



Technical Report

NetApp HCI for Virtual Desktop Infrastructure with VMware Horizon 7

Empower Your Power Users with 3D Graphics

Suresh Thoppay, NetApp
August 2019 | TR-4792

Abstract

This technical report provides guidance for the use of the NetApp® HCI 615C for 3D graphics workloads in a VMware Horizon environment powered by NVIDIA graphics processing units (GPUs) and virtualization software. We also provide the results from the preliminary testing of SPECviewperf 13 for the NetApp HCI 615C.

TABLE OF CONTENTS	1
1 Executive Summary.....	6
2 NetApp HCI.....	6
2.1 Storage Nodes	7
2.2 Compute Nodes	7
3 NVIDIA Licensing.....	8
3.1 GRID vPC	9
3.2 GRID vApps	9
3.3 Quadro vDWS.....	9
3.4 NVIDIA vCS	9
4 3D Workloads on VMware vSphere	9
4.1 Virtual Dedicated Graphics	9
4.2 Virtual Shared Graphics	11
5 VMware Horizon.....	19
5.1 Solution Reference Architecture	19
5.2 VMware Horizon Connection Servers	19
5.3 Horizon Client	19
5.4 App Volumes.....	20
5.5 User Environment Manager	20
5.6 SMB File Shares	20
5.7 Just-in-Time Management Platform	21
5.8 Universal Access Gateway	21
5.9 VMware vRealize Log Insight.....	21
5.10 VMware vRealize Operations.....	21
6 SPECviewperf 13 Benchmark.....	21
6.1 Overview	21
6.2 3DS Max (3dsmax-06)	23
6.3 Catia (catia-05)	29
6.4 Creo (creo-02).....	35
6.5 Energy (energy-02)	41
6.6 Maya (maya-05).....	47
6.7 Medical (medical-02).....	53
6.8 Showcase (showcase-02)	59
6.9 Siemens NX (snx-03)	64

6.10 Solidworks (sw-04).....	69
7 Integration with NetApp Private Cloud.....	74
8 Integration with NetApp Kubernetes Service	74
9 Summary	77
Appendix A: GPU in vSphere	78
Appendix B: T4 CUDA-Z Screenshots	79
Appendix C: T4 vGPU Settings	80
Where to Find Additional Information	84
Version History	86

LIST OF TABLES

Table 1) vDGA - H610C versus H615C.....	11
Table 2) T4 vGPU profiles.	14
Table 3) NVIDIA T4 vGPU profiles on single GPU.	17
Table 4) Incorrect vGPU profile mix in single GPU.....	17
Table 5) vGPU - H610C versus H615C.....	18
Table 6) Hardware list.	22
Table 7) Software list.....	23

LIST OF FIGURES

Figure 1) HCI components.	7
Figure 2) Front view of H615C.	7
Figure 3) NVIDIA Tesla GPUs.....	8
Figure 4) 3D Graphics on a VM.....	12
Figure 5) NVIDIA vGPU architecture.....	13
Figure 6) Shared Direct mode.	13
Figure 7) Selecting the GPU profile.....	16
Figure 8) Group VMs on GPU until full.	18
Figure 9) Solution reference architecture.	19
Figure 10) Enable HEVC.....	20
Figure 11) SPECviewperf 13 viewsets.	22
Figure 12) 3DS Max composite score.	24
Figure 13) 3DS Max vSphere utilization – 1x16Q.....	25
Figure 14) 3DS Max vSphere utilization - 12x4Q.	26
Figure 15) 3DS Max GPU utilization - 1x16Q.....	27
Figure 16) 3DS Max GPU utilization - 12x4Q.....	28

Figure 17) 3DS Max sample.....	29
Figure 18) Catia composite score.....	30
Figure 19) Catia vSphere CPU utilization – 1x16Q.	31
Figure 20) Catia vSphere CPU utilization – 12x4Q.	32
Figure 21) Catia GPU utilization - 1x16Q.	33
Figure 22) Catia GPU utilization - 12x4Q.	34
Figure 23) Catia sample.	35
Figure 24) Creo composite score.	36
Figure 25) Creo vSphere CPU utilization - 1x16Q.	37
Figure 26) Creo vSphere CPU utilization - 12x4Q.	38
Figure 27) Creo GPU utilization - 1x16Q.	39
Figure 28) Creo GPU utilization - 12x4Q.	40
Figure 29) Creo Sample.	40
Figure 30) Energy composite score.....	42
Figure 31) Energy vSphere CPU utilization - 1x16Q.	43
Figure 32) Energy vSphere CPU utilization - 12x4Q.	44
Figure 33) Energy GPU utilization - 1x16Q.	45
Figure 34) Energy GPU utilization - 12x4Q.	46
Figure 35) Energy Sample.....	47
Figure 36) Maya composite score.	48
Figure 37) Maya vSphere CPU utilization - 1x16Q.	49
Figure 38) Maya vSphere CPU utilization - 12x4Q.	50
Figure 39) Maya GPU utilization - 1x16Q.	51
Figure 40) Maya GPU utilization - 12x4Q.	52
Figure 41) Maya Sample.	52
Figure 42) Medical composite score.....	54
Figure 43) Medical vSphere CPU utilization - 1x16Q.	55
Figure 44) Medical vSphere CPU utilization - 12x4Q.	56
Figure 45) Medical GPU utilization - 1x16Q.	57
Figure 46) Medical GPU utilization - 12x4Q.	58
Figure 47) Medical sample.	58
Figure 48) Showcase composite score.	59
Figure 49) Showcase vSphere CPU utilization - 1x16Q.	60
Figure 50) Showcase vSphere CPU utilization - 12x4Q.	61
Figure 51) Showcase GPU utilization - 1x16Q.	62
Figure 52) Showcase GPU utilization - 12x4Q.	63
Figure 53) Showcase sample.	63
Figure 54) Siemens NX composite score.	64
Figure 55) Siemens NX vSphere CPU utilization - 1x16Q.	65
Figure 56) Siemens NX vSphere CPU utilization - 12x4Q.	66

Figure 57) Siemens NX GPU utilization - 1x16Q.....	67
Figure 58) Siemens NX GPU utilization - 12x4Q.....	68
Figure 59) Siemens NX Sample.....	68
Figure 60) Solidworks composite score.....	69
Figure 61) Solidworks vSphere CPU utilization - 1x16Q.....	70
Figure 62) Solidworks vSphere CPU utilization - 12x4Q.....	71
Figure 63) Solidworks GPU utilization - 1x16Q.....	72
Figure 64) Solidworks GPU utilization - 12x4Q.....	73
Figure 65) Solidworks sample.....	73
Figure 66.....	74
Figure 67) Hybrid Cloud Control.....	75
Figure 68) Enable NetApp Cloud Services.....	76
Figure 69) Bitfusion FlexDirect.....	77
Figure 70) T4 in vSphere 6.7 Update 1.....	78
Figure 71) T4 in vSphere 6.7 Update 2.....	78
Figure 72) T4 with NVIDIA vGPU.....	79
Figure 73) CUDA-Z screenshots.....	80

1 Executive Summary

Virtual desktops and hosted apps provide organizations with secure access to information from any device, including smart phones and tablets. NetApp and their partners have helped many customers to successfully implement virtual desktop infrastructure (VDI) environments. For the growing demand of hardware graphics acceleration, NetApp HCI has expanded its portfolio to address various business needs. This document is focused on a solution for virtualized 3D graphic workloads on various products from Autodesk, Dassault Systemes, Siemens, and so on.

NetApp HCI includes Intel's scalable second-generation processors, various memory configurations, and other components. You have the option to use NVIDIA M10 GPUs for scalability or NVIDIA T4 GPUs for the flexibility to run compute workloads in addition to VDI. To address storage resource demands, NetApp HCI provides iSCSI storage with efficiency features like thin provisioning, compression, and deduplication. NetApp HCI includes a Mellanox CX-4 dual-port 2x25GbE network controller and one 1GbE Baseboard Management Controller.

VMware Horizon 7 provides various implementation choices to meet the demands of virtual desktops and hosted apps. The session-sharing feature allows up to 10 users to collaborate and overcome the challenges of sharing video content.

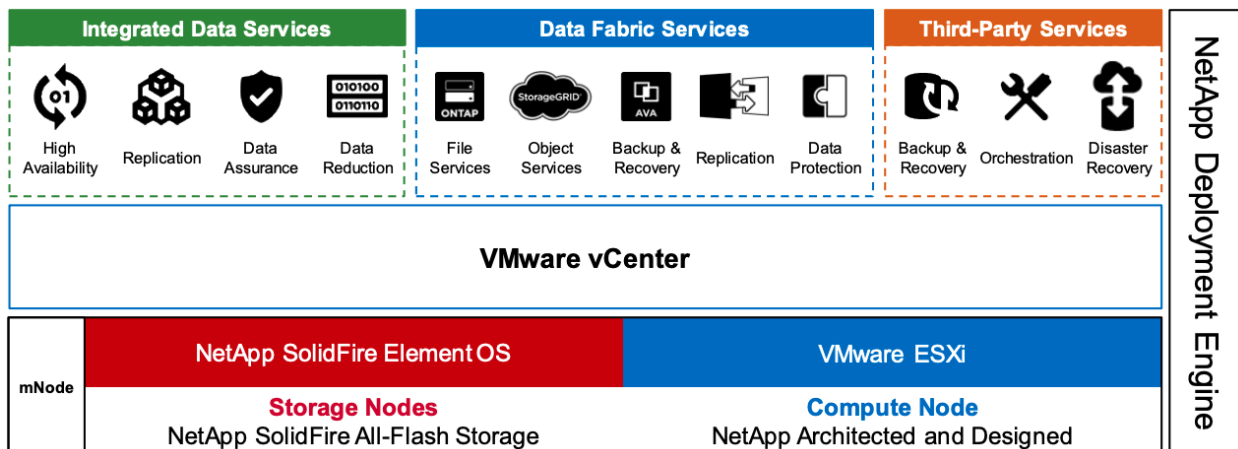
NetApp HCI not only addresses the challenges of virtualized 3D graphic workloads, it also enables you to quickly explore deep learning with the NetApp Kubernetes Service and NVIDIA GPU Cloud. With the NetApp Data Fabric, you can gain more value from NetApp HCI.

