NetApp Verified Architecture

NetApp HCI for Citrix Virtual Apps and Desktops with VMware vSphere

Suresh Thoppay, NetApp
June 2020 | NVA-1140 | Version 3.0

Abstract

This document describes the features of NetApp® HCI and NetApp ONTAP® storage that are relevant for Citrix Virtual Apps and Desktops running in a VMware vSphere environment. It discusses design considerations, best practices, sizing guidelines, and more.

NVIDIA CITRIX® NetApp®

TABLE OF CONTENTS

1 Executive Summary 6

2 Solution Overview 6

2.1 Target Audience 7

2.2 Solution Technology 7

2.3 Use Case Summary 9

3 NetApp HCI 9

3.1 Storage Nodes 10

3.2 Compute Nodes 10

4 NVIDIA Licensing 11

4.1 GRID Virtual PC 12

4.2 GRID Virtual Applications 12

4.3 Quadro Virtual Data Center Workstation 12

4.4 NVIDIA Virtual ComputeServer 12

5 GPU Usage with VMware vSphere 12

5.1 Virtual Dedicated Graphics 12

5.2 Virtual Shared Graphics 14

5.3 Virtual Shared Passthrough Graphics 15

6 Citrix Virtual Apps and Desktops 23

6.1 Citrix Workspace App (Receiver Client) 23

6.2 Citrix Gateway 23

6.3 Citrix StoreFront 23

6.4 Citrix Delivery Controller 24

6.5 Citrix Provisioning Services 24

6.6 Citrix Studio 25

6.7 Citrix License Server 26

6.8 Citrix Director 26

6.9 Virtual Delivery Agent 26

6.10 Database 27

6.11 File Shares 27

7 Technology Requirements 27

7.1 Hardware Requirements 28

7.2 Software Requirements 28

8 Deployment Steps 29
8.1 Prerequisites 29
8.2 Citrix License Server 31
8.3 Citrix Delivery Controller 33
8.4 Citrix StoreFront 48
8.5 Citrix Provisioning Server 57
8.6 Citrix Studio 76
8.7 Citrix Director 80
9 Solution Verification 85
9.1 Login VSI Test Results 86
9.2 Single-Server Login VSI Tests 86
9.3 NVIDIA nVector Test Results 91
9.4 Single-Server nVector SPECviewperf Workload Tests 102
10 Conclusion 123
Appendix A: NVIDIA nVector Tool Configuration Information 124
Where to Find Additional Information 126
NetApp 126
NVIDIA 126
Citrix 126
LakeSide 127
Version History 127

LIST OF TABLES

Table 1) vDGA: H610C compared to H615C. 14
Table 2) vGPU profiles for T4 and M10 GPU cards. 17
Table 3) NVIDIA T4 vGPU profiles on single GPU. 21
Table 4) Incorrect vGPU profile mix in single GPU. 22
Table 5) vGPU – H610C compared to H615C. 22
Table 6) Hardware requirements. 28
Table 7) Software requirements. 28
Table 8) NetApp HCI models. 29
Table 9) NetApp HCI H410C test summary. 88
Table 10) H610C Login VSI test summary. 89
Table 11) H615C Login VSI test summary. 90
Table 12) Login VSI test results summary. 91

LIST OF FIGURES

Figure 1) Citrix layers. 7
Figure 2) Solution architecture components. 8
Figure 3) HCI components. 10
Figure 4) Front view of H615C. 10
Figure 5) NVIDIA GRID GPUs. 11
Figure 6) 3D graphics on a VM. 15
Figure 7) NVIDIA vGPU architecture. 16
Figure 8) Citrix StoreFront. 24
Figure 9) Citrix PVS overview. 24
Figure 10) Citrix Studio. 25
Figure 11) Citrix Director. 26
Figure 12) H410C - Login VSI. 87
Figure 13) H410C Host CPU - Login VSI. 88
Figure 14) H610C - Login VSI. 89
Figure 15) H610C Host CPU - Login VSI. 89
Figure 16) H615C - Login VSI. 90
Figure 17) H615C Host CPU - Login VSI. 90
Figure 18) nVector test run. 92
Figure 19) H610C vSphere host CPU utilization. 93
Figure 20) Single Monitor H610C GPU utilization. 94
Figure 21) H615C Host CPU utilization. 94
Figure 22) H615C GPU utilization. 95
Figure 23) H615C end-user latency. 95
Figure 24) H615C average frame rate. 96
Figure 25) H610C host CPU utilization. 97
Figure 26) H610C GPU utilization. 97
Figure 27) H610C end-user latency. 98
Figure 28) H610C frame rates. 98
Figure 29) H610C image quality. 99
Figure 30) H615C host utilization. 100
Figure 31) H615C GPU utilization. 100
Figure 32) H615C end-user latency. 101
Figure 33) H615 frame rate. 101
Figure 34) H615C image quality. 102
Figure 35) SPECviewperf 13 viewsets. 103
Figure 36) Citrix policy for GPU testing. 104
Figure 37) 3ds Max composite scores. 105
Figure 38) 3ds Max sample. 106

Figure 39) Catia test score. 107
Figure 40) CATIA sample. 108
Figure 41) Creo test score. 109
Figure 42) Creo sample. 110
Figure 43) Energy test score. 111
Figure 44) Energy sample. 112
Figure 45) Maya test score. 113
Figure 46) Maya sample. 114
Figure 47) Medical test score. 116
Figure 48) Medical sample. 117
Figure 49) Showcase Test Result 118
Figure 50) Showcase sample. 119
Figure 51) Siemens test score. 120
Figure 52) Siemens NX sample. 121
Figure 53) SolidWorks test results. 122
Figure 54) SolidWorks sample. 123
Figure 55) Test summary results. 123

1 Executive Summary

Digital workspaces play an important role in the digital transformation journey of any enterprise. A digital workspace offers a unified, consistent user experience that can be accessed from any location. Users can access everything they need for work in a single interface, on the premises or in the cloud. A portable workspace allows users to work with any device they choose, which makes it attractive to the new generation of workers.

NetApp® HCI offers several models for hosting Citrix Virtual Apps and Desktops, including graphics processing unit (GPU) models targeted for the latest Windows 10 applications and 3D graphics workstation applications from Autodesk, Dassault systems, Siemens, and more. By offloading graphics workloads that are currently performed by CPUs and moving them to virtual GPUs, NVIDIA virtual GPU solutions free up CPU resources on server hardware and deliver a great user experience that rivals physical workstations.

NetApp HCI is a hybrid cloud infrastructure that can grow with business demands and can be easily repurposed as business needs change. And NetApp HCI storage provides powerful storage efficiency features such as global deduplication, compression, and thin provisioning.

The file services available in NetApp ONTAP® offer your team elegant solutions for the home directories, user profiles, and file shares, which can easily grow beyond traditional sizing limits. ONTAP includes easy-to-use NetApp Snapshot™ technology that enables self-service file recovery for end users. Adaptive quality-of-service (QoS) features allow you to automatically scale a throughput ceiling or floor to volume size, while maintaining the ratio of IOPS to TBs/GBs as the size of the volume changes.

NetApp HCI and ONTAP easily integrate into your existing automation framework. ONTAP has Ansible modules, PowerShell modules, and RESTful APIs for integration with other automation tools. Both products include many enterprise security features and monitoring options.

Citrix delivers software with flexible consumption and deployment models to meet your business needs, while the Citrix Director tool provides insights for the help desk.

2 Solution Overview

Today's mobile workforce wants to work from any location using their device of choice, which creates a significant challenge for administrators trying to keep up with security and management for these devices. Citrix Virtual Apps and Desktops (previously known as XenApp and XenDesktop) delivers fully functional virtual Windows desktops, Linux desktops, and applications that employees need to remain productive.

Citrix App Layering simplifies application and OS management by allowing administrators to manage apps, user profile data, and operating systems as individual layers. With this unique technology, IT can efficiently and securely deliver apps and desktops on the premises or from leading public cloud vendors. The Citrix App Layering technology containerizes desktop apps as virtual disks for streamlined delivery to end users. With user layers, administrators can easily deliver a persistent desktop experience to users. This system also delivers cost savings because of the efficiency of a pooled desktop environment.

Citrix High Definition eXperience (HDX) technology on Citrix Virtual Apps and Desktops gives users the best possible experience, whether on a local network, a mobile network, or a high-latency, wide-area network.

Citrix HDX 3D Pro optimizes the performance of professional-grade, graphics-intensive 3D applications for Windows and Linux virtual desktops. Slow-rendering graphics due to network lag are no longer an issue. Visually lossless deep compression enables users to collaborate in real time on pixel-perfect images.

While desktop virtualization technology has been around for many years, organizations still struggle with delivering a user experience that replicates what workers have enjoyed on physical PCs, the ability to

work from anywhere. NVIDIA virtual GPU solutions address the deficiencies of virtual desktop infrastructure (VDI) by bringing graphics and compute acceleration to the data center. With NetApp HCI accelerated by NVIDIA vGPU, you don't need a dedicated environment for a virtual desktop infrastructure (VDI). It can coexist with other workloads running on a cluster. Storage performance is guaranteed with minimum, maximum, and burst QoS policies. The NetApp HCI compute cluster can grow from two nodes to the maximum number of clusters supported by VMware vCenter.

2.1 Target Audience

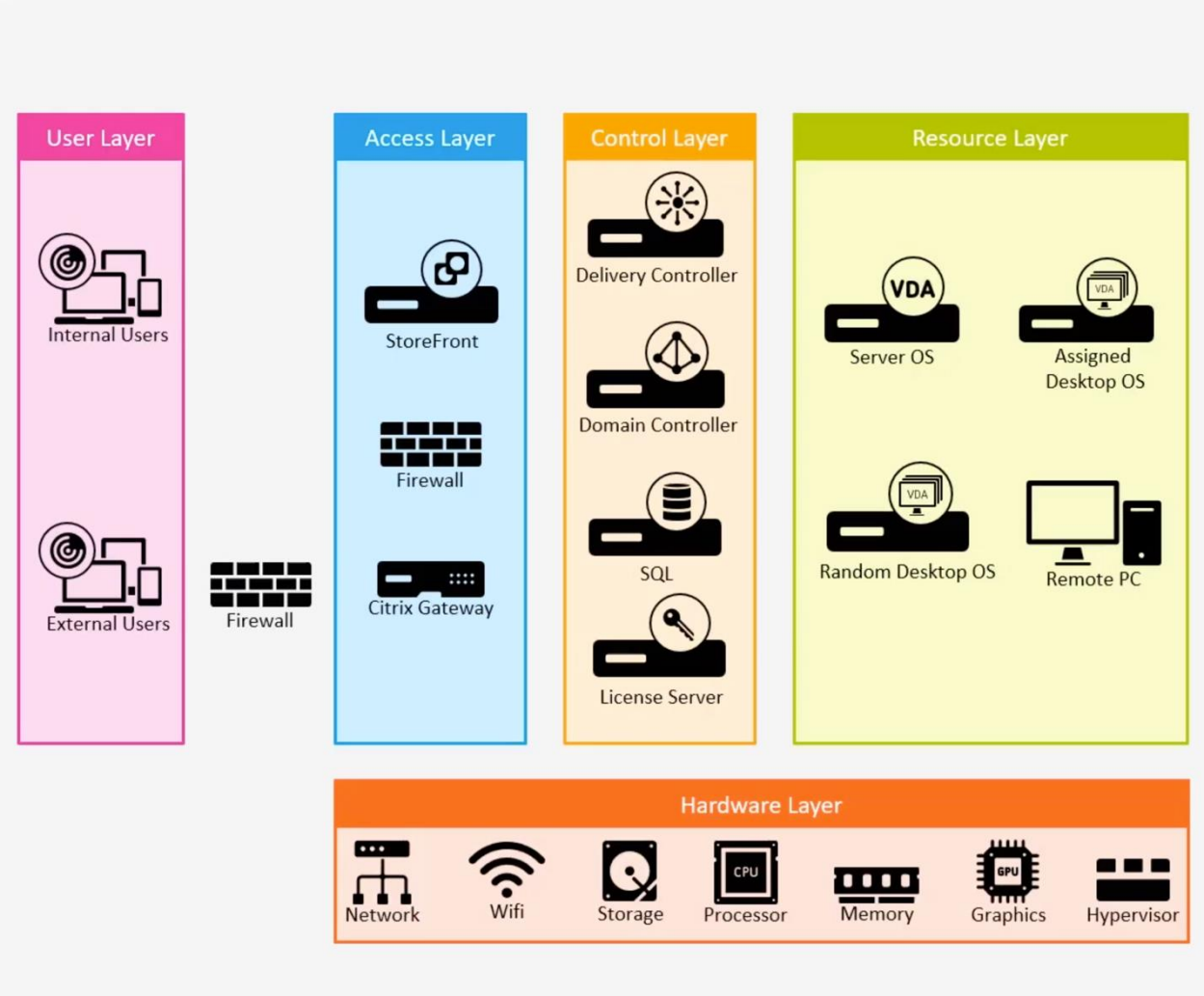
The target audience for this solution includes the following groups:

- Field consultants to help with design decisions for Citrix Virtual Apps and Desktops
- Executives and sales engineers to understand the value of the solution
- Professional services and IT managers to understand and identify the components of the solution
- Partners to assist customers who face similar challenges

2.2 Solution Technology

Citrix Virtual Apps and Desktops components are segregated into several layers, as shown in Figure 1.

Figure 1) Citrix layers.



Users accessing from on-premises locations (internal users) or from remote locations (external users) belong to the user layer. They access virtual apps and desktops using the StoreFront available on the access layer. External users connect to the StoreFront with the Citrix Gateway residing in a DMZ to provide secure access.

The delivery controller, which brokers user connections to desktops, resides in the control layer, along with other infrastructure components like the domain controller, database server, and license server. These components are typically deployed as virtual machines (VMs) on VMware vSphere with NetApp HCI.

The resource layer contains the VMs on which the virtual delivery agent (VDA) is installed. It includes:

- Server OS where applications are hosted
- Server OS for remote desktop sessions
- Single-session desktop OS on which applications are hosted
- Desktop OS for virtual desktops
- Existing physical machines that require remote access

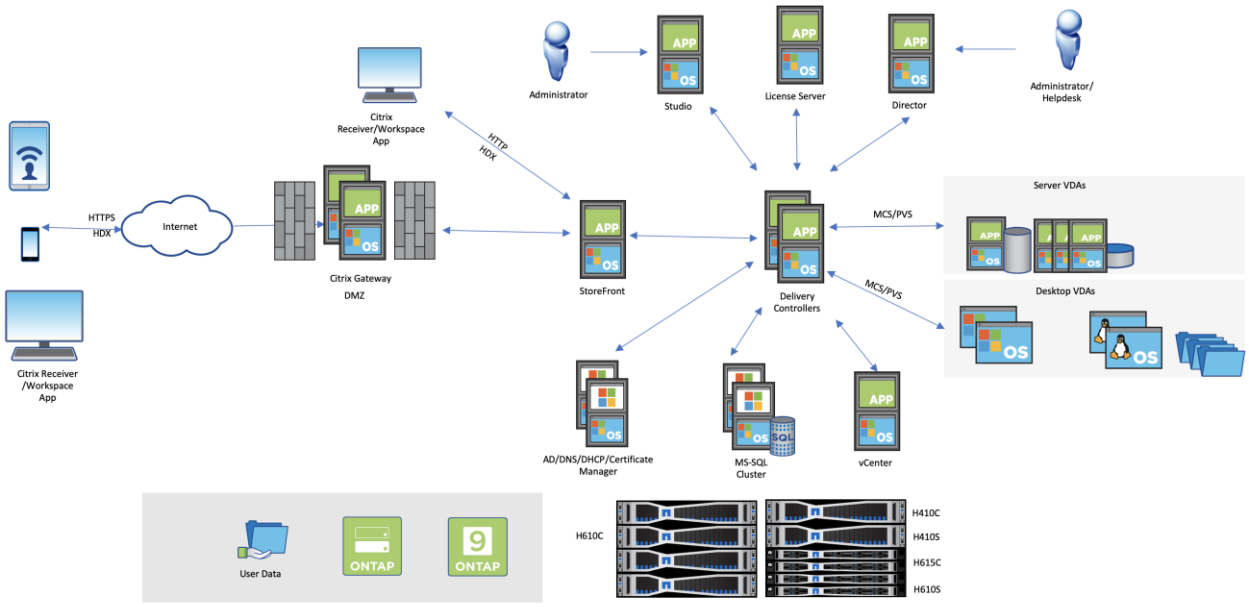
NetApp HCI belongs to the hardware layer, which provides compute, storage, memory, networking, and GPU resources, as well as the VMware hypervisor.

NetApp HCI consists of a mix of storage nodes and compute nodes. In the NetApp HCI H410 series, the storage and compute nodes are half-width blades that are located on a 2U chassis. H610S is a 1U storage node that provides up to 100,000 IOPS per node. NetApp HCI H610C compute node is a 2U compute node that contains two scalable, first-generation Intel processors and two NVIDIA Tesla M10 cards. NetApp HCI H615C is a 1U compute node that contains scalable second-generation Intel processors and a GPU with three NVIDIA Tesla T4 cards with the flexibility to run compute workloads in addition to graphics.

NetApp HCI provides the NetApp Deployment Engine (NDE) to automate the deployment and initial configuration of storage node clusters and compute clusters in a VMware vSphere environment.

Figure 2 shows the technical components of the solution.

Figure 2) Solution architecture components.



Citrix Director is a monitoring and troubleshooting console for helpdesk users and administrators. Citrix Studio is the primary management tool for managing Citrix Virtual Apps and Desktops.

The complete on-premises deployment, access, control, and resource layer resides on top of NetApp HCI in the hardware layer. As part of a hybrid-cloud deployment model, some parts of the resource layer can reside with the public cloud provider. For example, you can host Linux VMs on Amazon Web Services (AWS), or you can run multisession Windows 10 apps on Microsoft Azure.

In the Citrix Cloud deployment model, the access and control layer can be hosted on Citrix Cloud with resources on NetApp HCI, connected to Citrix Cloud by using Citrix Cloud Connector.

2.3 Use Case Summary

This solution can be applied to the following use cases:

- Freedom of choice for end-user devices
- Increased security of end devices
- Policy-based access control for user data
- Ease of management for multiple operating system versions
- Space-efficient storage solution for VDI
- High availability of file shares for user data
- Hosted application solution, including for graphics-intensive applications
- Remote access to existing graphics workstations

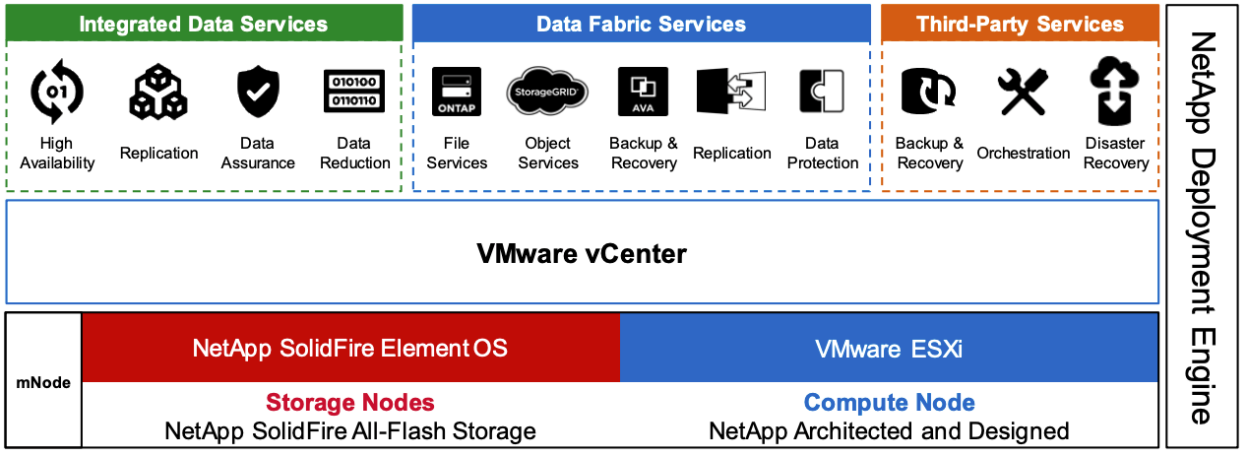
3 NetApp HCI

NetApp HCI is a hybrid cloud infrastructure that consists of a mix of storage nodes and compute nodes. It is available in either a two-rack unit or single-rack unit format, depending on the model. The installation and configuration required to deploy VMs are automated with the NDE. Compute clusters are managed with VMware vCenter, and storage clusters are managed by using the vCenter Plug-in deployed with NDE. A management VM called the mNode is deployed as part of NDE.

NetApp HCI handles the following functions:

- Version upgrades
- Pushing events to vCenter
- vCenter Plug-In management
- A VPN tunnel for support
- The NetApp Active IQ® collector
- The extension of NetApp Cloud Services to on the premises, enabling a hybrid cloud infrastructure

Figure 3) HCI components.



3.1 Storage Nodes

Storage nodes are available with either a half-width or full-width rack unit size. A minimum of four storage nodes is required to begin with, and a cluster can expand to up to 40 nodes. A storage cluster can be shared across multiple compute clusters. All the storage nodes contain a cache controller to improve write performance. A single node provides either 50K or 100K IOPS at a 4K block size.

NetApp HCI storage nodes run NetApp Element® software, which provides minimum, maximum, and burst QoS limits. The storage cluster supports a mix of storage nodes, although one storage node cannot exceed one-third of total capacity.

3.2 Compute Nodes

Compute nodes are available in half-width, full-width, and two rack-unit sizes. The NetApp HCI H410C and H610C are based on scalable Intel Skylake processors. The H615C is based on second-generation scalable Intel Cascade Lake processors. There are two compute models that contain GPUs: the H610C contains two NVIDIA M10 cards and the H615C contains three NVIDIA T4 cards.

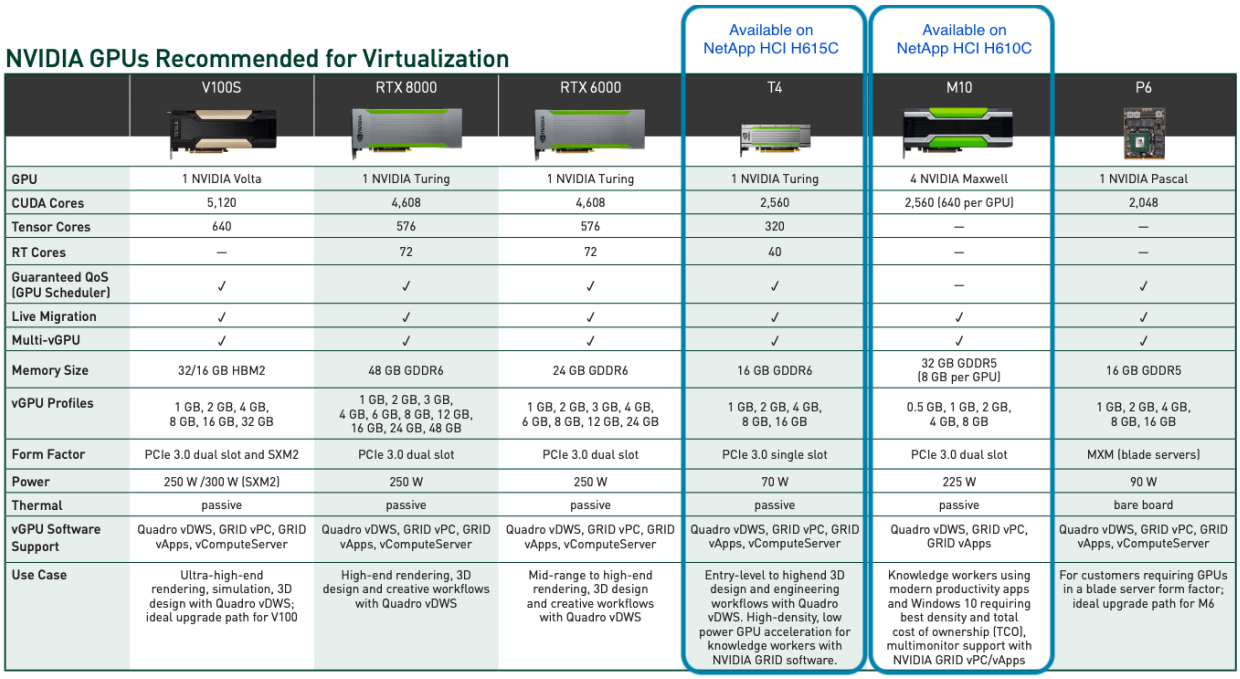
Figure 4) Front view of H615C.

The NVIDIA T4 has 40 RT cores that provide the computation power needed to deliver real-time ray tracing. The same server model used by designers and engineers can now also be used by artists to create photorealistic imagery that features light bouncing off surfaces just as it would in real life. This RTX-capable GPU produces real-time ray tracing performance of up to five Giga Rays per second. The NVIDIA T4, when combined with Quadro Virtual Data Center Workstation (Quadro vDWS) software, enables artists to create photorealistic designs with accurate shadows, reflections, and refractions on any device from any location.

Tensor cores enable you to run deep learning inferencing workloads. When running these workloads, an NVIDIA T4 powered with Quadro vDWS can perform up to 25 times faster than a VM driven by a CPU-only server. A NetApp H615C with three NVIDIA T4 cards in one rack unit is an ideal solution for graphics and compute-intensive workloads.

Figure 5 lists NVIDIA GPU cards and compares their features.

Figure 5) NVIDIA GRID GPUs.



The M10 GPU remains the best TCO solution for knowledge-worker use cases. However, the T4 makes a great alternative when IT wants to standardize on a GPU that can be used across multiple use cases, such as virtual workstations, graphics performance, real-time interactive rendering, and inferencing. With the T4, IT can take advantage of the same GPU resources to run mixed workloads—for example, running VDI during the day and repurposing the resources to run compute workloads at night.

The H610C compute node is two rack units in size; the H615C is one rack unit in size and consumes less power. The H615C supports H.264 and H.265 (High Efficiency Video Coding [HEVC]) 4:4:4 encoding and decoding. It also supports a VP9 decoder, which is becoming more mainstream; even the WebM container package served by YouTube uses the VP9 codec for video.

The number of nodes in a compute cluster is dictated by VMware; currently, it is 64. Mixing different models of compute nodes in a cluster is supported when Enhanced vMotion Compatibility (EVC) is enabled. For GPU nodes with default graphics settings (virtual shared graphics mode), compute models can be mixed in a cluster.

4 NVIDIA Licensing

When using an H610C or H615C, the license for the GPU must be procured from NVIDIA partners that are authorized to resell the licenses. You can find NVIDIA partners with the partner locator. Search for competencies such as virtual GPU (vGPU) or Tesla.

NVIDIA vGPU software is available in four editions:

- NVIDIA GRID Virtual PC (GRID vPC)
- NVIDIA GRID Virtual Applications (GRID vApps)
- NVIDIA Quadro Virtual Data Center Workstation (Quadro vDWS)
- NVIDIA Virtual ComputeServer (vComputeServer)

4.1 GRID Virtual PC

This product is ideal for users who want a virtual desktop that provides a great user experience for Microsoft Windows applications, browsers, high-definition video, and multi- monitor support. The NVIDIA GRID Virtual PC delivers a native experience in a virtual environment, allowing you to run all your PC applications at full performance.

4.2 GRID Virtual Applications

GRID vApps are for organizations deploying a Remote Desktop Session Host (RDSH) or other app-streaming or session-based solutions. Designed to deliver Microsoft Windows applications at full performance, Windows Server-hosted RDSH desktops are also supported by GRID vApps.

4.3 Quadro Virtual Data Center Workstation

This edition is ideal for mainstream and high-end designers who use powerful 3D content creation applications like Dassault CATIA, SOLIDWORKS, 3Dexcite, Siemens NX, PTC Creo, Schlumberger Petrel, or Autodesk Maya. NVIDIA Quadro vDWS allows users to access their professional graphics applications with full features and performance anywhere on any device.

4.4 NVIDIA Virtual ComputeServer

Many organizations run compute-intensive server workloads such as artificial intelligence (AI), deep learning (DL), and data science. For these use cases, NVIDIA vComputeServer software virtualizes the NVIDIA GPU, which accelerates compute-intensive server workloads with features such as error correction code, page retirement, peer-to-peer over NVLink, and multi-vGPU.

Note: A Quadro vDWS license enables you to use GRID vPC and NVIDIA vComputeServer.

5 GPU Usage with VMware vSphere

VMware vSphere became a trusted platform for virtualization by effectively using underlying resources and providing high availability for applications. The latest GPU driver can be downloaded from the NVIDIA site and installed on a vSphere host.

Note: You can use VMware Update Manager to deploy NVIDIA vGPU software on multiple hosts. Use the offline bundle .zip files to create a baseline of the type Host Extension.

VMs consume GPU resources in one of the following ways:

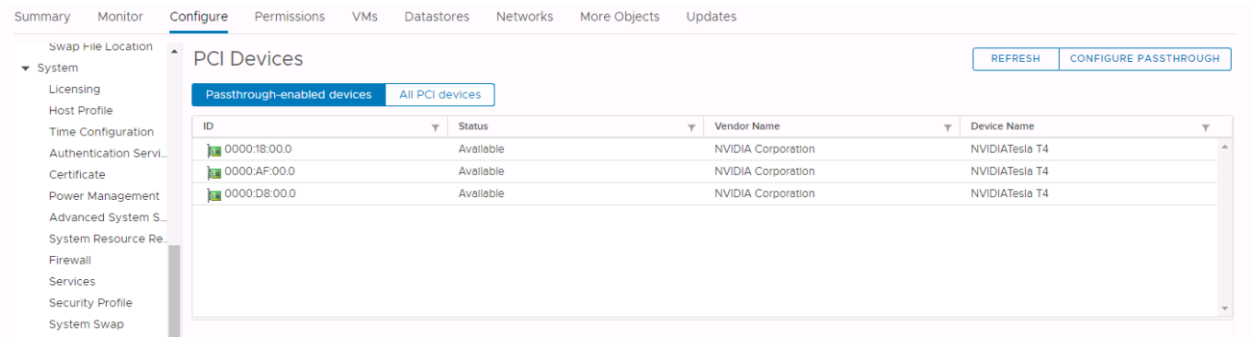
- Virtual Dedicated Graphics (vDGA)
- Virtual Shared Graphics (vSGA)
- Virtual Shared Passthrough Graphics (NVIDIA vGPU)

5.1 Virtual Dedicated Graphics

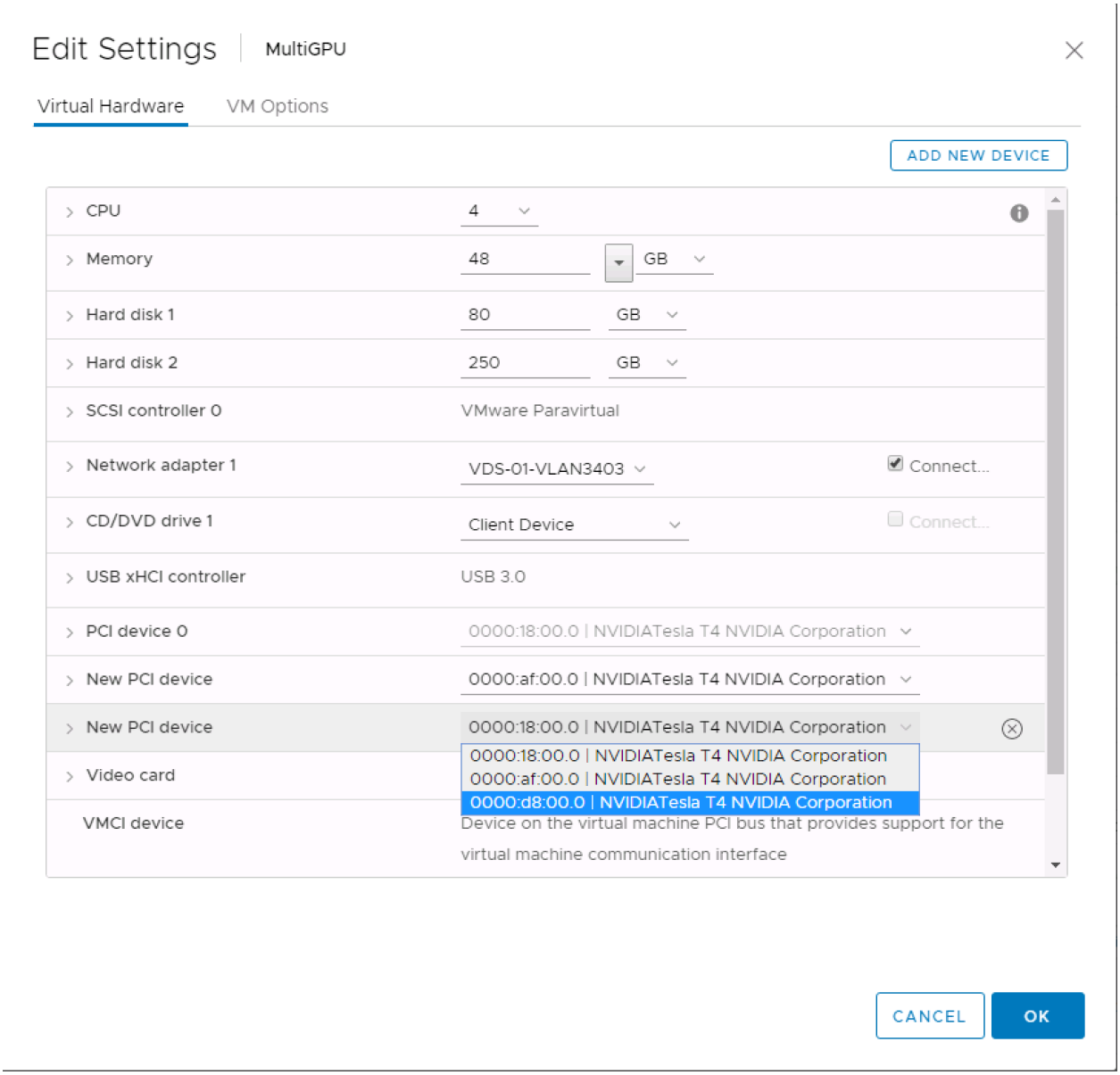
A VM has complete access to the GPU by using the PCI passthrough option with direct I/O. However, certain vSphere features, such as snapshots, vMotion, and so on are not supported. The VM performs at a native level.

To configure vDGA, follow these steps:

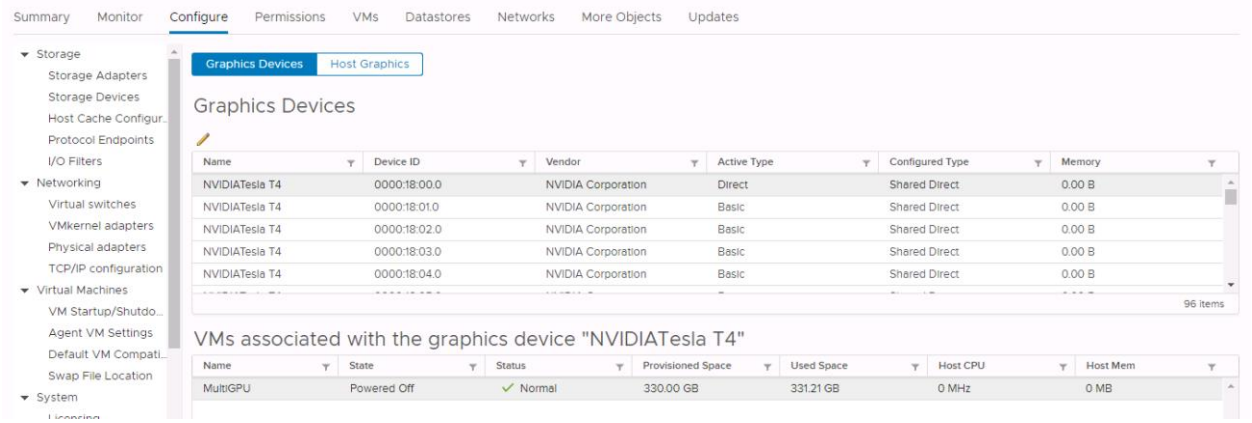
1. In the Hardware section, click PCI Devices and then click Passthrough-Enabled Devices.



2. Edit VM Settings and add the PCI device.



3. When the GPU is configured for vDGA mode, the Graphics Devices list shows the Active Type as Direct.



Before vSphere 6.7 update 2, vDGA was the only option of the three (vDGA, vSGA, or vGPU) listed that enabled the VM to access multiple GPUs. An NVIDIA vGPU driver is optional on the vSphere host. However, an OS driver is required inside the VM. An NVIDIA license server should be available to check licenses in and out. Without a license, only a very limited feature set is available.

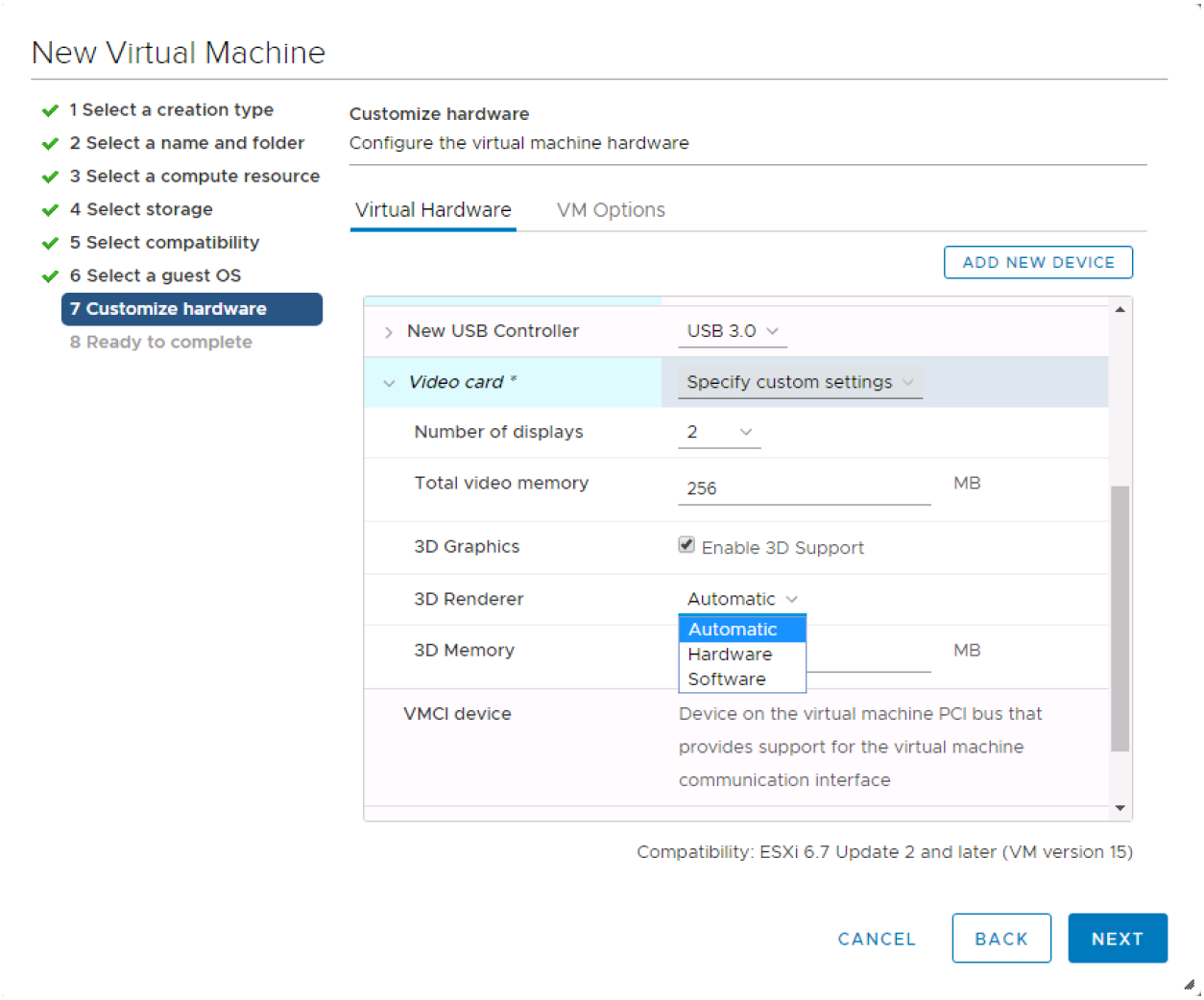
Table 1) vDGA: H610C compared to H615C.

Feature	H610C	H615C
Maximum number of passthrough devices per server	8	3

5.2 Virtual Shared Graphics

Virtual Shared Graphics is the default mode enabled on VMware vSphere. NVIDIA vGPU software is required on the vSphere host to use hardware-based rendering. When a GPU is not present, it uses CPU cycles to provide software-based rendering. To use virtual shared graphics, enable 3D support and install VMware Tools on the VM.

Figure 6) 3D graphics on a VM.



The VMware device driver provides limited support for DirectX and OpenGL There is also limited support for 4K monitors. Frame buffer memory is limited to 2GB.

H610C and H615C nodes can be part of same vSphere cluster with Virtual Shared Graphics mode when Enhanced vMotion Compatibility (EVC) is enabled.

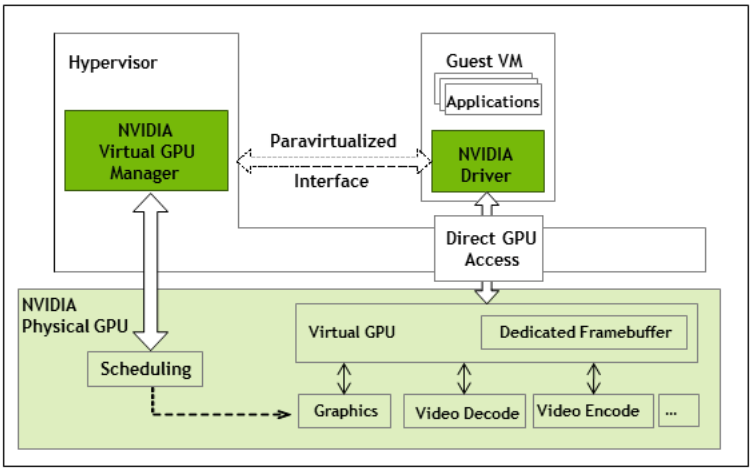
Note: vDGA uses PCI passthrough for a GPU card to a given VM. vDGA is not recommended because it does not support many vSphere features and provides low resource utilization. vSGA enables multiple VMs to leverage physical GPUs installed locally in the ESXi hosts to provide hardware-accelerated 3D graphics. vSGA is not recommended because graphics API support is limited and only some versions of DirectX and OpenGL are supported. Additionally, there is no Compute Unified Device Architecture (CUDA) support for vSGA. With vSGA, the VDI instance relies on the VMware vSGA driver, which receives access through an Xorg server running on the hypervisor. This is suboptimal from a performance standpoint. With vGPU, NVIDIA technology is used throughout, and a VDI instance gets the closest possible parity to running a native NVIDIA driver.

5.3 Virtual Shared Passthrough Graphics

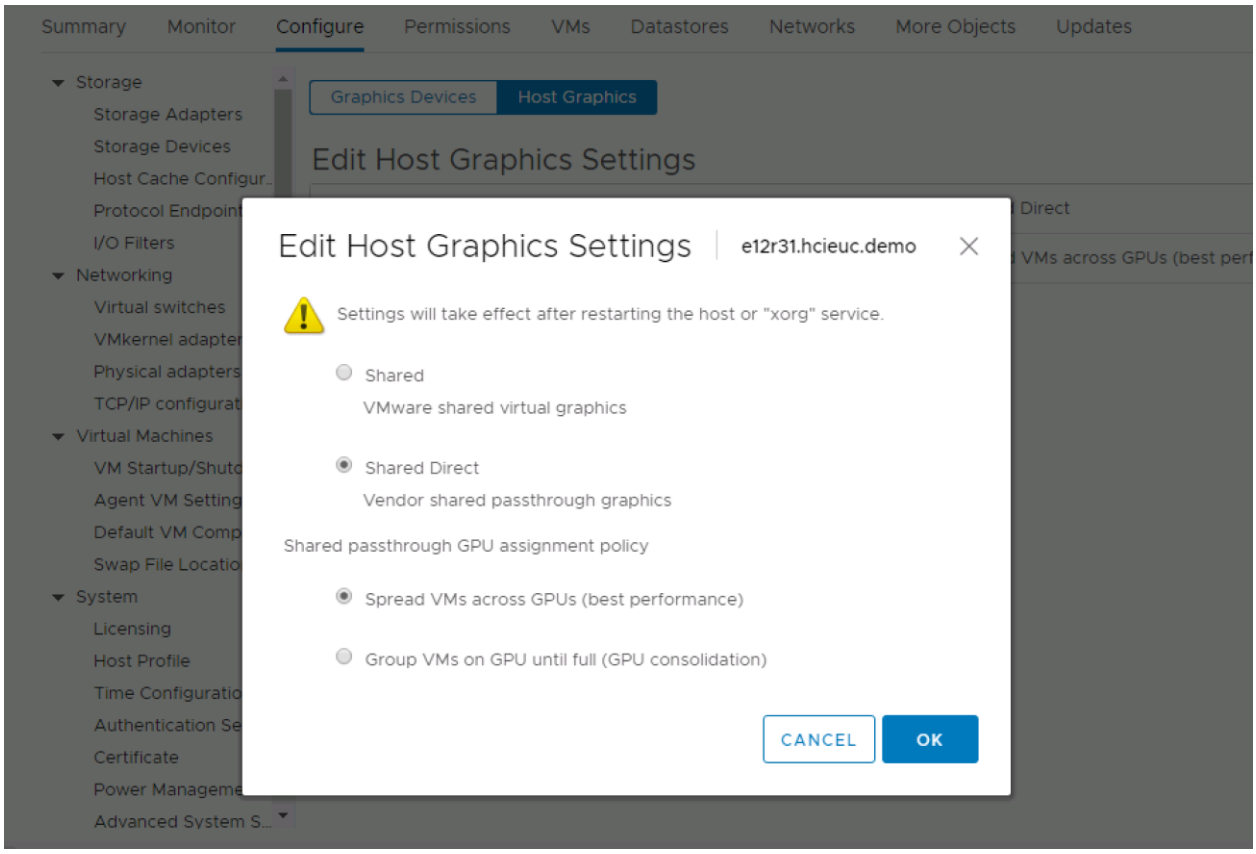
Virtual Shared Passthrough Graphics provides better utilization of GPU resources because each VM has its own dedicated frame buffer. However, the GPU compute, encoder, decoder, and so on are shared. In

a manner similar to a hypervisor sharing resources with VMs, NVIDIA GRID software manages the resource scheduling of GPU requests from a VM.

Figure 7) NVIDIA vGPU architecture.



To use Virtual Shared Passthrough Graphics mode, the host graphics setting must be changed to Shared Direct mode, as shown the following screenshot.



In VM Settings, add Shared PCI Device and choose the required vGPU profile (see Table 2) based on the frame buffer. You must reserve memory to enable direct access for the PCI device.

Table 2) vGPU profiles for T4 and M10 GPU cards.

Virtual GPU Type	Intended Use Case	Frame Buffer (MB)	Virtual Display Heads	Maximum Resolution per Display Head	Maximum vGPUs per GPU	Maximum vGPUs per H615C/H610C Server	Required License Edition
T4-16Q	Virtual workstations	16384	4	4096x2160	1	3	Quadro vDWS
T4-8Q	Virtual workstations	8192	4	4096x2160	2	6	Quadro vDWS
T4-4Q	Virtual workstations	4096	4	4096x2160	4	12	Quadro vDWS
T4-2Q	Virtual workstations	2048	4	4096x2160	8	24	Quadro vDWS
T4-1Q	Virtual desktops, virtual workstations	1024	2	4096x2160	16	48	Quadro vDWS
T4-16C	Training workloads	16384	1	4096x2160	1	3	vComputeServer or Quadro vDWS
T4-8C	Training workloads	8192	1	4096x2160	2	6	vComputeServer or Quadro vDWS
T4-4C	Inference workloads	4096	1	4096x2160	4	12	vComputeServer or Quadro vDWS
T4-2B	Virtual desktops	2048	2	4096x2160	8	24	GRID vPC or Quadro vDWS
T4-2B4	Virtual desktops	2048	4	2560x1600	8	24	GRID vPC or Quadro vDWS
T4-1B	Virtual desktops	1024	4	2560x1600	16	48	GRID vPC or Quadro vDWS

Virtual GPU Type	Intended Use Case	Frame Buffer (MB)	Virtual Display Heads	Maximum Resolution per Display Head	Maximum vGPUs per GPU	Maximum vGPUs per H615C/H610C Server	Required License Edition
T4-1B4	Virtual desktops	1024	1	4096x2160	16	48	GRID vPC or Quadro vDWS
T4-16A	Virtual applications	16384	1	1280x1024	1	3	GRID vApps
T4-8A	Virtual applications	8192	1	1280x1024	2	6	GRID vApps
T4-4A	Virtual applications	4096	1	1280x1024	4	12	GRID vApps
T4-2A	Virtual applications	2048	1	1280x1024	8	24	GRID vApps
T4-1A	Virtual applications	1024	1	1280x1024	16	48	GRID vApps
M10-8Q	Virtual workstations	8192	4	4096x2160	4	8	Quadro vDWS
M10-4Q	Virtual workstations	4096	4	4096x2160	8	16	Quadro vDWS
M10-2Q	Virtual workstations	2048	4	4096x2160	16	32	Quadro vDWS
M10-1Q	Virtual desktops, virtual workstations	1024	2	4096x2160	32	64	Quadro vDWS

Virtual GPU Type	Intended Use Case	Frame Buffer (MB)	Virtual Display Heads	Maximum Resolution per Display Head	Maximum vGPUs per GPU	Maximum vGPUs per H615C/H610C Server	Required License Edition
M10-0Q	Virtual desktops, virtual workstations	512	2	2560x1600	64	128	Quadro vDWS
M10-2B	Virtual desktops	2048	2	4096x2160	16	32	GRID vPC or Quadro vDWS
M10-2B4	Virtual desktops	2048	4	2560x1600	16	32	GRID vPC or Quadro vDWS
M10-1B	Virtual desktops	1024	4	2560x1600	32	64	GRID vPC or Quadro vDWS
M10-1B4	Virtual desktops	1024	1	4096x2160	32	64	GRID vPC or Quadro vDWS
M10-0B	Virtual desktops	512	2	2560x1600	64	128	GRID vPC or Quadro vDWS
M10-8A	Virtual applications	8192	1	1280x1024	4	8	GRID vApps
M10-4A	Virtual applications	4096	1	1280x1024	8	16	GRID vApps
M10-2A	Virtual applications	2048	1	1280x1024	16	32	GRID vApps
M10-1A	Virtual applications	1024	1	1280x1024	32	64	GRID vApps

Note: For NVIDIA vGPU mode, do not enable 3D support on the VM. For more information, see the NVIDIA vGPU User Guide. While using with T4 cards, consider choosing appropriate core and

speed in BIOS for Speed Select option. For vGPU with large frame buffers, consider selecting higher speed option compared to number of cores.

GRID vPC (profiles typically end with B) is used for virtual desktops, and GRID vApps (profiles end with A) is used for hosted apps. Most of the Q profiles (Quadro vDWS) support up to four 4K resolution monitors, which provides an enhanced user experience for image editing tools and support for professional graphics workloads.

The NVIDIA T4 GPU supports multiple profiles. NVIDIA recommends the GRID vPC 1GB profile to deliver an optimal experience for standard knowledge workers:

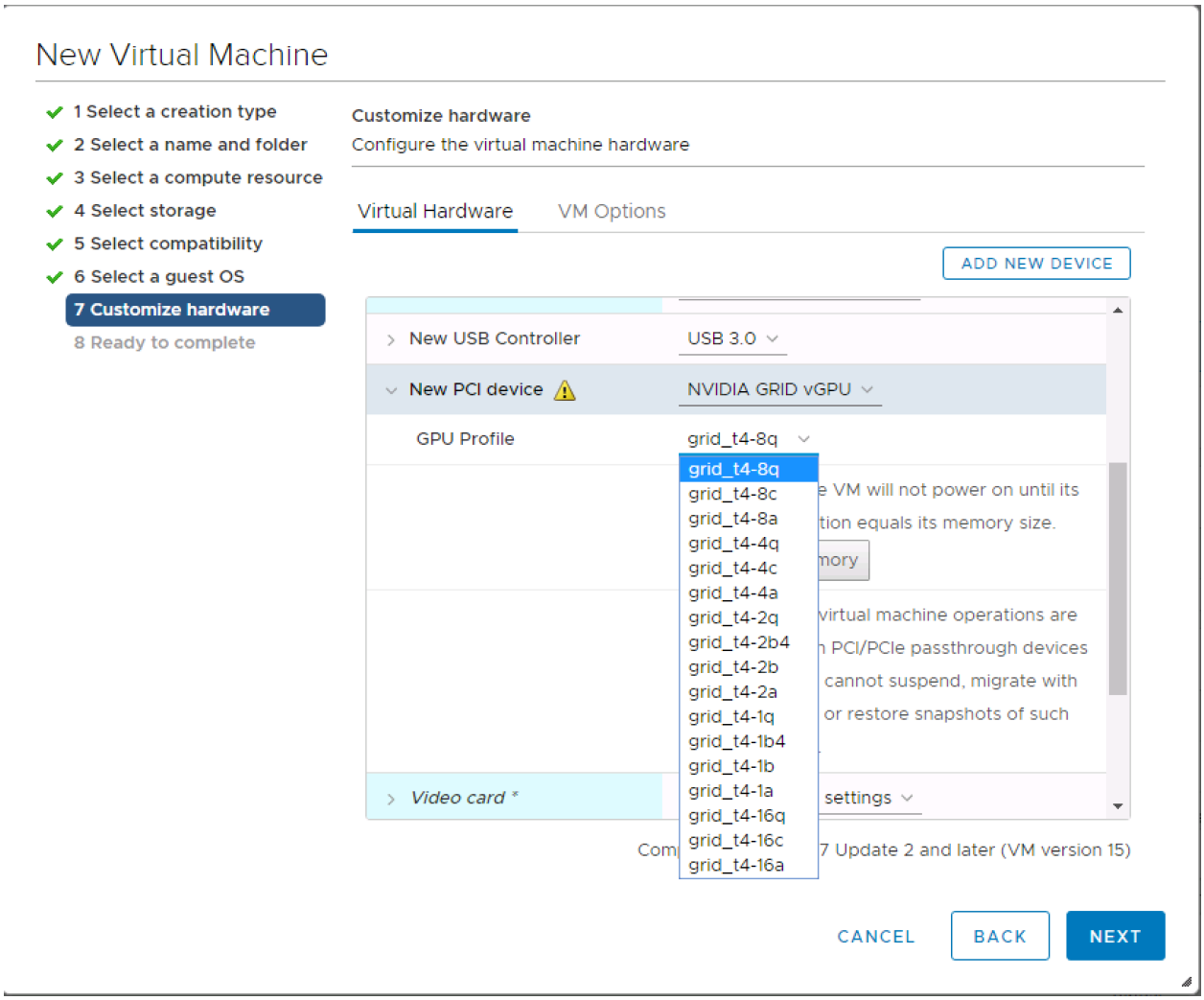
- Heavy application use, including browsing; using email; and creating complex documents, presentations, and spreadsheets.
- Applications supported include Windows 10, Office productivity apps, streaming video, and multimedia using the latest web standards like WebGL.
- Up to four 2K (2,560x1,600) resolution monitors.

On the other hand, users with any of the following characteristics should be assigned a GRID vPC 2GB profile for advanced knowledge workers:

- Users with specific requirements such as multiple high-resolution monitors to handle larger files and higher-resolution media.
- Applications supported include Windows 10, Office productivity apps, video and multimedia, and industry-specific apps like Bloomberg, Thomson Reuters Eikon, and DICOM viewers.
- Up to two 4096x2160 resolution monitors.

Creative and technical professionals running applications like Dassault Systèmes CATIA, Autodesk Revit, Siemens NX, Petrel, and so on might need high frame buffers, depending on the workload and size of the models being manipulated.

For compute workloads such as AI, machine learning, and data science, NVIDIA recommends a minimum of 4GB of frame buffer, with larger frame buffers for larger models.



NVIDIA supports the same vGPU profiles that are available on a GPU, as shown in Table 3.

Table 3) NVIDIA T4 vGPU profiles on single GPU.

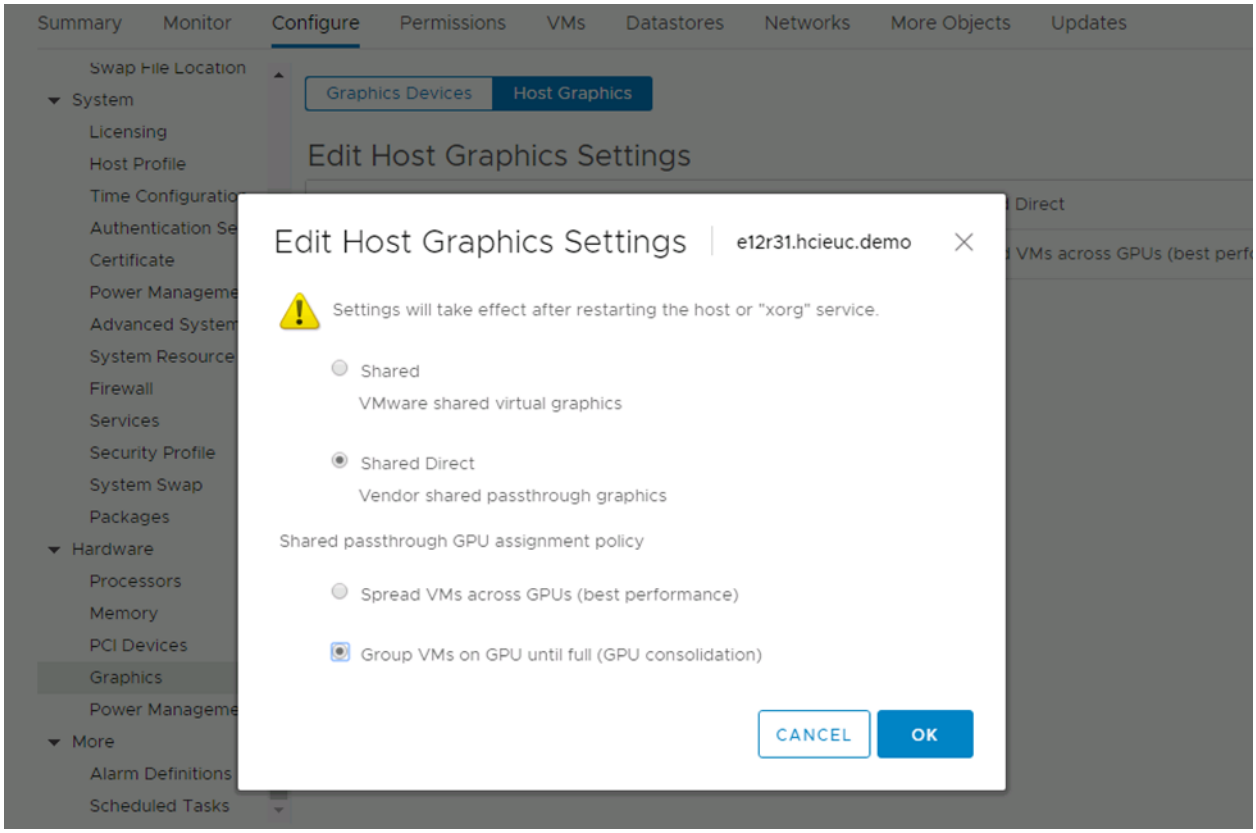
Tesla T4															
T4-16Q															
T4-8Q								T4-8Q							
T4-4Q				T4-4Q				T4-4Q				T4-4Q			
T4-2Q		T4-2Q		T4-2Q		T4-2Q		T4-2Q		T4-2Q		T4-2Q		T4-2Q	
T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q	T4-1Q

NVIDIA vGPU software does not support mixing profiles for a GPU accelerator with only one GPU. For example, if a VM with the 4Q profile is running on a GPU, it can only host another VM that has the same 4Q profile on the same GPU.

Table 4) Incorrect vGPU profile mix in single GPU.

T4-4Q	T4-2Q	T4-2Q

Because vGPU profiles on an H610C differ from the profiles on an H615C, a VM can't migrate from one host to another. Therefore, NetApp recommends having the same models in a cluster. If there is a strong need to mix the nodes in cluster, consider changing the GPU assignment policy to Group VMs on GPU. You need to have enough resources for a takeover if there is a node failure on either of the GPU nodes.



After the NVIDIA driver is installed on a VM, the VM console presents a blank screen for the vGPU profile. You must install VNC/Horizon Direct Connect prior to the NVIDIA driver to have console access.

VMware vSphere 6.7 update 1 and later provides support for vMotion for VMs with NVIDIA vGPU profiles.

Table 5) vGPU – H610C compared to H615C.

Frame Buffer	1xH610C (2xM10) 2 RU	1xH615C (3xT4) 1RU	2xH615C (6xT4) 2RU
1GB	64	48	96
2GB	32	24	48
4GB	16	12	24
8GB	8	6	12

Frame Buffer	1xH610C (2xM10) 2 RU	1xH615C (3xT4) 1RU	2xH615C (6xT4) 2RU
16GB	NA	3	6

The H615C can host 50% more density for the same rack unit size and provide double the performance for most workloads.

6 Citrix Virtual Apps and Desktops

This section covers the components of Citrix Virtual Apps and Desktops.

New versions of Citrix Virtual Apps and Desktops are delivered every quarter. Each new version is supported for six months from its release date, and users are expected to upgrade the software every three to six months. If a longer support term is required, consider using the Long-Term Service Release (LTSR) version. You can also use Citrix Cloud if you don't want to deal with version upgrades.

