GPU Resource Sharing for VDI and AI Workloads in a Healthcare Environment

Design Guide

Abstract

This Dell Technologies Validated Design Guide describes technical considerations and best practices for accelerating both VDI and AI healthcare workloads on a shared infrastructure stack by integrating VMware Horizon 8 software, Dell EMC VxRail, and NVIDIA virtual GPU technology.

Dell Technologies Solutions

Dell Technologies
Validated Design

Part Number: H18599.1
December 2021

Notes, cautions, and warnings

ⓘ **NOTE:** A NOTE indicates important information that helps you make better use of your product.

△ **CAUTION: A CAUTION indicates either potential damage to hardware or loss of data and tells you how to avoid the problem.**

⚠ **WARNING: A WARNING indicates a potential for property damage, personal injury, or death.**

Contents

1

Executive Summary

This chapter presents the following topics:

Topics:

Business challenge

Organizations in the healthcare industry are adopting virtual desktop infrastructure (VDI) to benefit from centralized management, better security and compliance, and worker mobility. In the healthcare industry, graphics processing units (GPUs) are used in VDI to enhance the quality of the visual experience for medical professionals in clinical imaging. The healthcare industry uses GPUs for use cases that require advanced graphic acceleration, such as PACS imaging technology and other graphics intensive use cases. These non-diagnostic use cases can include remote radiology, remote cardiology, and digital pathology.

The healthcare industry, such as academic medical centers and research organizations, also benefits from the use of GPUs for running artificial intelligence (AI) workloads, including learning and inferencing, that use parallel computation. The healthcare industry is poised to reap the benefits of AI to improve patient outcomes, reduce costs, and speed up diagnoses.

However, organizations in the healthcare industry need to find a cost-effective and optimal method for using GPUs. Not all workloads and use cases use GPUs to the fullest capacity, and processing demand can vary with time of day, depending on the workload. Healthcare organizations are looking for a solution that provides optimal and shared use of GPUs while running mixed workloads such as VDI and AI and that increases overall GPU utilization in the data center. Using this solution eliminates the cost of investing in dedicated GPU hardware for each workload.

Solution overview

Because of the broad capabilities of VDI from a security, operational simplicity, and total cost of ownership (TCO) perspective, it is often an integral part of the IT infrastructure of healthcare organizations. VDI environments are used by a range of organizational users, which increasingly includes users with requirements for graphics-rich environments that require GPU resources. GPU environments are also widely deployed for AI activities such as machine learning and deep learning. These AI activities are seeing a significant increase in usage across multiple settings, with healthcare, which is an intensely data-driven field, being one of the primary settings.

This design guide presents the results of work carried out by the Dell Technologies VDI Solutions team to demonstrate the sharing of GPU resources in a healthcare setting between graphics-accelerated VDI, including Digital Imaging and Communications in Medicine (DICOM) X-Ray viewing, and healthcare-related AI. The tests included the prediction of diseases from public-domain chest X-Ray datasets released by the US National Institutes of Health (NIH).

This process of sharing GPU resources is sometimes called “compute cycle harvesting” or “VDI by day, compute by night.” In this guide we have chosen to describe it as GPU resource sharing for VDI and AI workloads in a healthcare environment. The architecture and the performance testing described in this document for a shared VDI and AI infrastructure include scenarios where the VDI and AI workloads are run as dual workloads simultaneously, and scenarios where VDI and AI are run as single workloads consecutively. It should be noted that the mechanism for switching between the workloads when run consecutively, whether done manually or with automation, is not within the scope of this paper.

By implementing a common platform for VDI and AI workloads to increase GPU utilization, organizations in the healthcare industry can improve IT efficiency and reduce TCO. Dell Technologies VDI Solutions offers a tested and validated VMware Horizon solution based on the hyperconverged Dell EMC VxRail platform, configured with NVIDIA RTX 8000 GPUs that can run mixed workloads including VDI and AI. By sharing the GPU cards between the workloads to meet various business requirements, around the clock, you get the maximum return on investment (ROI) from your GPU investment.

This design guide provides:

- **A common infrastructure stack for VDI and AI**—We designed a solution stack based on a density-optimized configuration with Dell EMC VxRail V570F, NVIDIA RTX 8000 GPUs, Dell EMC PowerSwitch networking, VMware vSAN, VMware vSphere, and VMware Horizon, and NVIDIA RTX Virtual Workstation (vWS) software to support the combination of VDI and AI activities.
- **Optimal sizing for VDI**—We performed sizing for the VDI environment using Login VSI testing with a customized Login VSI healthcare workload.
- **A customized Login VSI workload for healthcare**—We customized a Login VSI Knowledge Worker workload and validated it with viewing activities based on clinical images from the NIH database, using MicroDicom as the image viewing software.
- **AI frameworks**—We designed and implemented AI frameworks based on standard toolsets, including TensorFlow, in collaboration with the Dell Technologies AI Solutions team.
- **AI workloads (learning and inferencing)**—We performed AI deep-learning activities, including learning and inferencing, to validate the AI/deep-learning capabilities of the environment.

Dell Technologies has extensive experience in VDI solutions and a broad selection of compute, storage, and networking for VDI. Tested and validated VMware Horizon-based Dell Technologies VDI Solutions configured with NVIDIA virtual GPUs offer exceptional graphics performance and predictable cost. Dell Technologies provides a single-vendor support experience.

As the hyperconverged platform for the solution, Dell EMC VxRail delivers a highly differentiated, turnkey experience, with fully automated lifecycle management, to offer a fast and simple path to IT outcomes. VMware Cloud Foundation leverages native integration between VxRail Manager and VMware Software-Defined Data Center (SDDC) Manager to build a fully automated deployment of SDDC architecture. VxRail provides the capability to host up to three NVIDIA RTX 8000 GPUs per VxRail, providing a common platform to run both VDI graphic acceleration and AI compute workloads that share GPU cards.

Solution benefits

The key benefits of a GPU-sharing environment include:

- **Lower TCO**—A solution for running both VDI and AI workloads that shares GPU hardware significantly improves the efficiency and utilization of the resources and therefore lowers the TCO of the infrastructure. The solution provides a platform that increases GPU utilization in data centers and offers optimal utilization of GPUs, thereby eliminating the cost to invest in new GPUs.
- **Improved mobility of personnel**—A GPU-accelerated VDI environment offers mobility for healthcare professionals, untethering them from physical PCs, workstations, and offices, improving productivity and user experience, while reducing IT costs.
- **VDI graphics and AI compute performance**—The NVIDIA RTX 8000 GPU, powered by the NVIDIA Turing architecture and the NVIDIA RTX platform, combines unparalleled performance and memory capacity to deliver a powerful graphics card solution for professional workflows. The RTX 8000 is suitable for deep-learning matrix arithmetic and computations, and comes with 48 GB of DDR6 memory, which is recommended for researching extra-large computational models. This GPU features 72 RT Cores for real-time ray tracing and 576 Tensor Cores for AI enhanced workflows, resulting in over 130 TFLOPS of deep-learning performance.

Document purpose

This document describes the business challenge, approach, and benefits of implementing an efficient infrastructure for the sharing of GPU resources for both VDI and AI workloads in a healthcare environment. The guide describes the solution architecture and the key hardware and software components of the architecture, and provides guidance on designing, managing, and scaling a shared VMware Horizon 8 environment on Dell EMC VxRail. It also summarizes the performance testing that the Dell Technologies VDI Solutions team performed and describes possible extensions to the architecture such as storage options.

Audience

This guide is for decision makers, managers, architects, developers, and technical administrators of Information Technology (IT) and Operational Technology environments, particularly in the healthcare field, who want to understand how to design and implement an infrastructure that supports the sharing of GPU resources for VDI and AI workloads.

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 VDI Solutions team by email or provide your comments by completing our documentation survey.

Authors: Dell Technologies VDI Solutions team

ⓘ **NOTE:** This website provides additional documentation for Dell Technologies VDI Solutions: Virtual Desktop Infrastructure.

2

Solution architecture

This section provides an architecture overview and guidance for managing and scaling a VMware Horizon 8 environment on Dell EMC VxRail when sharing VDI and AI workloads on the same platform.

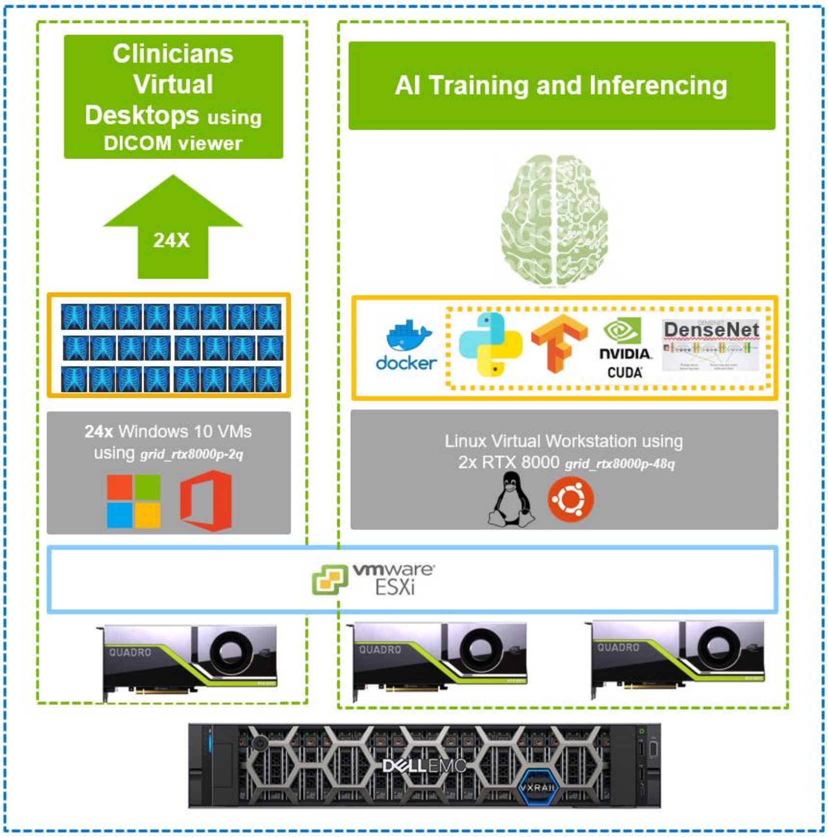
Topics:

Architecture overview

The following figure shows the architecture of the validated solution, focusing on the compute layer and how the graphics components are divided between the VDI and AI workloads.

This architecture aligns with the VMware Horizon block/pod design in which a pod is divided into multiple blocks as outlined here. Each block is made up of one or more vSphere clusters and an associated vCenter server appliance. The figure demonstrates how a single compute node is logically divided to support up to 24 users and a single AI virtual machine (VM) per VxRail.

The following figure shows a design where VDI and compute workloads run in parallel. When these workloads run consecutively, each workload has 3 GPUs assigned to it when running.



Clinicians Virtual Desktops using DICOM viewer
24X
24x Windows 10 VMs using grid_rtx8000p-2q
AI Training and Inferencing
docker
NVIDIA CUDA
DenseNet
Linux Virtual Workstation using 2x RTX 8000 grid_rtx8000p-48q
vmware ESXi
QUADRO
QUADRO
QUADRO
DELL EMC
VXRAIL

Figure 1. Logical division of a single compute node

The deployment option for this solution supports all of the cloning techniques available from VMware: full, linked, and instant.

A VMware vSAN-enabled vSphere Cluster can have a maximum of 64 nodes and 6,400 VMs per cluster. To expand from this limit, you can add clusters and balance the VMs and nodes across the new clusters.

NOTE: vSphere 7.0 Update 1 can scale up to 96 nodes per cluster, but the vSAN limitation is still 64.

Scaling the solution

Solutions based on Dell EMC VxRail provide flexibility as you scale, reducing the initial and future TCO. Add additional physical and virtual servers to the server pools to scale horizontally. Add additional resources such as faster CPUs, more memory or more disk capacity to scale vertically. The following figure shows an abstract view of how the solution can scale as you add VxRail to the solution to support additional users and AI virtual machines when supporting a mixed workload.

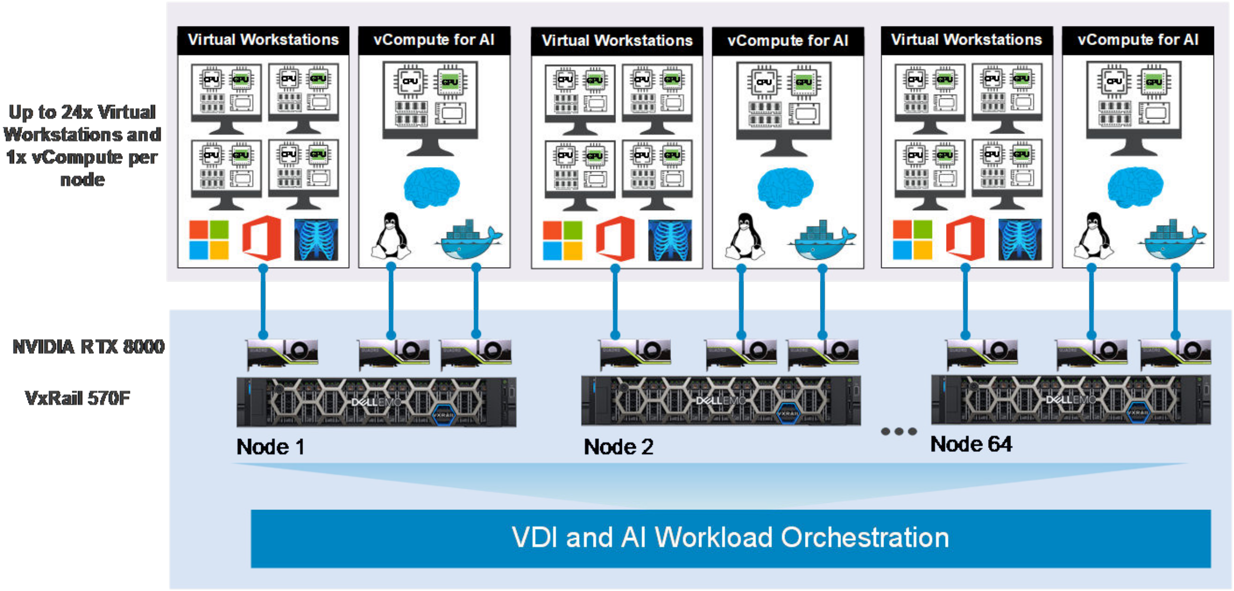


Figure 2. Scaling the solution

In the parallel use case, when the environment is reaching peak VDI user utilization, the workload orchestrator can power off the AI VMs and bring up additional VDI VMs to support the user workload at that time. This is demonstrated in the following figure where all the compute and related GPU resources are dedicated to the Virtual Workstation VDI workload.

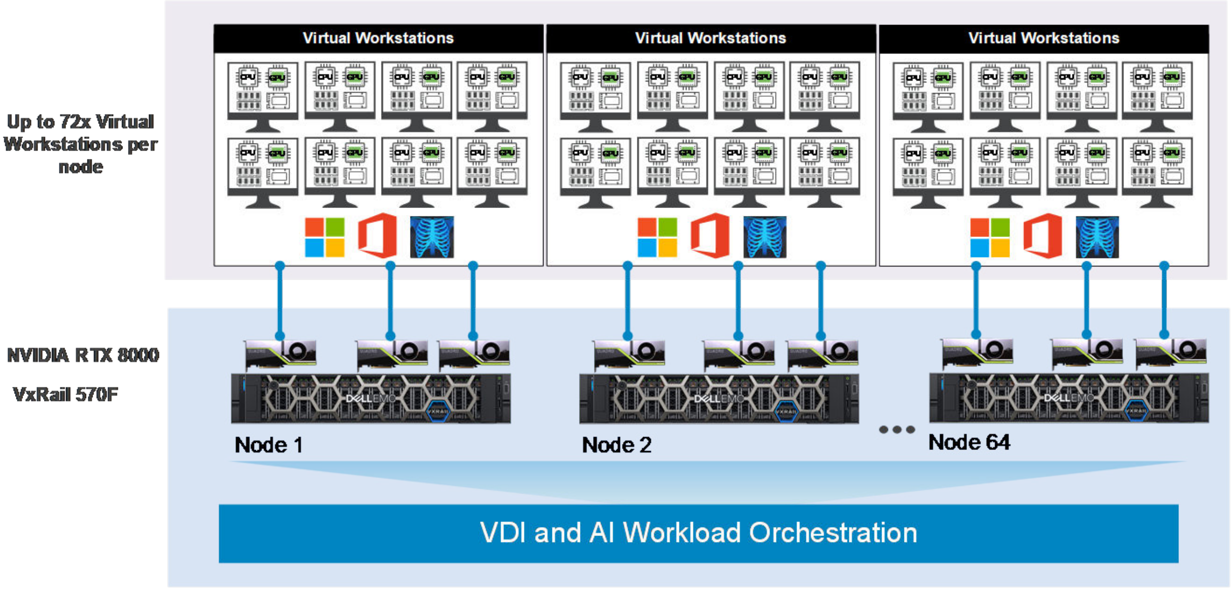


Figure 3. All compute and related GPU resources dedicated to VDI workload

Scaling out

Each component of the solution architecture scales independently, depending on the required number of supported users. You can add nodes at any time to expand the vSAN SDS pool in a modular fashion. The scaling limit for a vSphere cluster is restricted by the limits of vSAN at 64 nodes per block.

The boundary for a Horizon block is the vCenter. The number of virtual machines a vCenter (and therefore a block) can host depends on the type of Horizon 8 VMs being used. The recommended limits for a Horizon block at the time of writing are as follows:

- 12,000 full clone VMs
- 12,000 instant clone VMs
- 4,000 linked clone VMs

For the latest sizing guidance, see VMware Configuration Maximums. Additionally, see the VMware Knowledge Base article VMware Horizon 7 sizing limits and recommendations (2150348).

This design guide presents the use of instant clones in the following figures. We used design limits of 4,000 instant-clone VMs per block and up to 12,000 VMs per pod. VMware Horizon pools have a limit of 4,000 VMs, so additional pools are necessary when scaling above that.