2 NetApp HCI

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

Figure 1) HCI components.



2.1 Storage Nodes

Storage nodes are available with either a half-width or a full-width rack-unit size. Half-width rack units are populated into a two-rack-unit chassis that can contain either storage or compute nodes. A minimum of four storage nodes are required to begin with, and chassis 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 the quality-of-service (QoS) feature supporting minimum, maximum, and burst limits. The storage cluster supports a mix of storage nodes, although one storage node cannot exceed one third of total capacity.

2.2 Compute Nodes

Compute nodes are available in half-width, full-width, and two rack-unit sizes. The H410C and H610C are based on Intel scalable (Skylake) processors. The H615C is based on Intel second-generation scalable (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 2) Front view of H615C.







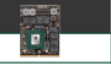


The NVIDIA T4 has 40 RT cores that provide the computation needed to deliver real-time ray tracing. The same server model used by designers and engineers can now also be used by artists to create photorealistic imagery that features light bouncing off surfaces just as it would in real life. This RTX-capable GPU produces real-time ray tracing performance of up to 5 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 anywhere.

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

[Figure 3](#) provides a list of NVIDIA GPU cards and a feature comparison.

Figure 3) NVIDIA Tesla GPUs.

NVIDIA GPUs Recommended for Virtualization					Available on NetApp HCI H615C	Available on NetApp HCI H610C	
	V100	RTX 8000	RTX 6000	P40	T4	M10	P6
							
GPU	1 NVIDIA Volta	1 NVIDIA Turing	1 NVIDIA Turing	1 NVIDIA Pascal	1 NVIDIA Turing	4 NVIDIA Maxwell	1 NVIDIA Pascal
CUDA Cores	5,120	4,608	4,608	3,840	2,560	2,560 (640 per GPU)	2,048
Tensor Cores	640	576	576	—	320	—	—
RT Cores	—	72	72	—	40	—	—
Guaranteed QoS (GPU Scheduler)	✓	✓	✓	✓	✓	—	✓
Live Migration	✓	✓	✓	✓	✓	✓	✓
Multi-vGPU	✓	✓	✓	✓	✓	✓	✓
Memory Size	32/16 GB HBM2	48 GB GDDR6	24 GB GDDR6	24 GB GDDR5	16 GB GDDR6	32 GB GDDR5 (8 GB per GPU)	16 GB GDDR5
vGPU Profiles	1 GB, 2 GB, 4 GB, 8 GB, 16 GB, 32 GB	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 16 GB, 24 GB, 48 GB	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 GB	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 GB	1 GB, 2 GB, 4 GB, 8 GB, 16 GB	0.5 GB, 1 GB, 2 GB, 4 GB, 8 GB	1 GB, 2 GB, 4 GB, 8 GB, 16 GB
Form Factor	PCIe 3.0 dual slot and SXM2 (rack servers)	PCIe 3.0 dual slot	PCIe 3.0 dual slot	PCIe 3.0 dual slot (rack servers)	PCIe 3.0 single slot (rack servers)	PCIe 3.0 dual slot (rack servers)	MXM (blade servers)
Power	250 W /300 W (SXM2)	295 W	295 W	250 W	70 W	225 W	90 W
Thermal	passive	active	active	passive	passive	passive	bare board
Use Case	Ultra-high-end rendering, simulation, 3D design with Quadro vDWS; ideal upgrade path for P100	High-end rendering, 3D design and creative workflows with Quadro vDWS	Mid-range to high-end rendering, 3D design and creative workflows with Quadro vDWS	Mid-range to high-end rendering, 3D design and engineering workflows with Quadro vDWS	Entry-level to high-end 3D design and engineering workflows with Quadro vDWS. High-density, low power GPU acceleration for knowledge workers with NVIDIA GRID software.	Knowledge workers using modern productivity apps and Windows 10 requiring best density and total cost of ownership (TCO), multimonitor support with NVIDIA GRID vPC/vApps	For customers requiring GPUs in a blade server form factor; ideal upgrade path for M6

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 different use cases. This includes using virtual workstations to improve graphics performance, real-time interactive rendering, or inferencing. With the T4, IT can take advantage of the same GPU resources to run mixed workloads, including running VDI during the day and repurposing the resources to run compute workloads at night.

The compute node H610C is two rack units in size whereas 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 because 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 and is currently 64. The mixing of different models of compute nodes in a cluster is supported when Enhanced vMotion Compatibility (EVC) is enabled. For the GPU nodes with default graphics settings (virtual shared graphics mode), compute models can be mixed in a cluster.

3 NVIDIA Licensing

When using a NetApp HCI H610C or H615C, the license for the GPU must be procured from NVIDIA partners that are authorized to re-sell the licenses. You can find NVIDIA partners with the [partner locator](#). Search for competencies such as vGPU or Tesla.

NVIDIA vGPU software is available in four editions:

- NVIDIA GRID Virtual PC (GRID vPC)
- GRID Virtual Applications (GRID vApps)
- NVIDIA Quadro vDWS
- NVIDIA Virtual ComputeServer (NVIDIA vCS)

3.1 GRID vPC

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

3.2 GRID vApps

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

3.3 Quadro vDWS

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. An NVIDIA Quadro Virtual Data Center workstation allows users to access their professional graphics applications with full features and performance anywhere on any device.

3.4 NVIDIA vCS

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 ECC, page retirement, peer-to-peer over NVLink, and multi-vGPU.

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

4 3D Workloads on VMware vSphere

VMware vSphere became a trusted platform for virtualization by effectively utilizing 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 file to create a baseline of the type Host Extension.

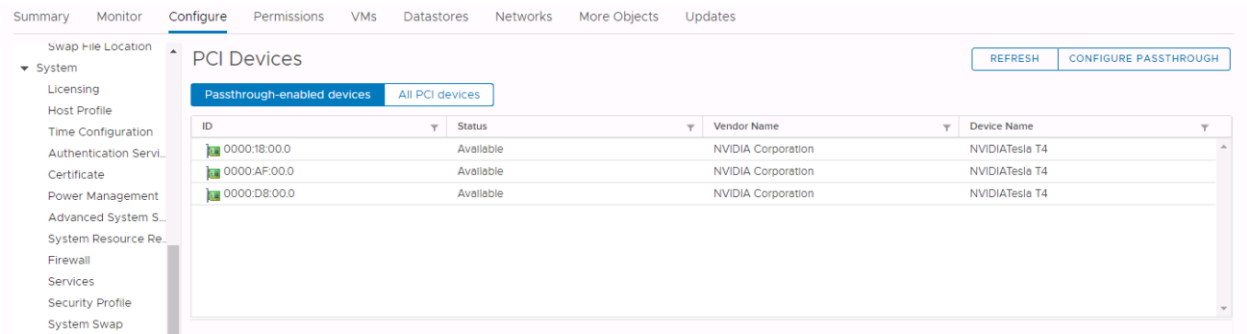
VMs consume GPU resources in one of the following ways:

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

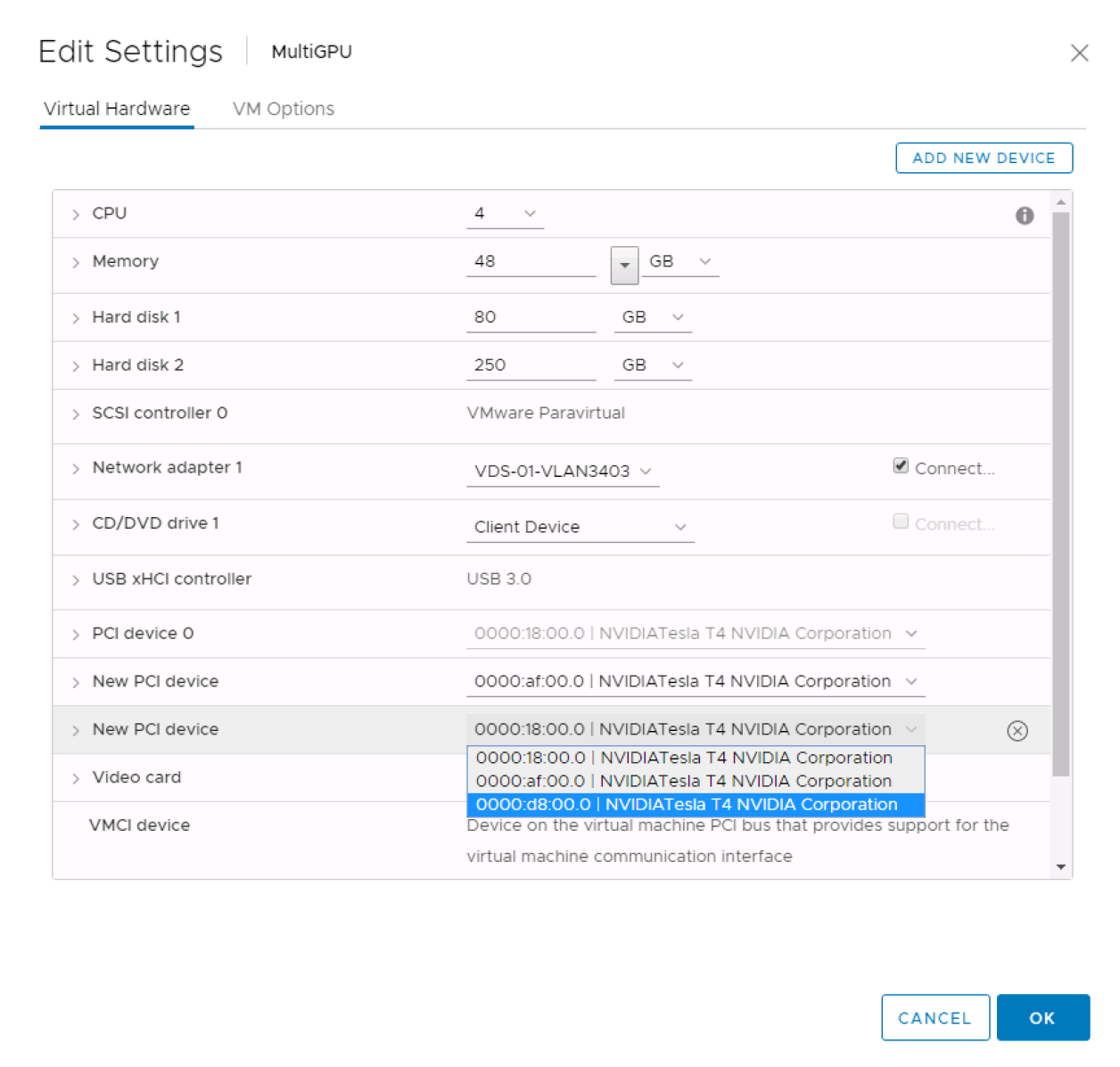
4.1 Virtual Dedicated Graphics

A VM has complete access to the GPU using the PCI pass-through option using Direct IO. However, certain vSphere features, like snapshots, vMotion, and so on, are not supported. The VM gets native performance results.

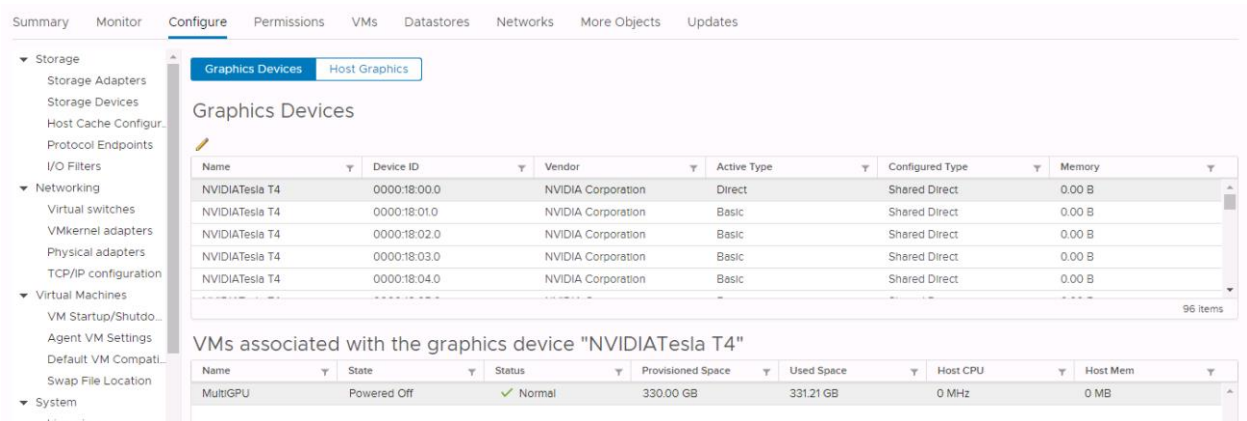
1. To configure vDGA, click PCI Devices under the Hardware section and click Configure Passthrough.



- Then, on the Edit Settings page, add the PCI device.



- When the GPU is configured for vDGA mode, the graphic devices listing shows the Active Type as Direct.



Until vSphere 6.7 update 2, vDGA was the only option among the three (vDGA, vSGA, or vGPU) listed where the VM can have access to 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 in and check out licenses. Without a license, a very limited feature set is available.

Table 1) vDGA - H610C versus H615C.

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

4.2 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 4) 3D Graphics on a VM.

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

Virtual Hardware VM Options

ADD NEW DEVICE

> New USB Controller	USB 3.0
Video card *	Specify custom settings
Number of displays	2
Total video memory	256 MB
3D Graphics	<input checked="" type="checkbox"/> Enable 3D Support
3D Renderer	Automatic
3D Memory	Automatic MB
VMCI device	Device on the virtual machine PCI bus that provides support for the virtual machine communication interface

Compatibility: ESXi 6.7 Update 2 and later (VM version 15)

CANCEL BACK NEXT

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 EVC is enabled.

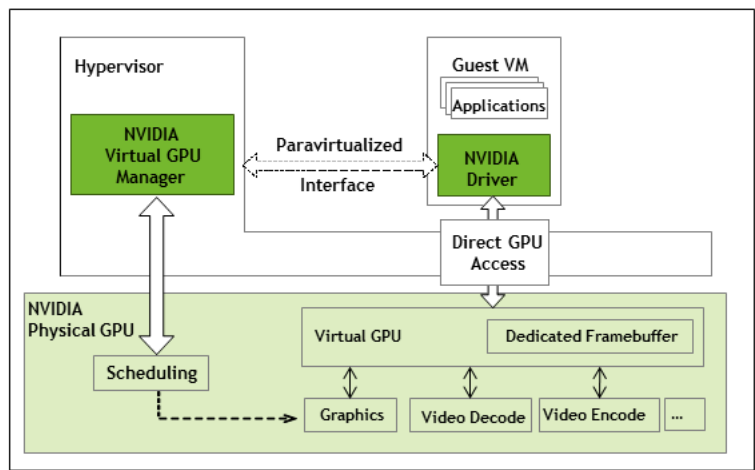
vDGA uses PCI pass through for a GPU card to a given VM. NetApp does not recommend vDGA because it does not support many vSphere features, and it provides low resource utilization. vSGA allows multiple VMs to use physical GPUs installed locally in the ESXi hosts and provide hardware-accelerated 3D graphics. NetApp does not recommend vSGA because graphics API support is limited and not all versions of DirectX and OpenGL are supported. Additionally, there is no CUDA support for vSGA. With vSGA, the VDI instance relies on a VMware vSGA driver that gets access through an Xorg server running on the hypervisor. This is suboptimal from a performance standpoint. A virtual GPU uses NVIDIA technology throughout, and a VDI instance provides the closest possible parity to running a native NVIDIA driver.

Virtual Shared Passthrough Graphics

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

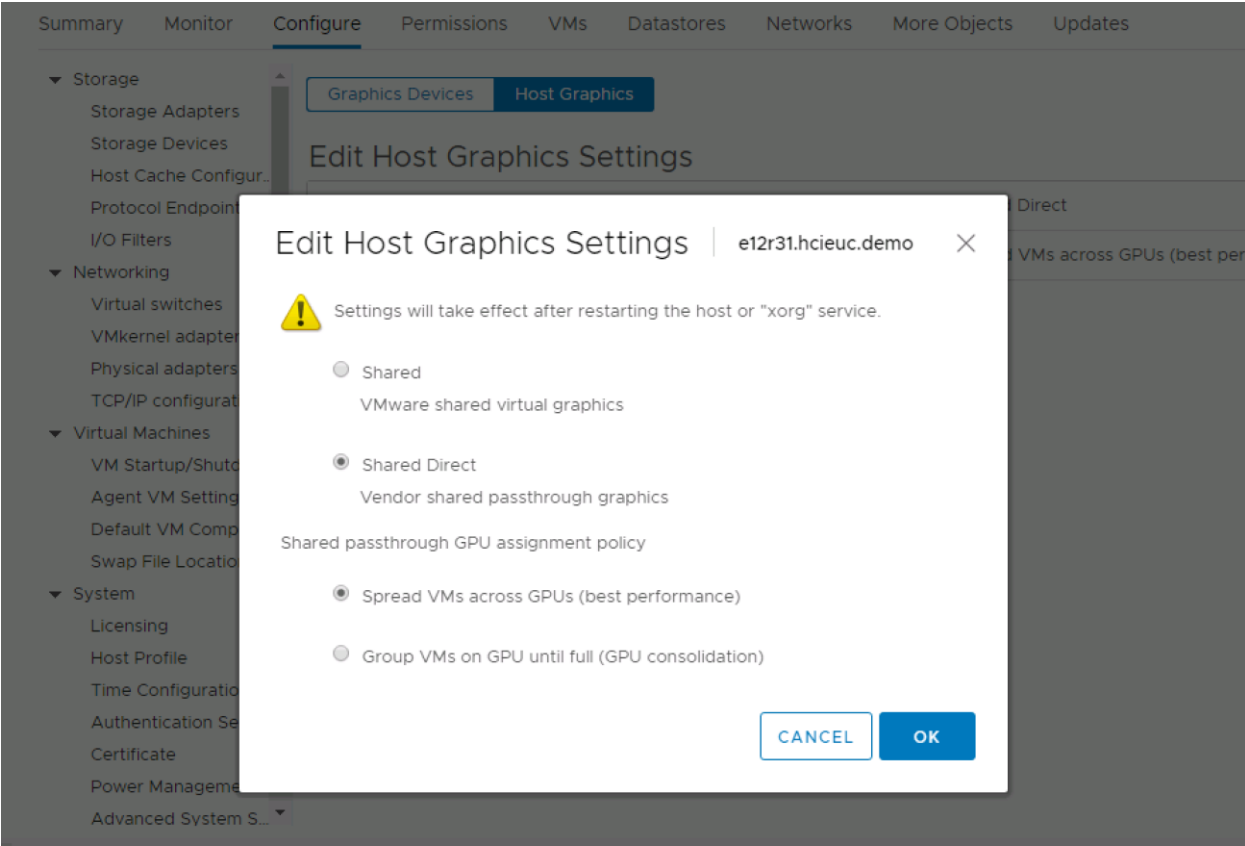
manner similar to a hypervisor sharing resources with VMs, NVIDIA GRID software manages the resource scheduling of GPU requests from a VM.