6.1 Citrix Workspace App (Receiver Client)

The Citrix Workspace App, which replaces the Receiver client, is client software that runs on virtually any device and operating platform, including Windows, Mac, Linux, iOS, and Android. The app must be downloaded onto user endpoints to access graphics applications that are hosted in the data center. The Citrix Workspace App provides client-side functionality to secure, optimize, and transport information to and from the endpoint or host over Citrix HDX. Citrix HDX is a set of technologies built into the networking protocol to provide a high-definition user experience regardless of device, network, or location. Framehawk is a display remoting technology for mobile workers on broadband wireless connections (Wi-FI and 4G/LTE cellular connections).

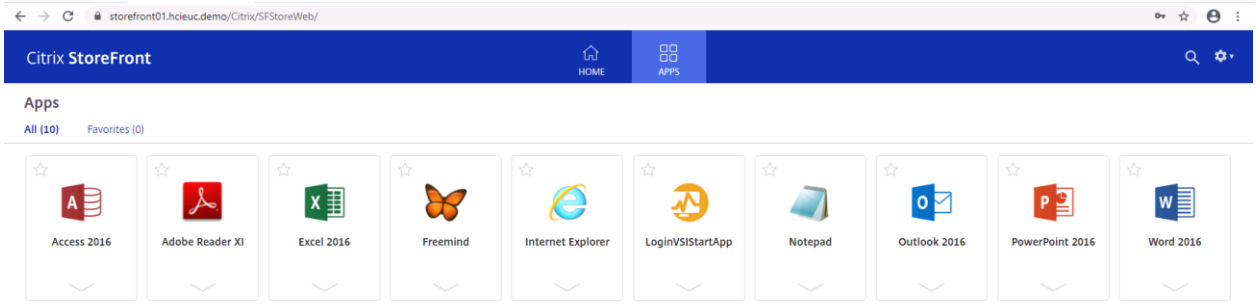
6.2 Citrix Gateway

Citrix Gateway, formerly known as NetScaler Unified Gateway, is an add-on component that provides secure access to virtual apps and desktops from remote locations. It can be deployed on the premises, or it can be consumed as a service from Citrix. It offers single sign-on access to web, virtual, cloud, and SaaS applications across multiple devices. It also supports multifactor authentication.

6.3 Citrix StoreFront

Citrix StoreFront is an enterprise app store that improves security and simplifies deployments, delivering a modern, unmatched near-native user experience across Citrix Receiver on any platform. StoreFront makes it easy to manage multisite and multi-version Citrix Virtual Apps and Desktops environments. Organizations can further customize with their own branding and deployments to meet specific business needs. Citrix StoreFront also enables mobility features for users to work with any device from any location. Citrix StoreFront can aggregate resources delivered to user from various sites avoiding duplicates.

Figure 8) Citrix StoreFront.



6.4 Citrix Delivery Controller

The Citrix Delivery Controller is responsible for managing user access and brokering and optimizing connections. Controllers also provide Machine Creation Services (MCS) that use the hypervisor cloning feature to create desktop and server images. A site must have at least one delivery controller. After you install the initial controller, you can add more controllers when you create a site, or possibly later. Having additional controllers at a site provides redundancy and scalability to improve overall responsiveness.

The Controller's broker service tracks which users are logged on and where, what session resources users have, and whether users need to reconnect to existing applications.

The Monitor Service collects historical data and places it in the Monitor database. This service uses TCP port 80 or 443. Data from the Controller services is stored in the site database.

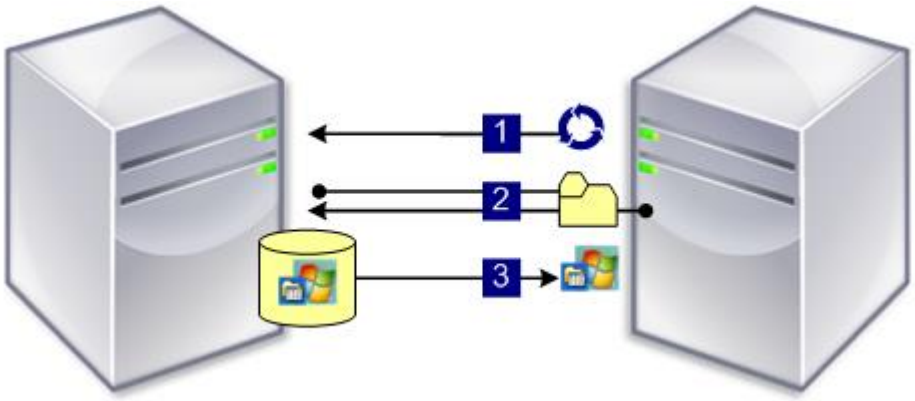
The Controller manages the state of desktops and starts and stops them based on demand and the administrative configuration. In some editions, the Controller allows you to install profile management to manage user personalization settings in virtualized or physical Windows environments.

6.5 Citrix Provisioning Services

Citrix Provisioning Services (PVS) is an optional component that can be used to deliver desktop or server images to physical machines as well as to VMs. It streams a shared disk image (vDisk) rather than copying images to individual machines. PVS enables organizations to reduce the number of disk images that they manage. All the write operations to the disk are redirected to a delta disk that is specific for each machine.

When a target device is turned on, it is set to boot from the network and communicate with a provisioning server. Unlike thin-client technology, processing takes place on the target device (step 1 in Figure 9).

Figure 9) Citrix PVS overview.



The target device downloads the boot file from a provisioning server (step 2), and then the target device boots. Based on the device boot configuration settings, the appropriate vDisk is located and then mounted on the provisioning server (step 3). The software on that vDisk is streamed to the target device as needed. To the target device, it appears like a regular hard drive on the system.

Instead of immediately pulling all vDisk contents down to the target device as in traditional imaging deployment solutions, the data is brought across the network in real time as needed. This approach allows a target device to get a completely new operating system and set of software in the time it takes to reboot, without requiring a visit to a workstation. This approach dramatically decreases the amount of network bandwidth required by traditional disk imaging tools, making it possible to support a larger number of target devices on your network without affecting overall network performance.

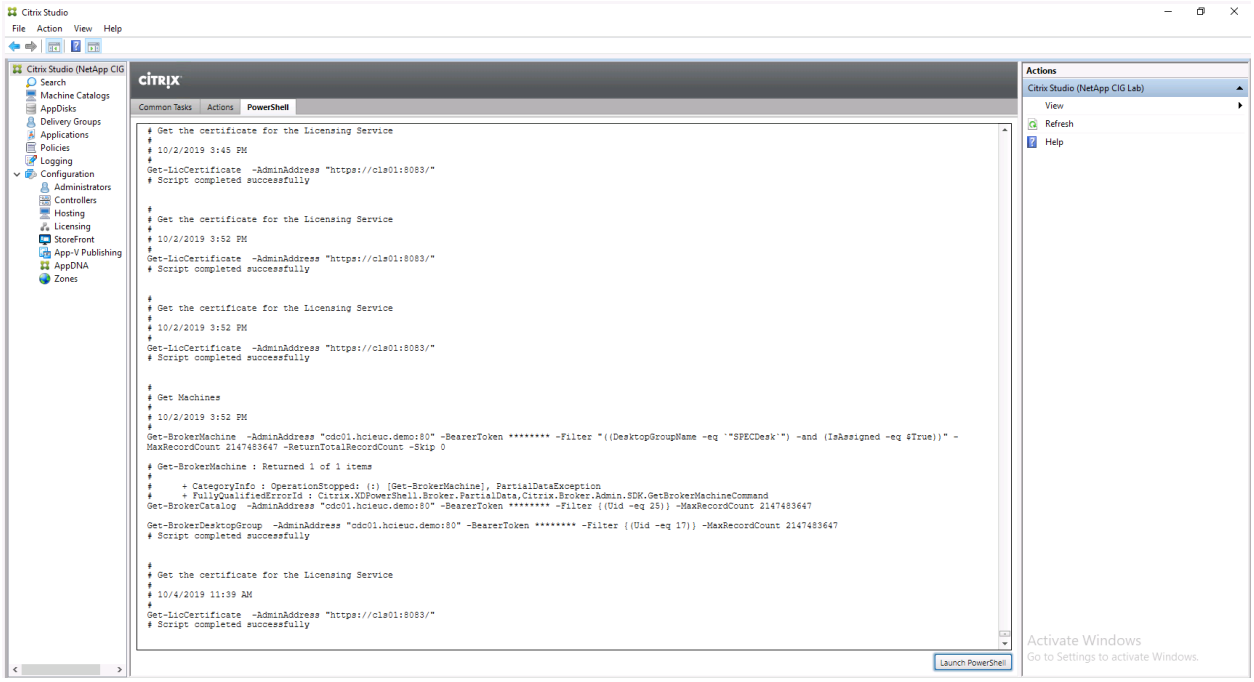
Traditionally, Citrix PVS used PXE services, which required some network configuration updates in order to make it work. Now, Citrix Provisioning allows you to include a Boot Device Manager ISO or hard disk partition that it can create automatically as part of virtual disk, which prevents the need for additional PXE network configurations.

6.6 Citrix Studio

Citrix Studio is the management console that enables you to configure and manage your Virtual Apps and Desktops deployment. This console eliminates the need for separate management consoles to manage delivery of applications and desktops. Citrix Studio provides wizards to guide you through environment setup, creating your workloads to host applications and desktops, and assigning applications and desktops to users. You can also use Studio to allocate and track Citrix licenses for your site.

Studio gets the information it displays from the broker service in the Controller by communicating over TCP port 80.

Figure 10) Citrix Studio.



Citrix Studio allows you to define policies that can be centrally managed for your virtual apps and desktops. For any actions you take in Citrix Studio, the corresponding PowerShell command line is displayed to facilitate automation.

6.7 Citrix License Server

Citrix License Server manages your Citrix product licenses. It communicates with the Controller to manage licensing for each user's session and with Citrix Studio to allocate license files. You must create at least one license server to store and manage your license files.

6.8 Citrix Director

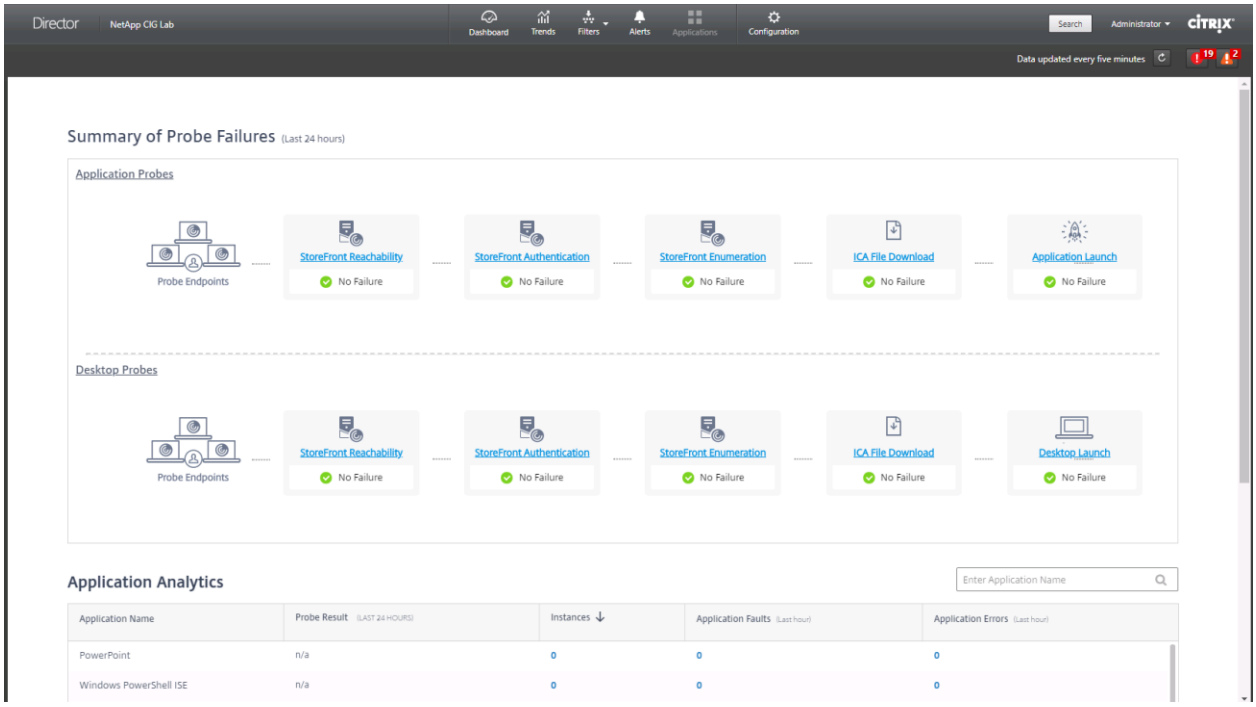
Citrix Director is a web-based tool that enables IT support and helps desk teams to monitor an environment, troubleshoot issues before they become system-critical, and perform support tasks for end users. You can use one Director deployment to connect to and monitor multiple Citrix Virtual Apps and Desktop Sites.

Citrix Director displays the following information:

- Real-time session data from the broker service in the Controller. This includes data that the broker service gets from the broker agent in the VDA.
- Historical site data from the monitor service in the Controller.
- Data about HDX traffic (also known as ICA traffic) captured by HDX Insight from the Citrix Gateway, if your deployment includes Citrix Gateway and your Citrix Virtual Apps and Desktops edition includes HDX Insight.

You can also view and interact with a user's sessions through Citrix Director by using Windows Remote Assistance.

Figure 11) Citrix Director.



6.9 Virtual Delivery Agent

VDA is installed on each physical machine or VM in your site that you make available to users. The VDA enables a machine to register with the Controller, which makes the machine and resources it is hosting available to users. VDAs establish and manage the connection between the machine and the user

device, verify that a Citrix license is available for the user or session, and apply whatever policies have been configured for the session.

The VDA communicates session information to the broker service in the Controller through the broker agent in the VDA. The broker agent hosts multiple plug-ins and collects real-time data.

The acronym VDA is often used to refer to the agent as well as the machine on which it is installed.

VDAs are available for Windows Server and Windows desktop operating systems. VDAs for Windows Server operating systems allow multiple users to connect to the server at one time. VDAs for Windows desktop operating systems allow only one user to connect to the desktop at a time. Linux VDAs are also available.

MCS I/O cache can be enabled for nonpersistent desktops to provide read/write caching and reduce the I/O requirement for the back-end storage system.

VDA will sense the presence of NVIDIA driver and install appropriate modules. So, always install VDA after NVIDIA drivers.

6.10 Database

At least one Microsoft SQL Server database is required for every Virtual Apps and Desktops site to store configuration and session information. This database stores the data collected and managed by the services that make up the Controller. Install the database in your data center and provide it with a persistent connection to the Controller. The site also uses a configuration logging database and a monitoring database. By default, those databases are installed in the same location as the site database, but you can change this.

6.11 File Shares

NetApp HCI provides file services by using NetApp ONTAP Select, which provides the following features useful for Citrix Virtual Apps and Desktops implementations:

- User home directories
- Storage virtual machines (SVMs)
- NetApp ONTAP FlexGroup technology
- NetApp FabricPool technology
- NetApp Data Availability Services
- Adaptive QoS
- Deduplication
- Data protection features:
  - NetApp RAID DP®
  - NetApp Snapshot™ technology
  - NetApp FlexClone® technology
  - NetApp SnapMirror® and SnapVault® technologies
- Self-restore from Snapshot copies

Citrix Virtual Apps and Desktops supports both Microsoft Windows and Linux environments. Therefore, ONTAP provides the SMB and NFS protocols with the secure multitenancy feature and the flexibility to use the same storage environment for multiple environments, such as test and production.

7 Technology Requirements

This section covers the technology requirements for the Citrix Virtual Apps and Desktops solution.

7.1 Hardware Requirements

Table 6 lists the hardware components that are required to implement the solution. The hardware components that are used in any particular implementation of the solution might vary based on customer requirements.

Table 6) Hardware requirements.

Hardware	Quantity	Description
NetApp HCI H410C	3	Infrastructure components
NetApp HCI H410C	7	Resource for around 700 non-GPU virtual desktops and apps
NetApp HCI H410S	4	Storage cluster to provide block storage
NetApp HCI H610C	3	Host around 120 virtual desktops for knowledge workers
NetApp HCI H615C	3	Host around 80 Quadro virtual workstations and 100 non-GPU virtual desktops
Mellanox SN2010	4	A 25Gb Ethernet switch
1Gb Ethernet switch	1	For Intelligent Platform Management Interface (IPMI) management

7.2 Software Requirements

Table 7 lists the software components that are required to implement the solution. The software components that are used in any particular implementation of the solution might vary based on customer requirements.

Table 7) Software requirements.

Software	Version
NetApp HCI NDE	1.7
NetApp ONTAP Select	9.7
Citrix Virtual Apps and Desktops (Premium Edition)	7 1909
• NVIDIA • vGPU Manager • Windows driver • Linux driver	• 9.1 • 430.46 • 431.79 • 430.46
Microsoft Windows Server	2016
VMware vSphere	6.7 Update 3
VMware vCenter	6.7 Update 3
SPECviewperf	13
Microsoft Windows	10 1903

8 Deployment Steps

8.1 Prerequisites

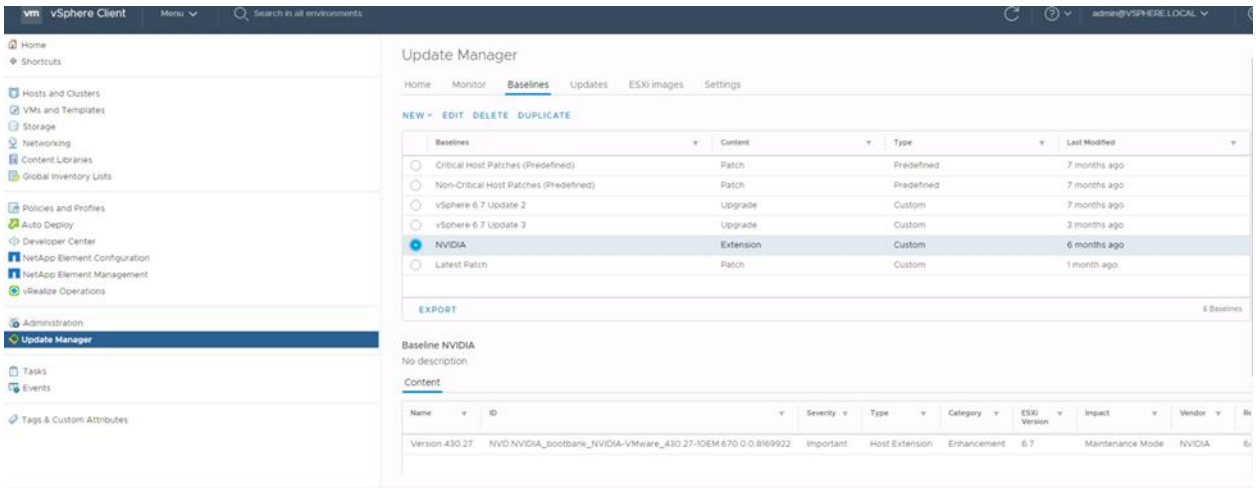
Deploy NetApp HCI by using NDE. For guidelines on cabling and switch configuration, accessing NDE, and post deployment tasks, see the NetApp HCI Deployment guide. Because you must add H410C, H610C, and H615C nodes on its own cluster, refer to the “Expanding Your NetApp HCI Installation” section of deployment guide to add additional compute resources that you use to add into the same cluster or on its own cluster.

Table 8 lists the models that were added in the lab environment.

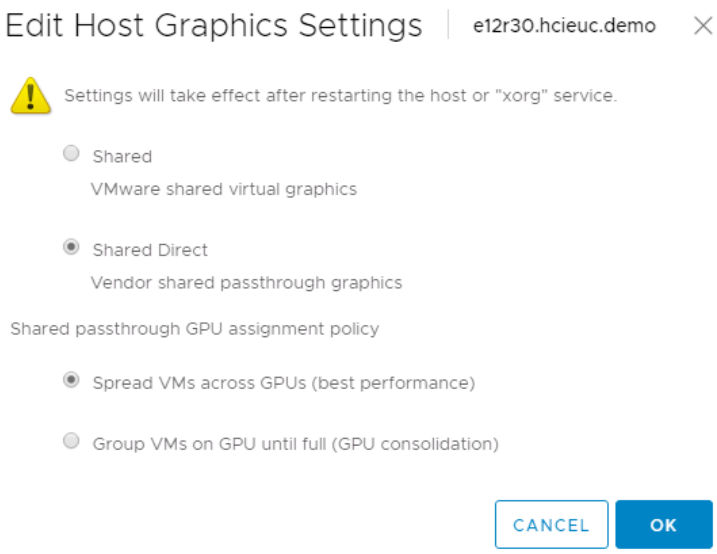
Table 8) NetApp HCI models.

Model	Number	Cluster
NetApp HCI H410S	4	Storage
NetApp HCI H410C	3	NetApp-HCI-Cluster-01 (Infrastructure)
NetApp HCI H410C	7	NetApp-HCI-Cluster-02 (VDI Workload)
NetApp HCI H610C (GPU)	3	NetApp-HCI-Cluster-03 (VDI Workload)
NetApp HCI H615C (GPU)	2	NetApp-HCI-Cluster-04 (VDI Workload)
NetApp HCI H610C (GPU)	3	NetApp-HCI-Cluster-05 (Launchers)
NetApp HCI H615C	1	H615C (VDI Workload)

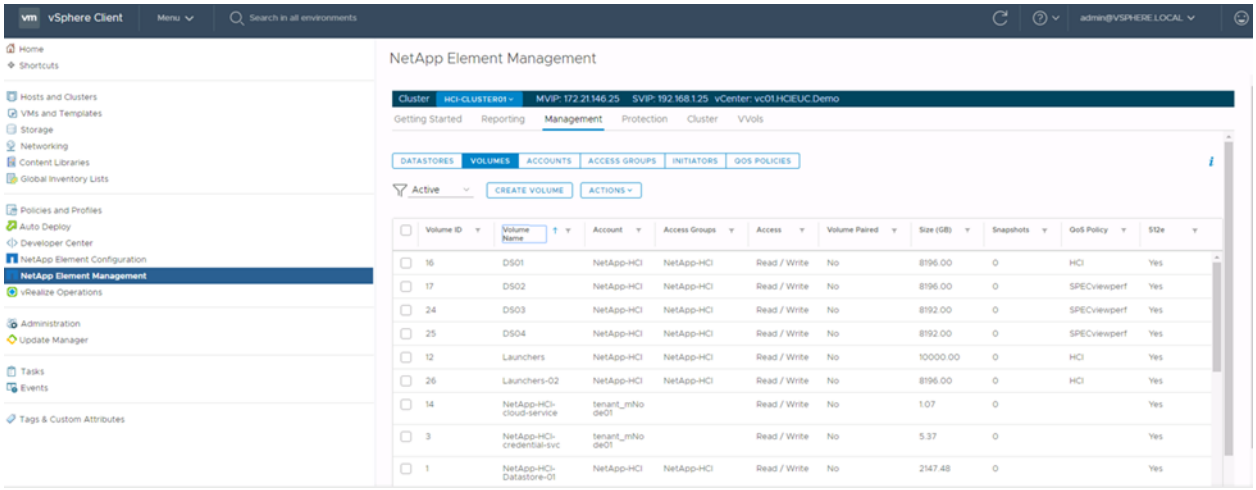
VMware Update Manager is used to upgrade the vSphere editions to vSphere 6.7 Update 3. NVIDIA GPU drivers on vSphere hosts on H610C and H615C clusters installed using the update manager as host extension.

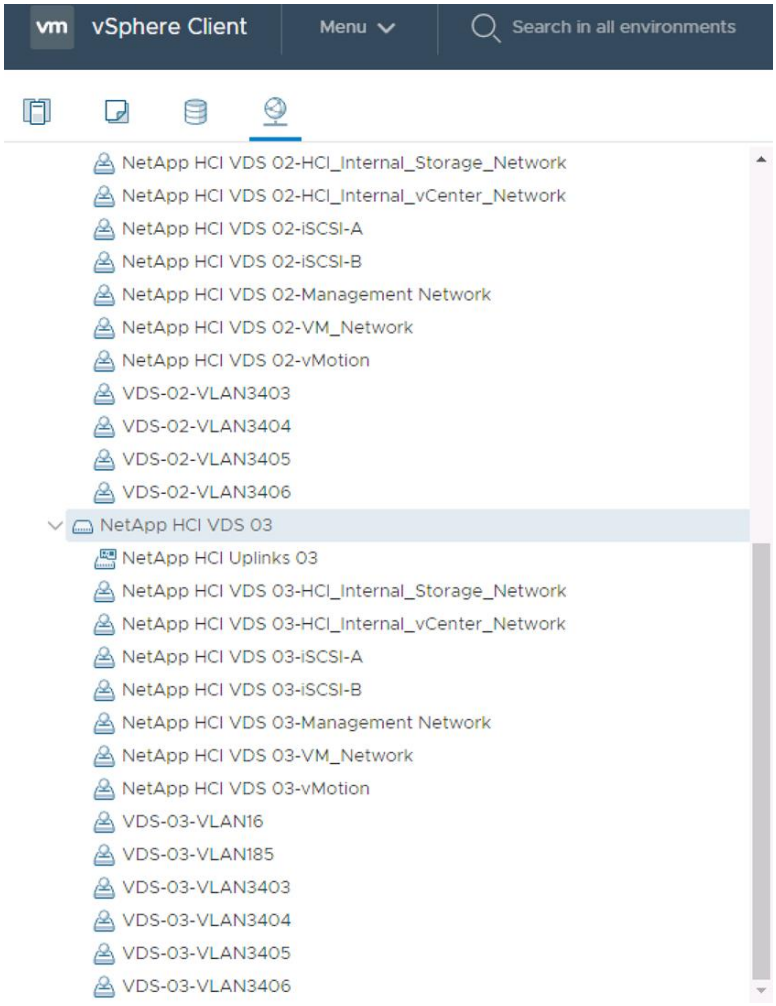


Modify host graphics settings to Shared direct to use vGPU profiles with virtual desktops. To enable vMotion of Virtual Desktops that hosts vGPU, the following vCenter Advanced settings (vgpu.hotmigrate.enabled) need to be set, as shown in the screenshot below.



Additional datastores are created to host infrastructure VMs and to host virtual desktops using NetApp HCI vCenter Plug-in.



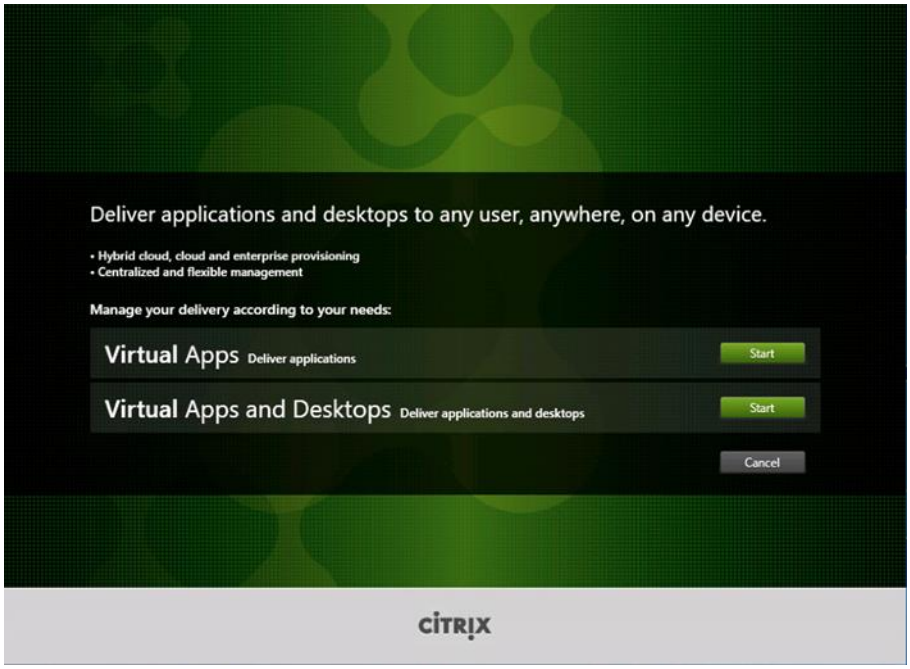


On VMware vCenter, additional port groups are created for the virtual desktop VLANs. We used VLAN 3403–3406 for virtual desktops.

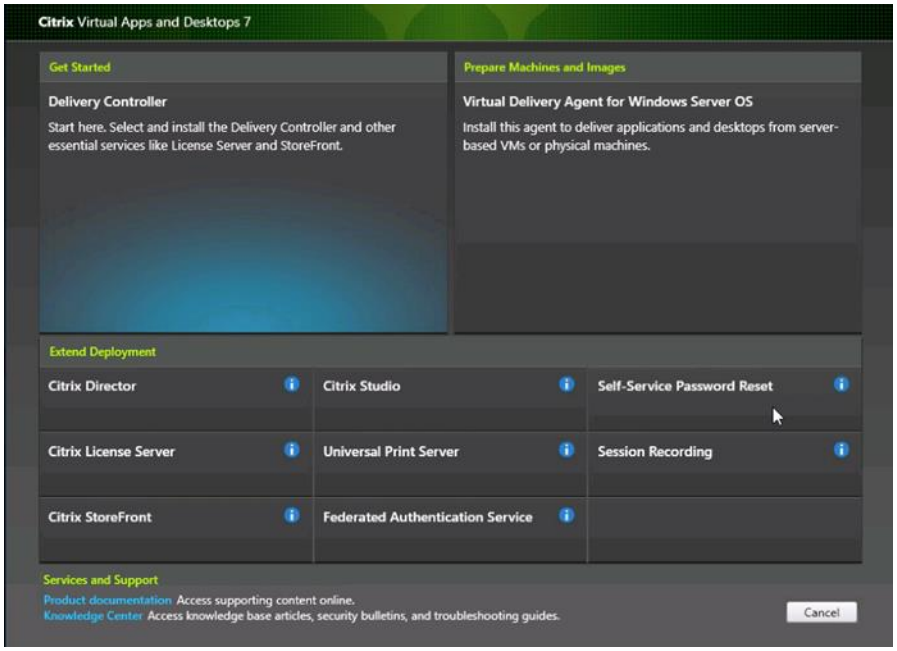
8.2 Citrix License Server

To install Citrix License Server, complete the following steps:

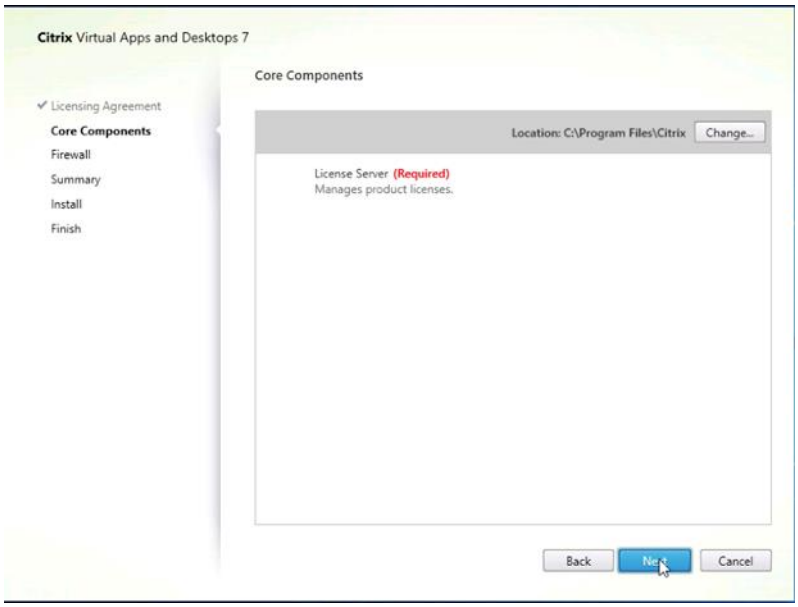
1. Deploy the supported Windows server by using ISO or from the VM template.
1. Mount the Citrix Virtual Apps and Desktops ISO file.
2. Run `AutoSelect.exe` at root of the drive.
3. Click Start on Virtual Apps and Desktops option.



4. Select Citrix License Server option.



5. If you agree, answer the license agreement, and click Next. You will receive the option to specify the installation folder.



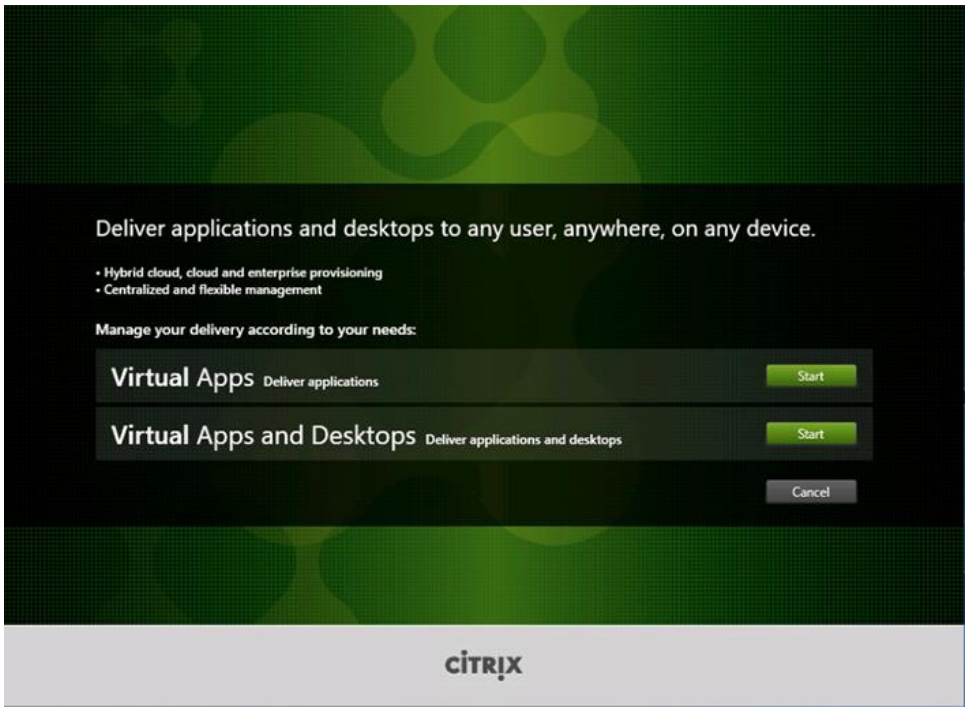
6. Answer and configure the remaining wizard pages and finish the installation.
7. On the Citrix Licensing portal, provide the license server details such as host name, FQDN, mac address, and IP address to retrieve the license file. For more information, see Activate, upgrade and manage Citrix licenses.

   You will use the downloaded license file after Citrix Studio is installed.

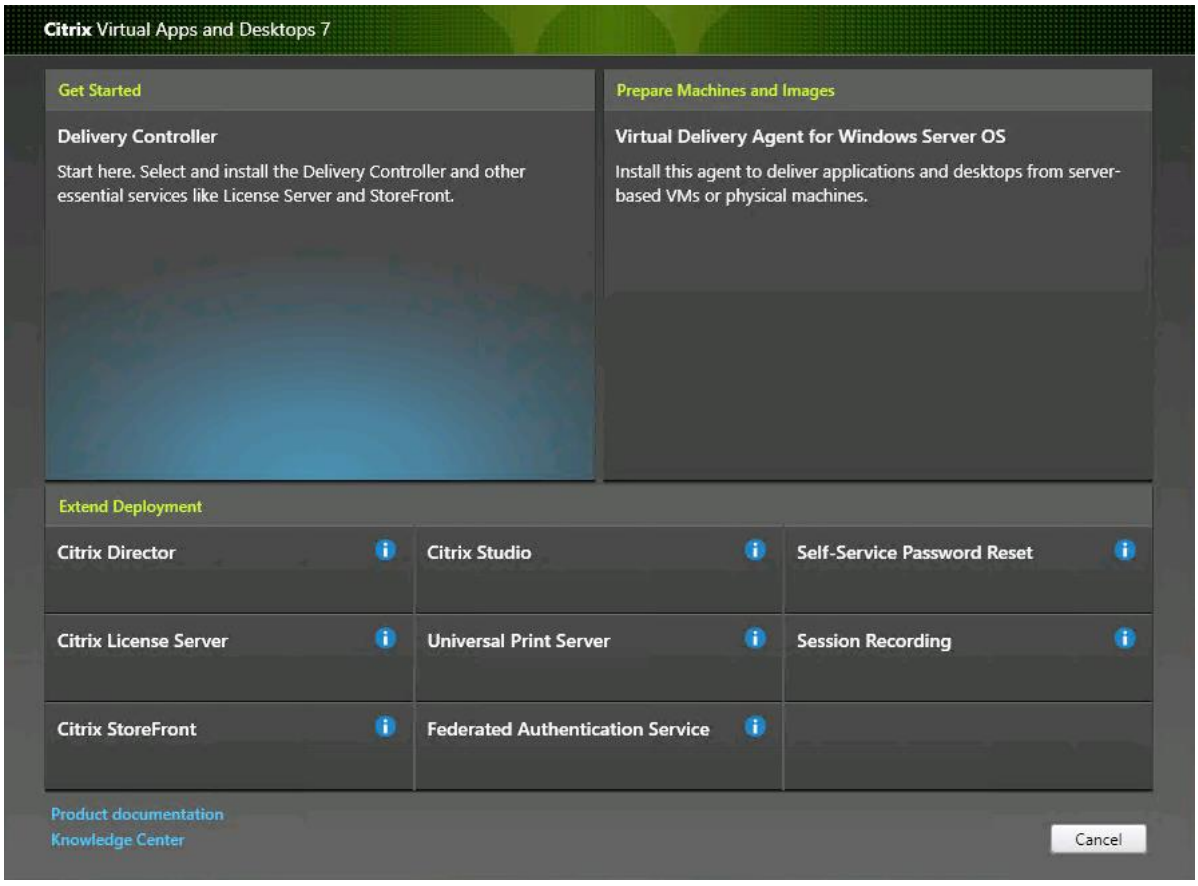
8.3 Citrix Delivery Controller

To install Citrix Delivery Controller, complete the following steps:

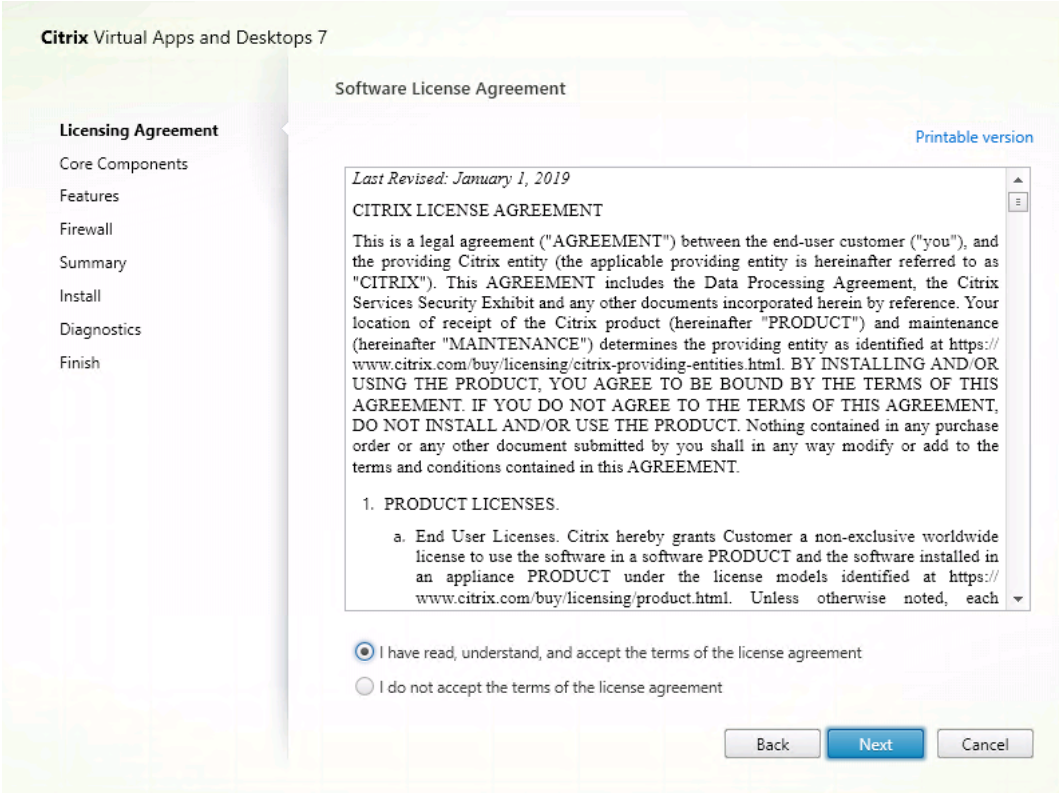
1. Deploy the supported Windows server by using ISO or from the VM template.
1. Mount the Citrix Virtual Apps and Desktops ISO file.
2. Run `AutoSelect.exe` at root of the drive.
3. Click Start on Virtual Apps and Desktops option.



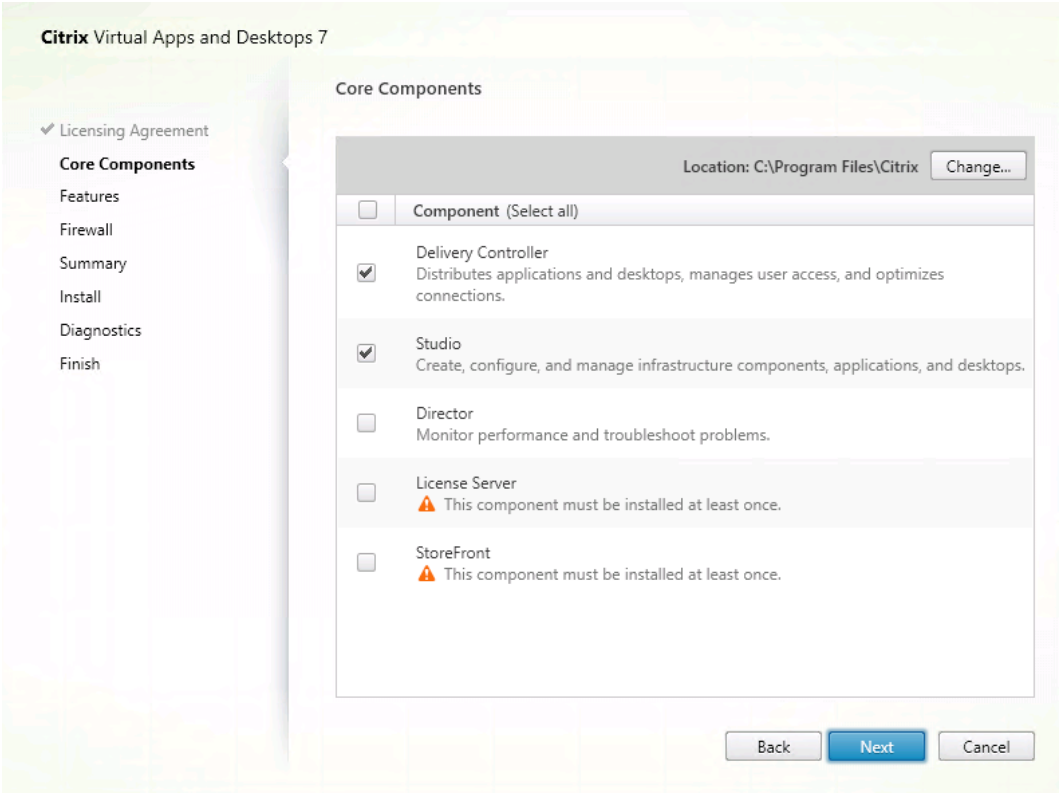
4. Click Delivery Controller to start the installation wizard.



5. Review the license agreement and select the appropriate option. Click Next.

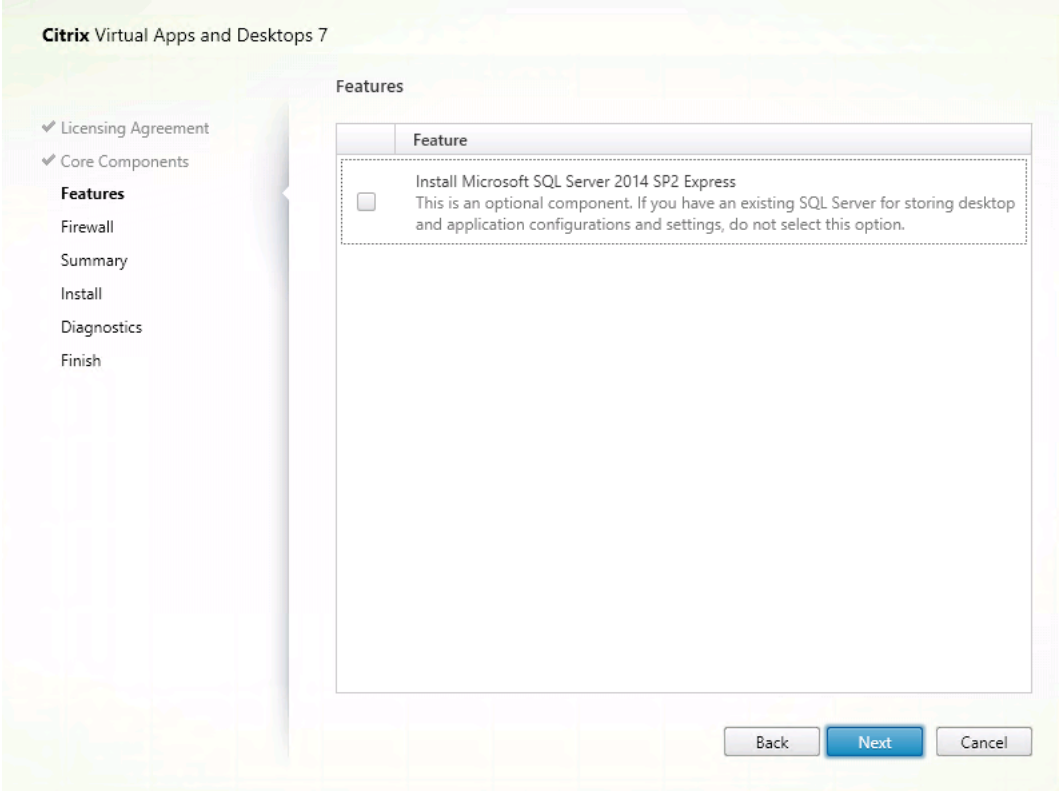


6. Review the installation location and select the components to be installed on this server. Click Next.

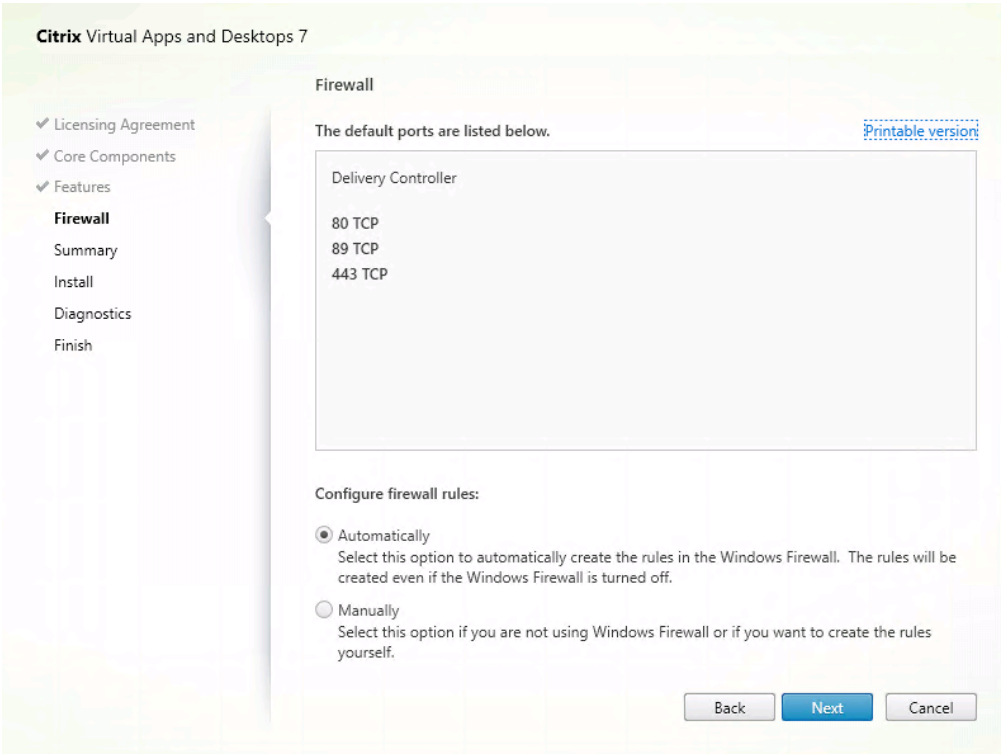


7. Select the feature to be installed. Click Next.

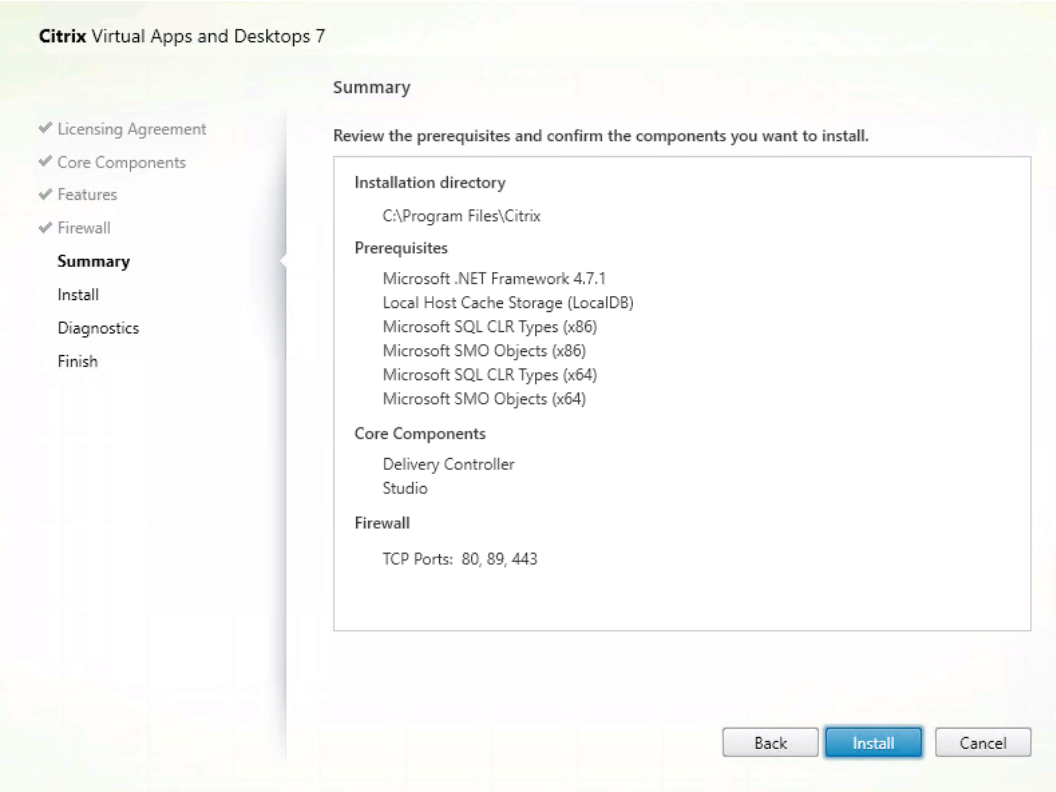
   Note: In this example, we cleared this the option because we have a dedicated SQL Server.



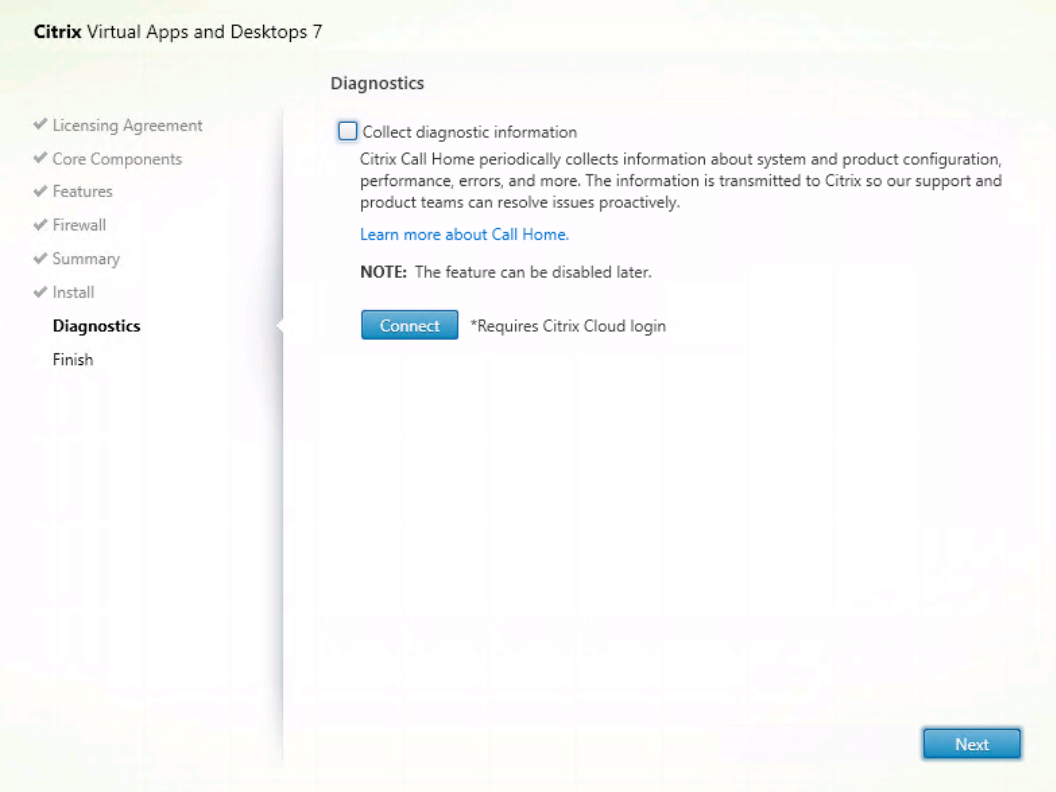
8. Review the firewall ports and select the appropriate option to enable them. Click Next.



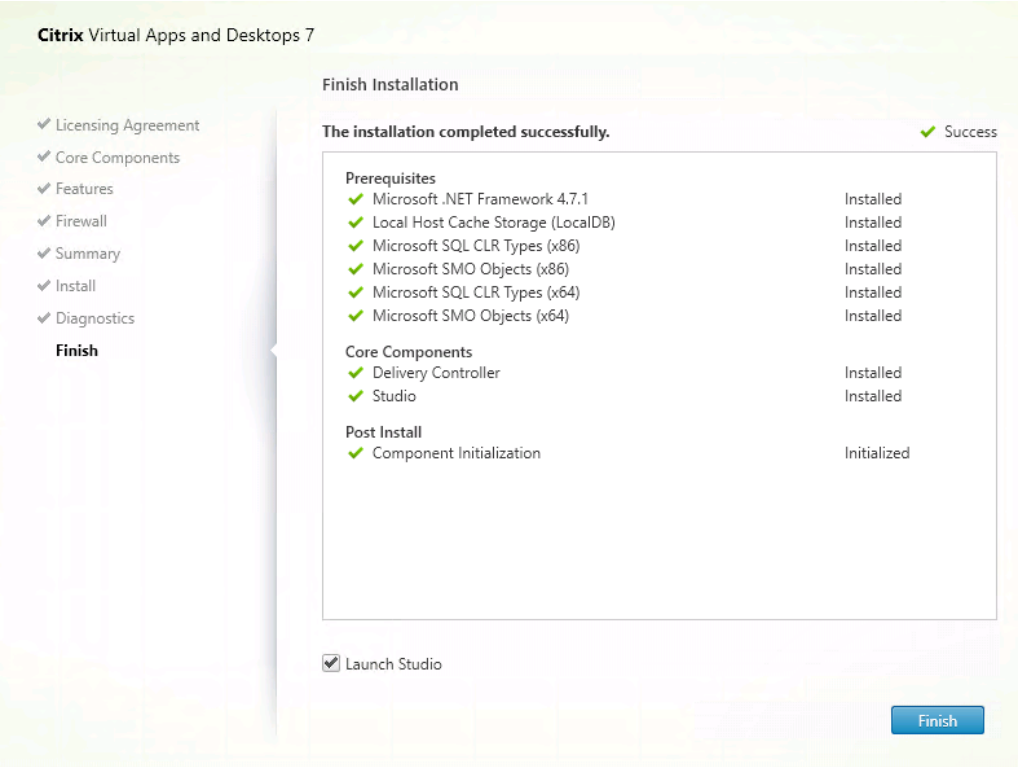
9. Review the summary and click Install.



10. Review and make corrections to collect diagnostic information. Click Next

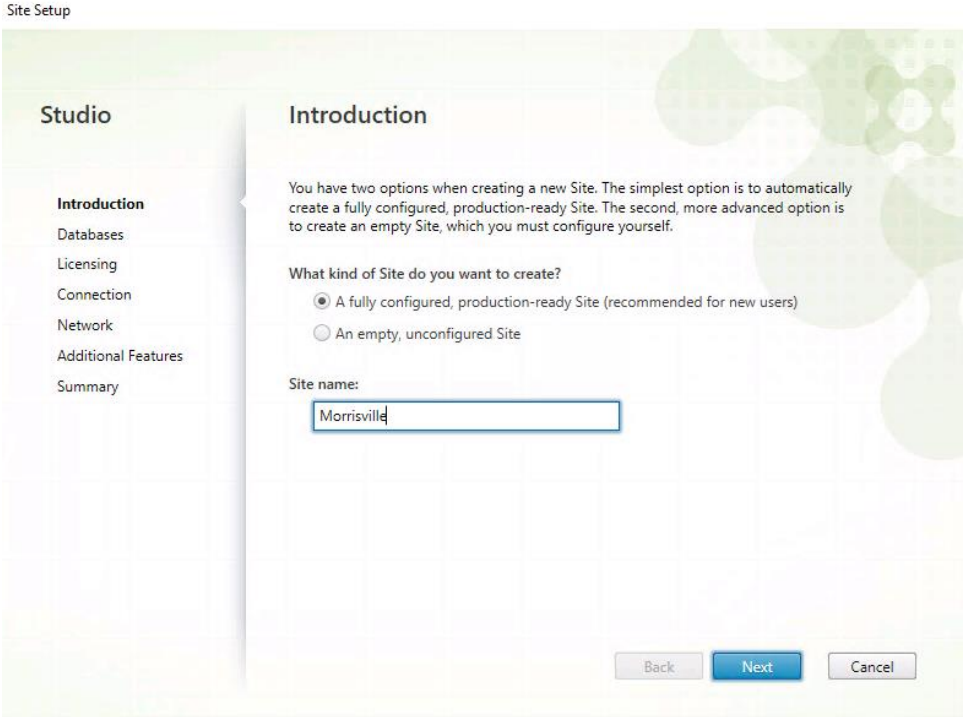


11. Review the option to launch Studio and click Finish.

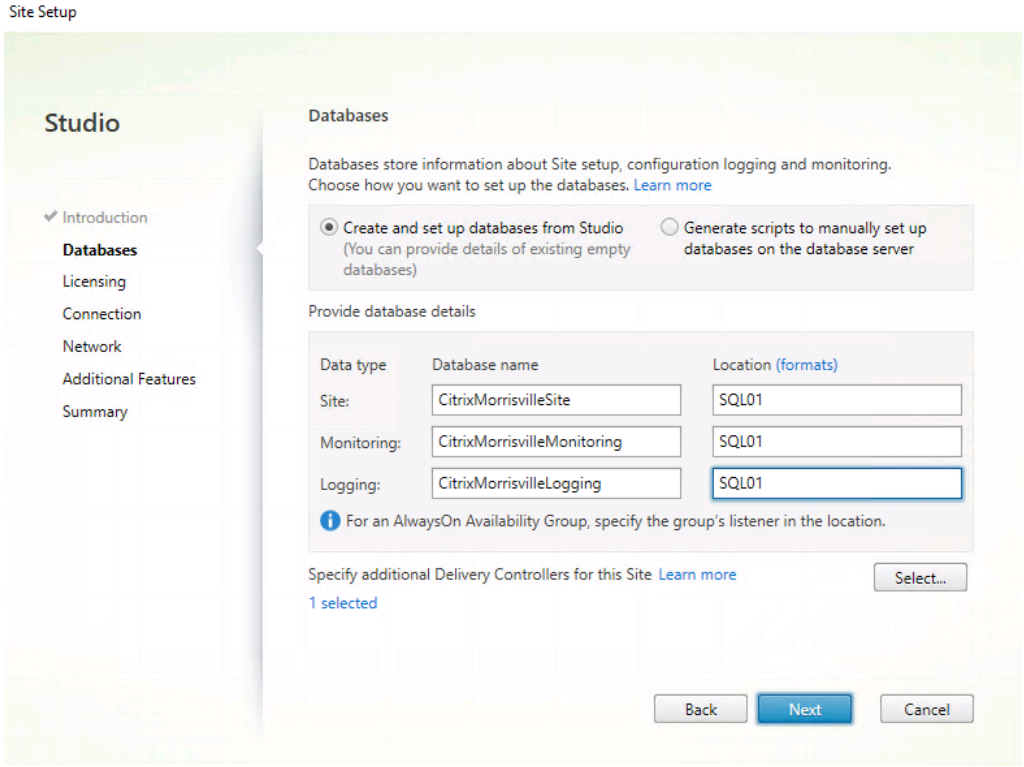


12. When the studio is launched, for the initial delivery controller, click Deliver Applications and Desktops to Your Users. For any subsequent delivery controllers, click Connect this Delivery Controller to an Existing Site.

