White Paper
Public

The bridge to possible

Integrate Cisco Intersight Managed Cisco UCS X-Series with NVIDIA GRID and VMware

Cisco Intersight Managed Cisco UCS X-Series Modular System with NVIDIA GRID and VMware

Last updated: October 3, 2022

Contents

This document introduces methods for integrating the Cisco UCS® X-Series Modular System with NVIDIA GRID T4 and A16 cards with VMware vSphere and Horizon products for virtual desktop infrastructure (VDI) that is ready for the future.

Cisco UCS X-Series with the Cisco Intersight platform

Cisco delivers the future-ready Cisco UCS X-Series Modular System (Figure 1) to support the deployment and management of virtual desktops and applications. Cisco UCS X-Series computing nodes combine the efficiency of a blade server and the flexibility of a rack server. The X-Series is the first system exclusively managed through Cisco Intersight™ software. This feature makes your VDI cloud operated, with complete hardware and software lifecycle management from a single interface regardless of location. The Cisco UCS X-Series with the Cisco Intersight platform carries forward the main capabilities Cisco has developed through its long history of virtual desktop deployments:

- Availability: VDI must always be on and available anytime and anywhere. Designed for high availability, the X-Series can sustain single failures and keep running. The Cisco Intersight platform offers an always-on connection to the Cisco® Technical Assistance Center (TAC), constantly monitoring your environment to help identify configuration or operational issues before they become problems. As the number of users increases, you can easily scale up and add new capabilities, without downtime.
- User productivity: Application access fuels user productivity. Support productivity with fast application deployment and upgrades, simplified application and desktop patching and management, and application migration support. Provision desktops instantly when new staff is hired.
- Flexible design: VDI applications range from call-center workers (task users) accessing a few applications, to professionals (knowledge workers) accessing many applications, to power users accessing graphics-intensive applications. The X-Series is flexible enough to house a complete VDI solution within the chassis (depending on storage needs) or as part of a converged infrastructure solution such as FlexPod or FlashStack. We test and document these solutions in Cisco Validated Designs, where every element is fully documented. The system can be purchased or paid for as it is used through Cisco Plus.
- Future ready: Cisco UCS X-Fabric technology is designed to accommodate new devices and interconnects as they become available, including the capability to extend the server's PCIe connectivity to attach additional graphics processing unit (GPU) accelerators for an even more compelling virtual desktop experience.

The Cisco UCS X-Series with the Cisco Intersight platform is future-ready, foundational infrastructure that can support your virtual desktops and applications today and for years to come. With Cisco Intersight software, you can simplify management of all your locations: data center, remote office, branch office, industrial, and edge.

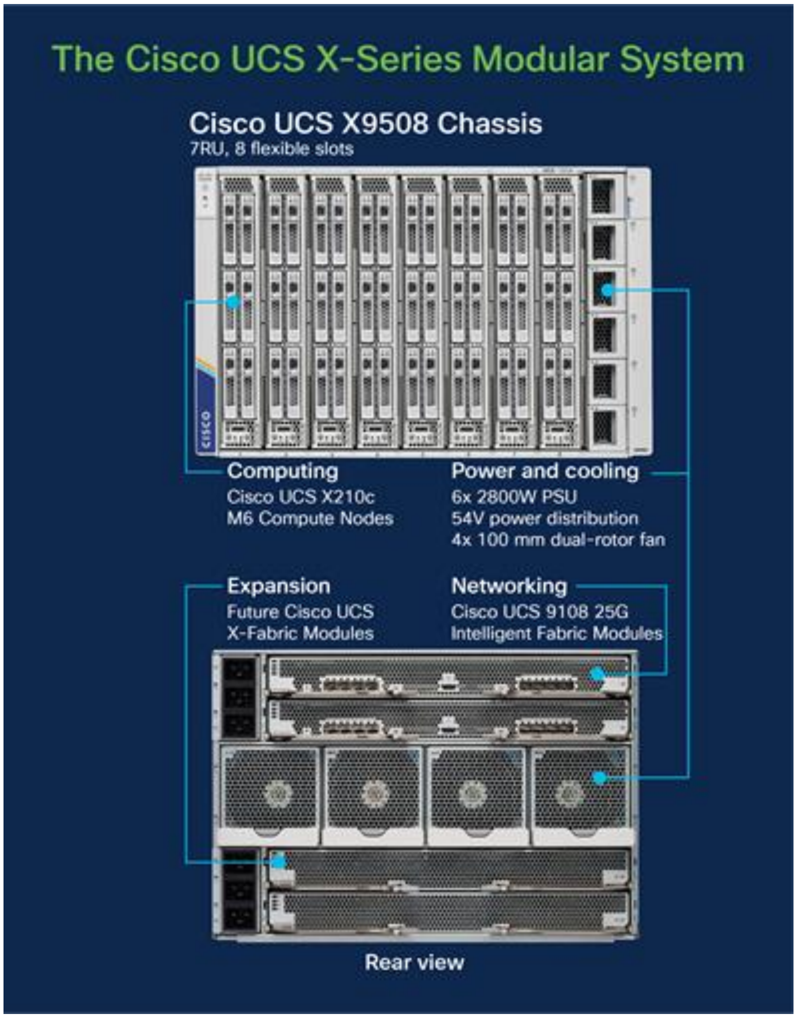


Figure 1.
Cisco UCS X-Series Modular System

Cisco UCS X-Series Modular System

The Cisco UCS X-Series includes the Cisco UCS 9508 X-Series Chassis, X210c M6 Compute Node, and X440p PCIe Node, discussed in this section.

Cisco UCS 9508 X-Series Chassis

The Cisco UCS 9508 X-Series Chassis (Figure 2) has the following main features:

- The seven-rack-unit (7RU) chassis has eight front-facing flexible slots. These can house a combination of computing nodes and a pool of future I/O resources, which may include GPU accelerators, disk storage, and nonvolatile memory.
- Two Cisco UCS 9108 Intelligent Fabric Modules (IFMs) at the top of the chassis connect the chassis to upstream Cisco UCS 6400 Series Fabric Interconnects. Each IFM provides up to 100 Gbps of unified fabric connectivity per computing node.
- Eight 25-Gbps SFP28 uplink ports carry unified fabric management traffic to the Cisco Intersight cloud-operations platform, Fibre Channel over Ethernet (FCoE) traffic, and production Ethernet traffic to the fabric interconnects.

- At the bottom are slots ready to house future I/O modules that can flexibly connect the computing modules with I/O devices. Cisco calls this connectivity Cisco UCS X-Fabric technology, with “X” as a variable that can evolve with new technology developments.
- Six 2800-watt (W) power supply units (PSUs) provide 54 volts (V) of power to the chassis with N, N+1, and N+N redundancy. A higher voltage allows efficient power delivery with less copper and reduced power loss.
- Four 100-mm dual counter-rotating fans deliver industry-leading airflow and power efficiency. Optimized thermal algorithms enable different cooling modes to best support the network environment. Cooling is modular so that future enhancements can potentially handle open- or closed-loop liquid cooling to support even higher-power processors.

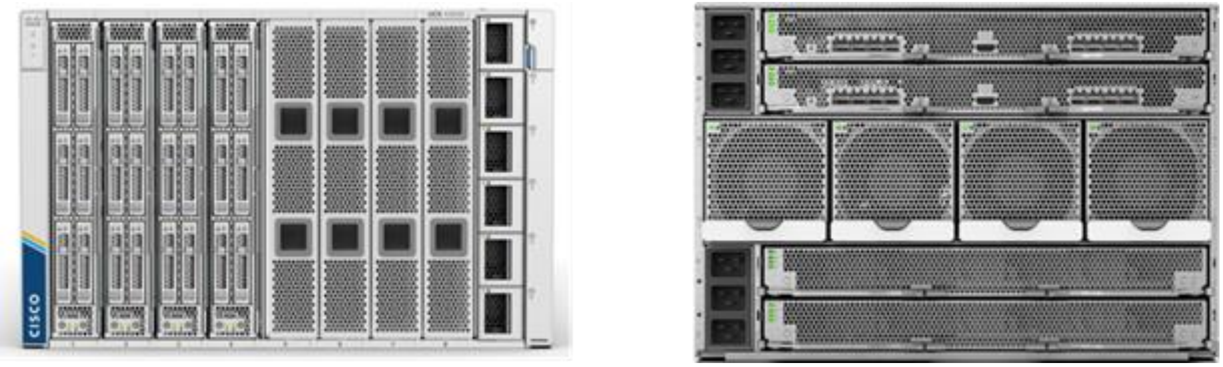


Figure 2.
Cisco UCS 9508 X-Series Chassis, front (left) and back (right)

Cisco UCS X210c M6 Compute Node

The Cisco UCS X210c M6 Compute Node (Figure 3) is the first computing device to integrate into the Cisco UCS X-Series Modular System. Up to eight computing nodes can reside in the 7RU Cisco UCS X9508 Chassis, offering one of the highest densities of computing, I/O, and storage resources per rack unit in the industry.

The Cisco UCS X210c M6 provides these main features:

- CPU: The CPU can contain up to two Third Generation (3rd Gen) Intel® Xeon® Scalable processors, with up to 40 cores per processor and 1.5 MB of Level 3 cache per core.
- Memory: The node can house up to thirty-two 256-GB DDR4 3200-megahertz (MHz) DIMMs, for up to 8 TB of main memory. Configuring up to sixteen 512-GB Intel Optane™ persistent memory DIMMs can yield up to 12 TB of memory.
- Storage: The node can include up to six hot-pluggable, solid-state disks (SSDs), or Non-Volatile Memory Express (NVMe) 2.5-inch drives with a choice of enterprise-class RAID or passthrough controllers with four lanes each of PCIe Gen 4 connectivity and up to two M.2 SATA drives for flexible boot and local storage capabilities.
- Modular LAN-on-motherboard (mLOM) virtual interface card (VIC): The Cisco UCS VIC 14425 can occupy the server’s mLOM slot, enabling up to 50 Gbps of unified fabric connectivity to each of the chassis IFMs for 100 Gbps of connectivity per server.
- Optional mezzanine VIC: The Cisco UCS VIC 14825 can occupy the server’s mezzanine slot at the bottom of the chassis. This card’s I/O connectors link to Cisco UCS X-Fabric technology that is

planned for future I/O expansion. An included bridge card extends this VIC's two 50-Gbps network connections through IFM connectors, bringing the total bandwidth to 100 Gbps per fabric: a total of 200 Gbps per server.

- Security: The server supports an optional Trusted Platform Module (TPM). Additional features include a secure boot field programmable gate array (FPGA) and Anti-Counterfeit Technology 2 (ACT2) anticounterfeit provisions.

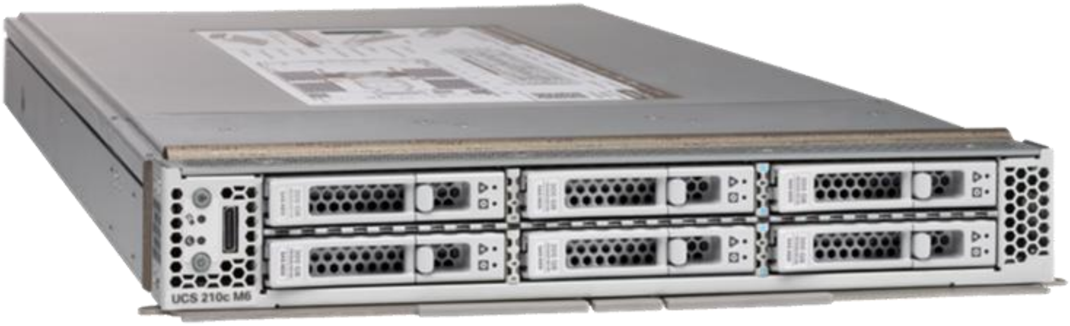


Figure 3.
Cisco UCS X210c M6 Compute Node

Cisco UCS X440p PCIe Node

The Cisco UCS X440p PCIe Node (Figure 4) is the first PCIe resource node to integrate into the Cisco UCS X-Series Modular System. The Cisco UCS X9508 Chassis has eight node slots, up to four of which can be X440p PCIe nodes when paired with a Cisco UCS X210c M6 Compute Node. The X440p PCIe Node supports two x16 full-height, full-length dual-slot PCIe cards, or four x8 full-height, full-length single-slot PCIe cards and requires both Cisco UCS 9416 X-Fabric Modules for PCIe connectivity. This configuration provides up to 16 GPUs per chassis to accelerate your applications with the X440p PCIe Nodes. If your application needs even more GPU acceleration, up to two additional GPUs can be added on each X210c M6 Compute Node.

The Cisco UCS X440p supports the following GPU options:

- NVIDIA A100 Tensor Core GPU
- NVIDIA A16 GPU
- NVIDIA A40 GPU
- NVIDIA T4 Tensor Core GPU

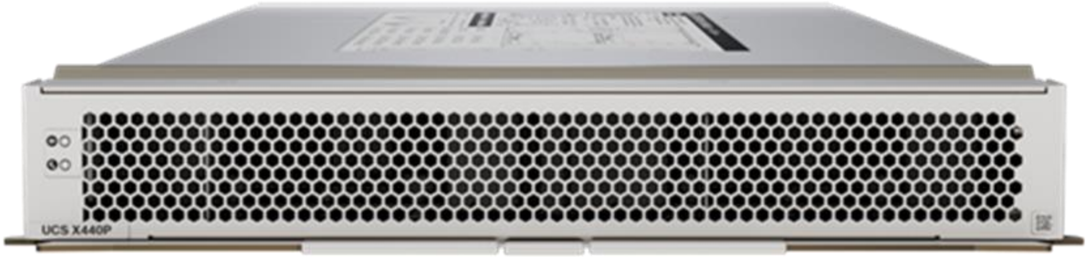


Figure 4.
Cisco UCS X440p PCIe Node

Why use NVIDIA GRID vGPU for graphics deployments on VMware Horizon

The NVIDIA GRID vGPU allows multiple virtual desktops to share a single physical GPU, and it allows multiple GPUs to reside on a single physical PCI card. All provide the 100 percent application compatibility of Virtual Dedicated Graphics Acceleration (vDGA) pass-through graphics, but with a lower cost because multiple desktops share a single graphics card simultaneously. With VMware Horizon, you can centralize, pool, and more easily manage traditionally complex and expensive distributed workstations and desktops. Now all your user groups can take advantage of the benefits of virtualization.

The GRID vGPU capability brings the full benefits of NVIDIA hardware-accelerated graphics to virtualized solutions. This technology provides exceptional graphics performance for virtual desktops equivalent to that of PCs with an onboard graphics processor. The GRID vGPU uses the industry's most advanced technology for sharing true GPU hardware acceleration among multiple virtual desktops—without compromising the graphics experience. Application features and compatibility are exactly the same as they would be at the user's desk.

With GRID vGPU technology, the graphics commands of each virtual machine are passed directly to the GPU, without translation by the hypervisor. By allowing multiple virtual machines to access the power of a single GPU in the virtualization server, enterprises can increase the number of users with access to true GPU-based graphics acceleration on virtual machines. The physical GPU in the server can be configured with a specific vGPU profile. Organizations have a great deal of flexibility in how best to configure their servers to meet the needs of various types of end users.

vGPU support allows businesses to use the power of the NVIDIA GRID technology to create a whole new class of virtual machines designed to provide end users with a rich, interactive graphics experience.

NVIDIA cards

For desktop virtualization applications, the NVIDIA T4 and A16 cards are optimal choices for high-performance graphics VDI (Table 1). For additional information about selecting GPUs for virtualized workloads, see the NVIDIA Technical Brief "NVIDIA Virtual GPU Positioning."

Table 1. NVIDIA graphic cards specifications

Hardware specification	T4	A16
GPU architecture	NVIDIA Turing	NVIDIA Ampere architecture
GPU memory	16-GB GDDR6	4 x 16-GB GDDR6
Memory bandwidth	300 GB/sec	4 x 200-GBps
Error-correcting code (ECC)	Yes	Yes
NVIDIA CUDA cores	2560	4 x 1280
NVIDIA Tensor cores	320	4 x 40 (Third generation)
NVIDIA RT cores	40	4 x 10 (Second generation)
FP32 | TF32 | TF32 (TFLOPS)	8.1	4 x 4.5 | 4 x 9 | 4 x 18
FP16 | FP16 (TFLOPS)	16.2	4 x 17.9 | 4 x 35.9

Hardware specification	T4	A16
INT8 | INT8 (TOPS)	130	4 x 35.9 | 4 x 71.8
INT4 | INT4 (TOPS)	260	
System interface	x16 PCIe Gen3	PCIe Gen4 (x16)
Maximum power consumption	70W	250W
Thermal solution	Passive	Passive
Form factor	Low-profile PCIe	Full height, full length (FHFL) dual slot
vGPU software support	NVIDIA Virtual PC (vPC), NVIDIA Virtual Applications (vApps), NVIDIA RTX Virtual Workstation (vWS), NVIDIA AI Enterprise, and NVIDIA Virtual Compute Server (vCS)	NVIDIA vPC, NVIDIA vApps, NVIDIA RTX vWS, NVIDIA AI Enterprise, and NVIDIA vCS
Computing APIs	CUDA, NVIDIA TensorRT, and ONNX	CUDA, DirectCompute, OpenCL, and OpenACC

Reference architecture

Figure 5. provides an overview of the environment used in testing.

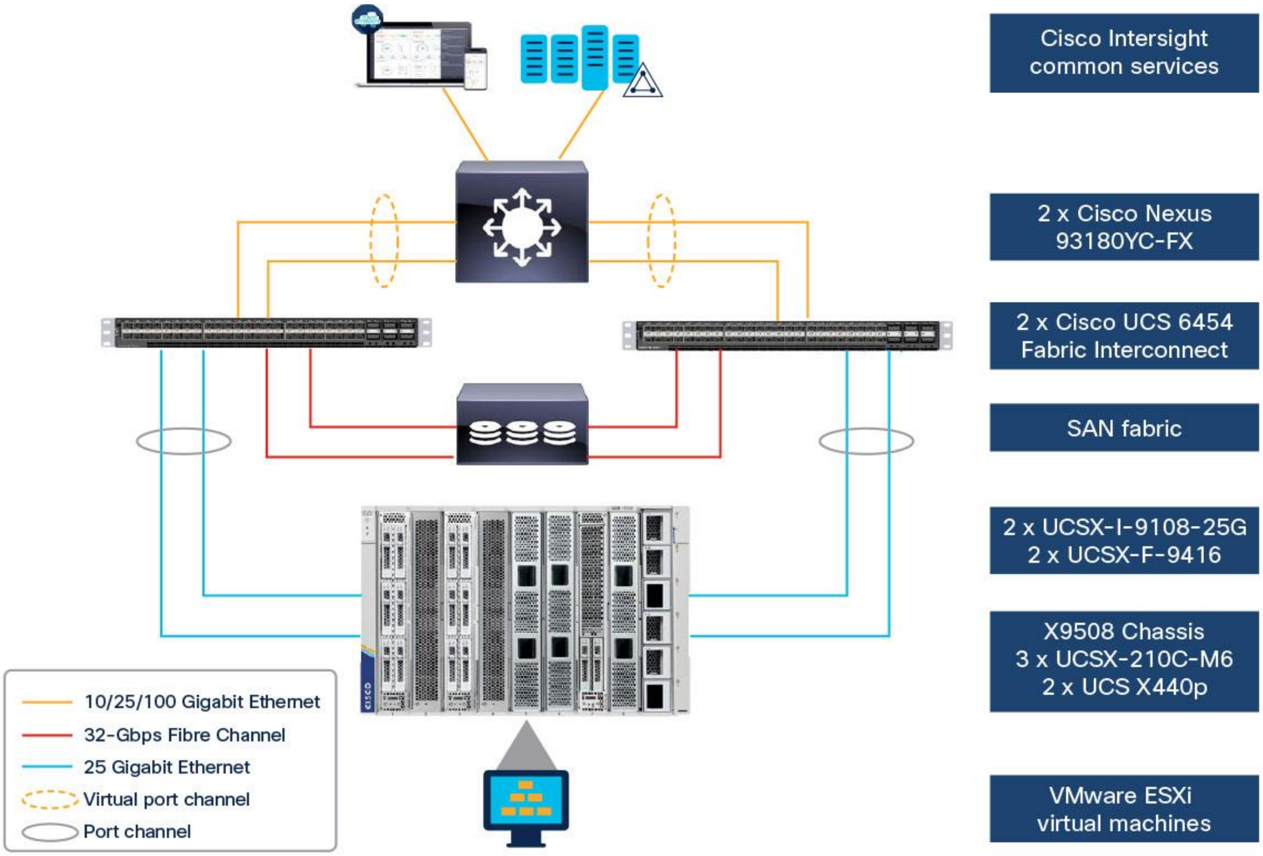


Figure 5.
Topology diagram

The following hardware components are used:

- Cisco UCS X210c M6 Compute Node (two Intel Xeon Scalable Platinum 8358 CPUs at 2.60 GHz) with 1 TB of memory (64 GB x 16 DIMMs at 3200 MHz)
- Cisco UCS X210c M6 Compute Node (two Intel Xeon Scalable Gold 6348 CPUs at 2.60 GHz) with 1 TB of memory (64 GB x 24 DIMMs at 3200 MHz)
- Cisco UCS X440p PCIe Node
- Cisco UCS VIC 14425 4 x 25-Gbps mLOM for X-Series compute node
- Cisco UCS X210c Compute Node front mezzanine card to support up to two NVIDIA T4 GPUs and two NVMe drives
- Cisco UCS PCI mezzanine card for X-Fabric*
- Two Cisco UCS 6454 Fabric interconnects
- NVIDIA Tesla T4 cards
- NVIDIA Ampere A16 cards
- Two Cisco Nexus® 93180YC-FX Switches (optional access switches)

The following software components are used:

- Cisco UCS Firmware Release 4.2(2a)
- VMware ESXi 7.0 Update 2a (17867351) for VDI hosts
- VMware Horizon 2111
- Microsoft Windows 10 64-bit
- Microsoft Windows 11 64-bit
- Microsoft Office 2021
- NVIDIA GRID 7.0 software and licenses:
- NVD-VGPU_510.73.06-1OEM.702.0.0.17630552_19796074.zip
- 512.78_grid_win10_win11_server2016_server2019_server2022_64bit_international.exe

* This rear mezzanine card enables connectivity from the X210c Compute Node to the X440p PCIe Node.

Configure Cisco UCS X-Series Modular System

This section describes the Cisco UCS X-Series Modular System configuration.

Hardware configuration

The Cisco UCS X-Series Modular System supports NVIDIA GPUs in two ways:

- Up to two NVIDIA T4 GPUs can be added to the Cisco UCS X210c M6 server's optional mezzanine VIC (UCSX-GPU-T4-MEZZ), shown in Figure 6.