Figure 5) NVIDIA vGPU architecture.



To use Virtual Shared Passthrough Graphics mode, the host graphics setting must be changed to Shared Direct mode, as shown in [Figure 6](#).

Figure 6) Shared Direct mode.



In VM Settings, add Shared PCI Device and pick the required vGPU profile ([Table 2](#)) based on the frame buffer. You must reserve memory to enable direct access for the PCI device. [Table 2](#) lists details for the T4 vGPU profiles that can be selected in the following screenshot.

Table 2) T4 vGPU profiles.

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 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 Virtual PC or Quadro vDWS
T4-2B4	Virtual desktops	2048	4	2560x1600	8	24	GRID Virtual PC 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 Server	Required License Edition
T4-1B	Virtual desktops	1024	4	2560x1600	16	48	GRID Virtual PC or Quadro vDWS
T4-1B4	Virtual desktops	1024	1	4096x2160	16	48	GRID Virtual PC or Quadro vDWS
T4-16A	Virtual applications	16384	1	1280x1024	1	3	GRID Virtual Application
T4-8A	Virtual applications	8192	1	1280x1024	2	6	GRID Virtual Application
T4-4A	Virtual applications	4096	1	1280x1024	4	12	GRID Virtual Application
T4-2A	Virtual applications	2048	1	1280x1024	8	24	GRID Virtual Application
TT4-1A	Virtual applications	1024	1	1280x1024	16	48	GRID Virtual Application

For NVIDIA vGPU mode, do not enable 3D support on the VM. For more information, see the NVIDIA vGPU User Guide.

A GRID virtual PC (profiles typically end with B) is used for virtual desktops, and a GRID virtual application (profiles end with A) is used for hosted apps. Most 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 Virtual PC (GRID vPC) 1GB profile to deliver an optimal experience for standard knowledge workers:

- Heavy application use like browsing, email, and creating complex documents, presentations, and spreadsheets.

- Applications including Microsoft Windows 10, Microsoft Office productivity apps, streaming video, and multimedia using the latest web standards like WebGL.
- Up to four 2K (2560×1600) 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 including Microsoft Windows 10, Microsoft Office productivity apps, video, multimedia, and industry specific apps like Bloomberg, Thomson Reuters Eikon, and DICOM viewers.
- Up to two 4096 x 2160 resolution monitors.

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

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

Figure 7) Selecting the GPU profile.

The screenshot shows the 'New Virtual Machine' wizard in VMware Workstation, specifically the 'Customize hardware' step. The 'Virtual Hardware' tab is selected, and a 'New PCI device' is being added. The 'GPU Profile' dropdown menu is open, showing a list of profiles. The 'grid_t4-8q' profile is selected. The 'Video card' section is also visible, showing a 'Video card *' entry. The 'Next' button is highlighted in blue.

NVIDIA supports the same vGPU profiles that are available on a GPU ([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

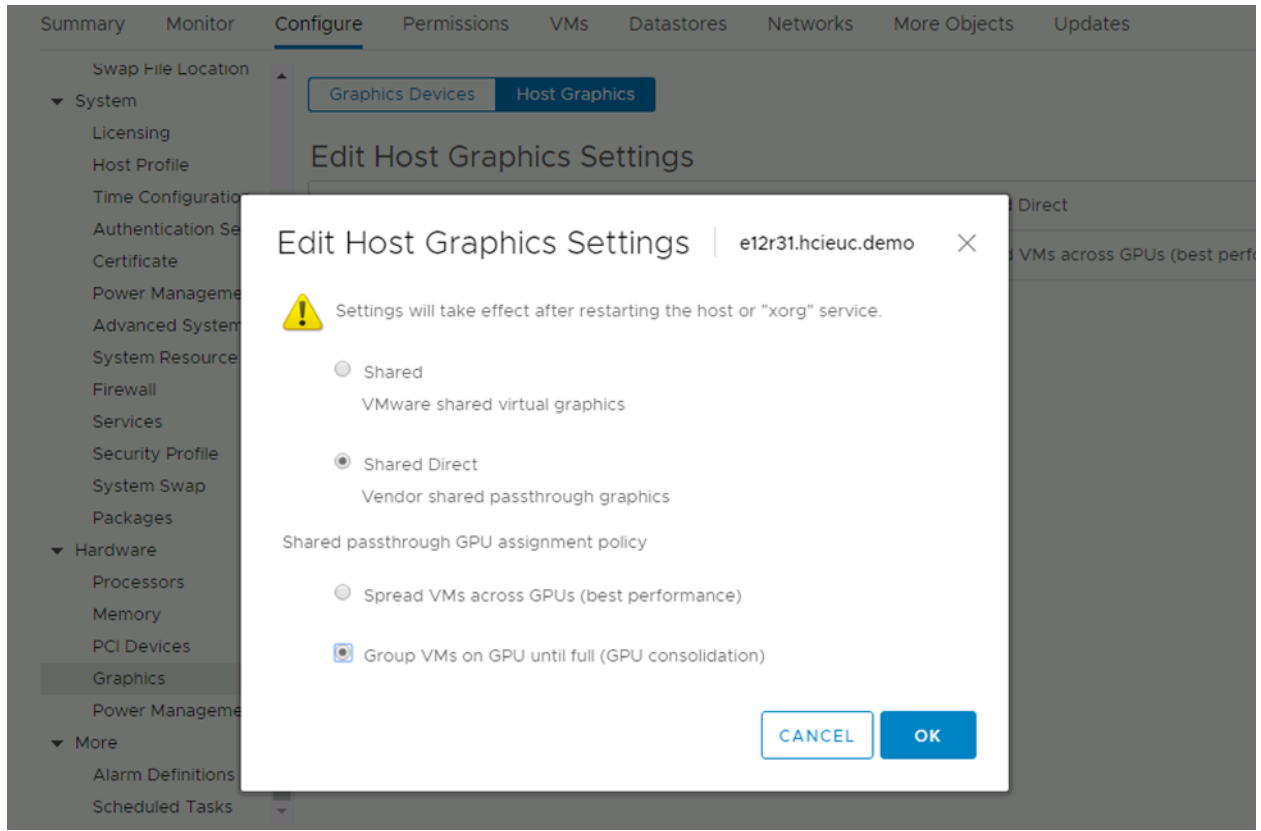
NVIDIA vGPU software does not support mixing profiles for a GPU accelerator with only one GPU. For example, if a VM with the 4Q profile is running on a GPU, it can only host another VM that has the same 4Q profile on the same GPU.

Table 4) Incorrect vGPU profile mix in single GPU.

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

Because vGPU profiles on a H610C differs from the profiles on a 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 desire to mix the nodes in cluster, consider changing the GPU assignment policy to Group VMs on GPU ([Figure 8](#)). You need to have enough resources for a take-over if there is a node failure on either of the GPU nodes.

Figure 8) Group VMs on GPU until full.



After the NVIDIA driver is installed on a VM, the VM console presents a blank screen for a 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 versus 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
16GB	NA	3	6

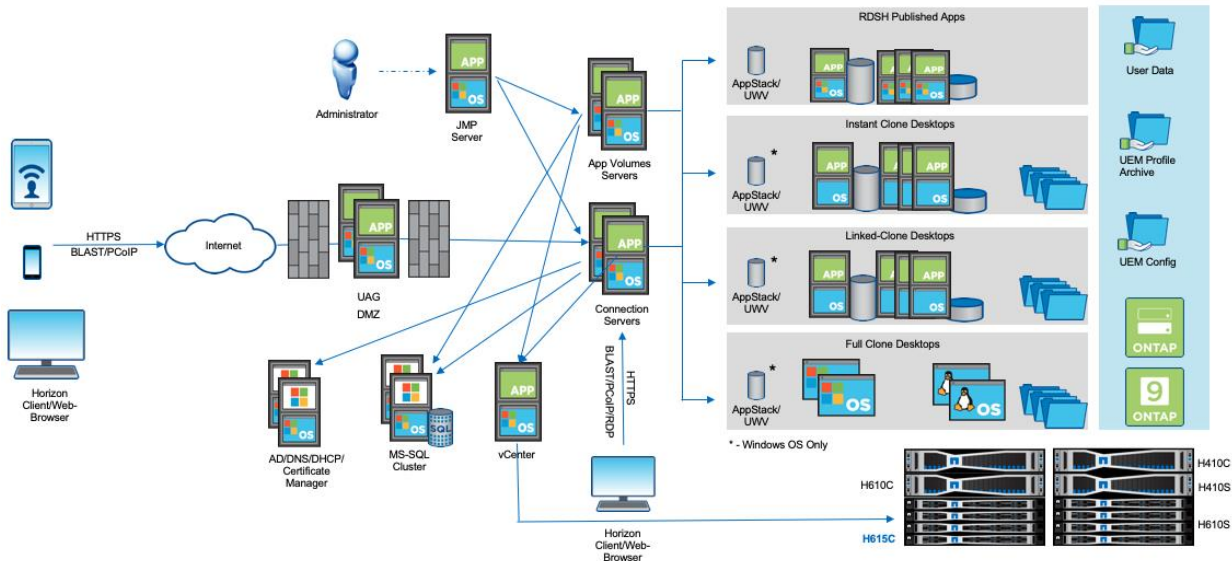
The H615C can host 50% more density for the same rack unit size and provide double the performance for most workloads.