The VMware Horizon management infrastructure and Knowledge-User VMs are located on separate vSphere clusters. Four management nodes are a suitable configuration to start with to provide redundancy and self-healing and can be scaled as appropriate. This logical model can be implemented using VMware Validated Designs or using VMware Cloud Foundation.

The following figure shows a 4,000-user pod supporting up to 4,000 Knowledge-User VMs with a single resource block and two vSphere clusters. With the above limits in mind, 56 compute nodes with 72 Knowledge-User VMs per node across two vSphere clusters would reach the maximum number of VMs for the block.

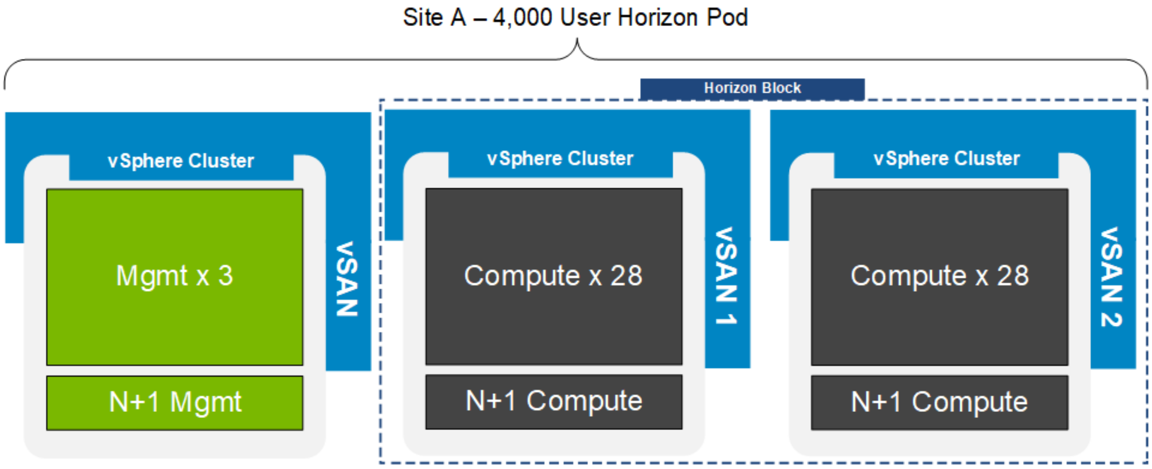


Figure 4. 4,000 user Horizon pod

The following figure shows a scale-out to a 12,000-user Horizon pod with three 4,000-user resource blocks and six vSphere clusters.

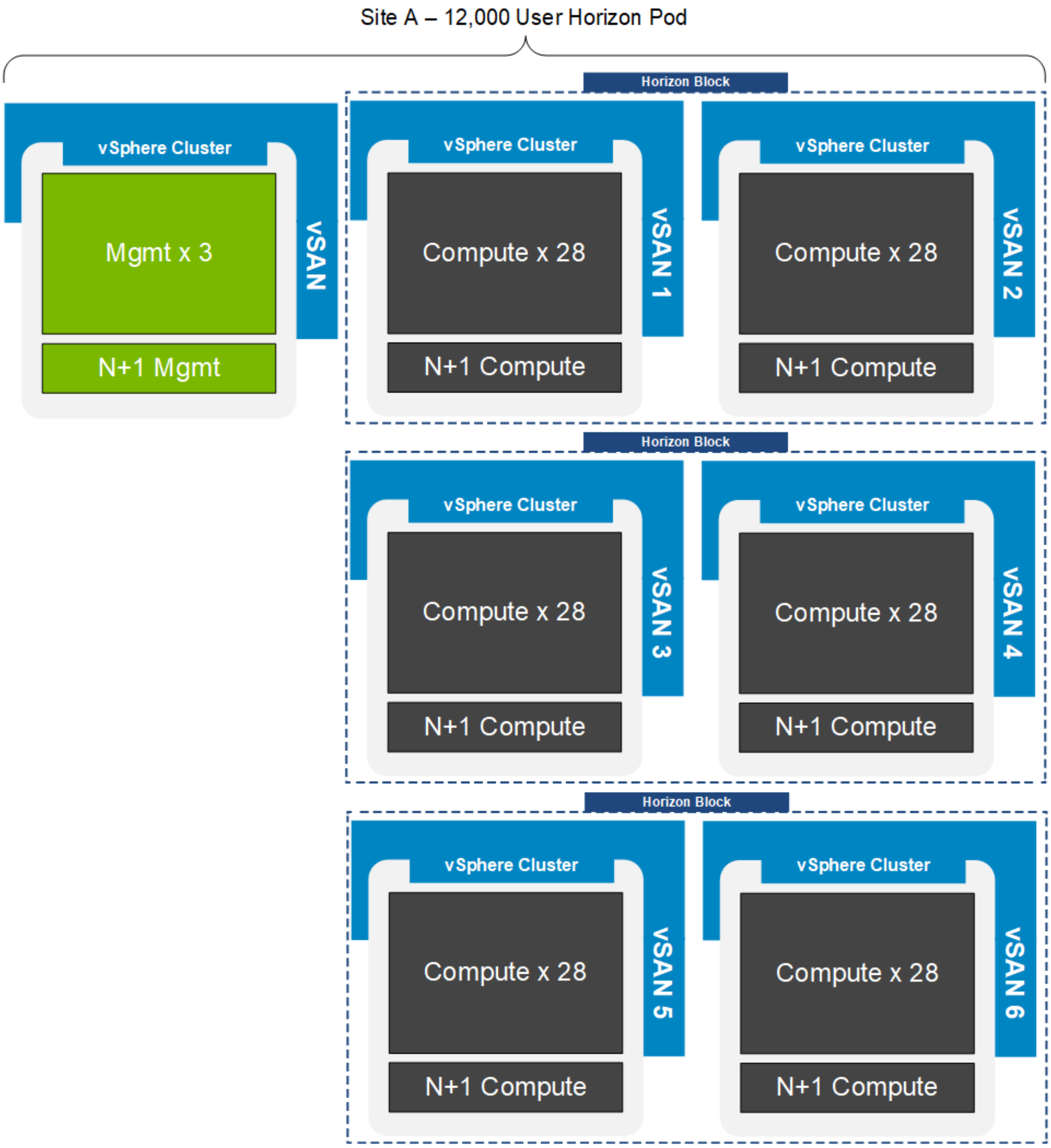


Figure 5. 12,000 user Horizon pod

Scaling up

Dell Technologies recommends a validated disk configuration for general-purpose VDI. These configurations leave drive slots available for future vertical expansion and ensure that you protect your investment as new technology transforms your organization.

NOTE: These configurations can accept additional or faster processors or memory than is stated here.

For more information about Horizon pod/block architecture and scaling, see the VMware Workspace ONE and VMware Horizon 7 Enterprise Edition On-premises Reference Architecture Guide.

Enterprise solution pods

The compute, management, and storage layers are converged into a single VxRail, hosting VMware vSphere. The number of nodes that are supported for a vSAN-enabled vSphere 7.0 cluster, which is 64, determines the recommended boundaries of an individual vSphere cluster.

Dell Technologies recommends that the VDI management infrastructure nodes be separated from the compute resources. In smaller environments, management and compute can be placed in the same vSphere HA cluster. In these smaller environments, a logically designated management node can optionally also be used for VDI VMs with an expected reduction of 30 percent of host resources for these nodes only. The 30 percent reduction accounts for management VM resource reservations and should be factored in when sizing. Compute hosts can be used interchangeably for Horizon Apps hosted applications and desktops, as required.

This design guide describes a single-site or single data center design. For multi-site or disaster recovery (DR) configurations, see the VMware Horizon Multi-Site Reference Architecture.

3

Key components

This section describes the key hardware and software components of the solution.

Topics:

- Dell Technologies VDI Solutions Optimized Configurations
- VMware vSAN software-defined storage
- NVIDIA Virtual GPU
- Physical network components
- VMware vSphere
- VMware Horizon
- Client components

Dell Technologies VDI Solutions Optimized Configurations

For graphics-intensive desktop deployments, we recommend the VDI-optimized 2U servers that support GPU hardware.

We have designated common configurations as Management-optimized, Density-optimized, and Virtual Workstation. These designations are referenced throughout this document and are outlined in the following table.

Table 1. Common configurations

Configuration	CPU	RAM	Disk	GPU	Description
Management-optimized	2 x Intel Xeon Silver 4214 (12 core @ 2.2 GHz)	192 GB (12 x 16 GB @ 2400 MHz)	4 TB + (Capacity)	None	Offers a scalable and value targeted configuration that meets the required compute and I/O demands
Density-optimized	2 x Intel Xeon Gold 6248 (20 core 2.5 GHz)	768 GB (12 x 64 GB @ 2933 MHz)	8 TB + (Capacity)	Up to 3 x full length, dual width (FLDW) (for example, the RTX 8000) Up to 6 x full length, single width (FLSW)	Offers an abundance of high-performance features and components selected to maximize user density
Virtual Workstation	2 x Intel Xeon Gold 6254 (18 core @ 3.1 GHz)	384 GB (12 x 32 GB @ 2933 MHz)	6 TB + (Capacity)	Up to 3 x FLDW Up to 6 x FLSW	Offers even higher performance at the tradeoff of user density. Typically for ISV or high-end graphics workloads.

VMware vSAN software-defined storage

VMware vSAN is available in all-flash or hybrid configurations.

After vSAN is enabled on a cluster, all disk devices presented to the hosts are pooled together to create a shared datastore that is accessible by all hosts in that vSAN cluster. You can then create VMs and assign storage policies to them. The storage policy dictates availability and performance.

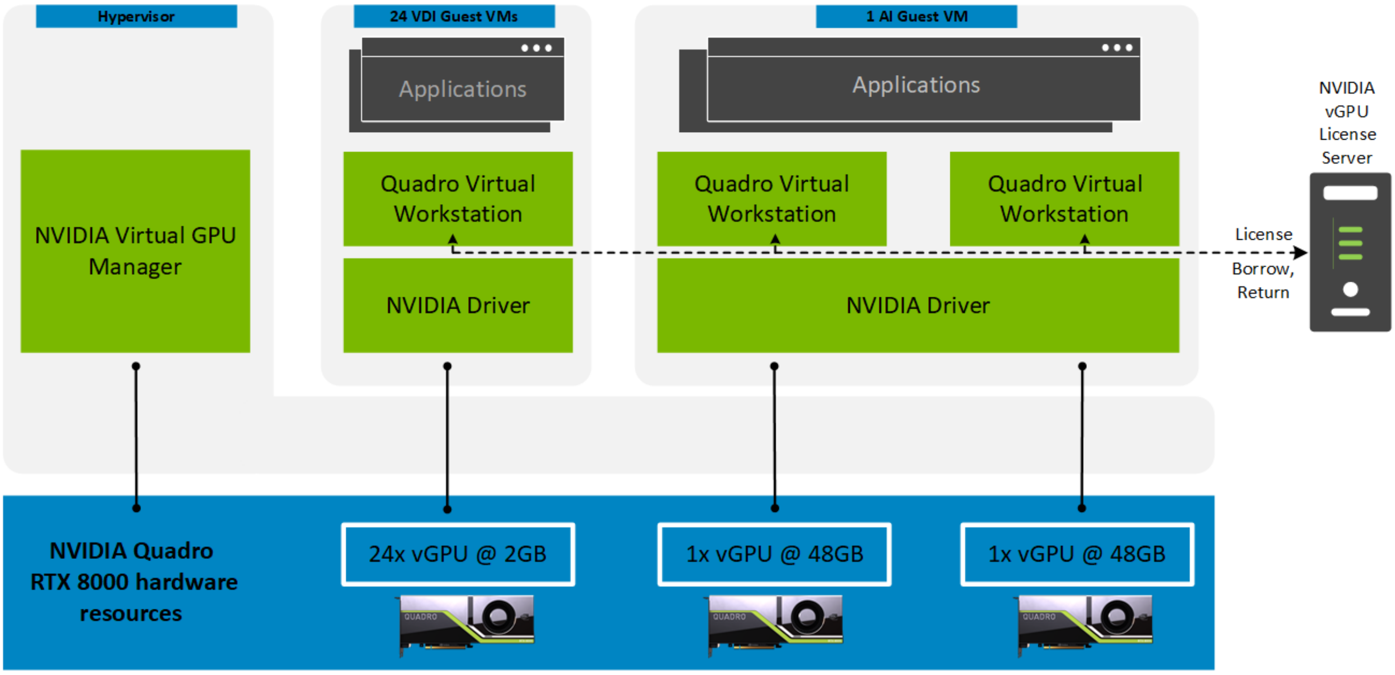
vSAN provides the following configuration options:

- **All-flash configuration**—Uses flash for both the cache tier and capacity tier to deliver enterprise performance and a resilient storage platform. In this configuration, the cache tier is fully dedicated to writes, allowing all reads to come directly from the capacity tier. The cache device protects the endurance of the capacity tier. All-flash configured solutions enable data reduction features to extend the capacity tier.
- **Hybrid configuration**—Uses flash-based devices for the cache tier and magnetic disks for the capacity tier. Hybrid configurations are ideal for clients looking for higher volume in the capacity tier. The performance of SSD and magnetic spinning disks is comparable in VDI applications if you use a sufficient number of magnetic spinning disks.

NVIDIA Virtual GPU

NVIDIA Virtual GPU (vGPU) brings the full benefit of NVIDIA hardware-accelerated graphics to virtualized solutions. This technology provides exceptional graphics performance for virtual desktops equivalent to local computers when sharing a GPU among multiple users. The following figures show how NVIDIA vGPU technology was used within this design guide. When sharing the compute and graphics resources, a single RTX 8000 GPU was used with 2GB vGPU framebuffers to support up to 24 users. The AI Guest VM on each VxRail in the cluster was assigned two vGPUs each with 48 GB of framebuffer that we mapped to the remaining two RTX 8000 GPUs within the VxRail.

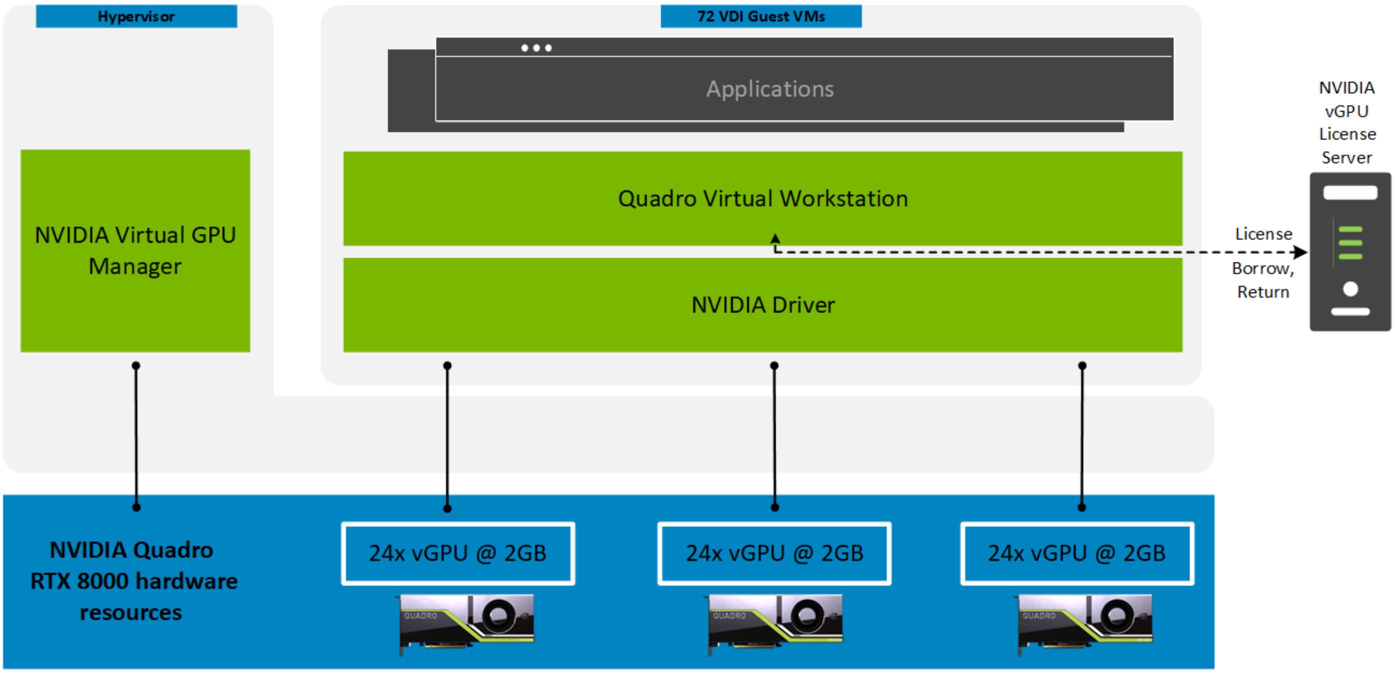
The following figure shows a shared VDI and AI use case:



Hypervisor
24 VDI Guest VMs
1 AI Guest VM
Applications
Applications
NVIDIA Virtual GPU Manager
Quadro Virtual Workstation
Quadro Virtual Workstation
Quadro Virtual Workstation
NVIDIA Driver
NVIDIA Driver
NVIDIA vGPU License Server
License Borrow, Return
NVIDIA Quadro RTX 8000 hardware resources
24x vGPU @ 2GB
1x vGPU @ 48GB
1x vGPU @ 48GB