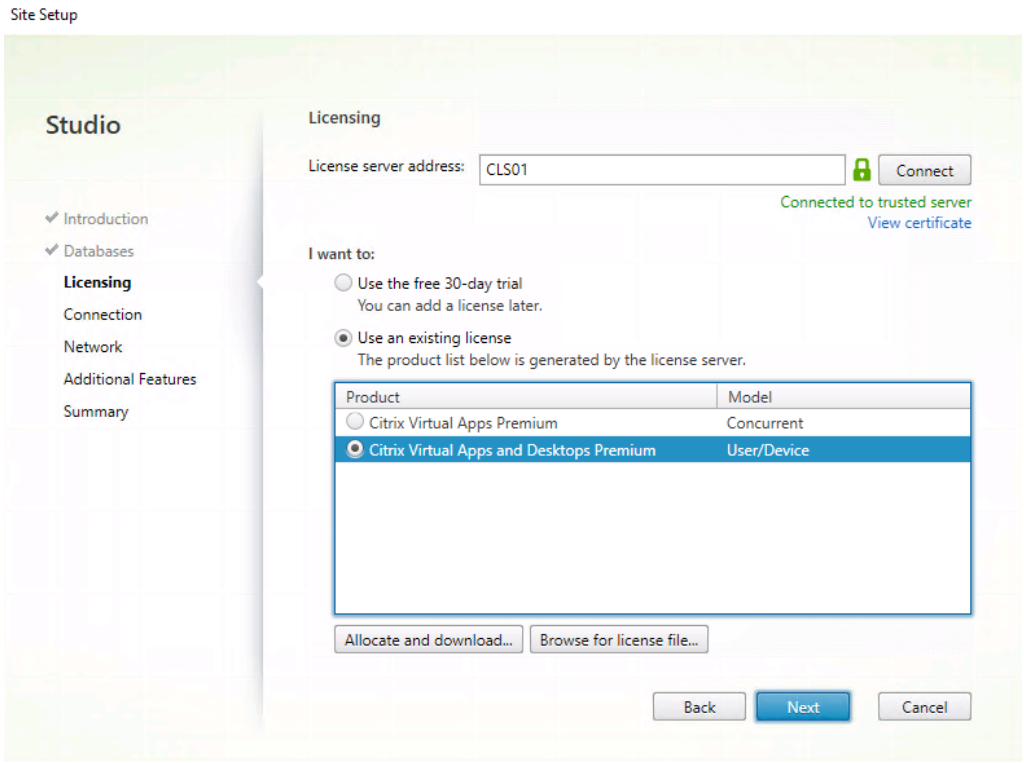
    Provide the site name and click Next.



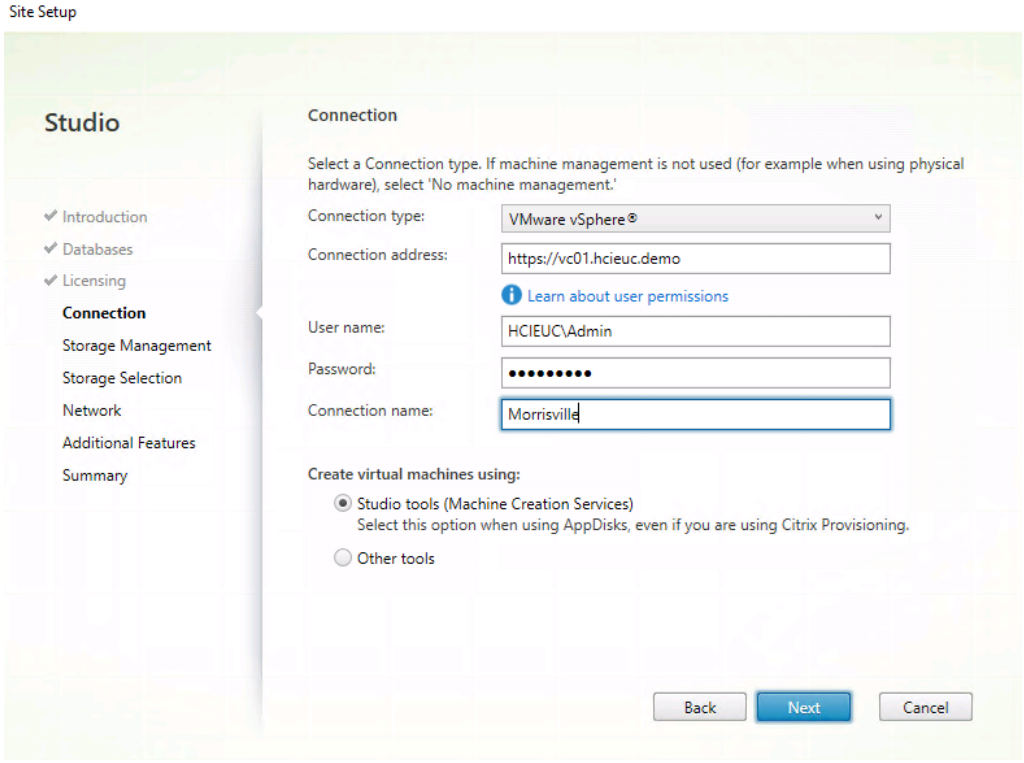
13. Review the database names and provide the SQL Server location information. Click Next



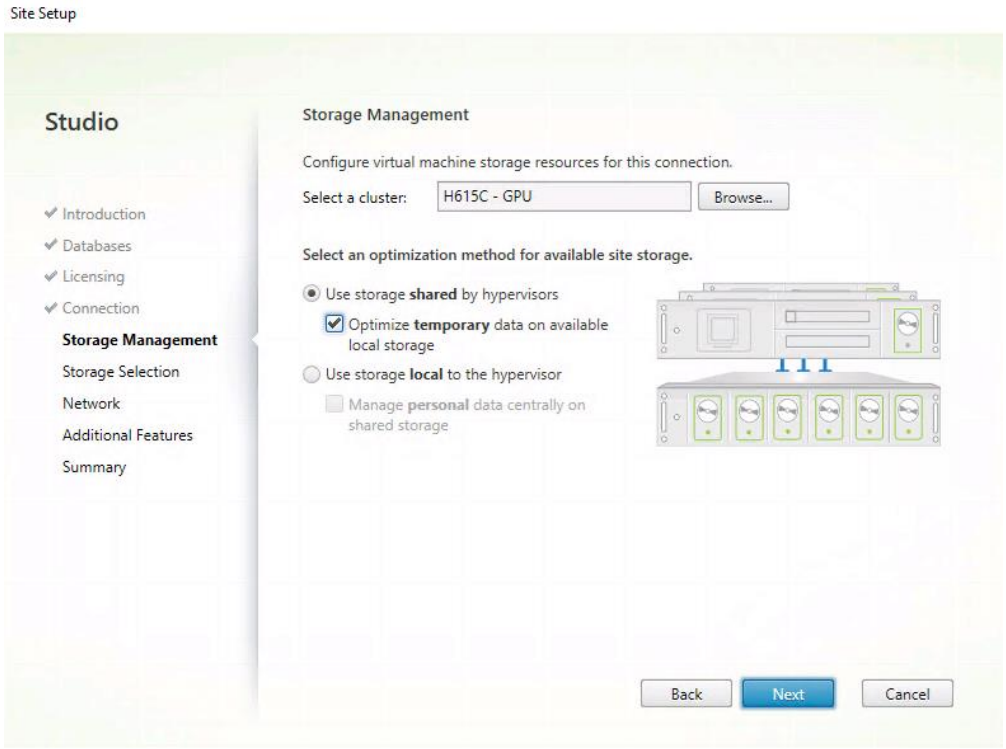
14. Provide the license server name and click Connect. Select the appropriate license information and click Next.



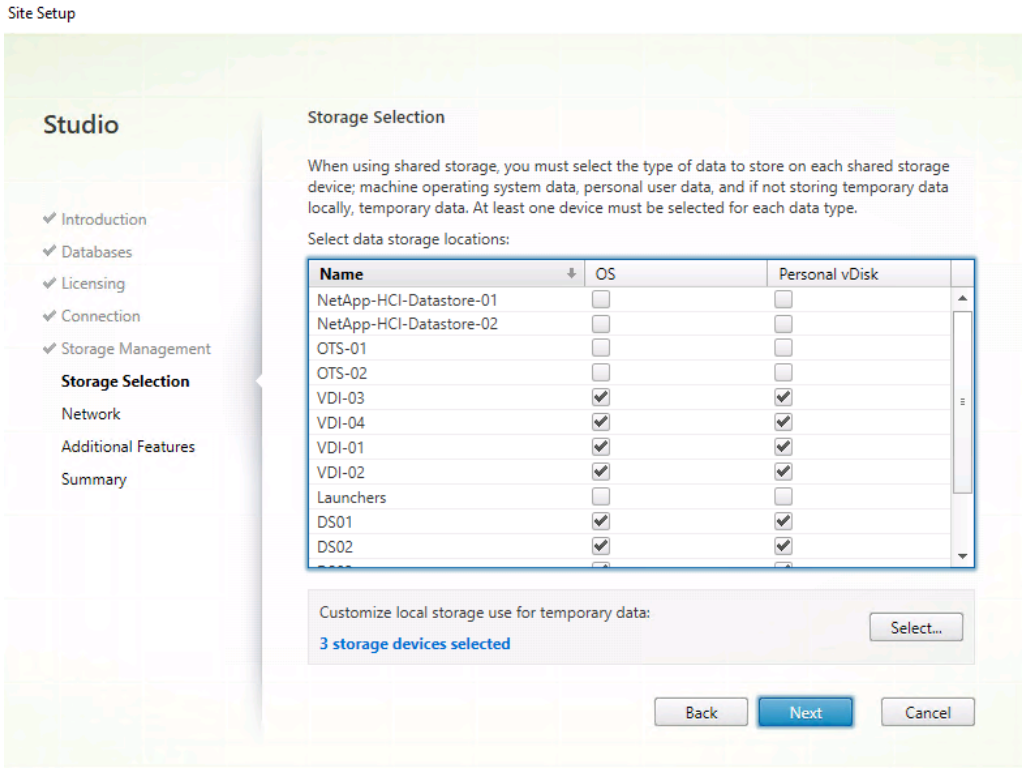
15. Select the appropriate connection type, provide the connection address, user credential, and connection name. Click Next.



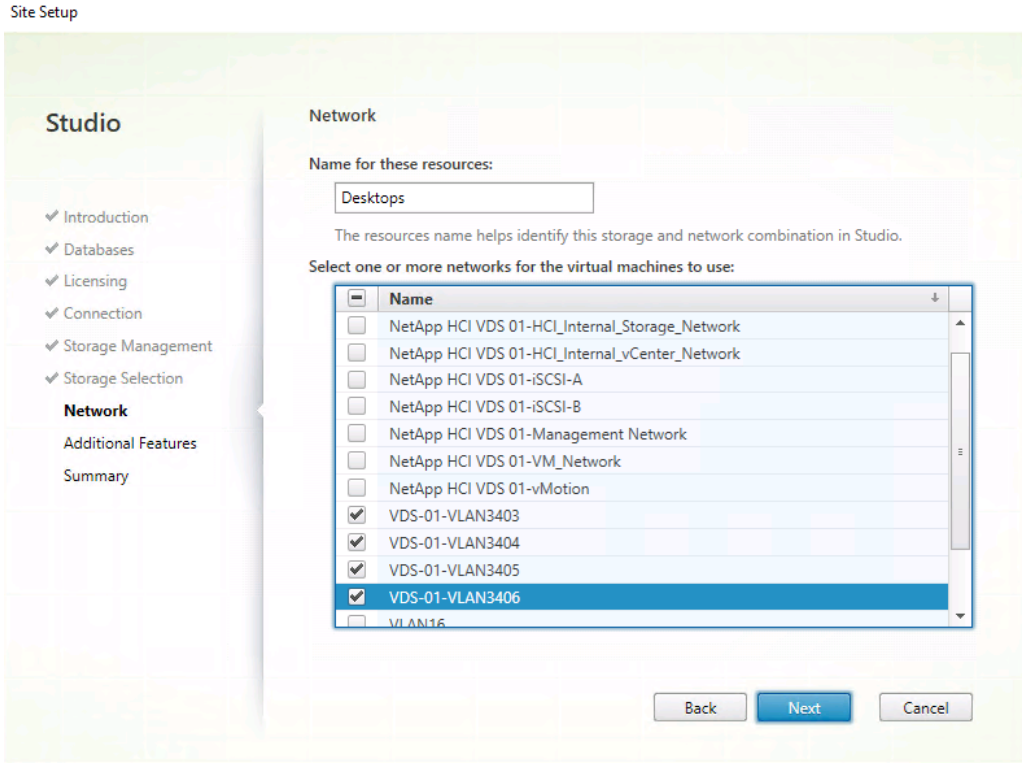
16. Select the vSphere cluster and review other options and make corrections to fit your requirement. Click Next.



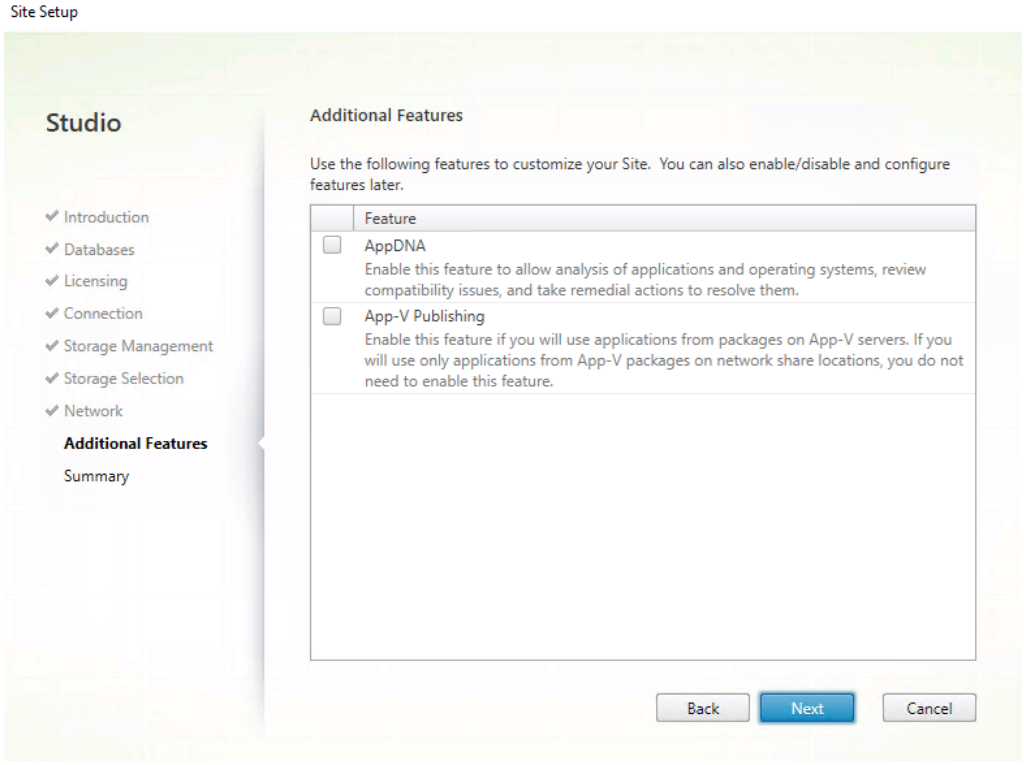
17. Select the datastores that will be used for VDI purposes. Click Next.



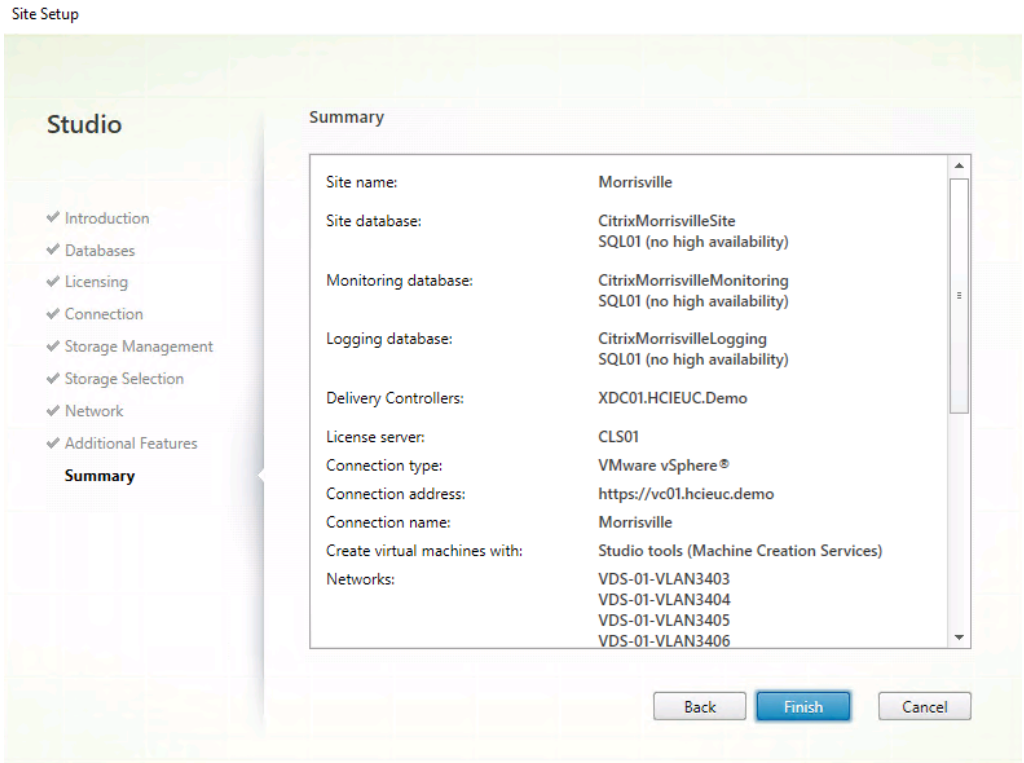
18. Provide a name for the network resource. Pick the vSphere Distributed Switch port group where the desktops will be connected. Click Next.



19. Review the options for additional features. Click Next.

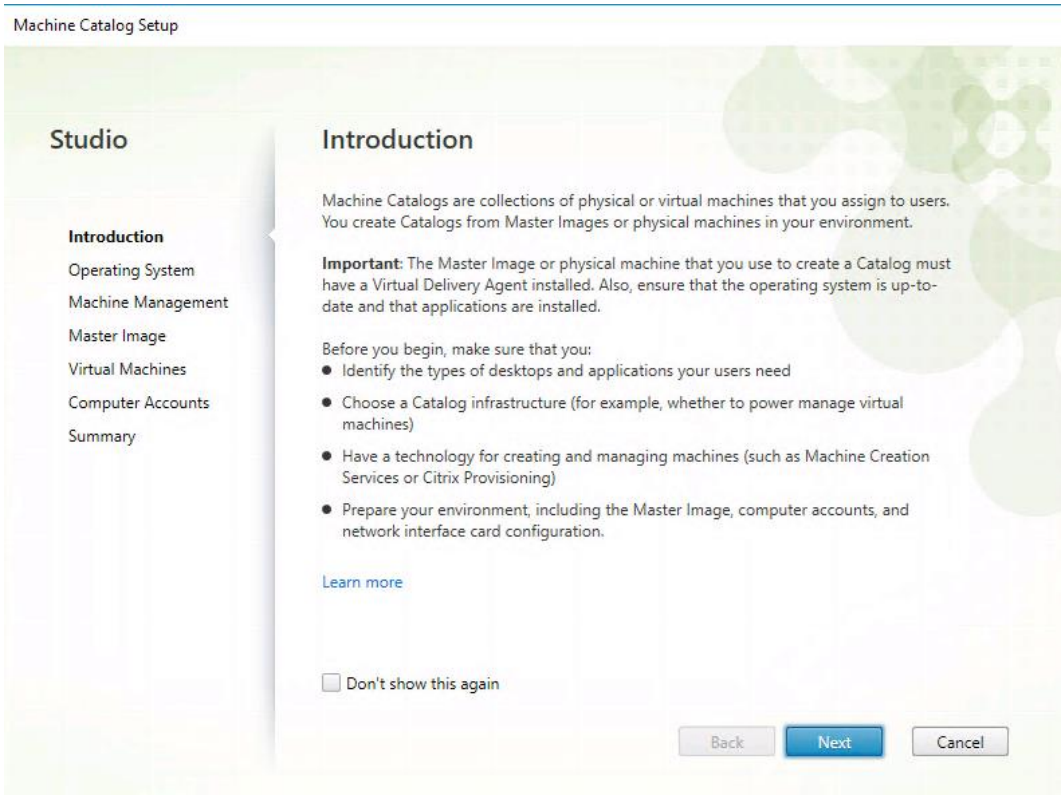


20. Review the summary and click Finish.

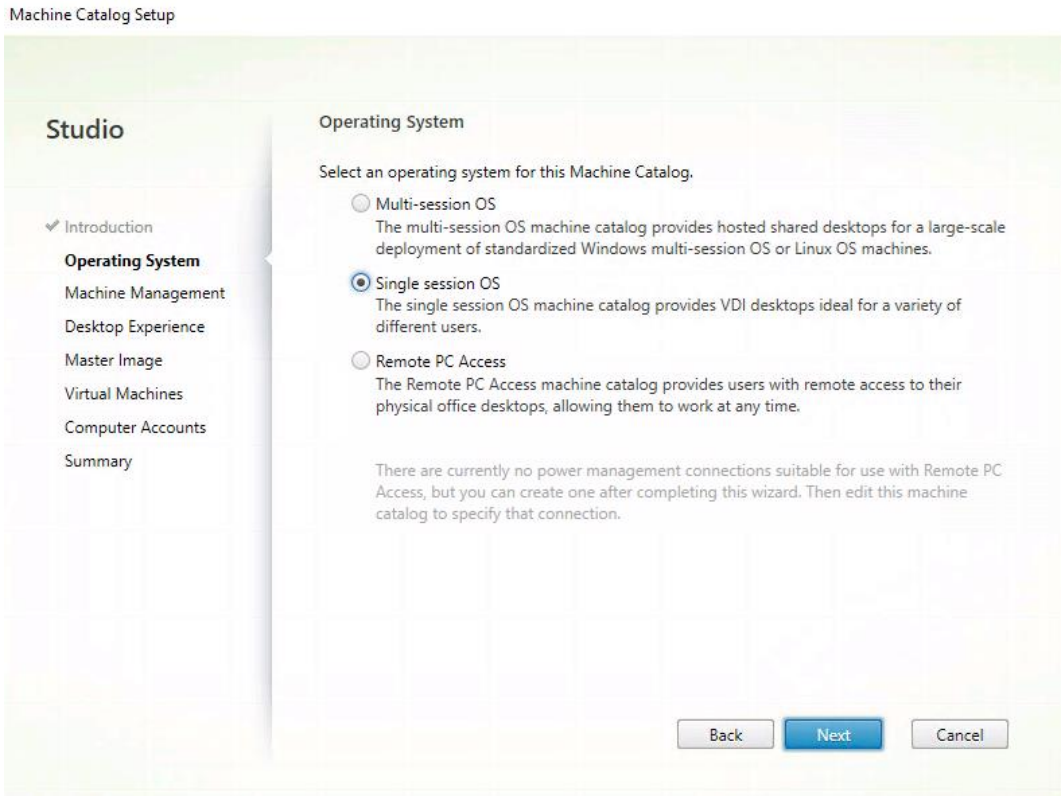


21. After it completes, Citrix Studio shows more options under the console tree.

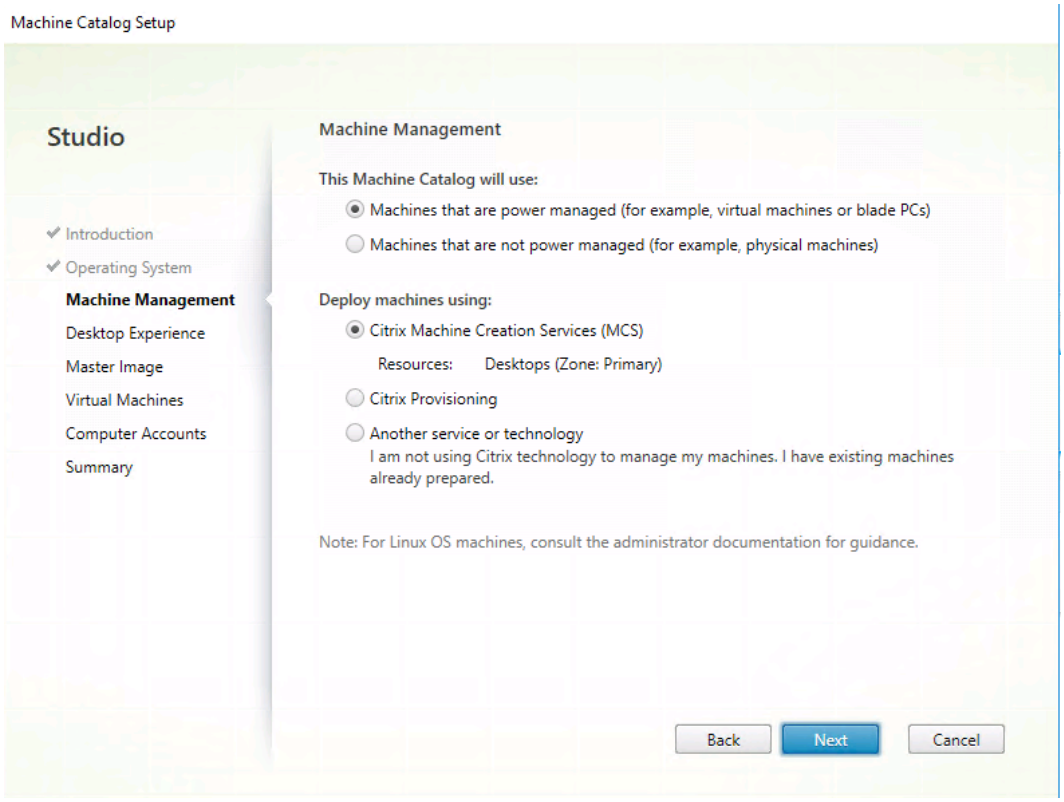
22. Click Option 2 to set up machines for desktops and applications or remote PC access. The Machine Catalog Setup wizard starts. Click Next.



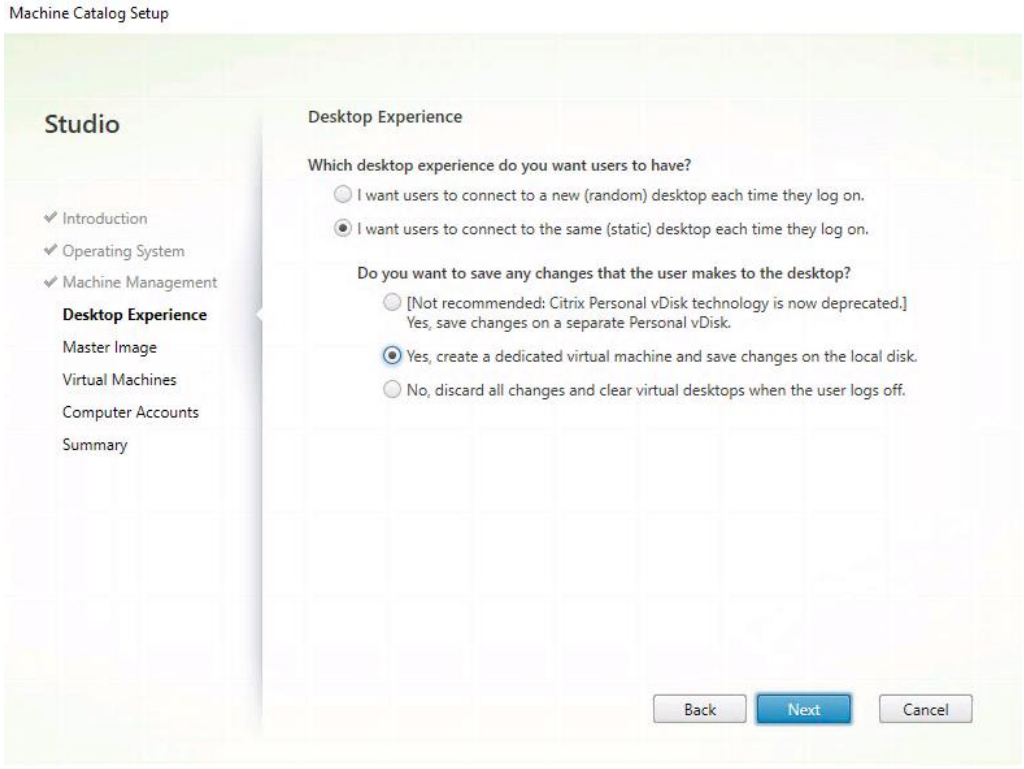
23. Select the option for the appropriate operating system. For a desktop operating system, select Single Session OS. For a server operating system (for virtual apps), select Multi-Session OS. Click Next.



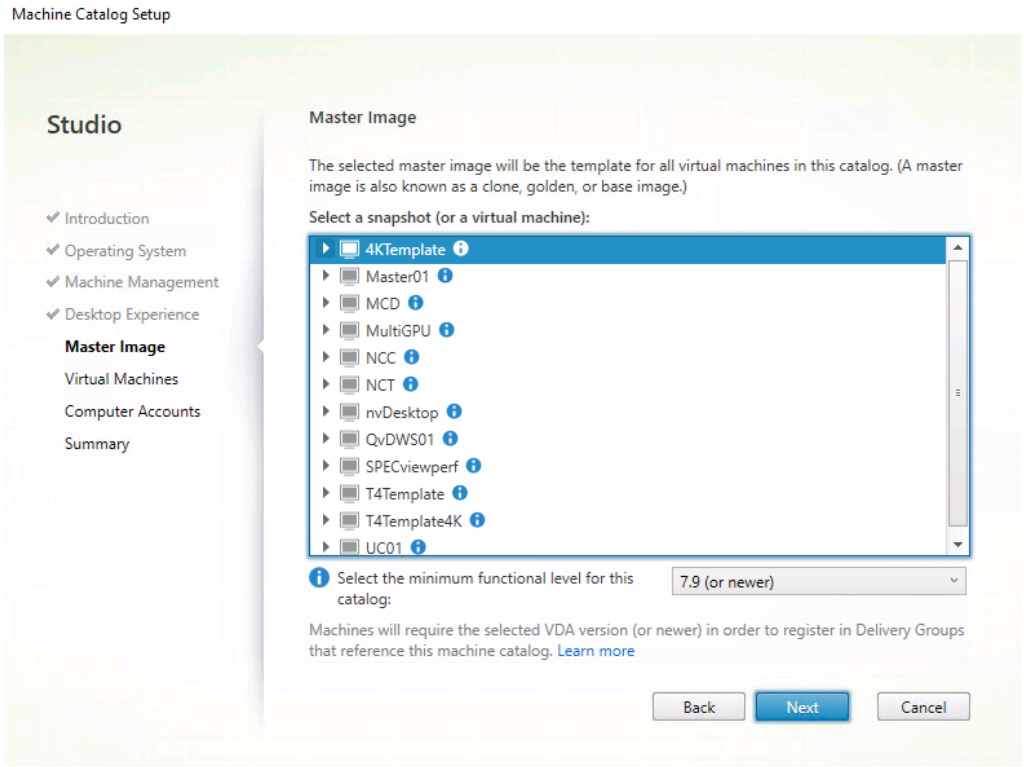
24. Select the appropriate option the machine management. Click Next.



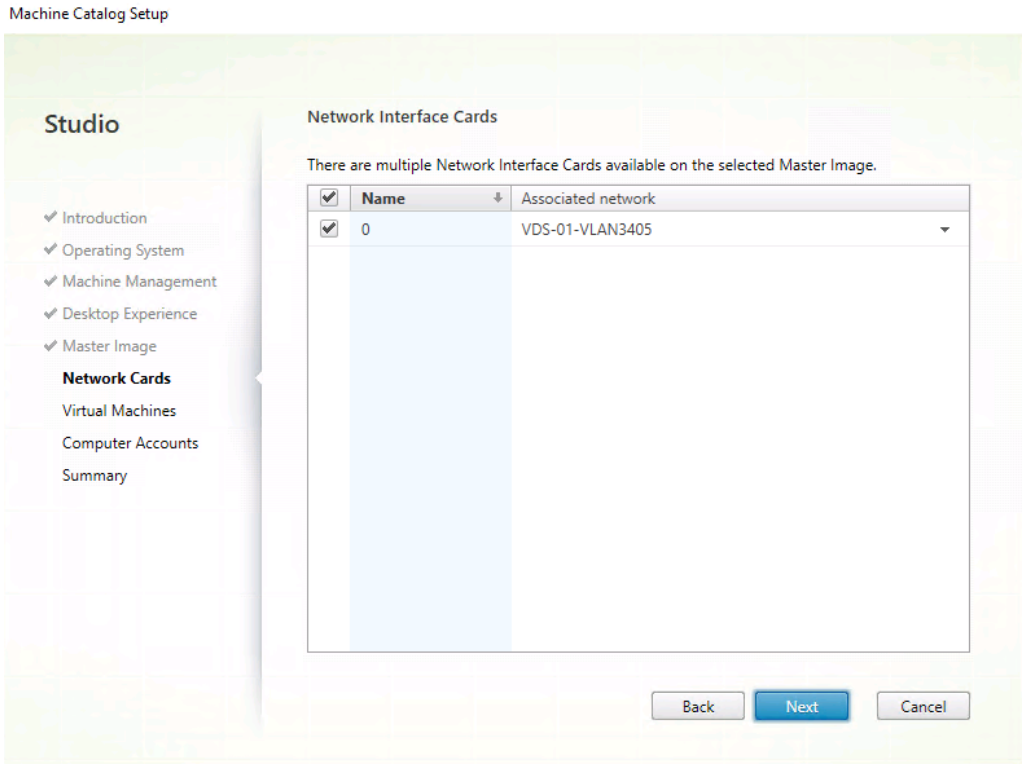
25. Select the option for persistent or nonpersistent desktops and then select how the user data disk needs to be handled. Click Next.



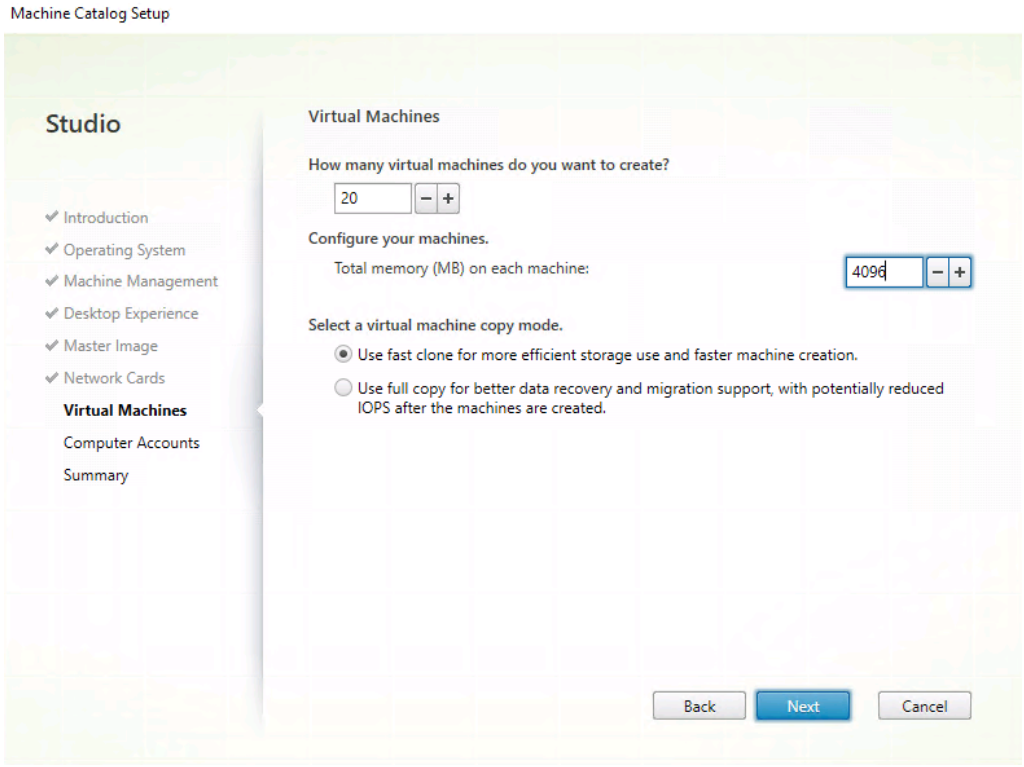
26. Select Master Image. Optionally, select the snapshot. Click Next.



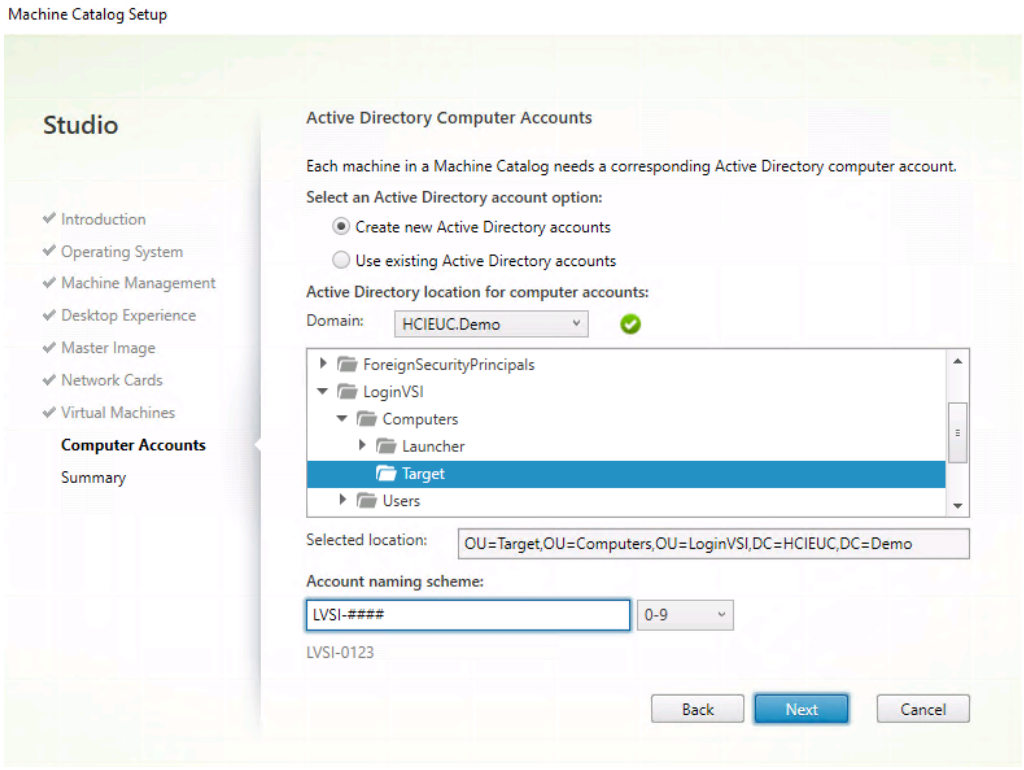
27. Associate the NIC to the vSphere Port group. Click Next.



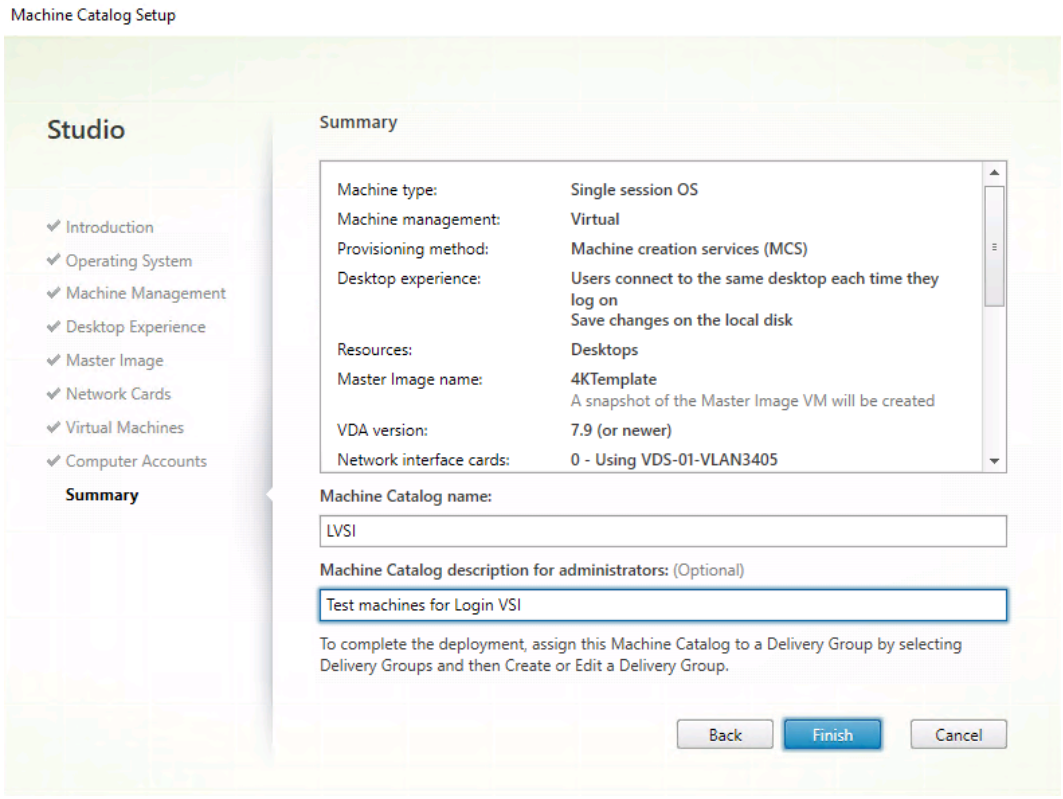
28. Enter the number of virtual desktops that must be created, the memory size, and whether to use the Fast Clone feature. Click Next.



29. Select the option to create a computer account or use an existing one. Select the OU. Provide a computer naming scheme. Based on your selection, the # symbol is replaced by a number or letter. Click Next.



30. Review the summary. Provide the machine catalog name and the optional description. Click Finish.



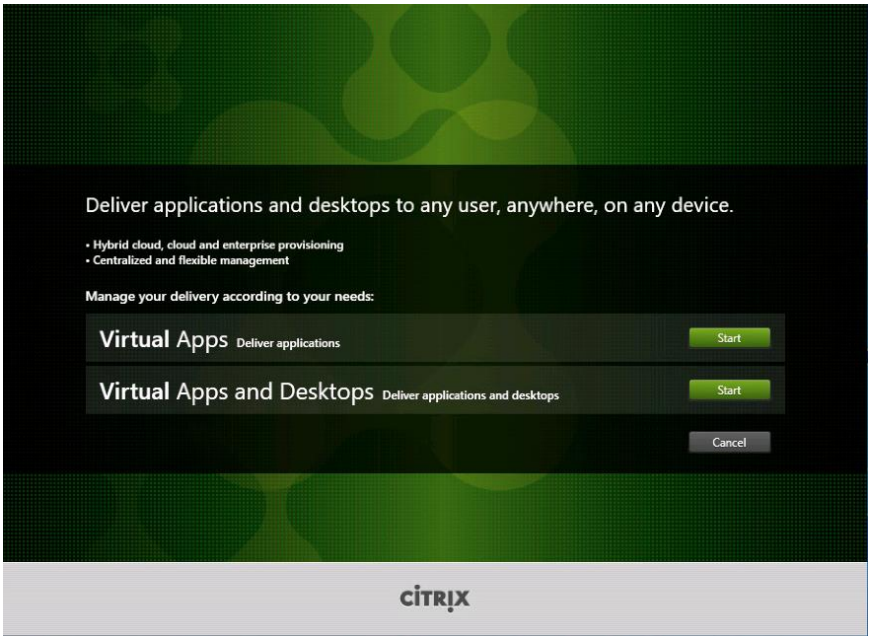
Note: Delivery Group must be created in order to assign the applications and desktops for the users.

8.4 Citrix StoreFront

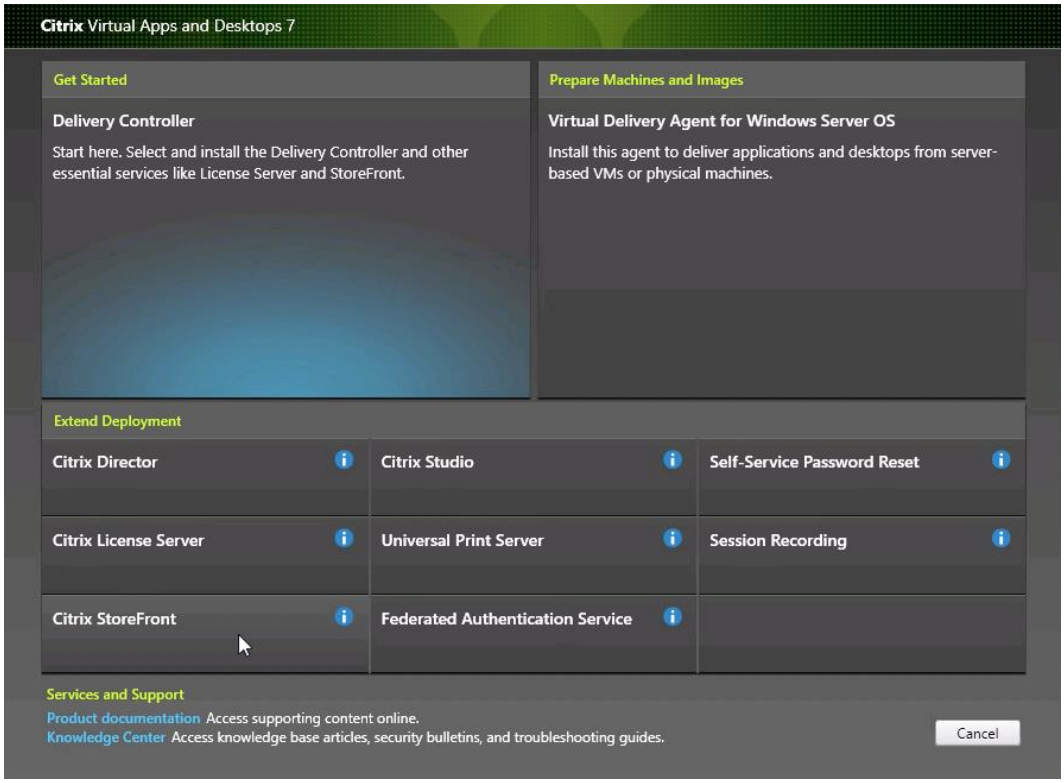
To install Citrix StoreFront, complete the following steps:

1. Deploy the supported Windows server by using ISO or from the VM template:
2. Mount the Citrix Virtual Apps and Desktops ISO file.
3. Run `AutoSelect.exe` at root of the drive.

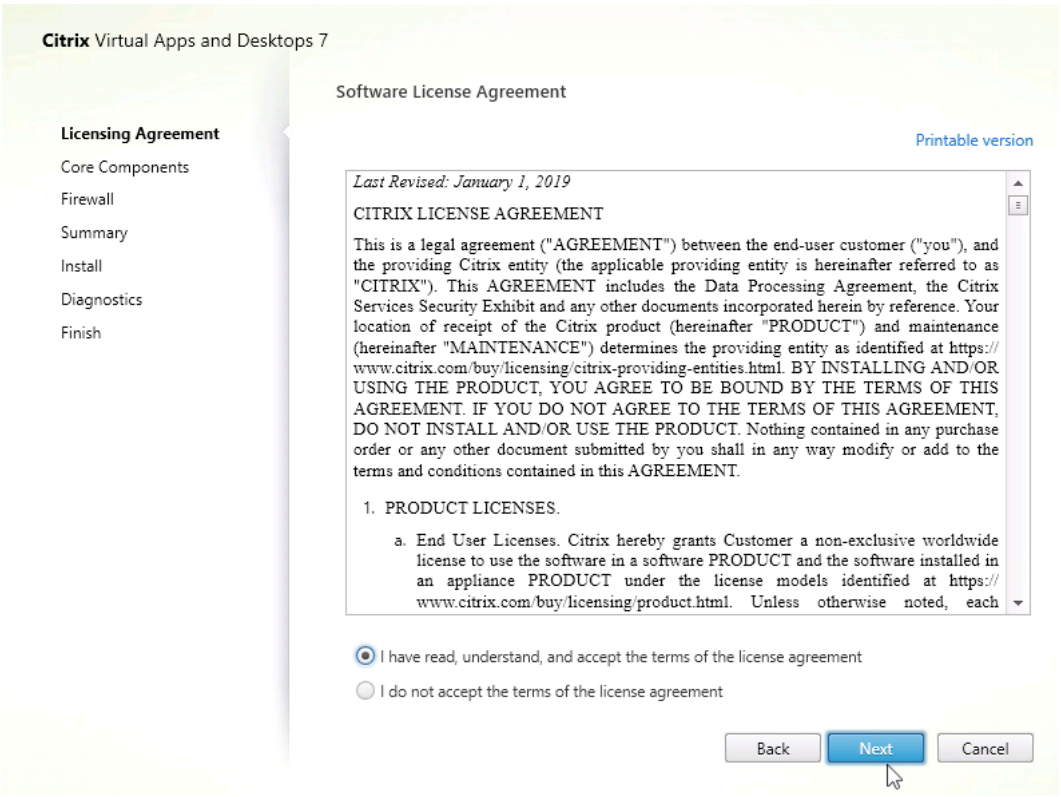
4. Select Start on the Virtual Apps and Desktops option.



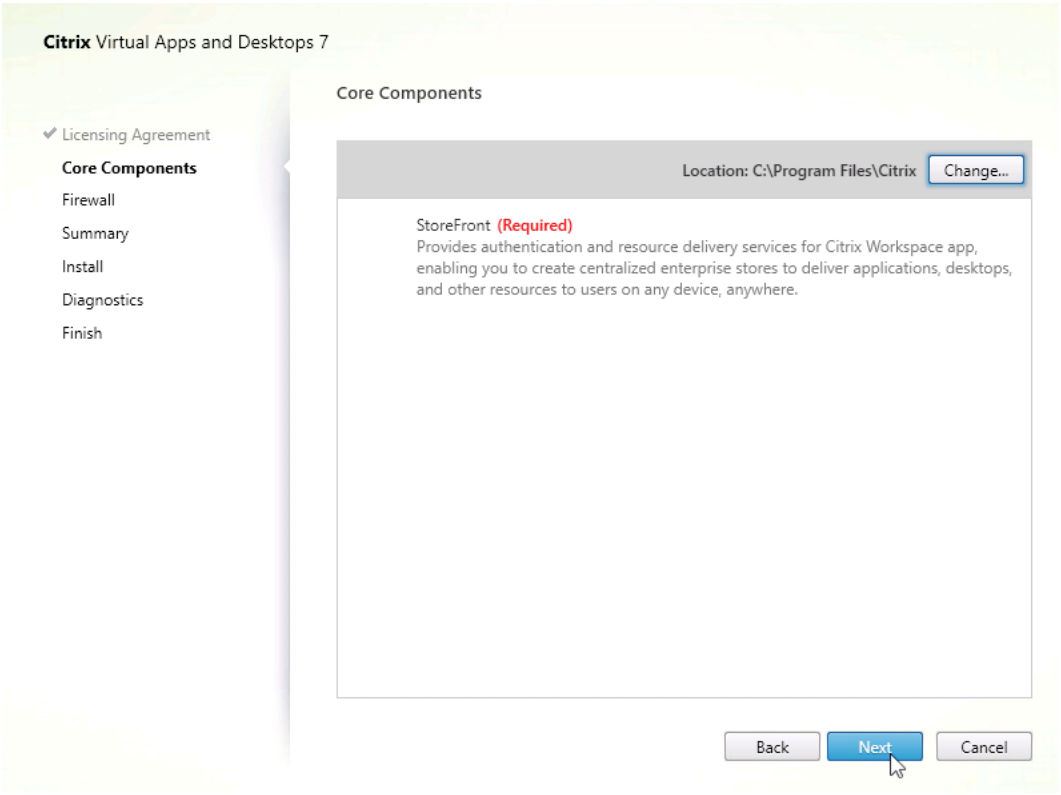
5. Click Citrix StoreFront.



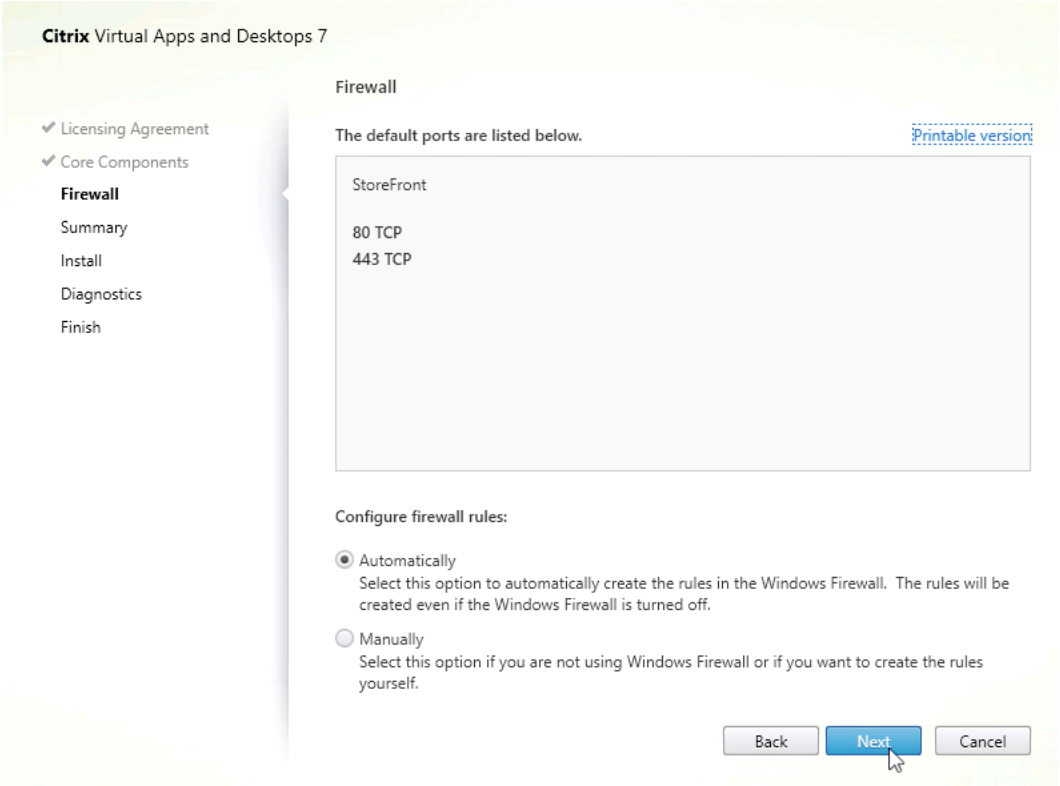
6. Review the license agreement and select the appropriate option. Click Next.



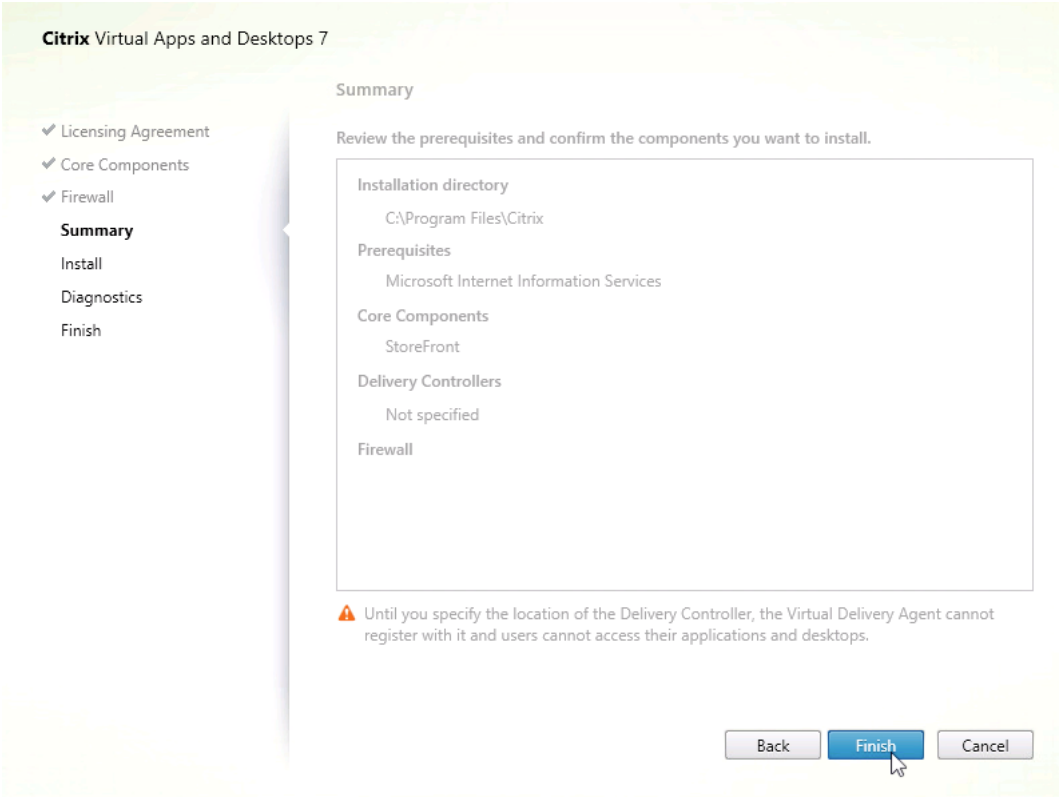
7. Review the installation location and click Next.



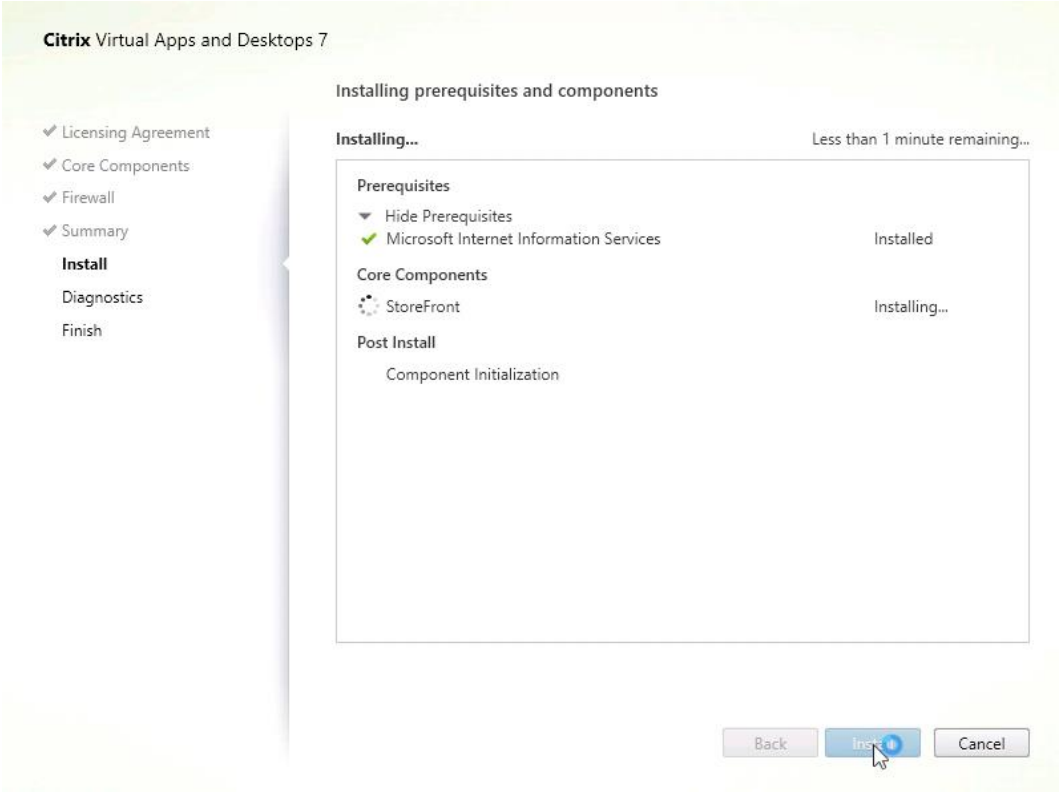
8. Review the firewall rules and select the option to configure the rules. Click Next



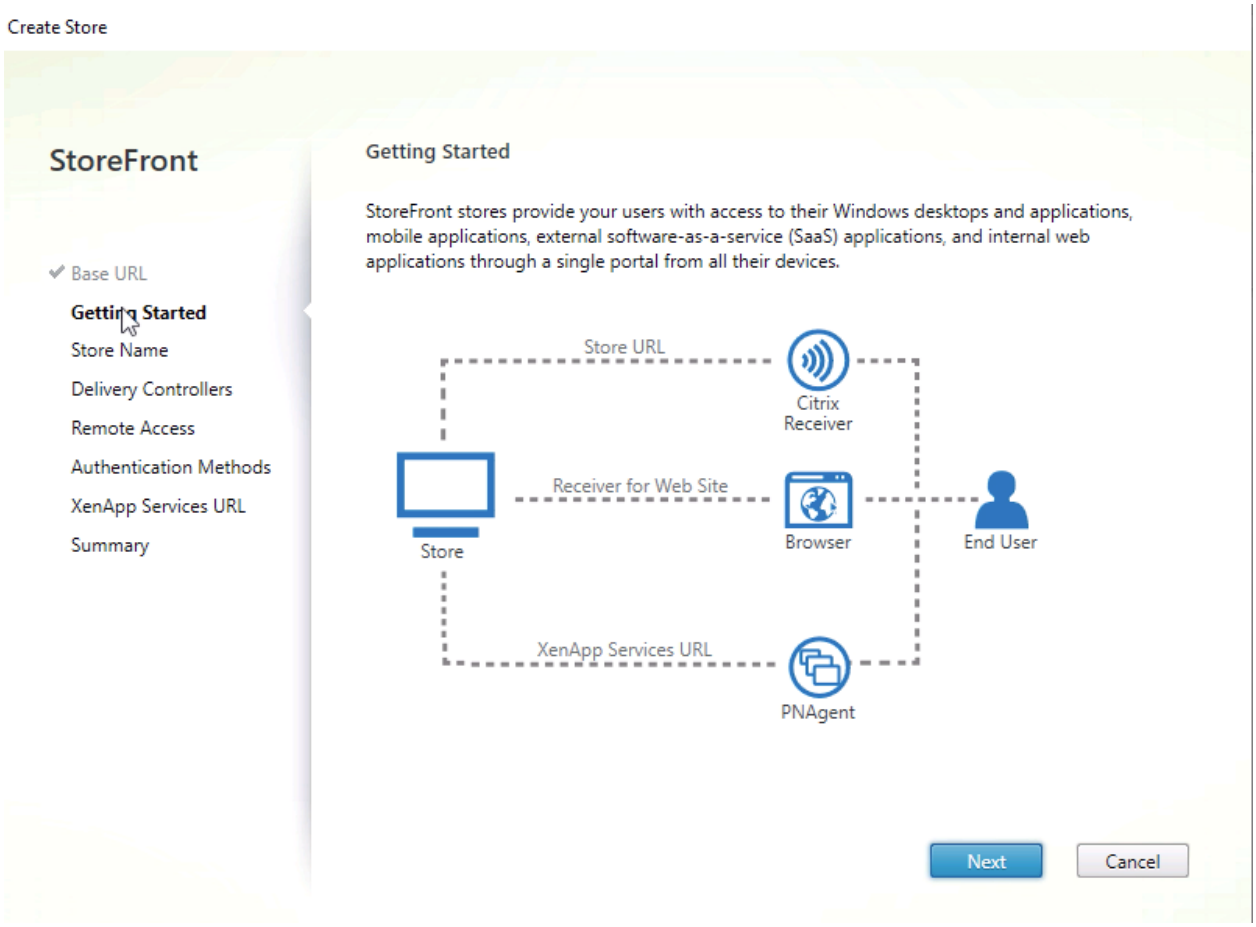
9. Review the summary and click Finish.