5 VMware Horizon

5.1 Solution Reference Architecture

A typical VMware Horizon reference architecture with NetApp HCI is shown in [Figure 9](#).

Figure 9) Solution reference architecture.



5.2 VMware Horizon Connection Servers

VMware Horizon is a connection broker for virtual desktops and hosted applications. VMware Horizon can perform automated provisioning of full desktop clones, linked clones, or instant clones. It can manage physical machines as well as virtual machines. Server-based computing is enabled with RDSHs that can be auto provisioned with linked clones or instant clones.

Instant access to applications can be a requirement for knowledge workers or in scenarios like hospitals or kiosks, where it is important to provide a native PC-like experience on any device. Hosted Applications on a remote desktop session server farm can help with these types of use cases. VMware Horizon 7.9 added support for VM-hosted applications, which enables you to host Windows 10 Universal Windows Platform (UWP) applications.

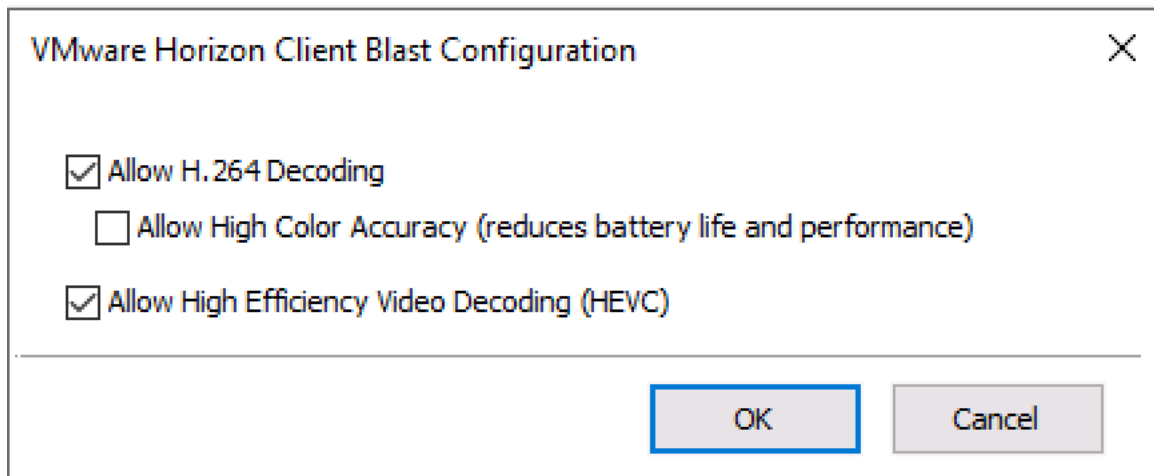
Note: VMware Horizon 7.9 added support for NVIDIA T4. That's the minimum version required for use with the NetApp HCI H615C.

Hosted Apps can help consolidate applications that require high rendering on a few H615C hosts. It can also be useful for Linux users who need to access application hosts on Windows.

5.3 Horizon Client

VMware Horizon supports various display protocols, including RDP, PCoIP, and BLAST. Support for BLAST Extreme with H.264 is enabled by default. VMware Horizon uses a hardware decoder if available. Otherwise, it uses software-based decoding. H.265, which is also referred to as HEVC, is an option on Horizon Client. Using HEVC reduces network traffic for high resolution displays. See [Figure 10](#) to enable HEVC.

Figure 10) Enable HEVC.



HTML clients should use the Google Chrome browser, which has support for HEVC. Horizon agents installed on virtual desktops perform encoding, whereas the Horizon Client on end devices like smart phones, tablets, laptops, and so on perform decoding.

5.4 App Volumes

The number of VM templates and images to manage tends to grow quickly because of the various operating system images, patches, and multiple versions of applications. App Volumes helps to decouple application and user data from operating system images. An application or set of applications are contained in a virtual disk. Those virtual disks can be mounted from the volume presented to the vSphere host or they can be mounted directly inside the VM. App Volume configuration is hosted on a Microsoft SQL database, and multiple App Volume servers can access the data and provide high availability.

5.5 User Environment Manager

User Environment Manager (UEM) manages user profiles and user application settings like the application layout, appearance, language preference, and so on. UEM stores configurations on an SMB file share. VMware Horizon client policies can be enforced using various conditions that can be defined within UEM. For example, USB redirection can be allowed in the office, but it is not allowed in other locations.

5.6 SMB File Shares

UEM requires SMB file shares to back up user profiles and to host configuration files. App Volumes requires SMB shares if virtual disks must be mounted inside the VM. NetApp HCI provides file services using NetApp ONTAP® Select, which provides the following features useful for VMware Horizon implementations:

- User home directories
- SVMs (Storage Virtual Machines)
- NetApp FlexGroup technology
- NetApp FabricPool technology
- NetApp Data Availability Services
- Adaptive QoS
- Deduplication
- Data protection features

- NetApp RAID-DP® technology
- NetApp Snapshot™ copies
- NetApp FlexClone® technology
- NetApp SnapMirror® and SnapVault® technology
- Self restore from Snapshot copies

5.7 Just-in-Time Management Platform

The Just-in-Time Management Platform (JMP) server hosts the new administrative interface for VMware Horizon that enables you to manage desktop and application provisioning, assign to users and groups, and associate policies defined with UEM.

5.8 Universal Access Gateway

The Universal Access Gateway provides users with secure access to their virtual desktops and applications from external networks.

5.9 VMware vRealize Log Insight

VMware vRealize Log Insight collects and indexes log files centrally. It allows administrators to quickly search multiple log files and identify any issues that might arise.

5.10 VMware vRealize Operations

VMware vRealize Operations (vROPS) provides central monitoring of infrastructure components. The NetApp HCI Plug-in for vROPS is available from Blue-Medora, and NVIDIA provides a management pack for monitoring NVIDIA GPUs. NVIDIA virtual GPU management support for vROPS provides a holistic view of your GPU-enabled environment at the cluster, host, or VM level, enabling real-time monitoring and proactive management.

6 SPECviewperf 13 Benchmark

6.1 Overview

The Standard Performance Evaluation Corporation (SPEC) is a non-profit 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 towards specific tools and requires vendor licenses. Here is a sample list of SPECapc tools:

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

SPECviewperf and SPECworkstation both measure graphics performance based on professional applications. These benchmarks measure 3D graphics performance using OpenGL and Direct X application programming interfaces. The benchmark's workloads are known as viewsets and represent graphics content and behavior from actual applications.

SPECviewperf is geared towards 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 11](#) shows SPECviewperf 13 viewsets.

Figure 11) SPECviewperf 13 viewsets.



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

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

The hardware and software used for testing are presented in [Table 6](#) and [Table 7](#).

Table 6) Hardware list.

Model	Count	Description
H410S	4	Storage cluster

Model	Count	Description
H410C	2	Infrastructure cluster
H615C (2x18C, 382 GB)	2	Desktop resource pool
Mellanox SN2010 switches	4	25Gb switch

Table 7) Software list.

Product	Description
Microsoft Windows 2016	Server OS for AD, DNS, Certificate Server, etc.
Microsoft Windows 2019	Server OS for VMware Horizon 7.9
VMware vSphere 6.7 Update 2 Build 13006603	Hypervisor
VMware vCenter 6.7 Update 2b	Management
NVIDIA vGPU Manager Windows Driver	9.0 430.27 431.02
VMware Horizon 7.9 VMware Horizon Client 5.1.0 build-14045148	Connection Broker Client Software to access desktops
SPECviewperf 13	Benchmarking Tool
Microsoft Windows 10 Pro Workstation Version 1903	Desktop OS

Note: Although we had access to 25Gb switches, we tested at 10Gb, which is more common.

For every viewset, we performed a test with 1 VM with the 16Q profile (full frame buffer on a single T4) and 12 VMs with the 4Q profile (full load on an H615C with three T4 cards). Performance characteristics of vSphere CPU utilization and GPU utilization during workload simulation are documented below. For a single VM test, we used the VM configuration of 18 vCPU with 16GB RAM. For a 12 VM test, we used the VM configuration of 6 vCPU with 16GB RAM. For the tests, we disabled the frame rate limit. The client VM configuration was 4 vCPUs with 4GB RAM and a GPU with the 1Q profile.