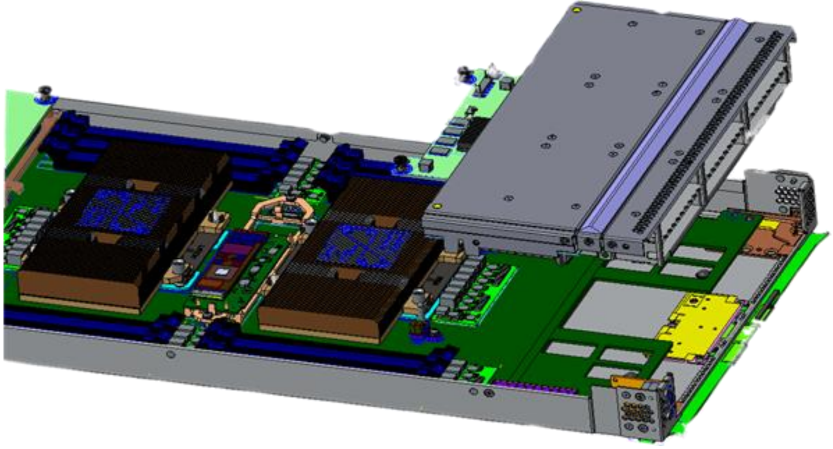


Figure 6.
UCSX-GPU-T4-MEZZ placement

- The Cisco UCS X440p PCIe Node provides two (riser A) or four (riser B) PCIe slots connected to an adjacent Cisco UCS X210c M6 Compute Node to support NVIDIA GPUs (Figures 7 and 8).

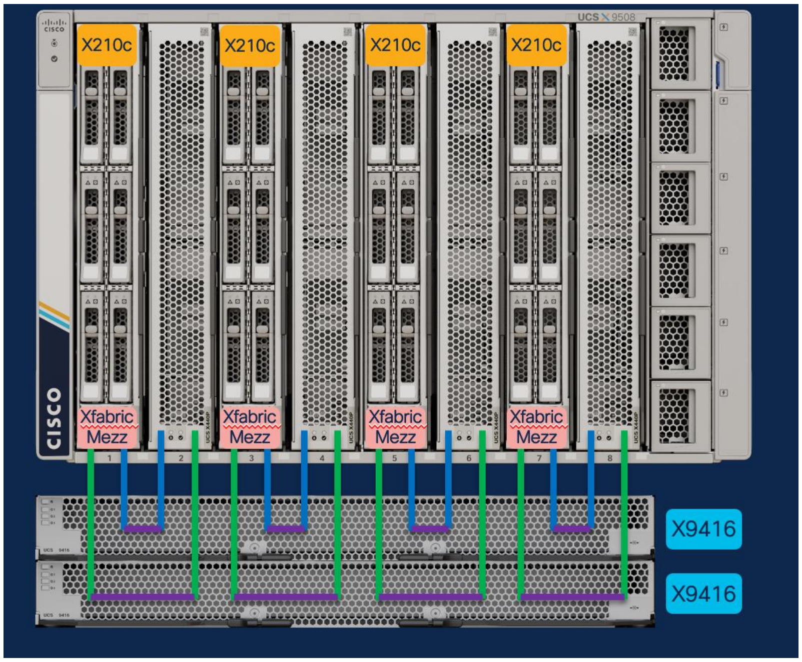


Figure 7.
Cisco UCS X440p PCIe Node placement

- Riser A: Supports up to two dual-width A16, A40, or A100 GPUs
- Riser B: Supports up to four single-width T4 GPUs

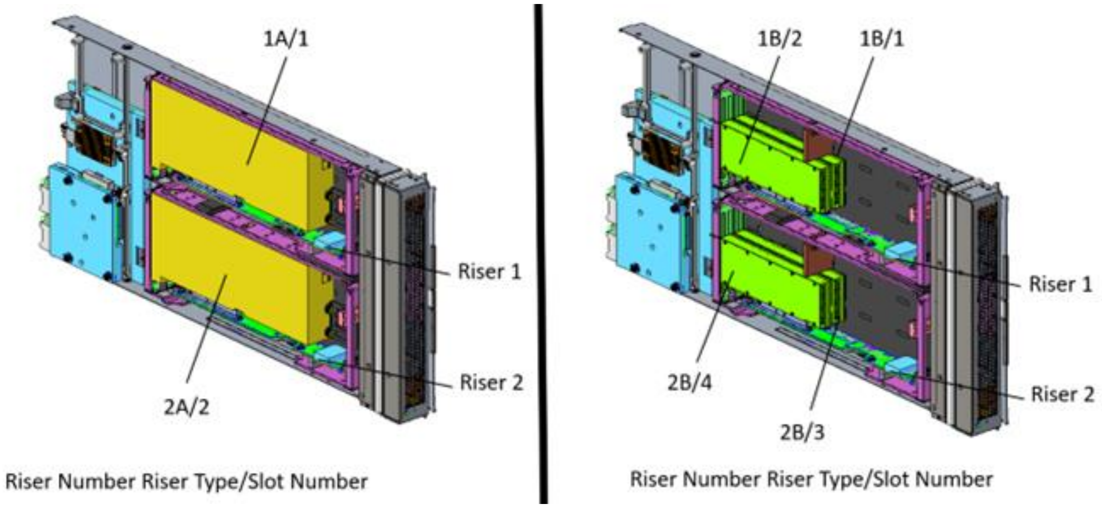

1A/1
Riser 1
Riser 2
2A/2
Riser Number Riser Type/Slot Number
1B/2
1B/1
Riser 1
Riser 2
2B/4
2B/3
Riser Number Riser Type/Slot Number

Figure 8.
GPU placement in risers A and B

Note: GPU models cannot be mixed on a server.

Cisco Intersight configuration

The Cisco Intersight platform provides an integrated and intuitive management experience for resources in the traditional data center and at the edge. Getting started with Cisco Intersight is quick and easy.

Figure 9. for configuring Cisco UCS with Cisco Intersight managed mode (IMM).

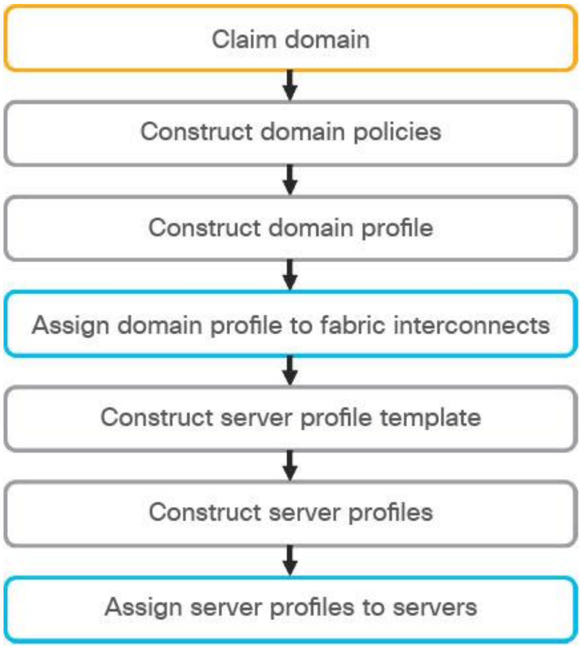
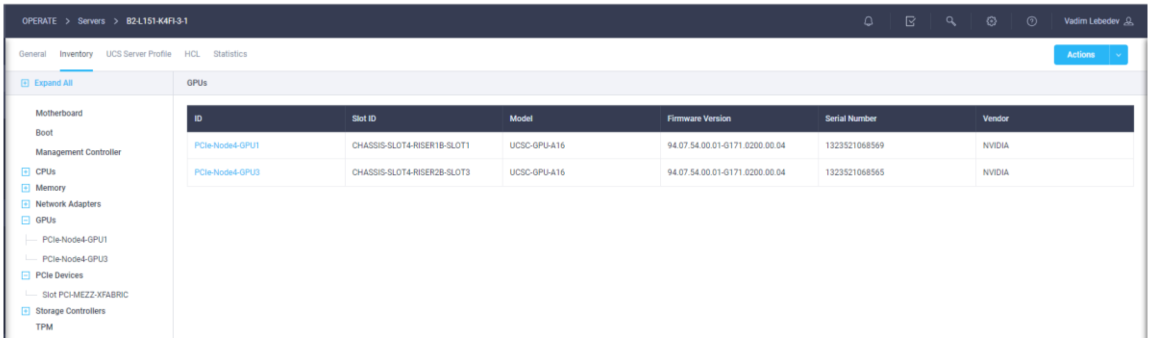

Claim domain
Construct domain policies
Construct domain profile
Assign domain profile to fabric interconnects
Construct server profile template
Construct server profiles
Assign server profiles to servers

Figure 9.
Cisco Intersight workflow

The detailed process is available in several guides available at Cisco.com:

- Getting Started with Intersight
- Deploy Cisco UCS X210c Compute Node with Cisco Intersight Management Mode for VDI

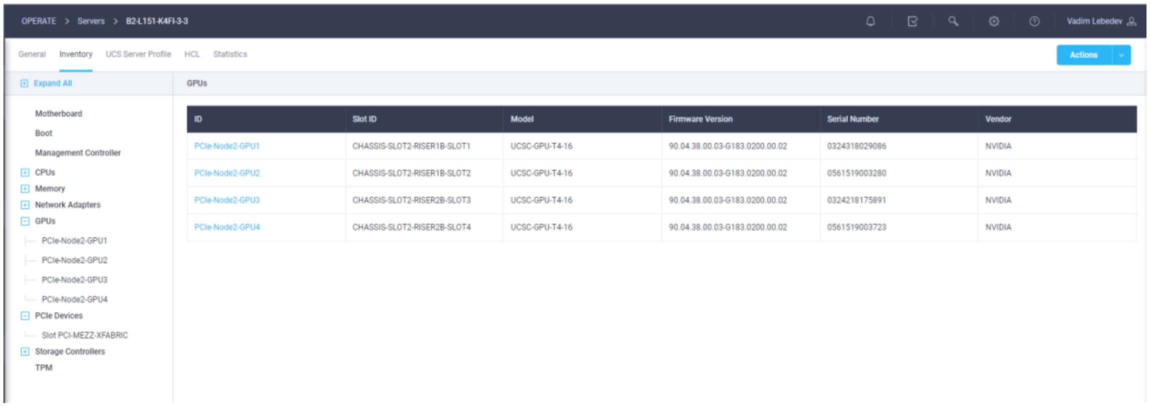
This environment uses one PCI node with two NVIDIA A16 cards, one PCI node with four NVIDIA T4 cards, and the compute node with two T4 cards (Figures 10, 11, and 12).



GPUs

ID	Slot ID	Model	Firmware Version	Serial Number	Vendor
PCIe-Node4-GPU1	CHASSIS-SLOT4-RISER1B-SLOT1	UCSC-GPU-A16	94.07.54.00.01-G171.0200.00.04	1323521068569	NVIDIA
PCIe-Node4-GPU3	CHASSIS-SLOT4-RISER2B-SLOT3	UCSC-GPU-A16	94.07.54.00.01-G171.0200.00.04	1323521068565	NVIDIA

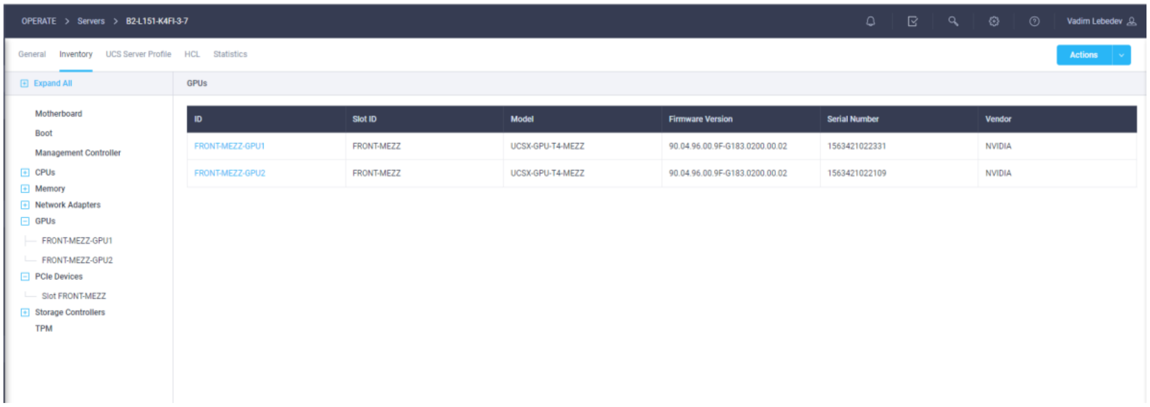
Figure 10.
Cisco Intersight views of the PCI node inventory with NVIDIA A16



GPUs

ID	Slot ID	Model	Firmware Version	Serial Number	Vendor
PCIe-Node2-GPU1	CHASSIS-SLOT2-RISER1B-SLOT1	UCSC-GPU-T4-16	90.04.38.00.03-G183.0200.00.02	0324318029086	NVIDIA
PCIe-Node2-GPU2	CHASSIS-SLOT2-RISER1B-SLOT2	UCSC-GPU-T4-16	90.04.38.00.03-G183.0200.00.02	0561519003280	NVIDIA
PCIe-Node2-GPU3	CHASSIS-SLOT2-RISER2B-SLOT3	UCSC-GPU-T4-16	90.04.38.00.03-G183.0200.00.02	0324218175891	NVIDIA
PCIe-Node2-GPU4	CHASSIS-SLOT2-RISER2B-SLOT4	UCSC-GPU-T4-16	90.04.38.00.03-G183.0200.00.02	0561519003723	NVIDIA

Figure 11.
Cisco Intersight views of the PCI node inventory with NVIDIA T4



GPUs

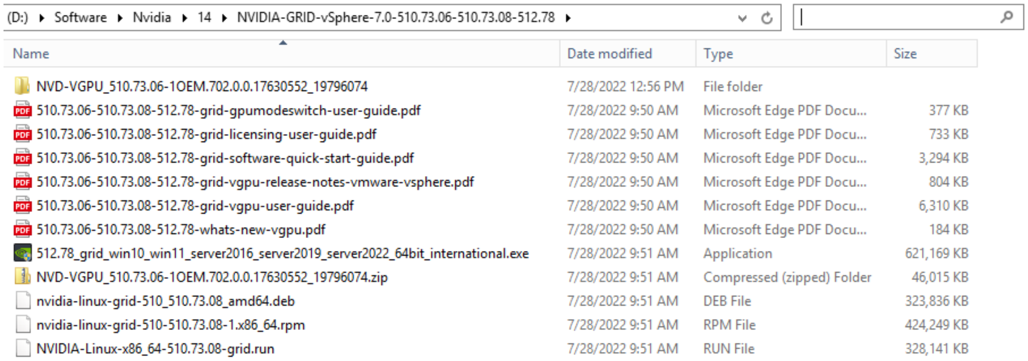
ID	Slot ID	Model	Firmware Version	Serial Number	Vendor
FRONT-MEZZ-GPU1	FRONT-MEZZ	UCSX-GPU-T4-MEZZ	90.04.96.00.9F-G183.0200.00.02	1563421022331	NVIDIA
FRONT-MEZZ-GPU2	FRONT-MEZZ	UCSX-GPU-T4-MEZZ	90.04.96.00.9F-G183.0200.00.02	1563421022109	NVIDIA

Figure 12.
Cisco Intersight views of the compute node inventory with NVIDIA T4

Install NVIDIA GRID software on the VMware ESXi host

This section summarizes the installation process for configuring an ESXi host and virtual machine for vGPU support (Figure 13).

Note: To be able to assign licenses to NVIDIA vGPUs, you must have at least one NVIDIA license server with an appropriate level of licenses.



(D:) ▸ Software ▸ Nvidia ▸ 14 ▸ NVIDIA-GRID-vSphere-7.0-510.73.06-510.73.08-512.78 ▸

Name	Date modified	Type	Size
NVD-VGPU_510.73.06-1OEM.702.0.0.17630552_19796074	7/28/2022 12:56 PM	File folder	
510.73.06-510.73.08-512.78-grid-gpumodeswitch-user-guide.pdf	7/28/2022 9:50 AM	Microsoft Edge PDF Docu...	377 KB
510.73.06-510.73.08-512.78-grid-licensing-user-guide.pdf	7/28/2022 9:50 AM	Microsoft Edge PDF Docu...	733 KB
510.73.06-510.73.08-512.78-grid-software-quick-start-guide.pdf	7/28/2022 9:50 AM	Microsoft Edge PDF Docu...	3,294 KB
510.73.06-510.73.08-512.78-grid-vgpu-release-notes-vmware-vsphere.pdf	7/28/2022 9:50 AM	Microsoft Edge PDF Docu...	804 KB
510.73.06-510.73.08-512.78-grid-vgpu-user-guide.pdf	7/28/2022 9:50 AM	Microsoft Edge PDF Docu...	6,310 KB
510.73.06-510.73.08-512.78-whats-new-vgpu.pdf	7/28/2022 9:50 AM	Microsoft Edge PDF Docu...	184 KB
512.78_grid_win10_win11_server2016_server2019_server2022_64bit_international.exe	7/28/2022 9:51 AM	Application	621,169 KB
NVD-VGPU_510.73.06-1OEM.702.0.0.17630552_19796074.zip	7/28/2022 9:51 AM	Compressed (zipped) Folder	46,015 KB
nvidia-linux-grid-510_510.73.08_amd64.deb	7/28/2022 9:51 AM	DEB File	323,836 KB
nvidia-linux-grid-510-510.73.08-1.x86_64.rpm	7/28/2022 9:51 AM	RPM File	424,249 KB
NVIDIA-Linux-x86_64-510.73.08-grid.run	7/28/2022 9:51 AM	RUN File	328,141 KB

Figure 13.
NVIDIA GRID vSphere 7.0 contents

1. Download the NVIDIA GRID GPU driver pack for VMware vSphere ESXi 7.0 U2.
2. Enable the ESXi shell and the Secure Shell (SSH) protocol on the vSphere host from the Troubleshooting Mode Options menu of the vSphere Configuration Console (Figure 14.).

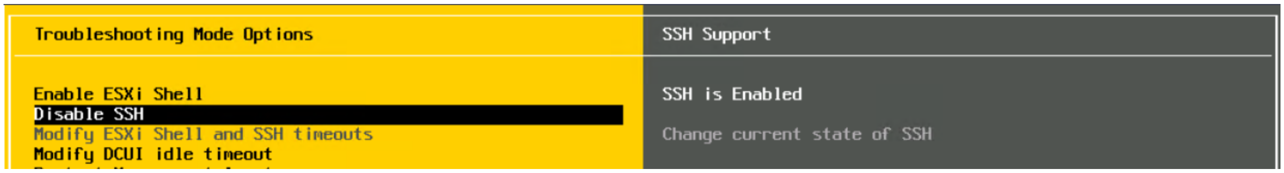

Troubleshooting Mode Options
Enable ESXi Shell
Disable SSH
Modify ESXi Shell and SSH timeouts
Modify DCUI idle timeout
SSH Support
SSH is Enabled
Change current state of SSH

Figure 14.
VMware ESXi Configuration Console

3. Upload the NVIDIA driver (vSphere Installation Bundle [VIB] file) to the /tmp directory on the ESXi host using a tool such as WinSCP. (Shared storage is preferred if you are installing drivers on multiple servers or using the VMware Update Manager.)
4. Log in as root to the vSphere console through SSH using a tool such as Putty.
5. The ESXi host must be in maintenance mode for you to install the VIB module. To place the host in maintenance mode, use this command:

```
esxcli system maintenanceMode set -enable true
```

```
Enter the following command to install the NVIDIA vGPU drivers:
```

```
esxcli software vib install -d /<path>/<filename>.zip
```

6. The command should return output similar to that shown here:

```
[root@K23-FCP-5:~] esxcli software vib install -d /vmfs/volumes/ESXTOP/NVD-
VGPU_510.73.06-1OEM.702.0.0.17630552_19796074.zip

Installation Result
```

```
 Message: Operation finished successfully.
Reboot Required: false
 VIBs Installed: NVIDIA_bootbank_NVIDIA-VMware_ESXi_7.0.2_Driver_510.73.06-
1OEM.702.0.0.17630552
 VIBs Removed:
 VIBs Skipped:
```

Note: Although the display shows “Reboot Required: false,” a reboot is necessary for the VIB file to load and for xorg to start.

7. Exit the ESXi host from maintenance mode and reboot the host by using the vSphere Web Client or by entering the following commands:

   ```
   #esxcli system maintenanceMode set -e false
   #reboot
   ```

8. After the host reboots successfully, verify that the kernel module has loaded successfully by entering the following command:

   ```
   esxcli software vib list | grep -i nvidia
   ```

   The command should return output similar to that shown here:

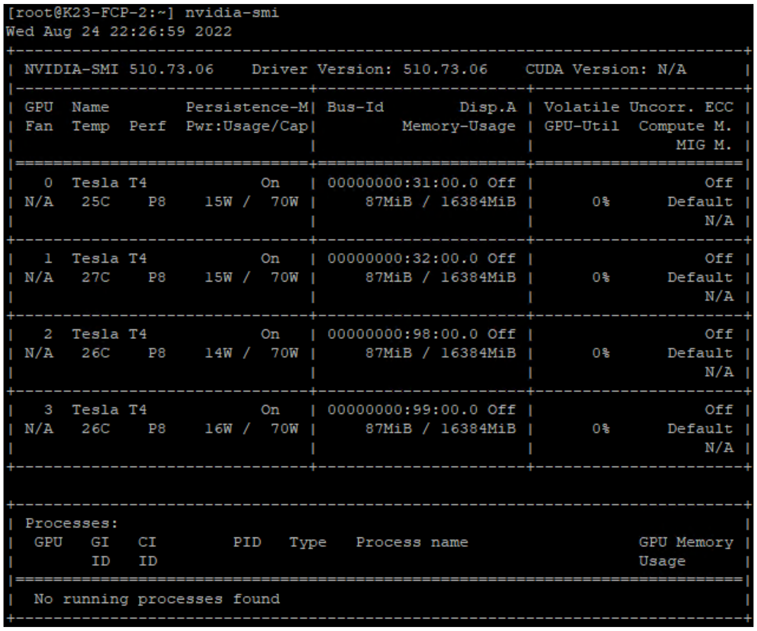
   ```
   [root@K23-FCP-5:~] esxcli software vib list | grep -i nvidia
   NVIDIA-VMware_ESXi_7.0.2_Driver  510.73.06-1OEM.702.0.0.17630552      NVIDIA
   VMwareAccepted    2022-08-04
   ```

   See the VMware knowledge base article for information about removing any existing NVIDIA drivers before installing new drivers: Installing and configuring the NVIDIA VIB on ESXi (2033434).