Figure 6. Shared VDI and AI use case

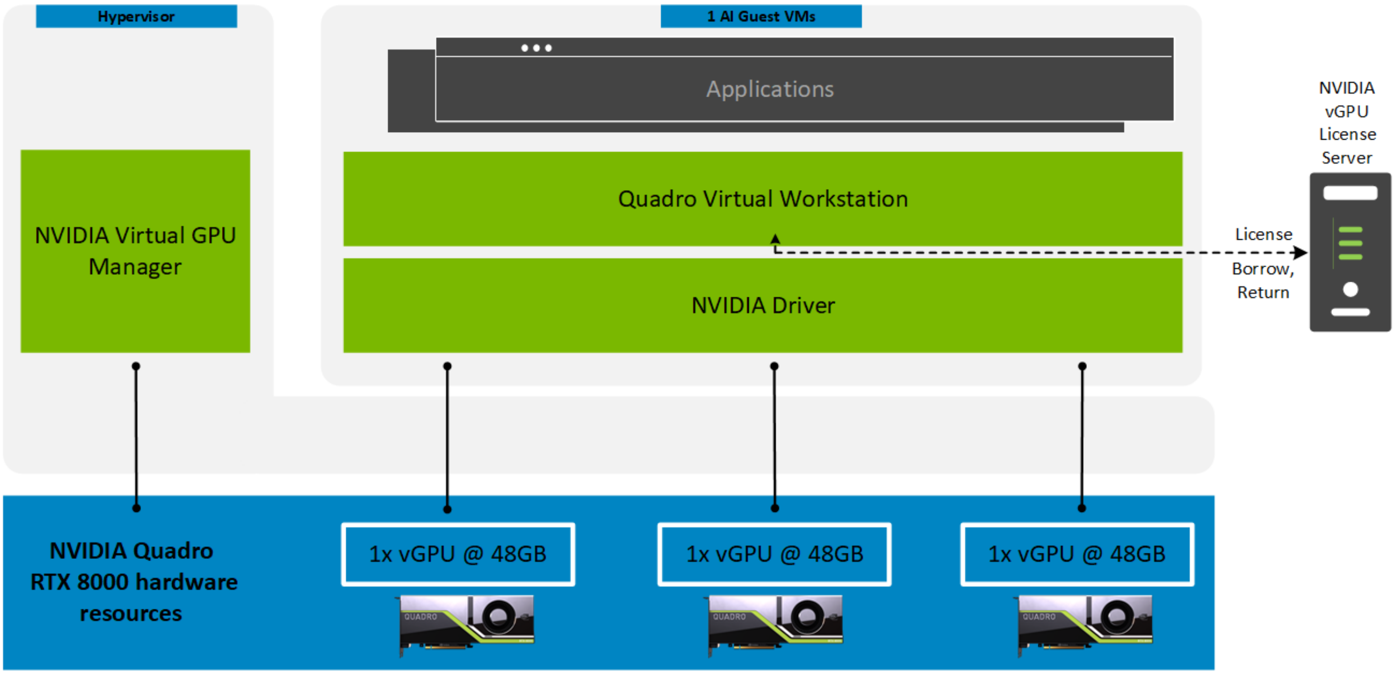
The following figure shows a VDI only use case:



Hypervisor
72 VDI Guest VMs
Applications
NVIDIA Virtual GPU Manager
Quadro Virtual Workstation
NVIDIA Driver
NVIDIA vGPU License Server
License Borrow, Return
NVIDIA Quadro RTX 8000 hardware resources
24x vGPU @ 2GB
24x vGPU @ 2GB
24x vGPU @ 2GB

Figure 7. VDI only use case

The following figure shows an AI only use case:



Hypervisor
1 AI Guest VMs
Applications
NVIDIA Virtual GPU Manager
Quadro Virtual Workstation
NVIDIA Driver
NVIDIA vGPU License Server
License Borrow, Return
NVIDIA Quadro RTX 8000 hardware resources
1x vGPU @ 48GB
1x vGPU @ 48GB
1x vGPU @ 48GB

Figure 8. AI only use case

NVIDIA vGPU is the industry's most advanced technology for sharing true GPU hardware acceleration between multiple virtual desktops without compromising the graphics experience.

NVIDIA vGPU offers multiple software variants to enable graphics for different virtualization techniques. This design guide uses the **NVIDIA RTX Virtual Workstation (RTX vWS)** software variant. RTX vWS is designed to provide workstation-grade performance in a virtual environment with support for up to four 4K or 5K monitors or up to two 8K monitors. We used this software to support a VDI workload with the appropriately sized frame buffer and features, and an AI workload that could take advantage of CUDA. CUDA is a parallel computing platform and programming model developed by NVIDIA that enables dramatic increases in computing performance by harnessing the power of GPUs.

Additional variants that are not used in this design guide are:

- **Virtual Applications**—Designed to deliver graphics-accelerated applications using Remote Desktop Session Host (RDSH)
- **Virtual PC**—Designed to provide full virtual desktops with up to dual 4K monitor support or single 5K monitor support
- **NVIDIA Virtual Compute Server (vCS)**—Designed to accelerate server virtualization so that the most compute-intensive workloads, such as artificial intelligence, deep learning, and data science, can be run in a VM

This design guide was configured with the following NVIDIA GPUs:

- **NVIDIA RTX 8000**—Select the Turing-based RTX 8000 for virtualized graphics performance for professional graphics and rendering workloads. The RTX 8000 has 48 GB of graphics frame buffer per card. Add up to three RTX 8000 GPU cards into your VxRail V570F to enable up 144 GB of frame buffer.

For additional GPU options, see the VxRail design guide on the VDI Info Hub.

Physical network components

Dell Technologies VDI Solutions allow for flexibility in networking selections. Although several other choices are available, VDI validations have been performed with the following hardware:

- **Dell EMC Networking S5248F (25 GbE ToR switch)**—The PowerSwitch S5248F switch provides optimum flexibility and cost-effectiveness for demanding compute and storage traffic environments. This ToR switch features 48 x 25 GbE SFP28 ports, 4 x 100 GbE QSFP28 ports, and 2 x 100 GbE QFSP28-DD ports. The S5248F-ON also supports Open Network Install Environment (ONIE) for zero-touch installation of alternate network operating systems.
- **Dell EMC Networking S4048 (10 GbE ToR switch)**—The PowerSwitch S4048 switch optimizes your network for virtualization with a high-density, ultra-low-latency ToR switch that features 48 x 10 GbE SFP+ and 6 x 40 GbE ports (or 72 x 10 GbE ports in breakout mode), and up to 720 Gbps performance. The S4048-ON also supports ONIE.

For more information on these switches, see Dell EMC Networking S-Series 10GbE switches and Dell EMC Networking S-Series 25GbE switches.

Designed for linear scaling, VxRail uses a leaf-spine network architecture, consisting of two network tiers: an L2 leaf and an L3 spine that is based on 40 GbE and non-blocking switches. This architecture maintains consistent performance without any throughput reduction.

VMware vSphere

VMware vSphere provides a powerful, flexible, and secure foundation for business agility that accelerates the digital transformation to cloud computing and promotes success in the digital economy.

vSphere provides the following benefits for VDI applications:

- **Improved Appliance Management**—The vCenter Server Appliance Management interface provides CPU and memory statistics, network and database statistics, disk space usage, and health data. These features reduce reliance on a command-line interface for simple monitoring and operational tasks.
- **VMware vCenter Server native high availability**—This solution for vCenter Server Appliance consists of active, passive, and witness nodes that are cloned from the existing vCenter Server instance. The vCenter HA cluster can be enabled, disabled, or destroyed at any time. Maintenance mode prevents planned maintenance from causing an unwanted failover. The vCenter Server database uses Native PostgreSQL synchronous replication, while key data outside the database uses a separate asynchronous file system replication.
- **Backup and Restore**—Native backup and restore for the vCenter Server Appliance enables users to back up vCenter Server appliances directly from the VAMI or API. The backup consists of a set of files that is streamed to a selected storage device using SCP, HTTP(S), or FTP(S) protocols.
- **VMware vSphere HA Support for NVIDIA vGPU-configured VMs**—vSphere HA protects VMs with the NVIDIA vGPU shared pass-through device. In the event of a failure, vSphere HA tries to restart the VMs on another host that has an identical NVIDIA vGPU profile. If no available healthy host meets this criterion, the VM fails to power on.
- **VMware vSAN Enterprise Edition**—Includes all-flash space-efficiency features (deduplication, compression, and erasure coding), software-defined, data-at-rest encryption, and stretched clusters for cost-efficient performance and greater hardware choice.
- **VMware Log Insight**—Provides log management, actionable dashboards, and refined analytics, which enable deep operational visibility and faster troubleshooting.

ⓘ **NOTE:** vSphere Enterprise Plus, or vSphere Desktop, which is used for deploying desktop virtualization, is recommended to support NVIDIA graphics cards for VDI use cases.

VMware Horizon

The architecture that this guide describes is based on VMware Horizon 8, which provides a complete end-to-end solution delivering Microsoft Windows virtual desktops to users on a wide variety of endpoint devices. Virtual desktops are dynamically assembled on demand, providing users with clean, yet personalized, desktops each time they log in.

VMware Horizon 8 provides a complete virtual desktop delivery system by integrating several distributed components with advanced configuration tools that simplify the creation and real-time management of the virtual desktop infrastructure.

For more information, see the Horizon Resources page and the Horizon License FAQ.

The core Horizon components include:

- **Horizon Connection Server (HCS)**—Installed on servers in the data center. The HCS brokers client connections, authenticates users, entitles users by mapping them to desktops and/or pools, establishes secure connections from clients to desktops, supports single sign-on, and sets and applies policies.
- **Horizon Administrator**—Provides administrator functions such as deployment and management of Horizon desktops and pools, setting and controlling user authentication, and more.
- **Horizon Agent**—Installed on all VMs, physical machines, and Terminal Service servers that are used as a source for Horizon desktops. On VMs, the agent is used to communicate with the Horizon client to provide services such as USB redirection, printer support, and more.
- **Horizon Client**—Installed on endpoints for creating connections to Horizon desktops that can be run from tablets, Windows, Linux, or Mac desktops or laptops, thin clients, and other devices.
- **Unified Access Gateway**—Provides a way to securely deliver connections that require a higher level of security to access, such as remote connections from the Internet.
- **Horizon Portal**—Provides access to links for downloading full Horizon clients. Enable the HTML access feature to run a Horizon desktop inside a supported browser.
- **vCenter Server**—Provides centralized management and configuration to the entire virtual desktop and host infrastructure. It facilitates configuration, provisioning, and management services.

Horizon clone technology

VMware Horizon 8 offers the following methods for cloning desktops:

- **Full clones**—Full clones are typically used in environments where dedicated resources are assigned to specific users. Full clones are typically not ideal for large-scale VDI deployments because full copies have no connection to the original VM. Updates must be performed on each VM with this approach. Additionally, full clones are not space-efficient as they will duplicate much of the same data to all of the VMs. Space efficiency technologies may be enabled at the storage layers, but that may incur additional CPU overhead and reduce VM and user density per node as well as overall performance.
- **Instant clones**—As of Horizon 8, instant clones are available with all licensing levels. These clones include enhancements and new capabilities such as “Smart Provisioning” to help reduce storage requirements and costs. This technology can provision a VM the instant that a user requests one. The result is a far easier approach to operating system updates and patch management, because the VM is created close to the login time. You can use the combination of Just-in-Time Management Platform (JMP) features such as App Volumes and Dynamic Environment Manager to emulate persistence. For more information on instant clones, see the VMware Horizon 7 Instant-Clone Desktops and RDSH Servers White Paper.
- **Linked clones**—Linked clones require fewer storage resources than full clones. This technology is appropriate for many VDI use cases. Differences between the parent VM and the clone are maintained in a delta file. While updates can be rolled out effectively, multiple VM rebuilds are required to correctly deploy a patch at the operating system level. Operating system updates are rolled out to the parent images, and then the Desktop pool is pointed to the new snapshot with the updates. A Horizon Composer instance is required with linked clones to manage the recompose functions of the pool.

  ⓘ **NOTE:** Linked clones are being deprecated in Horizon 8.0 and will not be available in future versions of VMware Horizon.

Client components

VDI enables centralized, secure access and support for clinical and non-clinical workers on a variety of client devices. This is particularly important as you scale your remote workforce. Users can access the virtual desktops through a variety of client devices as appropriate to the organization's needs. The following table lists the client components that Dell Technologies recommends:

Table 2. Recommended client components

Component	Description	Recommended use	More information
Wyse thin clients	• Highly secure thin client operating system with no sensitive data or personal information exposed on the local device • Dedicated to corporate use, prevents unauthorized software and viruses • Optimizes management and efficiency by delivering a controlled access to centralized data, applications, and resources • High-quality user experiences with desktop, All-in-One, and mobile form factors and a comprehensive ecosystem	• Security and manageability • Optimized to access virtualized desktops and cloud applications and deliver high-quality client computing experiences and enterprise-class security, while streamlining management through centralized control	https://www.delltechnologies.com/en-us/wyse/index.htm
Latitude laptops and 2-in-1s	• Biggest screens in a smaller footprint with a wide array of ports to connect peripherals and enjoy speakerphone experience • More responsive apps with Dell Optimizer and intelligent audio for better conference experience • Better connectivity including 4G LTE, Wi-Fi 6, and eSIM • 5G design on the Latitude 9510 • Smart antenna design on select products for better connections	• Mobility and space-saving devices • Allows users to be productive and stay connected with versatile, space-saving mobile solutions • Offers a modern portfolio built to prioritize customer experience and keep employees productive wherever they work with a selection of laptops, 2-in-1s, and ecosystem products	https://www.delltechnologies.com/en-us/latitude/index.htm
OptiPlex business desktops and All-in-Ones	• Intel 9th and 10th Gen core processors, providing two times system responsiveness with Intel Optane Memory • Flexible expansion options, including rich CPU, SSD, and PCIe NVMe • Many innovative form factors with versatile mounting options, including a zero-footprint modular desktop hidden in plain sight, and space-saving All-in-Ones • Rich interaction with display technology, including 4k UHD All-in-Ones and matching multi-monitor support	• Modular solution • Ideal for desk-centric and remote workers in fixed environments who require varying degrees of performance and expandability	https://www.delltechnologies.com/en-us/optiplex/index.htm
Precision workstations	• Complete workstation portfolio with towers, racks, and mobile form factors	• High-end graphics and extreme performance • Precision workstations designed to run processor- and graphic-	https://www.delltechnologies.com/en-us/precision/index.htm

Table 2. Recommended client components (continued)

Component	Description	Recommended use	More information
	● Powerful workstations for the most demanding applications, scalable storage, and RAID options ● Small, intelligent, and high performing mobile workstation portfolio	intensive applications and activities with mission-critical reliability such as analytics, simulations, and modeling	

4

Solution performance testing

This chapter presents the following topics:

Topics:

Introduction

For the performance testing, we designed ten different test cases to illustrate various combinations of AI and medical VDI workloads using both CPU and GPU resources:

- Four AI test cases, for AI training and AI model validation, both run with CPU only and with GPU enhancement
- Two traditional VDI test cases for a medical knowledge worker profile using a MicroDicom medical image viewer application, with the same test case run with CPU only and with GPU enhancement
- Four "dual workload" test cases that ran concurrent workloads for AI training and validation and VDI together, with and without GPUs in several combinations

The list of test cases is:

- AI test cases
  - AI Training GPU Enhanced
  - AI Validation GPU Enhanced
  - AI Training CPU Only
  - AI Validation CPU Only
- VDI test cases
  - VDI Medical Knowledge Worker GPU Enhanced
  - VDI Medical Knowledge Worker CPU Only
- Dual workload AI and VDI test cases
  - Dual Workload AI Training and VDI Medical Knowledge Worker Both GPU Enhanced
  - Dual Workload AI Validation and VDI Medical Knowledge Worker Both GPU Enhanced
  - Dual Workload AI Training CPU Only and VDI Medical Knowledge Worker GPU Enhanced
  - Dual Workload AI Validation CPU Only and VDI Medical Knowledge Worker GPU Enhanced

The following sections describe the test environment and methodology, and present a summary of the performance test results, with an emphasis on the dual workload cases that are the most applicable to the overall premise of GPU use in a healthcare environment.

Test environment and methodology

Test tools

To ensure the optimal combination of end-user experience and cost-per-user, the performance analysis and characterization on Dell Technologies VDI Solutions is carried out using a carefully designed, holistic methodology that monitors both hardware resource utilization parameters and user experience during load testing.

Login VSI is the industry standard tool for testing VDI environments and server-based computing. The tool installs a standard collection of desktop application software (for example, Microsoft Office, Adobe Acrobat Reader, and so on) on each VDI desktop. It then uses launcher systems to connect a specified number of users to available desktops within the environment.

After the user is connected, the workload is started by a login script, which starts the test script after the user environment is configured. Each launcher system can launch connections to several target machines (in this case VDI desktops).

For Login VSI, the launchers and virtual session indexer environment are configured and managed by a centralized management console. Additionally, the following login and boot paradigm is used:

- Users were logged in within a login timeframe of 1 hour. An exception to this login timeframe occurred when testing low-density solutions such as GPU/graphics-based configurations. With those configurations, users were logged in every 5 seconds.
- All desktops were pre-booted before logins commenced.
- The data collection interval for vSAN metrics was 5 minutes.

For the AI testing, AI training and AI model validation took place on the same VM. After training took place, the models obtained were then validated with the same VM hardware configuration used in training.

The methodology that we used for training and model validation was similar to the method that was used in the AI-assisted Radiology Using Distributed Deep Learning on Apache Spark and Analytics Zoo White Paper. Model weights were collected for 15 Epochs during training, with each subsequently validated with the same techniques as in the referenced paper. For example, the Average AUC-ROC[1] accuracy was calculated for all disease categories contained in the dataset. The model with the highest-performing AUC-ROC accuracy can be used for inferencing on real-world data. We validated against the entire dataset used for training and not on a withheld or holdout portion of the dataset, which is a common validation technique.

Resource monitoring

The team used several methods for resource and component monitoring during performance testing and in the deployed solutions where applicable.

We used VMware vCenter for VMware vSphere-based solutions to gather key data (CPU, GPU, memory, disk, and network usage) from each of the compute hosts during each test run. We exported this data to .csv files for single hosts and then consolidated it to show data from all hosts (when multiple hosts were tested). While the report does not include specific performance metrics for the Management host servers, we monitored these servers during testing to ensure that they were performing at an expected performance level with no bottlenecks.

We gathered GPU performance metrics directly from the vSphere client, either manually or using a script.

For resource utilization, we determined the user density at a reasonable system load. Testing to system failure was out of scope. To achieve a reasonable system load, we set target thresholds for system resources, as described in the following table. These thresholds reflect a system that is well utilized but not near failure.