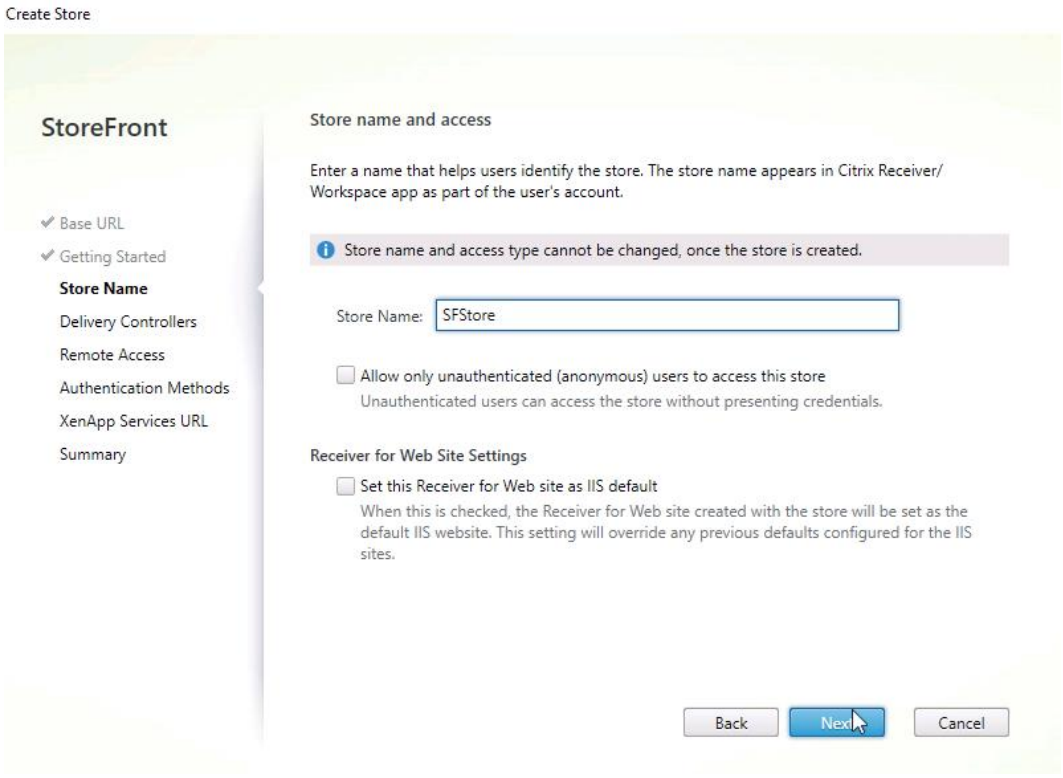
10. The installation progress is displayed. Click Next when it completes.



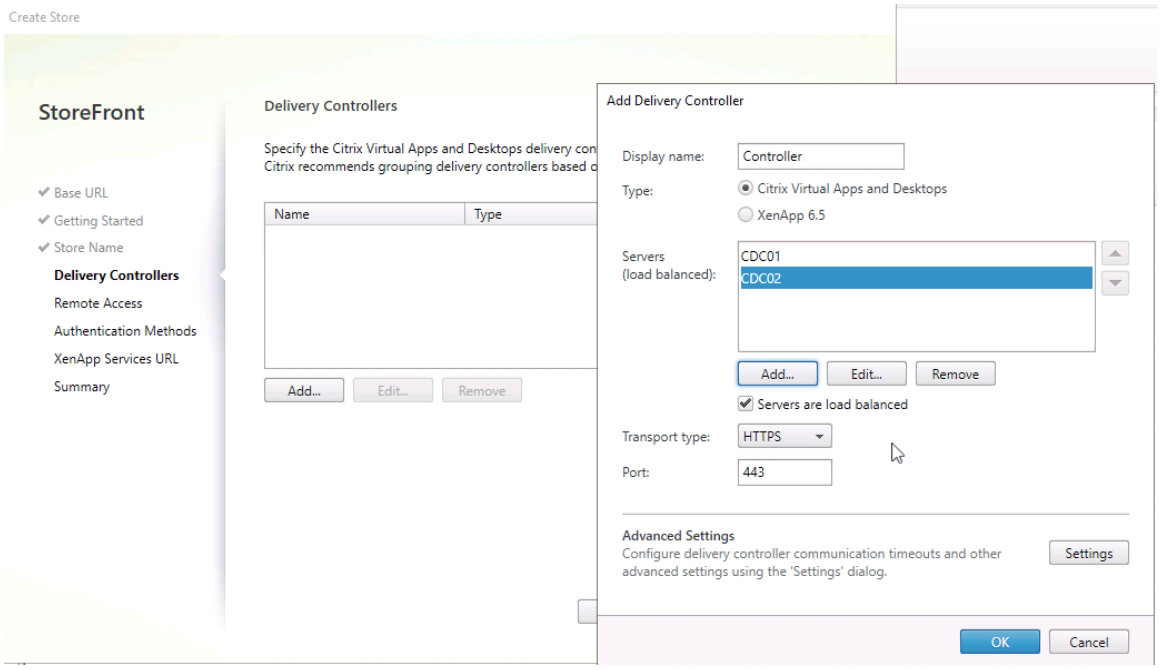
11. After the installation process is complete, follow the remaining wizard instructions. Launch the Citrix StoreFront icon, click Stores, and click Add Store under the action pane.
12. Accept the base URL and click Next.



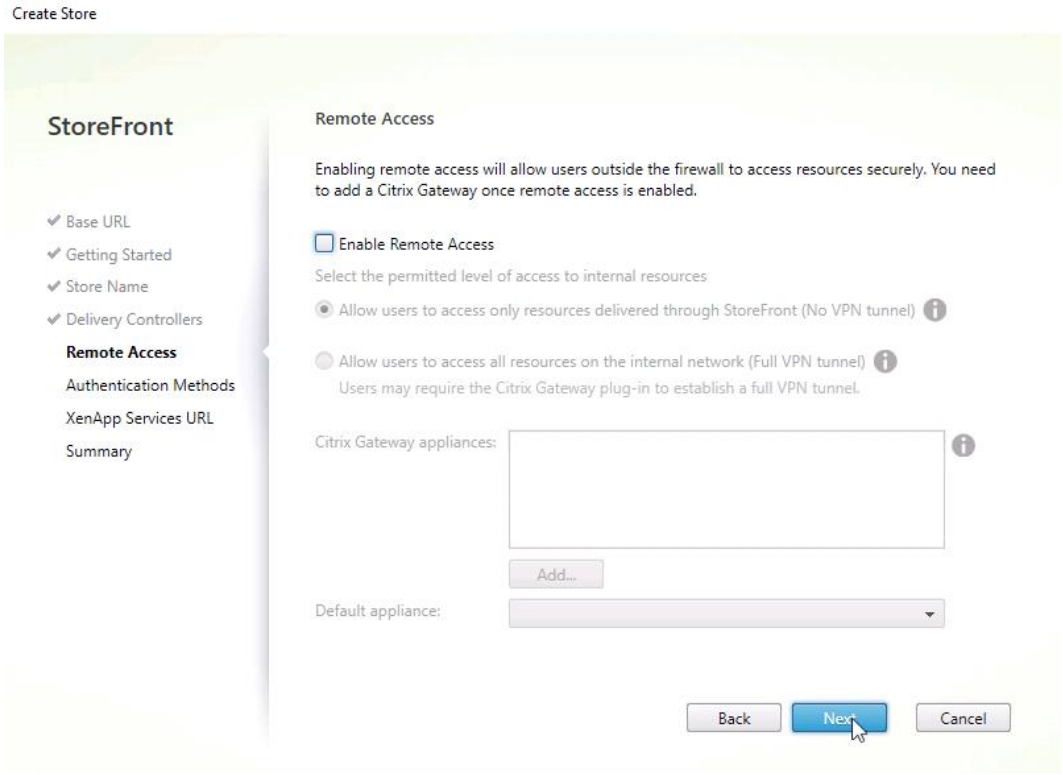
13. Click Next to go to the Store Name page. Provide a store name and select the other applicable options. Click Next.



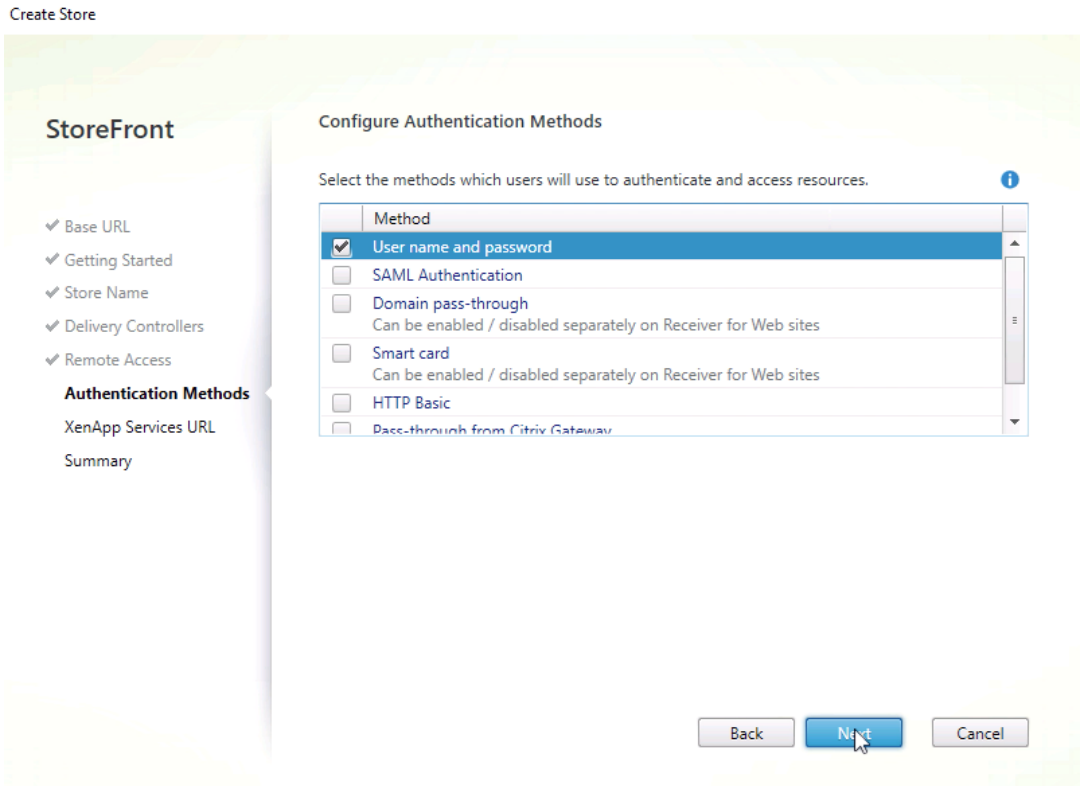
14. Click Add to add the Citrix Delivery Controller information. Under the Servers section, click Add to include the server details. Pick the appropriate transport type and click OK. Click Next.



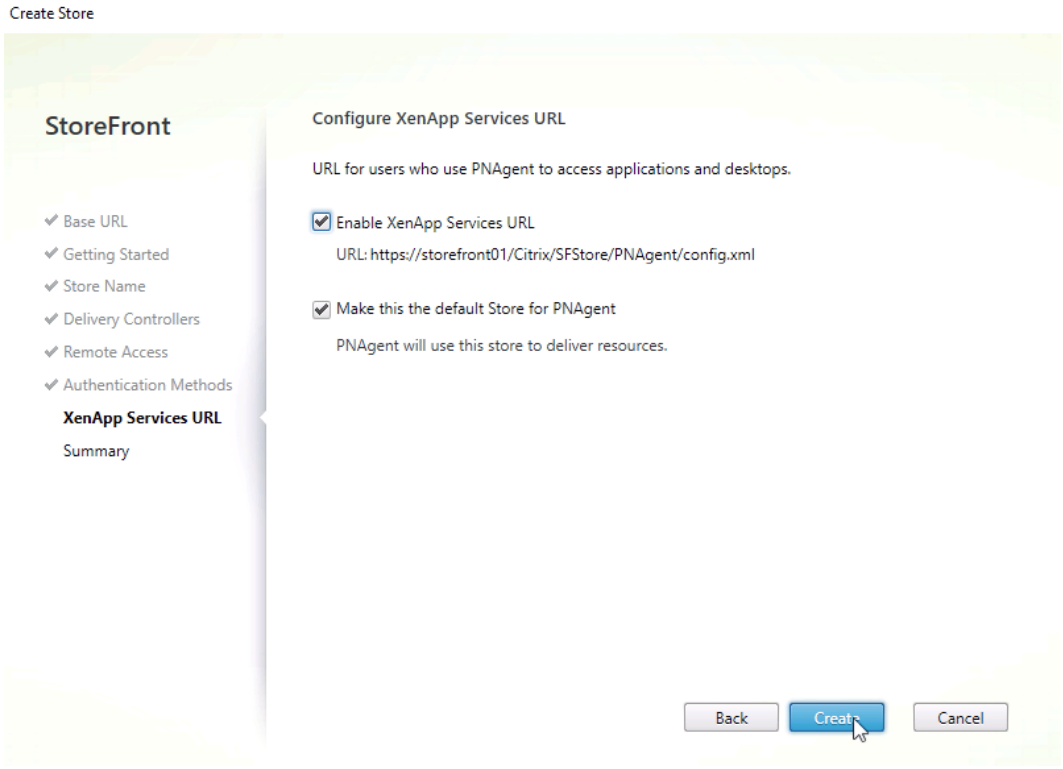
15. Select the appropriate remote access option and click Next.



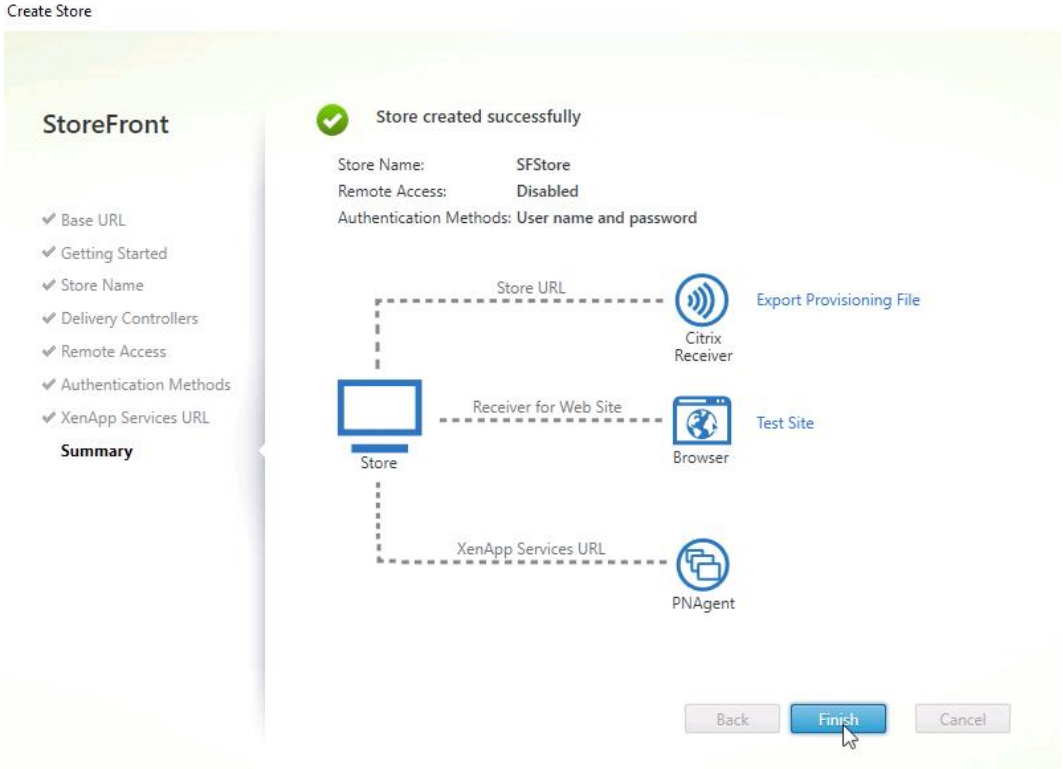
16. Select the authentication methods and click Next.



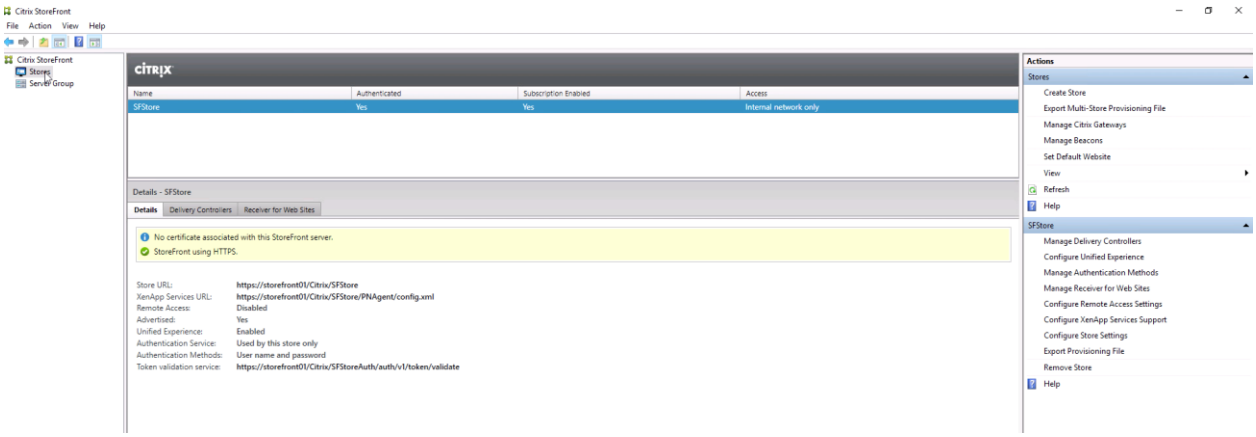
17. Make changes to XenApp Services URL options, if required, and click Create.



18. Review the summary and click Finish.



19. The newly created Store is now listed under the Stores section.

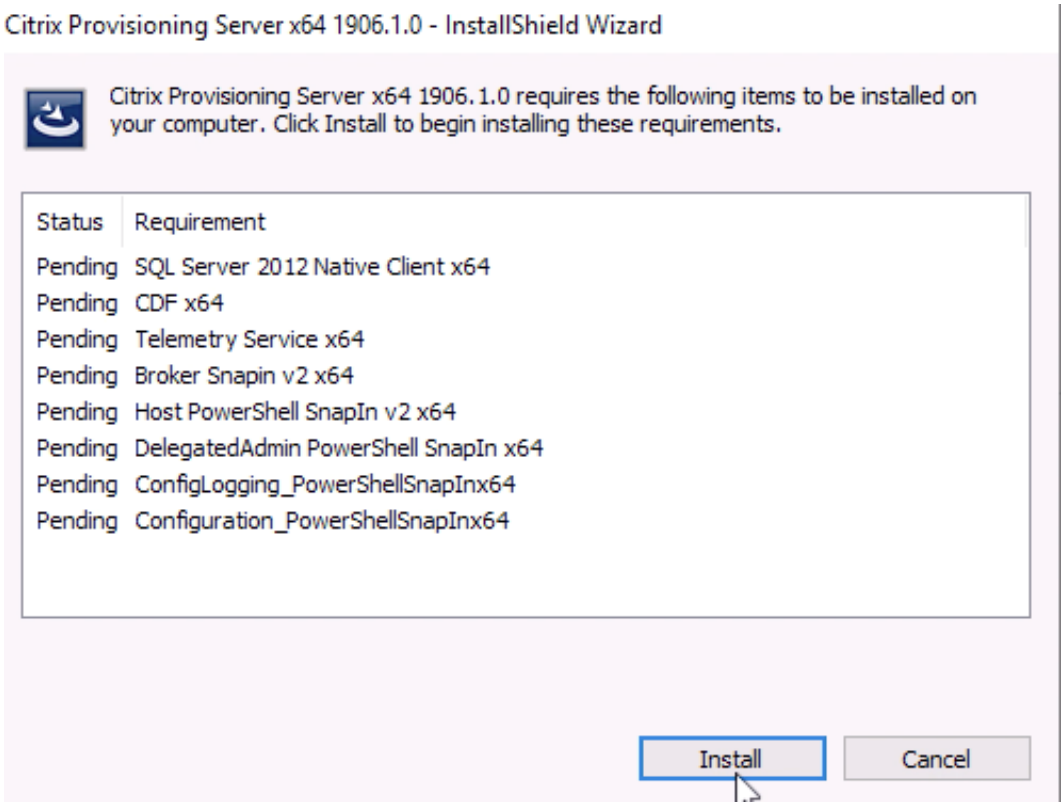


8.5 Citrix Provisioning Server

To install Citrix Provisioning Server, complete the following steps:

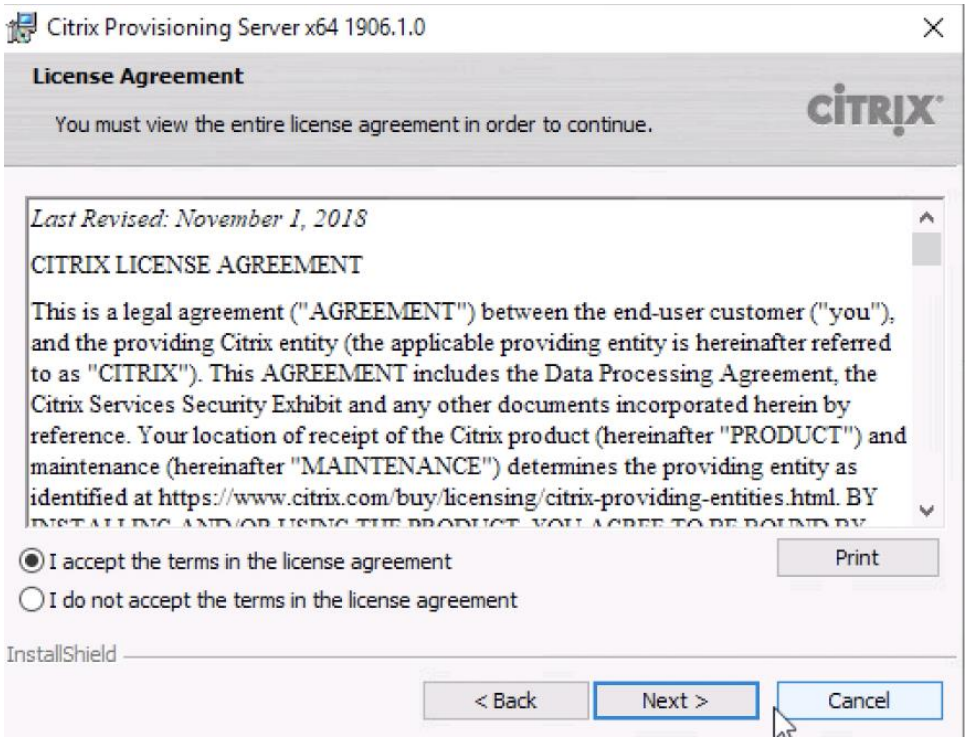
1. Deploy the supported Windows server by using ISO or from the VM template.
2. Mount the Citrix Provisioning ISO file.
3. Select Server Installation.

4. The Citrix Provisioning software prerequisites are verified; you are provided an option to install those. Click Install.

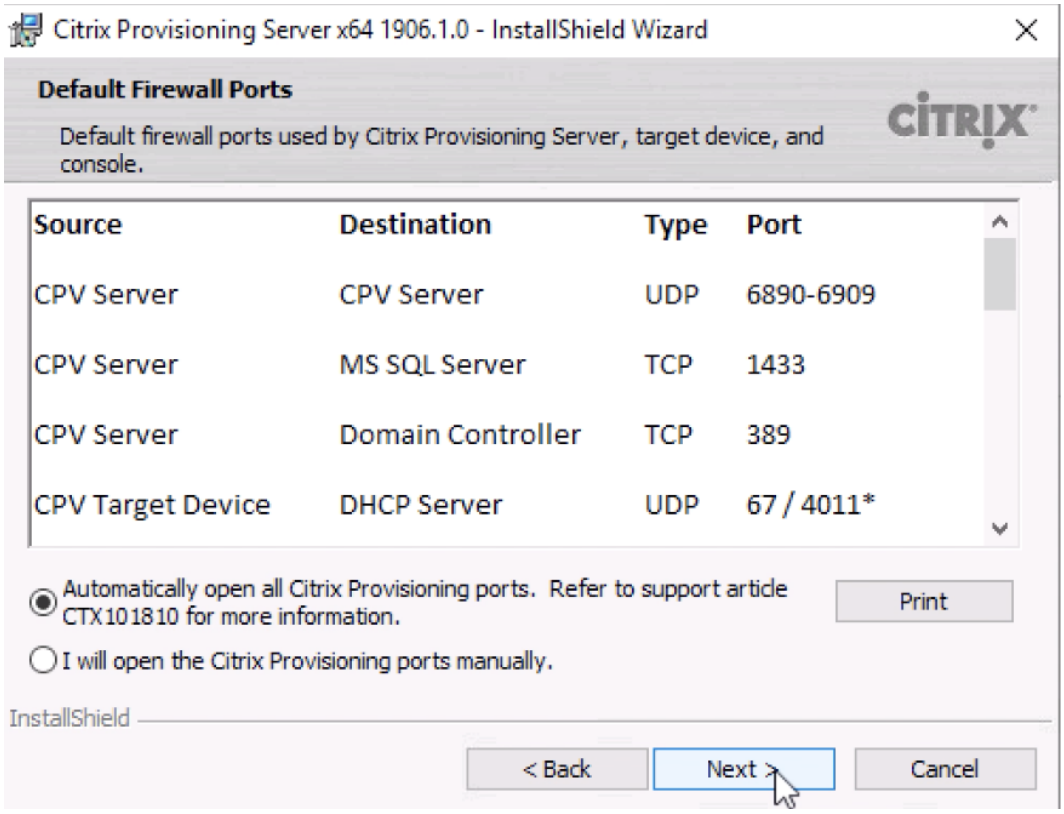


5. After the software is installed, the server installation wizard starts. Click Next.

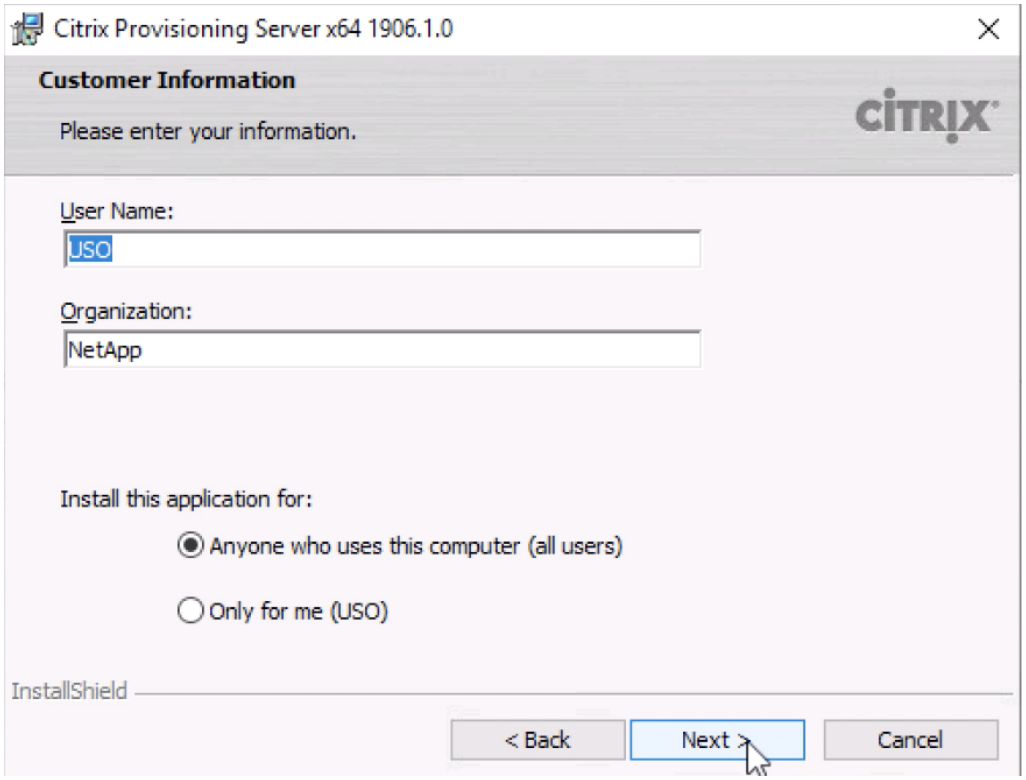
6. Review the license agreement and make the appropriate selection. Click Next.



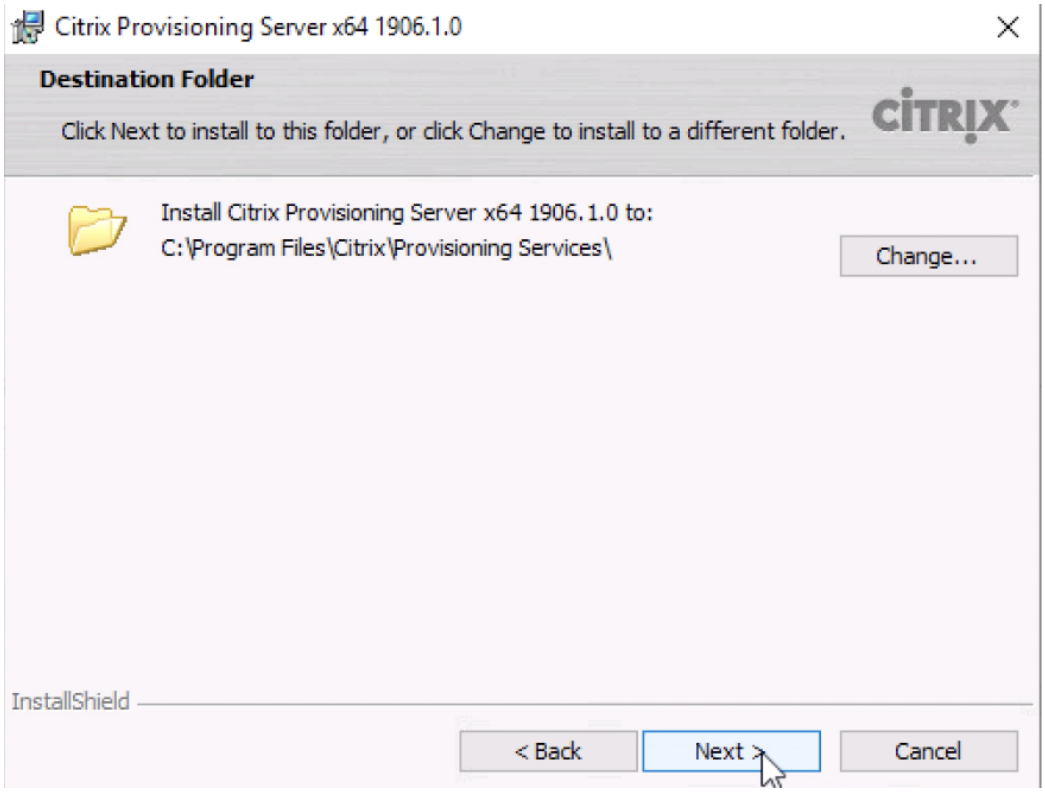
7. Review the firewall ports and select the option to configure the firewall ports. Click Next



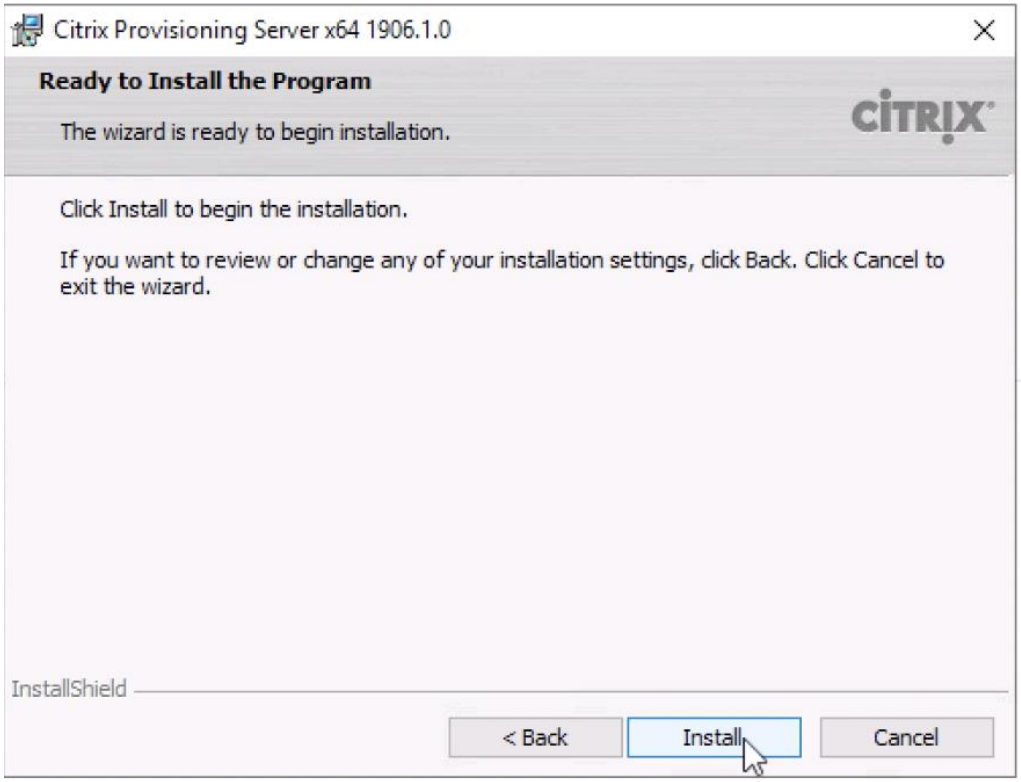
8. Provide user and organization name. Select the application available to users on this server. Click Next.



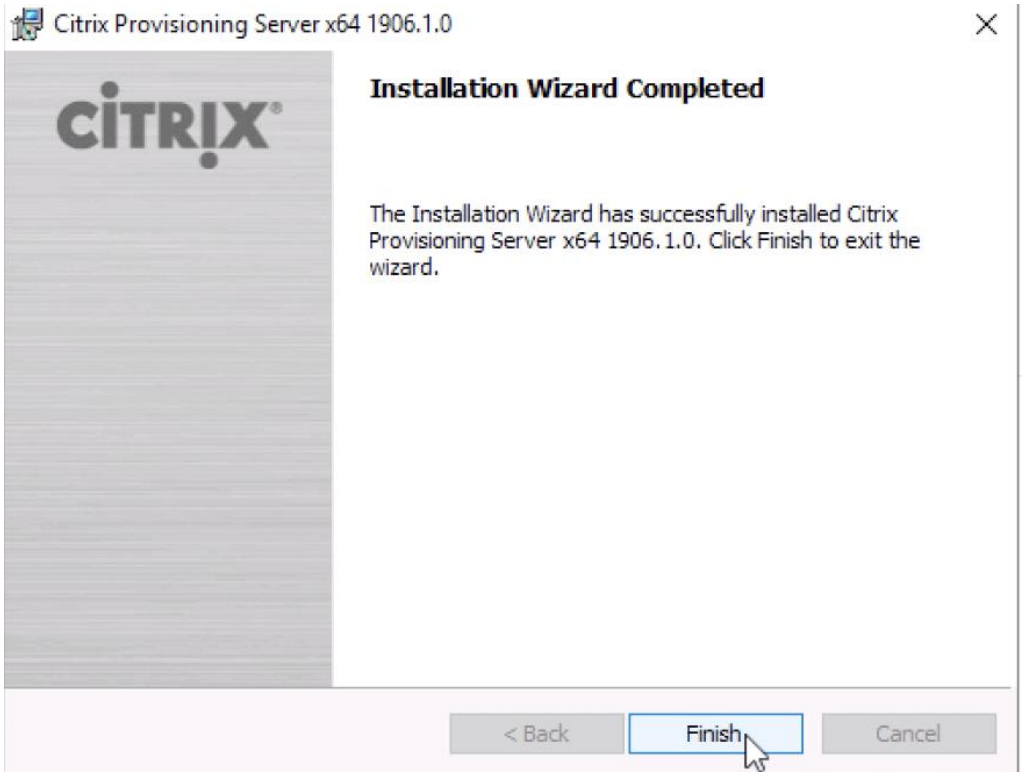
9. Review the destination folder and click Next.



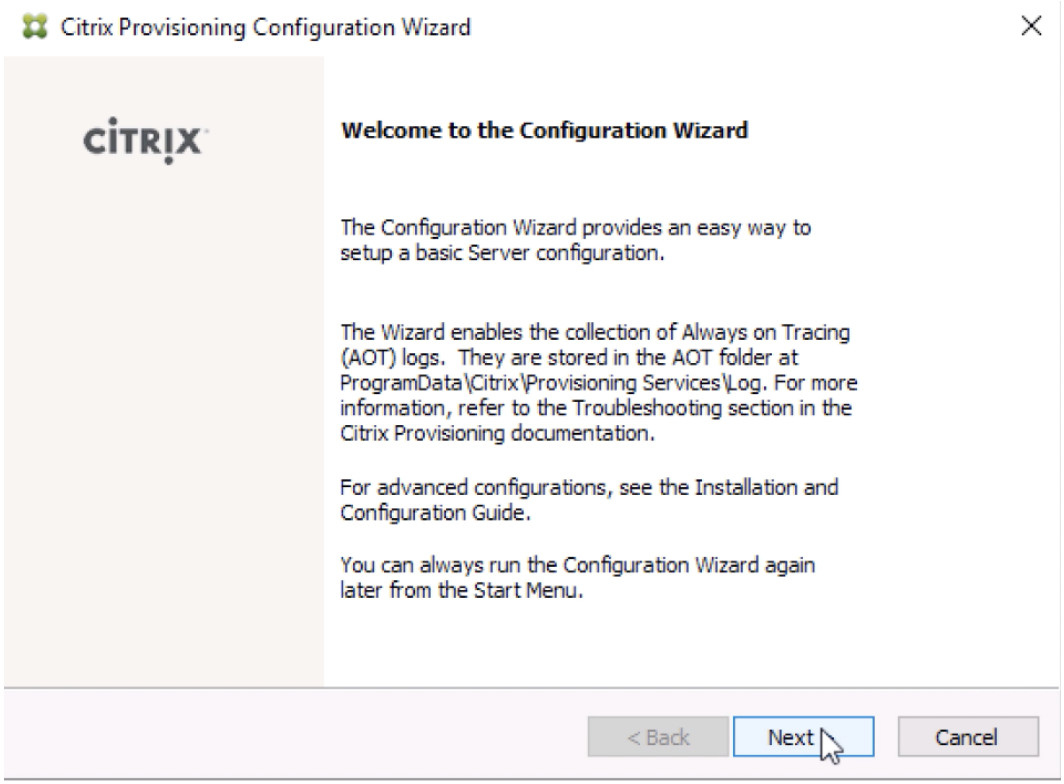
10. Click Install.



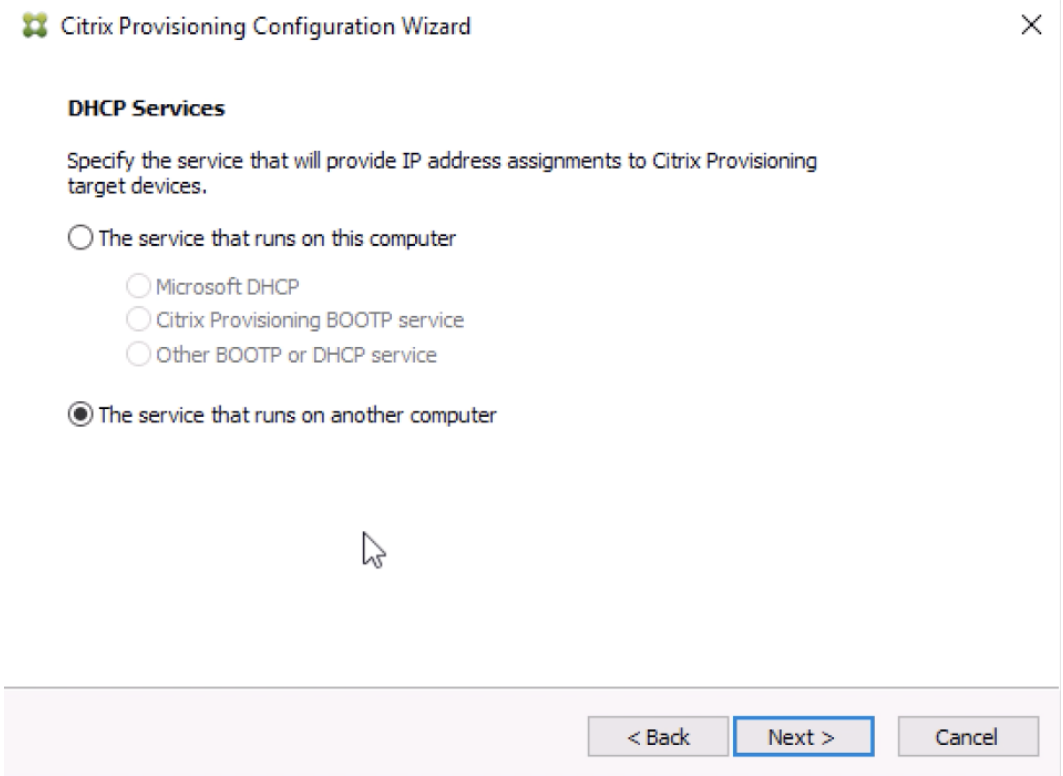
11. The installation progress is displayed. When the installation is complete, click Finish .



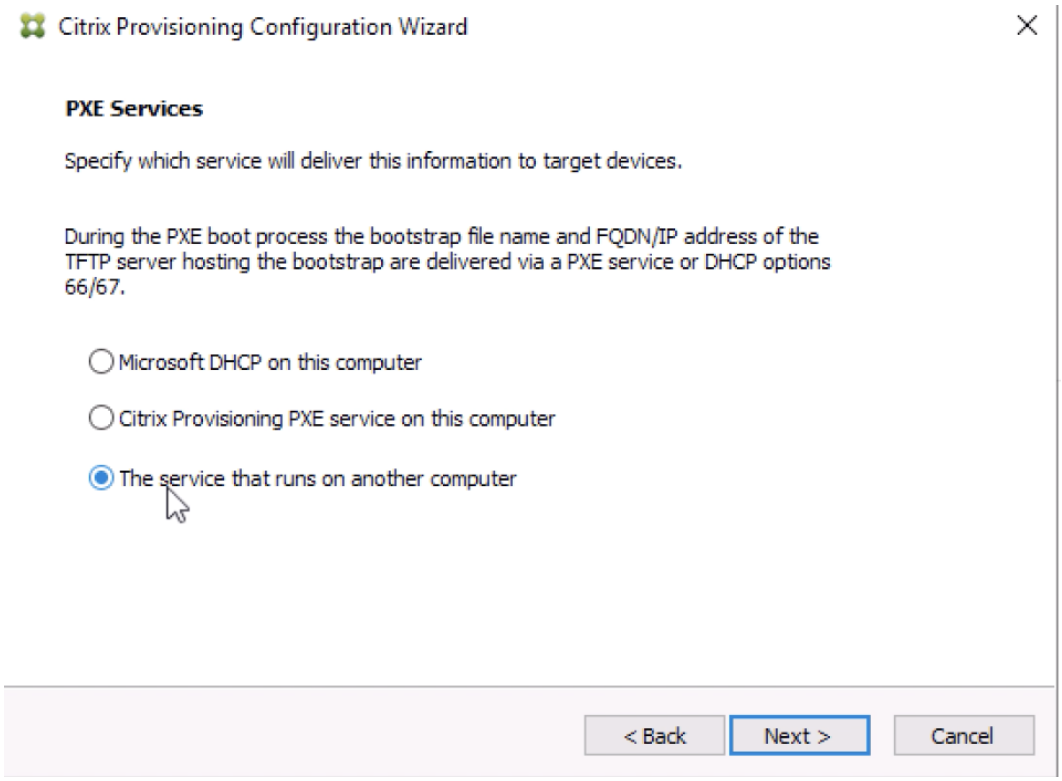
12. The Configuration Wizard starts automatically. On the Welcome page, click Next.



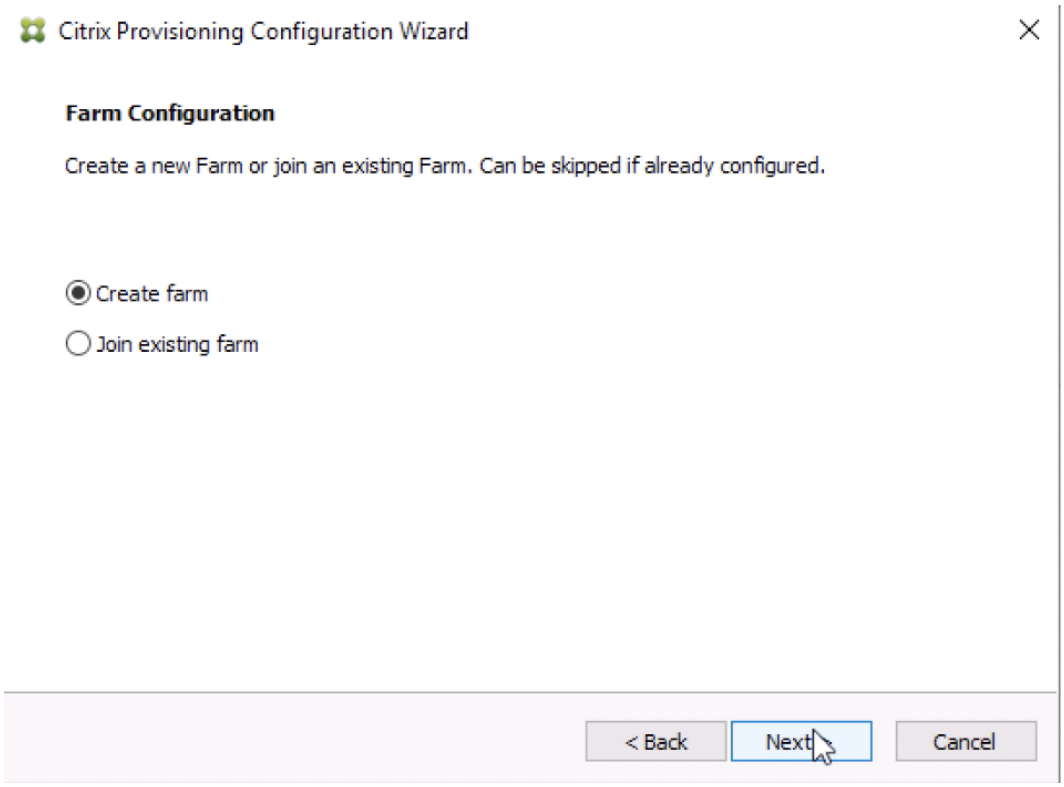
13. Select the DHCP option and click Next.



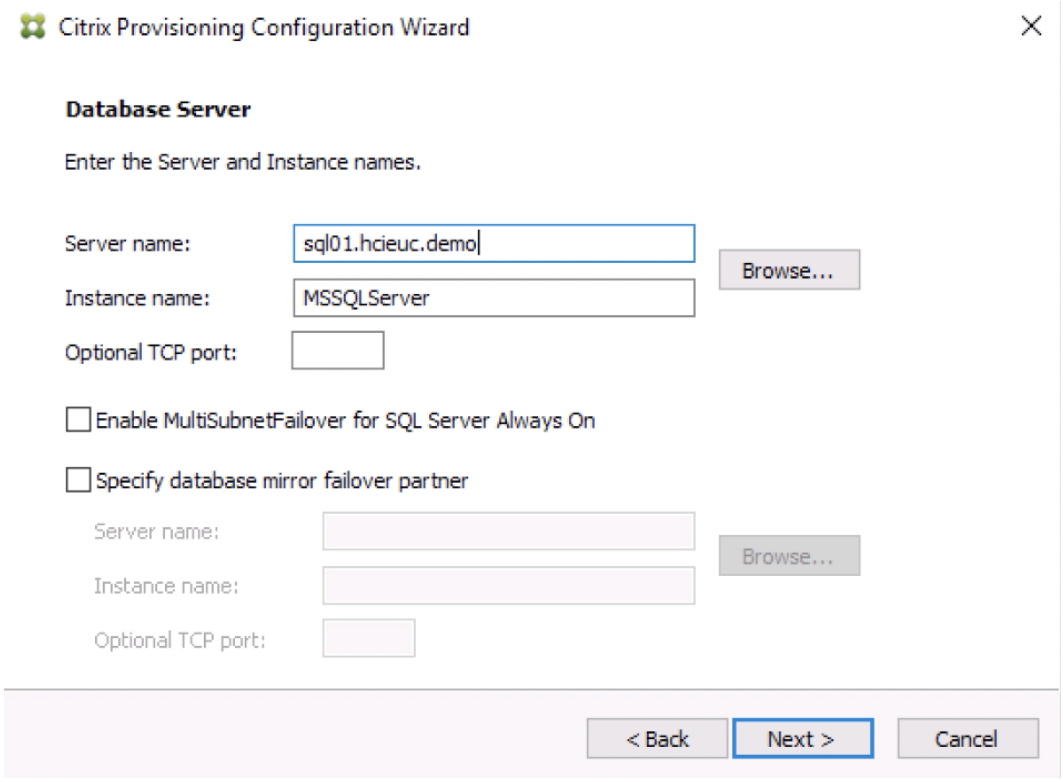
14. Select the PXE service option and click Next.



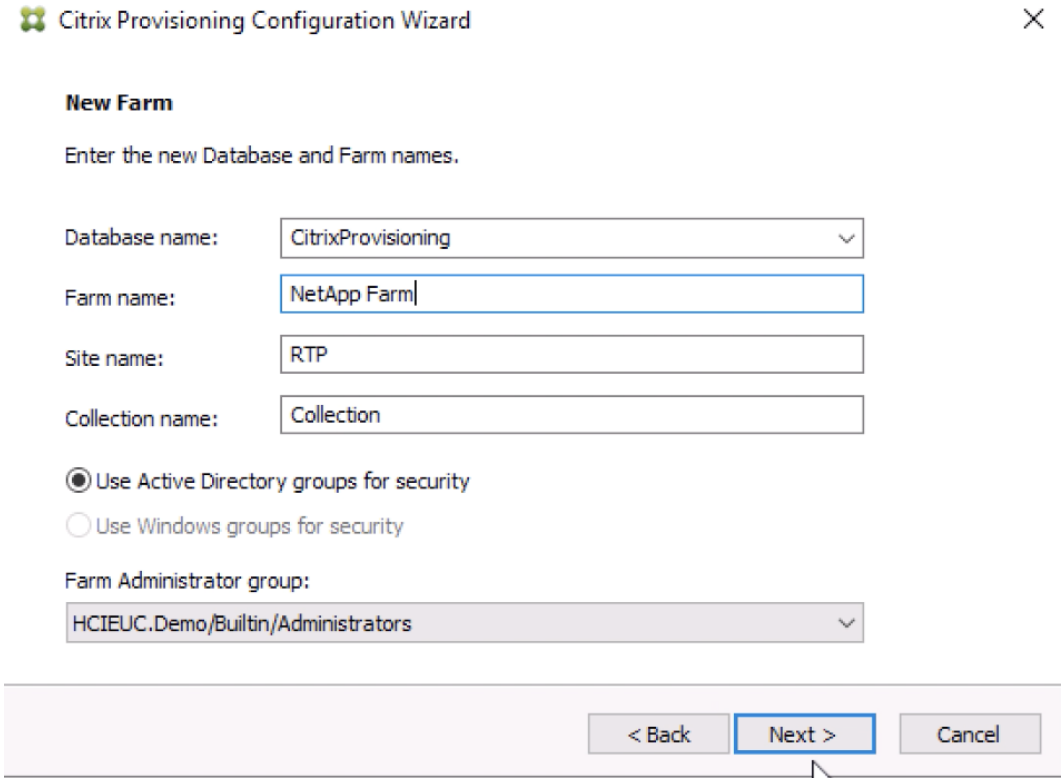
15. Select the Farm configuration option. Because this is the first server, select Create Farm. Click Next.



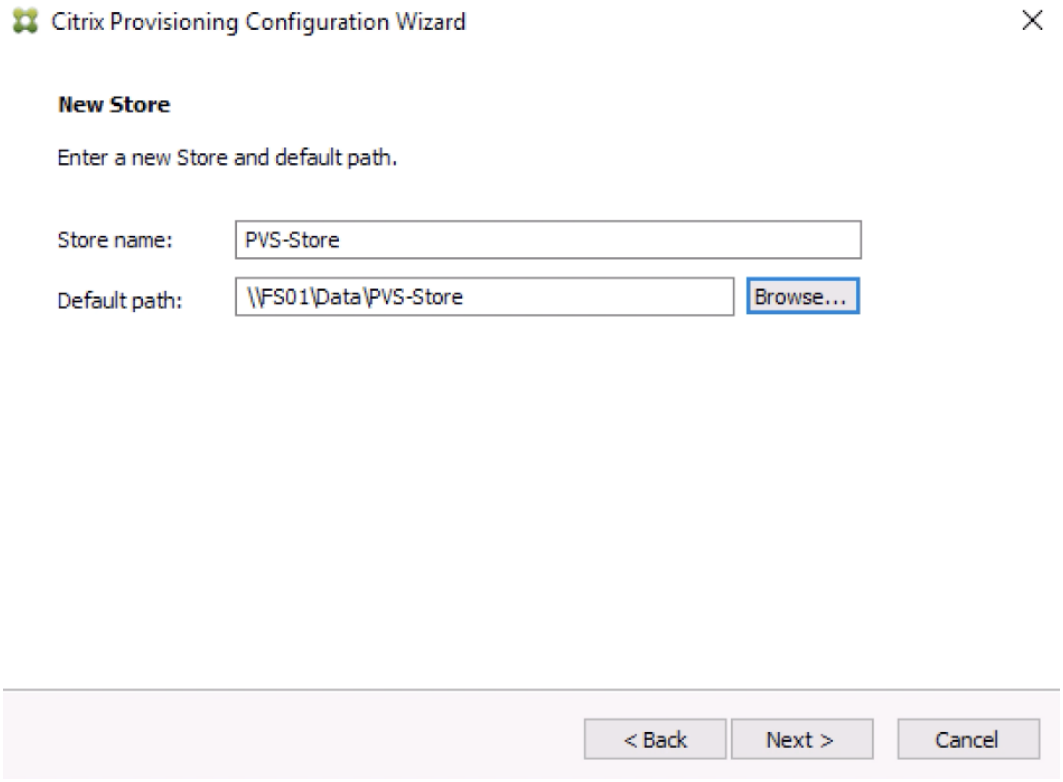
16. Enter the SQL Server database server information and click Next.



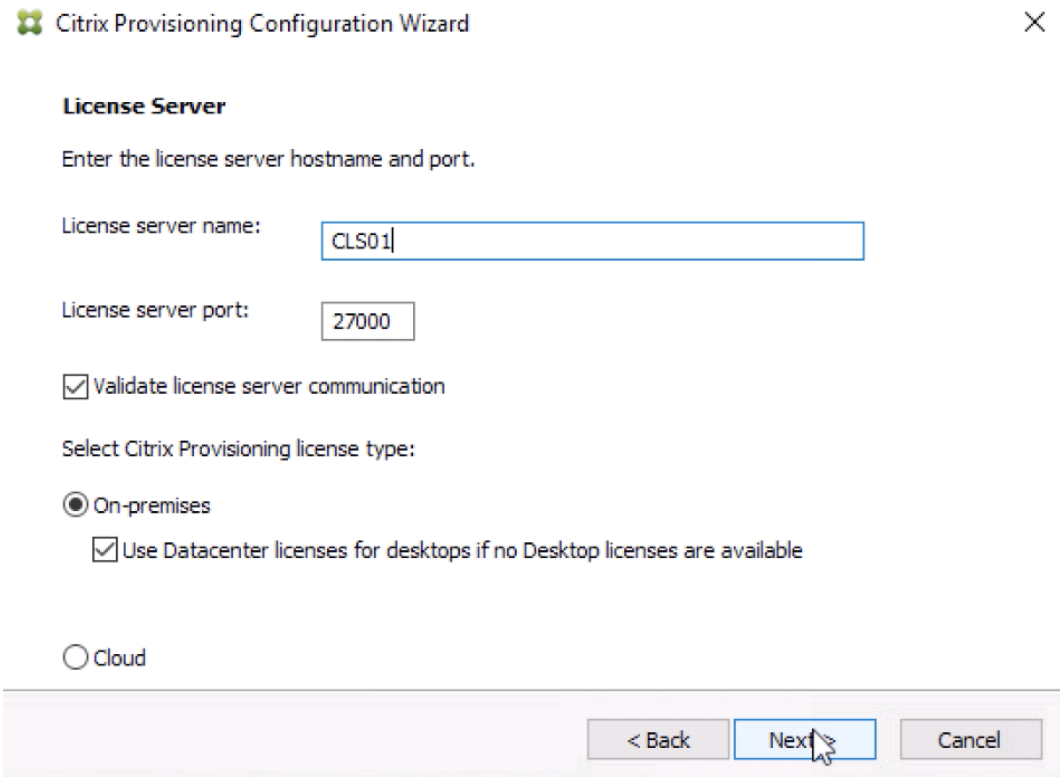
17. Enter the Farm information and click Next.



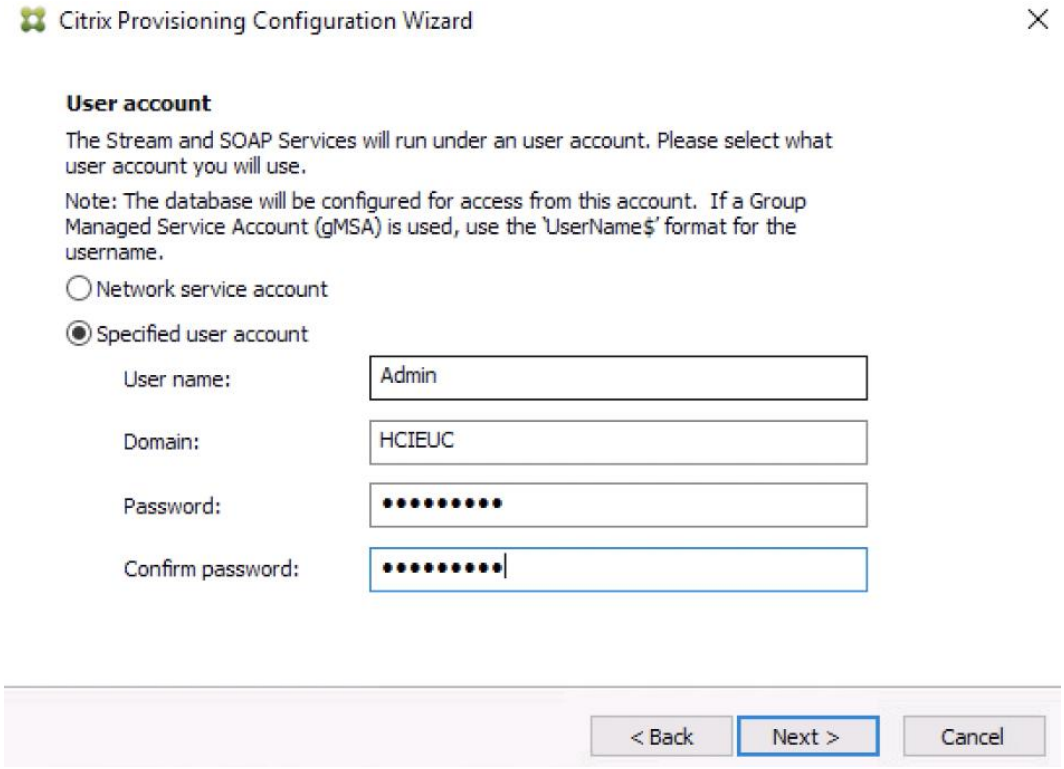
18. Enter the PVS Store details and click Next.