9. Confirm GRID GPU detection on the ESXi host. To determine the status of the GPU card’s CPU, the card’s memory, and the amount of disk space remaining on the card, enter the following command:

   ```
   nvidia-smi
   ```

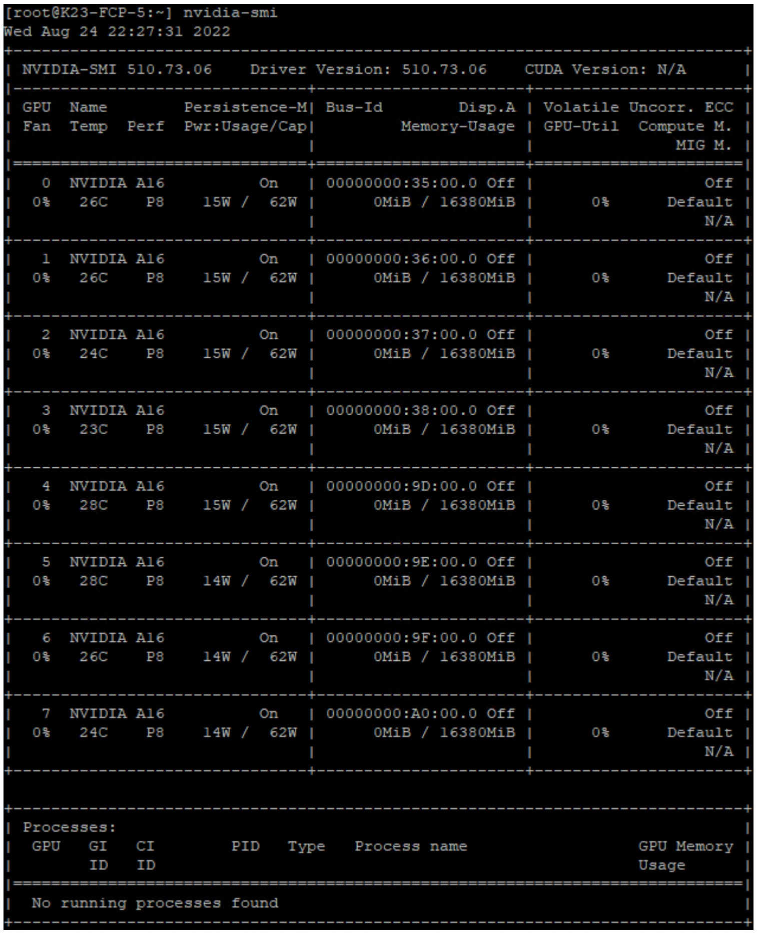
   The command should return output similar to that shown in Figure 15, 16, or 17, depending on the card used in your environment.



```
[root@K23-FCP-2:~] nvidia-smi
Wed Aug 24 22:26:59 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 510.73.06    Driver Version: 510.73.06    CUDA Version: N/A      |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla T4            On   | 00000000:31:00.0 Off |                  Off |
| N/A   25C    P8    15W /  70W |     87MiB / 16384MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  Tesla T4            On   | 00000000:32:00.0 Off |                  Off |
| N/A   27C    P8    15W /  70W |     87MiB / 16384MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   2  Tesla T4            On   | 00000000:98:00.0 Off |                  Off |
| N/A   26C    P8    14W /  70W |     87MiB / 16384MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla T4            On   | 00000000:99:00.0 Off |                  Off |
| N/A   26C    P8    16W /  70W |     87MiB / 16384MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

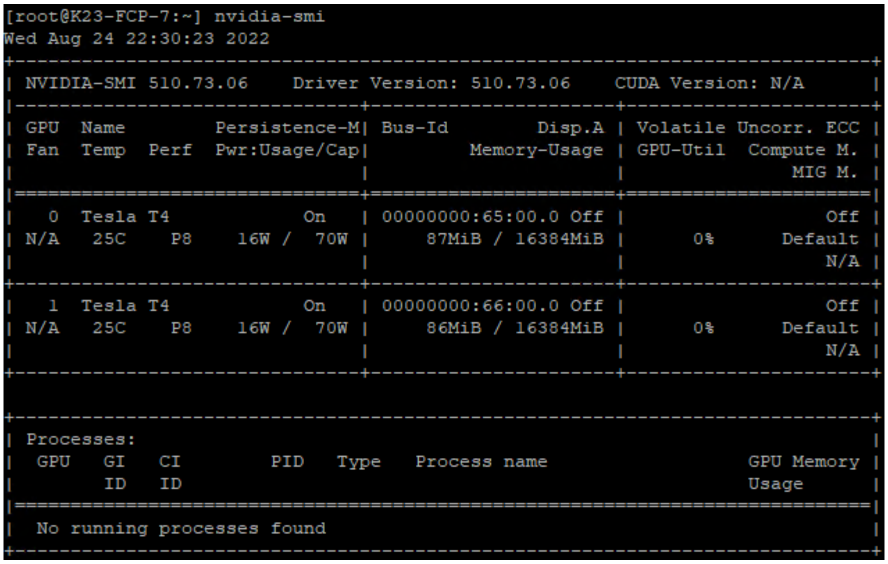
Figure 15.
NVIDIA System Management Interface inventory on VMware ESXi PCI node with NVIDIA T4



```
[root@K23-FCP-5:~] nvidia-smi
Wed Aug 24 22:27:31 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 510.73.06    Driver Version: 510.73.06    CUDA Version: N/A      |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  NVIDIA A16          On   | 00000000:35:00.0 Off |                  Off |
|  0%   26C    P8    15W /  62W |      0MiB / 16380MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  NVIDIA A16          On   | 00000000:36:00.0 Off |                  Off |
|  0%   26C    P8    15W /  62W |      0MiB / 16380MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   2  NVIDIA A16          On   | 00000000:37:00.0 Off |                  Off |
|  0%   24C    P8    15W /  62W |      0MiB / 16380MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  NVIDIA A16          On   | 00000000:38:00.0 Off |                  Off |
|  0%   23C    P8    15W /  62W |      0MiB / 16380MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   4  NVIDIA A16          On   | 00000000:9D:00.0 Off |                  Off |
|  0%   28C    P8    15W /  62W |      0MiB / 16380MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   5  NVIDIA A16          On   | 00000000:9E:00.0 Off |                  Off |
|  0%   28C    P8    14W /  62W |      0MiB / 16380MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   6  NVIDIA A16          On   | 00000000:9F:00.0 Off |                  Off |
|  0%   26C    P8    14W /  62W |      0MiB / 16380MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   7  NVIDIA A16          On   | 00000000:A0:00.0 Off |                  Off |
|  0%   24C    P8    14W /  62W |      0MiB / 16380MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Figure 16.
NVIDIA System Management Interface inventory on VMware ESXi PCI node with NVIDIA A16



```
[root@K23-FCP-7:~] nvidia-smi
Wed Aug 24 22:30:23 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 510.73.06    Driver Version: 510.73.06    CUDA Version: N/A      |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla T4            On   | 00000000:65:00.0 Off |                  Off |
| N/A   25C    P8    16W /  70W |     87MiB / 16384MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  Tesla T4            On   | 00000000:66:00.0 Off |                  Off |
| N/A   25C    P8    16W /  70W |     86MiB / 16384MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Figure 17.
NVIDIA System Management Interface inventory on VMware ESXi compute node with NVIDIA T4

Both cards, NVIDIA A16 and NVIDIA T4, support error correcting code (ECC) memory for improved data integrity.

However, the NVIDIA vGPU does not support ECC memory. If ECC memory is enabled, vGPU fails to start. Therefore, you must ensure that ECC memory is disabled on all GPUs if you are using the NVIDIA vGPU.

Disable ECC on your host by running this command:

```
nvidia-smi -e 0
```

Configure host graphics settings

After the GRID vGPU Manager has been installed, configure host graphics in vCenter on all hosts in the resource pool.

1. Select the ESXi host and click the Configure tab. From the list of options at the left, select Graphics under Hardware. Click Edit Host Graphics Settings (Figure 18).

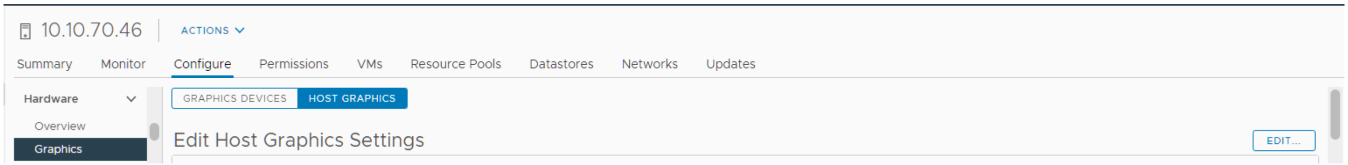
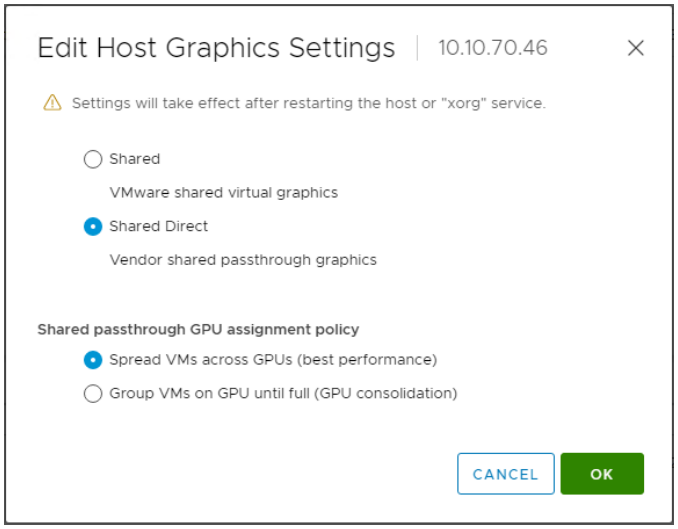


Figure 18.
Host graphics settings in vCenter

2. Select the following settings (Figure 19):
   - Shared Direct (Vendor shared passthrough graphics)
   - Spread VMs across GPUs (best performance)



Edit Host Graphics Settings | 10.10.70.46

⚠ Settings will take effect after restarting the host or "xorg" service.

○ Shared
VMware shared virtual graphics

● Shared Direct
Vendor shared passthrough graphics

Shared passthrough GPU assignment policy

● Spread VMs across GPUs (best performance)

○ Group VMs on GPU until full (GPU consolidation)

Figure 19.
Edit Host Graphics Settings window in vCenter

VMware ESXi host vGPU performance monitoring in vCenter

It is possible to monitor NVIDIA GPU use through the vSphere Web Client.

1. Use the vSphere Client to connect to the vCenter Server system.
2. Navigate to the ESXi host and open the Monitor tab.
3. Select the Performance tab and click Advanced.
4. Click Chart Options.
5. Under Chart Metrics, select GPU and under Counters select cards to visualize in the chart (Figure 20).

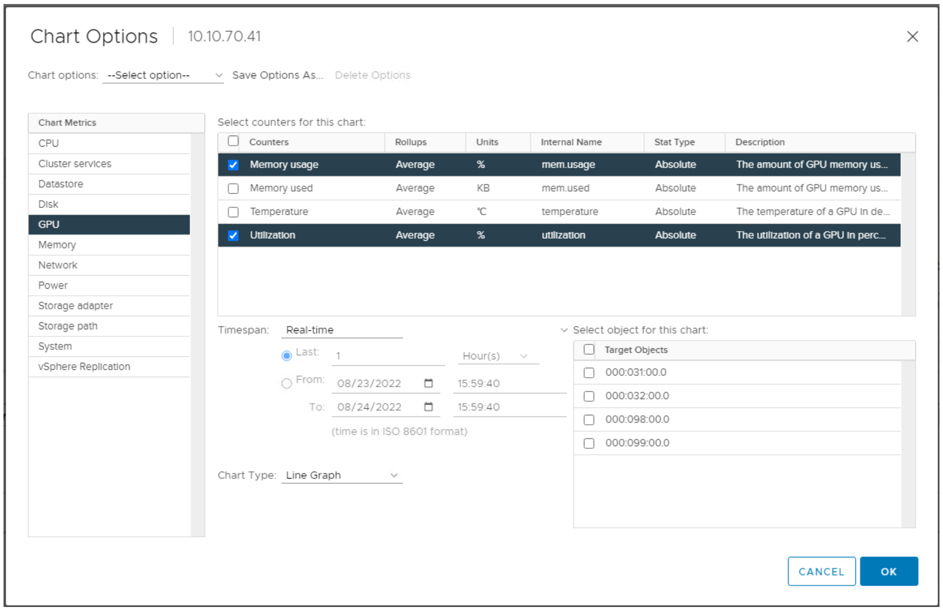


Chart Options | 10.10.70.41

Chart options: --Select option-- Save Options As... Delete Options

Chart Metrics
CPU
Cluster services
Datastore
Disk
GPU
Memory
Network
Power
Storage adapter
Storage path
System
vSphere Replication

Select counters for this chart:

	Counters	Rollups	Units	Internal Name	Stat Type	Description
☑	Memory usage	Average	%	mem.usage	Absolute	The amount of GPU memory us...
□	Memory used	Average	KB	mem.used	Absolute	The amount of GPU memory us...
□	Temperature	Average	℃	temperature	Absolute	The temperature of a GPU in de...
☑	Utilization	Average	%	utilization	Absolute	The utilization of a GPU in perc...

Timespan: Real-time

● Last: 1 Hour(s)

○ From: 08/23/2022 15:59:40

To: 08/24/2022 15:59:40

(time is in ISO 8601 format)

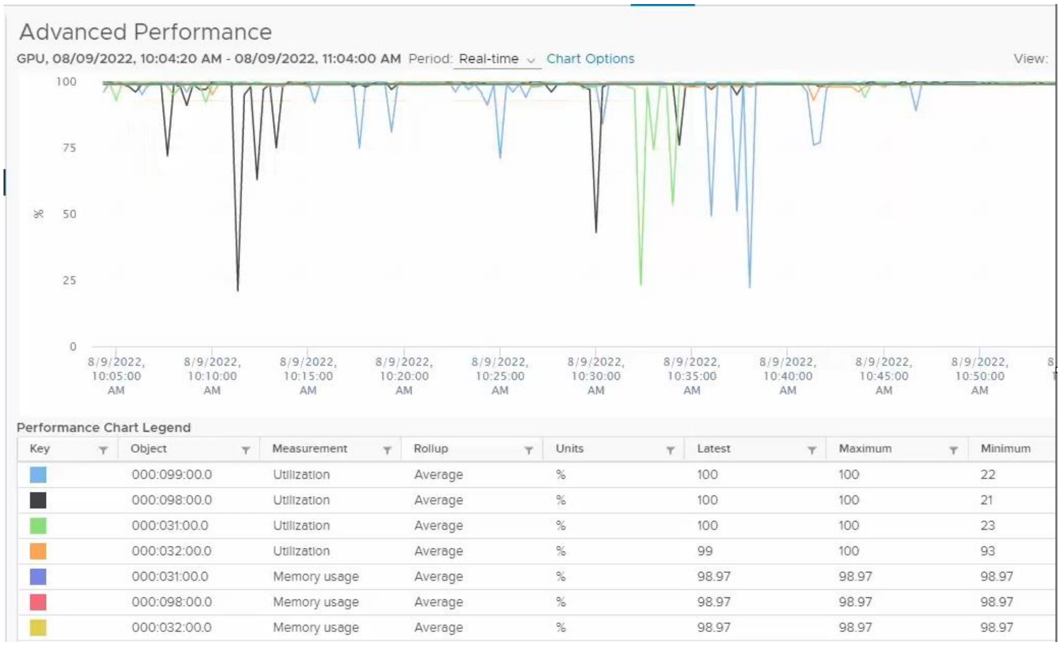
Chart Type: Line Graph

Select object for this chart:

	Target Objects
□	000:031:00.0
□	000:032:00.0
□	000:098:00.0
□	000:099:00.0

Figure 20.
Host GPU performance chart configuration in vSphere Web Client

6. The performance chart is displayed (Figure 21).



Advanced Performance

GPU, 08/09/2022, 10:04:20 AM - 08/09/2022, 11:04:00 AM Period: Real-time Chart Options View:

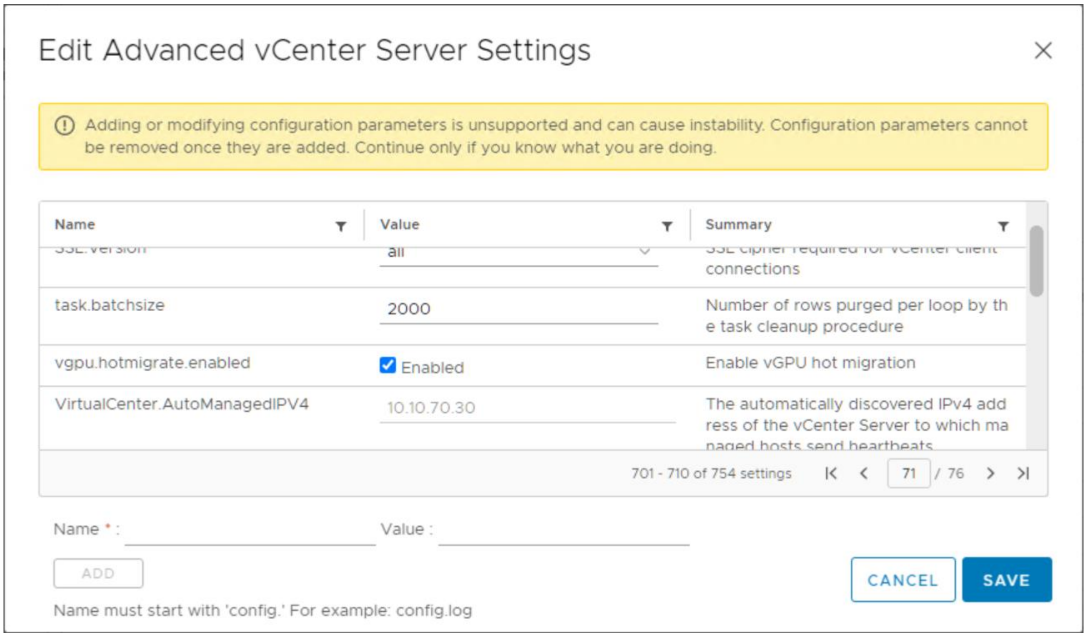
Performance Chart Legend

Key	Object	Measurement	Rollup	Units	Latest	Maximum	Minimum
🟦	000:099:00.0	Utilization	Average	%	100	100	22
⬛	000:098:00.0	Utilization	Average	%	100	100	21
🟩	000:031:00.0	Utilization	Average	%	100	100	23
🟧	000:032:00.0	Utilization	Average	%	99	100	93
🟪	000:031:00.0	Memory usage	Average	%	98.97	98.97	98.97
🟥	000:098:00.0	Memory usage	Average	%	98.97	98.97	98.97
🟨	000:032:00.0	Memory usage	Average	%	98.97	98.97	98.97

Figure 21.
Host GPU performance during test in vSphere Web Client

VMware vMotion migration

You can use the VMware vMotion Migration wizard to migrate a powered-on virtual machine from one computing resource to another by using vMotion. To enable this function, set the Advanced vCenter Server setting vgpu.hotmigrate.enabled to true (Figure 22). For more information about support and restrictions, refer to the VMware documentation.



Edit Advanced vCenter Server Settings

Adding or modifying configuration parameters is unsupported and can cause instability. Configuration parameters cannot be removed once they are added. Continue only if you know what you are doing.

Name	Value	Summary
SSL.version	all	SSL cipher required for vCenter client connections
task.batchsize	2000	Number of rows purged per loop by the task cleanup procedure
vgpu.hotmigrate.enabled	☑ Enabled	Enable vGPU hot migration
VirtualCenter.AutoManagedIPV4	10.10.70.30	The automatically discovered IPv4 address of the vCenter Server to which managed hosts send heartbeats

701 - 710 of 754 settings   71 / 76

Name * :   Value :

Name must start with 'config.' For example: config.log

Figure 22.
VMware vCenter Server Advanced setting to enable vGPU migration

NVIDIA Tesla T4 and A16 profile specifications

The Tesla T4 card has a single physical GPU, and the Ampere A16 card has four physical GPUs. Each physical GPU can support several types of vGPU. Each type of vGPU has a fixed amount of frame buffer space, a fixed number of supported display heads, and a fixed maximum resolution, and each is designed for a different class of workload. Table 2 lists the vGPU types supported by GRID GPUs.

For more information, see the Virtual GPU Software User Guide.

Table 2. User specification profiles for NVIDIA T4 and A16 graphic cards

vGPU profile option	A-series application streaming	B-series virtual desktops for business professionals and knowledge workers	C-series computing-intensive server workloads	Q-series virtual workstations (Quadro technology)
1 GB	A16-1A	A16-1B T4-1B		A16-1Q T4-1Q
2 GB	A16-2A	A16-2B T4-2B		A16-2Q T4-2Q