Table 3. Resource utilization thresholds

Metrics	Target threshold
Average CPU usage	85%
Average CPU core utilization	85%
Average CPU readiness	10%
Average memory utilization (active)	85%
Consumed memory	<100%
Memory ballooning	None
Memory swapping	None
Network throughput	85%
Storage latency	20 milliseconds (ms)
Spare storage capacity	15%

Test configuration

The following table describes the hardware and software components of the infrastructure that we used for the performance analysis and characterization tests:

[1] Area under the curve (AUC) Receiver operating characteristics (ROC)

Table 4. Validated hardware configurations

Category	Platform	CPU	Memory	Network
Compute host hardware	4 x VxRail V570F (Density Optimized)	2 x 6248 at 2.5 GHz, 20-core processors	768 GB memory at 2933 MT/s 64 GB x 12 DDR-4	Broadcom Adv. Dual 25 Gb Ethernet
Management host hardware	4 x VxRail E460F (Management)	2 x Intel Xeon CPU E5-2698 v4 at 2.20 GHz	512 GB Memory at 2400 DDR-4 (32 GB x 16)	Intel 2P X520/2P I350 rNDC

The following table describes the hardware and software components that we used:

Table 5. Hardware and software components

Category	Component
NVIDIA GPU	3 NVIDIA RTX 8000 GPU cards installed in one host
Storage	Compute ● vSAN VxRail ● HBA330-Adp ● BOSS S1 (2 x 240 GB SATA M.2 SSD) ● 2 x 800 GB WI SSDs + 2 x 1.92 TB RI SSDs per host ● 2 vSAN disk groups Management ● HBA330 Mini ● 2 x 400 GB WI SSDs, 2 x 1.92 TB RI SSDs
Network	PowerSwitch S5248-ON switch
Patches	All Spectre/Meltdown/L1TF patches were applied to the parent image and ESXi hosts as required.
Protocol	BLAST Extreme H.264 + Switch Codec
Broker	VMware Horizon 8.0.0 build - 16592062
Hypervisor	VMware vSphere ESXi, 7.0.0, 15843807 VMware vCenter 7.0.0 15952498
SQL	Microsoft SQL Server 2019
vGPU software version	NVIDIA vGPU 11
Desktop operating system	Microsoft Windows 10 Enterprise 64-Bit, 1909 version
Microsoft Office version	Office 2019
Management operating system	Microsoft Window Server 2019
Login VSI version	Login VSI 4.1.40.1
Antivirus software	Windows Defender

AI training and validation configuration, dataset, and model

The following tables show the AI VM and operating system configurations and the dataset that we chose for our AI training and validation. Other configurations of the AI VM are possible that may improve performance. However, we didn't extensively explore these.

Table 6. AI VM configurations

Configuration	CPU	Memory	Disk	GPU PCI 0*	GPU PCI 1*	GPU PCI 2*
AI VM	20 vCPUs, 1 core per socket, 20 sockets	256 GB reserved	560 GB HD	Rtx8000 Grid_rtx8000p-48q	Rtx8000 Grid_rtx8000p-48q	Rtx8000 Grid_rtx8000p-48q

*In the testing scenarios that follow, we performed testing with one, two, or all of the GPUs assigned to the AI VM.

Table 7. Operating system configurations

Operating system	Version	Kernel	NVIDIA driver	NVIDIA-SMI CUDA	CUDA
Ubuntu	18.04.4 LTS	5.4.0-47-generic	440.87	10.2	11.0.182

Table 8. Dataset used for training and validation

Data set	Images	Description
ChestXRay14	112,120 png images	Chest x-ray dataset, containing over 100,000 frontal view x-ray images with 14 labeled diseases categories.

Table 9. Deep-learning convolution neural network (DLCNN) model topology

Neural network topology	Description
DenseNet	DenseNet alleviates the vanishing-gradient problem, strengthens feature propagation, encourages feature reuse, and substantially reduces the number of parameters.

Graphics VM configuration

The following table shows the configuration of the graphics VM:

Table 10. Graphics VM configuration

Configuration	CPU	Memory	Disk	GPU PCI 0
Graphics VM	4 vCPUs	8 GB reserved	60 GB HD	Rtx8000 Grid_rtx8000p-2q

Medical dataset used for graphics

The following table shows the medical dataset that we used for the VDI graphics tests:

Table 11. Medical dataset used

Dataset	Images	Description
DICOM Library for abdomen	361 DICOM images	CT anonymized images of abdomen dataset, containing 361 DICOM images

Test results

The following test results include the performance of the platform during the deletion and re-creation of the instant clone VMs after all users logged out when the test run was completed. The different phases of the test cycle are displayed in the test results graphs in Appendix A.

The following tables summarize the test results for the various workloads and configurations. For graphs and analysis of the detailed test results for the dual workload AI and VDI test cases, see Appendix A.

The user densities we tested with for the "VDI Medical Knowledge Worker GPU enhanced" workload were as follows:

- Single workload—72 users
- Dual workload—24 users

The following table summarizes the test results for single workload test cases:

Table 12. Summary of test results for single workload test cases

Test case	AI duration	GPU A	GPU B	GPU C	Average host CPU utilization	Average GPU card utilization	Average vSAN read latency	Average vSAN write latency	LVSI UI metrics	Average AUC
AI Training GPU enhanced	1 hr 5 min	AI	AI	AI	28.6%	GPU A 81% GPU B 81% GPU C 81%	0.3 ms	1.8 ms	N/A	N/A
AI Validation GPU enhanced	38 min	AI	N/A	N/A	13%	GPU A 30% GPU B N/A GPU C N/A	0.4 ms	1.3 ms	N/A	0.80
AI Training CPU only	47 hr 24 min	N/A	N/A	N/A	42%	N/A	0.2 ms	0.6 ms	N/A	N/A
AI Validation CPU only	1 hr 35 min	N/A	N/A	N/A	30%	N/A	0.2 ms	1.0 ms	N/A	0.72
VDI Medical Knowledge Worker GPU enhanced	N/A	VDI	VDI	VDI	84%	GPU A 25% GPU B 25% GPU C 25%	0.1 ms	1.02 ms	Base: 806 Avg max: 1226 Threshold: 1806	N/A
VDI Medical Knowledge Worker CPU only	N/A	N/A	N/A	N/A	73%	N/A	0.2 ms	1.0 ms	Base: 792 Avg max: 1216 Threshold: 1792	N/A

The following table summarizes the test results for dual workload test cases:

Table 13. Summary of test results for dual workload test cases

Test case	AI duration	GPU A	GPU B	GPU C	Average host CPU utilization	Average GPU card utilization	Average vSAN read latency	Average vSAN write latency	LVSI UI metrics	Average AUC
Dual Workload AI Training and VDI Medical Knowledge Worker, both GPU enhanced	4 hr 43 min	VDI	AI	AI	38%	GPU A 28% GPU B 28% GPU C 29%	0.2 ms	0.8 ms	Base: 795 Avg max: 821 Threshold: 1795	N/A
Dual Workload AI Validation and VDI Medical Knowledge Worker, both GPU enhanced	48 min 30 sec	VDI	AI	N/A	38%	GPU A 27% GPU B 13% GPU C N/A	0.1 ms	0.77 ms	Base: 791 Avg max: 796 Threshold: 1791	0.75
Dual Workload AI Training CPU only and VDI Medical Knowledge Worker GPU enhanced	50 hr 30 min	N/A	N/A	VDI	71%	GPU A N/A GPU B N/A GPU C 27%	0.25 ms	1.30 ms	Base: 830 Avg max: 897 Threshold: 1831	N/A
Dual Workload AI Validation CPU only and VDI Medical Knowledge Worker GPU enhanced	1 hr 44 min	N/A	N/A	VDI	53%	GPU A N/A GPU B N/A GPU C 27%	0.2 ms	1.0 ms	Base: 920 Avg max: 894 Threshold: 1920	0.69

Test conclusions

The charts in this section show the key conclusions and takeaways from the test results.

AI duration and AUC accuracy

The following figure shows that in the dual workload GPU-enabled workloads, AI and VDI together perform significantly better for model training times and AUC accuracy compared to CPU-only configurations.

This chart depicts both the duration (left Y axis) and AUC accuracy (right Y axis) for four test cases. AI training and validation without GPU for both single and dual workload test cases are used as baseline measurements. GPU enhancement significantly reduces duration and increases accuracy. The AI workload is 28 times faster with GPUs and the dual AI and VDI medical workload is 10 times faster with GPUs. At the same time, in both instances model accuracy is higher with GPU acceleration.

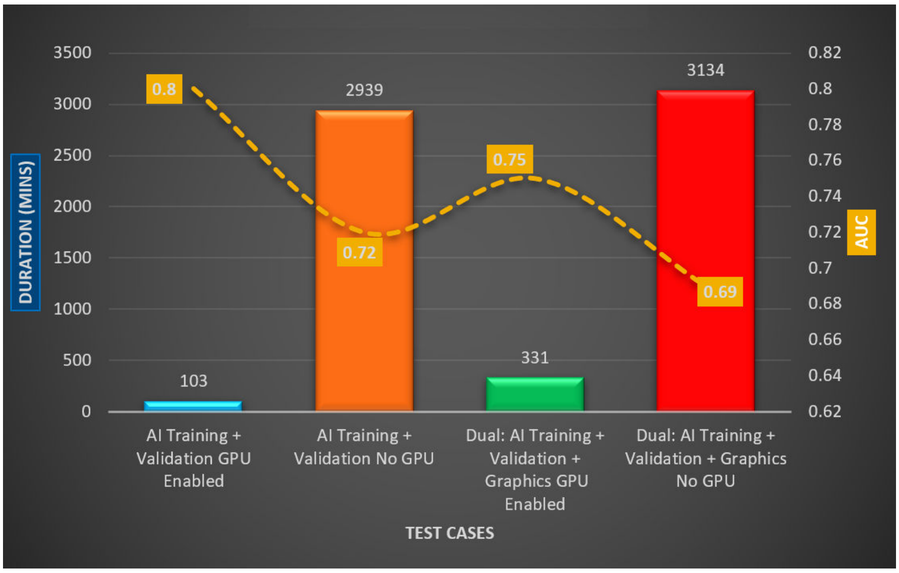


Figure 9. Duration and AUC accuracy

Dual AI and VDI workloads details

The following figures show the model AUC accuracy and Login VSI base response times respectively.

These dual workloads can co-exist and are not impacted by “noisy neighbor” type issues. Both the AI and VDI metrics show that the dual test case scenario performed well for model accuracy and VDI user experience response times.

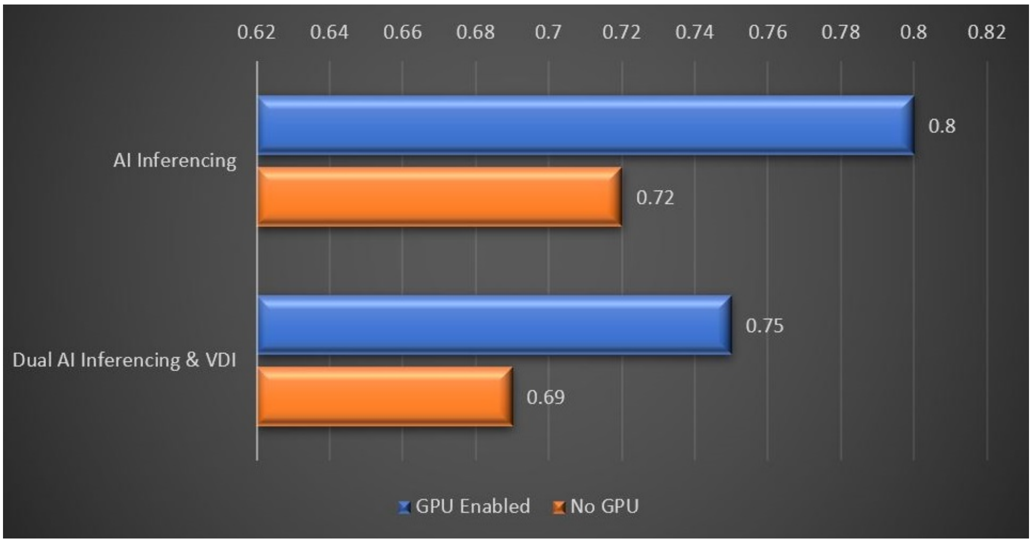


Figure 10. Model AUC accuracy for AI metrics

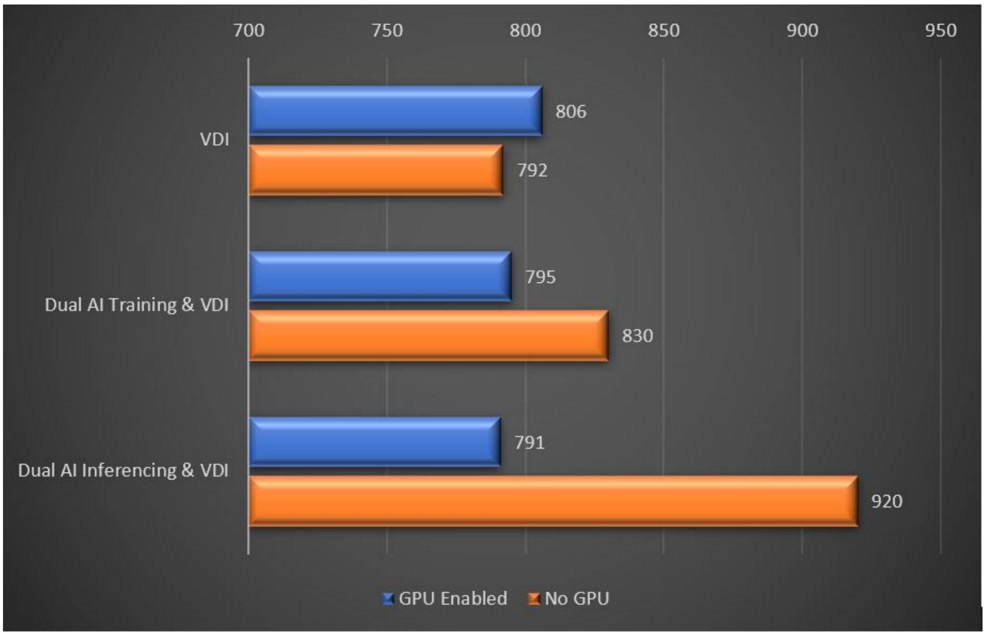


Figure 11. Login VSI base response time (ms) for VDI metrics

Workload vGPU and vCPU utilization profiles

The following figures show several views of resource utilization with a combined view of CPU and GPU resource utilization. This can provide insight into the impact of switching workloads between VDI and AI in a single use case operation when compared with running dual workloads in parallel, and how either case can impact overall utilization.

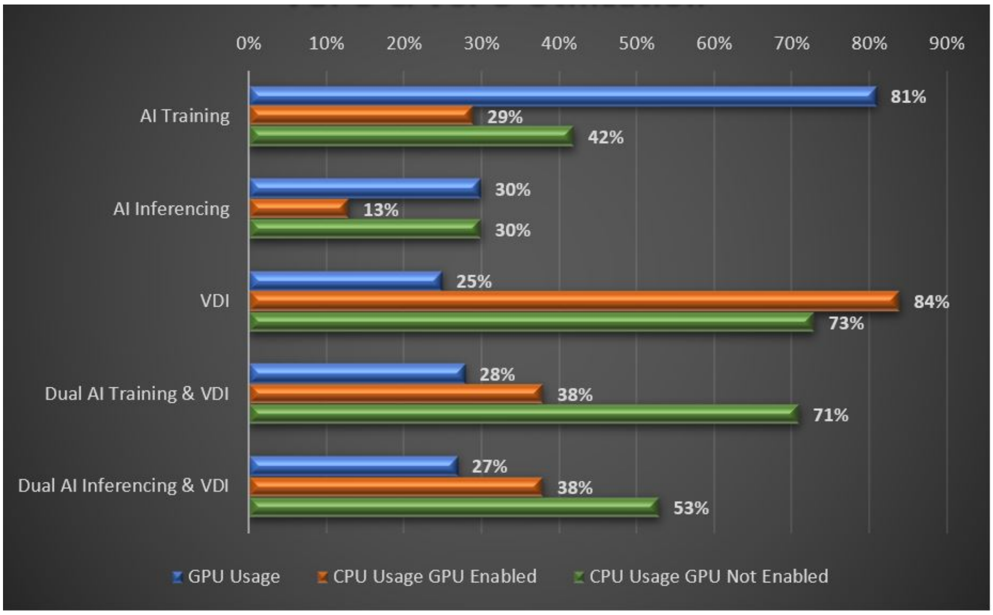


Figure 12. vGPU and vCPU usage percentage

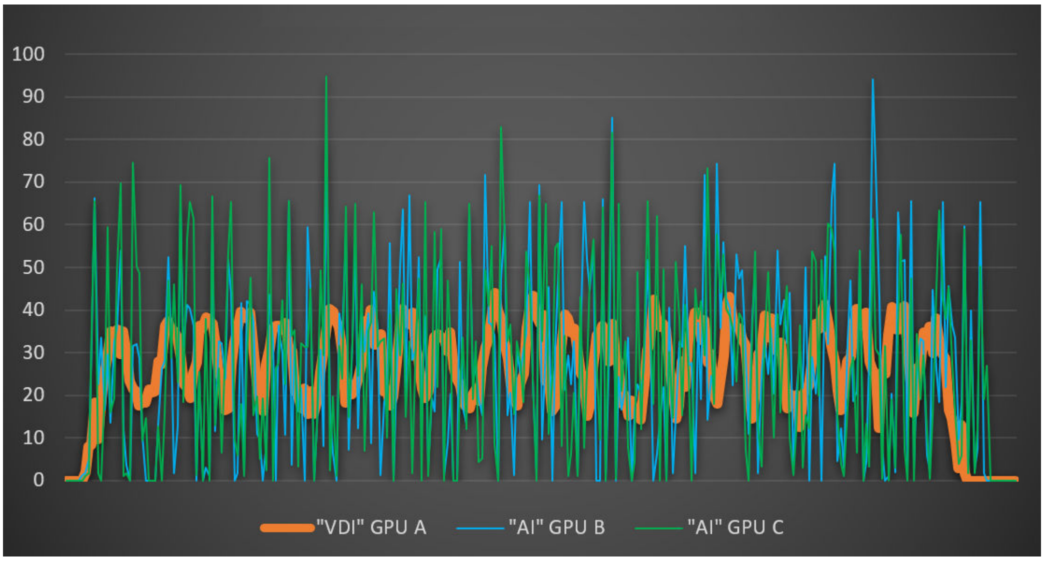


Figure 13. GPU utilization percentage for dual workloads

Duration of workloads

The following figure shows how the duration of workload execution is significantly reduced by applying GPU resources.