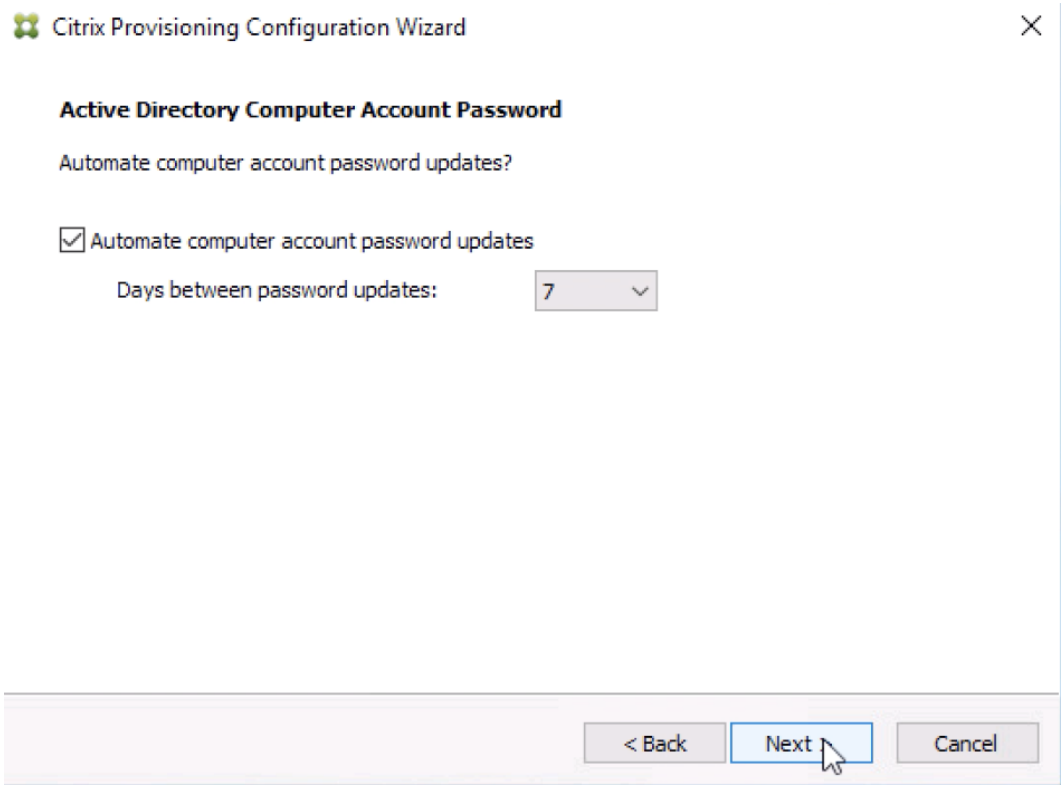
19. Enter the Citrix License Server information and click Next.



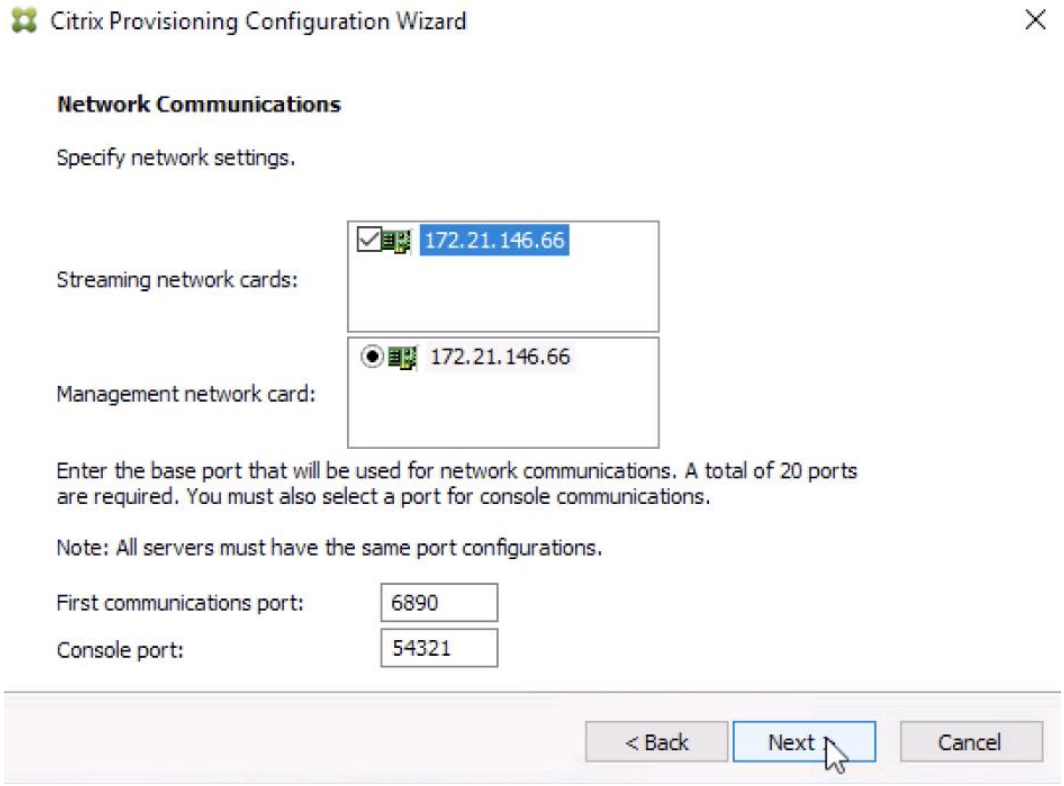
20. Enter the user account information for the service account and click Next.



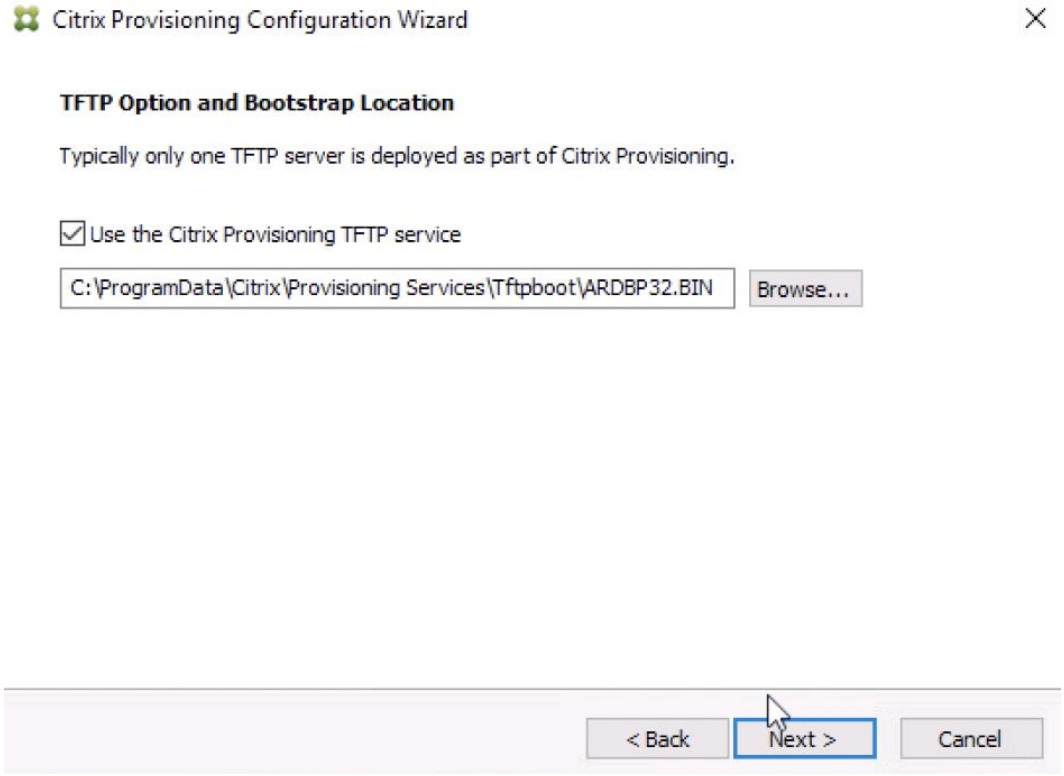
21. Select the option for your computer account password update and its frequency. Click Next.



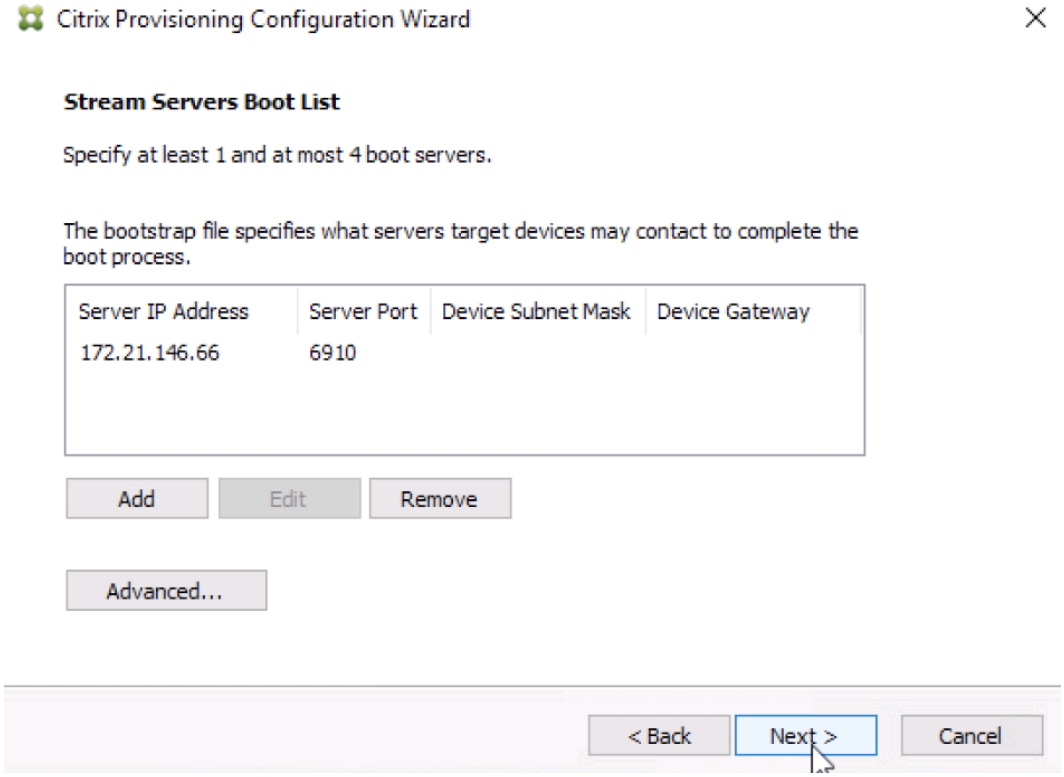
22. Select the networks and review the port information. Click Next.



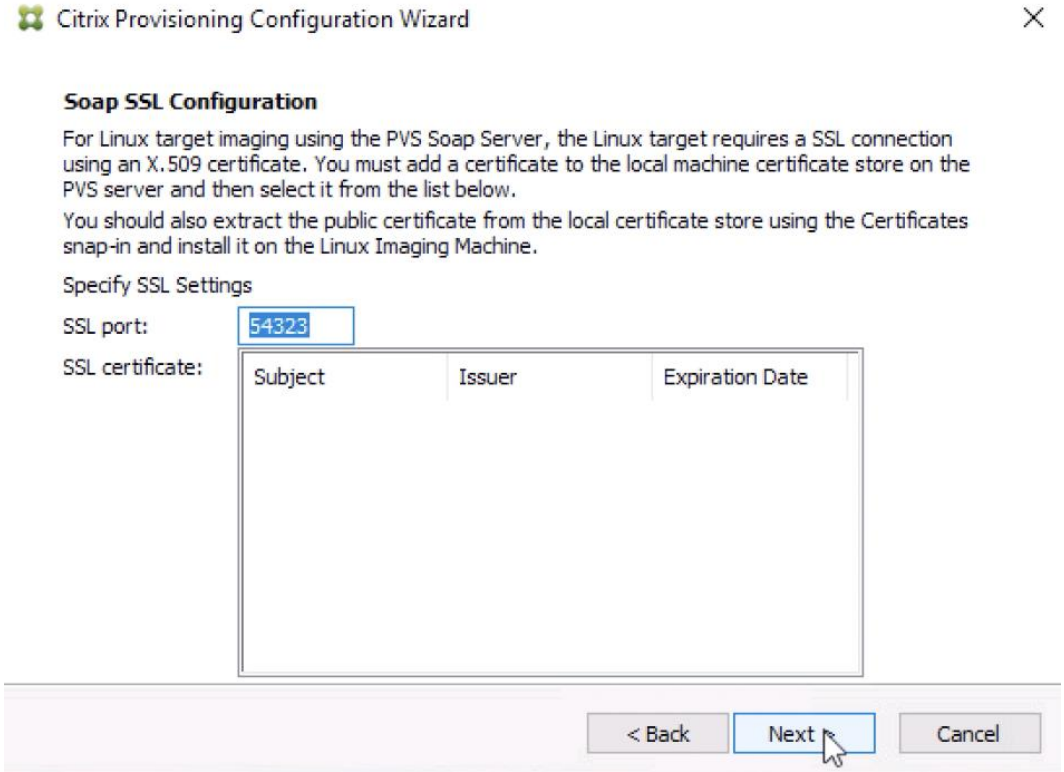
23. Verify the TFTP option and Bootstrap location and click Next.



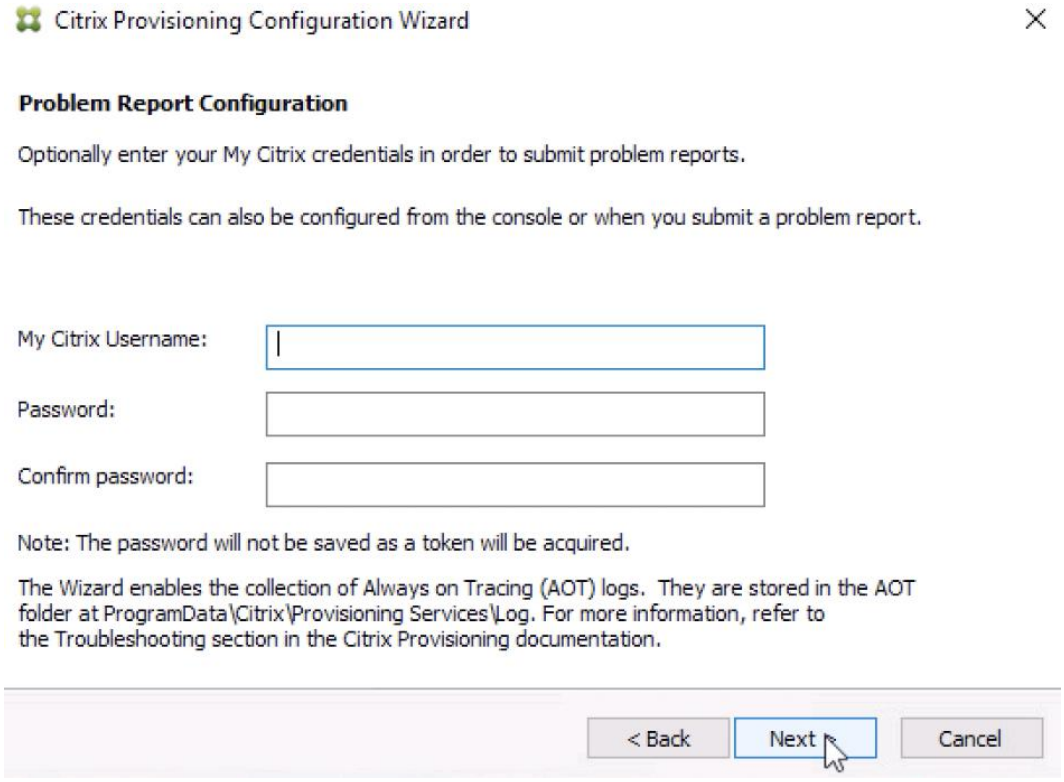
24. Verify the stream server list and click Next.



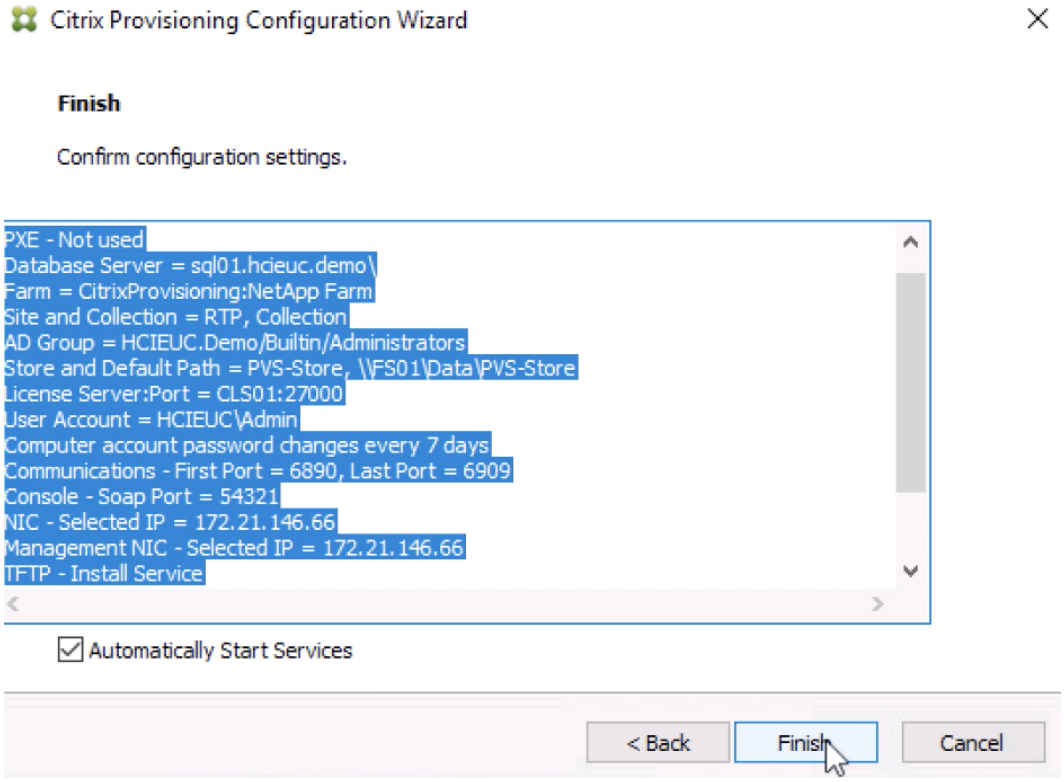
25. Verify the SOAP SSL configuration and click Next



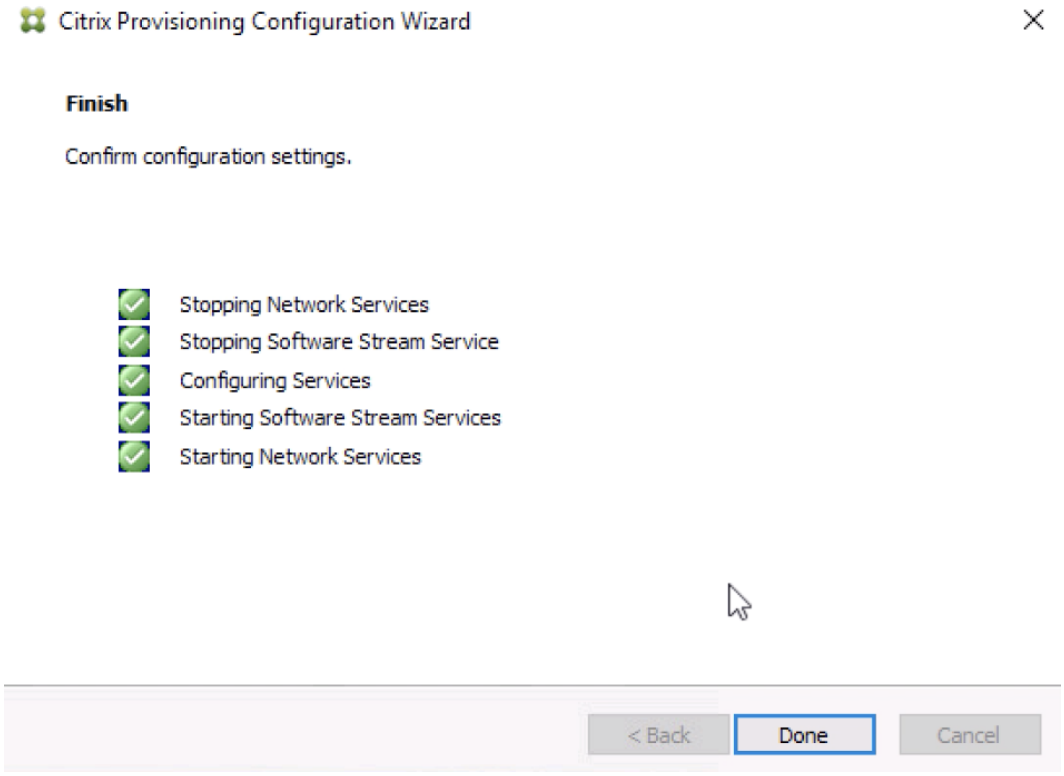
26. For problem reporting, enter your My Citrix credentials and click Next.



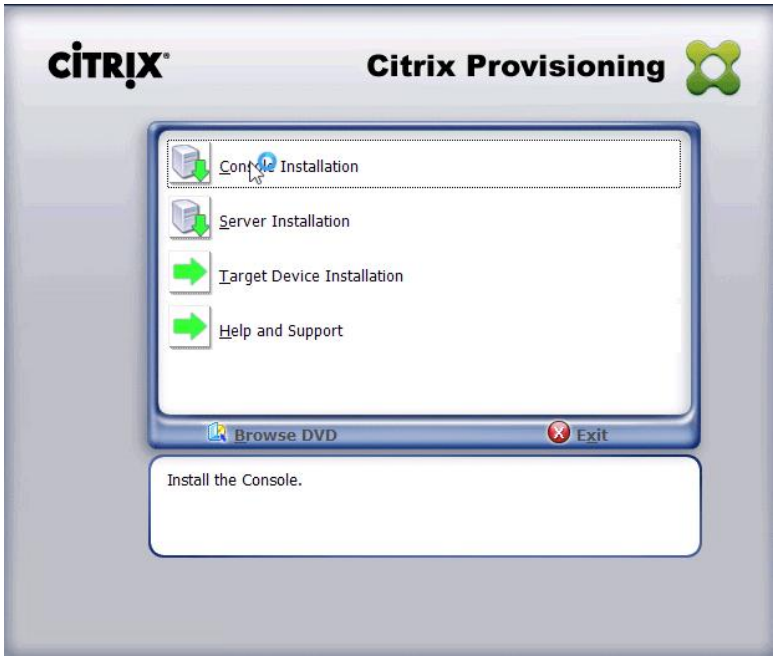
27. Review the configuration settings and click Finish.



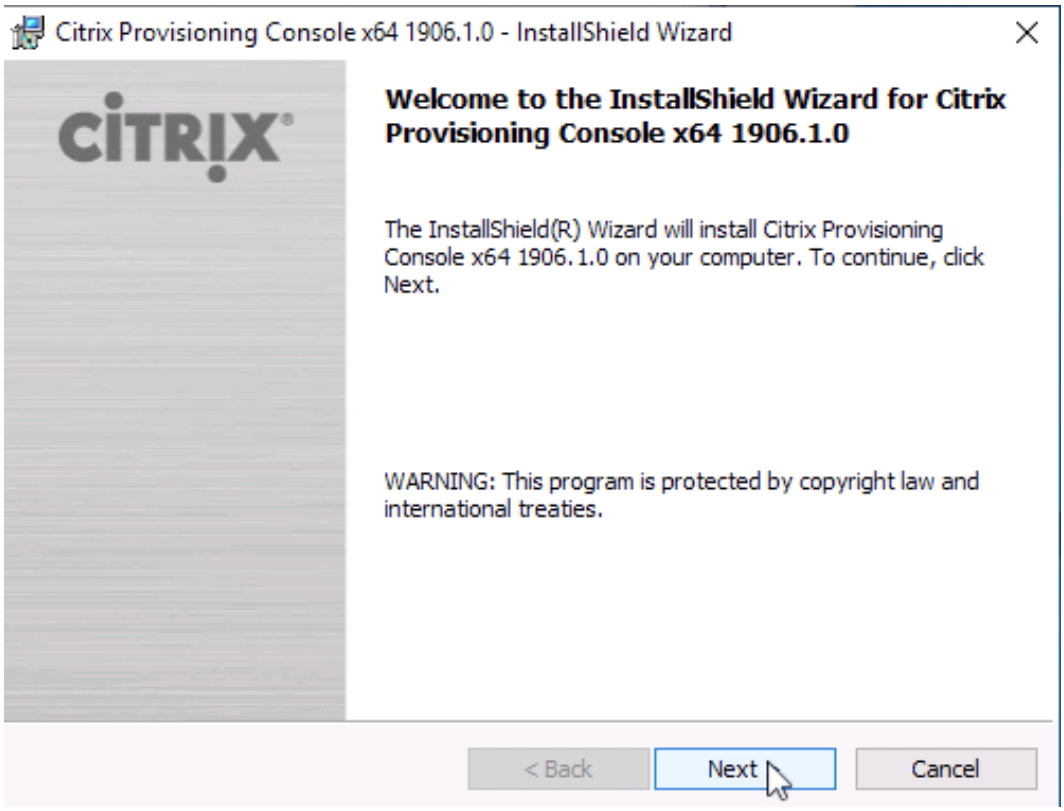
28. After the configuration statuses are enabled, click Done.



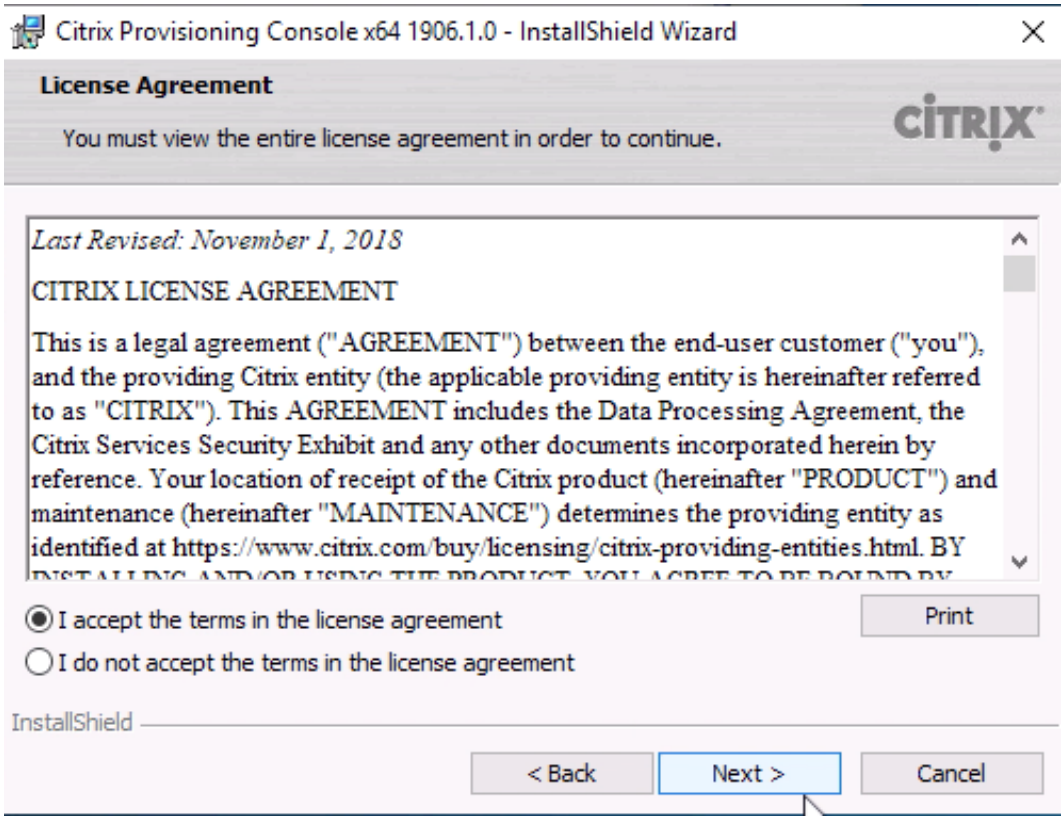
29. Mount Citrix Provisioning ISO image and double-click on it.
30. Start the PVS Console installation by clicking on Console Installation.



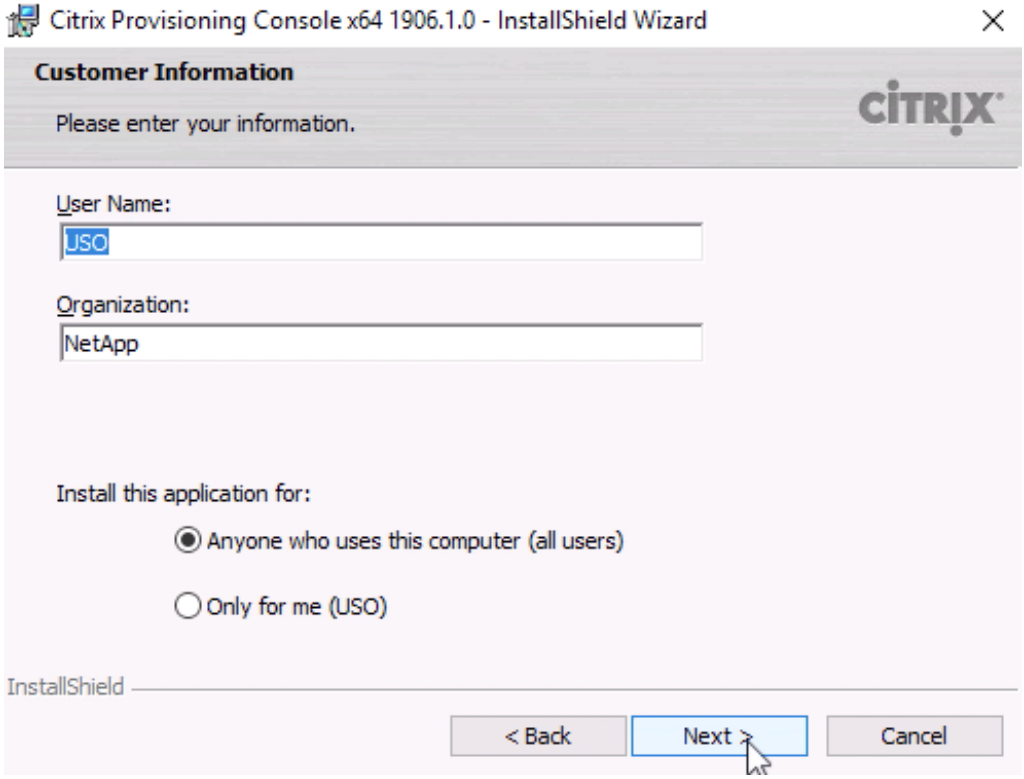
31. The installation wizard is launched. On the welcome page. click Next.



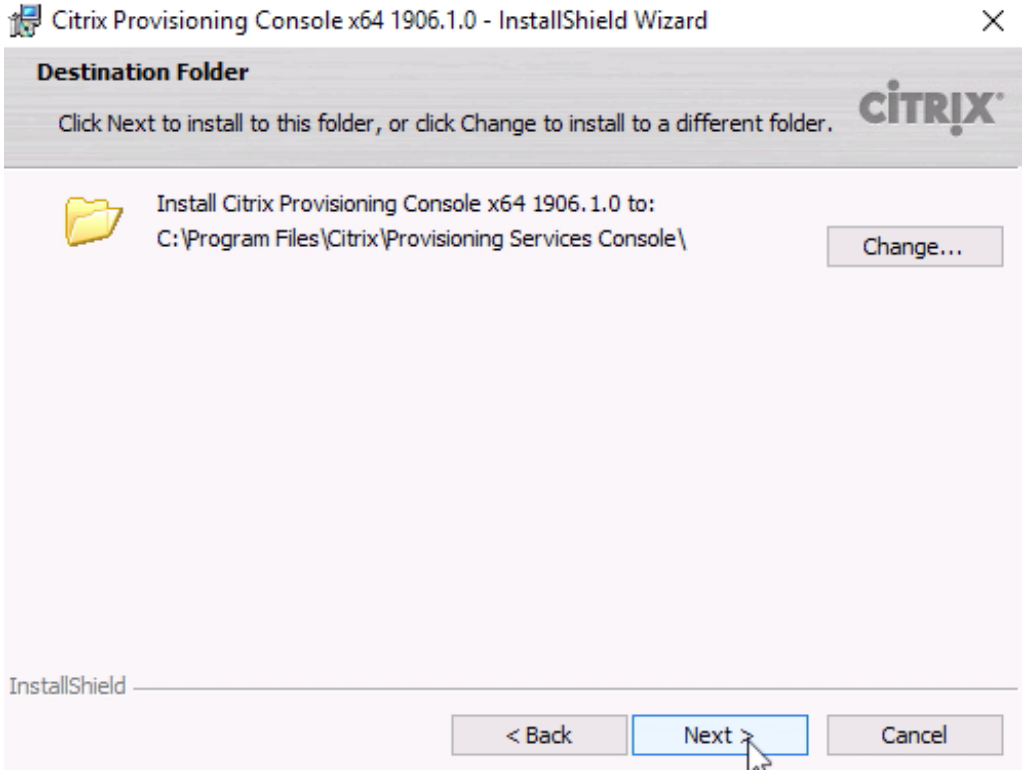
32. Review Citrix PVS Console license agreement and select the appropriate option. Click Next.



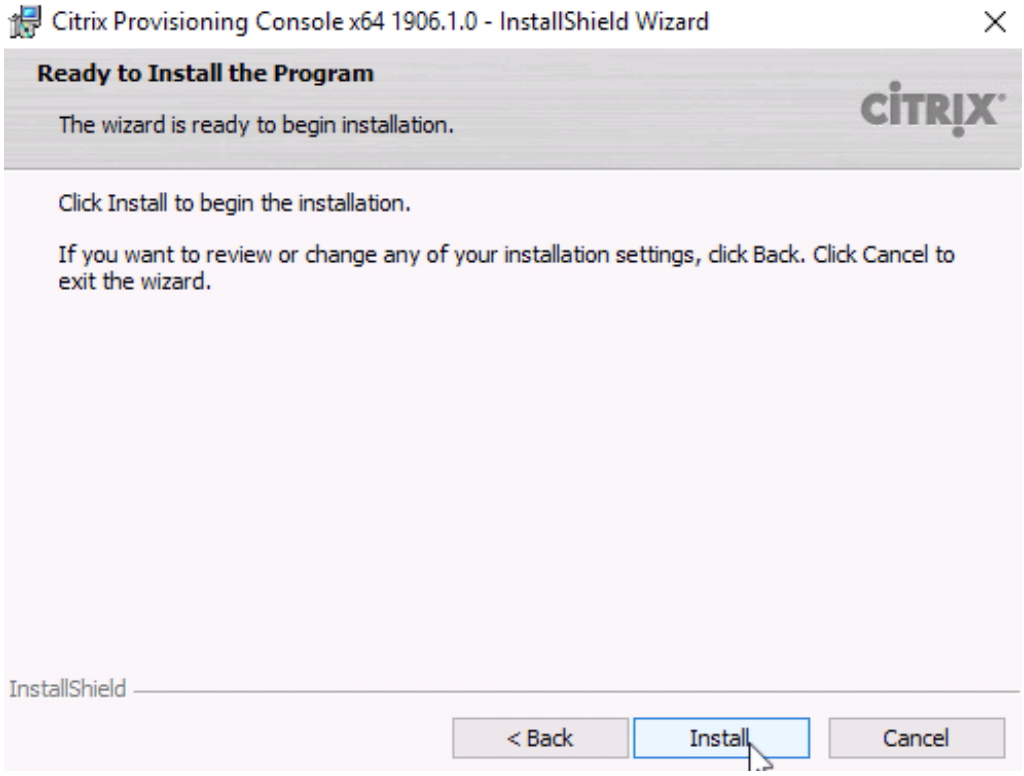
33. Enter the user and organization information. Select the option for who this application is available. Click Next.



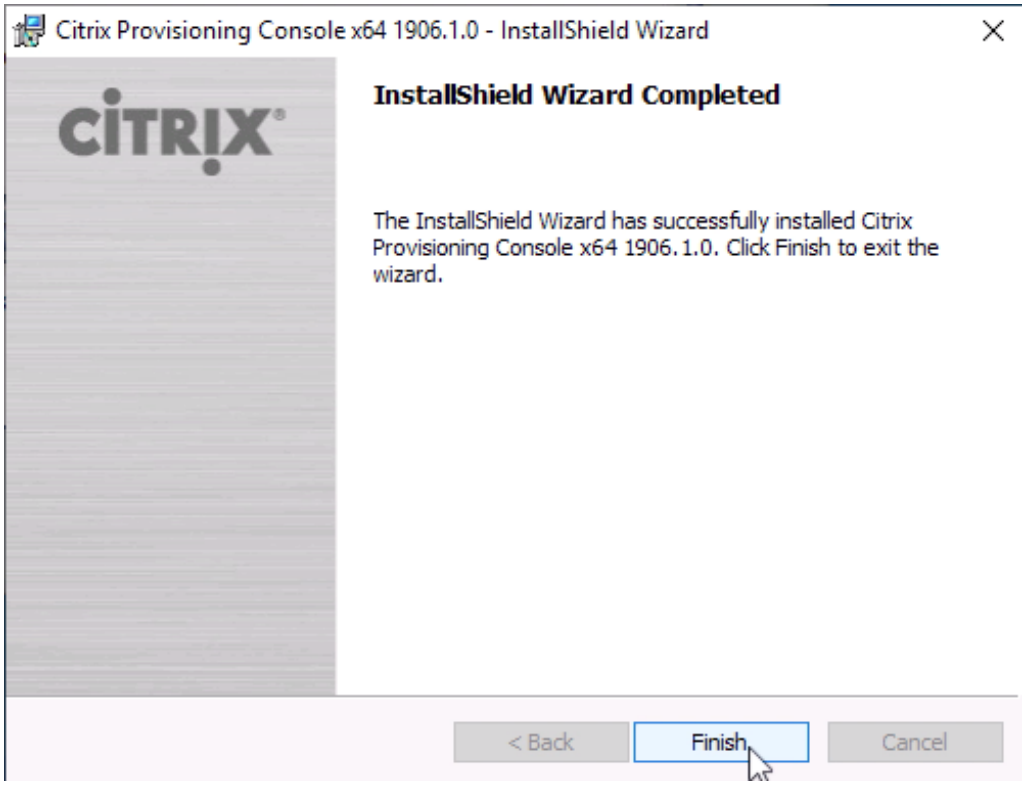
34. Verify the installation location and click Next.



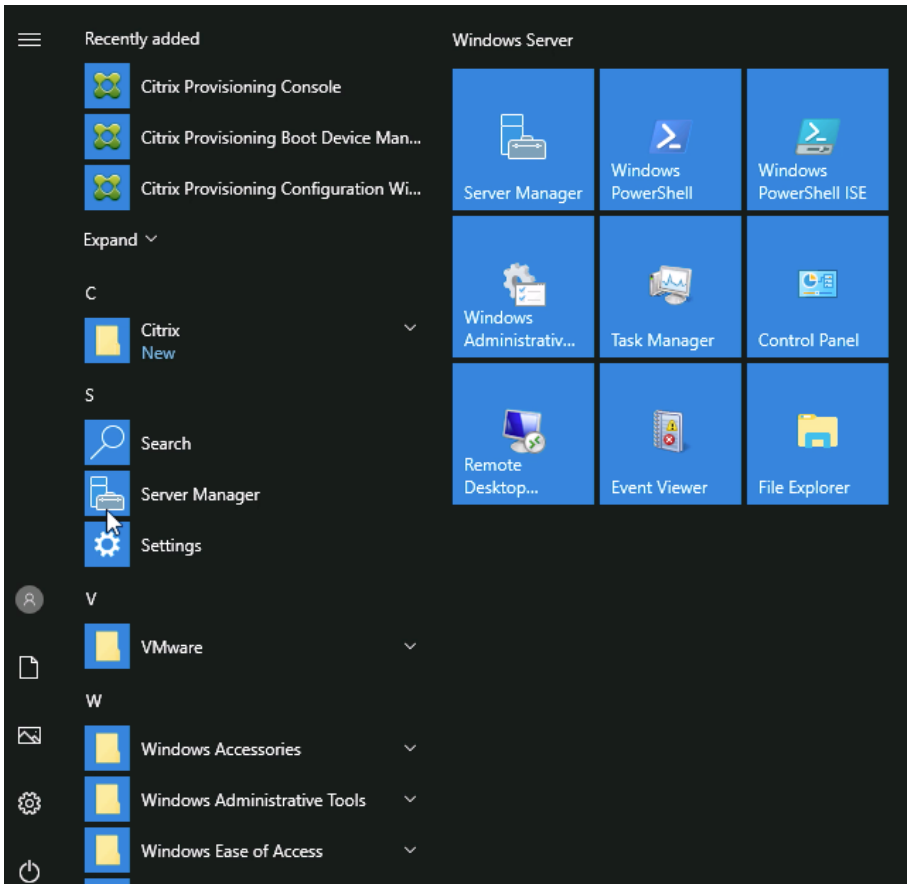
35. Click Install.



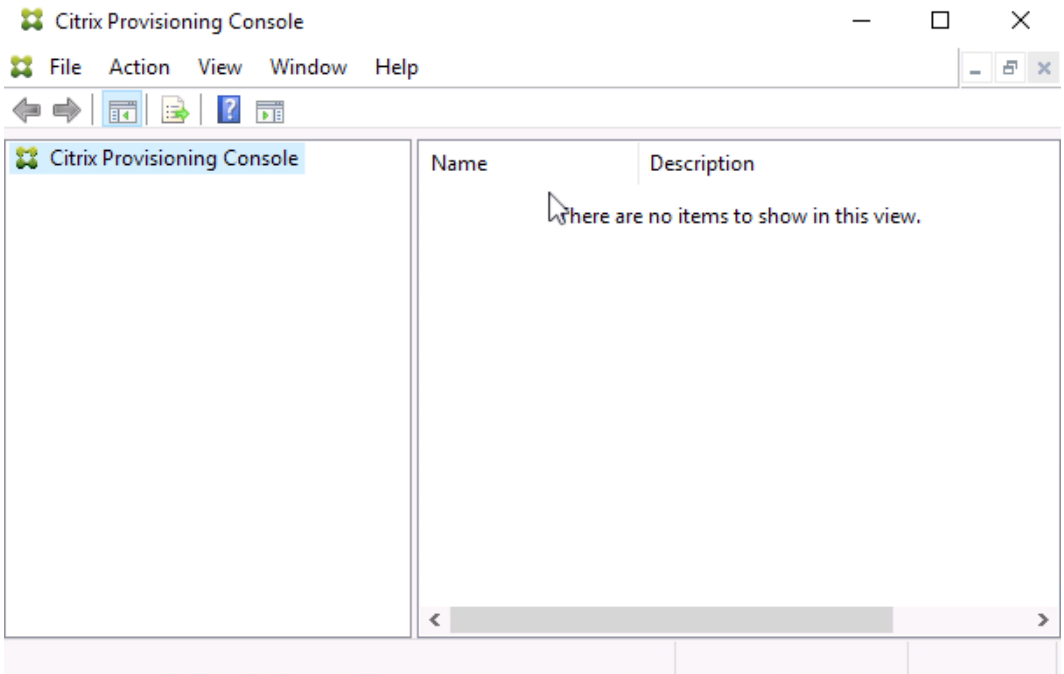
36. Click Finish.



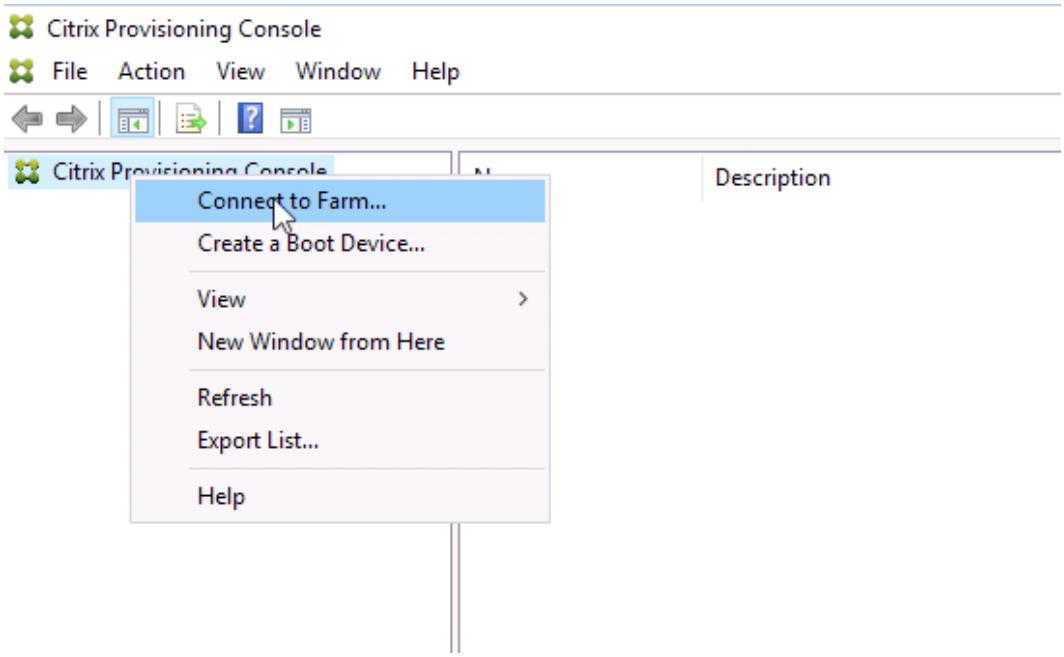
37. From the Start menu, launch PVS Console.



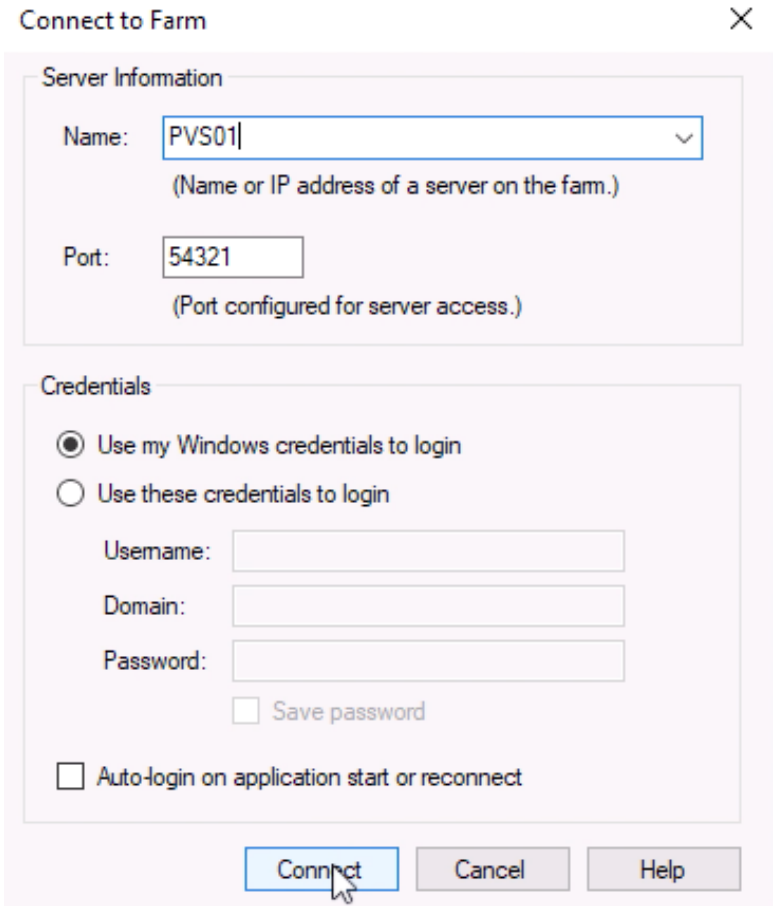
38. PVS Console MMC is displayed.



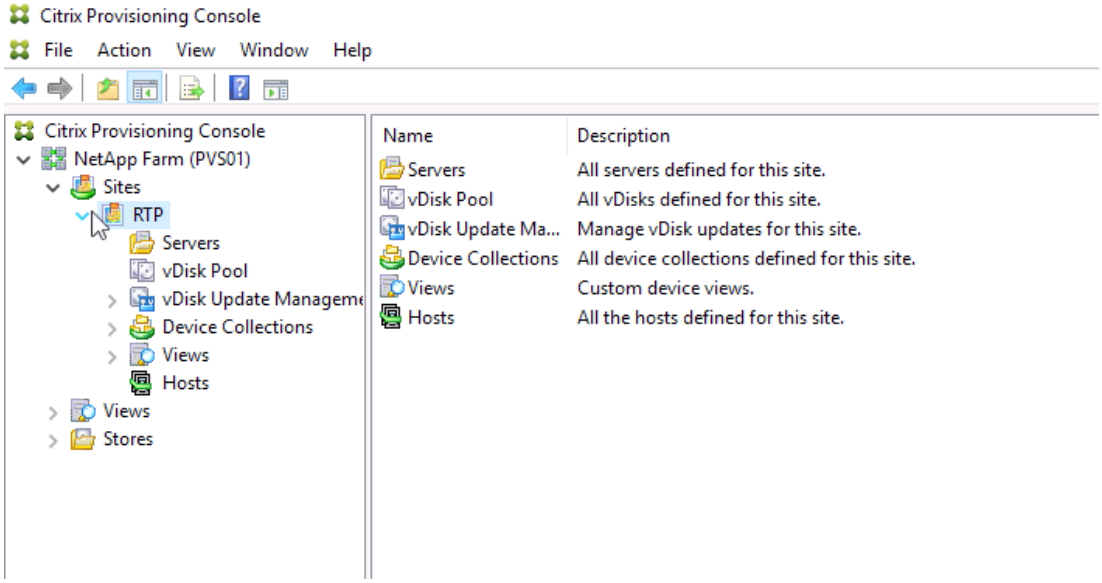
39. Right-click Citrix PVS or click the Action menu and select Connect to Farm



40. Enter the name of the server that was just installed and click Connect.



41. Expand the farm to view the sites, views, and stores.



Follow these high-level steps to provision Citrix Virtual Apps and Desktops using PVS:

1. Create the master image VM. Before you deploy VDA, deploy PVS Target Agent.
2. From the Start menu, launch the Imaging wizard.

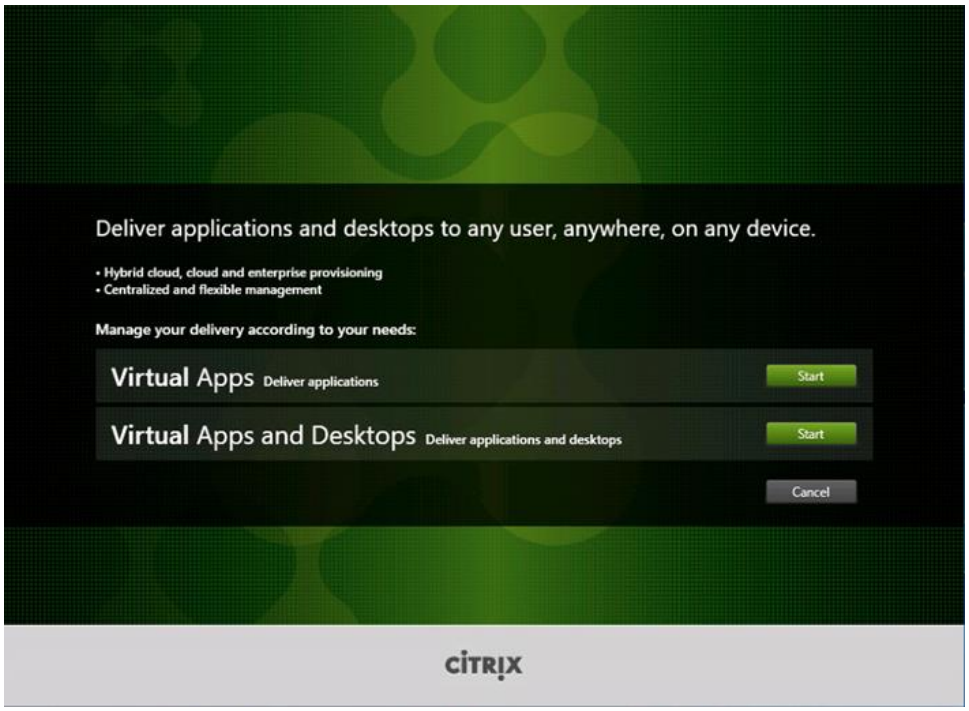
   Note: Provide a single PVS server name rather than a load-balanced IP.
3. Follow the wizard instructions. Before you click the confirmation to reboot, modify the boot option to show the BIOS setting on next boot so that it can be booted with network or BDM ISO.
4. After you log in, the imaging process resumes.
5. On the newly created vDisk, enable load balancing and change the access mode to standard image. Modify the maximum RAM size when you pick the cache type as “Cache in Device RAM with Overflow to Hard Disk”.
6. Clone the master image VM, remove disks, and convert to template. Keep the template on the cluster where you wish to deploy the target devices.
7. Right-click Site and select the Citrix Virtual Desktops Setup wizard.
8. Follow the wizard instructions to complete the creation of the machine catalog.
9. Create Delivery Group on Citrix Studio to assign Virtual Apps and Desktops to the users.

8.6 Citrix Studio

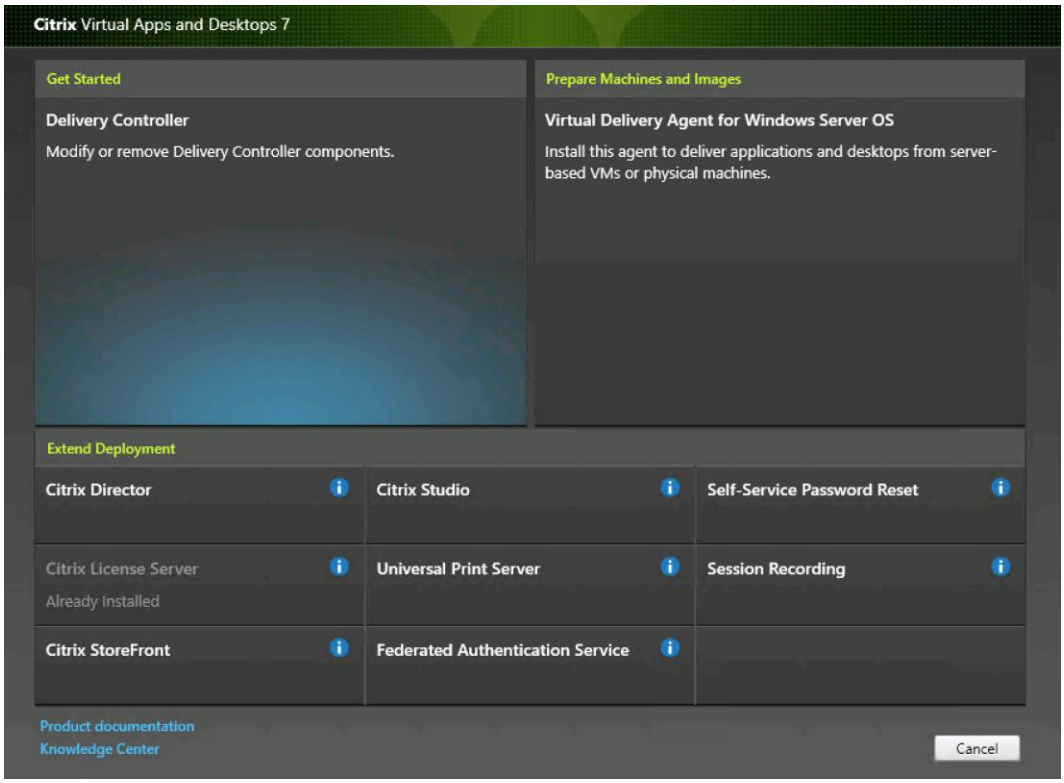
To install Citrix Studio, complete the following steps:

Note: Citrix Studio can be installed on any machine from where you like to manage the Citrix environment.

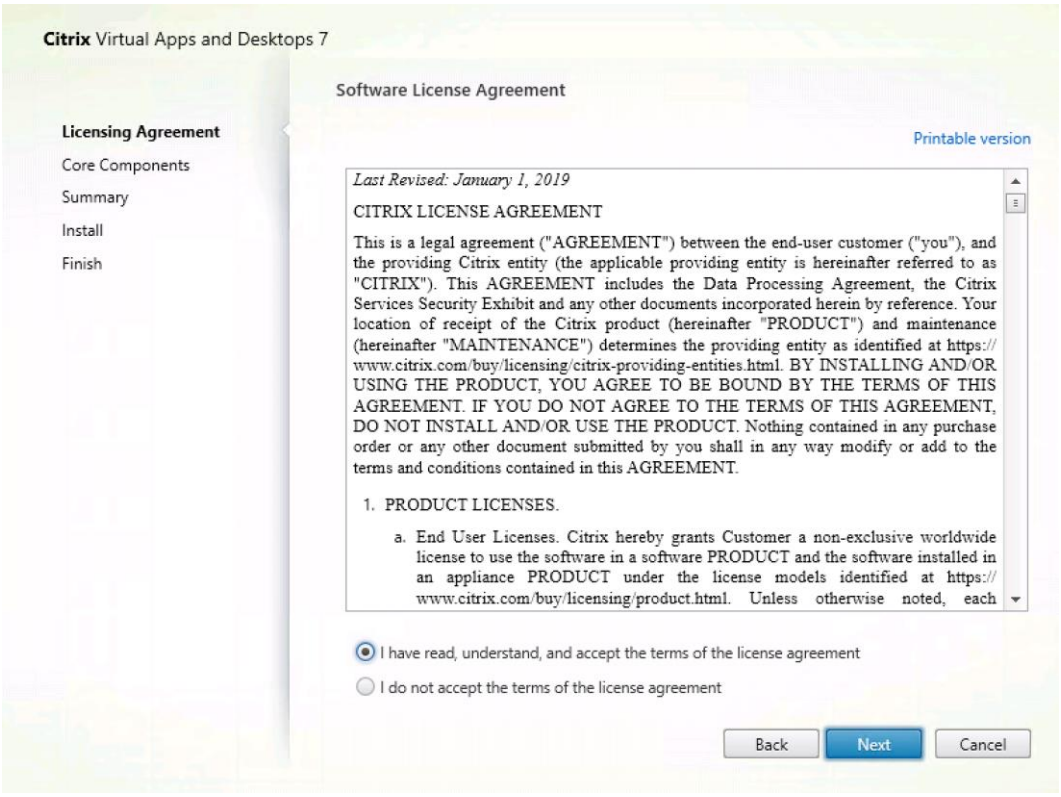
1. Mount the Citrix Virtual Apps and Desktops ISO file.
2. Run `AutoSelect.exe` at the root of the drive.
3. Select Start on the Virtual Apps and Desktops option.



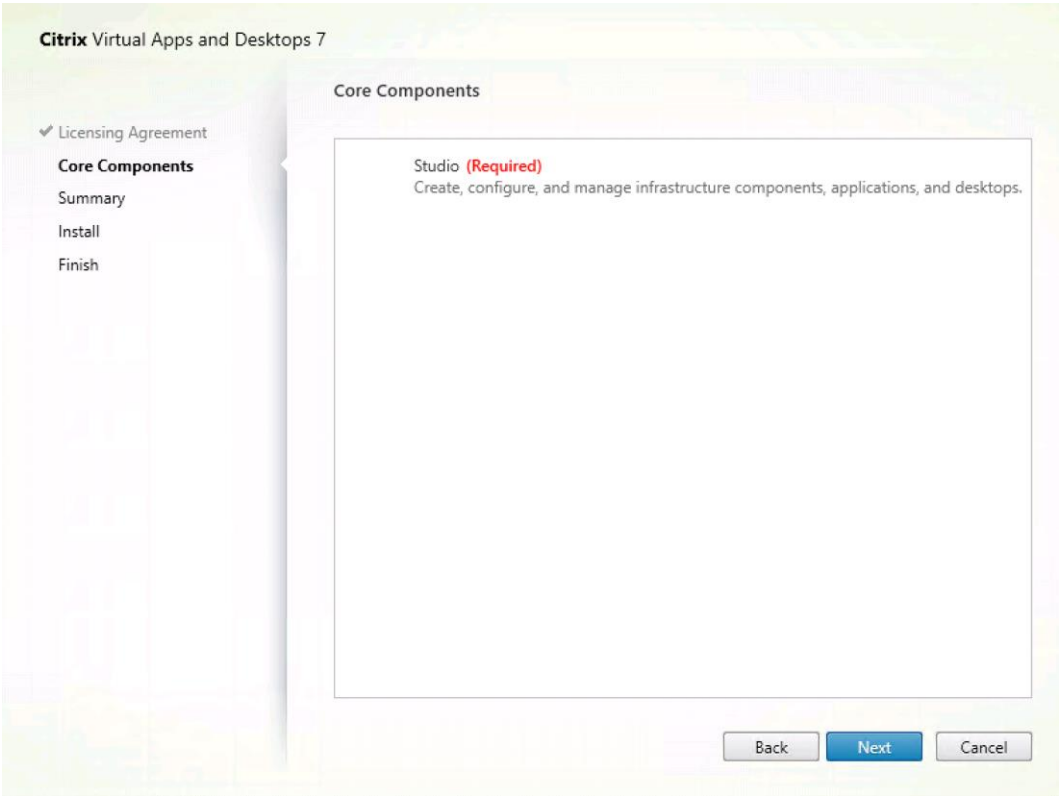
4. Click Citrix Studio.