vGPU profile option	A-series application streaming	B-series virtual desktops for business professionals and knowledge workers	C-series computing-intensive server workloads	Q-series virtual workstations (Quadro technology)
4 GB	A16-4A		A16-4C T4-4C	A16-4Q T4-4Q
8 GB	A16-8A		A16-8C T4-8C	A16-8Q T4-8Q
16 GB	A16-16A		A16-16C T4-16QC	A16-16Q T4-16Q

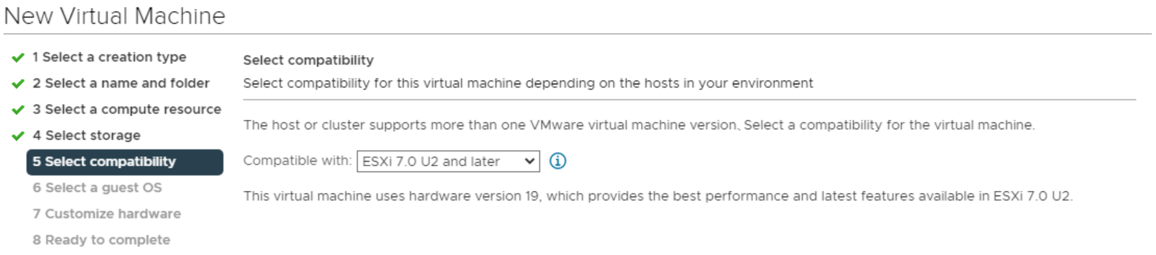
Note the following configurations were used in the evaluation:

- SPECviewperf testing: Two vCPUs with 16 GB of memory for 2Q and 4Q
- Login VSI testing: Two vCPUs with 4 GB of memory for 2B

Creating a new virtual machine in VMware vSphere Web Client

Create a new virtual machine in the vSphere Web Client.

1. Choose “ESXi 7.0 U2 and later” from the “Compatible with” drop-down menu to use the latest features, including the mapping of shared PCI devices, which is required for the vGPU feature. “ESXi 7.0 U2 and later” is used for this study, which provides the latest features available in ESXi 7.0 U2 and virtual machine hardware Release 19 (Figure 23).



New Virtual Machine

✓ 1 Select a creation type
✓ 2 Select a name and folder
✓ 3 Select a compute resource
✓ 4 Select storage
5 Select compatibility
6 Select a guest OS
7 Customize hardware
8 Ready to complete

Select compatibility

Select compatibility for this virtual machine depending on the hosts in your environment

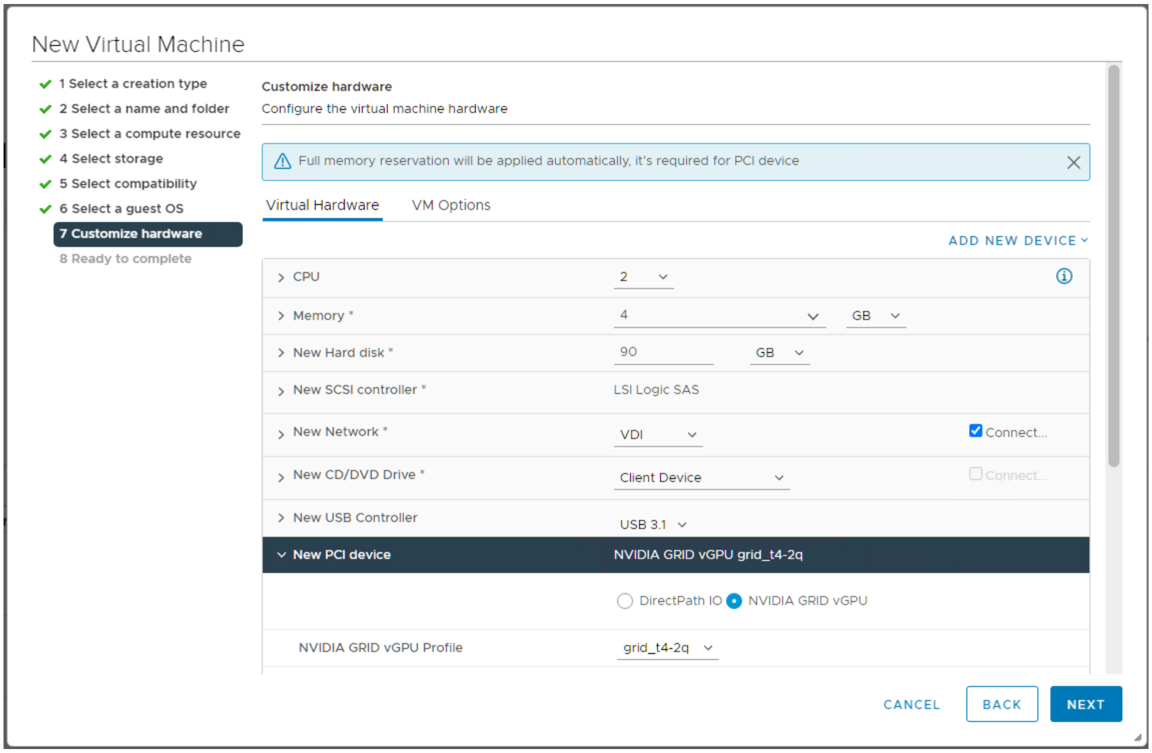
The host or cluster supports more than one VMware virtual machine version. Select a compatibility for the virtual machine.

Compatible with: ESXi 7.0 U2 and later

This virtual machine uses hardware version 19, which provides the best performance and latest features available in ESXi 7.0 U2.

Figure 23.
Selecting the virtual machine version and compatibility

2. To customize the hardware of the new virtual machine, add a new shared PCI device and select the appropriate GPU profile (Figure 24).



New Virtual Machine

✓ 1 Select a creation type
✓ 2 Select a name and folder
✓ 3 Select a compute resource
✓ 4 Select storage
✓ 5 Select compatibility
✓ 6 Select a guest OS
7 Customize hardware
8 Ready to complete

Customize hardware
Configure the virtual machine hardware

Full memory reservation will be applied automatically, it's required for PCI device

Virtual Hardware | VM Options

ADD NEW DEVICE

CPU	2	
Memory *	4 GB	
New Hard disk *	90 GB	
New SCSI controller *	LSI Logic SAS	
New Network *	VDI	☑ Connect...
New CD/DVD Drive *	Client Device	□ Connect...
New USB Controller	USB 3.1	
New PCI device	NVIDIA GRID vGPU grid_t4-2q	
	○ DirectPath IO ● NVIDIA GRID vGPU	
NVIDIA GRID vGPU Profile	grid_t4-2q	

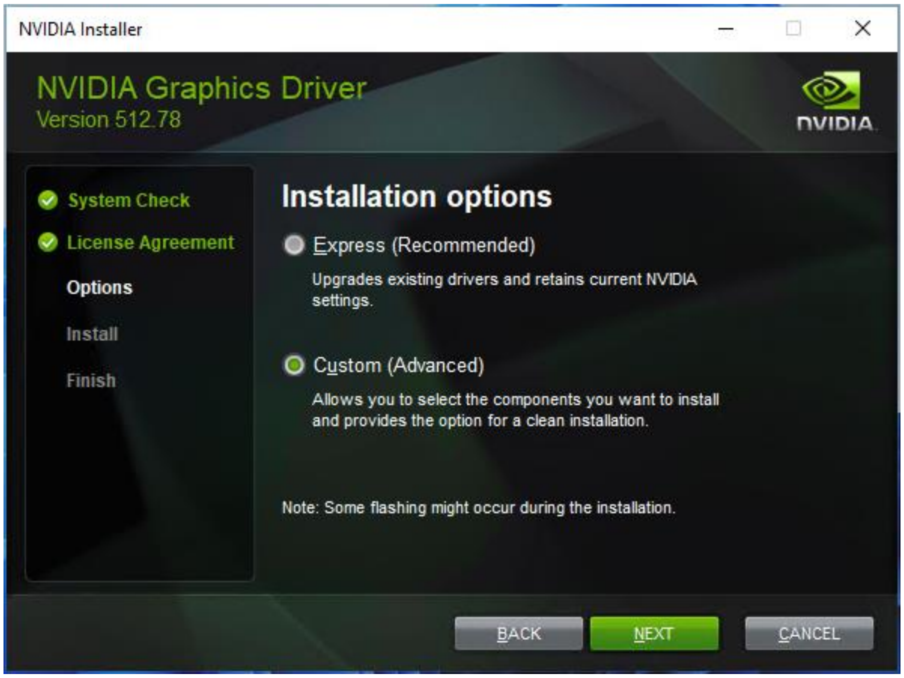
Figure 24.
Adding a shared PCI device to the virtual machine to attach the GPU profile

Install and configure the NVIDIA vGPU software driver

To fully enable vGPU operation, the NVIDIA driver must be installed. Use the following procedure to install and configure the NVIDIA GRID vGPU drivers on the desktop virtual machine.

Note: Before the NVIDIA driver is installed on the guest virtual machine, be sure that remote desktop connections have been enabled. After this step, console access to the virtual machine may not be available when you are connecting from a vSphere Client.

1. Install the graphics drivers using the Express/Custom option (Figure 25.). After the installation has been completed successfully (Figure 26.), restart the virtual machine.



Installation options

○ Express (Recommended)

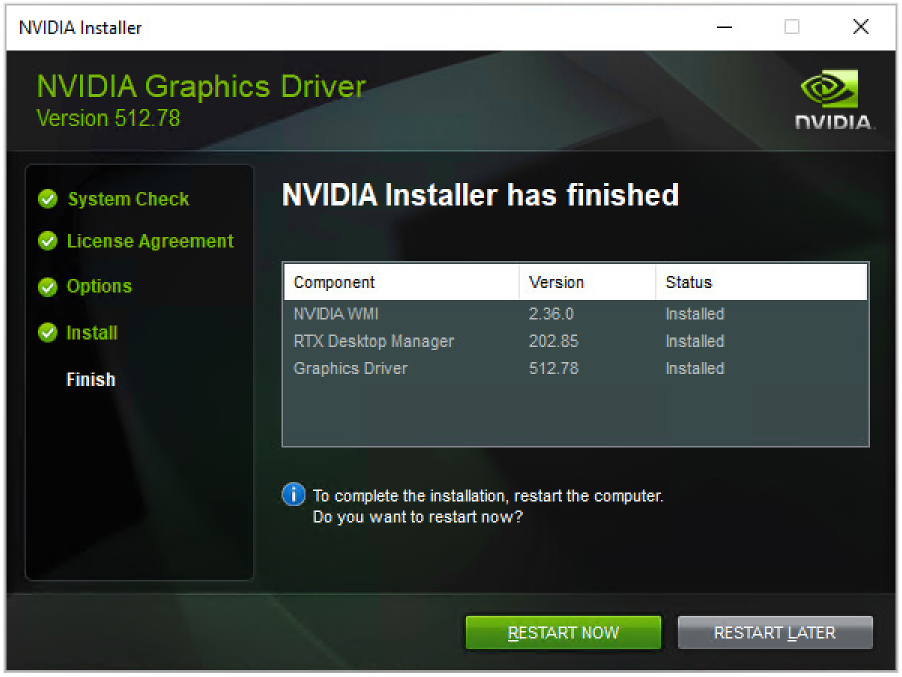
Upgrades existing drivers and retains current NVIDIA settings.

● Custom (Advanced)

Allows you to select the components you want to install and provides the option for a clean installation.

Note: Some flashing might occur during the installation.

Figure 25.
NVIDIA Graphics Driver Installer Options screen



NVIDIA Installer has finished

Component	Version	Status
NVIDIA WMI	2.36.0	Installed
RTX Desktop Manager	202.85	Installed
Graphics Driver	512.78	Installed

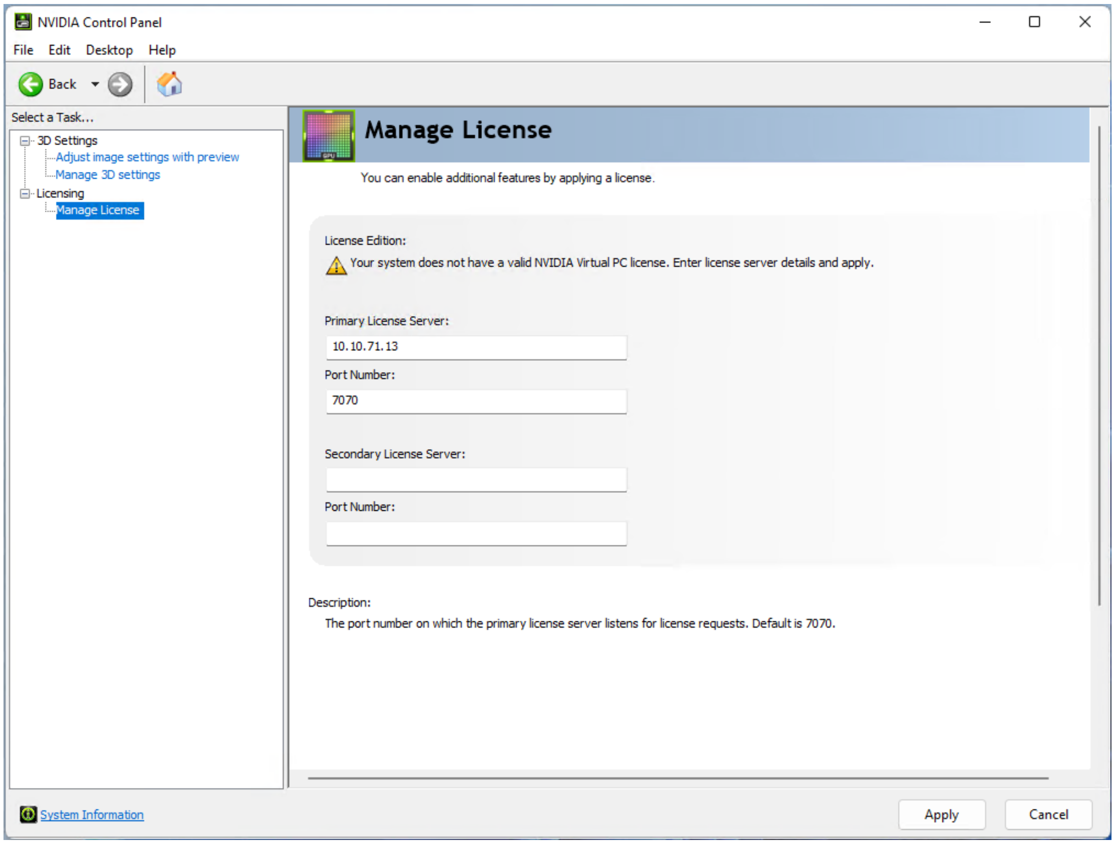
To complete the installation, restart the computer.
Do you want to restart now?

Figure 26.
NVIDIA Graphics Driver Installer Finish screen

1. Configure the NVIDIA license server address in a virtual machine or a master image after the driver has been successfully installed.

Note: The license settings persist across reboots. These settings can also be preloaded through registry keys.

2. Open the NVIDIA control panel, select Manage License, and enter your license server address and port. Apply the settings (Figure 27).



Manage License

You can enable additional features by applying a license.

License Edition:

Your system does not have a valid NVIDIA Virtual PC license. Enter license server details and apply.

Primary License Server: 10.10.71.13

Port Number: 7070

Secondary License Server:

Port Number:

Description:

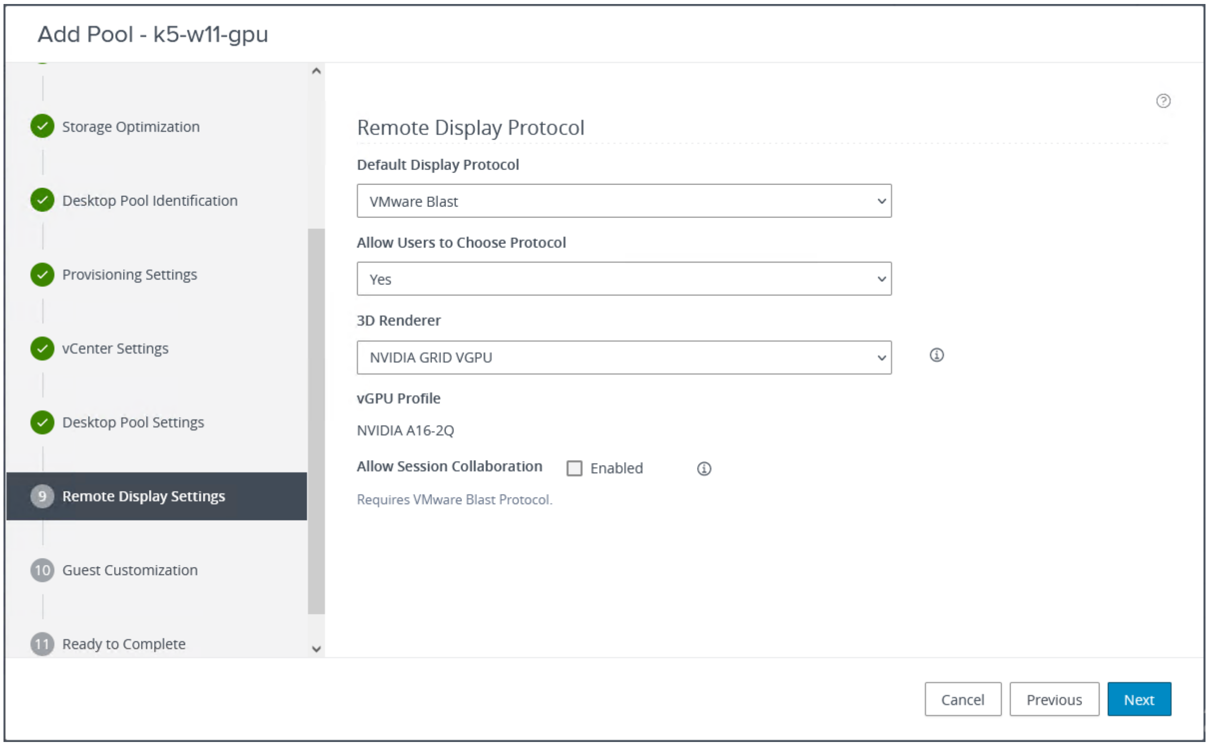
The port number on which the primary license server listens for license requests. Default is 7070.

Figure 27.
NVIDIA Control Panel

Create the VMware Horizon 2111 pool

Each Horizon desktop pool configuration depends on the specific use case. The automated desktop pool based on instant clone desktops with floating assignments was created as part of the Login VSI testing. The virtual machines are deployed as instant clones from the master image template.

