

These PowerEdge servers are NVIDIA-Certified Systems and have been proven through a rigorous suite of functional and performance tests. The test results confirm that these servers can deliver high performance both in single-node and networked multinode cluster training and inference benchmarks. Also, these servers are certified to be compatible with NVIDIA AI Enterprise through additional testing and validation.

Dell PowerScale storage

PowerScale storage helps unlock the structure within data and address the challenges of unstructured data management. PowerScale is the next evolution of OneFS—the operating system powering the scale-out NAS platform. The PowerScale family includes Dell Isilon nodes and PowerScale nodes, with PowerScale OneFS running across all of them. The software-defined architecture of OneFS provides simplicity at scale, intelligent insights, and the ability to have data anywhere it needs to be. Whether hosting file shares or home directories, or delivering high-performance data access for applications such as analytics, video rendering, and life sciences, PowerScale can seamlessly scale performance, capacity, and efficiency to handle any unstructured data workload. The new PowerScale all-flash platforms co-exist seamlessly in the same cluster with your existing Isilon nodes to drive your traditional and modern applications.

In this validated design, we use PowerScale as storage for the data lake—the data repository for unstructured data that you can use for neural network training. PowerScale All-Flash Scale-out NAS storage is ideal, delivering the analytics performance and extreme concurrency at scale to consistently feed the most data-hungry deep learning algorithms.

GPUDirect Storage, an NVIDIA technology, enables a direct data path between local or remote storage, like NVMe or NVMe over Fabric (NVMe-oF), and GPU memory. GPUDirect Storage avoids extra copies through a bounce buffer in the CPU's memory. It

enables a direct memory access (DMA) engine near the storage to move data on a direct path into or out of GPU memory – all without burdening the CPU or GPU. PowerScale supports GPUDirect Storage.

Dell PowerStore storage

PowerStore is a modern storage appliance designed for the data era. The single architecture of PowerStore for block, file, and VMware vVols uses the latest technologies to support an enterprise-class variety of traditional and modern workloads – from relational databases, to ERP and EMR applications, cloud native applications, and file-based workloads such as content repositories and home directories. The ability to accommodate application, multiprotocol network, and multiformat storage diversity (physical and virtual volumes, containers, and traditional files) in a single 2U appliance provides business-enabling flexibility and helps IT simplify and consolidate their infrastructure.

Administrators can choose to deploy PowerStoreOS in a bare-metal configuration directly on the PowerStore hardware or in a VM running on PowerStore's optional integrated VMware hypervisor, providing yet another layer of isolation, intelligence, and abstraction.

Dell PowerSwitch switches

In this validated design, we use the following Dell switches:

- PowerSwitch S5232F-ON switch for 100 GbE network connectivity. Using the S5332F-ON switch, you can build a high-performance, cost-efficient data center leaf/spine fabric featuring 32 x 100 GbE QSFP28 ports. It supports Open Network Install Environment (ONIE) for zero-touch installation of network operating systems.
- PowerSwitch S5248F-ON switch for 25 GbE network connectivity. It features 48 x 25 GbE SFP28 ports, 4 x 100 GbE QSFP28 ports, and 2 x 100 GbE QFSP28-DD ports. It supports ONIE for zero-touch installation of network operating systems.
- PowerSwitch N3248TE-ON or PowerSwitch S4148T switch for 1 GbE out-of-band (OOB) connectivity.

Solution architecture

With VMware vSphere support for virtualized GPUs, IT administrators can run AI workloads such as neural network training, inference, or model development along with their standard data center applications. The following figure shows the high-level architecture for this validated design with PowerEdge R760 servers, each with two NVIDIA A100 GPUs and a ConnectX network adapter, as part of a VMware vSphere cluster. The VMs with vGPU run containers from NVIDIA AI Enterprise. This validated design allows AI workloads to run either as VM or Kubernetes pods in Tanzu Kubernetes clusters.