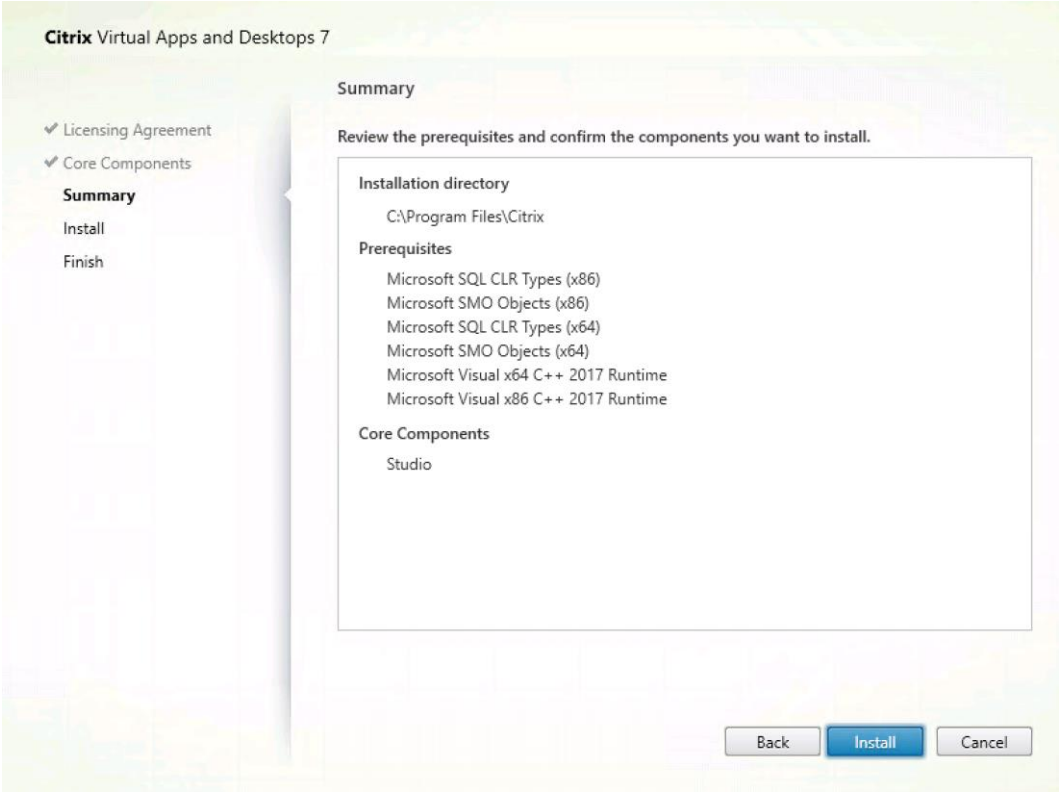
5. Review the license agreement and select appropriate option. Click Next.



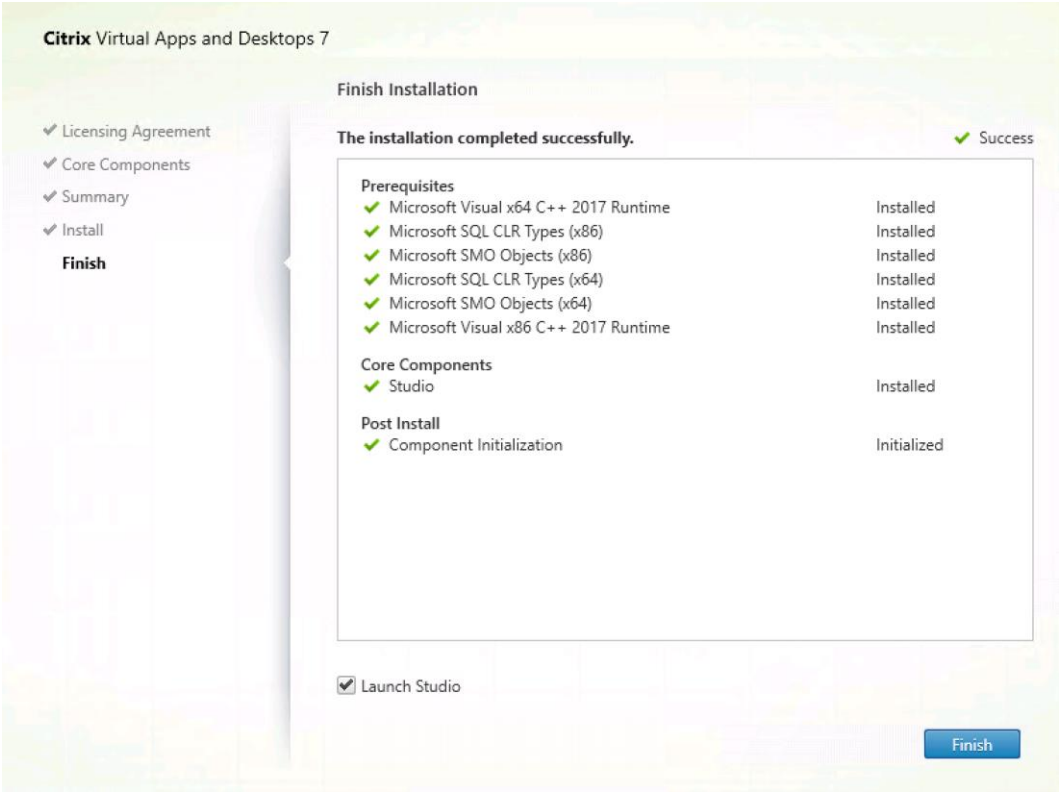
6. Click Next.



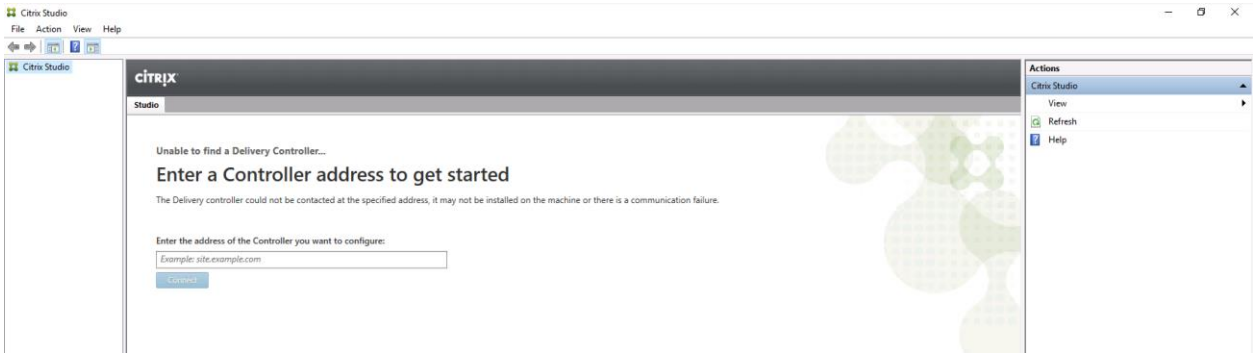
7. Review the summary and click Install.



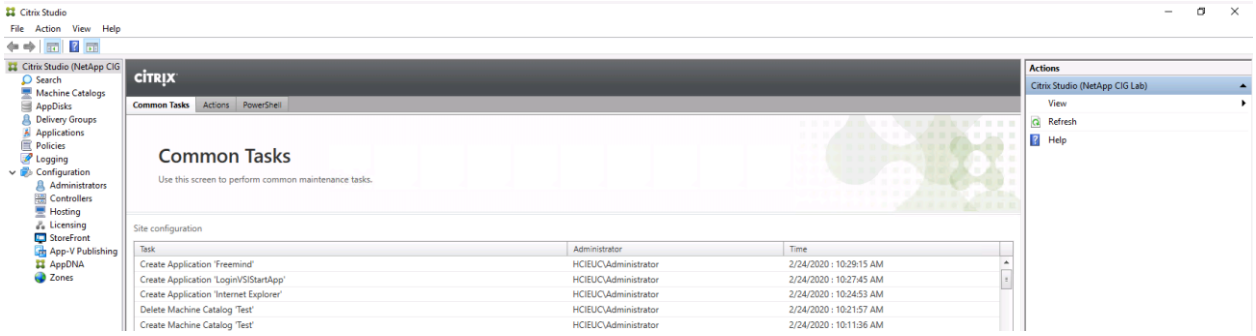
8. Click Finish.



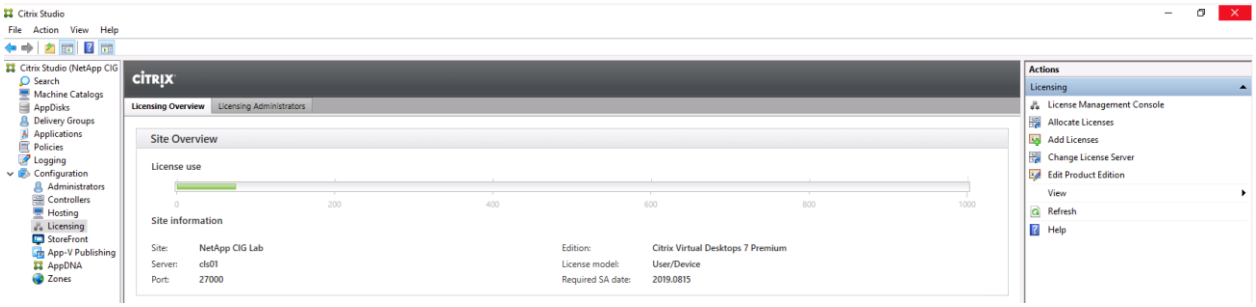
9. Citrix Studio is now launched. You must provide the Citrix Delivery Controller address. Provide the IP address or FQDN of Citrix Delivery Controller. Click Connect.



The user interface to manage the Citrix Virtual Apps and Desktops is displayed.



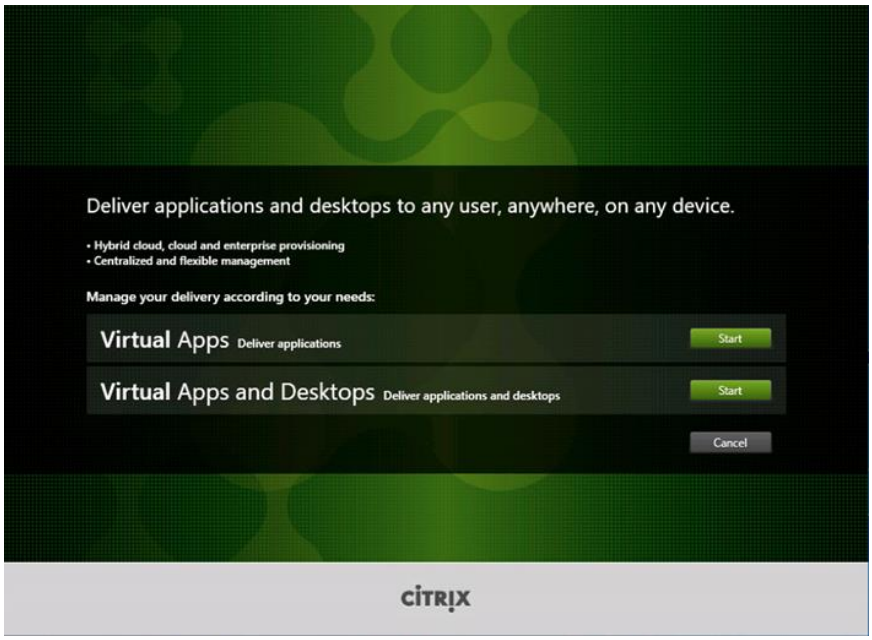
10. Select Configuration > Licensing. To add the licensing file, click Add Licenses under the Action pane.



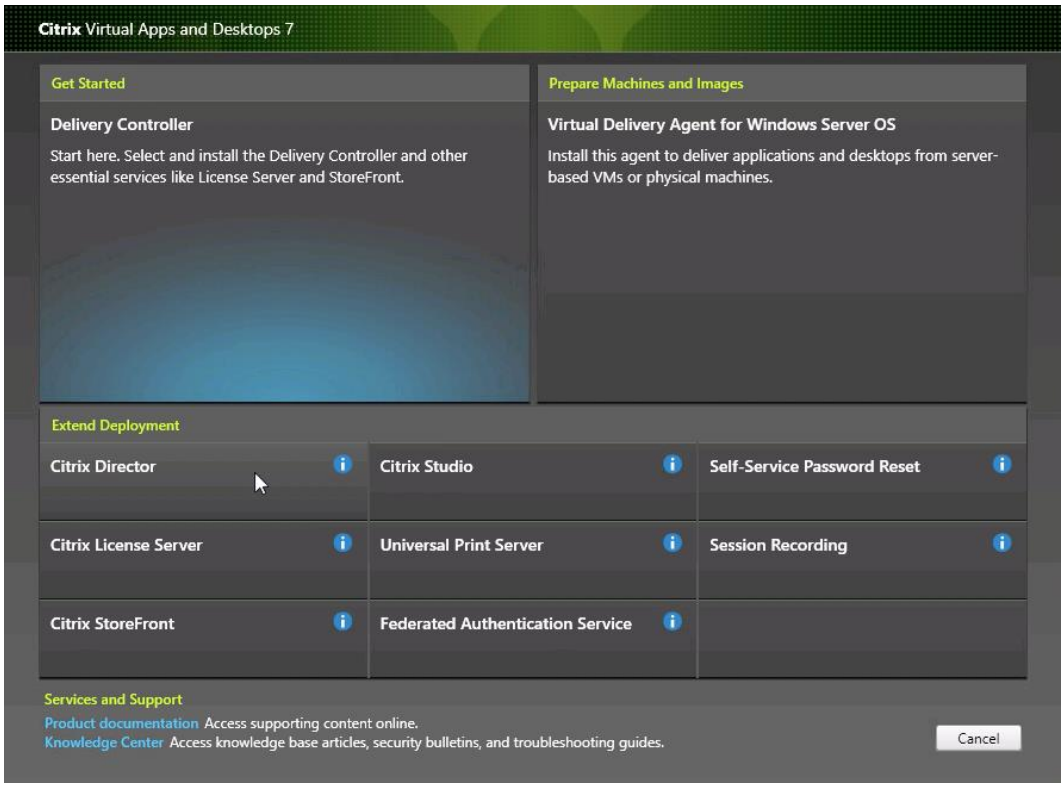
8.7 Citrix Director

To install Citrix Director, complete the following steps:

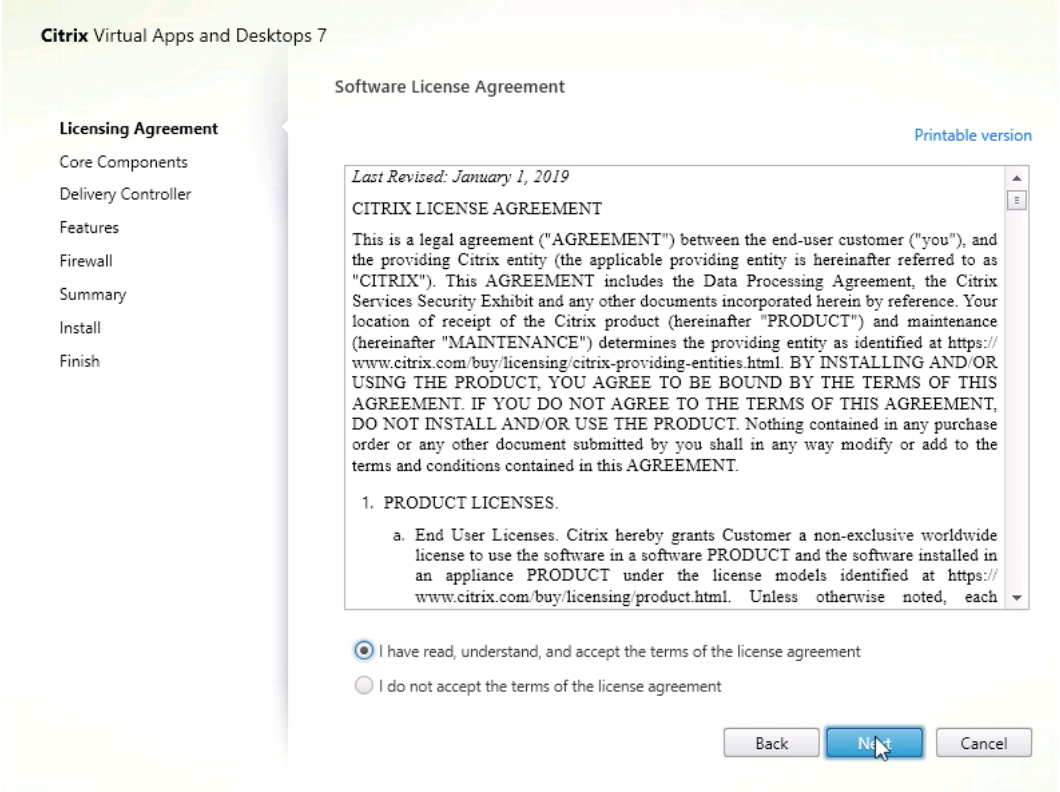
1. Deploy the supported Windows server by using ISO or from the VM template.
2. Mount the Citrix Virtual Apps and Desktops ISO file.
3. Run `AutoSelect.exe` at root of the drive.
4. Select Start in the Virtual Apps and Desktops option.



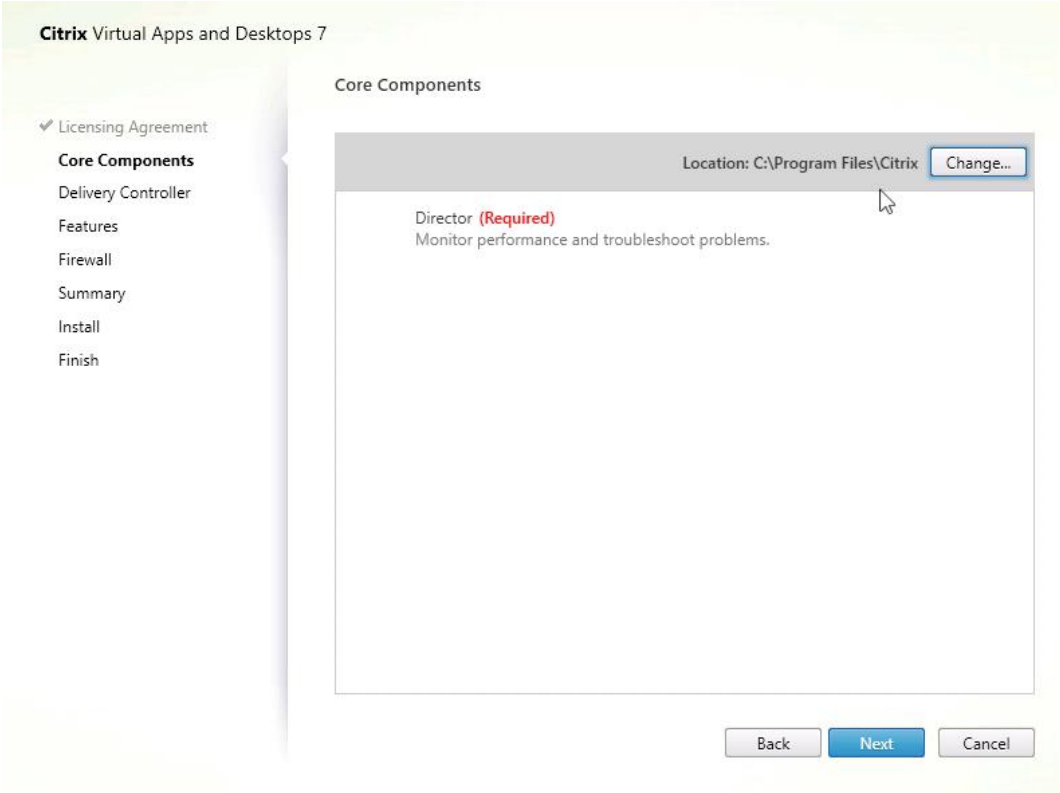
5. Click Citrix Director to launch the installation wizard.



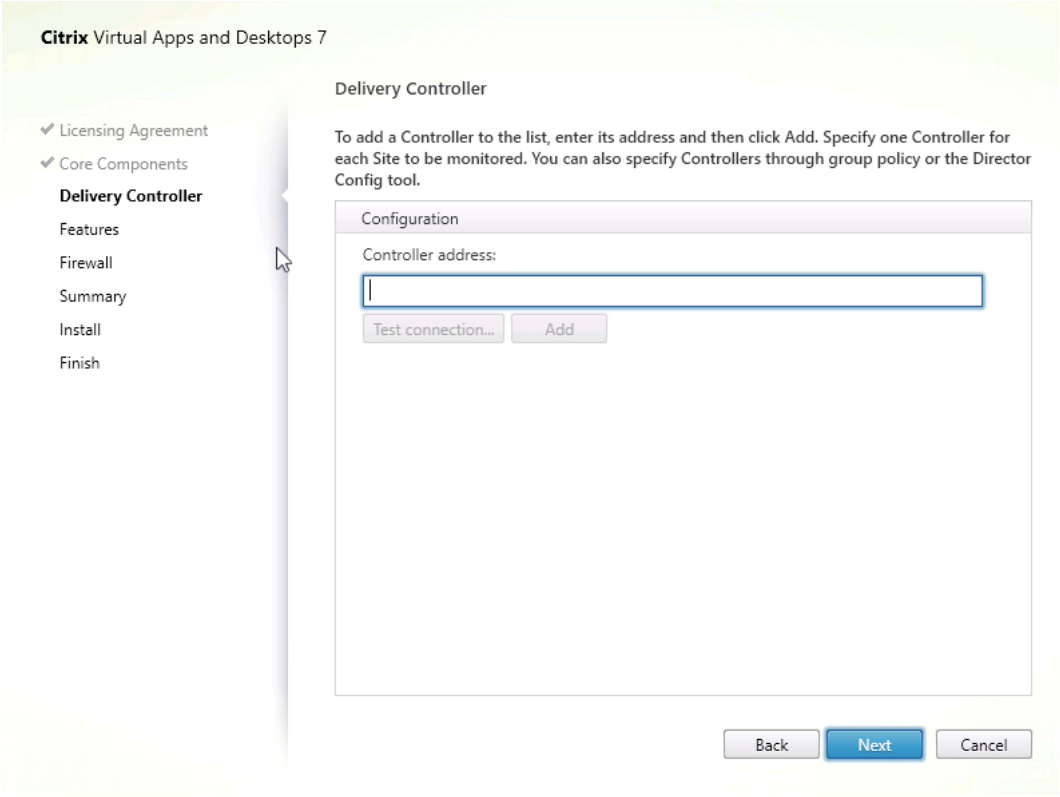
6. Review the license agreement and select the appropriate option. Click Next.



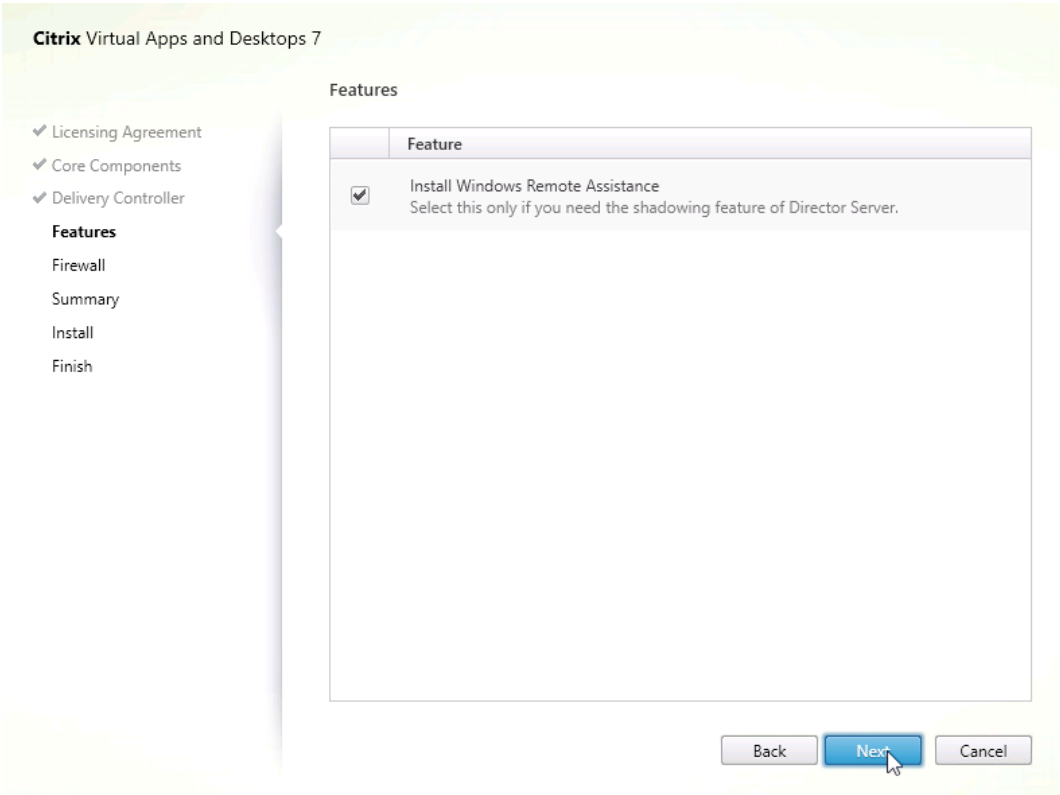
7. Verify the installation location and click Next.



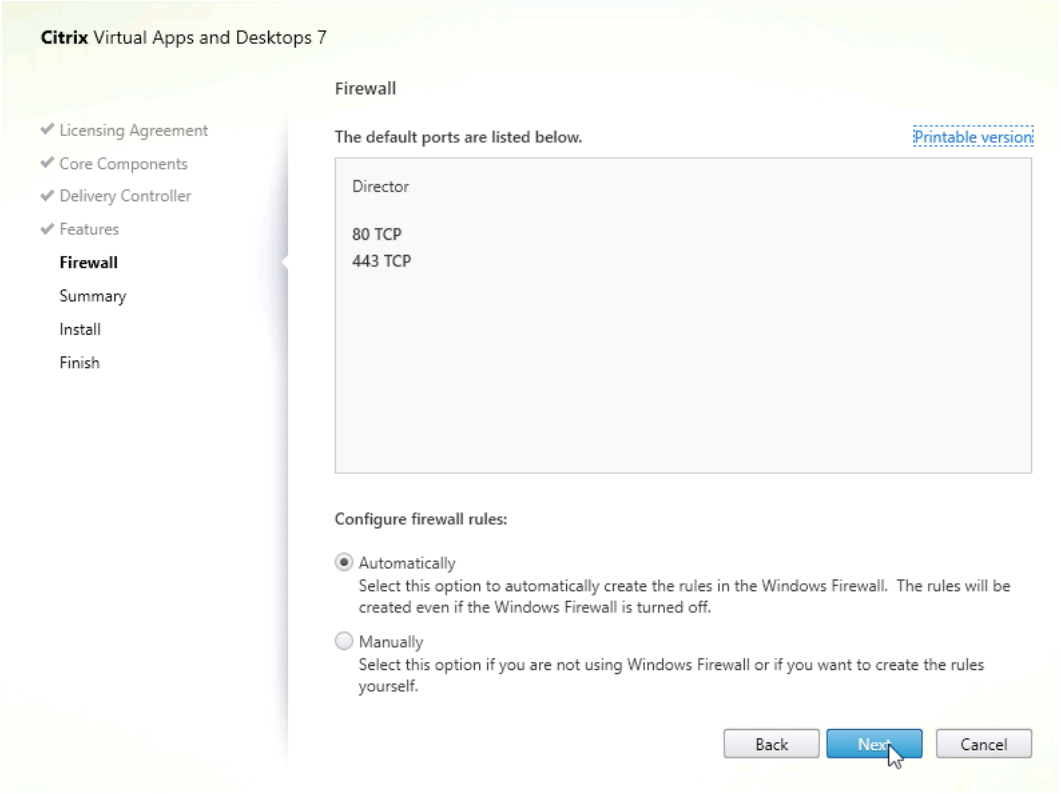
8. Add the Citrix Delivery Controller address and test the connection. Click Next.



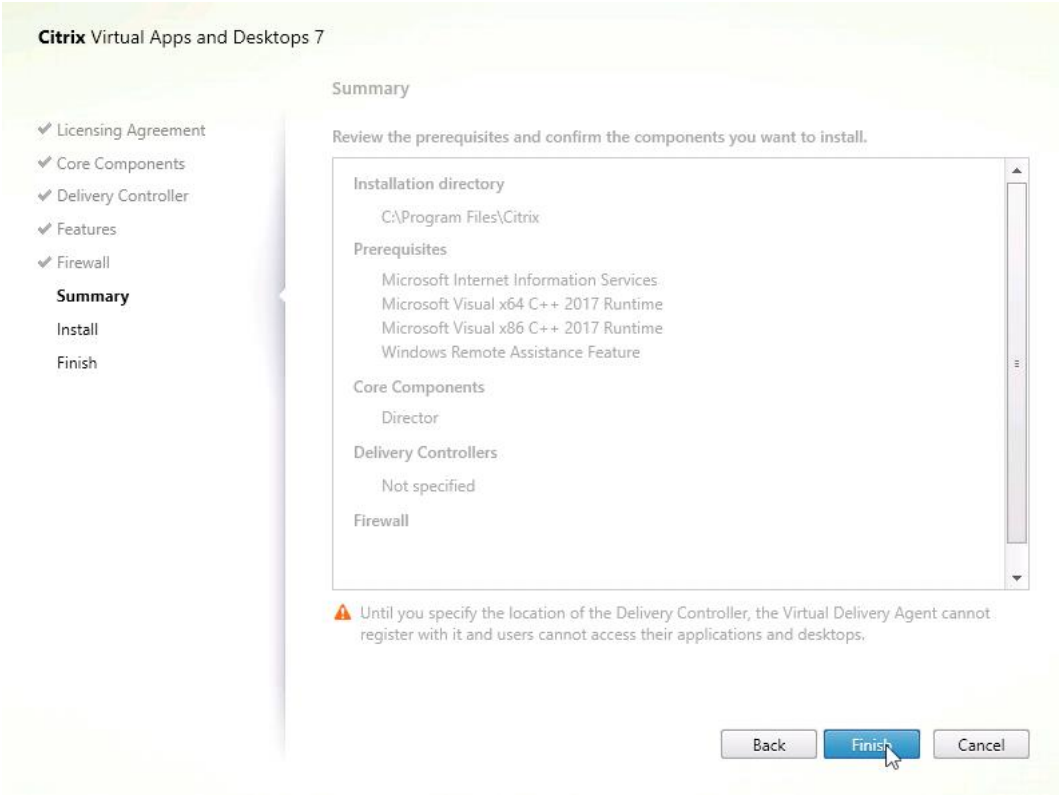
9. Review the features to be installed and click Next.



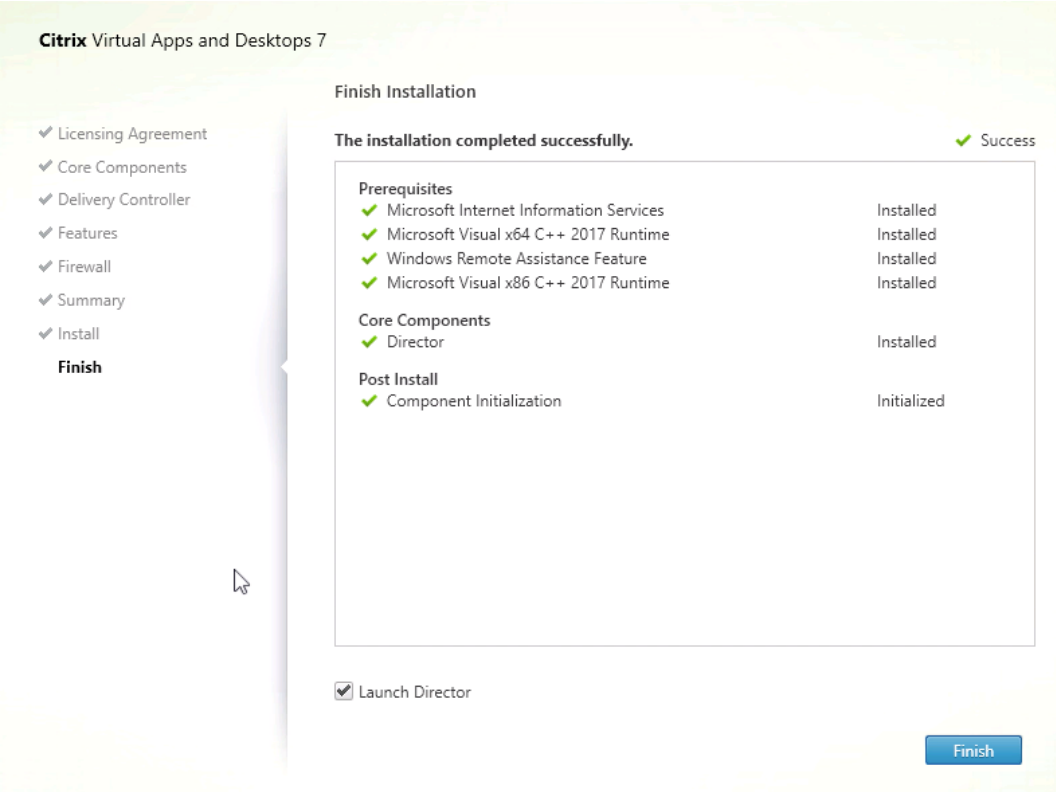
10. Review the firewall ports and select the option for how to configure the firewall rules. Click Next.



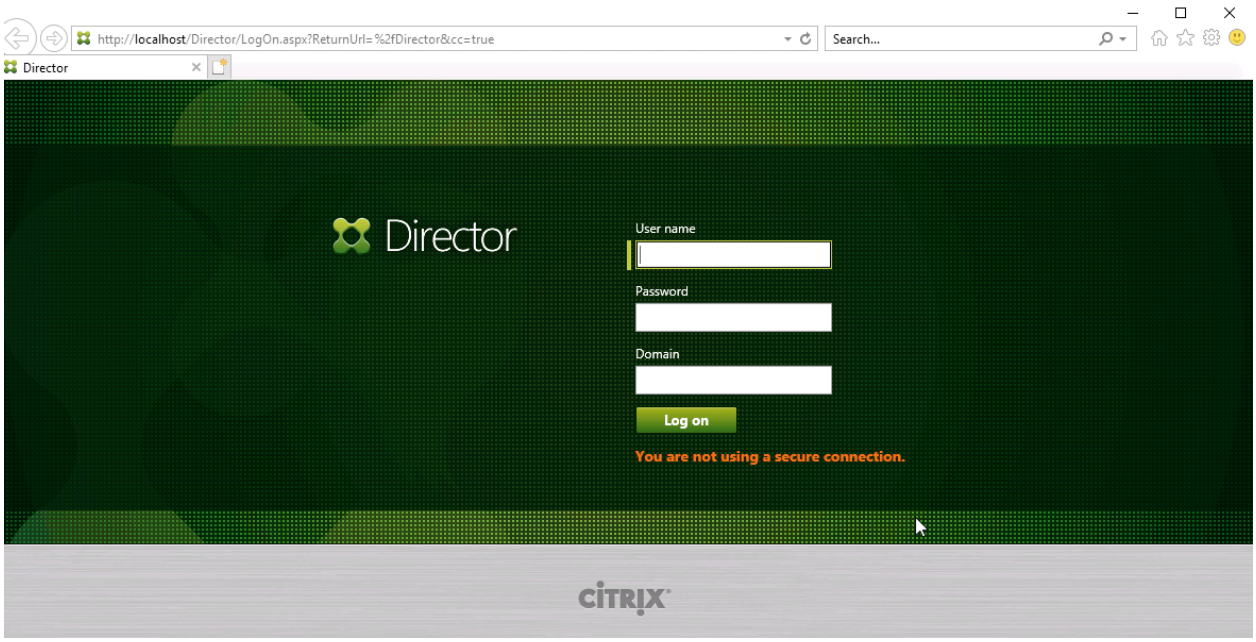
11. Review the summary and click Finish.



12. The installation progress is displayed. When it completes, click Finish.



13. Using default internet browser, Director is launched.



9 Solution Verification

In the NetApp Research Triangle Lab, we validated the following test cases:

- Single-server benchmarking testing with Login Virtual Session Indexer (Login VSI)
- 1,000 users-simulated workloads with Login VSI
- Single-server GPU benchmarking with NVIDIA nVector Knowledge worker workload
- Single-server simulated graphics benchmarking with NVIDIA nVector SPECviewperf13 workload

9.1 Login VSI Test Results

Login VSI is the industry-standard load-testing tool for testing the performance and scalability of centralized Windows desktop environments such as server-based computing (SBC) and VDI. Login VSI is used for testing and benchmarking by all major hardware and software vendors and is recommended by both leading IT analysts and the technical community. Login VSI is vendor independent and works with standardized user workloads; therefore, conclusions based on Login VSI test data are objective, verifiable, and replicable. SBC-oriented and VDI-oriented vendor organizations that are committed to enhancing end-user experience in the most efficient way use Login VSI as an objective method of testing, benchmarking, and improving the performance and scalability of their solutions. VSImax provides the proof (vendor independent, industry standard, and easy to understand) to innovative technology vendors to demonstrate the power, the scalability, and the gains of their solutions. Login VSI-based test results are published in technical white papers and presented at conferences. Login VSI is used by end-user organizations, system integrators, hosting providers, and testing companies. For more information about Login VSI or for a free test license, refer to the Login VSI website.

Login VSI simulates typical user behavior in workplace environment which includes normal operations with Microsoft Office applications, Adobe Acrobat, Internet Explorer, Photo Viewer, Notepad, 7-Zip, and so on. Humans are unpredictable and typically very random. Login VSI also adds random operations like typing random text, copying, and opening random files, adds wait times and idle times to simulate talking over phone, discussion with coworker, coffee and bathroom breaks.

9.2 Single-Server Login VSI Tests

The purpose of single-server Login VSI tests is to identify the number of virtual desktops that can be hosted on a specific NetApp HCI model while simulating an industry-standard knowledge worker workload.

Login VSI gradually increases the simulated user until the system is saturated. After the system is saturated, the response time of the applications increases, which might not be a positive experience for the end users.

Login VSI provides VSI Index, which is the average of weighted time of Notepad start, Notepad file open, zip high compression, zip low compression, and CPU time to calculate a large array of random data. VSI Index is considered the average response time of the virtual desktop.

As the workload starts, it provides the best user experience, and gradually the response time increases until the system is saturated. The baseline is calculated based on an average of the 13 lowest response times; by default, the threshold is 1,000ms. VSI Max is the session count which is one session before the session that exceeds the baseline by 1,000ms (threshold value).

Because each user is assigned dedicated virtual desktops, the session count provides the number of desktops that is in use.

For testing purposed, we used Citrix MCS to deploy virtual desktops by using the master image Windows 10 1903 edition with Microsoft Office 2016, Login VSI Target components (Adobe Acrobat, JAVA, Freemind, and so on), Citrix Virtual Apps and Desktops Agent 1909. Microsoft .Net Framework installed and fine-tuned using Base Image Script Framework along with Citrix Optimizer.

The virtual desktops are hosted on a single vSphere host and the desktops are configured with two vCPU, 4GB RAM, VMXNet3 Ethernet Card, Paravirtual SAS disks of 75GB. The user profiles are stored on SMB File share provided by ONTAP data management software.

NetApp HCI H410C

The test results for NetApp HCI H410C are shown in Figure 12 and Figure 13. The test consisted of the following configuration: Intel Xeon Gold 5120 running at 2.2GHz with two sockets and 14 cores per socket.

Figure 12) H410C - Login VSI.

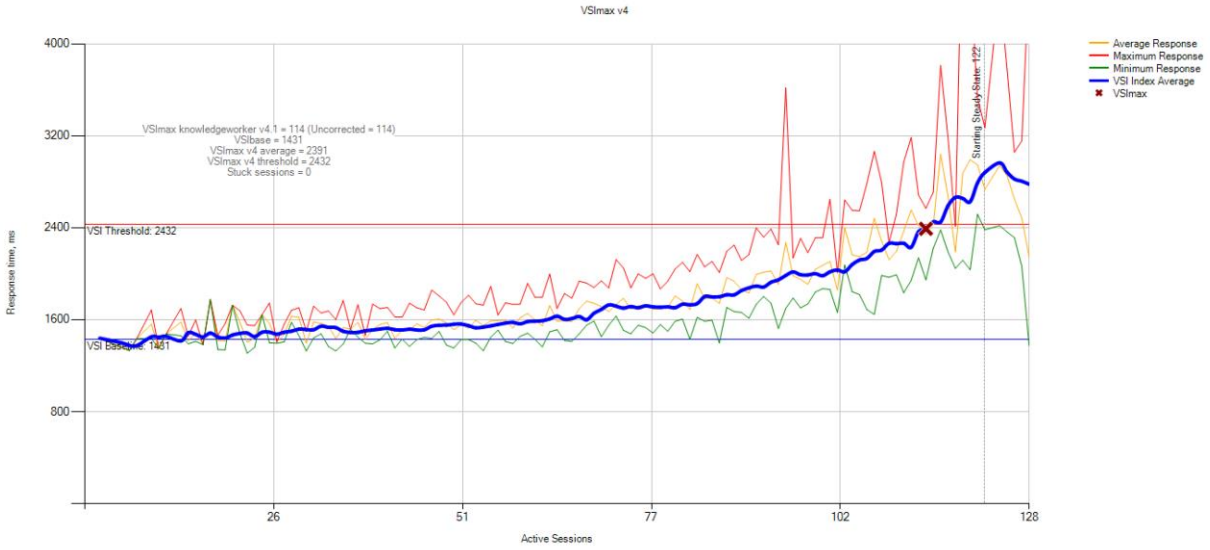


Figure 13) H410C Host CPU - Login VSI.

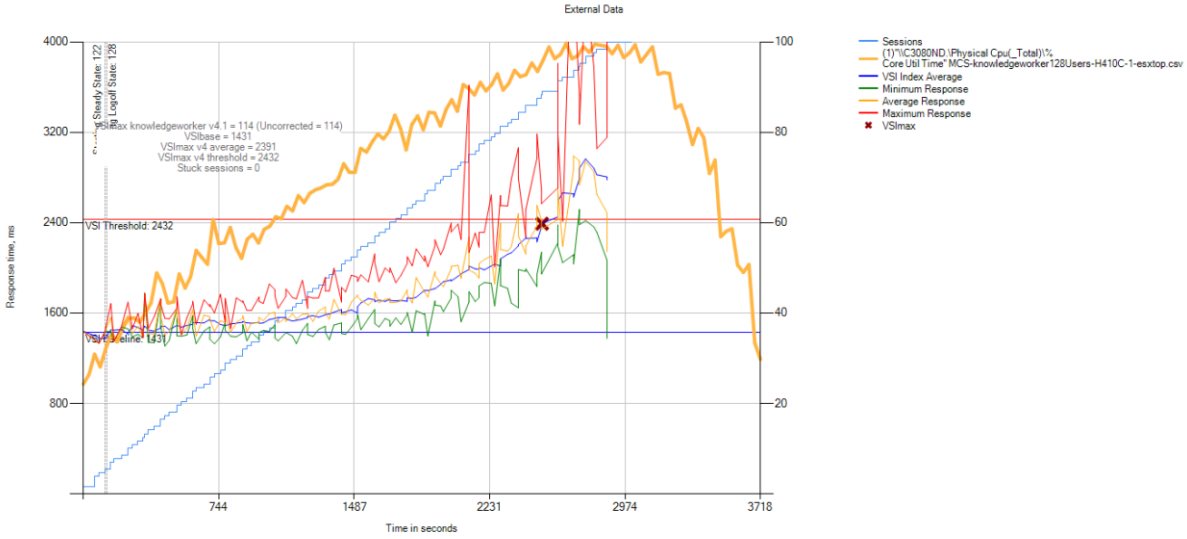


Table 9) NetApp HCI H410C test summary.

Model	Login VSI Baseline	Login VSI Max	80% of VSI Max
NetApp HCI H410C	1,431	114	91

For an improved user experience, we recommend running to a maximum of 80% load. Therefore, for H410C model, the number of virtual CPU per core is (91/(2*14)*2) 6.5 for knowledge worker workload.

NetApp HCI H610C

H610C has two NVIDIA Tesla M10 cards. The maximum number of virtual desktops using vGPU is determined by the vGPU profile used. For non-GPU knowledge worker workload, the Login VSI test results are shown in Figure 14 and Figure 15.

Figure 14) H610C - Login VSI.

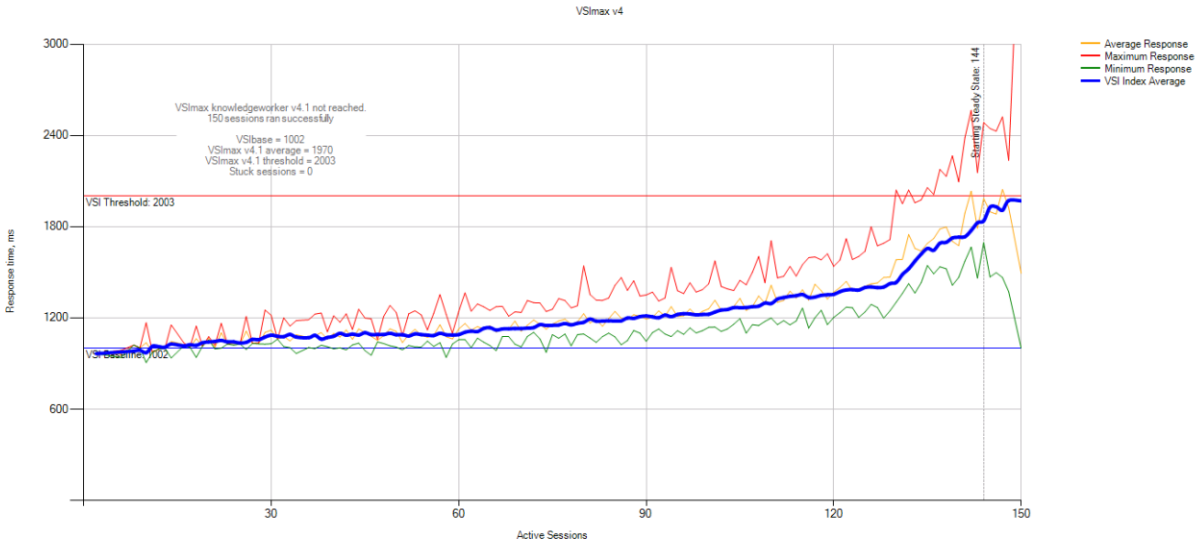


Figure 15) H610C Host CPU - Login VSI.

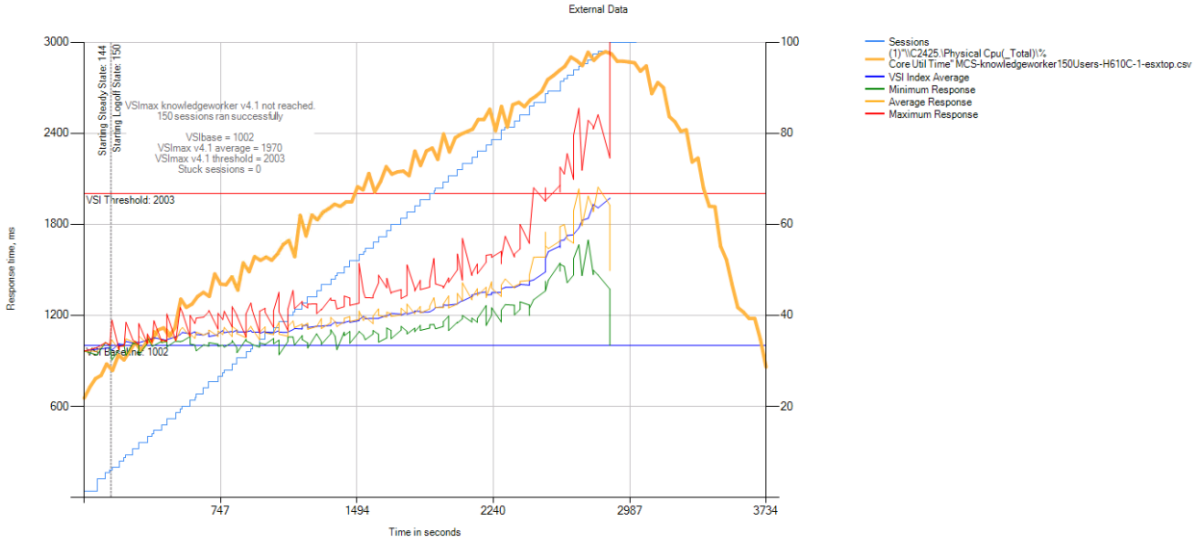


Table 10) H610C Login VSI test summary.

Model	Login VSI Baseline	Login VSI Max	80% of VSI Max
NetApp HCI H610C	1002	150	120

H610C comes with Intel Xeon Gold 6130 CPU running at 2.10GHz on two sockets with each having 16 cores. Therefore, the number of vCPU per core for H610C is (120/(2*16) * 2) 7.5 for knowledge worker workload

NetApp HCI H615C

H615C contains Intel Second Gen Scalable processors and provides various options to choose on the number of cores and CPU speed. One model contains three NVIDIA T4 cards. The maximum number of virtual desktops using vGPU is determined by the vGPU profile type used. The Login VSI test results shown in Figure 16 and Figure 17 are executed on the server with Intel Xeon Gold 6252 with 2.1GHz on two sockets each with 24 cores

Figure 16) H615C - Login VSI.

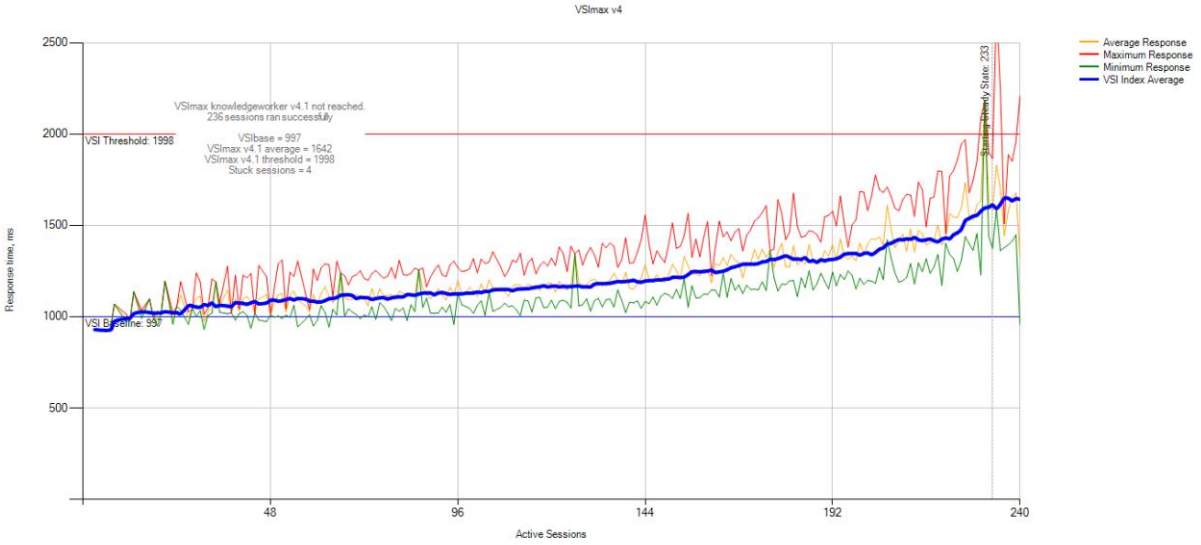


Figure 17) H615C Host CPU - Login VSI.

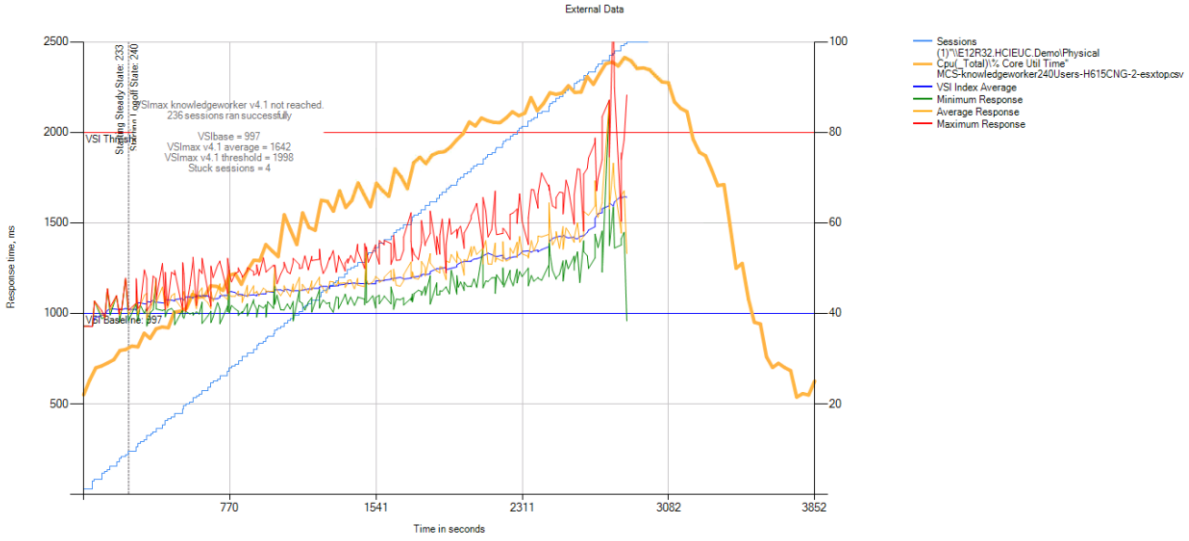


Table 11) H615C Login VSI test summary.

Model	Login VSI Baseline	Login VSI Max	80% of VSI Max
NetApp HCI H615C	997	240	192

The number of vCPUs per core for this model of H615C is (192/(2*24) *2) eight for the knowledge worker workload.

In summary, the single-server Login VSI tests for various NetApp HCI model is provided in Table 12.

Table 12) Login VSI test results summary.

Test Case	Item	H410C	H610C	H610C - GPU	H615C - NonGPU
Windows 1903 Static Personal Virtual Desktops (MCS)	Login VSI Baseline	1431	1002	1085	997
	Login VSI Max	114	150	64*	240
Windows 2019 RDSH Shared Desktops (MCS)	Login VSI Baseline	897	823	860	782
	Login VSI Max	140	134	119	240
Windows 1903 Static Virtual Desktops (PVS)	Login VSI Baseline	1085	1045	1022	957
	Login VSI Max	100	112	64*	203
Windows 2019 RDSH Shared Desktops (PVS)	Login VSI Baseline	829	782	813	779
	Login VSI Max	138	164	154	220
* Limited by the vGPU profile used.					

Note: For RDSH server VMs, we hosted eight VMs per server. Each VM had vCPU equal to NUMA boundary. It was configured to use MCS IO or PVS cache. PVS Static Virtual Desktops use case also used cache.

9.3 NVIDIA nVector Test Results

Single-Server nVector Knowledge Worker Tests

NVIDIA nVector is a benchmarking tool that is designed to provide insight to the quality of the virtual desktop user experience. The unique NVIDIA approach focuses on the experience of the user on the endpoint device rather than the response time on the virtual desktop. This focus on the end user's perspective allows administrators to better understand how fast and smooth the experience feels for the user, as well as image quality.

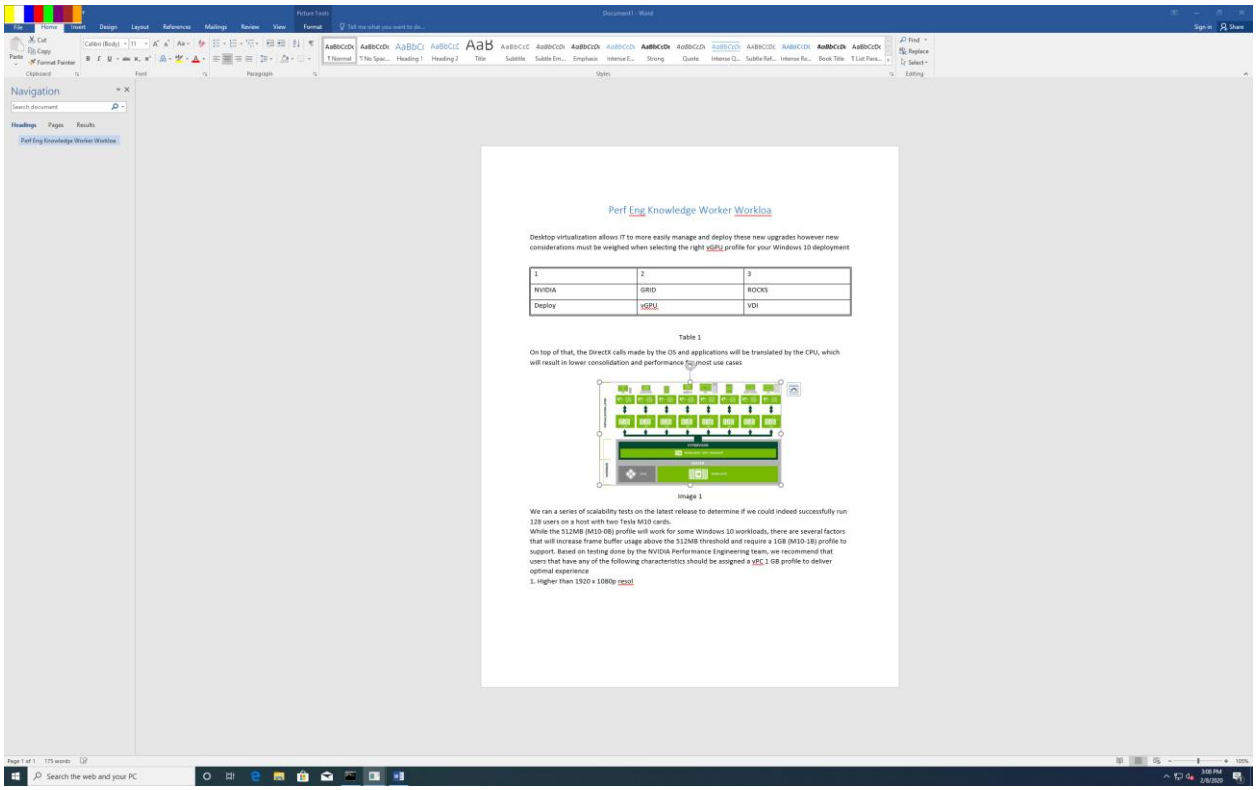
This tool measures the quality of the user experience across three specific metrics:

- **End-user latency** indicates how remote the session feels or how interactive the session is (how fast).
- **Frame rate** is the number of frames that are sent to the end user (how smooth).
- **Image quality** is how much the image was impacted and manipulated by the remote protocol.

End-User Latency

During the end-user latency test, a sequence of colors will be noticed in corner of the screen, as shown in Figure 18.

Figure 18) nVector test run.



The color palette is used to measure the end-user latency. The tool reads the color palette from the target as well as the client machine to determine the time drift to produce the same colors.

Frame Rate

The frame rate is the number of frames delivered to the client. The smoothness of a video is provided by higher frame rate. This metric is available from the remoting protocol counters and 90th percentile is considered.

Image Quality

Imagine the difference between watching a video in 480p versus 4k. This metric measures the impact of remoting on the image quality. The quality is determined by using Structural Similarity Index (SSIM), which quantifies image quality degradation by measuring the difference between two images. NVIDIA nVector applies this method by comparing a frame rendered on the target VM with the frame shown to the end user.

The NVIDIA nVector tool uses vGPU both at the client machine as well as on the target virtual desktops. The tool provisions virtual desktops with defined resources, power management, enforcing policies, launching clients, monitoring and gathering metrics, and simulating workloads, including support for multiple monitors.