In creating the VMware Horizon desktop pool, choose VMware Blast and select the NVIDIA GRID VGPU option for the 3D render. The NVIDIA vGPU profile attached to a desktop master image will be shown (Figure 28).



Add Pool - k5-w11-gpu

Remote Display Protocol

Default Display Protocol: VMware Blast

Allow Users to Choose Protocol: Yes

3D Renderer: NVIDIA GRID VGPU

vGPU Profile: NVIDIA A16-2Q

Allow Session Collaboration: □ Enabled

Requires VMware Blast Protocol.

Figure 28.
VMware Horizon instant desktop pool creation with NVIDIA vGPU profile attached to a Microsoft Windows 11 master image

SPECviewperf results with NVIDIA Virtual Workstation profiles

SPECviewperf 2020 is the latest version of the benchmark that measures the 3D graphics performance of systems running under the OpenGL and Direct X APIs. The benchmark's workloads, called viewsets, represent graphics content and behavior from actual applications.

SPECviewperf uses these viewsets:

- 3ds Max (3dsmax-07)
- CATIA (catia-06)
- Creo (creo-03
- Energy (energy-03)
- Maya (maya-06)
- Medical (medical-03)
- Siemens NX (snx-04)
- SolidWorks (solidworks -07)

The following features introduced in SPECviewperf 2020 v1.0 were used for the tests:

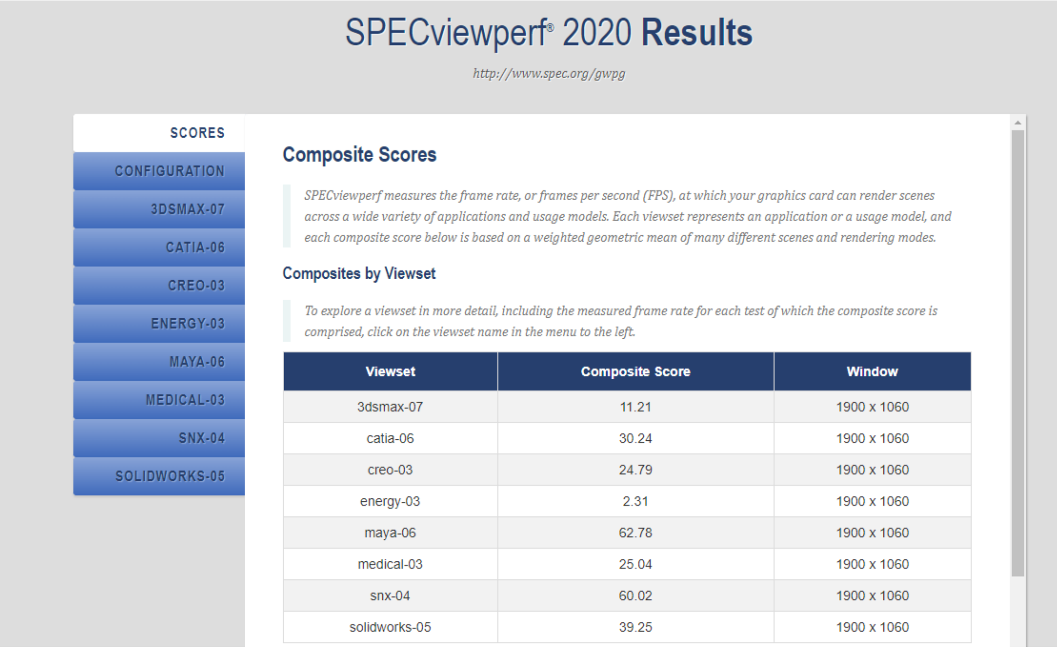
- New viewsets are created from API traces of the latest versions of 3ds Max, Catia, Maya, and Solidworks applications.
- Updated models in the viewsets are based on 3ds Max, Catia, Creo, Solidworks, and real-world medical applications.
- All viewsets support both 2K and 4K resolution displays.
- User interface improvements include better interrogation and assessment of underlying hardware, clickable thumbnails of screen grabs, and a new results manager.
- The benchmark can be run using command-line options.

SPECviewperf hardware and operating system requirements are as follows:

- Microsoft Windows 10 Version 1709 (Fall Creators Update / RS3) or Windows 11 or later
- 16 GB or more of system RAM
- 80 GB of available disk space
- Minimum screen resolution of 1920 × 1080 for submissions published on the SPEC website
- OpenGL 4.5 (for catia-06, creo-03, energy-03, maya-06, medical-03, snx-04, and solidworks-07) and DirectX 12 API support (for 3dsmax-07)
- GPU with 2 GB or more dedicated GPU memory

Cisco UCS X210c M6 server with mezzanine VIC (T4-2Q)

Figures 29 through 33 show the results for Microsoft Windows 10 with four vCPUs and 16 GB of memory and the T4-2Q NVIDIA GRID profile on a Cisco UCS X210c M6 server with dual Intel Xeon Gold 6348 2.60-GHz 28-core processors and 1 TB of 3200-MHz RAM.



SPECviewperf® 2020 **Results**

http://www.spec.org/gwpg

Composite Scores

SPECviewperf measures the frame rate, or frames per second (FPS), at which your graphics card can render scenes across a wide variety of applications and usage models. Each viewset represents an application or a usage model, and each composite score below is based on a weighted geometric mean of many different scenes and rendering modes.

Composites by Viewset

To explore a viewset in more detail, including the measured frame rate for each test of which the composite score is comprised, click on the viewset name in the menu to the left.

Viewset	Composite Score	Window
3dsmax-07	11.21	1900 x 1060
catia-06	30.24	1900 x 1060
creo-03	24.79	1900 x 1060
energy-03	2.31	1900 x 1060
maya-06	62.78	1900 x 1060
medical-03	25.04	1900 x 1060
snx-04	60.02	1900 x 1060
solidworks-05	39.25	1900 x 1060

Figure 29.
SPECviewperf composite scores

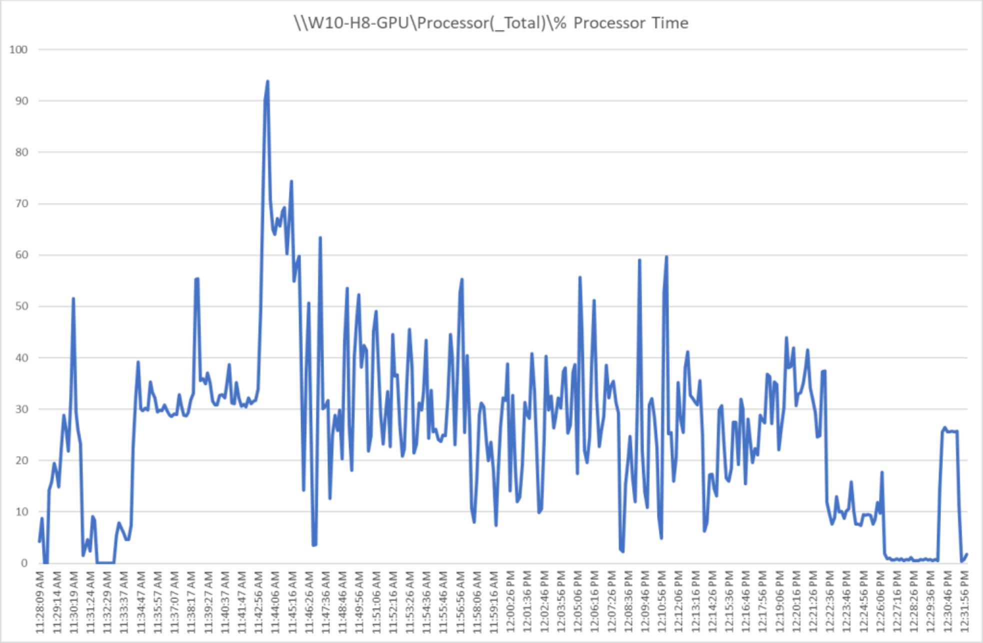


Figure 30.
Perfmon virtual machine CPU utilization

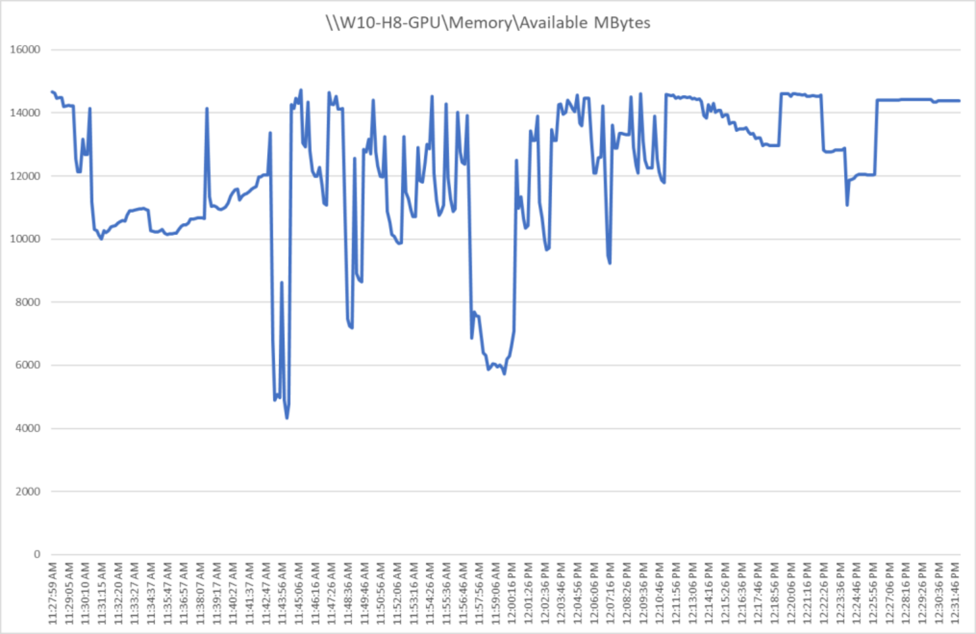


Figure 31.
Perfmon virtual machine memory utilization

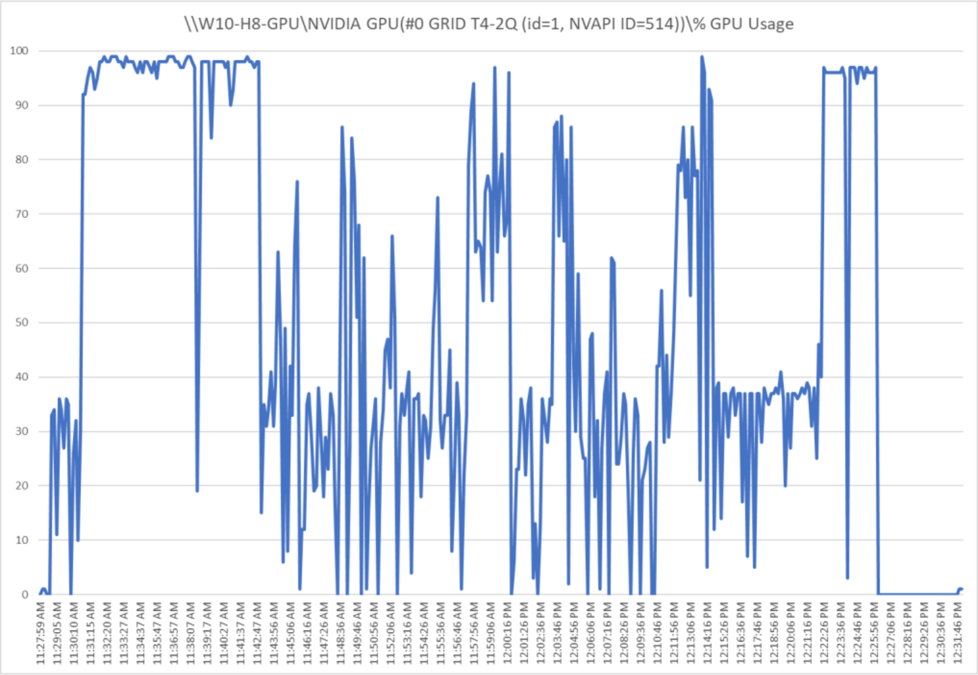


Figure 32.
Perfmon GPU utilization

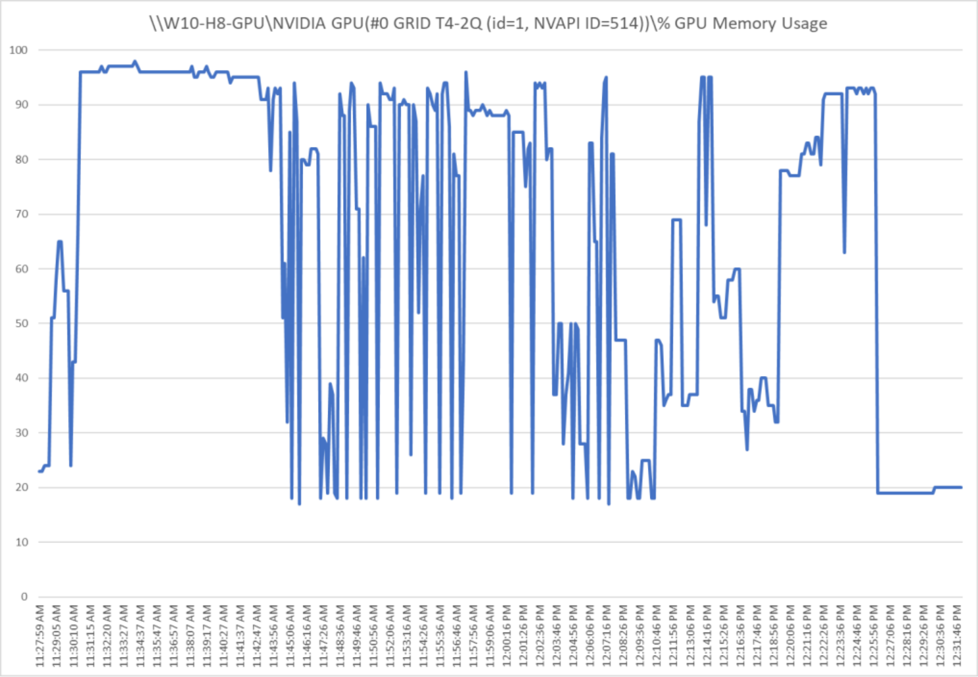
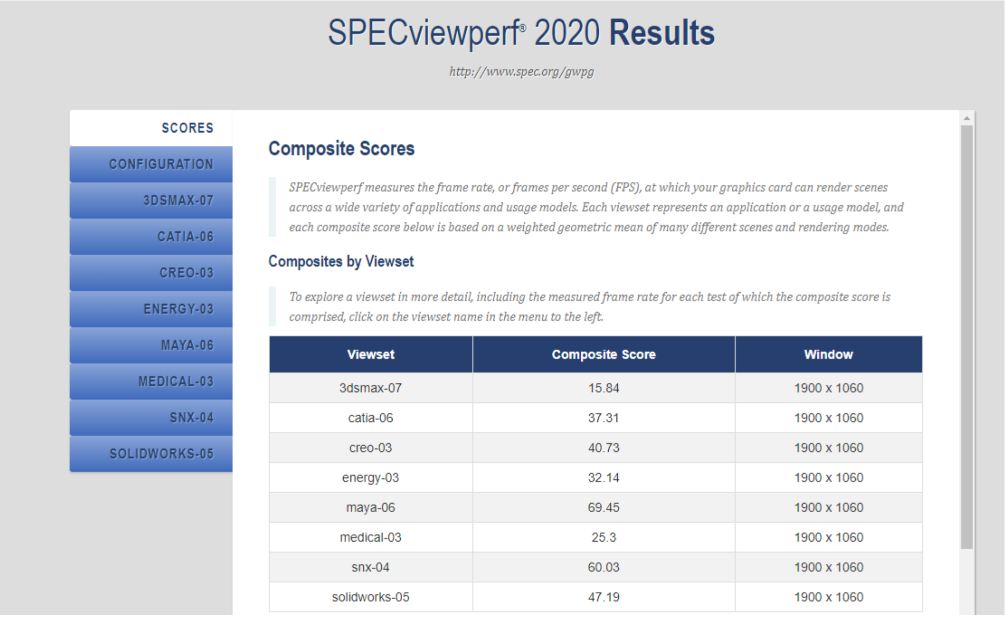


Figure 33.
Perfmon GPU memory utilization

Cisco UCS X210c M6 server with mezzanine VIC (T4-4Q)

Figures 34 through 38 show the results for Microsoft Windows 10 with four vCPUs and 16 GB of memory and the T4-4Q NVIDIA GRID vGPU profile on a Cisco UCS X210c M6 server with dual Intel Xeon Gold 6348 2.60-GHz 28-core processors and 1 TB of 3200-MHz RAM.



SPECviewperf® 2020 **Results**

http://www.spec.org/gwpg

Composite Scores

SPECviewperf measures the frame rate, or frames per second (FPS), at which your graphics card can render scenes across a wide variety of applications and usage models. Each viewset represents an application or a usage model, and each composite score below is based on a weighted geometric mean of many different scenes and rendering modes.

Composites by Viewset

To explore a viewset in more detail, including the measured frame rate for each test of which the composite score is comprised, click on the viewset name in the menu to the left.

Viewset	Composite Score	Window
3dsmax-07	15.84	1900 x 1060
catia-06	37.31	1900 x 1060
creo-03	40.73	1900 x 1060
energy-03	32.14	1900 x 1060
maya-06	69.45	1900 x 1060
medical-03	25.3	1900 x 1060
snx-04	60.03	1900 x 1060
solidworks-05	47.19	1900 x 1060

Figure 34.
SPECviewperf composite scores

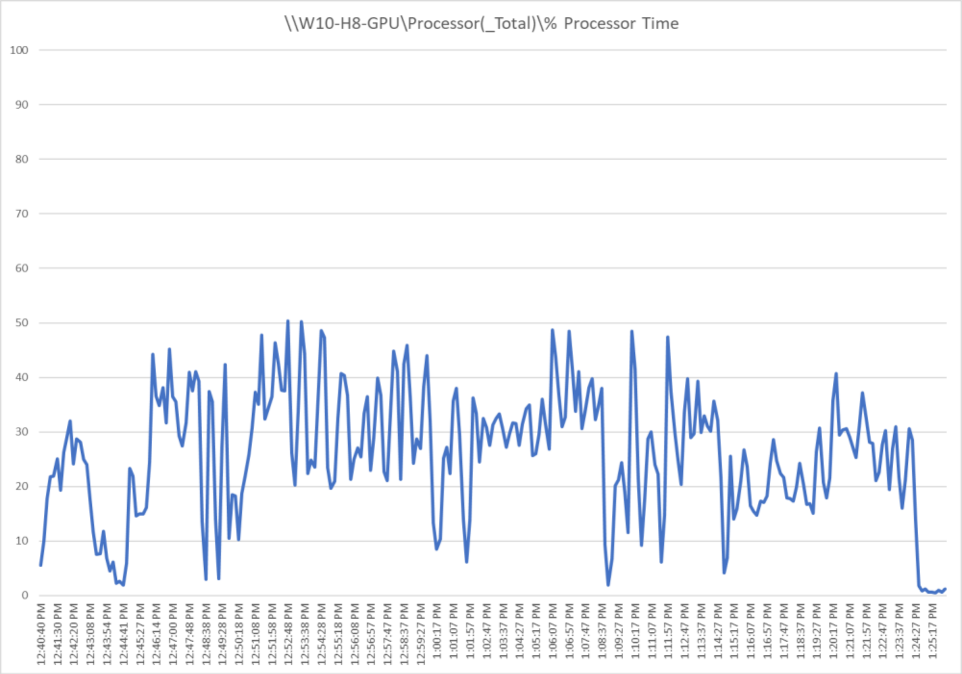


Figure 35.
Perfmon virtual machine CPU utilization

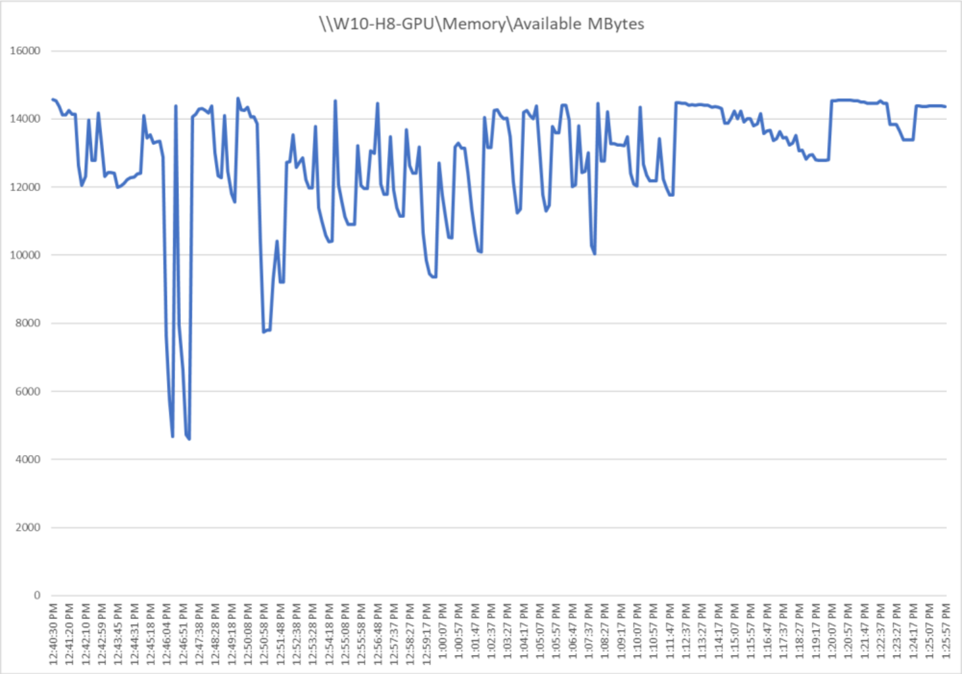


Figure 36.
Perfmon virtual machine memory utilization

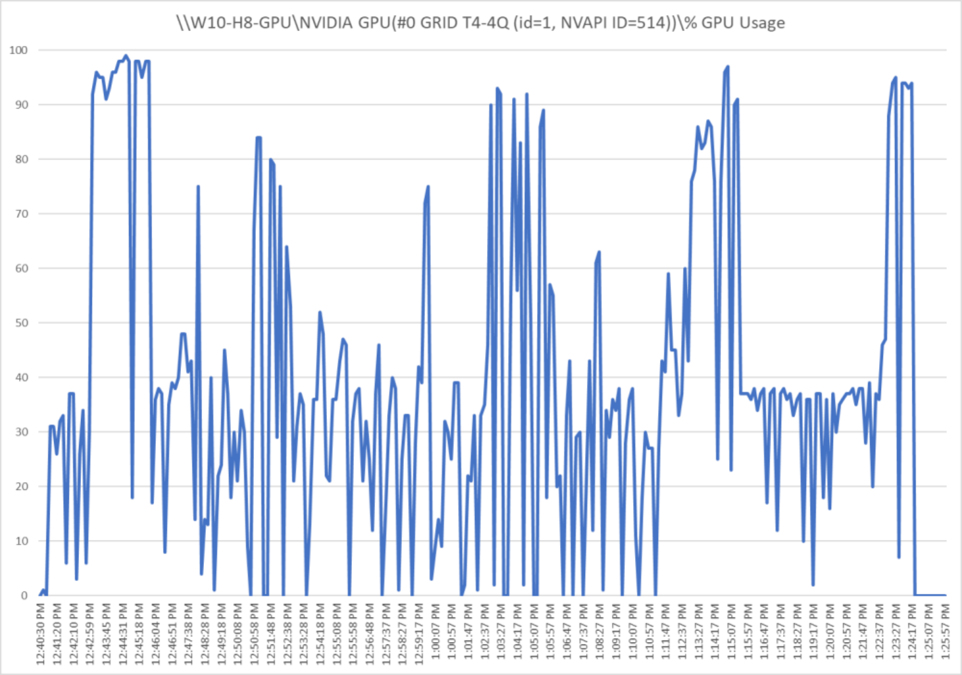


Figure 37.
Perfmon GPU utilization