SPECviewperf measures the frame rate, measured in 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, and each composite score represented in the graphs below is based on a weighted geometric mean of many different scenes and rendering modes.

6.2 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 lesser-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.

The following viewset tests were run:

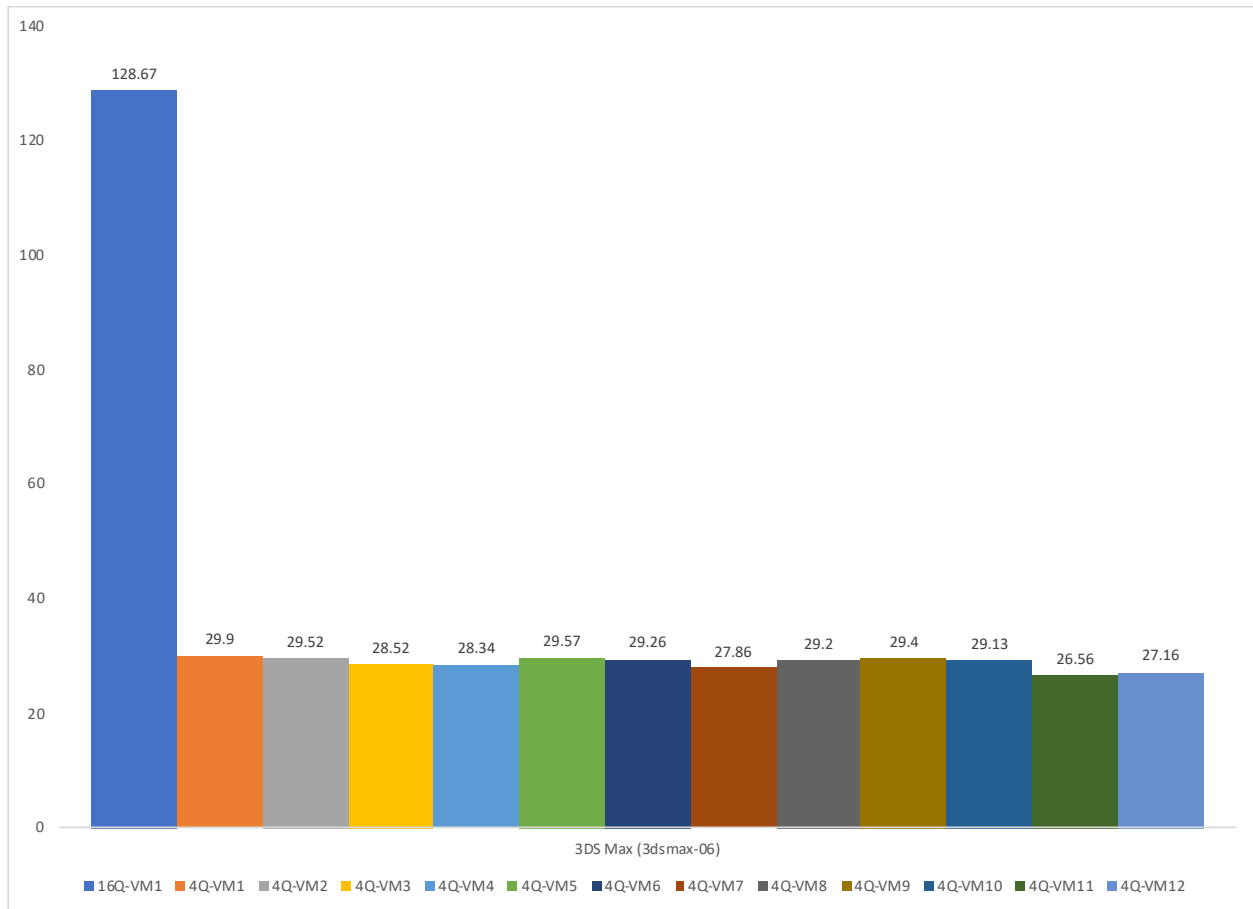
- Architectural model, shaded
- Architectural model, graphite
- Space model, wireframe

- 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 details, see the [3ds Max viewset \(3dsmax-06\) page](#).

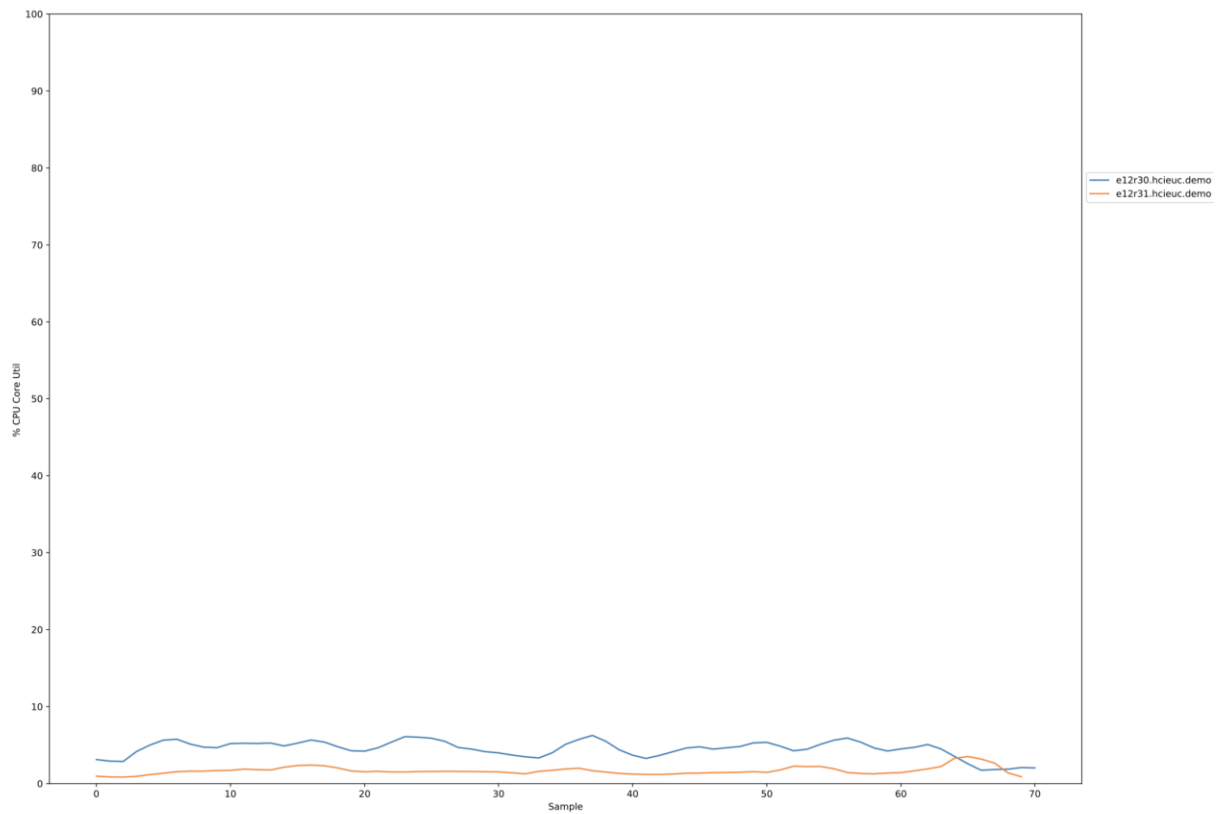
The composite score for the SPECviewperf 13 3DS Max viewset is shown in [Figure 12](#).

Figure 12) 3DS Max composite score.



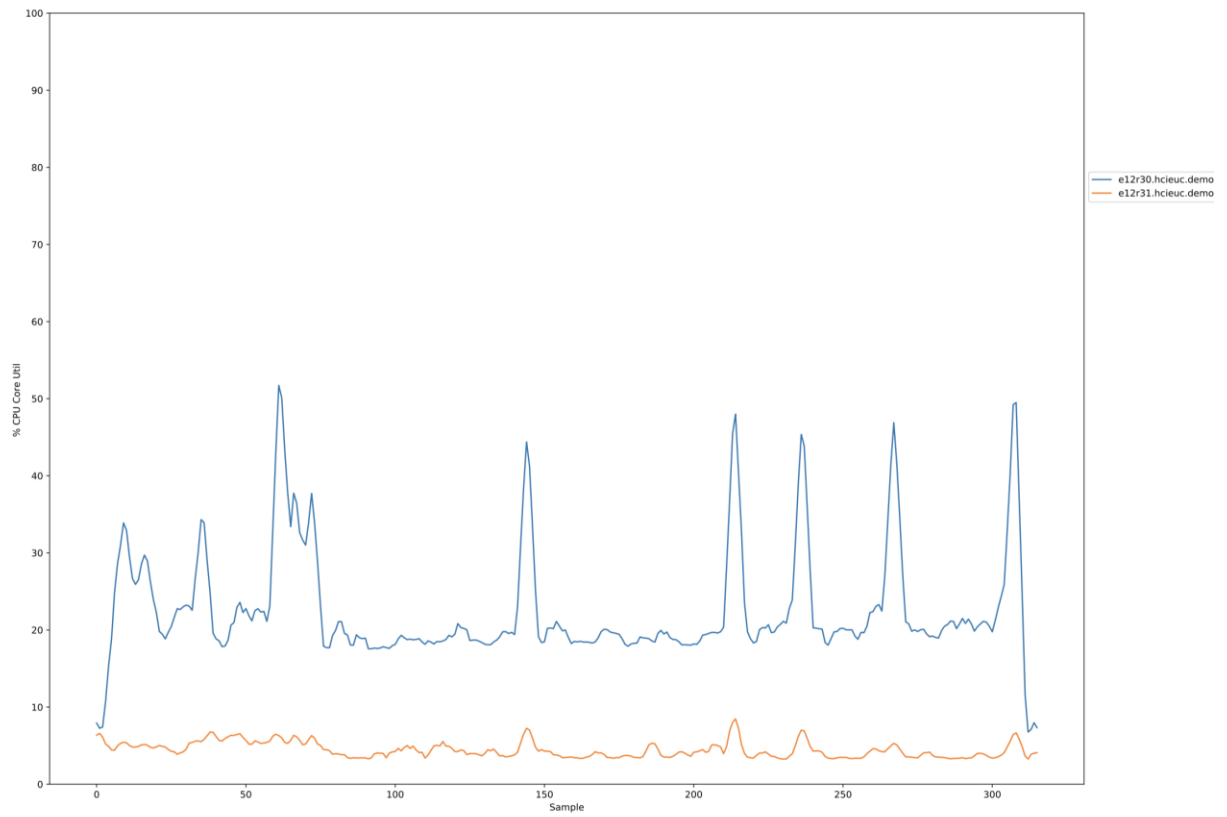
The vSphere host percent core utilization for a single VM is less than 10%, as shown in [Figure 13](#). The viewset ran on an e12r30 host and the client ran on an e12r31 host.