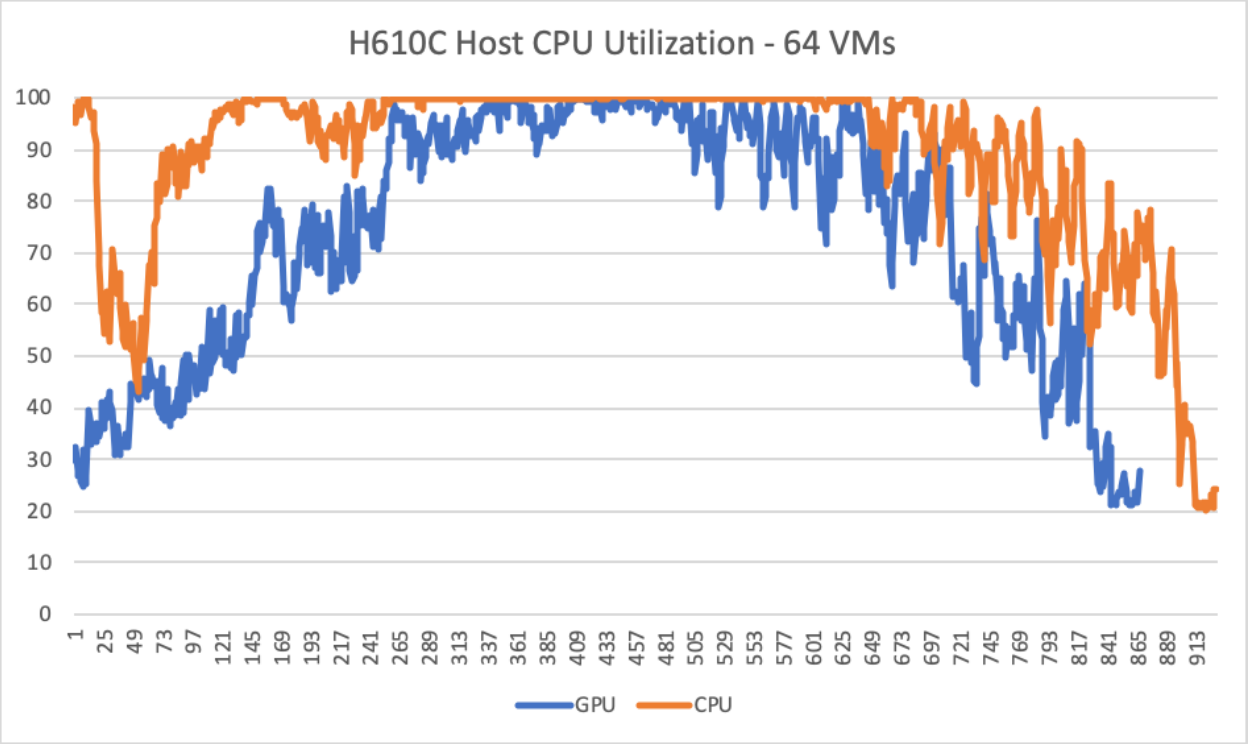
Many users are moving toward the 4K monitors to gain a better user experience; therefore, we validated single-monitor use cases with a 1B vGPU profile and two monitors with a 2B vGPU profile on NetApp HCI H610C and H615C. We reran the tests with the same master VM with no vGPU profiles attached and on the same vSphere host using CPU for the same knowledge worker workload with Citrix Virtual Apps and Desktops. For a sample json file for the test run, see Appendix A: NVIDIA nVector Tool Configuration Information.

Single 4K Monitor with NetApp HCI H610C

The minimum vGPU profile required to run a single monitor with 4K resolution is profiles ending with 1B.

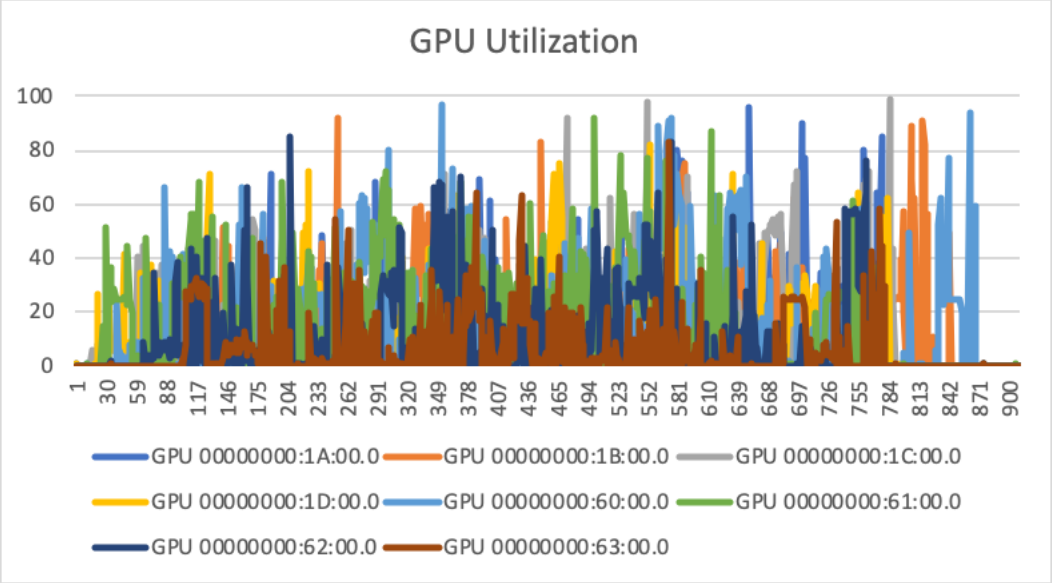
With two M10 cards on NetApp HCI H610C, the maximum number of virtual desktops that can have M10-1B vGPU profile is 64.

Figure 19) H610C vSphere host CPU utilization.



The average host CPU utilization is around 74%, with GPU assisted and 89% with CPU alone. When all the sessions are active, the host CPU utilization reaches its peak. Therefore, we recommend using m10-2b profile for a 4K resolution or the minimize number of m10-1b profile VMs per host. If you need to run 64 m10-1b VMs per host, consider reducing the resolution or validate for your workload before going to production.

Figure 20) Single Monitor H610C GPU utilization.



The average GPU utilization was around 60% and all the GPUs are used during the GPU test.

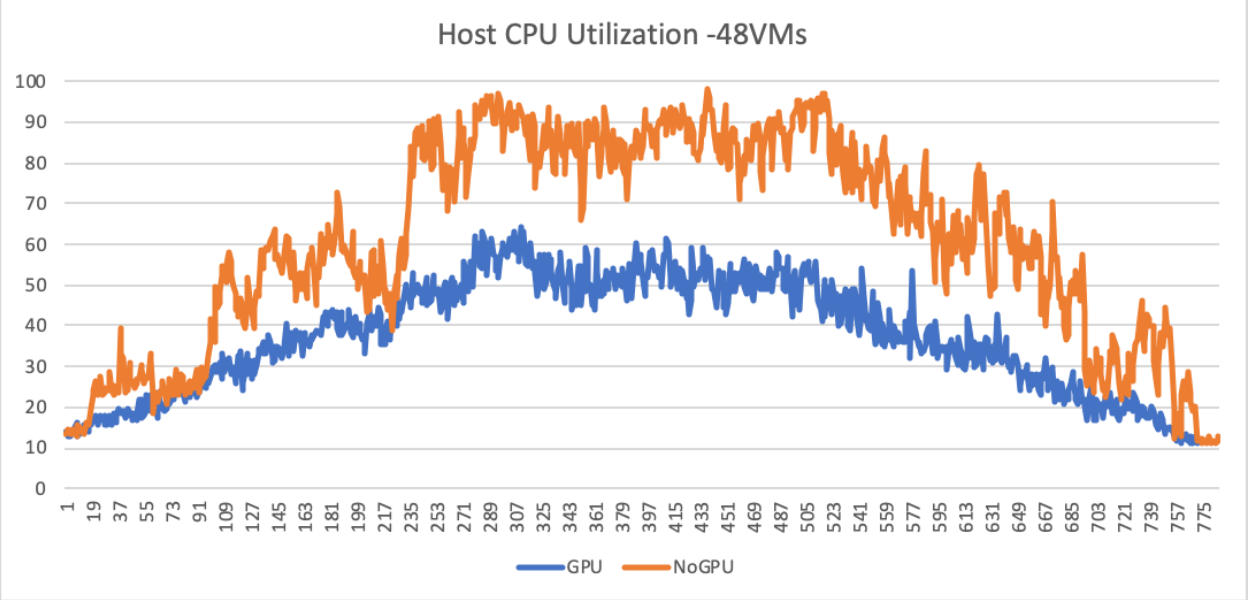
Single 4K Monitor with NetApp HCI H615C

The NetApp HCI H615C is running with Intel Second Generation Scalable Processor and has three Tesla T4 cards. A total of 48 VMs can be hosted on H615C while running with T4-1B vGPU profile at 4K resolution.

With the current vGPU driver used for this testing, the vGPU profile that supports two virtual heads is T4-2B. Using that vGPU profile, 24 virtual machines can be hosted per H615C server.

The host CPU utilization during the knowledge worker test with (compared to a VM without vGPU profile) is shown in Figure 21.

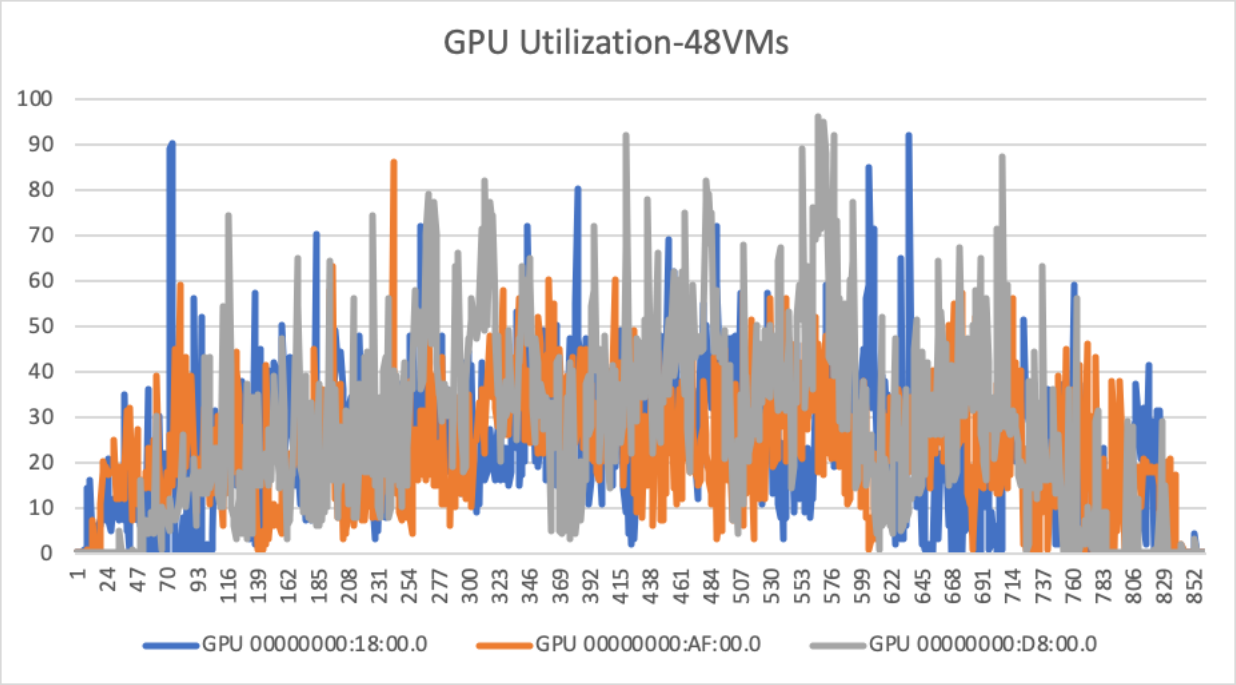
Figure 21) H615C Host CPU utilization.



The average host CPU utilization with vGPU is approximately 38%; without GPU utilization, it's approximately 61%. That's an approximately 61% savings, which can be used for hosting other non-GPU VMs.

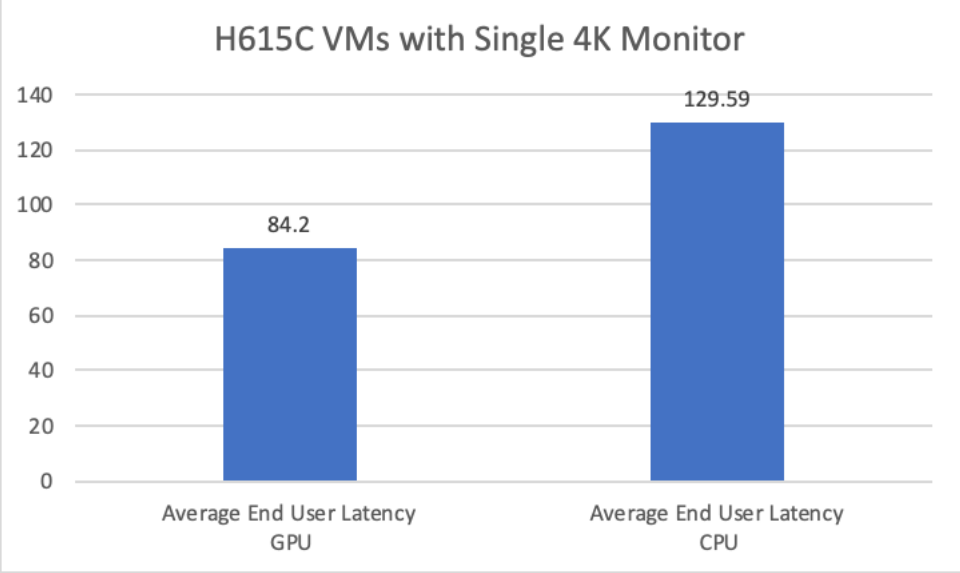
All the three GPUs are equally consumed and the average GPU utilization is approximately 45%.

Figure 22) H615C GPU utilization.



Compared to CPU, the latency perceived by the end user is approximately 54% lower with GPU.

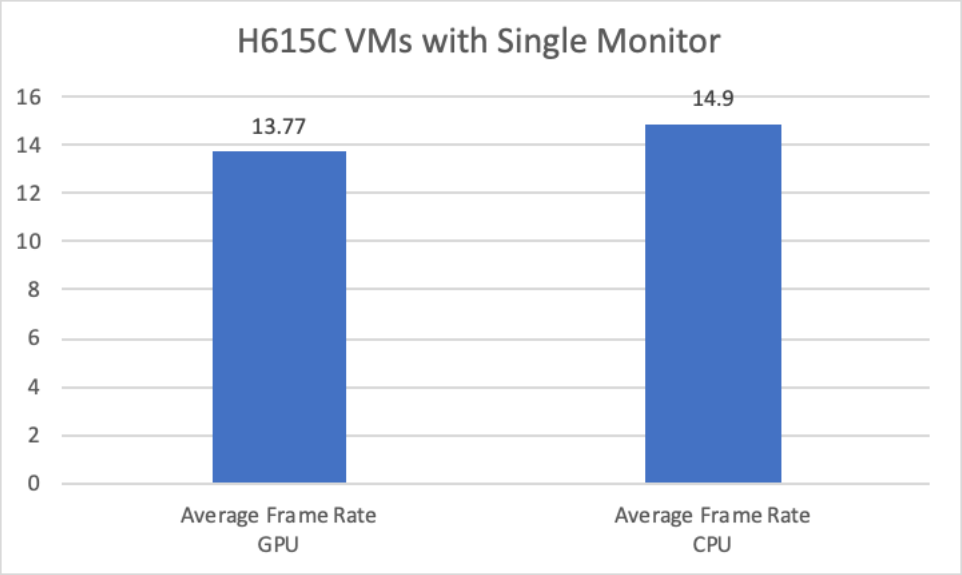
Figure 23) H615C end-user latency.



The number of frames per second is larger with CPU. There might be an opportunity to fine-tune vGPU.

Note: This test is out of the box; no modifications were made.

Figure 24) H615C average frame rate.



The image quality was identical in both cases and the value was 0.99.

In summary, H615C with vGPU delivered the following results:

Note: These results are compared to using CPU alone.

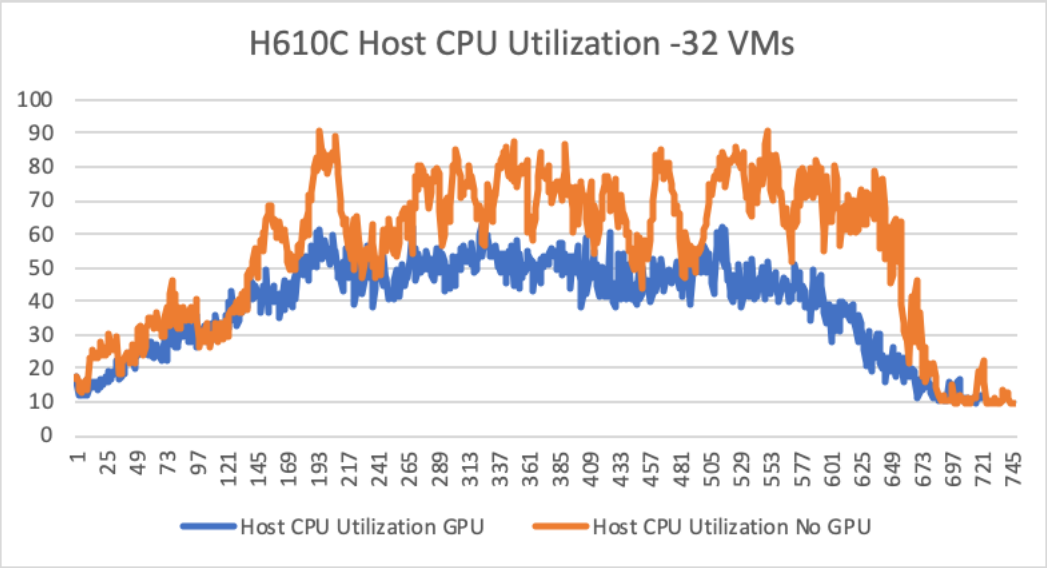
- 61% less host CPU utilization
- 54% lower average end-user latency
- The average frame rate and image quality were almost equal

Two 4K Monitors with NetApp HCI H610C

Many users prefer to run with two monitors, therefore, we decided to test this configuration. With H610C, the M10-2B vGPU profile supported two virtual monitors with 4K resolution. With this profile, we can have 32 VMs per host.

Note: The vSphere host core utilization is compared with VMs running with and without GPU.

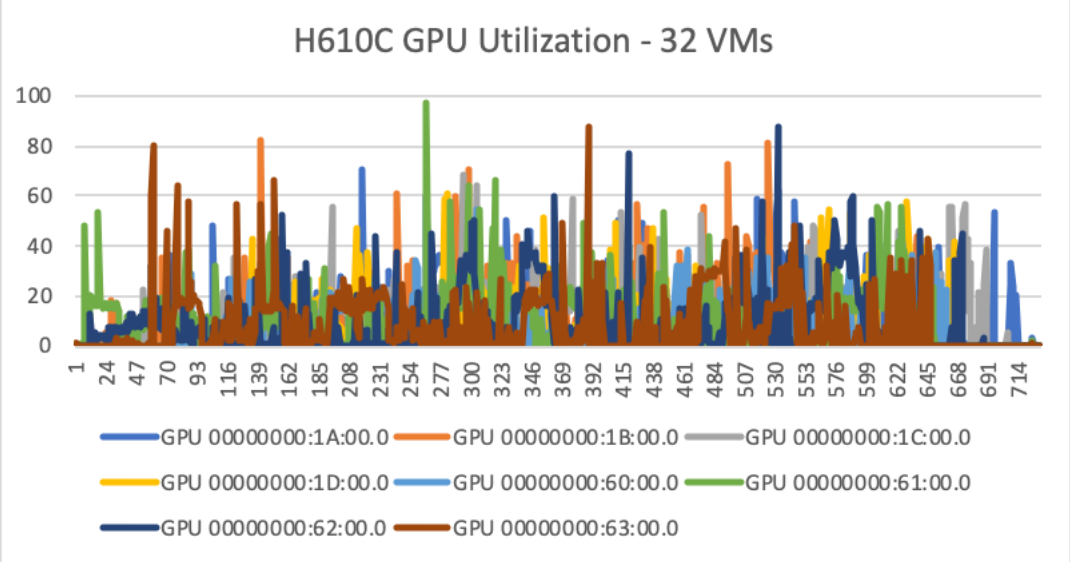
Figure 25) H610C host CPU utilization.



The average host CPU utilization was approximately 39% with GPU and 55% without GPU. GPU provided approximately 40% savings.

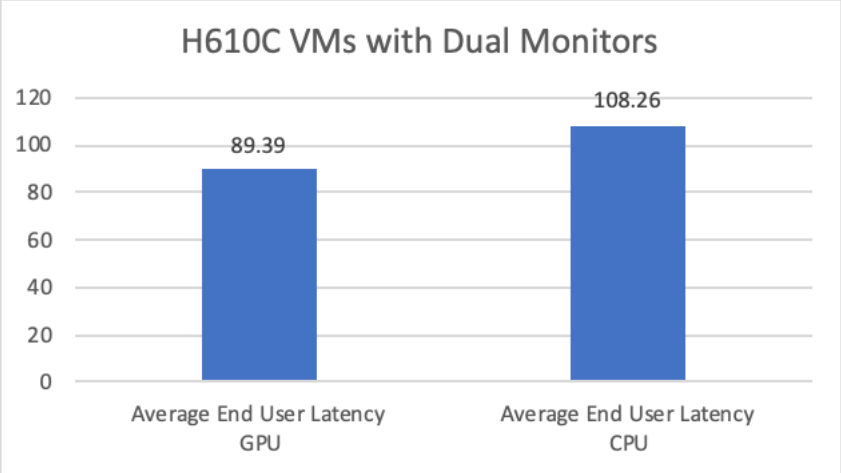
All the GPUs showed similar performance results with an average of about 40%.

Figure 26) H610C GPU utilization.



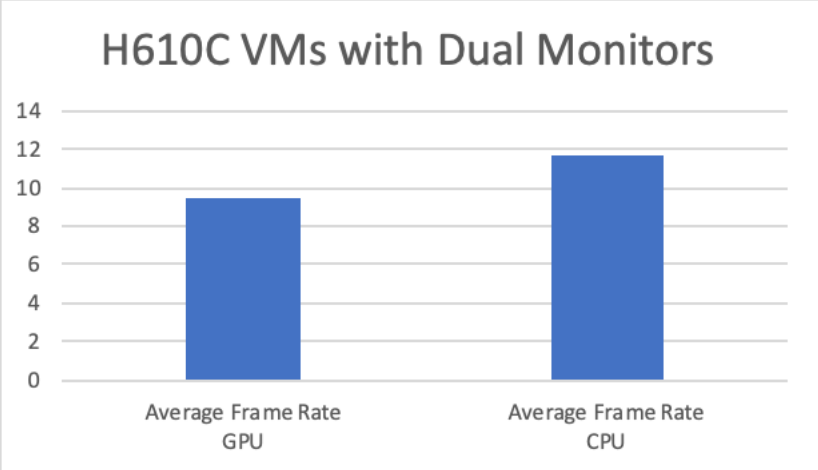
The average end-user latency improved approximately 21% with GPU.

Figure 27) H610C end-user latency.



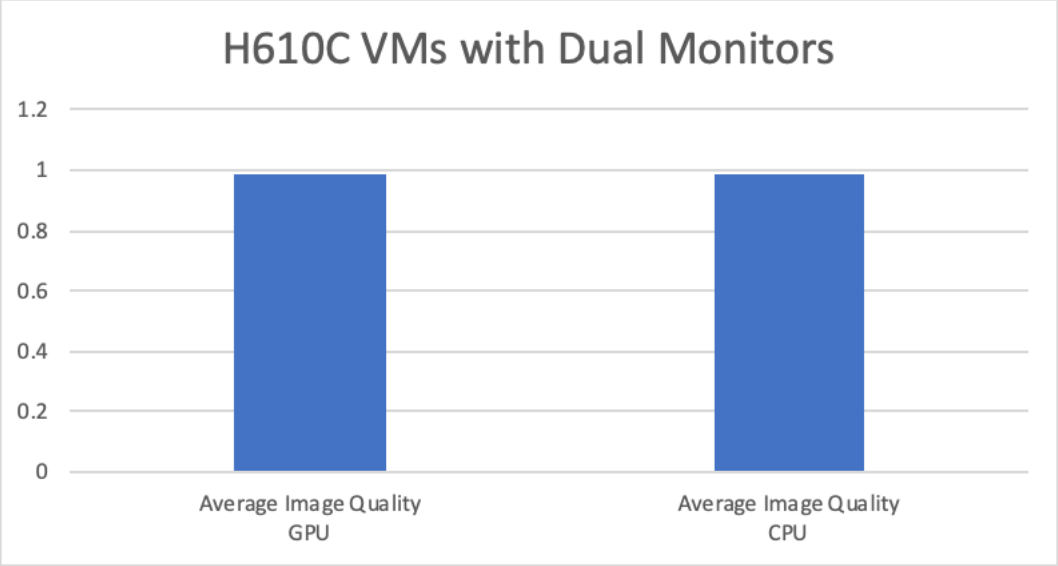
Compared to GPU, the frame rates improved with CPU. This was out-of-box testing, so it might require some fine-tuning.

Figure 28) H610C frame rates.



The average image quality was the same on both monitors.

Figure 29) H610C image quality.



In summary, H610C VMs with a M10-2B profile for dual 4K monitors delivered the following results:

Note: These results are compared to CPU-only workload.

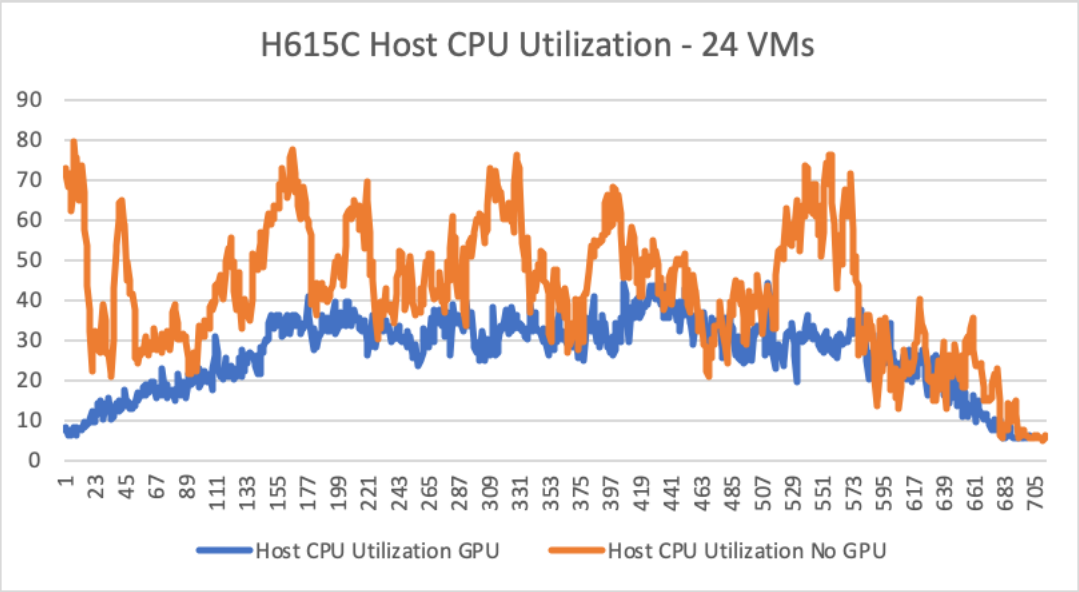
- 40% better host CPU utilization
- 21% better average end-user latency
- 23% less frame rates
- Same image quality

Two 4K Monitors with NetApp HCI H615C

For HCI H615C with NVIDIA Grid version 9.1, to run VMs with dual 4K monitors, a minimum of a T4-2B vGPU profile is required. With that profile, 24 VMs can be deployed per host.

During the GPU versus CPU test, the vSphere percent core CPU utilization delivered the results shown in Figure 30.

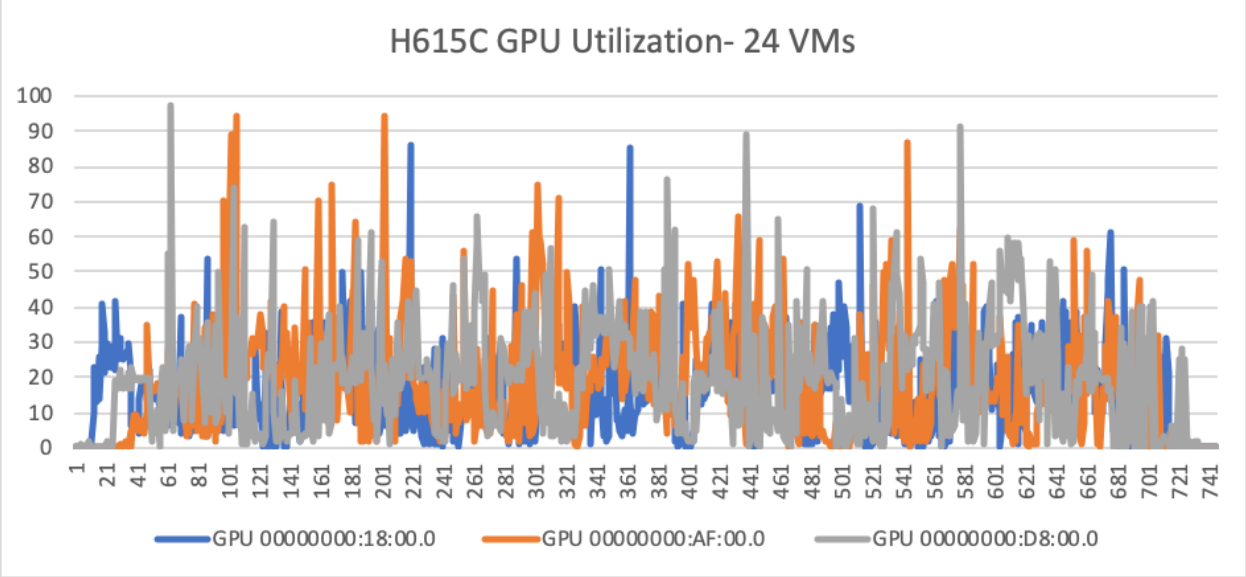
Figure 30) H615C host utilization.



The average host CPU utilization with GPU was approximately 27%; with just CPU, it was approximately 42%. Using GPU saved approximately 57% host CPU utilization.

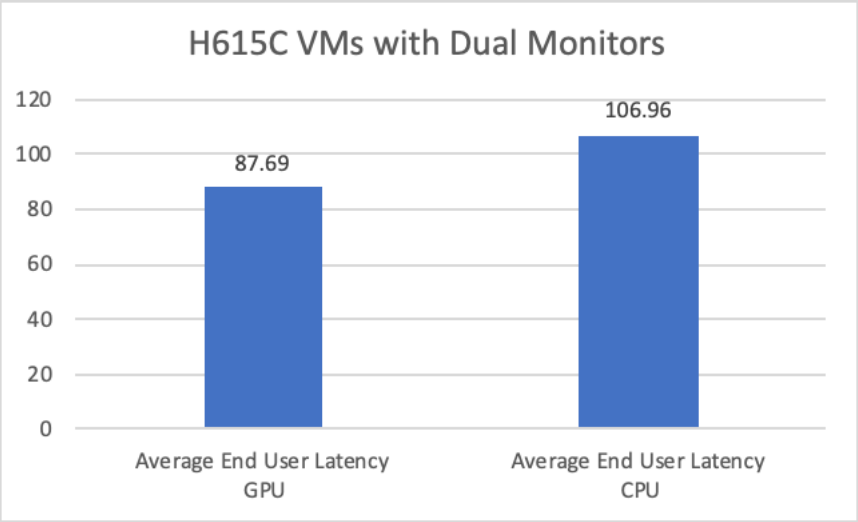
The GPU workload was evenly spread across all three GPUs using encoder and decoder.

Figure 31) H615C GPU utilization.



The average end-user latency was improved approximately 22% with GPU.

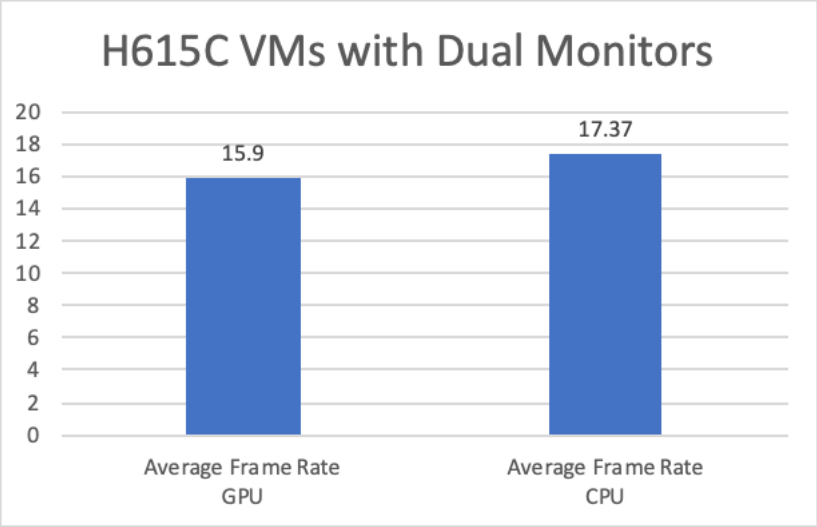
Figure 32) H615C end-user latency.



The average frame rate was improved approximately 9% with CPU alone.

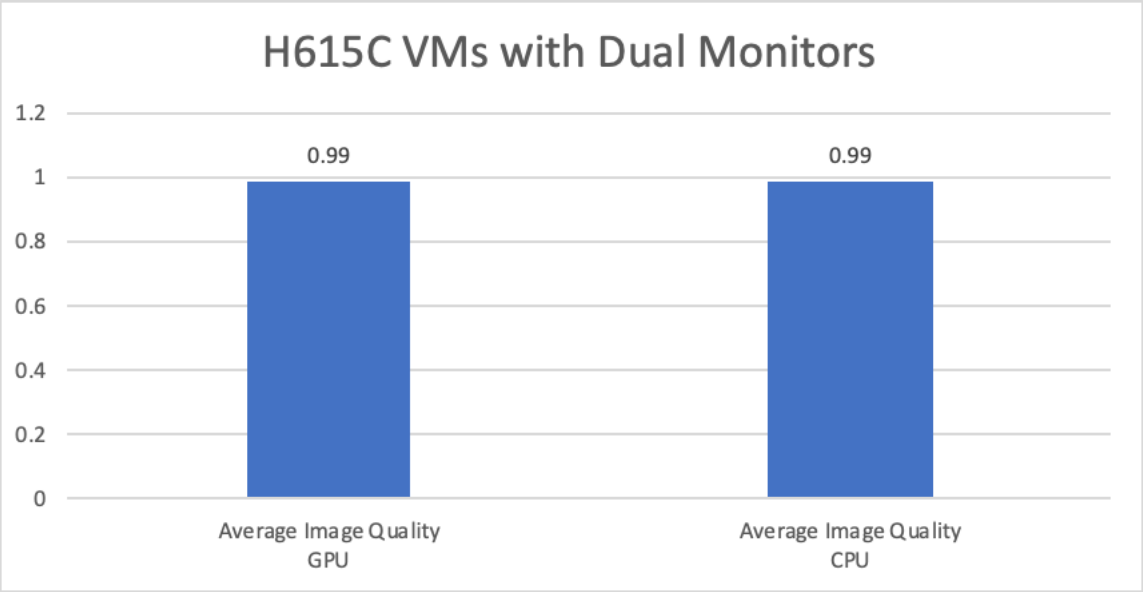
Note: This test was performed with the default settings.

Figure 33) H615 frame rate.



The image quality was the same on both monitors.

Figure 34) H615C image quality.



In summary, the H615C VMs with T4-2B vGPU profile for dual 4K monitors delivered the following results:

Note: These results are compared to using CPU alone.

- 57% better host CPU utilization
- 22% better average end-user latency
- 9% less frame rate
- Same image quality

9.4 Single-Server nVector SPECviewperf Workload Tests

The Standard Performance Evaluation Corporation (SPEC) is a nonprofit corporation formed to establish, maintain, and endorse standardized benchmarks and tools. For graphics and workstation performance, they have endorsed Application Performance Characterization (SPECapc) tools geared toward specific tools and required vendor licenses. Here is a sample list of SPECapc tools:

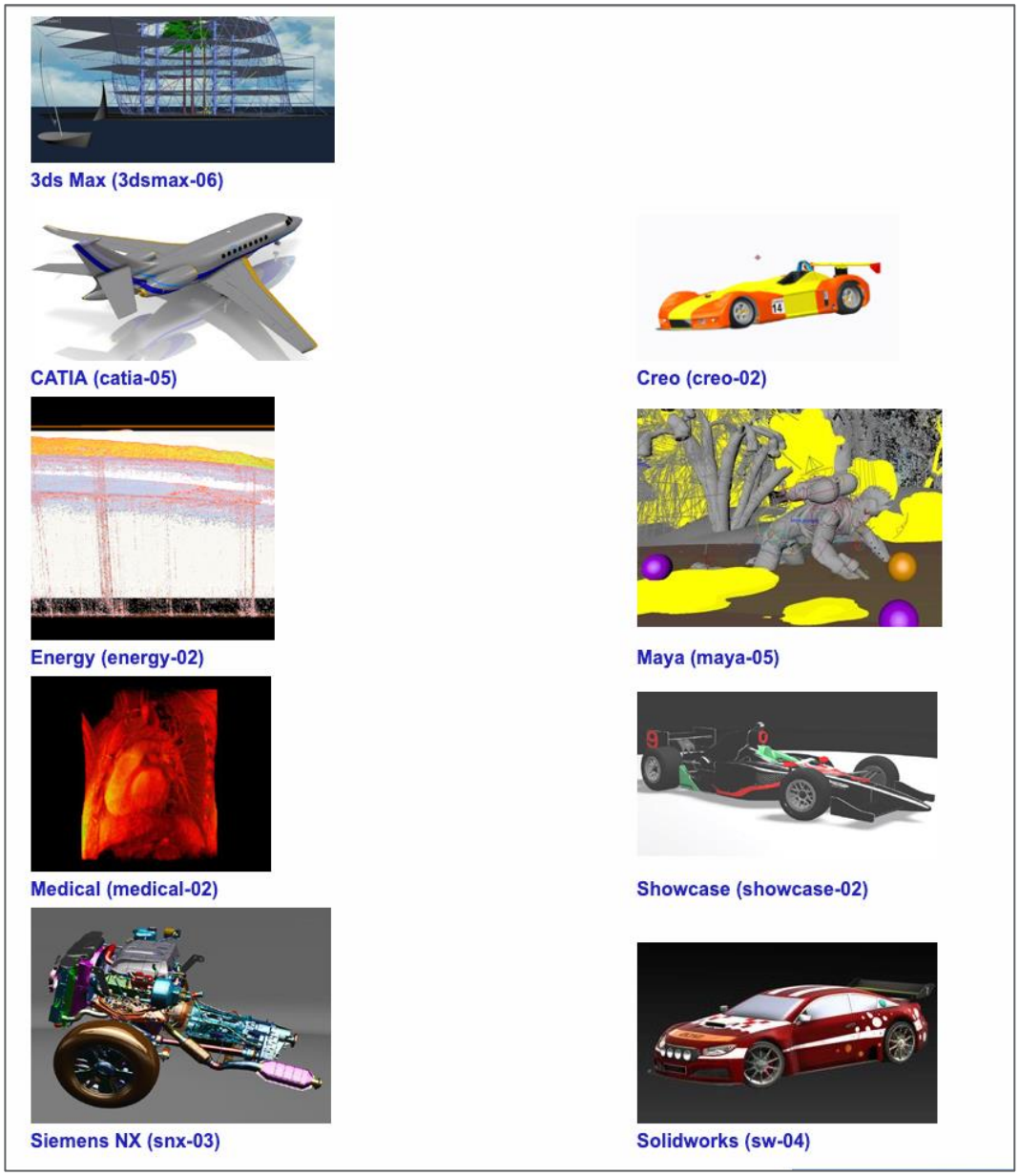
- SPECapc for 3ds Max 2015
- SPECapc for Maya 2017
- SPECapc for PTC Creo 3.0
- SPECapc for Siemens NX 9.0 and 10.0
- SPECapc for SolidWorks 2017

SPECviewperf and SPECworkstation both measure graphics performance based on professional applications. These benchmarks measure 3D graphics performance using the OpenGL and Direct X APIs. The benchmarks' workloads, known as viewsets, represent graphics content and behavior from actual applications.

SPECviewperf is geared toward GPU cards. SPECworkstation also measures all key aspects of a workstation, such as CPU, memory, storage, and so on. The test duration for SPECworkstation is longer than for SPECviewperf.

Figure 35 shows the viewset of SPECviewperf 13.

Figure 35) SPECviewperf 13 viewsets.



Note: The 3ds Max and Showcase viewsets are not available when running 4K tests.

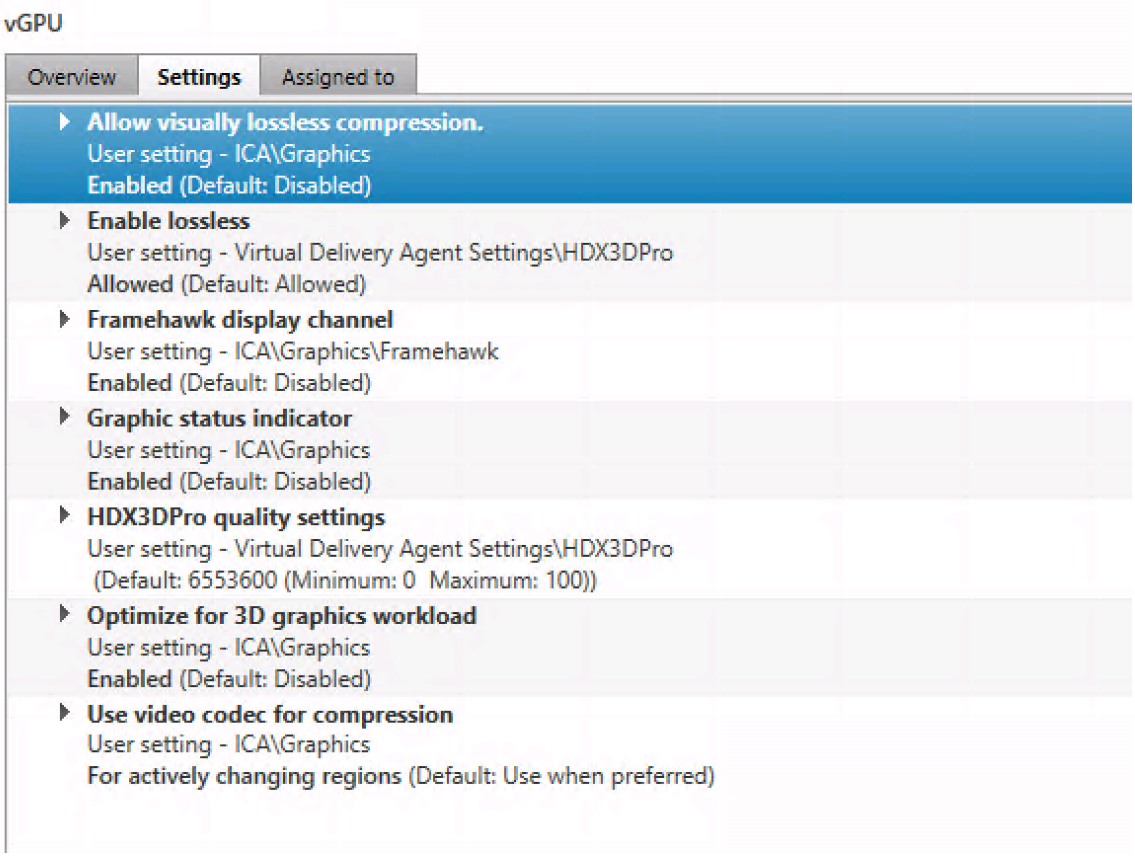
We performed the tests using the NVIDIA nVector toolset, which orchestrates the creation of VMs and client machines and sets the encoding options, executes the tests, and captures the performance data.

For every viewset, we performed the test with three VMs configured with the 16Q profile with the maximum frame buffer at full scale on a single NetApp HCI H615C server. We also used 12 VMs with the 4Q profile and the minimum frame buffer required for workloads at full scale on a single NetApp HCI H615C server.

For the 16Q profile test, we used a VM configuration of 18 vCPUs and 16GB of RAM. For the 12 VM test, we used a VM configuration of 6 vCPU and 16GB of RAM. In addition, we disabled the frame-rate limit. The client VM configuration is 4 vCPU, 4GB of RAM, and a GPU with the 1Q profile.

The Citrix policy applied to the GPUs is shown in Figure 36.

Figure 36) Citrix policy for GPU testing.



SPECviewperf measures the frame rate, or frames per second (FPS), at which a graphics card can render scenes across a wide variety of applications and usage models. Each viewset represents an application or usage model.

3ds Max (3dsmax-06)

The 3ds Max viewset was created from traces of the graphics workload generated by Autodesk 3ds Max 2016. The styles of rendering in the viewset reflect those most commonly used in major markets, including realistic, shaded, and wireframe. Some less commonly used but interesting rendering models such as facets, graphite, and clay are also incorporated. The animations in the viewset are a combination of model spin and camera fly-through, depending on the model.

We performed the following viewset tests:

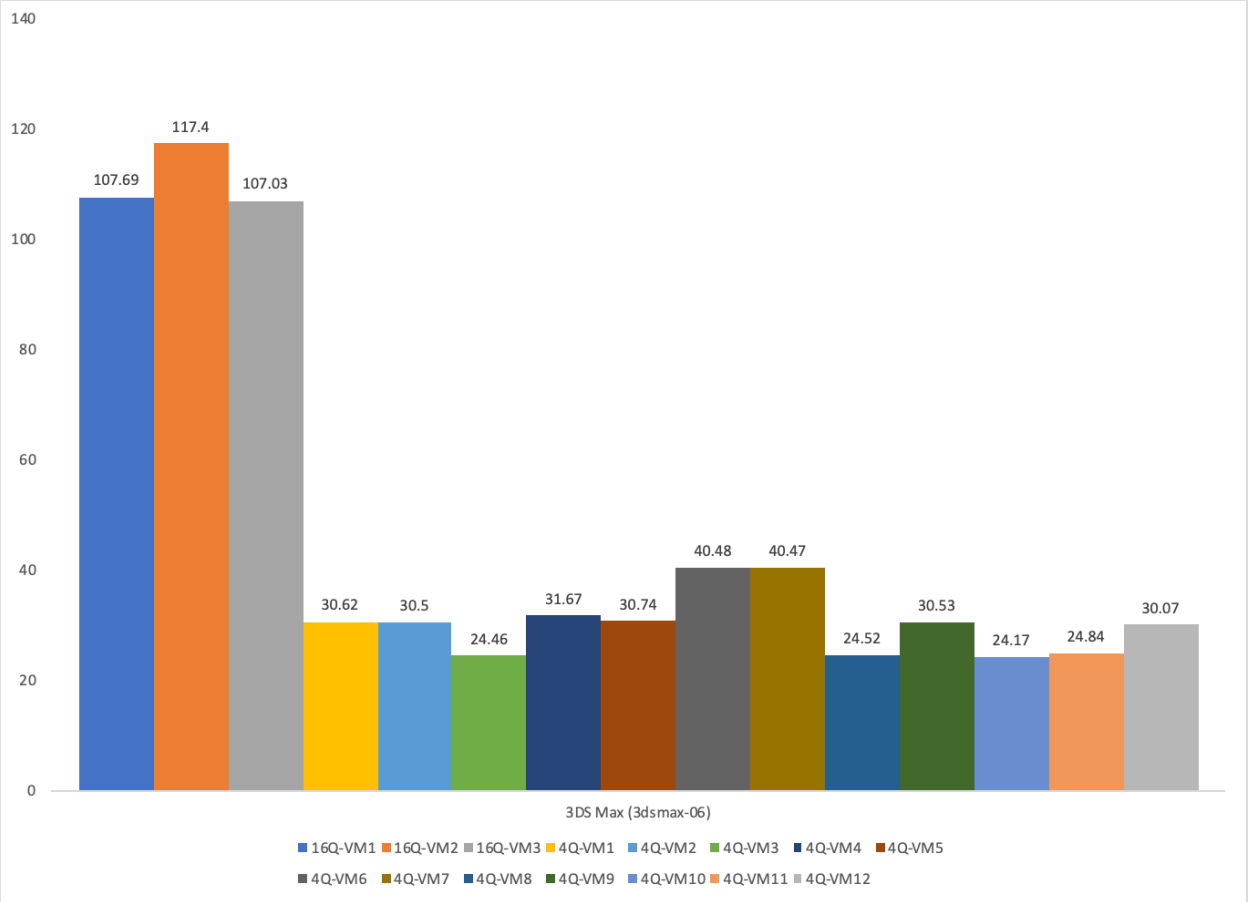
- Architectural model, shaded
- Architectural model, graphite
- Space model, wireframe
- Space model, clay
- Underwater model, wireframe
- Underwater model, shaded
- Hugh fish model, shaded
- Office model, realistic
- Office model, shaded

- Office model, realistic, with materials

For more information, see the SPEC page on the 3ds Max viewset.

The composite score of the 3ds Max for T4-4Q and T4-16Q vGPU profiles is shown in Figure 37.

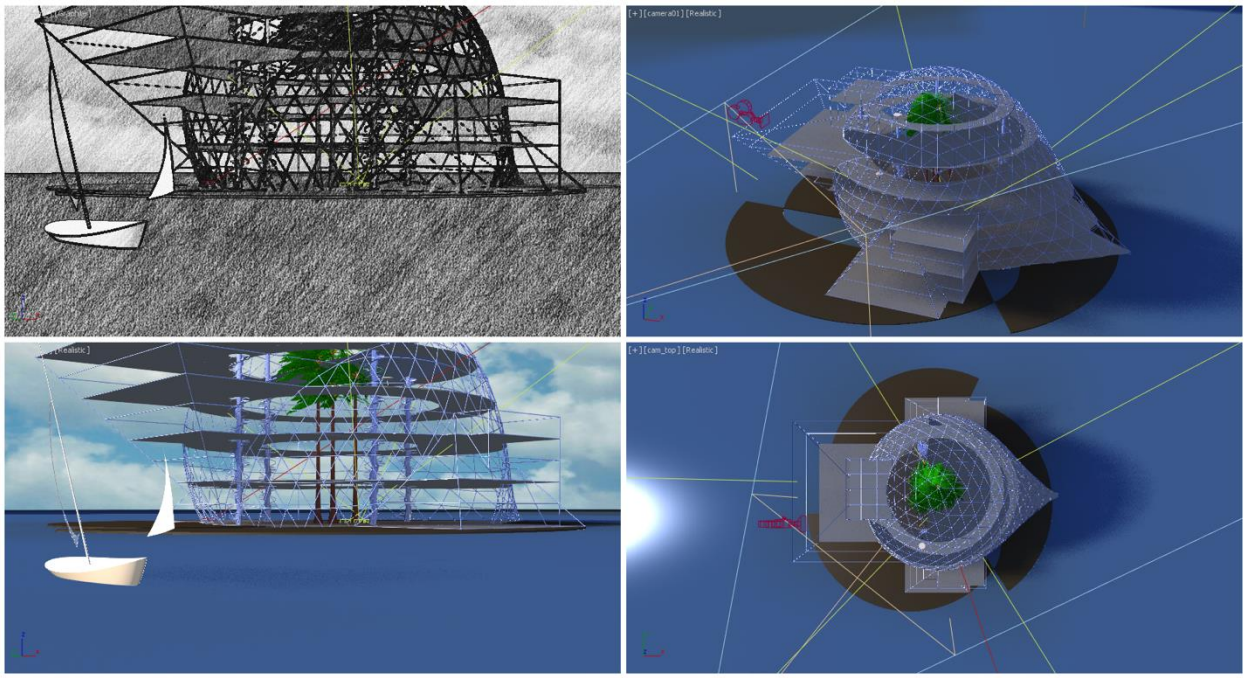
Figure 37) 3ds Max composite scores.



We can clearly see that the 16Q profile performed much better than the 4Q profile. Either the workload was benefiting from extra frame buffers or dedicated GPU compute resource and encoder/decoder. With variations on the 4Q profile, we can sense that it was related to shared GPU compute, encoder, decoder rather than frame buffer.

Figure 38 is a sample screenshot captured during the test.

Figure 38) 3ds Max sample.



CATIA (catia-05)

The catia-05 viewset was created from the traces of the graphics workload generated by the CATIA V6 R2012 application from Dassault Systèmes. Model sizes range from 5.1 to 21 million vertices.

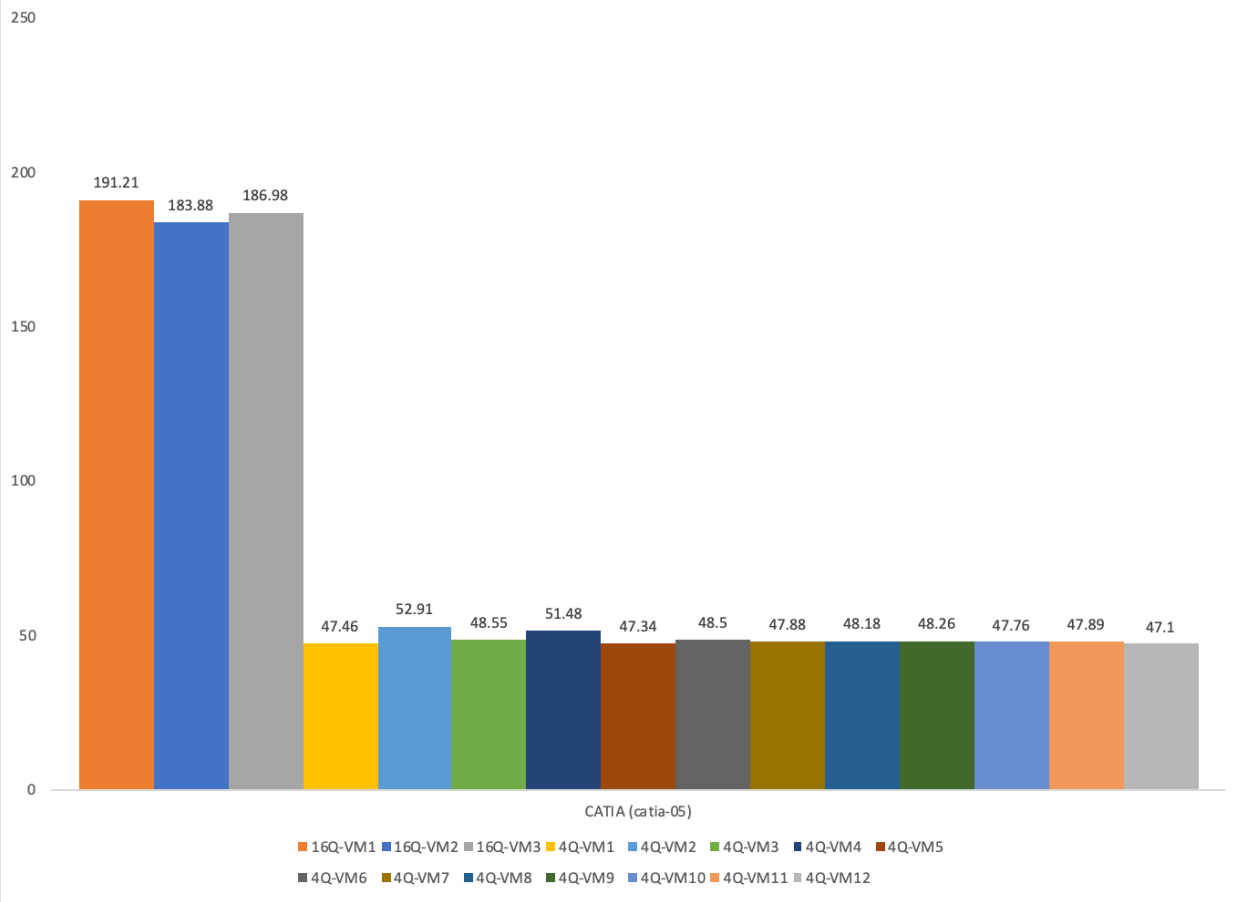
The viewset includes numerous rendering modes supported by the application, including wireframe, antialiasing, shaded, shaded with edges, depth of field, and ambient occlusion.

Viewset tests included the following:

- Race car shaded with ambient occlusion and depth of field effect
- Race car shaded with pencil effect
- Race car shaded with ambient occlusion
- Airplane shaded with ambient occlusion and depth-of-field effect
- Airplane shaded with pencil effect
- Airplane shaded
- Airplane shaded with edges
- Airplane shaded with ambient occlusion
- SUV1 vehicle shaded with ground reflection and ambient occlusion
- SUV2 vehicle shaded with ground shadow
- SUV2 vehicle shaded with ground reflection and ambient occlusion
- Jet plane shaded with ground reflection and ambient occlusion
- Jet plane shaded with edges with ground reflection and ambient occlusion

The test score of Catia for T4-4Q and T4-16Q is shown in Figure 39.

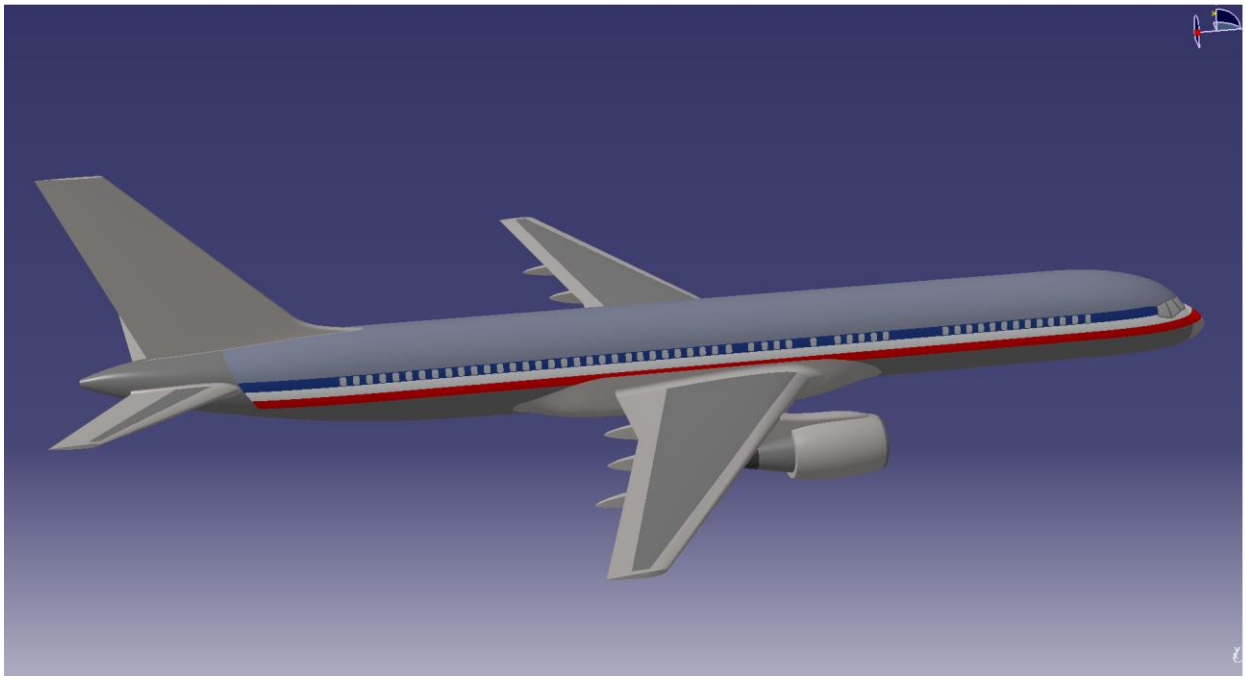
Figure 39) Catia test score.



The Catia composite score was much better with 16Q profile compared to 4Q. With 4Q, the score was almost similar, this workload was more dependent on frame buffers.

Figure 40 shows a sample screenshot captured during the test.

Figure 40) CATIA sample.



Creo (creo-02)

The creo-02 viewset was created from traces of the graphics workload generated by the Creo 3 and Creo 4 applications from PTC. Model sizes range from 20 to 48 million vertices. The viewsets include numerous rendering modes supported by the application.