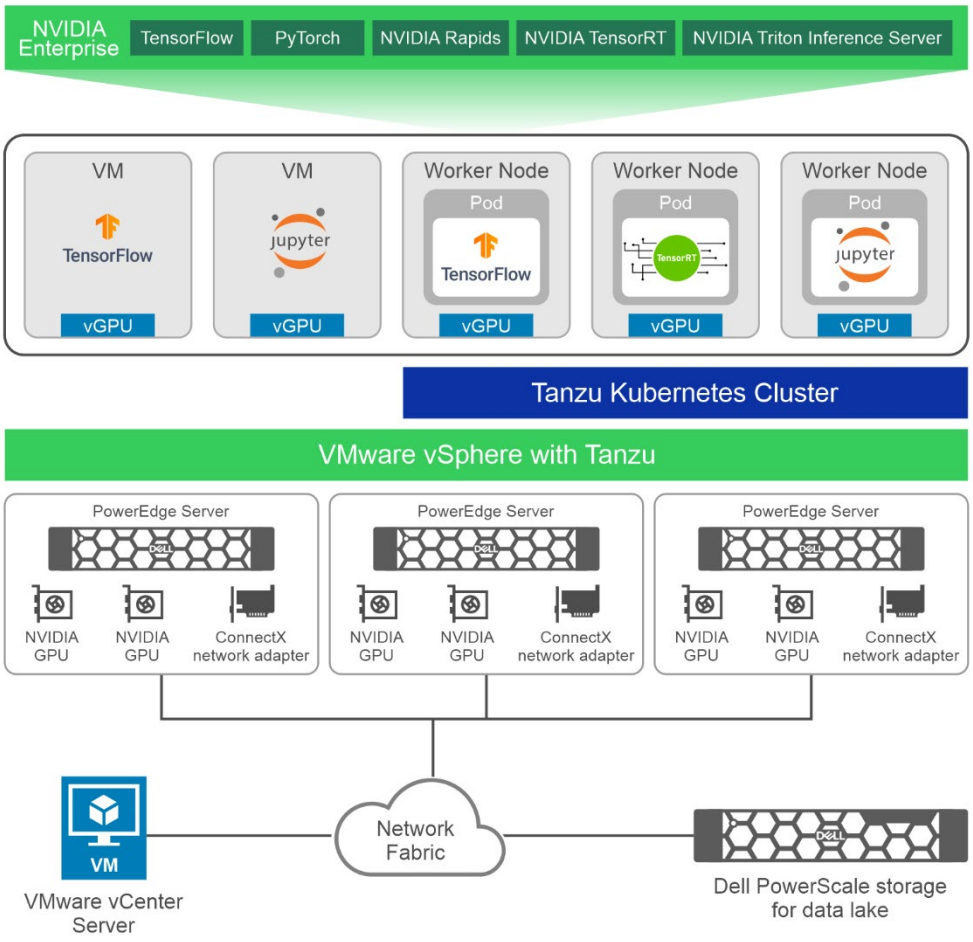

Figure 3. High-level architecture with PowerEdge R760 servers and PowerScale storage

Key aspects of this validated design include:

- **Compute server**—The PowerEdge R760 and R7625 servers are part of this validated design.
- **GPUs**—NVIDIA A100 and A30 GPUs can be used for AI and machine learning. We recommend the A100 GPU for large neural network training models that require high performance and the A30 GPU for AI inference and mainstream enterprise workloads. The number of GPUs supported in a server depends on the server model as shown in Table 2.

- **Storage**—vSAN is the recommended storage for VMs. We recommend PowerStore storage for data lake storage, that is, storing data that are required for neural network training. PowerScale storage can also be used both for storing data for AI workloads in an NFS partition.
- **Network infrastructure**—Customers can have either a 25 Gb Ethernet network infrastructure or a 100 Gb Ethernet network infrastructure. We recommend 25 GbE for workloads that can use existing network infrastructure without needing to invest in 100 Gb network infrastructure. This design is suited for neural network training jobs that can run on a single node (using at most two GPUs), and for model development and inference jobs that take advantage of GPU partitioning.