Figure 13) 3DS Max vSphere utilization – 1x16Q.



With 12 VMs, the vSphere percent core utilization stayed below 60% ([Figure 14](#)).

Figure 14) 3DS Max vSphere utilization - 12x4Q.



GPU utilization with a single VM and with 12 VMs is shown in [Figure 15](#).

Figure 15) 3DS Max GPU utilization - 1x16Q.

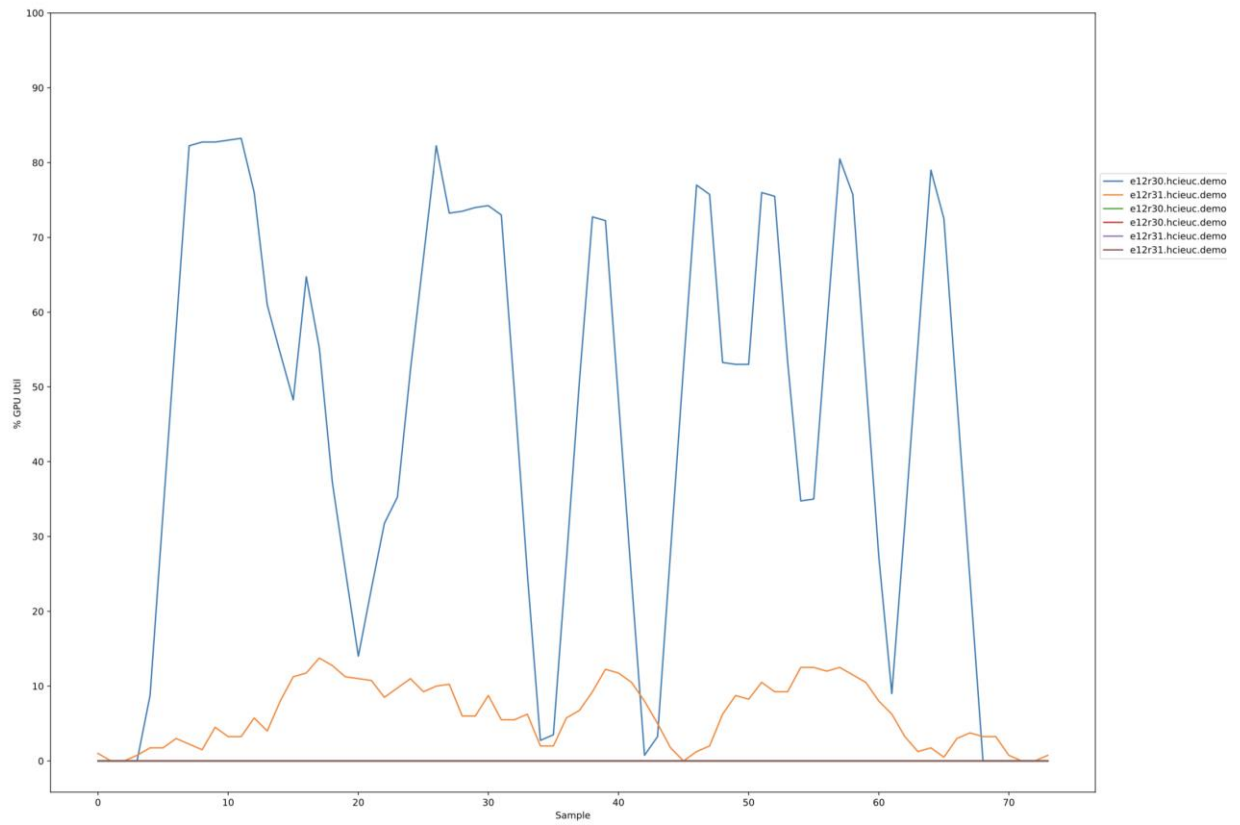
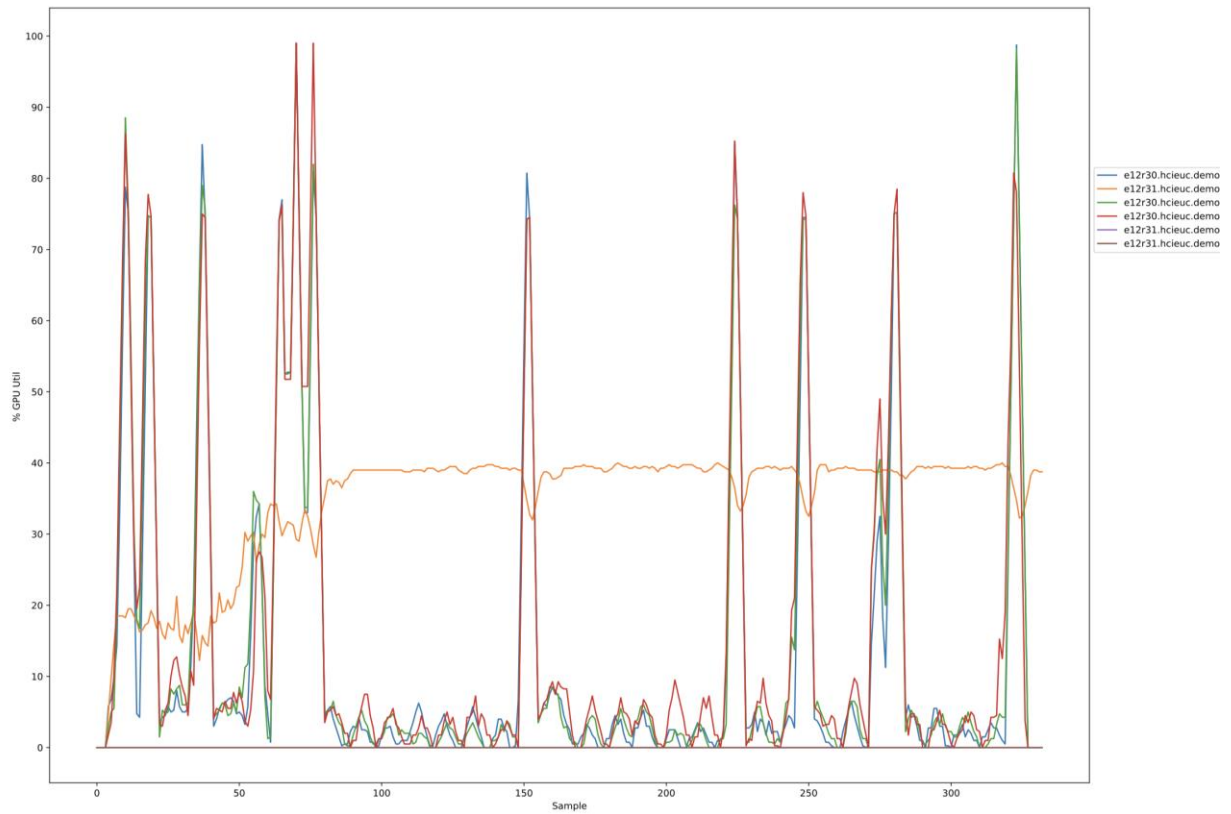