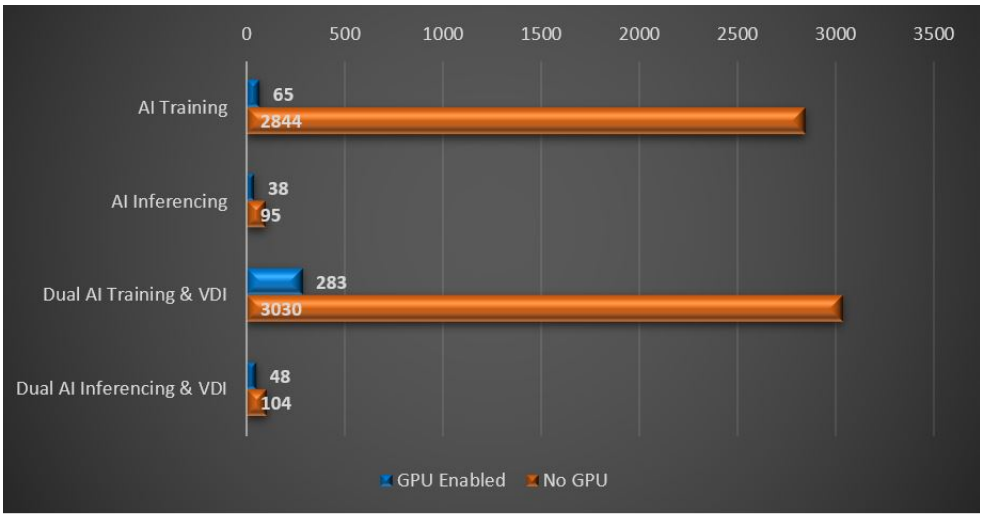


Figure 14. Baseline durations (in minutes)

5

Design guidance

This chapter presents the following topics:

Topics:

Platform configurations

With several configurations to choose from, consider these basic differences:

- The Density Optimized configuration provides a good balance of performance and scalability for various general-purpose VDI workloads.
- The Virtual Workstation configuration provides the highest levels of performance for more specialized VDI workloads, which means it can be used with ISV and high-end computing workloads.

CPU

Dell Technologies VDI Solutions validation test results suggest that you can use CPU oversubscription to effectively size VDI user density. To use a CPU configuration other than those that have been validated, consider the following guidance to achieve comparable results:

- For architectures with Intel Xeon scalable Cascade Lake processors:
  - **Medical Knowledge Workers**—1.8 users per core. For example, 28 users with dual eight-core processors
  - **Medical Knowledge Workers + AI Appliance**—0.6 users per core. For example, 10 users with dual eight-core processors
- AMD and Intel CPUs are not vMotion compatible within the same VMware vSphere Cluster. If using a mixed CPU vendor environment, ensure that CPUs from the same vendor are in the same cluster. For more information, see the VMware EVC and CPU Compatibility FAQ (1005764) knowledgebase article.
- For graphics:
  - For high-end graphics configurations with NVIDIA vWS graphics enabled, choose higher clock speeds over higher core counts. Many applications that benefit from high-end graphics are engineered with single-threaded CPU components. Higher clock speeds benefit users more in these workloads.
  - Most graphics configurations do not experience high CPU oversubscription because vGPU resources are likely to be the resource constraint in the appliance.
  - VMware has released and updated the per-CPU licensing model that requires a license per-CPU for up to 32 physical cores. This design guide recommends using processors with fewer than 32 cores to avoid additional licensing requirements. For more information, see Update to VMware's per-CPU Pricing Model.

Memory

Best-practice recommendations for memory allocation and configuration include:

- Do not overcommit memory when sizing because memory is often not the constraining resource. Overcommitting memory increases the possibility of performance degradation if contention for memory resources, such as swapping and ballooning of memory, occurs. Overcommitted memory can also affect storage performance when swap files are created.

- Populate memory in units of six DIMMs per CPU to yield the highest performance. Dell EMC PowerEdge servers using Intel Xeon Scalable processors have six memory channels per CPU, which are controlled by two internal memory controllers, each handling three memory channels. To ensure that your environment has the optimal memory configuration, use a balanced configuration, where each CPU supports a maximum of 12 DIMMs (or 24 DIMMs for a dual-CPU server). The most effective configuration is 12 DIMMs (6 per processor) with Intel Xeon Scalable processors.

NVIDIA vGPU considerations

Best practices for sizing and configuring solutions requiring graphics accelerators include:

- vPC licenses support up to 2 GB of frame buffer and up to two 4K monitors or a single 5K monitor to cover most traditional VDI users. Maximum node density for graphics-accelerated use can typically be calculated as the available frame buffer per node divided by the frame buffer size.
- The addition of GPU cards does not necessarily reduce CPU utilization. Instead, it enhances the user experience and offloads specific operations best performed by the GPU.
- Dell Technologies recommends using the BLAST protocol for vGPU enabled desktops. NVIDIA GPUs are equipped with encoders that support BLAST.
- Virtual Workstations are typically configured with at least 2 GB video buffer.
- When configuring NVIDIA M10 GPU cards in a solution, Dell Technologies recommends a maximum memory capacity of 768 GB, due to limitations in the Maxwell architecture. Pascal and Turing architectures do not have the same limitation.

Sizing considerations

This section provides various general best practices for sizing your deployment.

- **User density**—If concurrency is a concern, calculate how many users will use the environment at the peak of utilization. For example, if only 80 percent are using the environment at any time, the environment must support only that number of users (plus a failure capacity).
- **Disaster recovery**—For DR planning, Dell Technologies recommends that you implement a dual or multi-site solution. The goal is to keep the environment online and, in case of an outage, to perform an environment recovery with minimum disruption to the business.
- **Management and compute clusters**—For small environments, it may be appropriate to use a combined management and compute cluster. For environments deployed at a larger scale, Dell Technologies recommends using separate management and compute layers. When creating a management cluster for a large-scale deployment, consider using the VxRail E560F to reduce the data center footprint. With a more flexible platform that accommodates a wider variety of VDI application workloads, the VxRail V570F is preferred for compute clusters.
- **Network isolation**—The design shown in this document illustrates a two-NIC configuration per appliance with all the traffic separated logically using VLAN. When designing for larger-scale deployments, consider physically separating the management and VDI traffic from the vSAN traffic for traffic isolation and to improve network performance and scalability.
- **FTT**—Dell Technologies recommends sizing storage with NumberOfFailuresToTolerate (FTT) set to 1, which means that you must double the amount of total storage to accommodate the mirroring of each VMDK.
- **Slack space**—Dell Technologies recommends adding an additional 30 percent of slack space to prevent automatic rebalancing of storage, which impacts performance. Automatic balancing occurs when the storage reaches 80 percent of the full threshold. Therefore, we recommend 70 percent to reserve a 10 percent buffer.
- **All-flash compared to hybrid:**
  - Hybrid and all-flash configurations have shown to provide similar performance results when testing VDI configurations. Because hybrid uses spinning drives, consider the durability of the disks.
  - Only all-flash configurations offer deduplication and compression for vSAN. Dell Technologies recommends all-flash configurations for simplified data management.
  - All-flash configurations need considerably less storage capacity than hybrid configurations to produce similar FTT, as shown in the following table.

    ⓘ **NOTE:** Enabling certain data efficiency features can incur processing overhead.

Table 14. FTT comparisons

Configuration	VM size	FTM	FTT	Overhead	Capacity required	Hosts required
Hybrid	50 GB	RAID-1 (Mirrored)	1	2 x	100 GB	3

Table 14. FTT comparisons (continued)

Configuration	VM size	FTM	FTT	Overhead	Capacity required	Hosts required
All-flash	50 GB	RAID-5 (3+1) (Erasure coding)	1	1.33 x	66.5 GB	4
Hybrid	50 GB	RAID-1 (Mirrored)	2	3 x	150 GB	4
All-flash	50 GB	RAID-6 (4+2) (Erasure coding)	2	1.5 x	75 GB	6

ⓘ **NOTE:** For more information about multi-site design considerations for Horizon, see the VMware Workspace ONE and VMware Horizon Reference Architecture.

Design assessment

Before deploying the solution, assess your environment to validate design considerations and ensure that you are designing your architecture to meet or exceed the performance of your current environment. Dell Technologies Professional Services offers an assessment service for all VDI needs.

6

Design enhancements

This chapter presents the following topics:

Topics:

- File workload guidance
- Data center infrastructure

File workload guidance

The increased growth in the amount of data that is stored in file shares and user home directories across IT environments in recent years has resulted in an increased focus on the need to better manage this unstructured data. As a result, many organizations are deploying dedicated file workload solutions with capabilities such as cloud file tiering and single file system namespaces across their IT infrastructure, including for file workloads in a VDI environment.

Dell Technologies provides several solutions for different types of file workloads.

Dell EMC PowerStore storage

Dell EMC PowerStore T storage is simple, unified storage that enables flexible growth with intelligent scale-up and scale-out capabilities and public cloud integration.

Dell EMC PowerStore T is ideal for general-purpose NAS/SAN mixed workload consolidation, smaller file workloads (including small to midsized VDI environments), and transactional databases.

The following figure shows an example of a 4,000-user VDI deployment using Dell EMC PowerStore T storage for file shares:

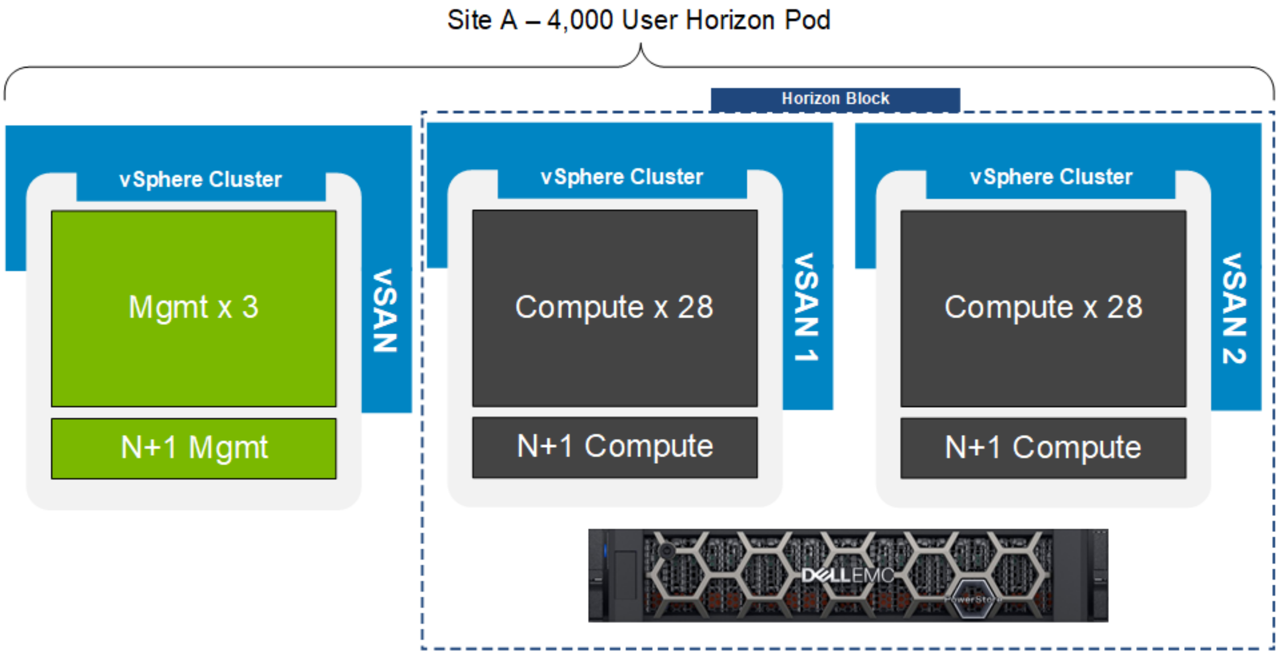


Figure 15. 4,000-user pod on Dell EMC PowerStore T

When you are deploying Dell EMC PowerStore T in a VDI environment, Dell Technologies recommends that you deploy a separate PowerStore T storage system with each VMware Horizon block. This logical mapping will allow for isolated maintenance outages as the environment grows and there are multiple Horizon blocks within a pod architecture.

Each PowerStore T system can scale up to four appliances per cluster. This structure provides the greatest scalability, resiliency, and flexibility when deploying and maintaining file services for the overall user pod. As unstructured data storage needs grow over time, the capacity of each PowerStore T storage system can be scaled up or out independently with minimal user impact. You have the choice to deploy alternative architectures to the one suggested here, but you should carefully consider the tradeoffs.

For guidance about selecting an appropriate Dell EMC PowerStore T storage solution for your file workload requirements, see the Dell EMC PowerStore website.

Dell EMC PowerScale file storage

Dell EMC PowerScale storage is a scale-out NAS solution for any file workload.

The PowerScale system is ideal for a wide range of file workloads (including large-scale enterprise VDI environments requiring a single file system namespace), high-performance computing (HPC), archiving, and infrastructure consolidation.

The following figure shows an example of an 8,000-user VDI deployment using PowerScale scale-out storage with a single namespace:

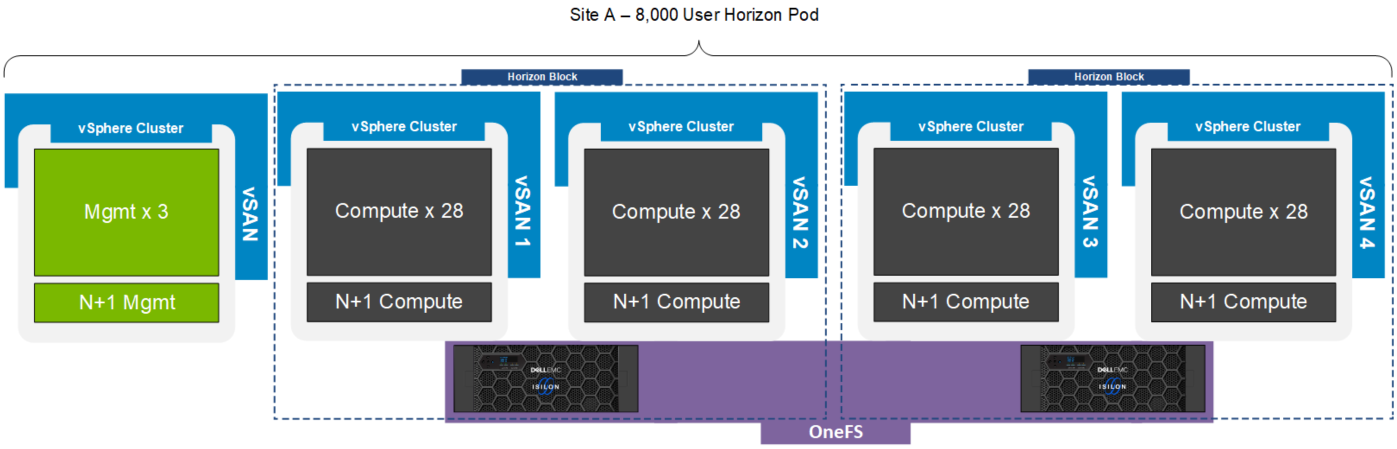


Figure 16. 8,000-user pod on PowerScale system

When you are deploying a PowerScale storage system in a VDI environment, Dell Technologies recommends that you deploy a separate PowerScale node with a VMware Horizon block. This logical mapping will allow for isolated maintenance outages as the environment grows and there are multiple Horizon blocks within a pod architecture.

This structure provides the greatest scalability, resiliency, and flexibility for deploying and maintaining file services for the overall user pod. As unstructured data-storage needs grow over time, you can scale up the capacity of each PowerScale storage system independently with minimal user impact. In addition to scaling up each PowerScale chassis, you can also scale out a PowerScale system by using the Dell EMC OneFS operating system. Thus, multiple PowerScale systems can provide a single volume and namespace that all user pods in a data center can access.

As shown in the previous figure, you can scale out the system as the VDI environment grows. You can deploy alternative architectures to the architecture suggested here, but first consider the tradeoffs carefully.

For guidance about selecting an appropriate PowerScale storage solution for your file workload requirements, see the Dell EMC PowerScale website.

Data center infrastructure

Enterprise equipment requires power to operate, racks to enable streamlined management, and cooling to maintain reliable operations.

Careful selection of the infrastructure solutions that provide these capabilities is vital to ensure uptime, scalability, energy efficiency, and ease of management. Dell Technologies provides a wide range of data center infrastructure solutions:

- **Dell EMC Netshelter SX racks**—Deploy server, storage, and networking equipment and other IT hardware while optimizing power, cooling, cabling, and systems management.
- **Dell EMC Keyboard Video Mouse (KVM) and Keyboard Monitor Mouse (KMM) solutions**—Manage 8 to 1,024 local and remote servers running various operating systems across the enterprise.

- **Dell EMC Smart-UPS**—Deliver reliable power and protect IT equipment, including servers, storage, networking, point-of-sale, and medical equipment.
- **APC Rack Power Distribution Units (PDUs)**—Provide reliable power distribution that is designed to increase manageability and efficiency in your data center.

7

Conclusion

This design guide describes technical considerations and best practices for sharing GPU resources and accelerating VDI and AI workloads on the same infrastructure in a healthcare environment, in what may be considered as a "VDI by day, compute by night" model.

The design integrates VMware Horizon 8 software on Dell EMC VxRail, with VMware vSAN storage and VMware vSphere virtualization, and uses NVIDIA RTX 8000 GPUs and NVIDIA vWS software to create GPU-accelerated virtual application and desktop environments.