  We recommend 100 GbE for workloads that require large-scale model training using large datasets (typically high-resolution video or image-based datasets).
- **Virtualization and container orchestration**—GPUs can be virtualized and made available to VMs running on VMware ESXi servers deployed on PowerEdge servers. For containerized workloads, Tanzu Kubernetes Grid service is enabled on vSphere cluster. Kubernetes worker nodes can be created with virtual GPU resources. AI workloads can be deployed as pods on deployment services running on Tanzu Kubernetes clusters.
- **Management with VMware vCenter**—VMware vCenter Server can be deployed as a VM in your data center. vCenter is critical to the deployment, operation, and maintenance of a vSAN environment. For this reason, Dell Technologies recommends that vCenter be deployed on a highly available management cluster, which exists outside of the compute cluster.

For more information about the validated design, including detailed recommended configurations, design considerations, and deployment overview, see the Virtualizing GPUs for AI with VMware and NVIDIA design guide.

Validation results

Based on the architecture described in this white paper, the Dell Integrated Solutions team used the Dell Technologies HPC & AI Innovation Lab to conduct validation and performance studies of virtualized GPUs on VMware vSphere.

We also characterized the performance of MIGs using both ResNet model training and inference. The results also showed increased GPU utilization when using GPU partitions for workloads, such as inference, that does not require an entire GPU.

For more information about the performance studies, see the Virtualizing GPUs for AI with VMware and NVIDIA design guide.

Conclusion

Data science professionals have been pushing the limits of IT infrastructure and creating shadow IT environments for years. Dell Technologies, VMware, and NVIDIA have responded with numerous advancements in computation, data storage options, and high-speed networking to meet that challenge with fully integrated solutions. In this white paper, we show how you can manage these advances in hardware technology more efficiently, simultaneously preserving the impressive performance gains in a virtualized environment using the latest version of VMware virtualization software.

Nowhere have the performance advancements for data science been as rapid as the development of hardware accelerators based on the technology once used primarily for graphics processing. The newest generation of Ampere GPUs from NVIDIA provides up to 20-times higher performance over the prior generation of NVIDIA GPUs. The NVIDIA A100 GPU can be partitioned up to seven GPU instances to dynamically adjust to shifting demands. With NVIDIA AI Enterprise software suite, data scientists have the right tools and frameworks for their entire data science life cycle.

The Dell Validated Design for Virtualizing GPUs for AI with VMware and NVIDIA provides a combination of virtualized GPUs, AI enterprise software, and container orchestration that gives IT professionals consistent tools and processes to manage their entire data center, while allowing the data scientist to focus on data preparation and model development without worrying about the infrastructure.

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on the solution and the solution documentation. Contact the Dell Technologies Solutions team by email.

Note: For links to additional documentation for this solution, see the Dell Technologies Solutions Info Hub for Artificial Intelligence and Data Analytics Workloads.

References

Dell Technologies links

The following Dell Technologies links and documentation provide additional and relevant information. Access to these documents depends on your login credentials. If you do not have access to a document, contact your Dell Technologies representative.

- HPC & AI Innovation Lab
- Dell PowerSwitch Data Center Switches
- Dell Technologies Customer Solution Centers
- Virtualizing GPUs for AI with VMware and NVIDIA Design Guide
- Virtualizing GPUs for AI with VMware and NVIDIA Implementation Guide
- Dell Technologies Solutions Info Hub for Artificial Intelligence and Data Analytics Workloads

VMware links

The following VMware links and documentation provide additional and relevant information:

- vSphere 8 - What's New
- VMware vSAN Design Guide
- Blog: vSphere 7 with Multi-Instance GPUs (MIG) on the NVIDIA A100 for Machine Learning Applications – Part 1 and Part 2
- VMware Tanzu Overview
- VMware Tanzu for Kubernetes Operations Documentation
- VMware vSphere Product Line Comparison

NVIDIA links

The following NVIDIA links and documentation provide additional and relevant information:

- NVIDIA AI Enterprise Documentation, for comprehensive documentation about NVIDIA AI Enterprise
- NVIDIA A100 Tensor Core GPU
- NVIDIA Multi-Instance GPU User Guide
- NVIDIA-Certified Systems for Enterprises
- NVIDIA GPU Cloud (NGC) Overview
- NVIDIA Multi-Instance GPU and NVIDIA Virtual Compute Server Technical Brief
- NVIDIA AI Enterprise Packaging, Pricing, and Licensing Guide