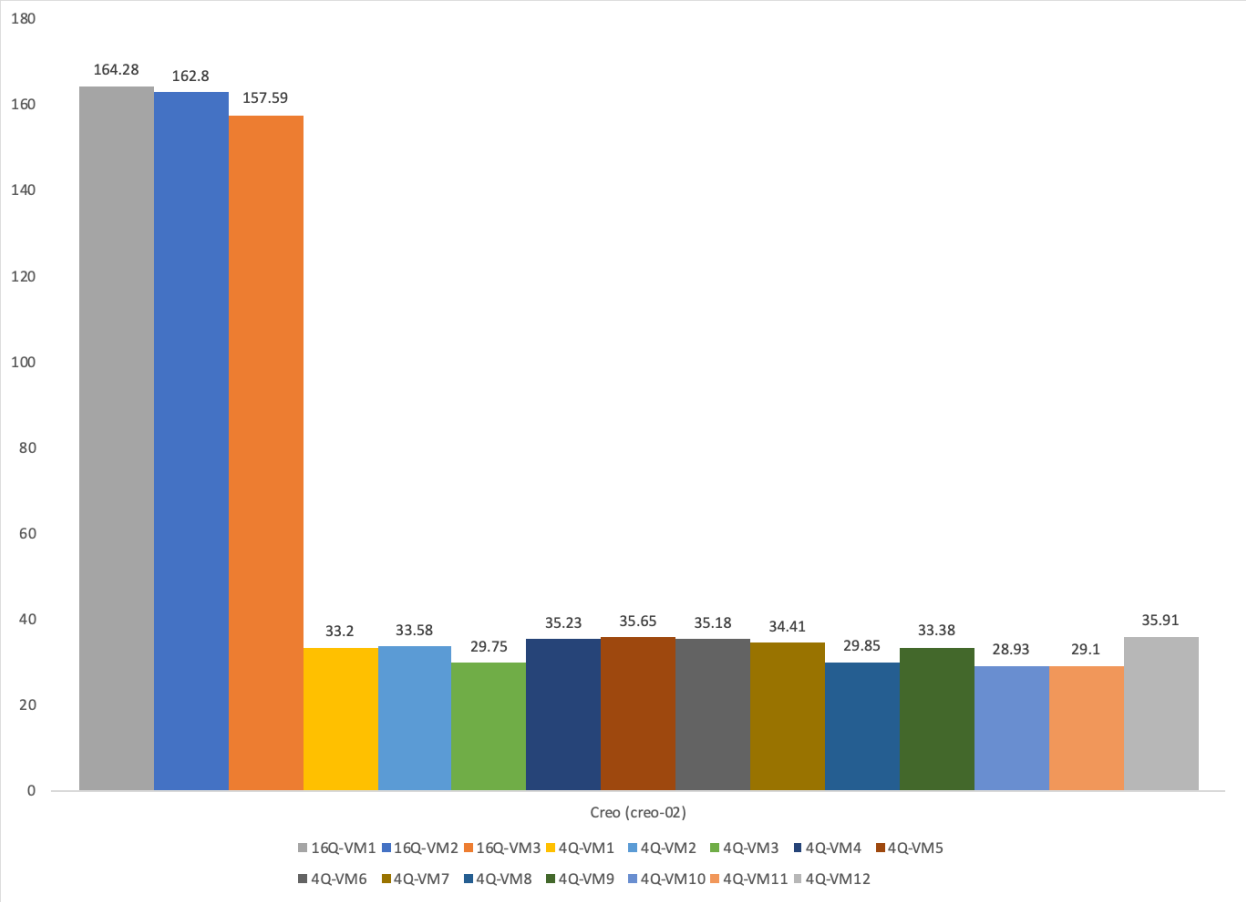
Viewset tests included the following:

- Worldcar in shaded mode, with environment mapped reflections, texture space bump mapping, image background, and screen-space ambient occlusion
- Worldcar in shaded mode, with reflections, bump mapping, image background, ambient occlusion, and 4x multisampled antialiasing
- Worldcar in shaded mode, with reflections, bump mapping, image background, ambient occlusion, and 8x multisampled antialiasing
- Worldcar in shaded mode
- Engine in shaded mode
- Motorcycle in shaded mode and 4x multisampled antialiasing
- Worldcar in shaded-with-edges mode and 4x multisampled antialiasing
- Engine in shaded-with-edges mode
- Motorcycle in shaded-with-edges mode
- Four bombers in shaded-with-edges mode and 8x multisampled antialiasing (traced from PTC Creo 4)
- Four engines in wireframe mode and 4x multisampled antialiasing
- Four bombers in wireframe mode (traced from PTC Creo 4)
- Worldcar in hidden-line mode
- Motorcycle in hidden-line mode and 8x multisampled antialiasing
- Engine in no-hidden-edge mode

- Four bombers in no-hidden-edge mode and 8x multisampled antialiasing (traced from PTC Creo 4)

The test score of Creo with T4-4Q and T4-16Q profiles are shown in Figure 41.

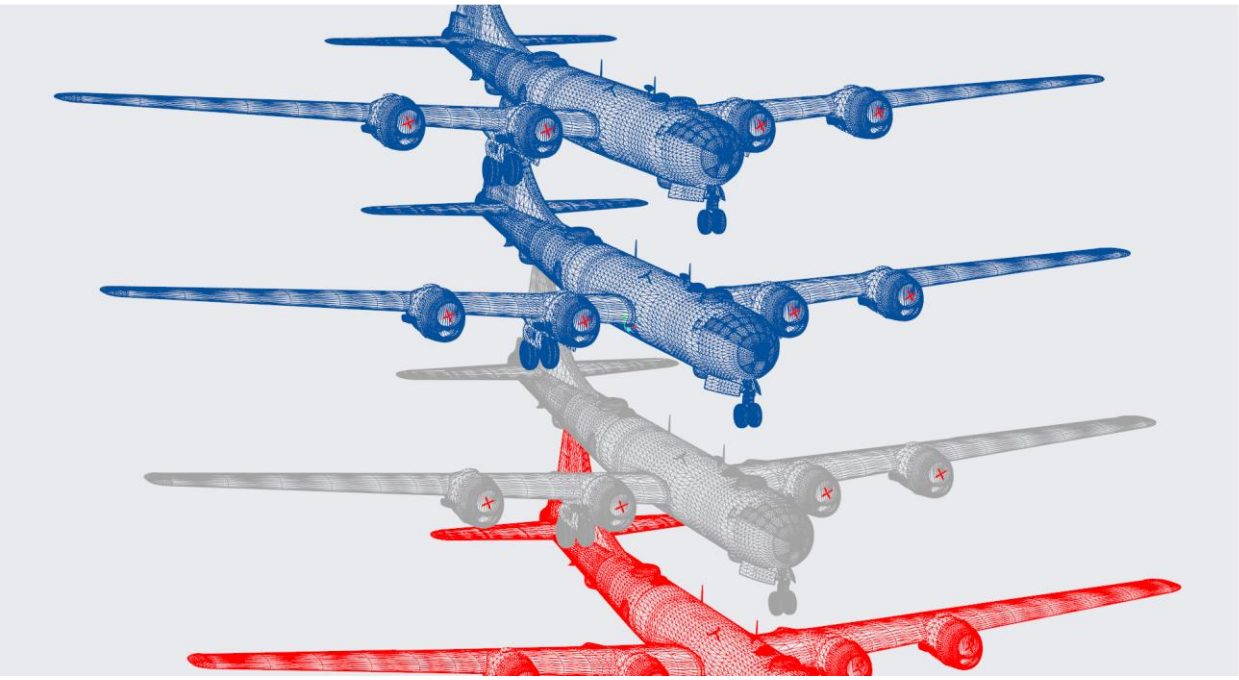
Figure 41) Creo test score.



Creo test results demonstrated that 16Q performed better than 4Q profile. 16Q profiles were run on dedicated GPU resources, which included compute, encoder, and decoder engines. With 4Q profile, it was shared with three other VMs.

Figure 42 is a sample screenshot that was captured during the test.

Figure 42) Creo sample.



Energy (energy-02)

The energy-02 viewset is based on rendering techniques used by the open-source OpendTect seismic visualization application. In a manner similar to medical imaging such as MRI or CT, geophysical surveys generate image slices through the subsurface that are built into a 3D grid. Volume rendering provides a 2D projection of this 3D volumetric grid for further analysis and interpretation.

At every frame, the bounding cube faces of the volume are tessellated and rendered with a fragment shader that performs a ray cast from the eye position through the volume, accumulating transparently lit, color-mapped values. This process is continued until either the pixel becomes fully opaque or the volume is exited.

The voxel in the 3D grid is a single scalar value. A transfer function—simply a 1D lookup table—maps the 3D density value to color and alpha values. For lighting calculations, the gradients are computed on the fly using the central differences at each voxel. These state changes exercise various parts of the graphics subsystem. This viewset makes use of hardware support for 3D textures and therefore trilinear interpolation.

In addition to the volume rendering, the test includes both inline and crossline planes (slices in the X and Y planes). Also, for some subtests, horizons are present; these are geological strata boundaries of interest that are generated by exploration geophysicists and rendered using textured triangle strips.

The 3D datasets used in this viewset are real-world seismic datasets found at the SEG wiki. These datasets were translated from their native SEG-Y format and compressed using JPEG-2000.

We performed the following viewset tests:

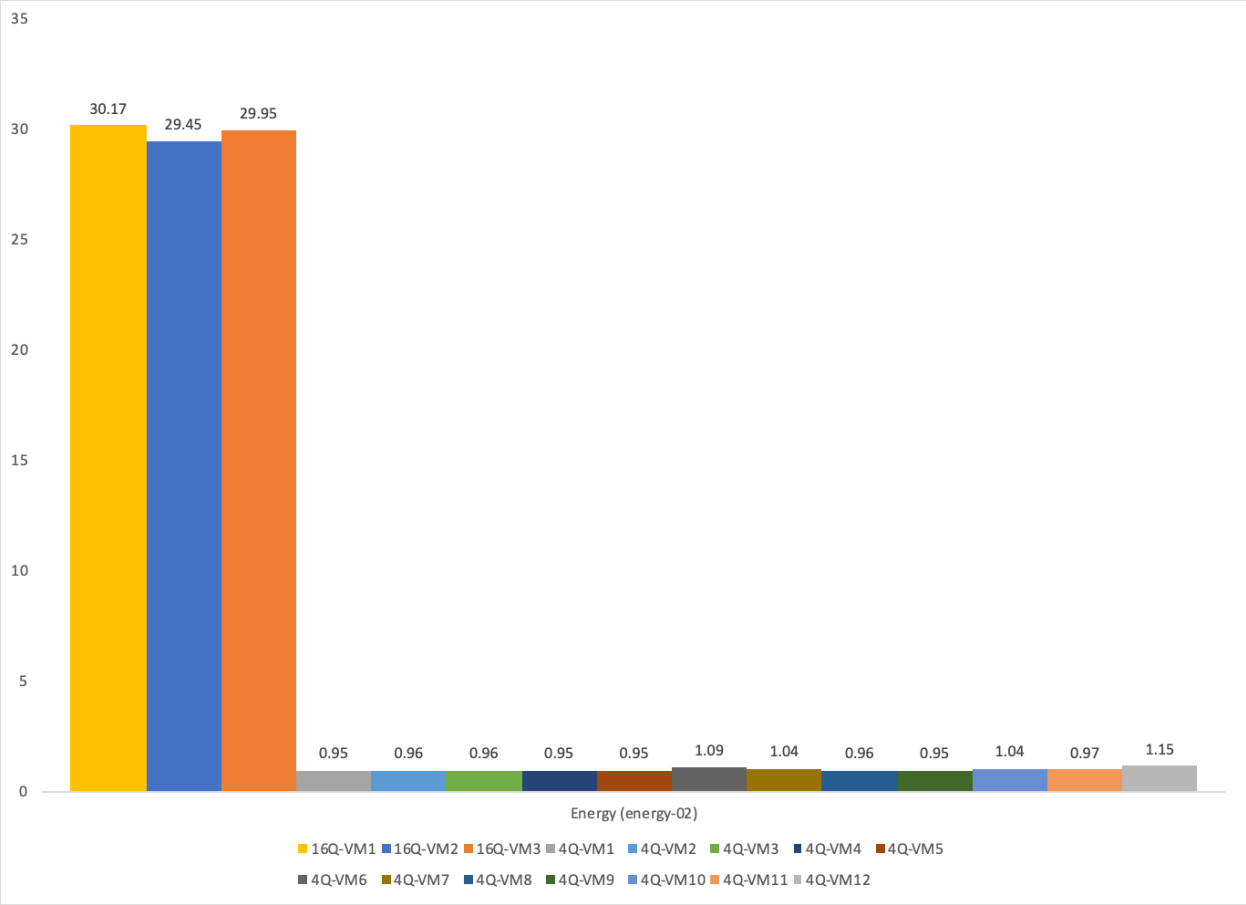
- Blake Ridge volume (1307x95x1300) and horizons
- F3 Netherlands volume (950x450x462) and horizons
- Opunake volume (1949x731x1130)
- Blake Ridge volume (with animated clipping plane) and horizons

- F3 Netherlands volume (with animated clipping plane) and horizons
- Opunake volume (with animated clipping plane)

For the Energy viewset, the frame-buffer size affects the composite score significantly.

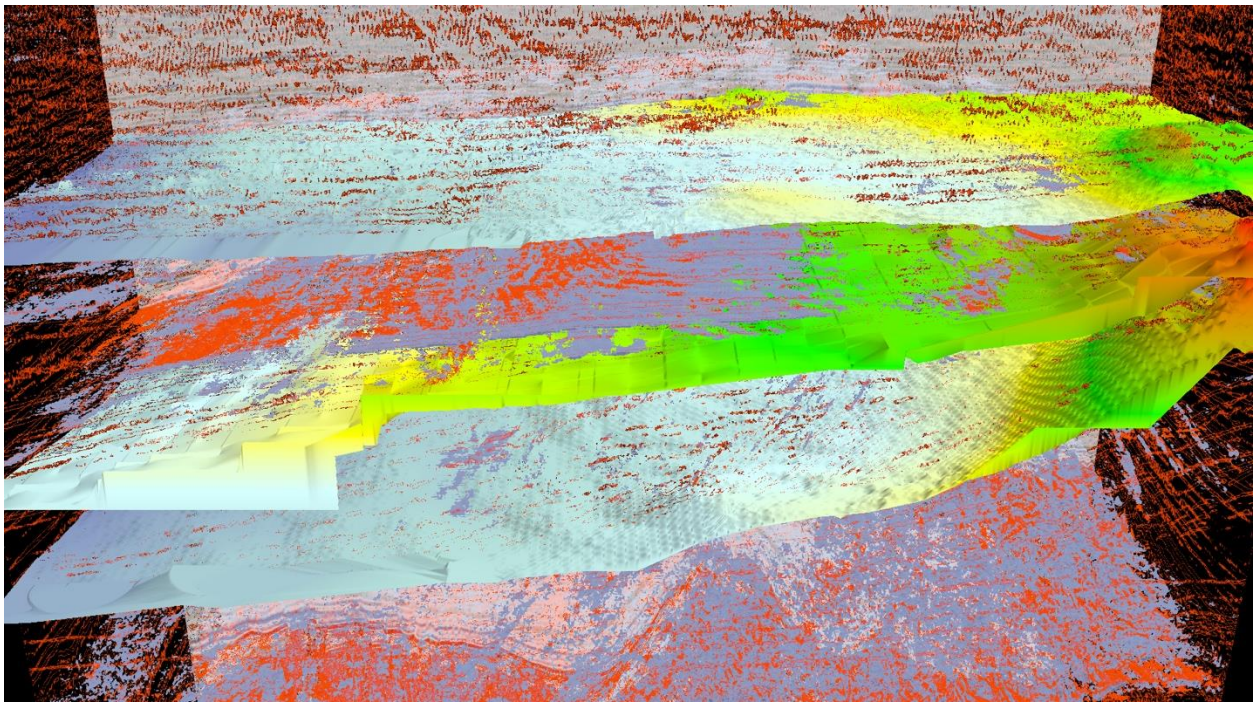
The composite score of both vGPU profiles are shown in Figure 43.

Figure 43) Energy test score.



Energy composite scores clearly shows, with 16Q profile it performed much better compared to 4Q. Consider using high-frame buffers for these use cases.

Figure 44) Energy sample.



Maya (maya-05)

The maya-05 viewset was created from traces of the graphics workload generated by the Maya 2017 application from Autodesk.

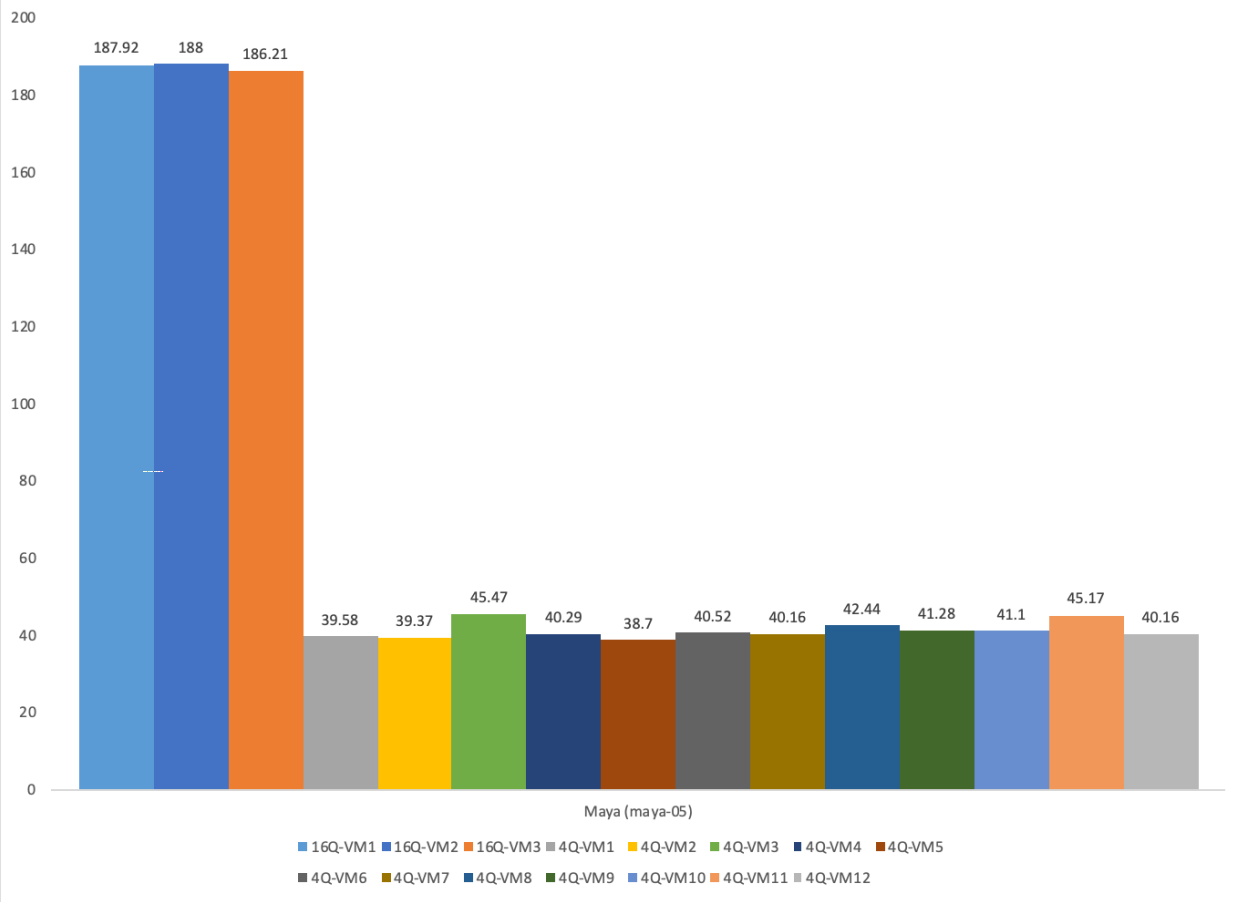
This viewset includes numerous rendering modes supported by the application, including shaded mode, ambient occlusion, multisample antialiasing, and transparency. All tests are rendered using Viewport 2.0.

We performed the following viewset tests:

- Toy store, smooth shaded with wireframe on shaded mode, ambient occlusion, and 4x multisample antialiasing
- Toy store, wireframe mode and 8x multisample antialiasing
- Jungle escape, smooth shaded with hardware texture mode and ambient occlusion
- Jungle escape, smooth shaded with hardware texture mode
- Sven space, smooth shaded with hardware texture mode
- Sven space, smooth shaded, ambient occlusion, and 4x multisample antialiasing
- HSM satellite, smooth shaded and 8x multisample antialiasing
- Ship splash, smooth shaded with all lights
- Ship splash, wireframe mode and 4x multisample antialiasing
- Ship splash, smooth shaded with hardware texture mode, ambient occlusion, and 8x multisample antialiasing

The composite score of Maya for both vGPU profiles is shown in Figure 45.

Figure 45) Maya test score.



Maya composite scores were also better with higher frame buffer profiles.

Figure 46) Maya sample.

Medical (medical-02)

The medical-02 viewset uses the Tuvok rendering core of the ImageVis3D volume visualization program. It renders a 2D projection of a 3D volumetric grid. A typical 3D grid in this viewset is a group of 3D slices acquired by a scanner such as a CT or MRI machine.

Two rendering modes are represented: slice-based rendering and ray casting.

For slice-based rendering, a series of coplanar slices aligned with the current viewing angle are computed on the CPU. They are then sent to the graphics hardware for texturing and further calculations, such as transfer function lookup, lighting, and clipping to reveal internal structures. Finally, the slices are blended together before the image is displayed.

For ray casting, rays are cast through the volume, accumulating transparently lit, colored pixels until full opacity or the bounds of the volume are reached.

For both slice-based and ray-cast rendering, the volumes are potentially subdivided into 512x512x512 3D volumes. This technique, known as bricking, typically results in better rendering performance on a wider range of GPU hardware.

The voxel in the 3D grid is a single scalar value. A transfer function—either a 1D or a 2D lookup table—maps the 3D density value to color and alpha values. For 2D tables, the second axis is defined as the magnitude of the gradient at each sample. For lighting calculations, the gradients are computed on the fly using the central differences at each voxel. These state changes exercise various parts of the graphics subsystem. This viewset makes use of hardware support for 3D textures and therefore trilinear interpolation.

The following descriptions and weighting are for the four datasets in this viewset:

- A 4D heart dataset composed of multiple 3D volumes iterated over time. These were obtained from a phase-contrast MRI scanner. The 80MB dataset was contributed by the Department of Radiology at the Stanford School of Medicine and Lucile Packard Children's Hospital. Each volume consists of 256x256x32 16-bit samples.

- A stag beetle dataset provided by the Technical University of Vienna. The dataset size is 650MB and represents a workload with larger memory requirements. The volume consists of 832x832x494 16-bit samples.
- An MRI scan of the head of a member of the SPECgpc committee, who has released the data for use in SPECviewperf. The volume consists of 232x256x192 16-bit samples.
- A CT scan of the right upper thorax and arm of the same member of the SPECgpc committee, who has also released this data for use in SPECviewperf. The volume consists of 512x512x102 16-bit samples.

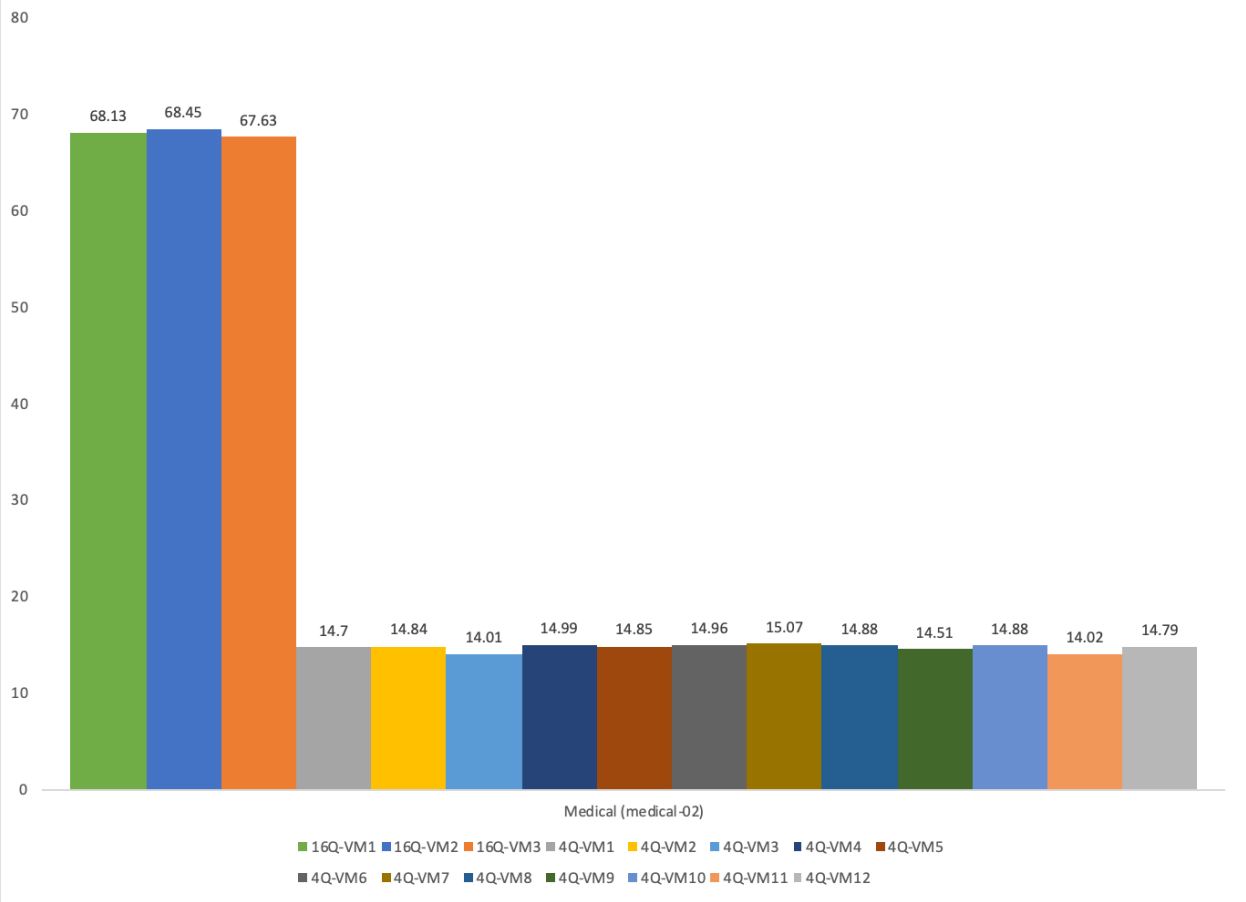
The tests in the viewset were derived from those four datasets as follows:

- 4D heart, 1D transfer function, slice-based rendering
- 4D heart, 1D transfer function, ray casting
- Stag beetle, 1D transfer function, slice-based rendering
- Stag beetle, 1D transfer function, ray casting
- Head MRI, 2D transfer, ray casting
- Head MRI, 2D transfer, ray casting, clipping plane
- Thorax CT, 2D transfer, ray casting
- Thorax CT, 2D transfer, ray casting, clipping plane

The Tuvok rendering core is licensed under the MIT open-source license; see the GitHub site for Tuvok for more information. Tuvok includes a Hilbert Curve implementation, which is copyright 1998, Rice University. Tuvok also includes LZ4, which is licensed under the BSD 2-Clause license.

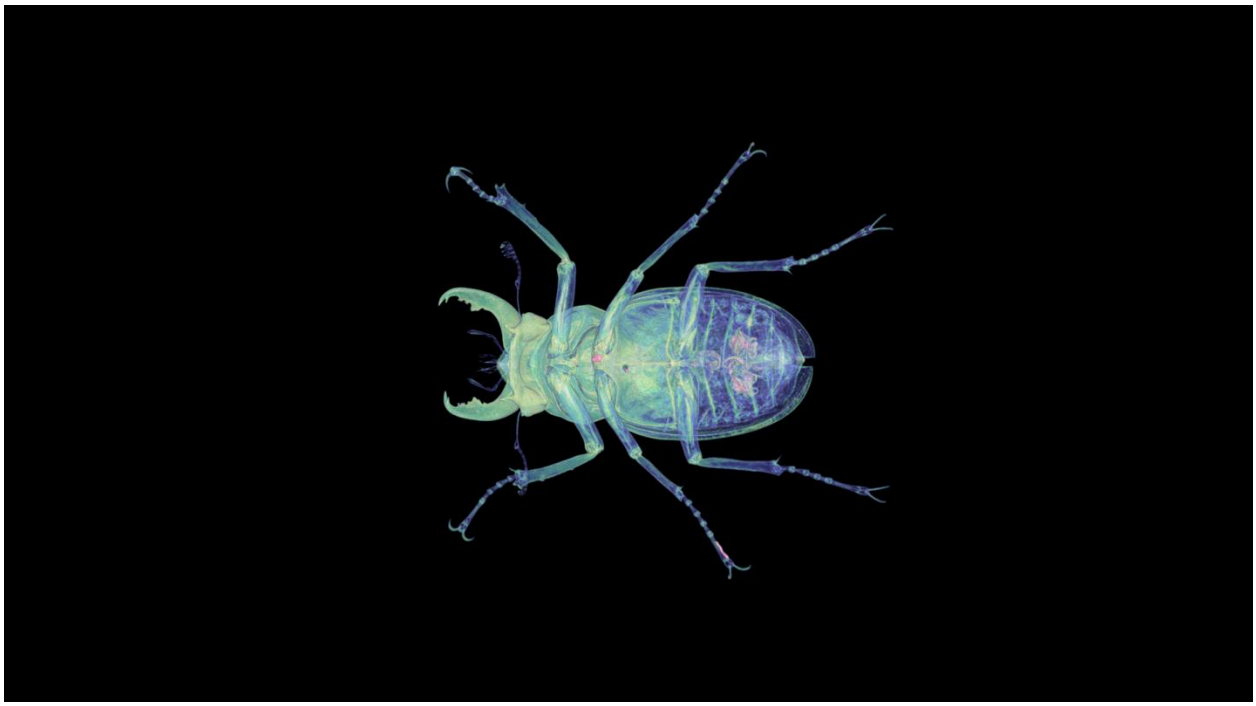
The composite score of both vGPU profiles is shown in Figure 47.

Figure 47) Medical test score.



Medical viewset also benefited from a large frame buffer.

Figure 48) Medical sample.



Showcase (showcase-02)

The showcase-02 viewset was created from traces of Autodesk's Showcase 2013 application. The model used in the viewset contains eight million vertices.

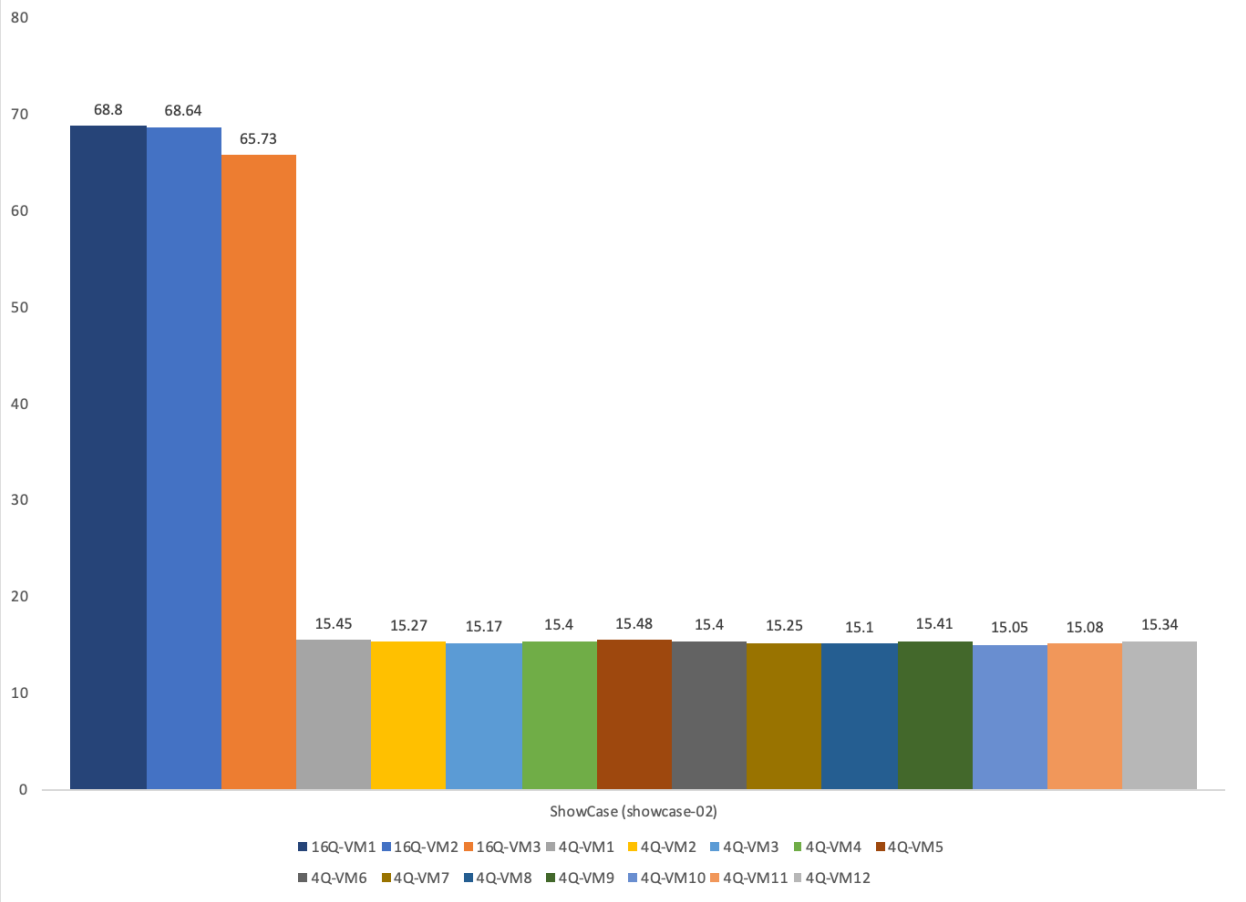
The viewset features DX rendering. Rendering modes included in the viewset include shading, projected shadows, and self-shadows.

The following tests were included in the viewset:

- Shaded with self-shadows
- Shaded with self-shadows and projected shadows
- Shaded
- Shaded with projected shadows

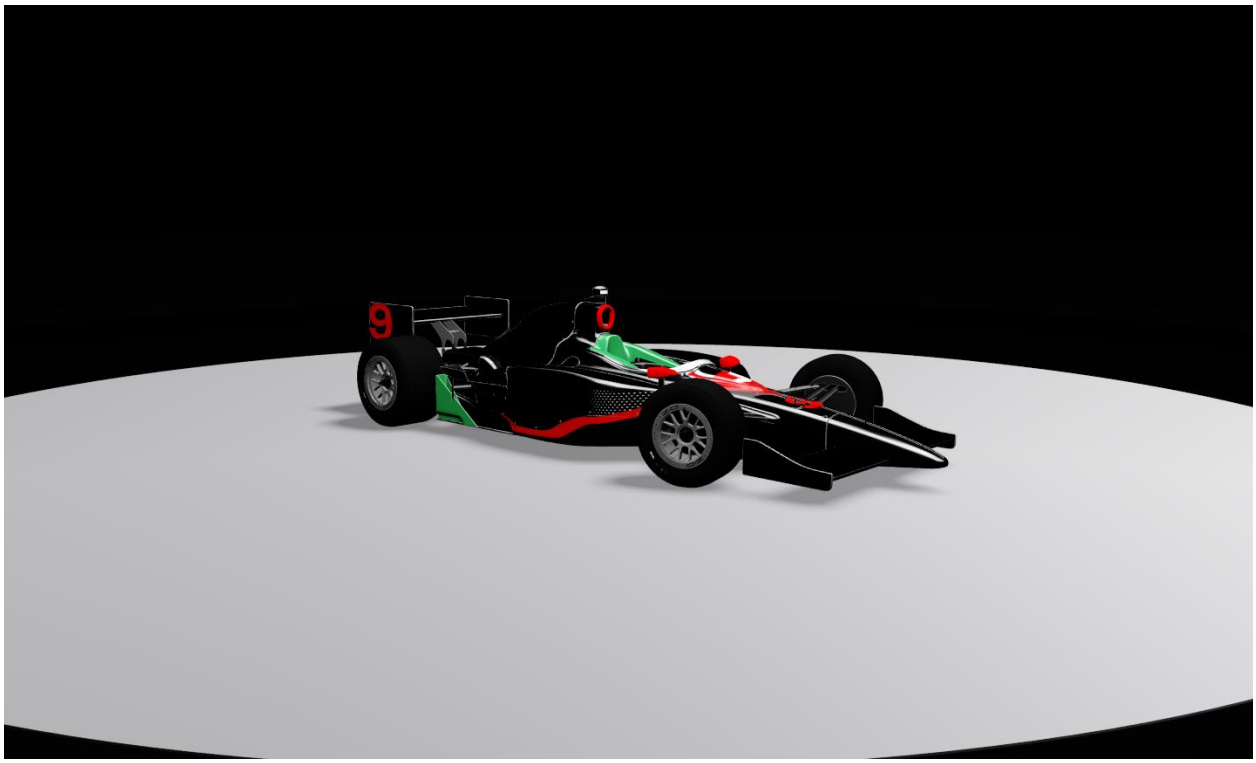
The composite score of the 4Q and 16Q vGPU profile is shown in Figure 49.

Figure 49) Showcase Test Result



The Showcase composite score was better with the 16Q profile. With the 4Q profile, the composite scores are almost similar, which indicates that it was more dependent on the frame buffer than GPU compute, encoder, and decoder.

Figure 50) Showcase sample.



Siemens NX (snx-03)

The snx-03 viewset was created from traces of the graphics workload generated by the NX 8.0 application from Siemens PLM. Model sizes range from 7.15 to 8.45 million vertices.

The viewset includes numerous rendering modes supported by the application, including wireframe, antialiasing, shaded, shaded with edges, and studio mode.

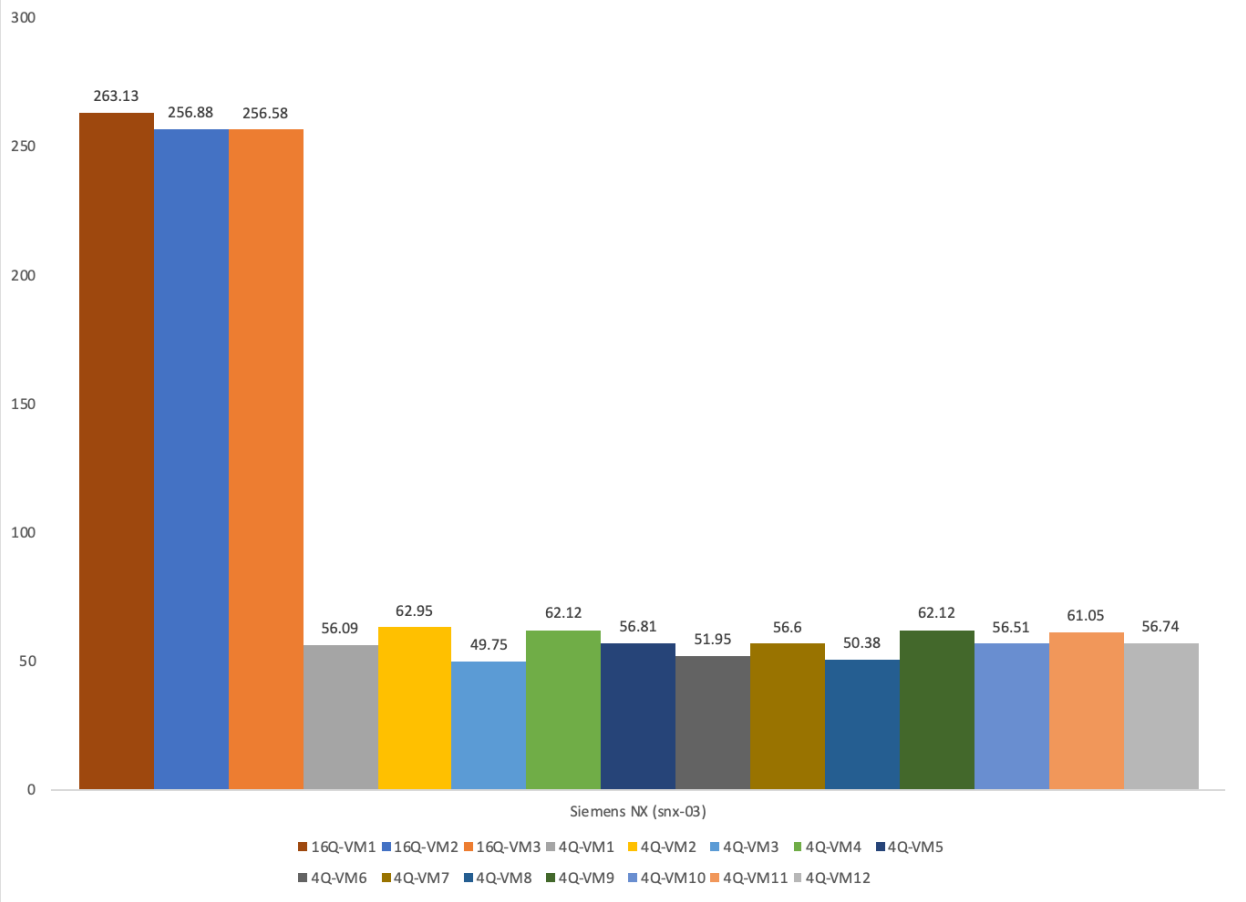
Viewset Tests

The following tests were included in the viewset:

- Powertrain in advanced studio mode
- Powertrain in shaded mode
- Powertrain in shaded-with-edges mode
- Powertrain in studio mode
- Powertrain in wireframe mode
- SUV in advanced studio mode
- SUV in shaded mode
- SUV in shaded-with-edges mode
- SUV in studio mode
- SUV in wireframe mode

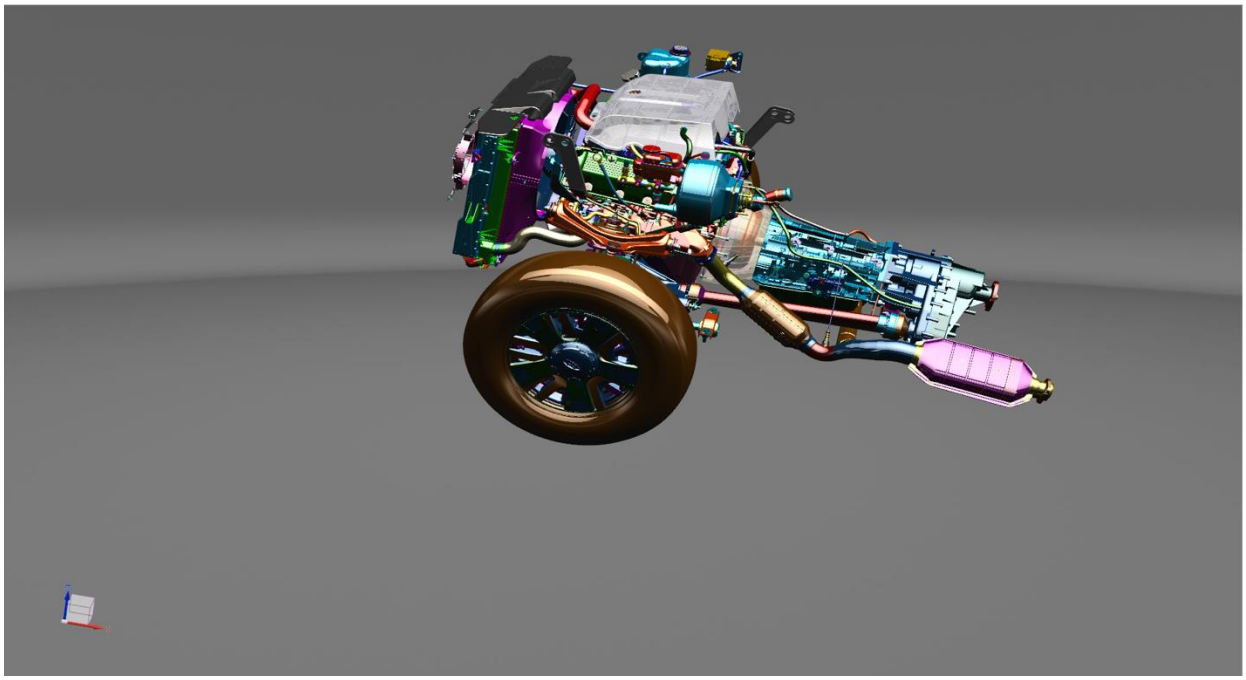
The Siemens composite score is shown in Figure 51.

Figure 51) Siemens test score.



Siemens NX used any frame buffers that were available. Variations in the 4Q profile shows that it was using shared resources such as GPU compute, encoder, and decoder.

Figure 52) Siemens NX sample.



SolidWorks (sw-04)

The sw-04 viewset was created from traces of Dassault Systèmes SolidWorks 2013 SP1 application. Models used in the viewset range in size from 2.1 to 21 million vertices.

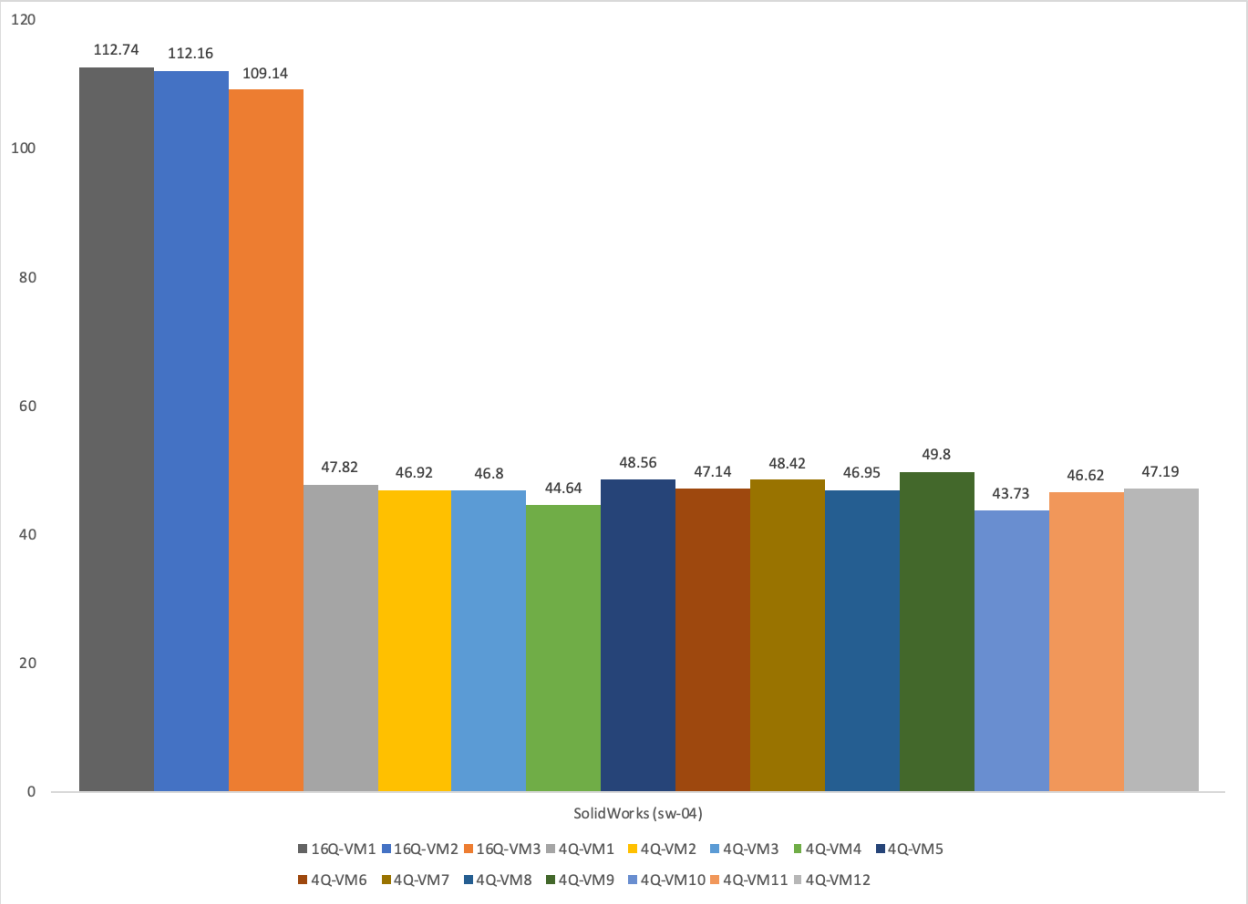
The viewset includes numerous rendering modes supported by the application, including shaded mode, shaded with edges, ambient occlusion, shaders, and environment maps.

The following tests were included in the viewset:

- Vehicle in shaded mode—normal shader with environment cubemap
- Vehicle in shaded mode—bump parallax mapping with environment cubemap
- Vehicle in shaded mode—ambient occlusion enabled with normal shader and environment map
- Vehicle in shaded-with-edges mode—normal shader with environment cubemap
- Vehicle in wireframe mode
- Rally car in shaded mode—ambient occlusion enabled with normal shader and environment map
- Rally car in shaded mode—normal shader with environment cubemap
- Rally car in shaded-with-edges mode—normal shader with environment cubemap
- Tesla tower in shaded mode—ambient occlusion enabled with normal shader and environment map
- Tesla tower in shaded mode—normal shader with environment cubemap
- Tesla tower in shaded-with-edges mode—normal shader with environment cubemap

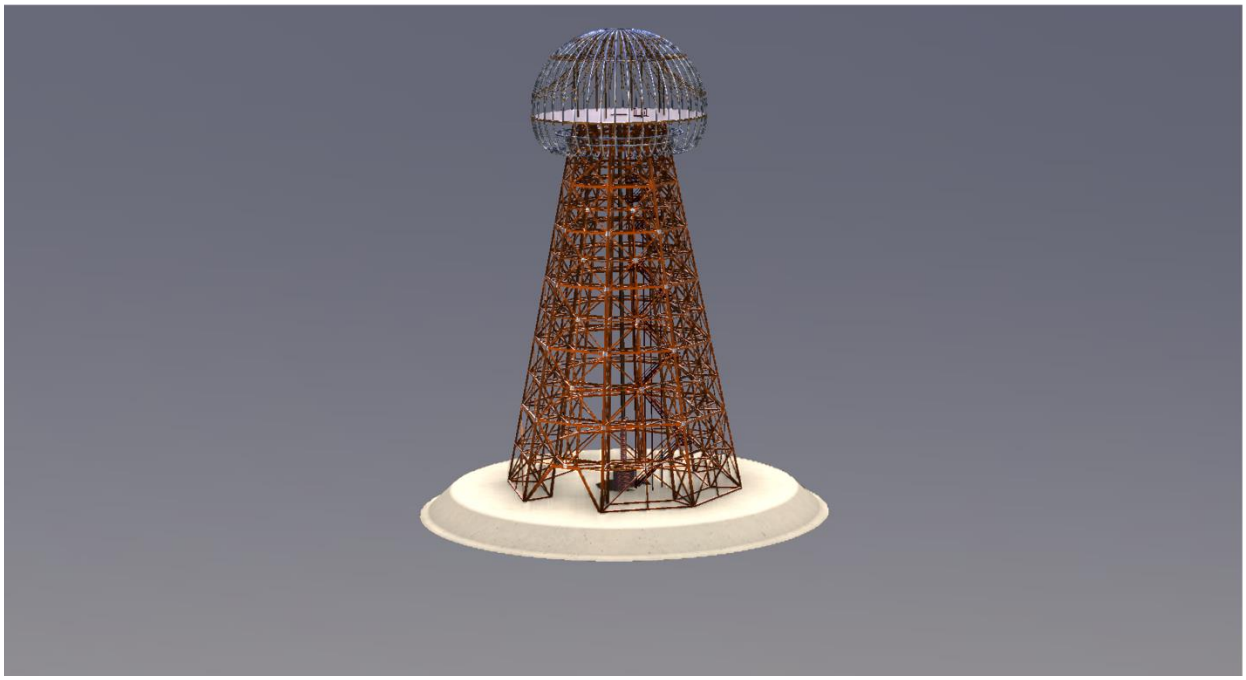
The composite score of our test case is shown in Figure 53.

Figure 53) SolidWorks test results.



As with many viewsets, SolidWorks also benefited by using large frame buffer profile.

Figure 54) SolidWorks sample.



In summary, compared with T4-4Q, T4-16Q vGPU performed much better. The Energy viewset composite scores are much lower with 4Q and can benefit with large frame buffers.

Figure 55) Test summary results.

Viewset	16Q-VM1	16Q-VM2	16Q-VM3	4Q-VM1	4Q-VM2	4Q-VM3	4Q-VM4	4Q-VM5	4Q-VM6	4Q-VM7	4Q-VM8	4Q-VM9	4Q-VM10	4Q-VM11	4Q-VM12
3DS Max (3dsmax-06)	107.69	117.4	107.03	30.62	30.5	24.46	31.67	30.74	40.48	40.47	24.52	30.53	24.17	24.84	30.07
CATIA (catia-05)	191.21	183.88	186.98	47.46	52.91	48.55	51.48	47.34	48.5	47.88	48.18	48.26	47.76	47.89	47.1
Creo (creo-02)	164.28	162.8	157.59	33.2	33.58	29.75	35.23	35.65	35.18	34.41	29.85	33.38	28.93	29.1	35.91
Energy (energy-02)	30.17	29.45	29.95	0.95	0.96	0.96	0.95	0.95	1.09	1.04	0.96	0.95	1.04	0.97	1.15
Maya (maya-05)	187.92	188	186.21	39.58	39.37	45.47	40.29	38.7	40.52	40.16	42.44	41.28	41.1	45.17	40.16
Medical (medical-02)	68.13	68.45	67.63	14.7	14.84	14.01	14.99	14.85	14.96	15.07	14.88	14.51	14.88	14.02	14.79
ShowCase (showcase-02)	68.8	68.64	65.73	15.45	15.27	15.17	15.4	15.48	15.4	15.25	15.1	15.41	15.05	15.08	15.34
Siemens NX (snx-03)	263.13	256.88	256.58	56.09	62.95	49.75	62.12	56.81	51.95	56.6	50.38	62.12	56.51	61.05	56.74
SolidWorks (sw-04)	112.74	112.16	109.14	47.82	46.92	46.8	44.64	48.56	47.14	48.42	46.95	49.8	43.73	46.62	47.19

10 Conclusion

NetApp HCI provides the flexibility to start Citrix Virtual Apps and Desktops with a few nodes and grow as your demands increase. NetApp HCI all-flash storage offers storage saving by using global in-line deduplication, compression, and thin provisioning. H610S with NVMe drives reduce the latencies further down.

NetApp HCI Compute offers a selection of GPU cards to choose based on the use cases of Citrix Virtual Apps and Desktops. GPU reduces host CPU utilization and improves end-user latency.

NetApp HCI minimum, maximum, and burst QoS features guarantee performance for your workloads. ONTAP SMB file shares make it easier to set up user home folders, and FlexGroup allows you to scale based on capacity and performance needs.

ActiveIQ provides a SaaS-based portal to manage both NetApp HCI and ONTAP based systems centrally.

NetApp HCI can be easily repurposed as workload demands change to provide a robust return on your investment. For more information, or to see a demonstration, contact your NetApp representative.

Appendix A: NVIDIA nVector Tool Configuration Information

The sample Config.JSON for the SPECviewperf test is as follows:

```
{
        "vcip":"x.x.x.x",
        "vcuser":"administrator@vsphere.local",
        "vcpwd":"********",
        "esxuser":"root",
        "esxpwd":"*******",
        "domain":"HCIEUC",
        "domainadmin":"Administrator",
        "domainadminpwd":"********",
        "desktopuserprefix":"SPEC",
        "desktoppwd":"********",
        "datacenter":"NetApp-HCI-Datacenter-01",
        "iterations":"1",
    "vdiversion":"1906",
    "vdiclientversion":"1906",
    "vgpumanager":"430.27",
    "vgpudriver":"431.02",
        "vdicluster":"H615C - GPU",
        "vdidatastore":"DS02",
        "clientcluster":"H615C - GPU",
        "clientdatastore":"DS01",
        "guestcustomizationspec":"Windows 10",
        "provisionedservers":"0",
        "scale":"3",
        "parentdesktopname":"QvDWS01",
        "desktopprefix":"SPECDesk",
        "parentclientname":"NCC",
        "clientprefix":"CLI",
        "vcpu":"18",
        "vram":"16384",
        "gridprofile":"t4-16q",
      "workload": {
        "workloadname":"specviewperf",
        "workloadexecutable": "specviewperf.exe",
        "workloadpath":"C:\\Users\\Public\\Downloads",
        "workloadconfigname": "specviewperf.cfg",
        "workloadconfigpath":"C:\\Users\\Public\\Downloads",
        "resultsfolder":"SPECViewPerf_results",
        "params":""
        },
        "protocol":"hdx3dpro",
        "frameratelimiter":"0",
        "vsync":"0",
        "fps":"0",
        "viewport":"0",
        "viewportsettings":"OpenGL",
        "vgpupolicy":"breadthfirst",
        "resolution":"1920x1080",
        "numofscreens":"1",
        "logesxtop":"1",
        "lognvsmi":"1",
        "logwinperf":"0",
        "logpowerstats":"0",
        "logclockstats":"0",
        "logvgpu":"1",
    "cleanup":"0",
    "staggeredstartinterval":"0",
    "imagequality":"0",
        "hypervisor":"esx",
    "isdirectconnect":"0",
        "vditype":"XenDesktop",
        "broker":"cdc01.hcieuc.demo",
        "schedulingpolicy": "0x00",
        "gputype" : "nvidia",
        "enduserlatency" : "1"
}
```

The sample Config.JSON for the Knowledge Worker test is as follows:

```
{
    "vcip":"x.x.x.x",
    "vcuser":"administrator@vsphere.local",
    "vcpwd":"********",
    "esxuser":"root",
    "esxpwd":"**********",
    "domain":"HCIEUC",
    "domainadmin":"Administrator",
    "domainadminpwd":"********",
    "desktopuserprefix":"SPEC",
    "desktoppwd":"********",
    "datacenter":"NetApp-HCI-Datacenter-01",
    "iterations":"",
    "vdiversion":"1909",
    "vdiclientversion":"1909",
    "vgpumanager":"430.27",
    "vgpudriver":"431.02",
    "vdicluster":"H615C - GPU",
    "vdidatastore":"DS02",
    "clientcluster":"NetApp-HCI-Cluster-03",
    "clientdatastore":"DS03",
    "guestcustomizationspec":"Windows 10",
    "provisionedservers":"0",
    "scale":"24",
    "parentdesktopname":"4KTemplate",
    "desktopprefix":"H615C",
    "parentclientname":"nvClient4K",
    "clientprefix":"End",
    "vcpu":"4",
    "vram":"8192",
    "gridprofile":"",
      "workload": {
        "workloadname":"KnowledgeWorker",
        "workloadexecutable": "kwMain.exe",
        "workloadpath":"C:\\Users\\Public\\Downloads",
        "workloadconfigname": "",
        "workloadconfigpath":"",
        "resultsfolder":"KW_results",
        "params":"172.21.146.29 1 XenDesktop None 2"
        },
    "protocol":"hdx3dpro",
    "frameratelimiter":"0",
    "vsync":"0",
    "fps":"0",
    "viewport":"0",
    "viewportsettings":"OpenGL",
    "vgpupolicy":"breadthfirst",
    "resolution":"3840x2160",
    "numofscreens":"2",
    "logesxtop":"1",
    "lognvsmi":"1",
    "logwinperf":"0",
    "logpowerstats":"0",
    "logclockstats":"0",
    "logvgpu":"1",
    "cleanup":"0",
    "staggeredstartinterval":"35",
    "imagequality":"1",
    "hypervisor":"esx",
    "isdirectconnect":"0",
    "vditype":"XenDesktop",
    "broker":"cdc01.hcieuc.demo",
    "schedulingpolicy": "0x00",
    "gputype" : "nvidia",
    "enduserlatency" : "1"
}
```

Where to Find Additional Information

To learn more about the information that is described in this document, review the following documents and/or websites.

NetApp

- NetApp HCI Theory of Operations
  https://www.netapp.com/us/media/wp-7261.pdf
- VMware End-User Computing with NetApp HCI and NVIDIA GPUs
  https://www.netapp.com/us/media/nva-1129-design.pdf
- NetApp HCI for End-User Computing with VMware and NVIDIA GPUs
  https://www.netapp.com/us/media/nva-1129-deploy.pdf
- NetApp HCI for Virtual Desktop Infrastructure with VMware Horizon 7
  https://www.netapp.com/us/media/tr-4792.pdf
- NetApp Cloud Services
  https://cloud.netapp.com/home

NVIDIA

- NVIDIA Tesla GPUs for virtualization
  https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/solutions/resources/documents1/tesla-gpu-linecard-virtualization-us-nvidia-669786-r7.pdf
- NVIDIA GRID: Deployment Best Practices for the Digital Workspace
  http://images.nvidia.com/content/pdf/grid/guides/vgpu-profile-sizing-guidance-for-windows-10.pdf
- Virtual Workstation 101
  https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/solutions/resources/documents1/Virtual-Workstation-101-Technology-Brief.pdf
- NVIDIA Virtual GPU Packaging, Pricing and Licensing
  https://images.nvidia.com/content/grid/pdf/Virtual-GPU-Packaging-and-Licensing-Guide.pdf
- NVIDIA RTX
  https://www.nvidia.com/en-us/design-visualization/technologies/rtx/
- NVIDIA T4 for Virtualization
  https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/solutions/resources/documents1/TechBrief_T4.pdf
- NVIDIA GRID Deployment Guide for Citrix XenDesktop 7.12 on VMware vSphere 6
  http://images.nvidia.com/content/pdf/grid/guides/NVIDIA_GRID_Citrix_XexDesktop_Deployment_Guide.pdf
- NVIDIA Management and Monitoring
  https://www.nvidia.com/en-us/data-center/virtualization/it-management/
- NVIDIA GPU Cloud
  https://ngc.nvidia.com/catalog/landing

Citrix

- Citrix Tech Zone
  https://docs.citrix.com/en-us/tech-zone
- Citrix Virtual Apps and Desktops
  https://www.citrix.com/products/citrix-virtual-apps-and-desktops/
- Technical Overview of Citrix Application Layering
  https://www.citrix.com/content/dam/citrix/en_us/documents/white-paper/citrix-application-layering-technical-overview.pdf

- Citrix VDI Handbook and Best Practices
  https://docs.citrix.com/en-us/xenapp-and-xendesktop/7-15-ltsr/citrix-vdi-best-practices.html
- Citrix Virtual Apps and Desktops – Technical Overview
  https://docs.citrix.com/en-us/xenapp-and-xendesktop/7-15-ltsr/technical-overview.html
- Citrix Provisioning Services product overview
  https://docs.citrix.com/en-us/provisioning/7-15/overview.html
- Choosing the Provisioning Model for Image Management
  https://docs.citrix.com/en-us/tech-zone/design/design-decisions/image-management.html
- Citrix Blogs
  https://www.citrix.com/blogs/
- Citrix Cloud Services
  https://www.cloud.com/cloud-services.html

LakeSide

- GPU Assessment
  https://www.lakesidesoftware.com/assessments/nvidia

Version History

Version	Date	Document Version History
Version 1.0	October 2019	Initial draft by Suresh Thoppay
Version 2.0	March 2020	Added deployment steps, Login VSI, and NVIDIA nVector test results for a single server.
Version 3.0	June 2020	Added single-server login VSI test results for RDSH and Citrix Provisioning.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

Trademark Information

NVA-1140-0620