By implementing a shared common platform for VDI and AI workloads that increases GPU utilization, organizations in the healthcare industry can lower their TCO, improve the mobility of their employees, and experience high performance for both VDI graphics and AI workloads, thereby improving the overall efficiency of their day-to-day operating environment.

Modern IT infrastructure from Dell Technologies helps facilitate next-generation medical imaging applications by providing critical solutions and partnerships that allow healthcare organizations to balance clinical requirements with the resources needed to deliver a unified medical imaging solution. Shifting from disconnected departmental PACS instances to interoperable medical imaging solutions allows clinicians, researchers, and healthcare systems to:

- Ingest and analyze data in real time
- Retain images based on regulatory and compliance requirements, thereby limiting potential liability exposure
- Seamlessly share data with collaborators, patients, payers, and other healthcare-life science organizations, thereby improving outcomes.

Dell Technologies has extensive experience in VDI solutions and a wide selection of compute, storage, and networking for VDI. Our tested and validated Dell Technologies VDI Solutions, based on VMware Horizon and configured with NVIDIA vGPUs, offer exceptional graphics performance and predictable cost, and provide a single vendor support experience.

With VDI solutions from Dell Technologies, you can streamline the design and implementation process, and be assured that you have a solution that is optimized for performance, density, and cost-effectiveness.

References

The documentation and reference links in this section provide additional information. This list includes references cited throughout this document.

Dell Technologies documentation

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

- Dell Technologies VDI Solutions Info Hub
- AI-assisted Radiology Using Distributed Deep Learning on Apache Spark and Analytics Zoo
- Dell Technologies - Technology Solutions for Care Providers
- Dell EMC VxRail InfoHub
- TechBook—Dell EMC VxRail System
- Planning Guide—Dell EMC VxRail Network Planning
- Planning Guide—VMware Cloud Foundation 3.x on VxRail
- Dell EMC PowerSwitch S-Series 25GbE switches
- Dell EMC PowerSwitch S-Series 10GbE switches
- Dell Latitude Laptops and Two-in-Ones
- Dell OptiPlex Business Desktops and All-in-Ones
- Dell Precision Workstations
- Wyse Thin Clients
- Dell EMC PowerProtect DD Series Appliances
- Dell EMC PowerStore
- Dell EMC PowerScale

VMware documentation

The following links provide additional information from VMware:

- VMware Horizon 8 2006 Configuration Limits
- VMware Horizon 7 Sizing and Limitations and Recommendations
- VMware Horizon 7 Deployment Guide for Healthcare
- VMware Workspace ONE and VMware Horizon 7 Enterprise Edition On-premises Reference Architecture
- VMware Horizon 7 Enterprise Edition Multi-Site Reference Architecture
- VMware Horizon Resources Page
- VMware Horizon License FAQ
- VMware Horizon 7 Instant-Clone Desktops and RDSH Servers White Paper
- VMware EVC and CPU Compatibility FAQ
- VMware Update to VMware's per-CPU Pricing Model
- VMware Data Protection for a VMware Horizon VDI Environment using Dell EMC Data Protection Suite Operations Guide

NVIDIA documentation

The following link provides additional information from NVIDIA:

- NVIDIA Virtual GPU Software Quick Start Guide

MicroDicom

The following link provides additional information from MicroDicom:

- MicroDicom DICOM viewer for Windows

A

Appendix: Solution performance testing details

This appendix contains further details on the test results for the four dual workload test cases. The graphs shown here have time marked on the X axis and the metric marked on the Y axis.

Performance metrics

The graphs in this section present the following metrics:

- CPU utilization
- GPU utilization
- GPU memory
- vSAN cluster latency
- Login VSI user experience

These represent the key results. Additional metrics that we collected during the performance testing included CPU core utilization and readiness, CPU active and consumed memory, network usage, and cluster disk IOPS.

Test results for dual workload test cases

The four dual workload test cases that this section describes are:

- Dual Workload AI Training and VDI Medical Knowledge Worker Both GPU Enhanced
- Dual Workload AI Validation and VDI Medical Knowledge Worker Both GPU Enhanced
- Dual Workload AI Training CPU Only and VDI Medical Knowledge Worker GPU Enhanced
- Dual Workload AI Validation CPU Only and VDI Medical Knowledge Worker GPU Enhanced

Dual Workload AI Training and VDI Medical Knowledge Worker Both GPU Enhanced

CPU utilization

The following figure shows the average CPU utilization for VDI steady-state and AI training. The CPU average utilization during the steady-state phase was approximately 38 percent. The numerous CPU usage peaks are the VDI medical workload phases on top of a latent AI training CPU requirement of 25 percent. Minor CPU contention was seen during these phases along with elevated CPU readiness averaging 0.1 percent during the steady-state phase and peaking at 0.7 percent during the logout phase.

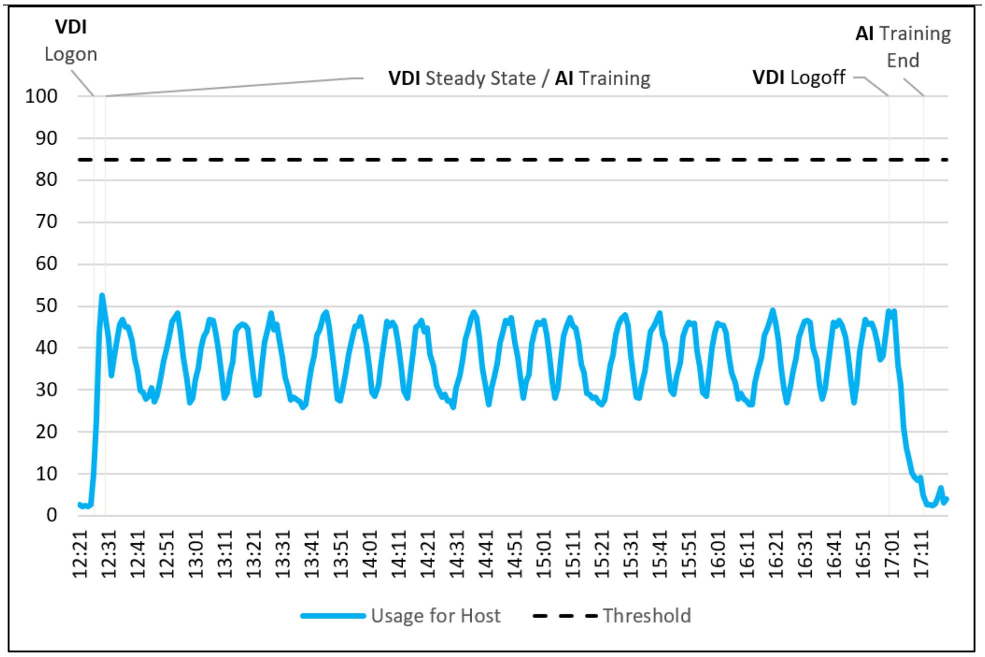


Figure 17. Host CPU usage percentage

GPU utilization

The following figure shows the peak GPU utilization for VDI steady-state and AI training. The two AI GPU cards show similar utilization throughout the training. AI GPU requirements show large peaks and troughs, but overall usage is similar to the VDI-assigned GPU of 28 percent. GPU B and C are not driven to the average usage of 81 percent when all three GPUs are assigned to AI. The VDI GPU workload utilization pattern can be seen clearly for GPU A following the various workload phases and averaged at 28 percent during the steady-state phase.

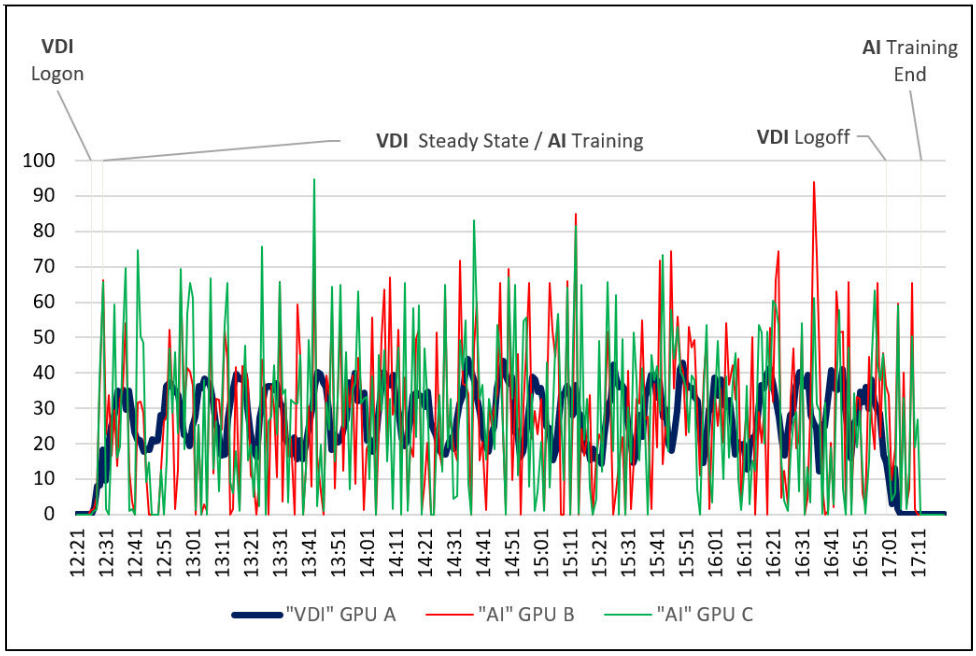


Figure 18. GPU usage percentage

GPU memory

The following figure shows the peak GPU memory utilization for VDI steady-state and AI training. The three GPU cards showed the same memory usage, consuming 100 percent of available memory during the login period for steady-state and training. Memory usage dips during VDI clone recreation for GPU A before consuming 100 percent of available GPU memory when the VMs are re-created.

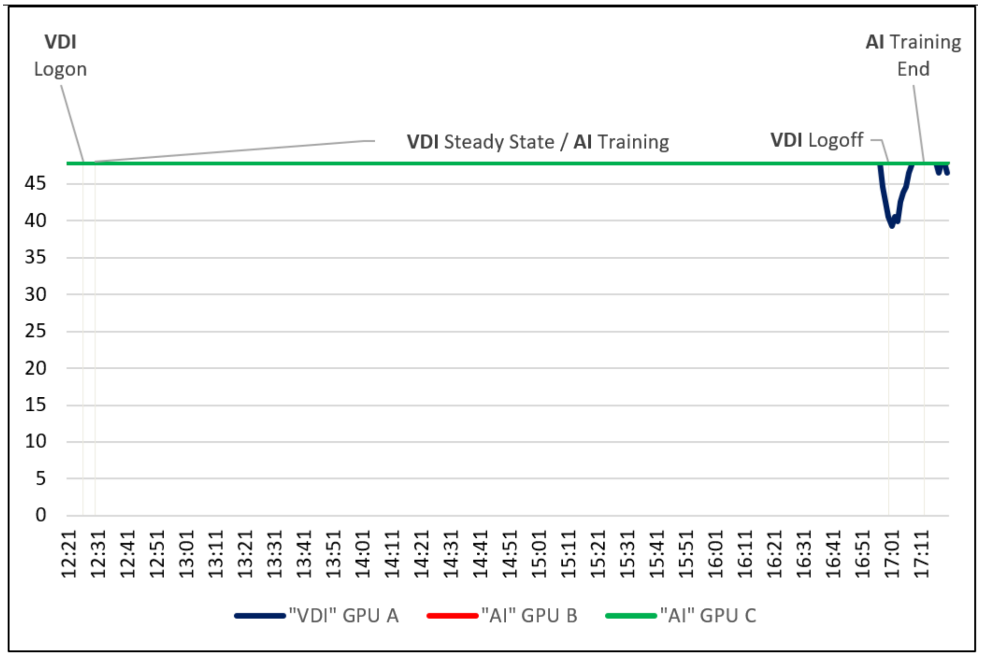


Figure 19. GPU memory usage (GB)

vSAN cluster latency

The following figure shows the vSAN cluster read and write latency for VDI steady-state and AI training. Cluster latency remained at an acceptable level with cluster read latency averaging 0.2 ms and cluster write latency averaging 0.8 ms during steady-state. No anomalies were seen during training.

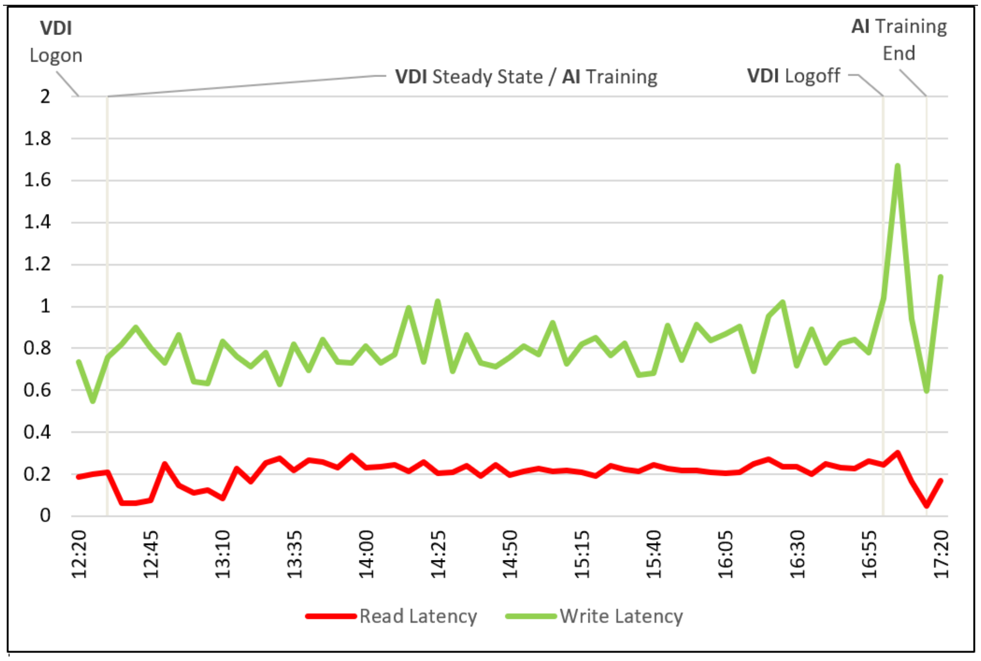


Figure 20. Cluster latency (ms)

Login VSI user experience

The following figure shows the Login VSI base, average, and maximum user experience response times. The VSI base is at an acceptable score. The VSI index average shows that the system responds appropriately as additional load is added and the system behaves in a predictable manner. The response times are tightly grouped, meaning that there is a consistent user experience across VDI sessions.

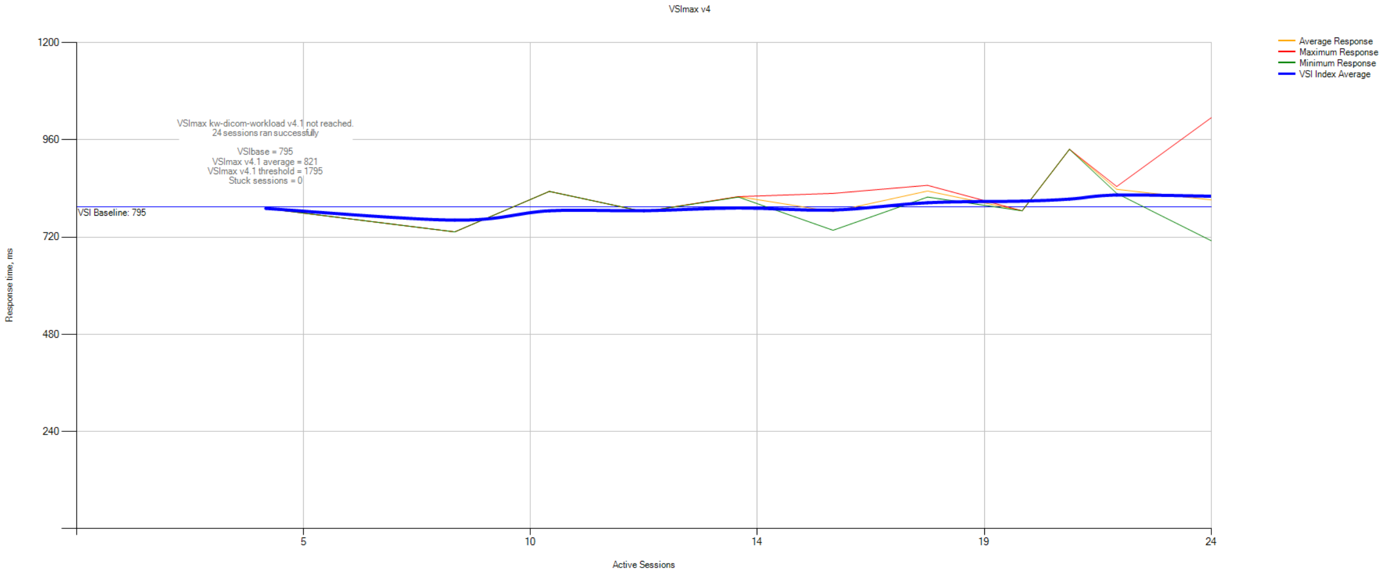


Figure 21. Login VSI user experience

Dual Workload AI Validation and VDI Medical Knowledge Worker Both GPU Enhanced

CPU utilization

The following figure shows the average CPU utilization for VDI steady-state and AI validation. The numerous CPU usage peaks are the result of the VDI medical workload phases on top of a latent AI Validation CPU requirement of 13 percent. CPU readiness averaged at 0.1 percent during steady-state and peaked at 0.7 percent during the log out phase.