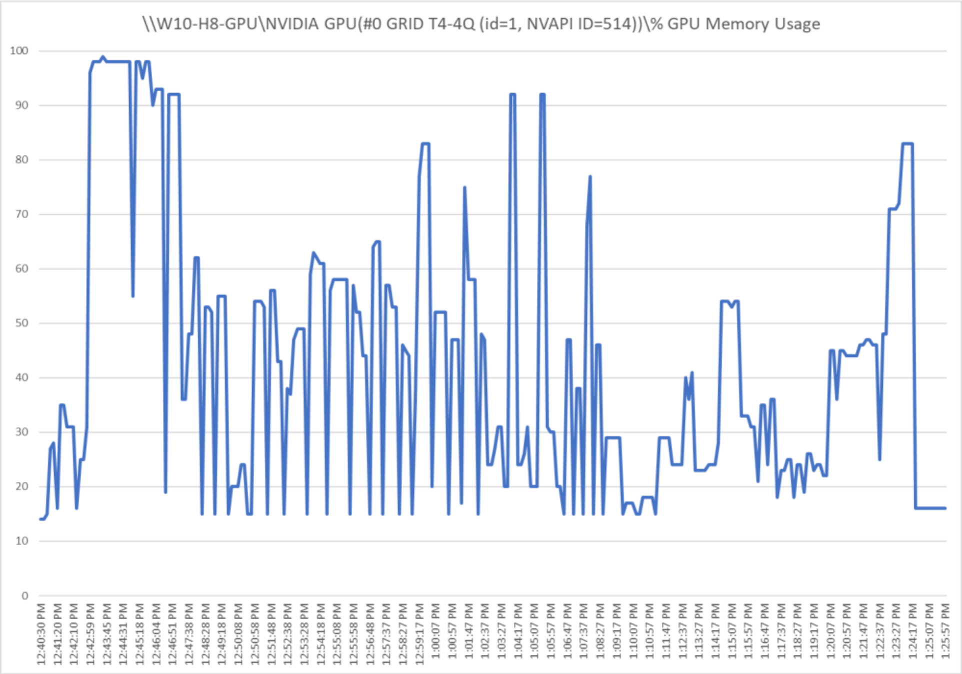
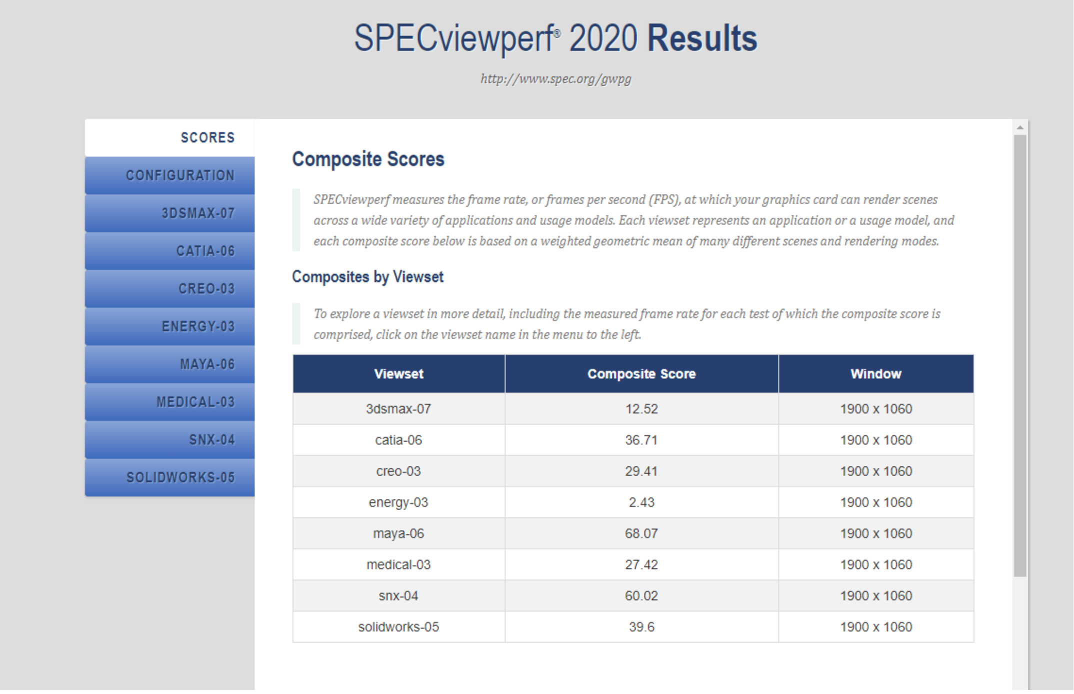


Figure 38.
Perfmon GPU memory utilization

Cisco UCS X210c M6 server with PCI node (T4-2Q)

Figures 39 through 43 show the results for Microsoft Windows 10 with four vCPUs and 16 GB of memory and the T4-2Q NVIDIA GRID vGPU profile on a Cisco UCS X210c M6 server with dual Intel Xeon Platinum 8358 2.60-GHz 32-core processors and 2 TB of 3200-MHz RAM.



SPECviewperf® 2020 **Results**

http://www.spec.org/gwpg

Composite Scores

SPECviewperf measures the frame rate, or frames per second (FPS), at which your graphics card can render scenes across a wide variety of applications and usage models. Each viewset represents an application or a usage model, and each composite score below is based on a weighted geometric mean of many different scenes and rendering modes.

Composites by Viewset

To explore a viewset in more detail, including the measured frame rate for each test of which the composite score is comprised, click on the viewset name in the menu to the left.

Viewset	Composite Score	Window
3dsmax-07	12.52	1900 x 1060
catia-06	36.71	1900 x 1060
creo-03	29.41	1900 x 1060
energy-03	2.43	1900 x 1060
maya-06	68.07	1900 x 1060
medical-03	27.42	1900 x 1060
snx-04	60.02	1900 x 1060
solidworks-05	39.6	1900 x 1060

Figure 39.
SPECviewperf composite scores

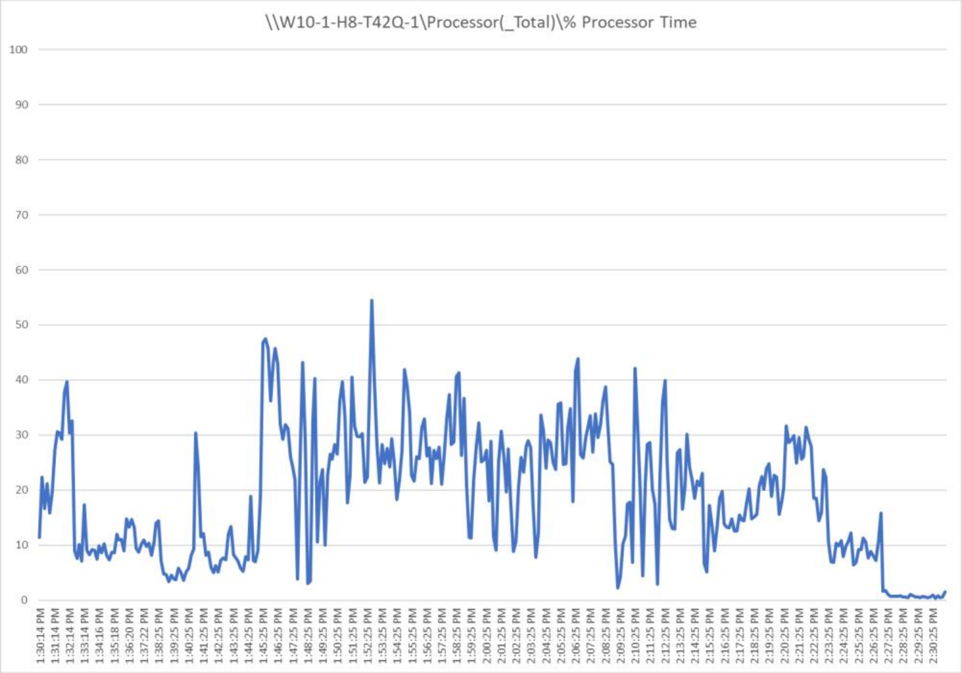


Figure 40.
Perfmon virtual machine CPU utilization

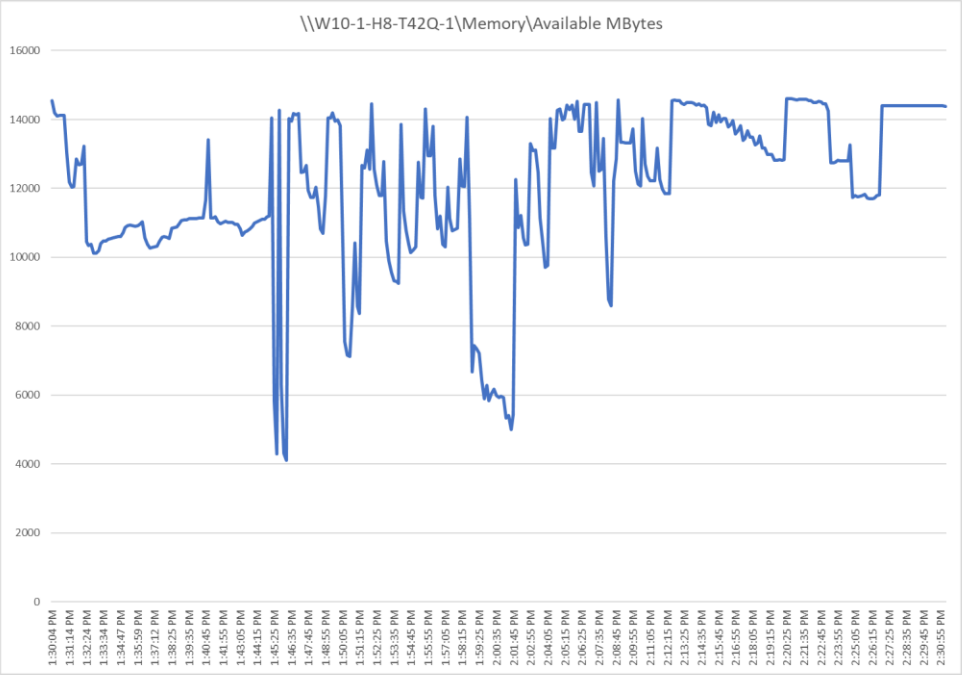


Figure 41.
Perfmon virtual machine memory utilization

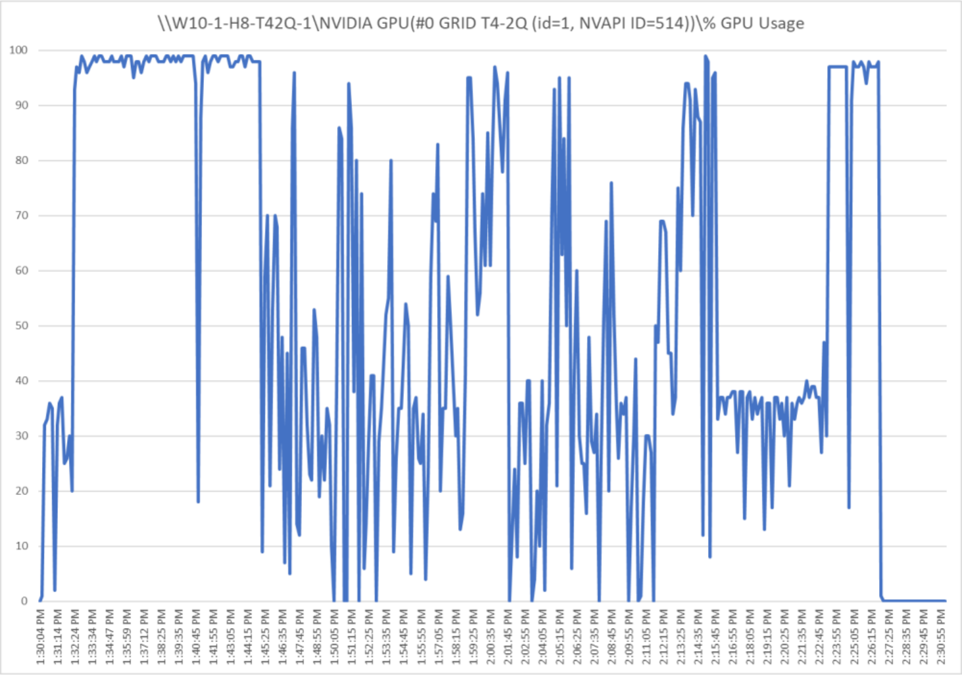


Figure 42.
Perfmon GPU utilization

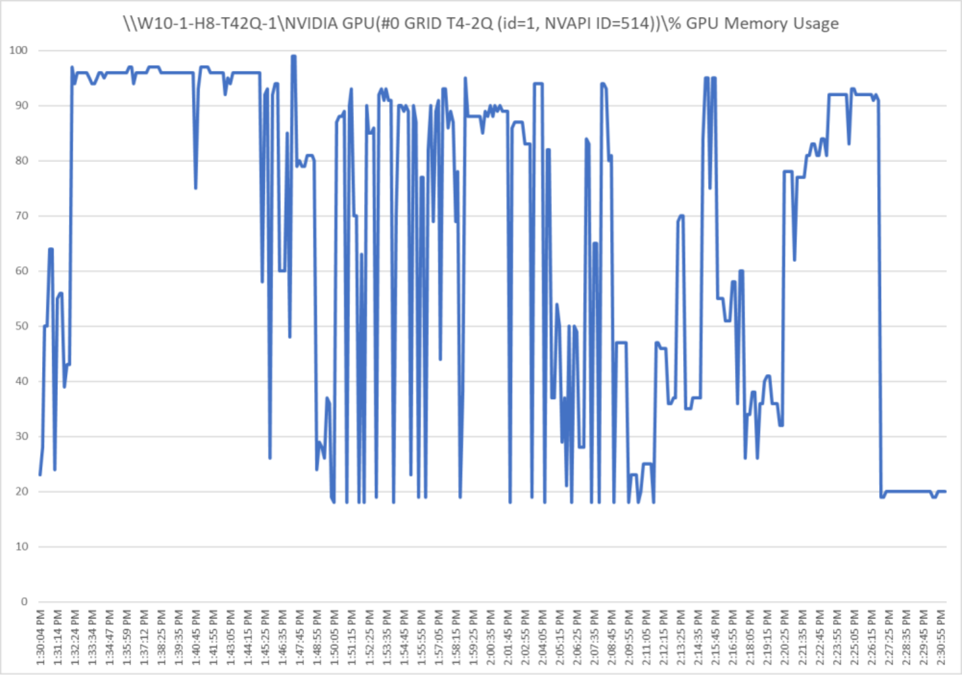
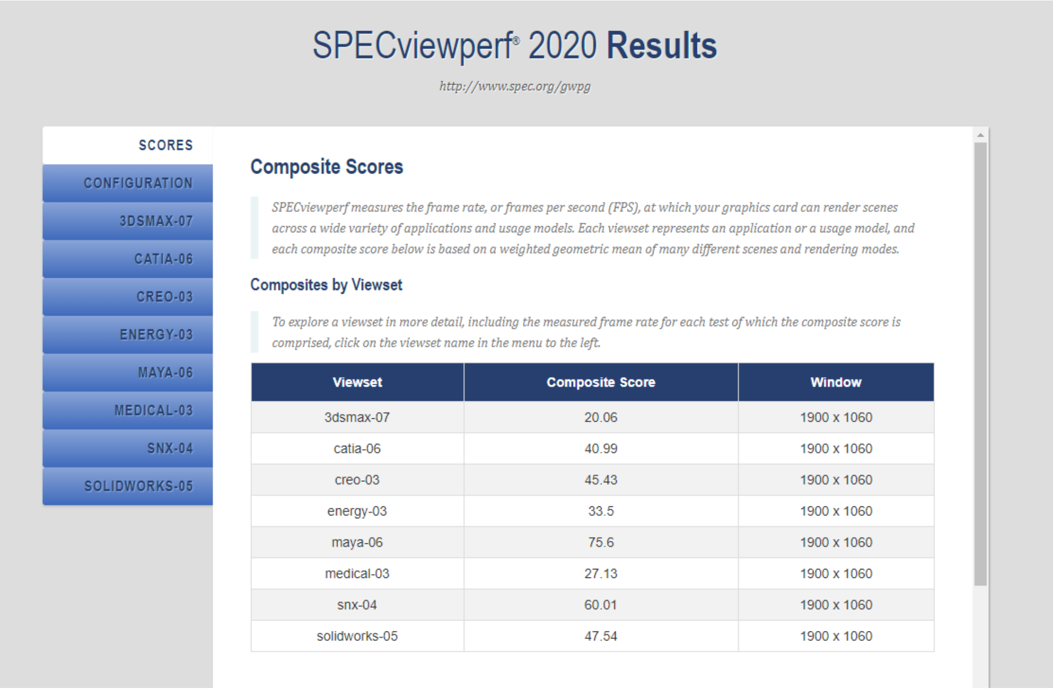


Figure 43.
Perfmon GPU memory utilization

Cisco UCS X210c M6 server with PCI node (T4-4Q)

Figures 44 through 48 show the results for Microsoft Windows 10 with four vCPUs and 16 GB of memory and the T4-4Q NVIDIA GRID vGPU profile on a Cisco UCS X210c M6 server with dual Intel Xeon Platinum 8358 2.60-GHz 32-core processors and 2 TB of 3200-MHz RAM.



Composite Scores

SPECviewperf measures the frame rate, or frames per second (FPS), at which your graphics card can render scenes across a wide variety of applications and usage models. Each viewset represents an application or a usage model, and each composite score below is based on a weighted geometric mean of many different scenes and rendering modes.

Composites by Viewset

To explore a viewset in more detail, including the measured frame rate for each test of which the composite score is comprised, click on the viewset name in the menu to the left.

Viewset	Composite Score	Window
3dsmax-07	20.06	1900 x 1060
catia-06	40.99	1900 x 1060
creo-03	45.43	1900 x 1060
energy-03	33.5	1900 x 1060
maya-06	75.6	1900 x 1060
medical-03	27.13	1900 x 1060
snx-04	60.01	1900 x 1060
solidworks-05	47.54	1900 x 1060

Figure 44.
SPECviewperf composite scores

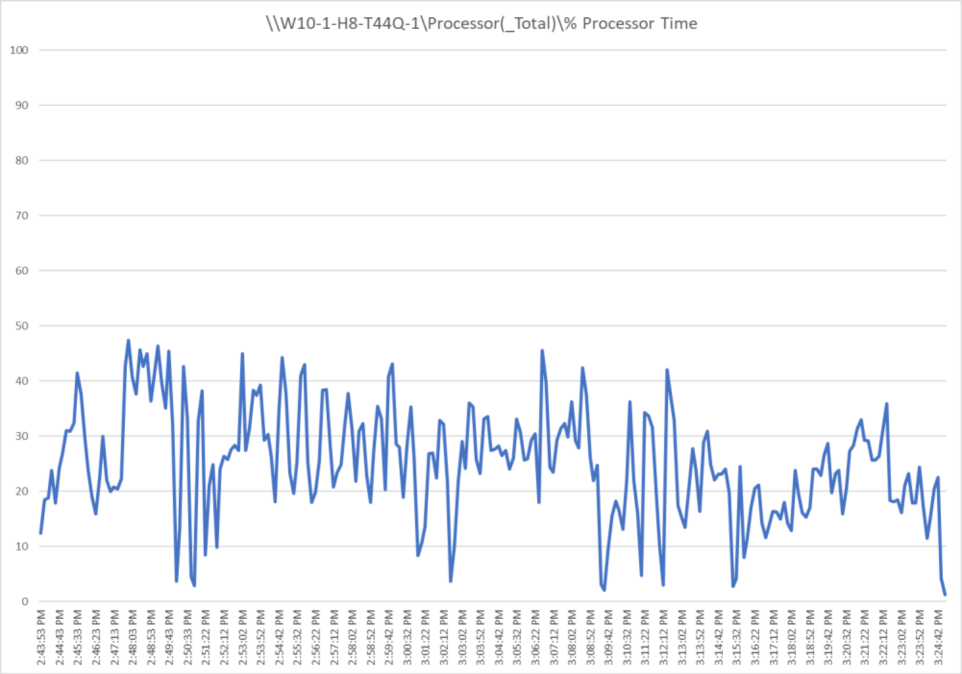


Figure 45.
Perfmon virtual machine CPU utilization

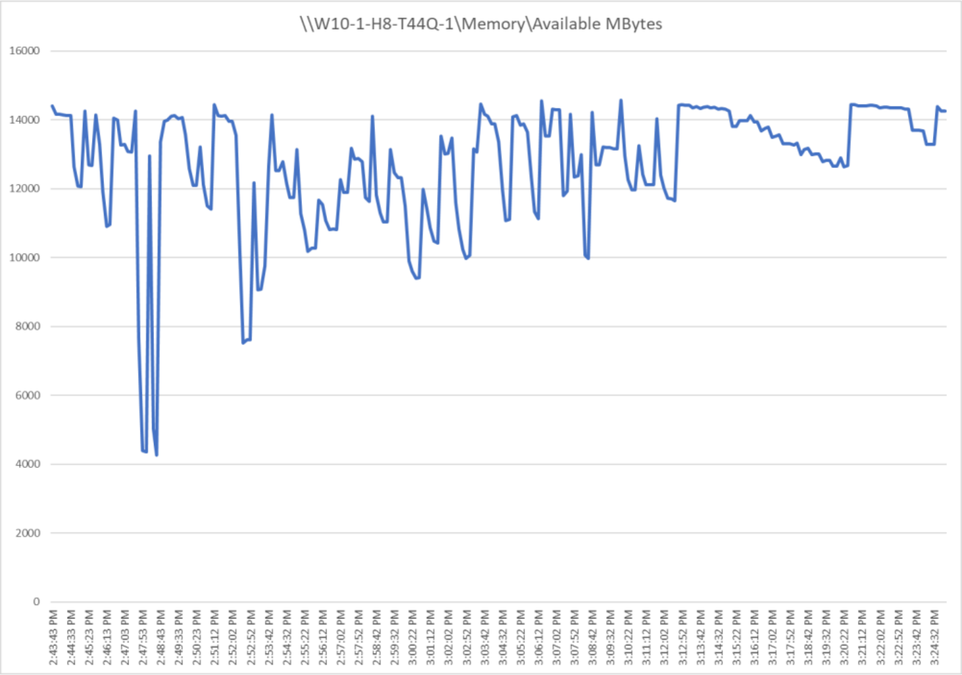


Figure 46.
Perfmon virtual memory utilization

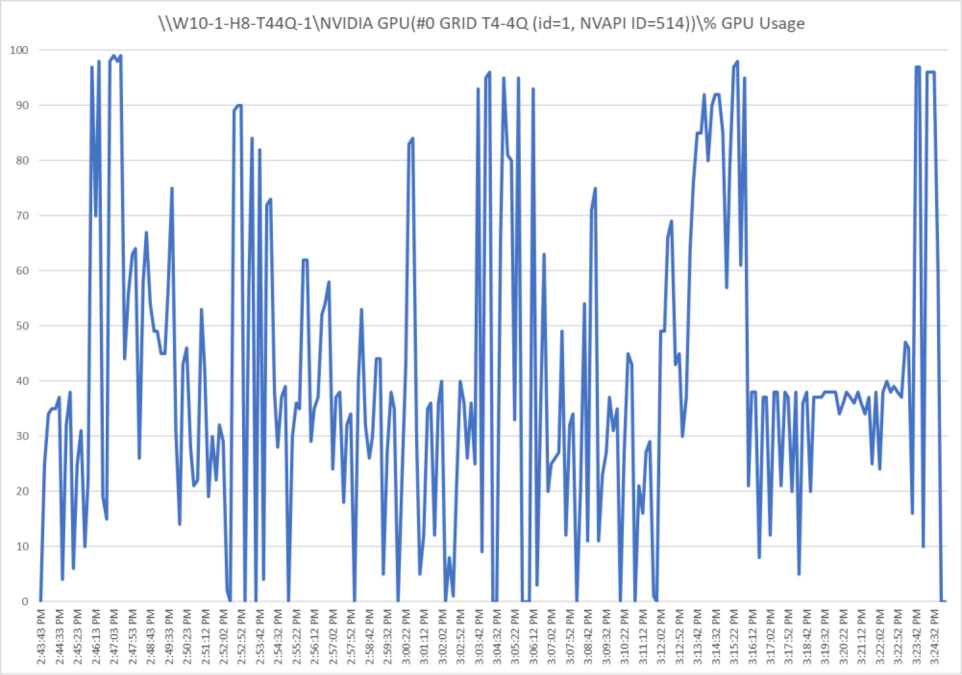


Figure 47.
Perfmon GPU utilization

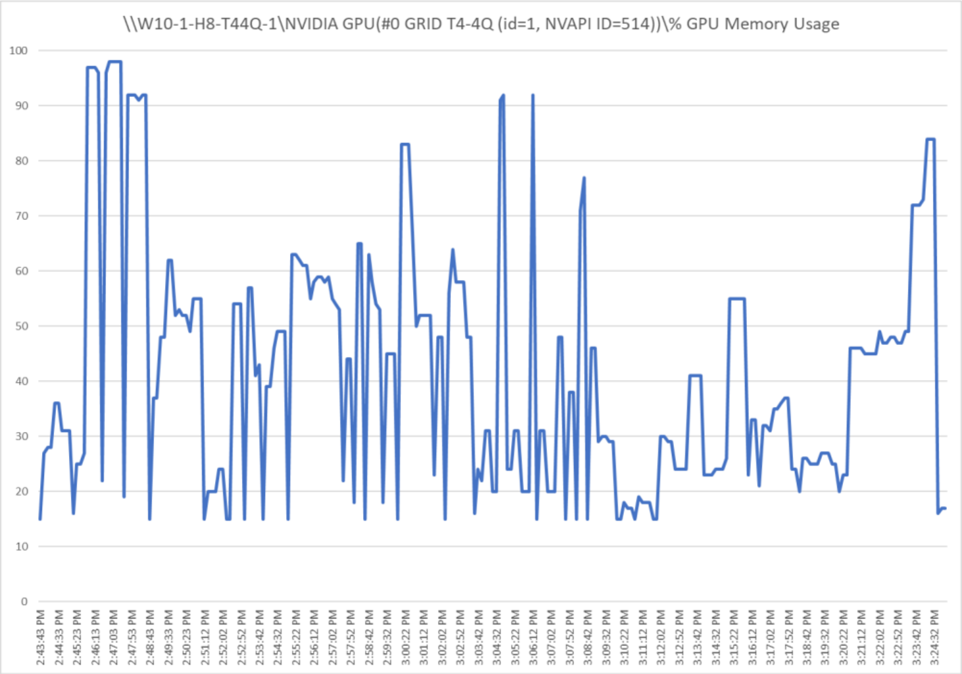
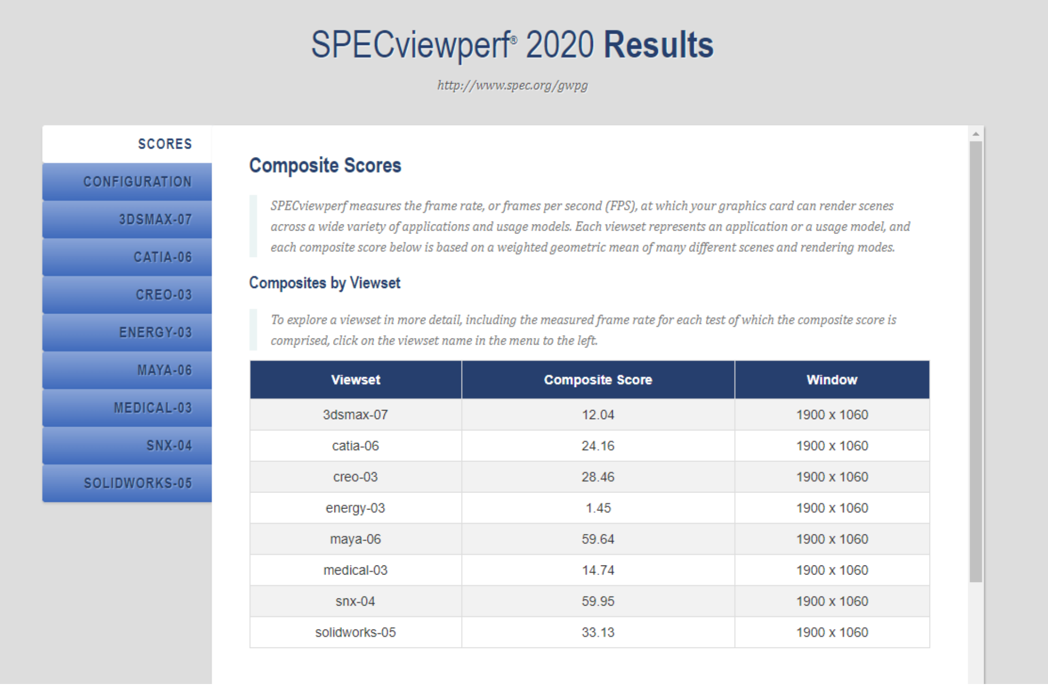


Figure 48.
Perfmon GPU memory utilization

Cisco UCS X210c M6 server with PCI node (A16-2Q)

Figures 49 through 53 show the results for Microsoft Windows 10 with four vCPUs and 16 GB of memory and the A16-2Q NVIDIA GRID vGPU profile on a Cisco UCS X210c M6 server with dual Intel Xeon Gold 6348 2.60-GHz 28-core processors and 1 TB of 3200-MHz RAM.



SPECviewperf® 2020 **Results**

http://www.spec.org/gwpg

Composite Scores

> *SPECviewperf measures the frame rate, or frames per second (FPS), at which your graphics card can render scenes across a wide variety of applications and usage models. Each viewset represents an application or a usage model, and each composite score below is based on a weighted geometric mean of many different scenes and rendering modes.*

Composites by Viewset

> *To explore a viewset in more detail, including the measured frame rate for each test of which the composite score is comprised, click on the viewset name in the menu to the left.*

Viewset	Composite Score	Window
3dsmax-07	12.04	1900 x 1060
catia-06	24.16	1900 x 1060
creo-03	28.46	1900 x 1060
energy-03	1.45	1900 x 1060
maya-06	59.64	1900 x 1060
medical-03	14.74	1900 x 1060
snx-04	59.95	1900 x 1060
solidworks-05	33.13	1900 x 1060

Figure 49.
SPECviewperf composite scores

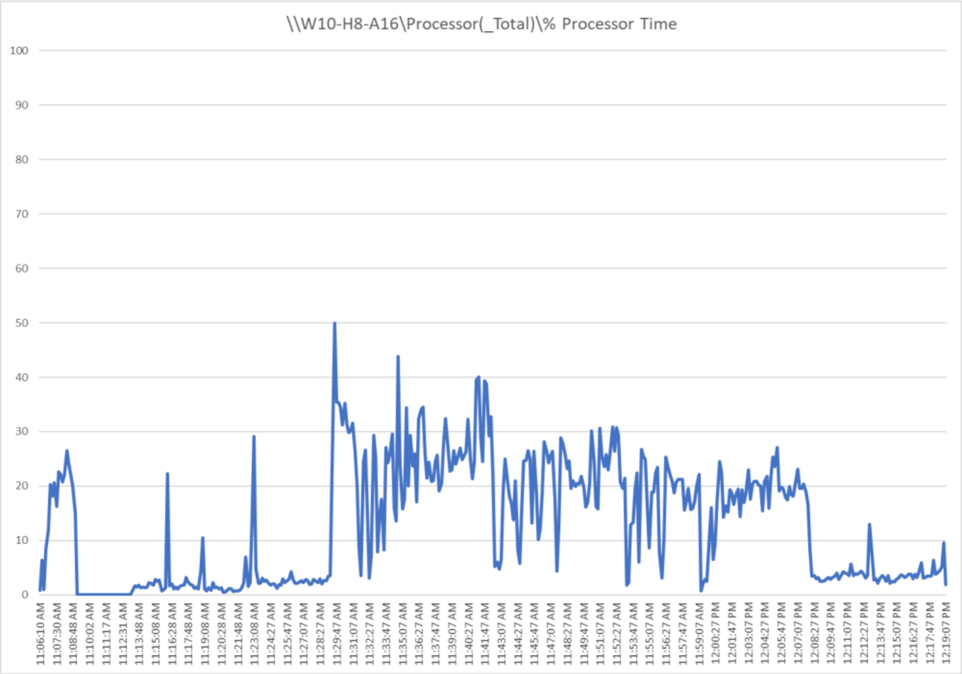


Figure 50.
Perfmon virtual machine CPU utilization

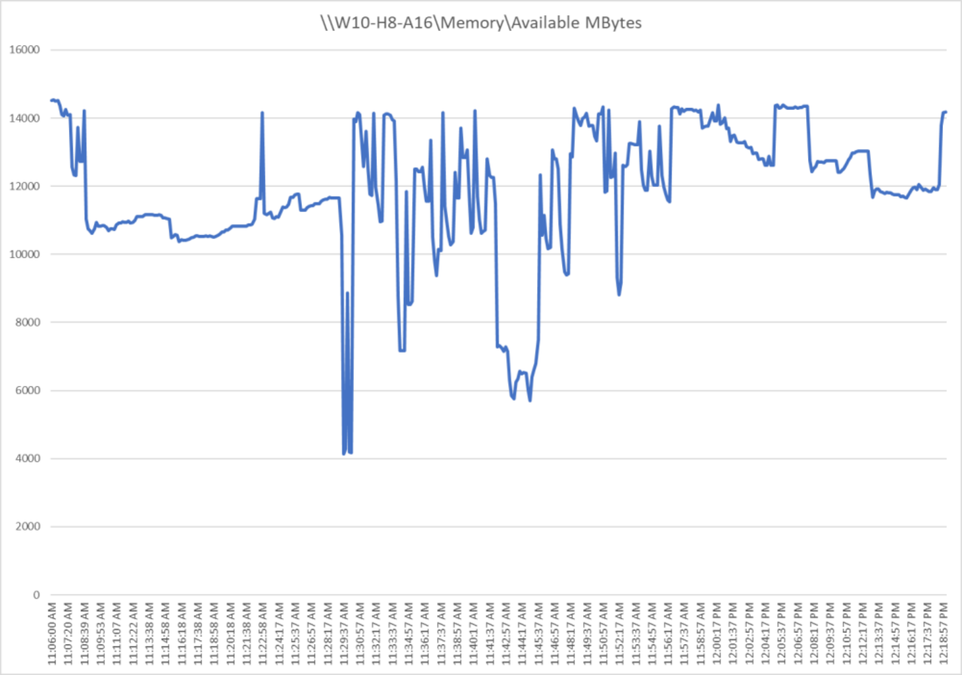


Figure 51.
Perfmon virtual machine memory utilization

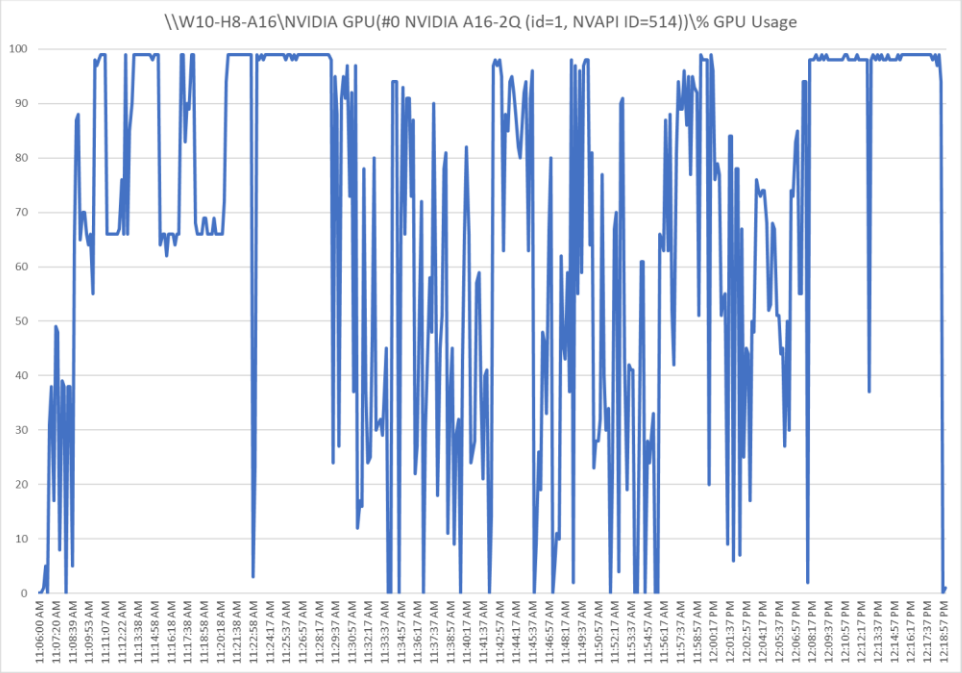


Figure 52.
Perfmon GPU utilization

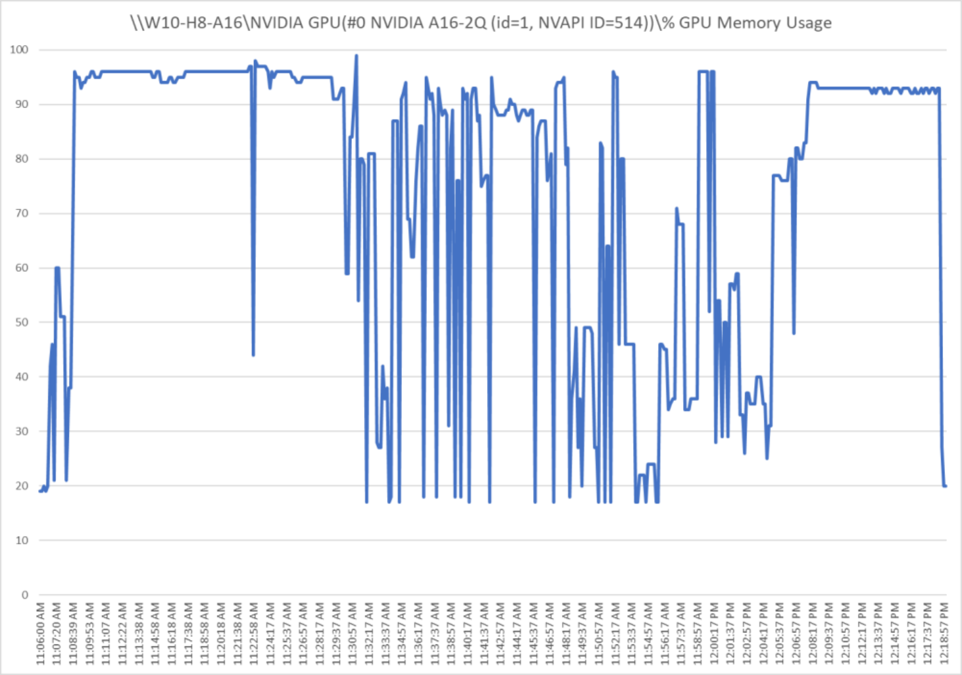
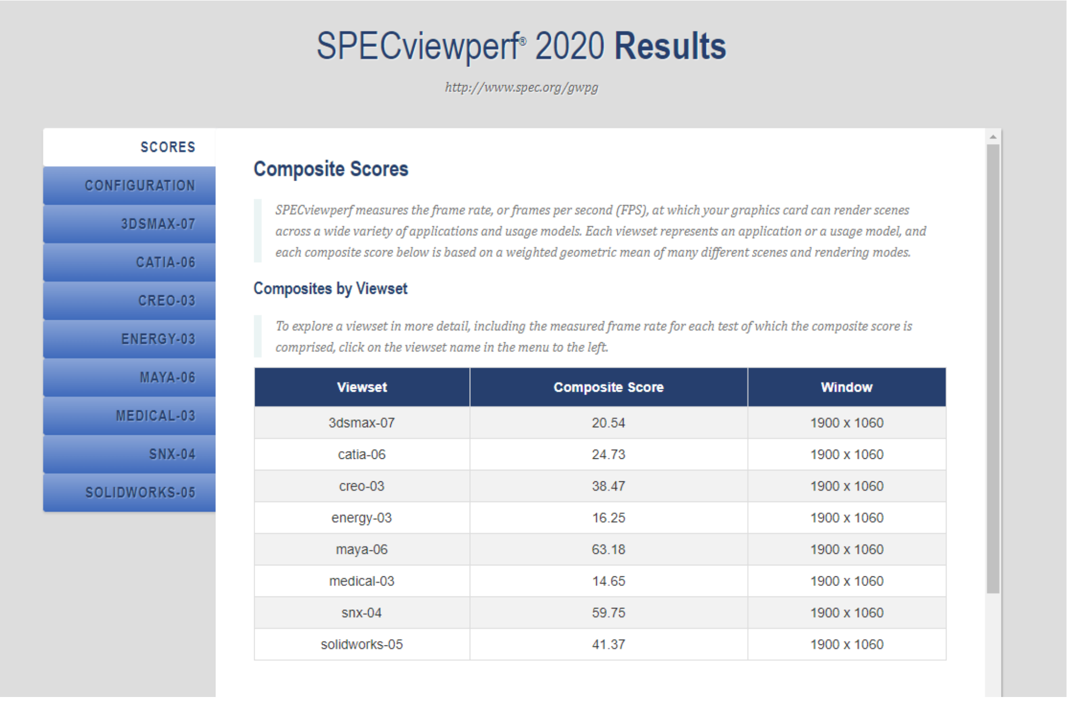


Figure 53.
Perfmon GPU memory utilization

Cisco UCS X210c M6 server with PCI node (A16-4Q)

Figures 54 through 58 show the results for Microsoft Windows 10 (OS build 18363) with four vCPUs and 16 GB of memory and the A16-4Q NVIDIA GRID vGPU profile on a Cisco UCS X210c M6 server with dual Intel Xeon Gold 6348 2.60-GHz 28-core processors and 1 TB of 3200-MHz RAM.



Composite Scores