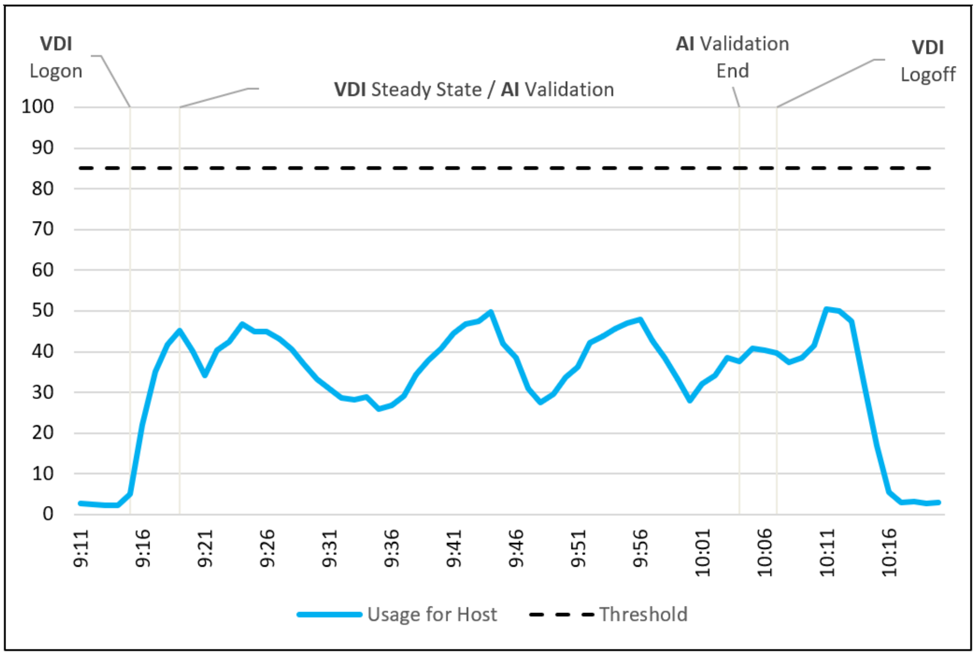


Figure 22. Host CPU utilization percentage

GPU utilization

The following figure shows the peak GPU utilization for VDI steady-state and AI validation. The two GPU cards show distinctly different GPU utilization patterns. Fifteen peaks can be seen for the GPU B which possibly coincides with the 15 Epoch models being validated.

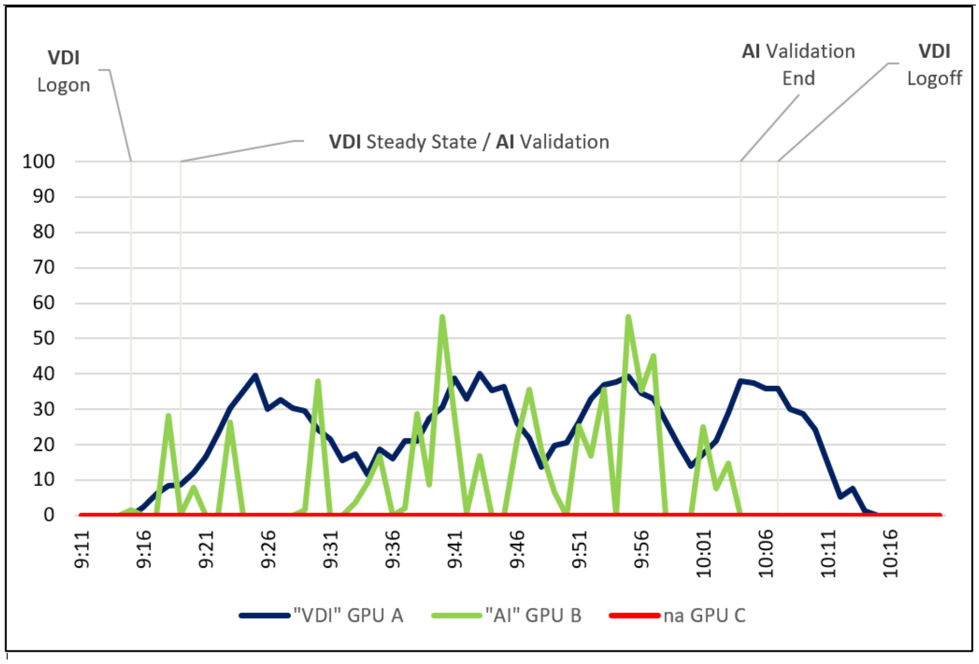


Figure 23. GPU utilization percentage

GPU memory

The following figure shows the peak GPU memory utilization for VDI steady-state and AI validation. The three GPU cards showed the same memory usage, consuming 100 percent of available memory during the login phase for steady-state and training. Memory usage dipped during VDI clone re-creation for GPU A before consuming 100 percent of available GPU memory when the VMs were re-created.

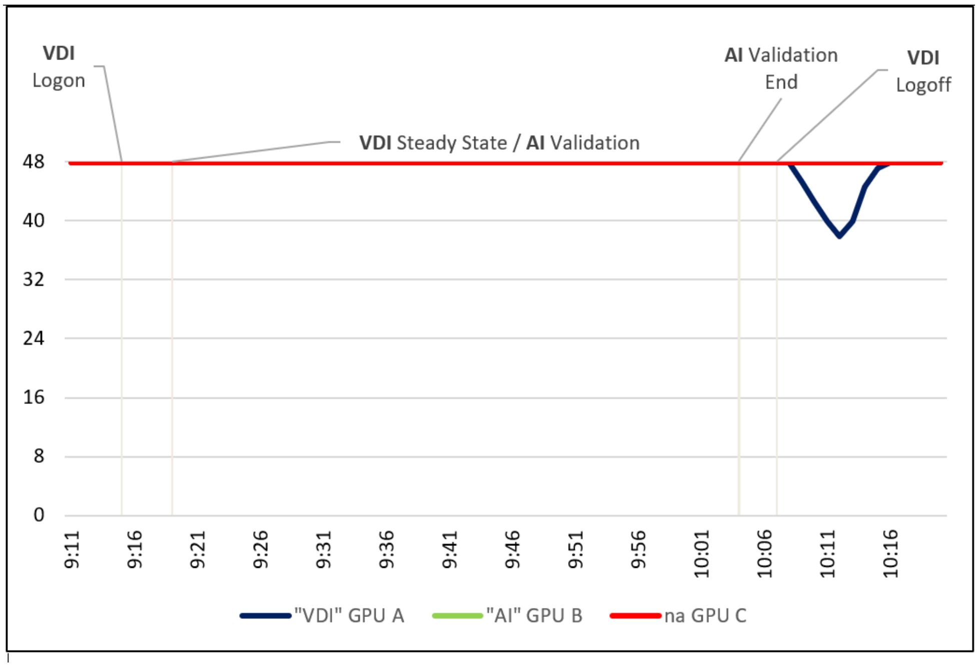


Figure 24. GPU memory (GB)

vSAN cluster latency

The following figure shows the vSAN cluster read and write latency for VDI steady-state and AI validation. Cluster latency remained at an acceptable level with cluster read latency averaging 0.1 ms and cluster write latency averaging 0.8 ms during steady-state. No anomalies were seen during training.

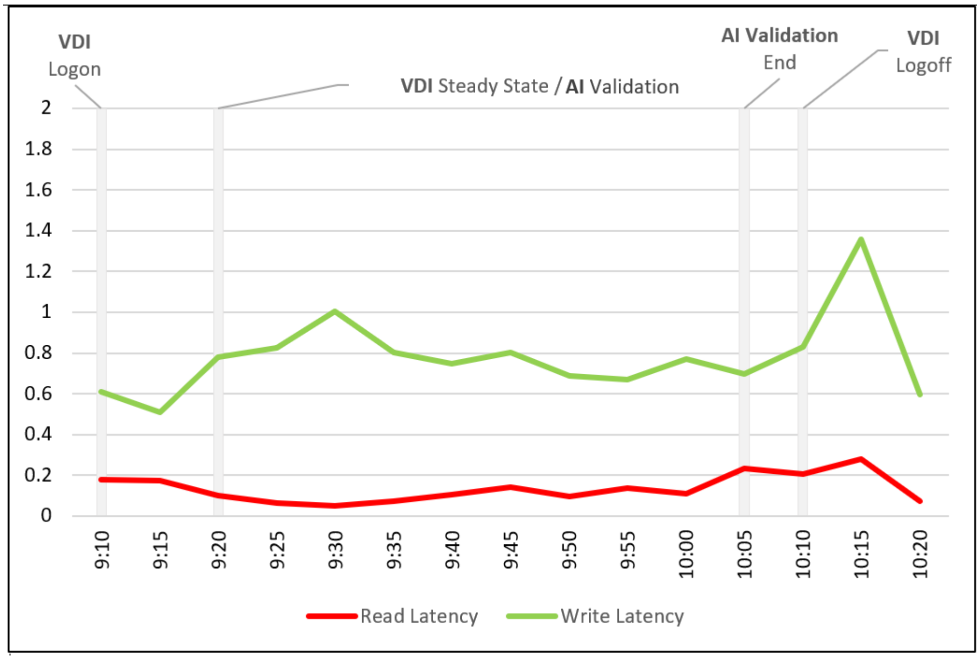


Figure 25. Cluster latency (ms)

Login VSI user experience

The following figure shows the Login VSI base, average, and maximum user experience response times. The VSI base was at an acceptable score. The VSI index average shows that the system responded appropriately as additional load was added and the system behaved in a predictable manner. The response times are tightly grouped, meaning that there is a consistent user experience across VDI sessions.

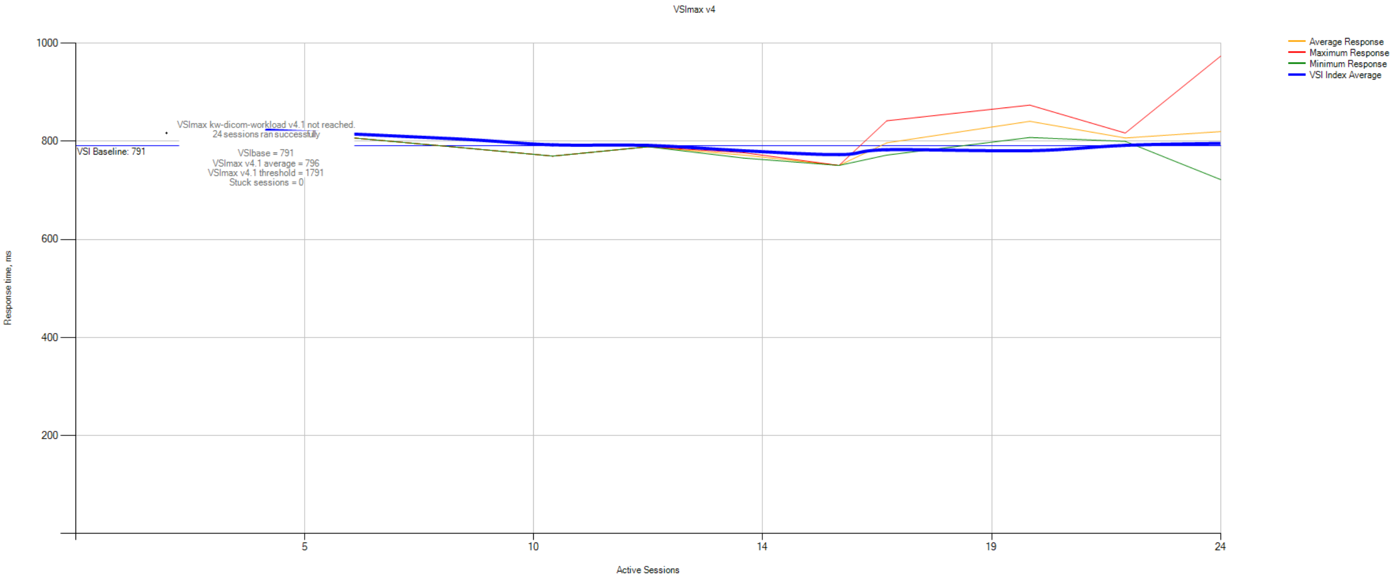


Figure 26. Login VSI user experience

Dual Workload AI Training CPU Only and VDI Medical Knowledge Worker GPU Enhanced

CPU utilization

The following figure shows the average CPU utilization for VDI steady-state and AI training. The numerous CPU usage peaks are the result of the VDI medical workload phases on top of a latent AI validation CPU requirement of 30 percent. CPU readiness averaged at 0.1 percent during steady-state and peaked at 100 percent spike during the logout phase.

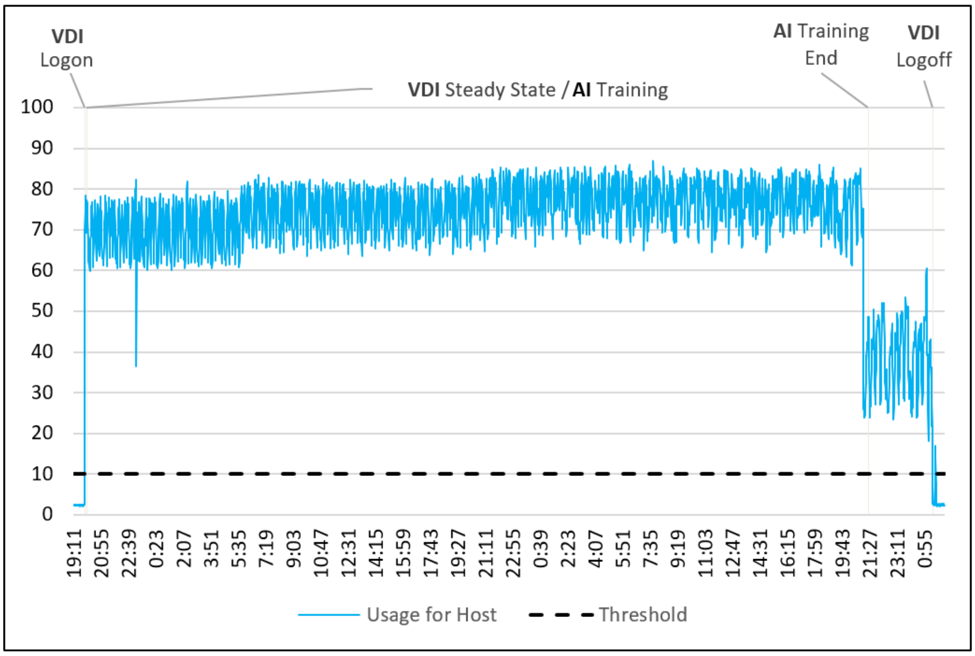


Figure 27. CPU utilization percentage

GPU utilization

The following figure shows the peak GPU utilization for VDI steady-state and AI training. A single GPU card was used for VDI graphics during the testing, with a uniform usage pattern.

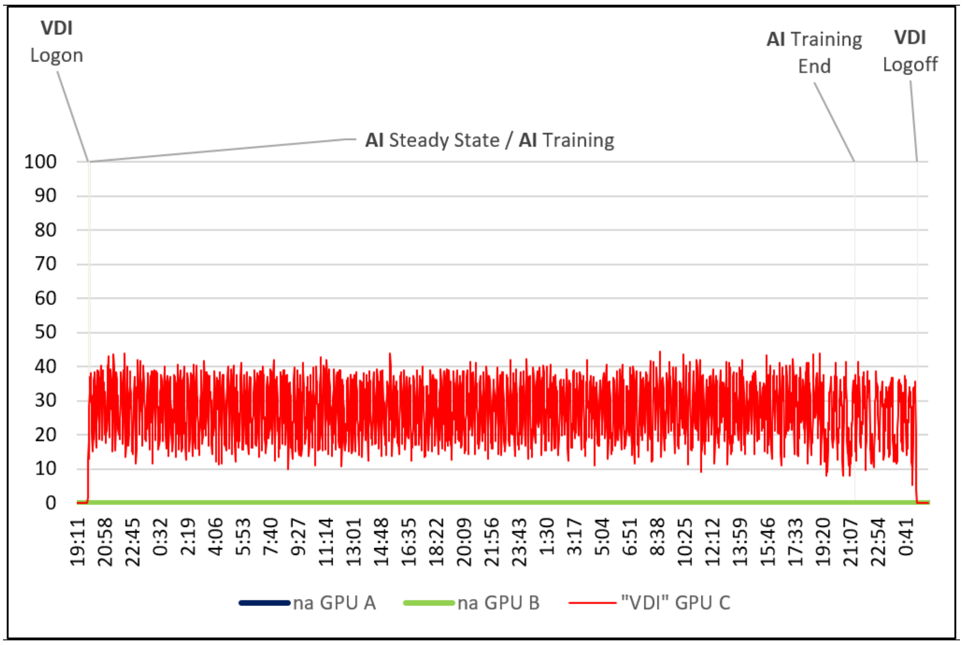


Figure 28. GPU utilization percentage

GPU memory

The following figure shows the peak GPU memory utilization for VDI steady-state and AI training. The GPU card consumed 100 percent of available memory during the login phase for steady-state and training.

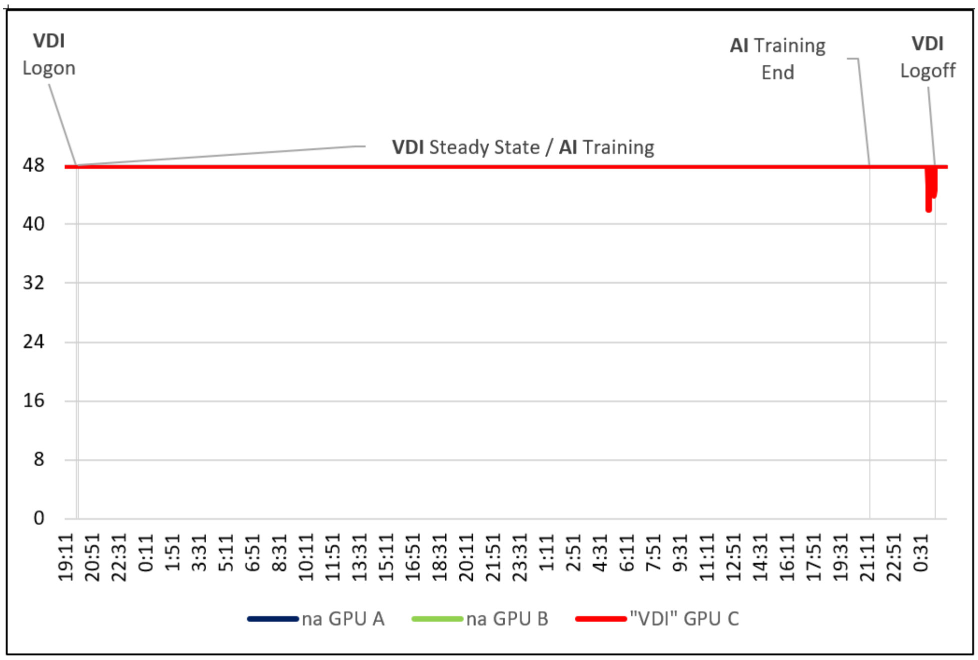


Figure 29. GPU memory (GB)

vSAN cluster latency

The following figure shows the vSAN cluster read and write latency for VDI steady-state and AI training. Cluster latency remained at an acceptable level with cluster read latency averaging 0.2 ms and cluster write latency averaging 1.3 ms during steady-state. No anomalies were seen during training. However, the write latency had an unusual pattern during steady-state, although that pattern was contained within a 0.5 ms to 2 ms envelope.