SPECviewperf measures the frame rate, or frames per second (FPS), at which your graphics card can render scenes across a wide variety of applications and usage models. Each viewset represents an application or a usage model, and each composite score below is based on a weighted geometric mean of many different scenes and rendering modes.

Composites by Viewset

To explore a viewset in more detail, including the measured frame rate for each test of which the composite score is comprised, click on the viewset name in the menu to the left.

Viewset	Composite Score	Window
3dsmax-07	20.54	1900 x 1060
catia-06	24.73	1900 x 1060
creo-03	38.47	1900 x 1060
energy-03	16.25	1900 x 1060
maya-06	63.18	1900 x 1060
medical-03	14.65	1900 x 1060
snx-04	59.75	1900 x 1060
solidworks-05	41.37	1900 x 1060

Figure 54.
SPECviewperf composite scores

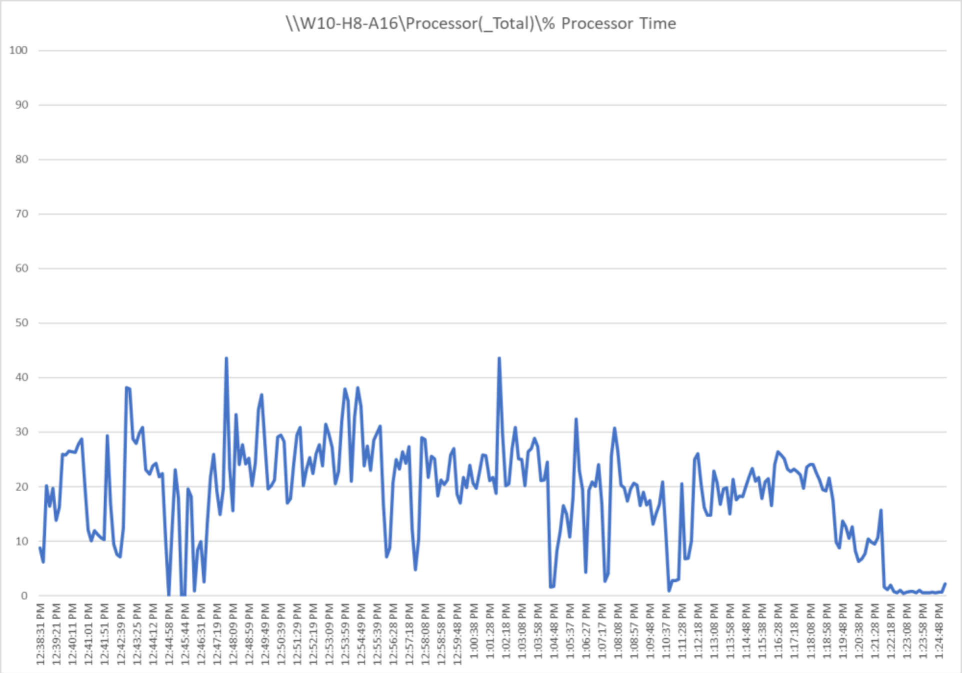


Figure 55.
Perfmon virtual machine CPU Utilization

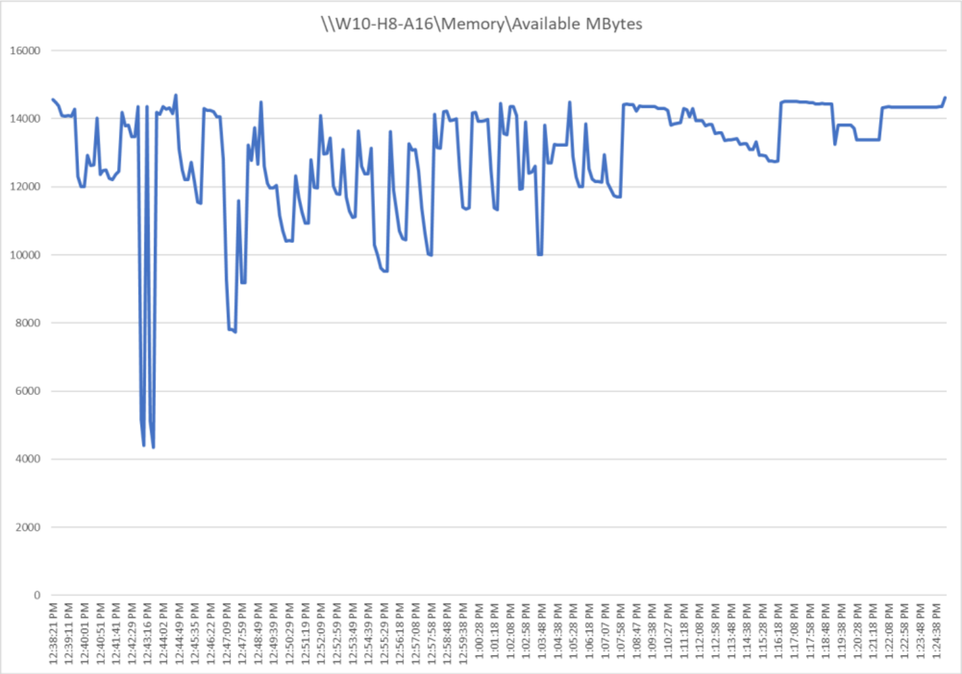


Figure 56.
Perfmon virtual machine memory utilization

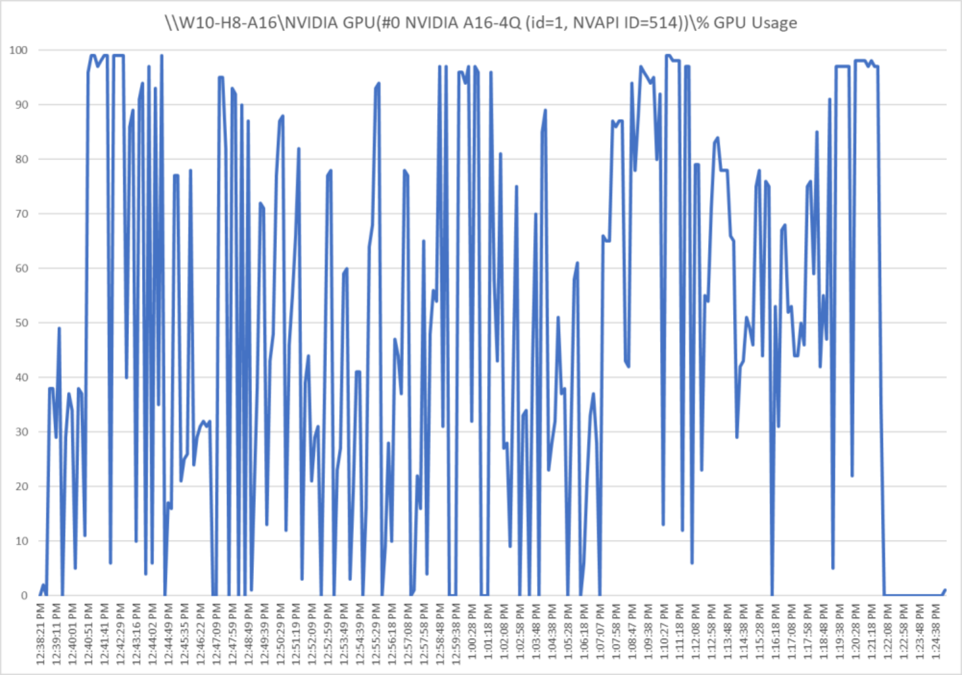


Figure 57.
Perfmon GPU utilization

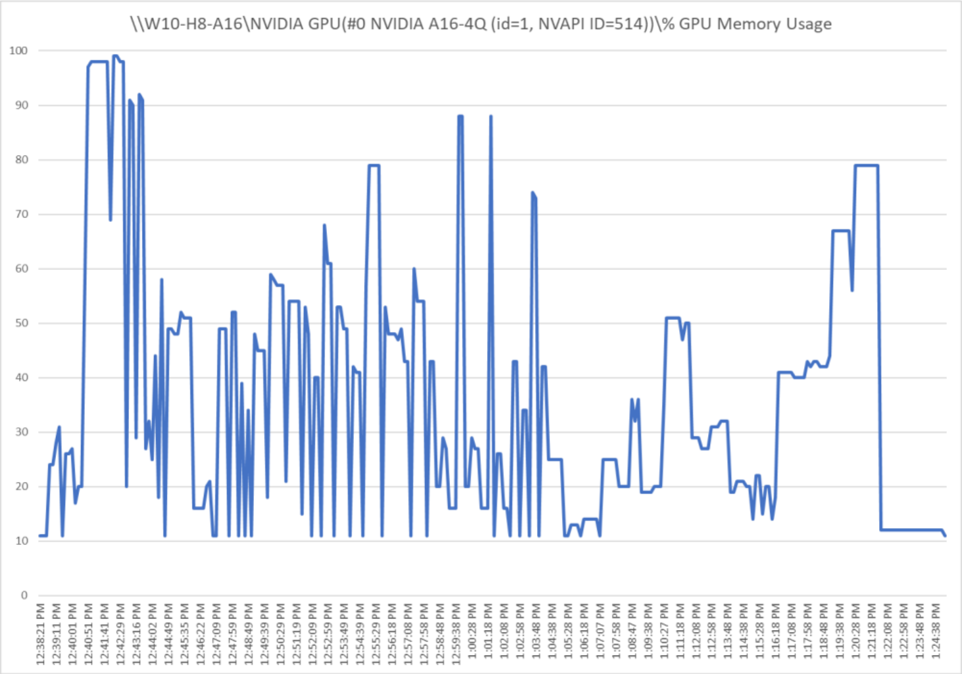


Figure 58.
Perfmon GPU memory utilization

Login VSI knowledge worker results with vPC profiles

Login VSI is designed to perform benchmarks for VDI workloads through system saturation. Login VSI loads the system with simulated user workloads using well-known desktop applications such as Microsoft Office, Internet Explorer, and Adobe PDF Reader. By gradually increasing the number of simulated users, the system will eventually be saturated. After the system is saturated, the response time of the applications will increase significantly. This latency in application response times show a clear indication whether the system is (close to being) overloaded. As a result, by nearly overloading a system it is possible to find out what its true maximum user capacity is.

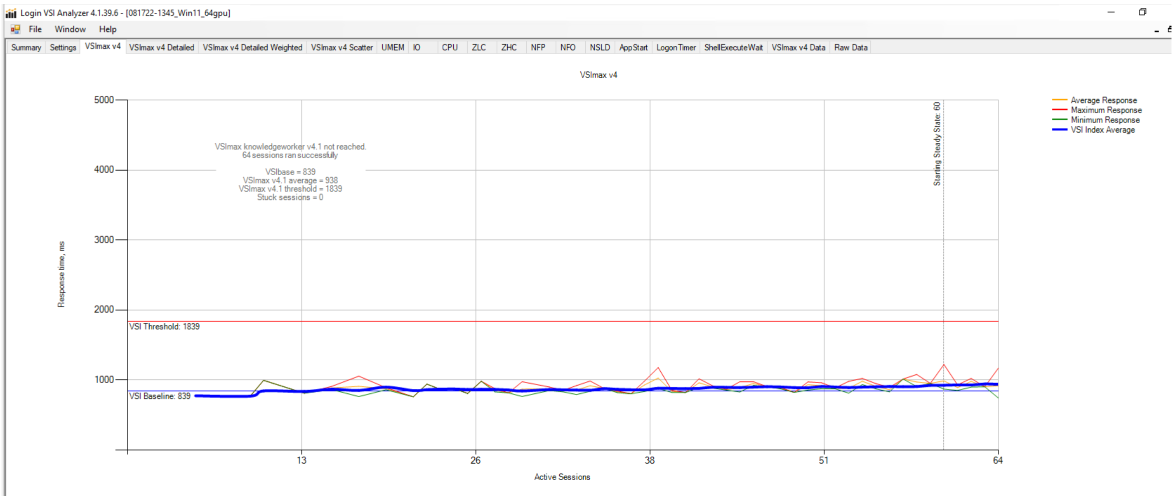
After a test is performed, the response times can be analyzed to calculate the maximum active session or desktop capacity. Within Login VSI, this time is calculated as VSImax. As the system comes closer to its saturation point, response times will rise. When you review the average response time, you will see that the response times escalate at the saturation point.

This section provides examples of the tests run on a single server designed to fully utilize the NVIDIA cards with 64 users.

Single server with a two NVIDIA A16 knowledge worker workload for an instant-clone single-session OS with random sessions with 64 users and the 2B profile

Figures 59 through 62 show the results for a knowledge worker workload on a Cisco UCS X210c M6 server with dual Intel Xeon Platinum 8358 2.60-GHz 32-core processors and 2 TB of 3200-MHz RAM running Microsoft Windows 11 64-bit and Office 2021 nonpersistent instant-clone virtual machines with two vCPUs and 4 GB of RAM.

Note: No VMware optimizations were applied to the desktop image.



VSImax knowledgeworker v4.1 not reached.
64 sessions ran successfully

VSIbase = 839
VSImax v4.1 average = 938
VSImax v4.1 threshold = 1839
Stuck sessions = 0

Figure 59.
Login VSI response chart

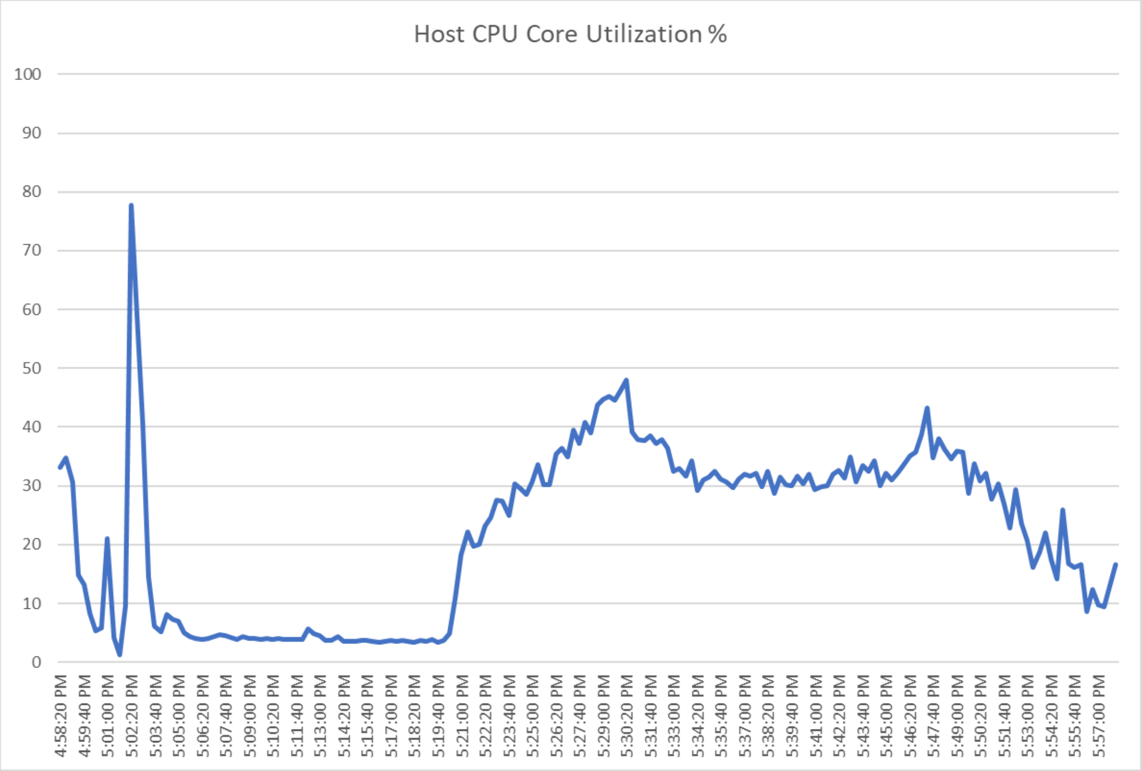


Figure 60.
Host CPU utilization

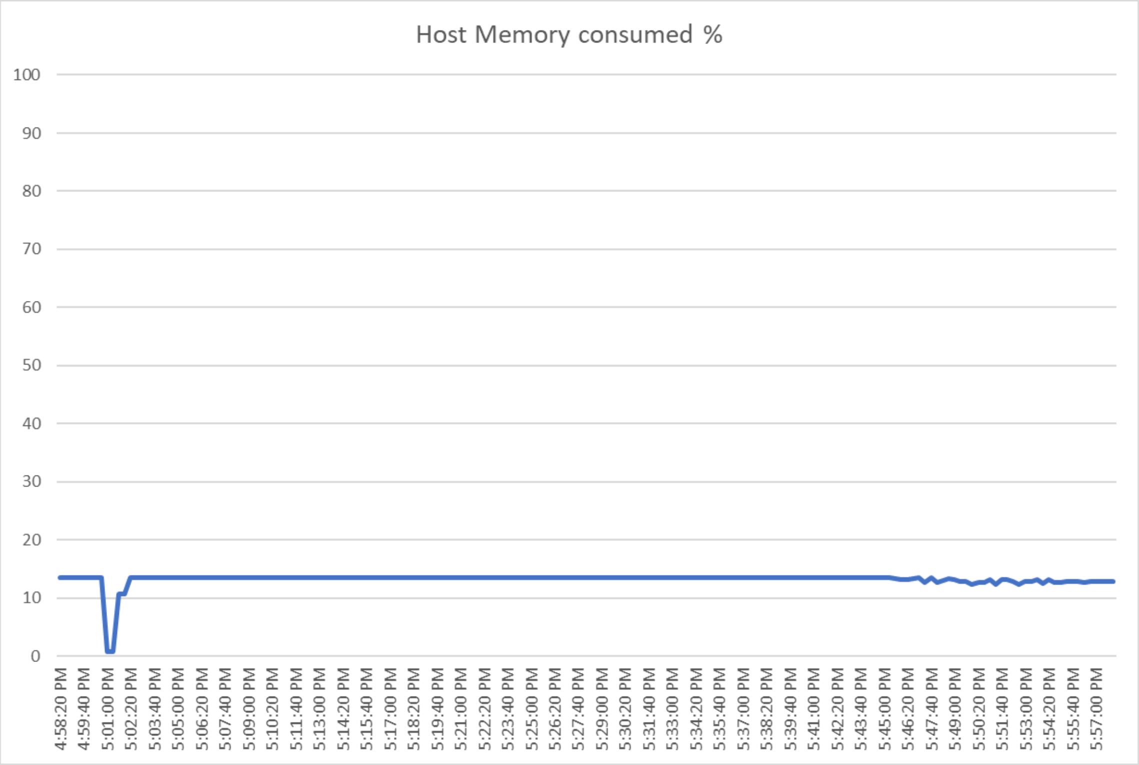


Figure 61.
Host memory utilization

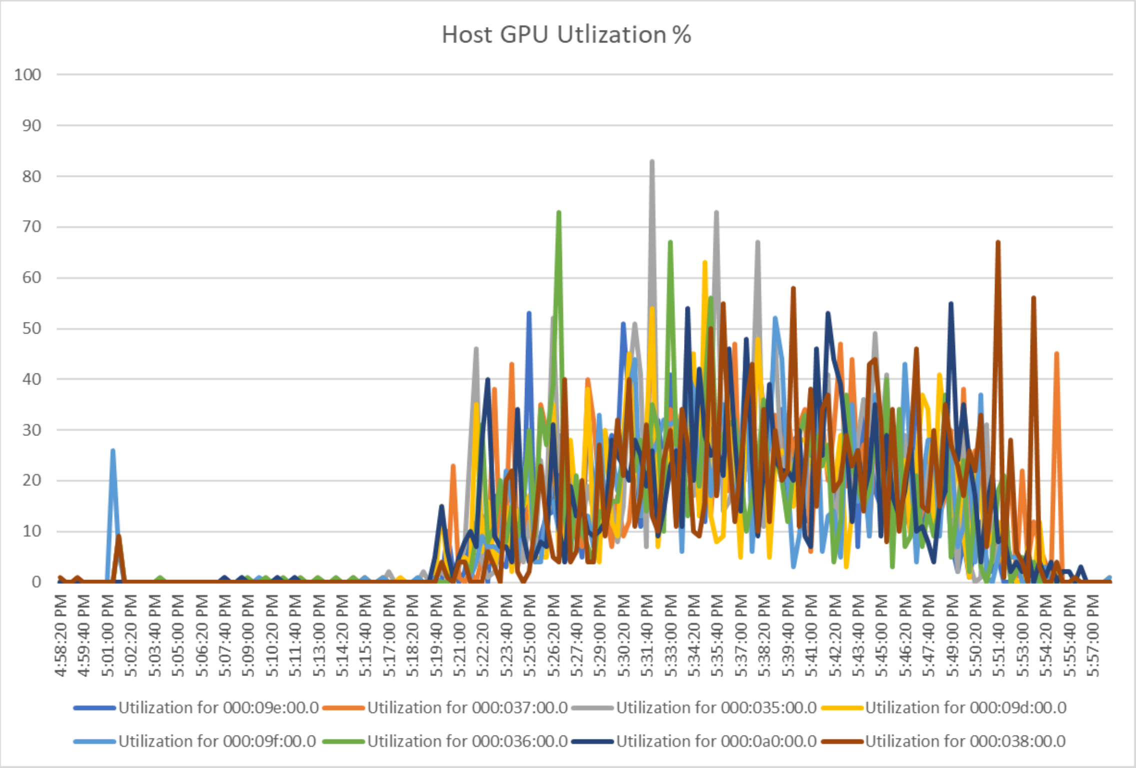
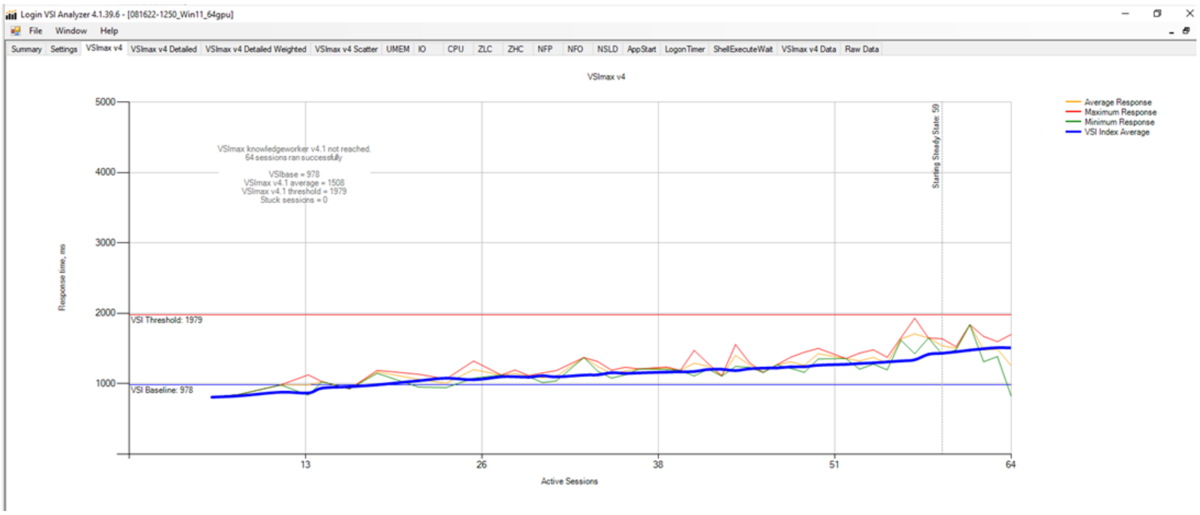


Figure 62.
Host GPU utilization

Single server with a four NVIDIA T4 knowledge worker workload for an instant-clone single-session OS with random sessions with 64 users and the 1B profile

Figures 63 through 66 show the results for a knowledge worker workload on a Cisco UCS X210c M6 server with dual Intel Xeon Gold 6348 2.60-GHz 28-core processors and 1 TB of 3200-MHz RAM running Microsoft Windows 11 64-bit and Office 2021 nonpersistent instant-clone virtual machines with two vCPUs and 4 GB of RAM.

Note: No VMware optimizations were applied to the desktop image.



VSImax knowledgeworker v4.1 not reached.
64 sessions ran successfully

VSIbase = 978
VSImax v4.1 average = 1508
VSImax v4.1 threshold = 1979
Stuck sessions = 0

Figure 63.
Login VSI response chart

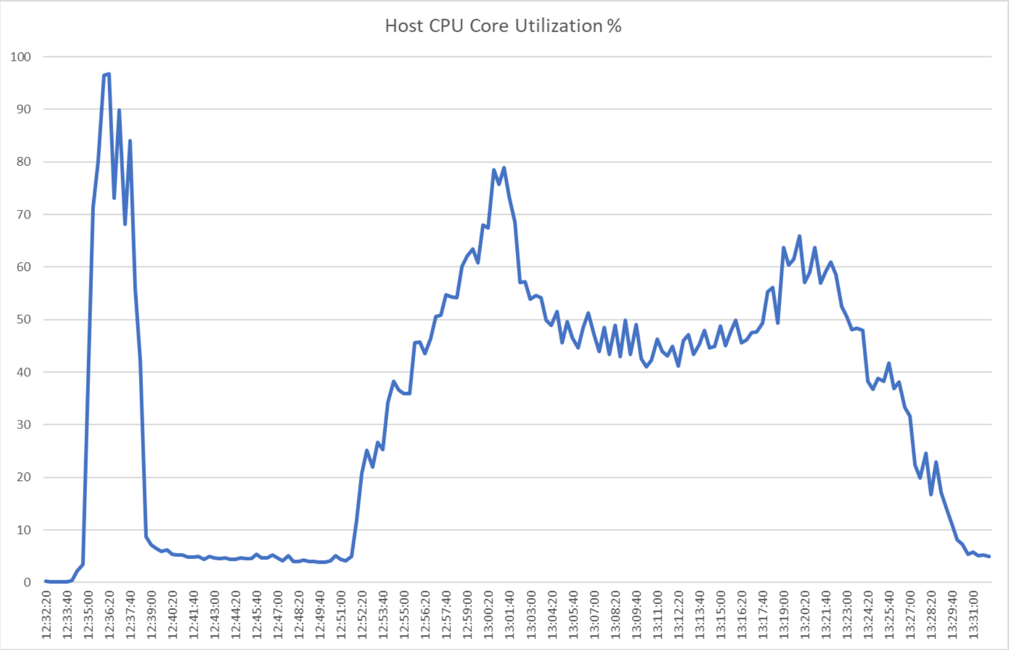


Figure 64.
Host CPU utilization

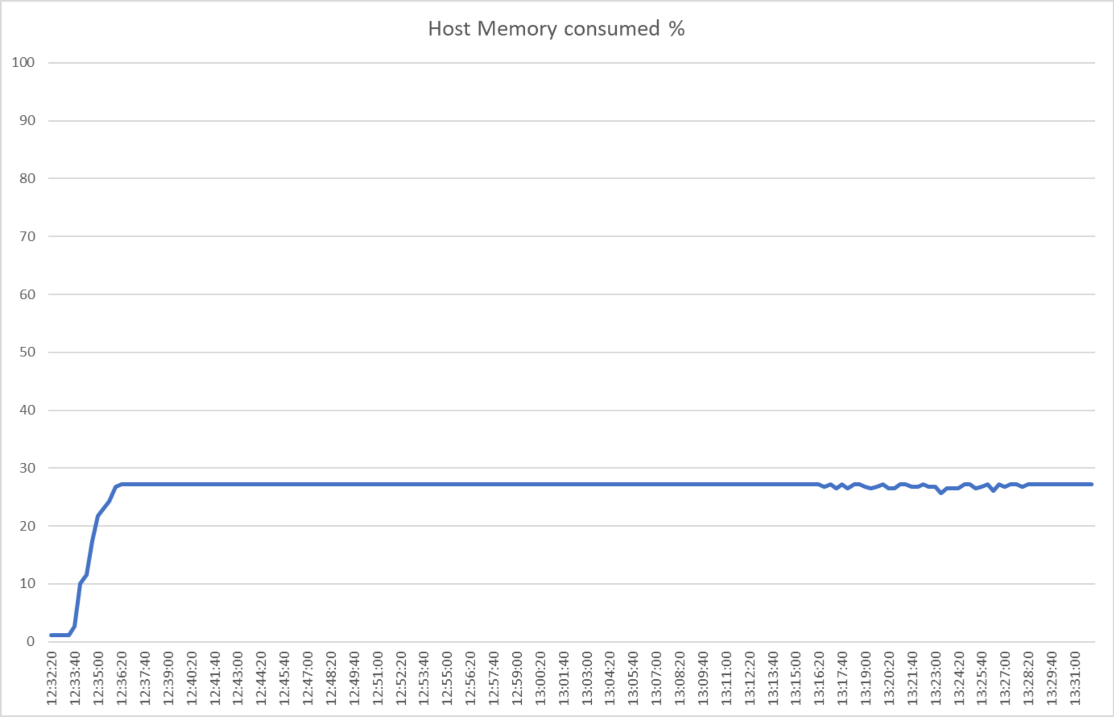


Figure 65.
Host memory utilization

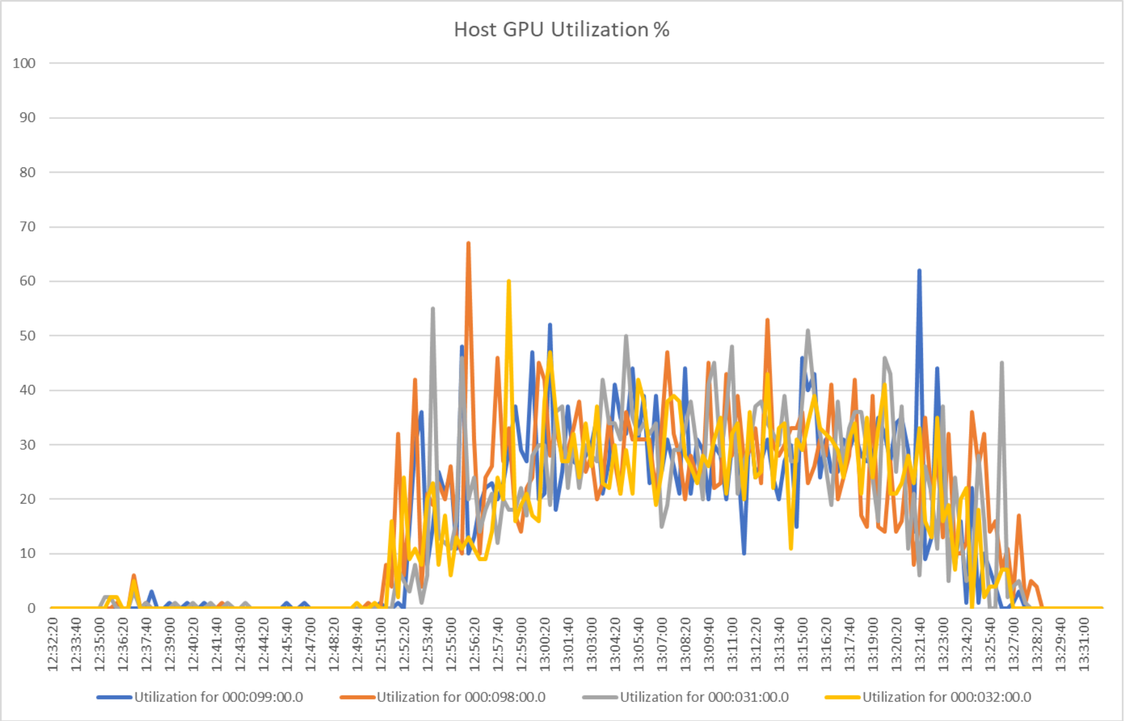


Figure 66.
Host GPU utilization

Conclusion

The Cisco UCS X-Series is designed with modularity and workload flexibility. It allows you to manage the different lifecycles of CPU and GPU components easily and independently. The Cisco UCS X210c M6 Compute Node with Intel Xeon CPUs provides a highly capable platform for enterprise end-user computing deployments, and the Cisco UCS X440p PCIe Node allows you to add up to four GPUs to a Cisco UCS X210c Compute Node with Cisco UCS X-Fabric technology. This new Cisco architecture simplifies the addition, removal, and upgrading of GPUs to computing nodes, and in conjunction with the NVIDIA GRID-enabled graphics cards, allows end-users to benefit more readily from GPU-accelerated VDI.

For more information

For additional information about topics discussed in this document, see the following resources:

- cisco.com/go/ucsx
- cisco.com/go/intersight
- cisco.com/go/vdi

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Printed in USA 222200.2 10/22