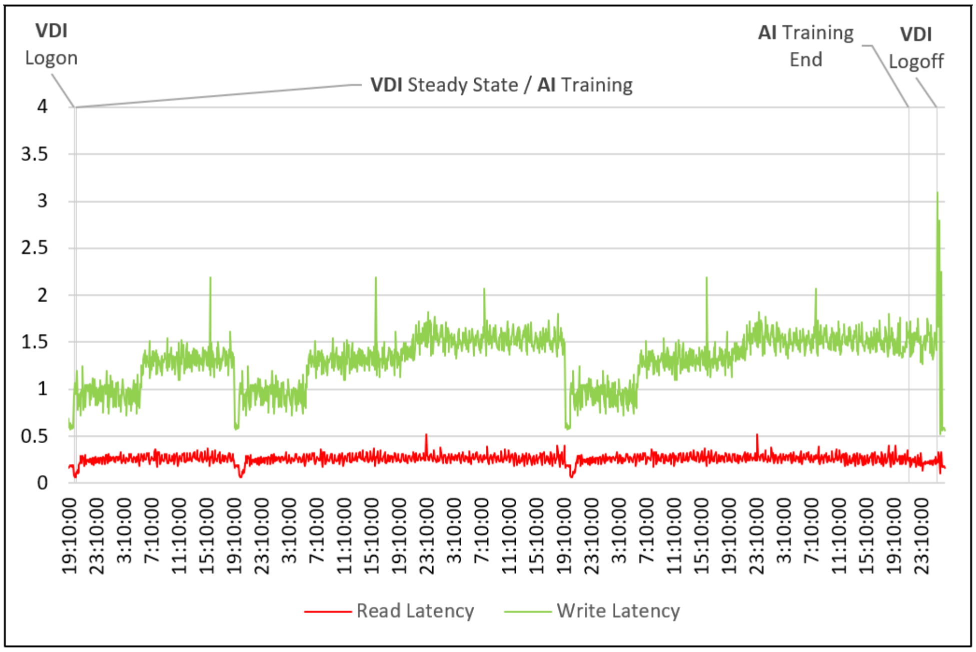


Figure 30. Cluster latency (ms)

Login VSI user experience

The following figure shows the Login VSI base, average, and maximum user experience response times. The VSI base was at an acceptable score. The VSI index average shows that the system responded appropriately as additional load was added and the system behaved in a predictable manner. The response times are tightly grouped, meaning that there is a consistent user experience across VDI sessions.

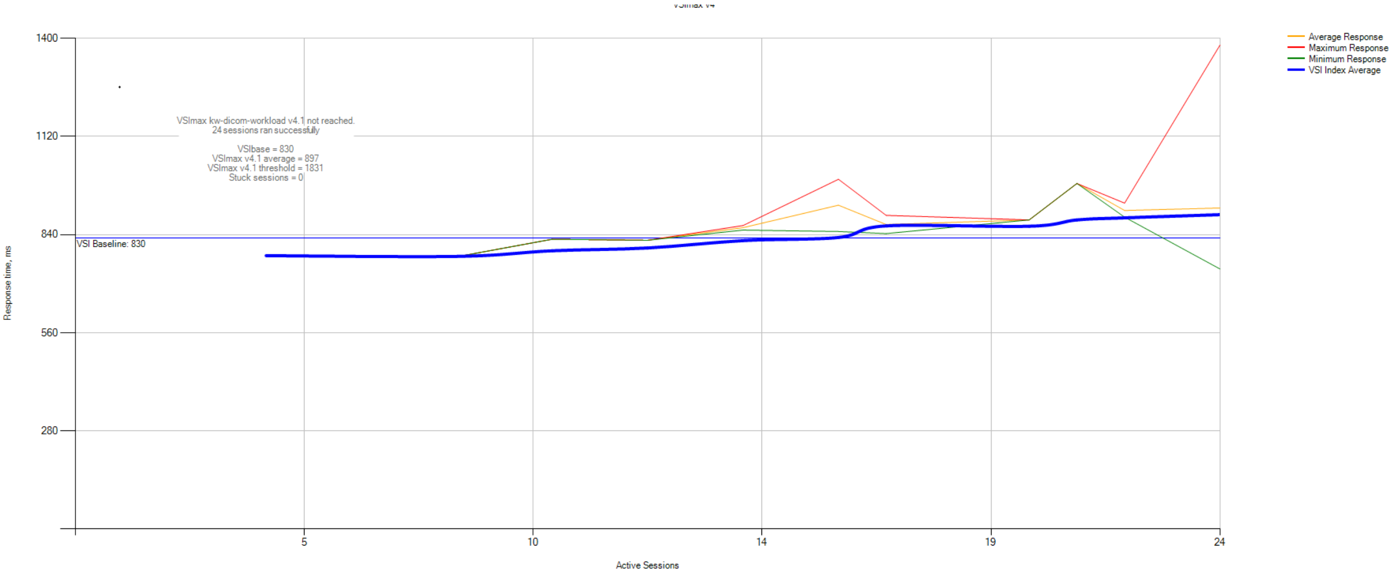


Figure 31. Login VSI user experience

Dual Workload AI Validation CPU Only and VDI Medical Knowledge Worker GPU Enhanced

CPU utilization

The following figure shows the average CPU utilization for VDI steady-state and AI validation. The numerous CPU usage peaks are a result of the VDI medical workload phases on top of a latent AI validation CPU requirement of 20 percent. CPU readiness averaged 0.29 percent during steady-state and peaked at 100 percent spike during the logout phase.

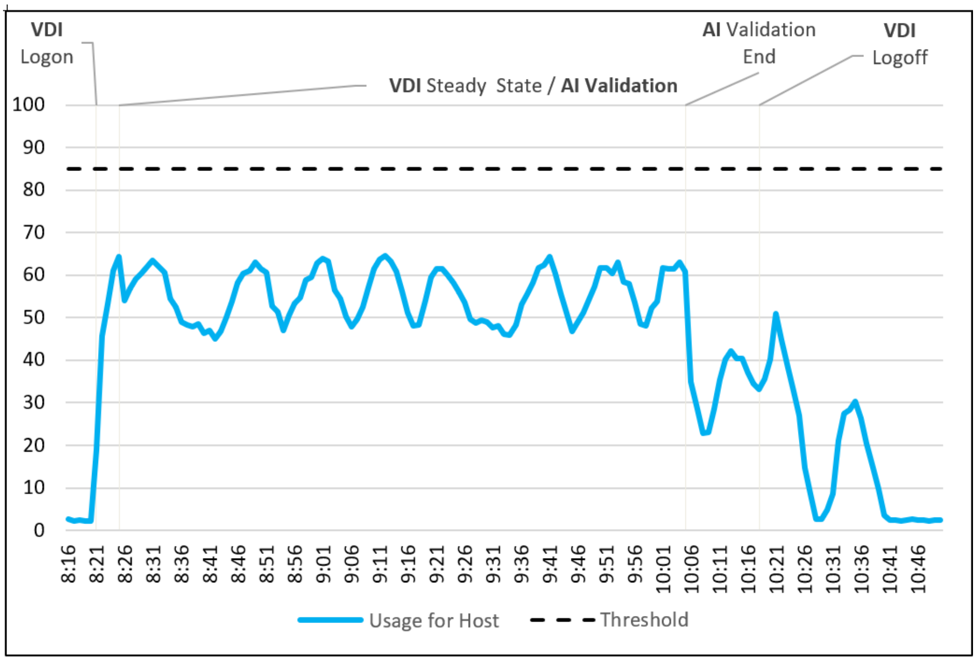


Figure 32. CPU utilization percentage

GPU utilization

The following figure shows the peak GPU utilization for VDI steady-state and AI validation. A single GPU card was used for VDI graphics during testing and shows a modulating but steady usage pattern.

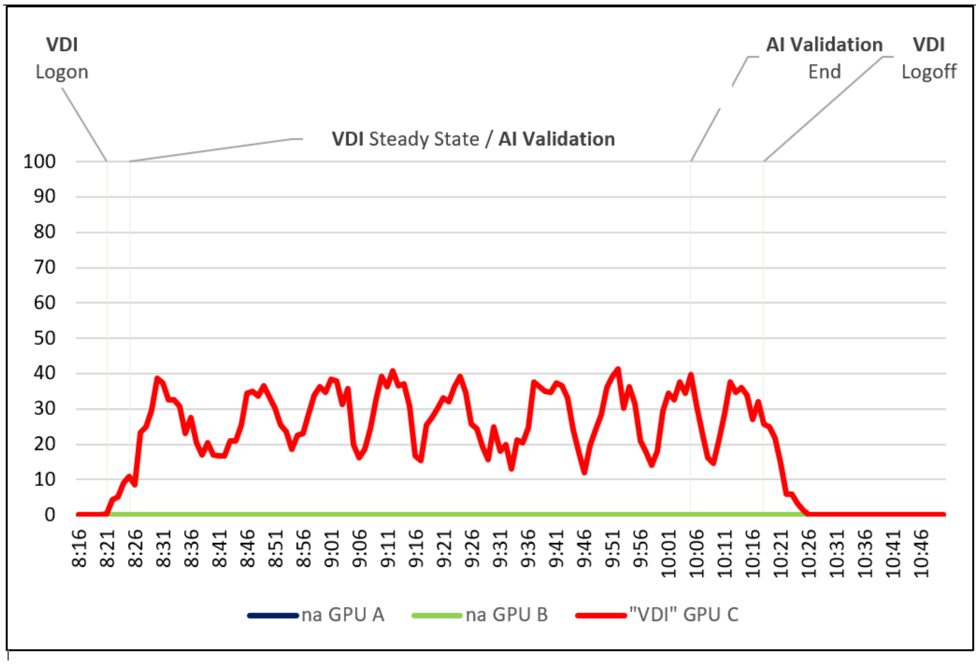


Figure 33. GPU utilization percentage

GPU memory

The following figure shows the peak GPU memory utilization for VDI steady-state and AI validation. The GPU card consumed about 50 percent of available memory during the login phase for steady-state and training.

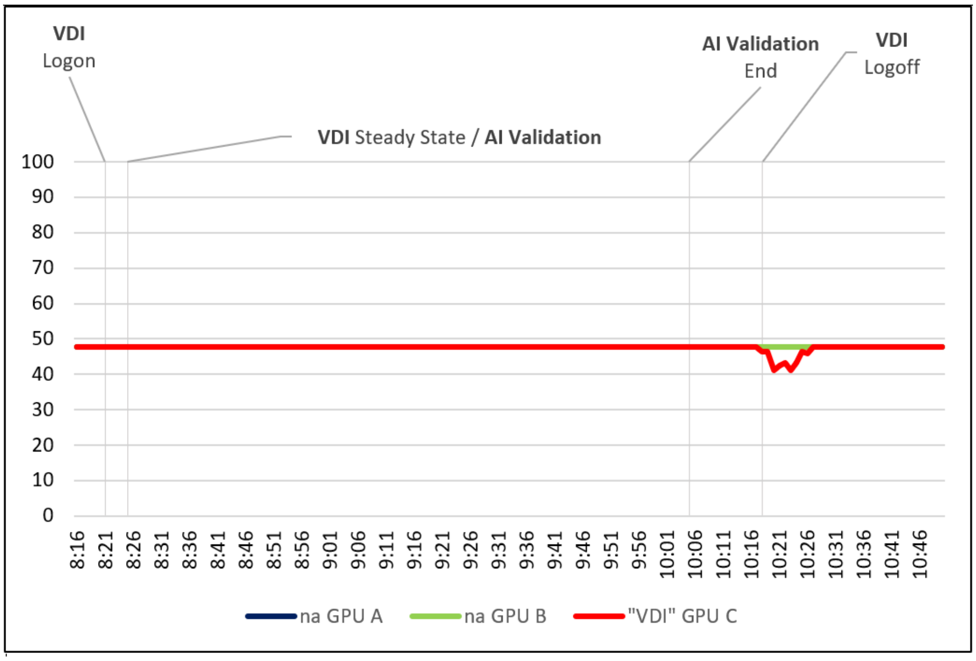


Figure 34. GPU memory (GB)

vSAN cluster latency

The following figure shows the vSAN cluster read and write latency for VDI steady-state and AI validation. Cluster latency remained at an acceptable level, with cluster read latency averaging 0.2 ms and cluster write latency averaging 1.0 ms during steady-state. No anomalies were seen during validation.

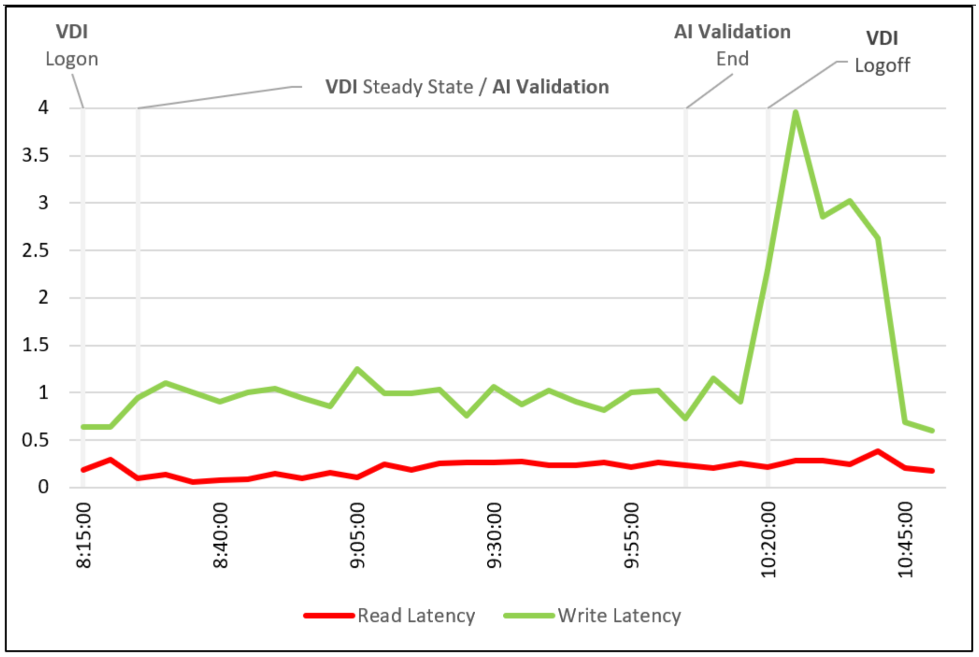


Figure 35. Cluster latency (ms)

Login VSI user experience

The following figure shows the Login VSI base, average, and maximum user experience response times. The VSI base is at an acceptable score. The VSI index average shows that the system responded appropriately as additional load was added and the system behaved in a predictable manner. The response times are tightly grouped, meaning that there is a consistent user experience across VDI sessions.

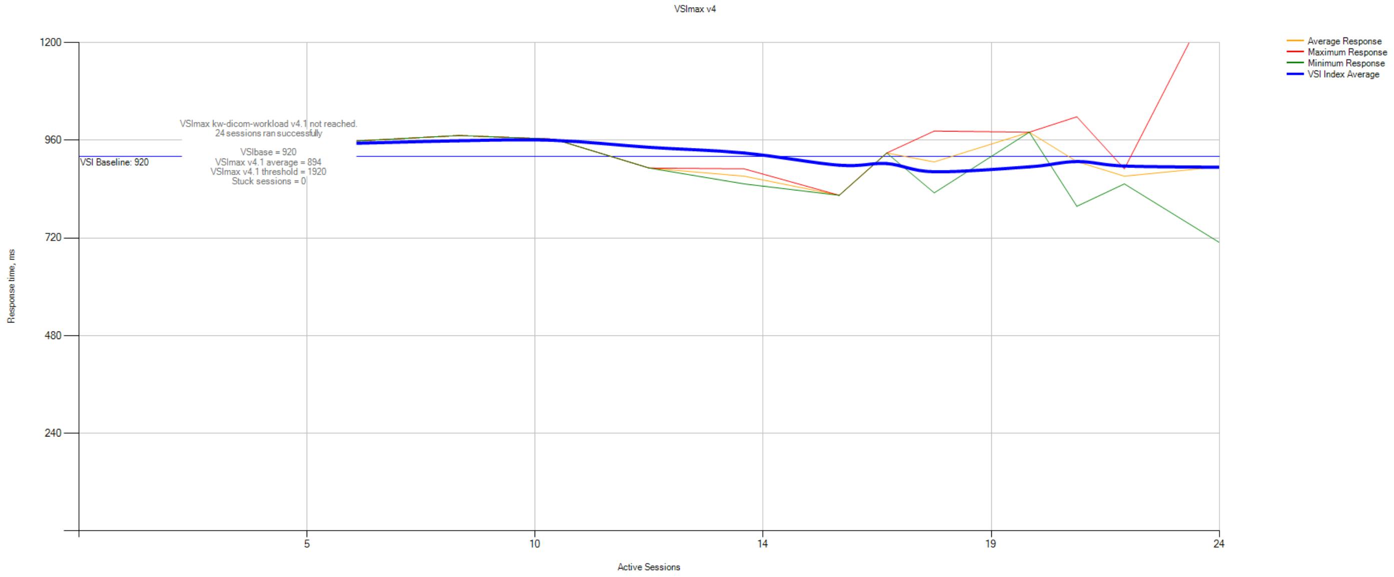


Figure 36. Login VSI user experience