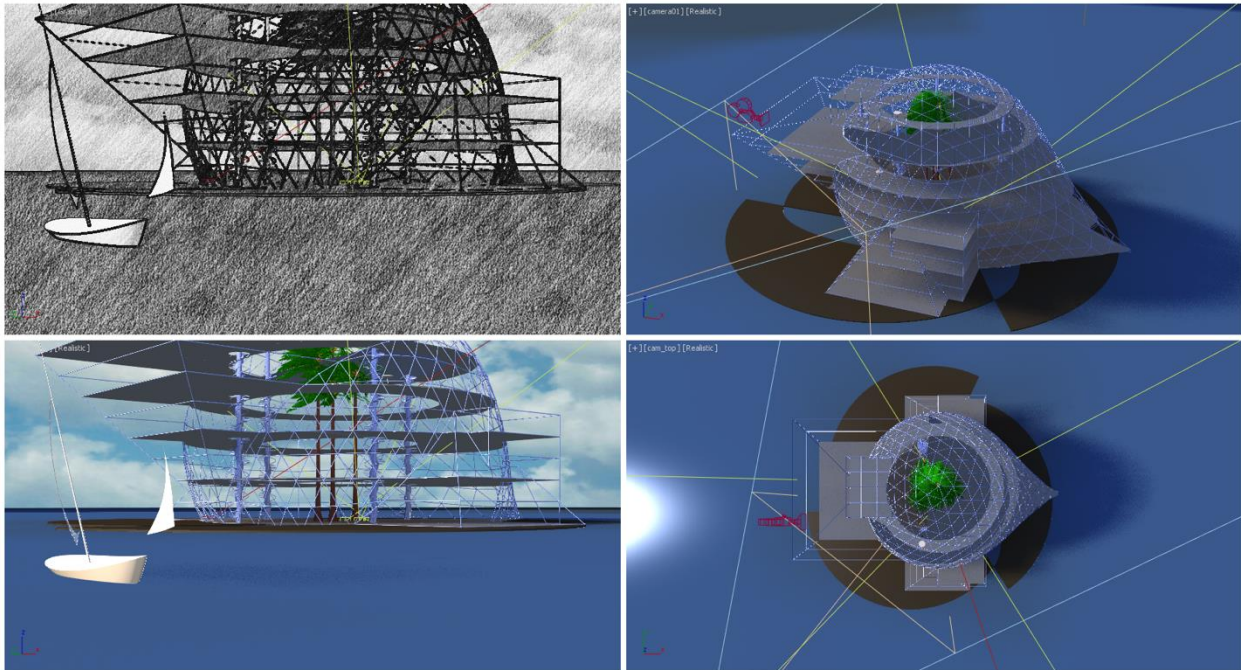
Figure 16) 3DS Max GPU utilization - 12x4Q.



With the depth-first vGPU allocation policy, all 12 client VMs were hosted on same GPU.

[Figure 17](#) provides a sample screenshot captured during the test.

Figure 17) 3DS Max sample.



6.3 Catia (catia-05)

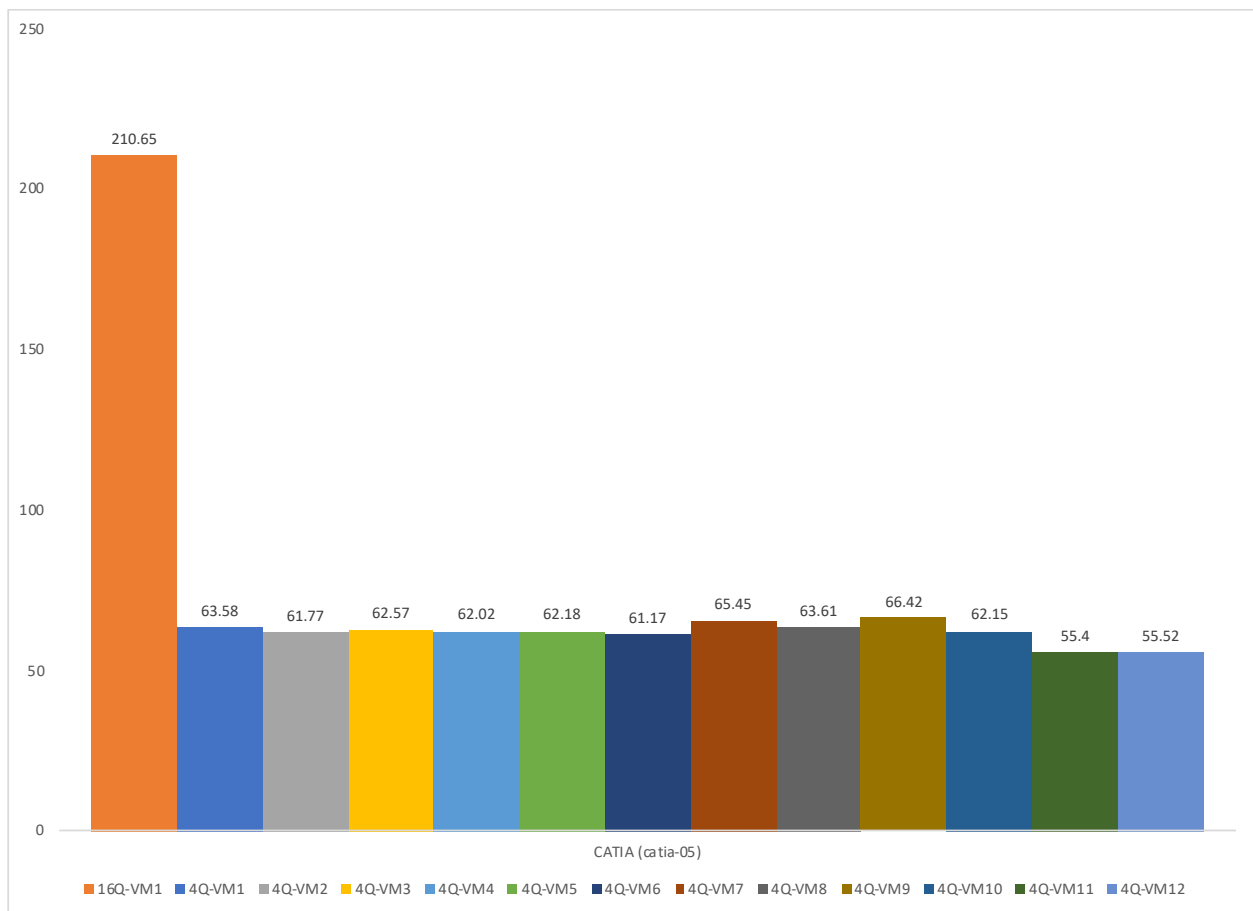
The catia-05 viewset was created from the traces of the graphics workload generated by the CATIA V6 R2012 application from Dassault Systemes. Model sizes ranged from 5.1 to 21 million vertices.

The viewset included numerous rendering modes supported by the application, including wireframe, anti-aliasing, shaded, shaded with edges, depth of field, and ambient occlusion.

The following viewset tests were run:

- 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

Figure 18) Catia composite score.



The percent core utilization of vSphere host stayed below 10% for a single VM ([Figure 19](#)) and was less than 80% for 12 VMs ([Figure 20](#)).

Figure 19) Catia vSphere CPU utilization – 1x16Q.

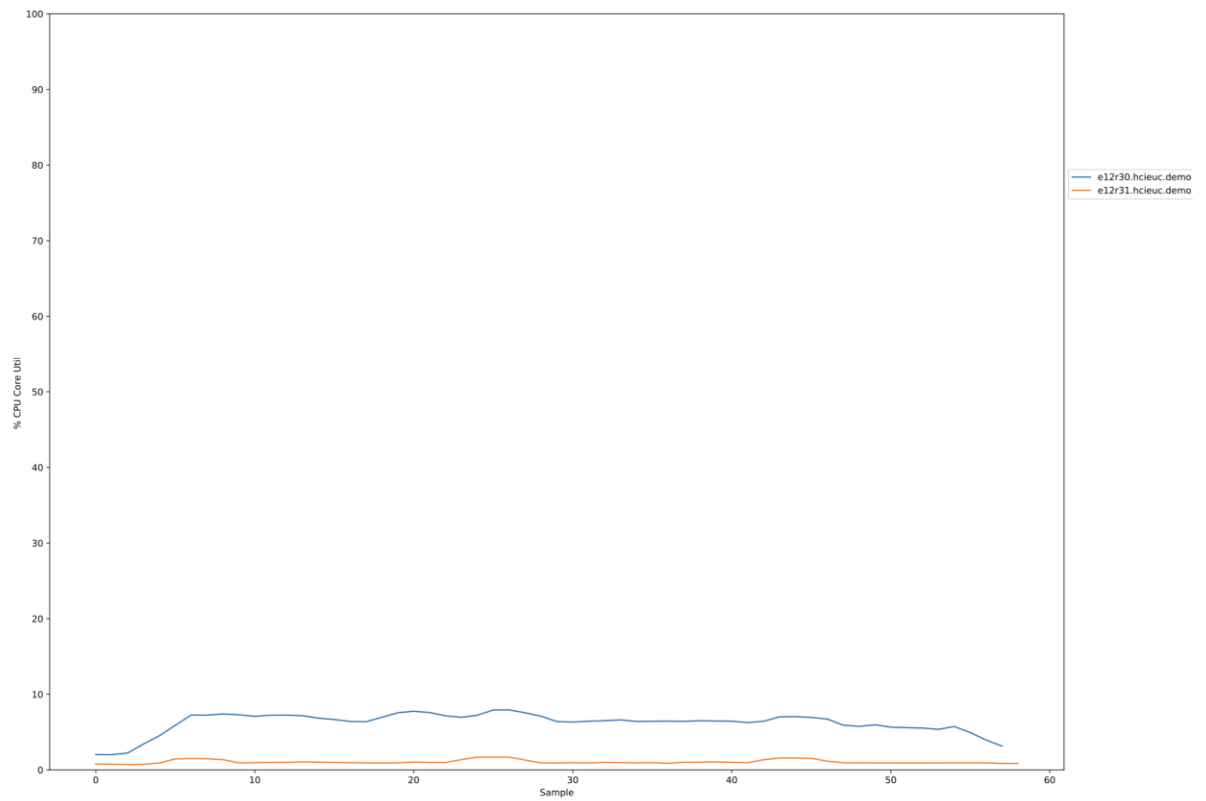
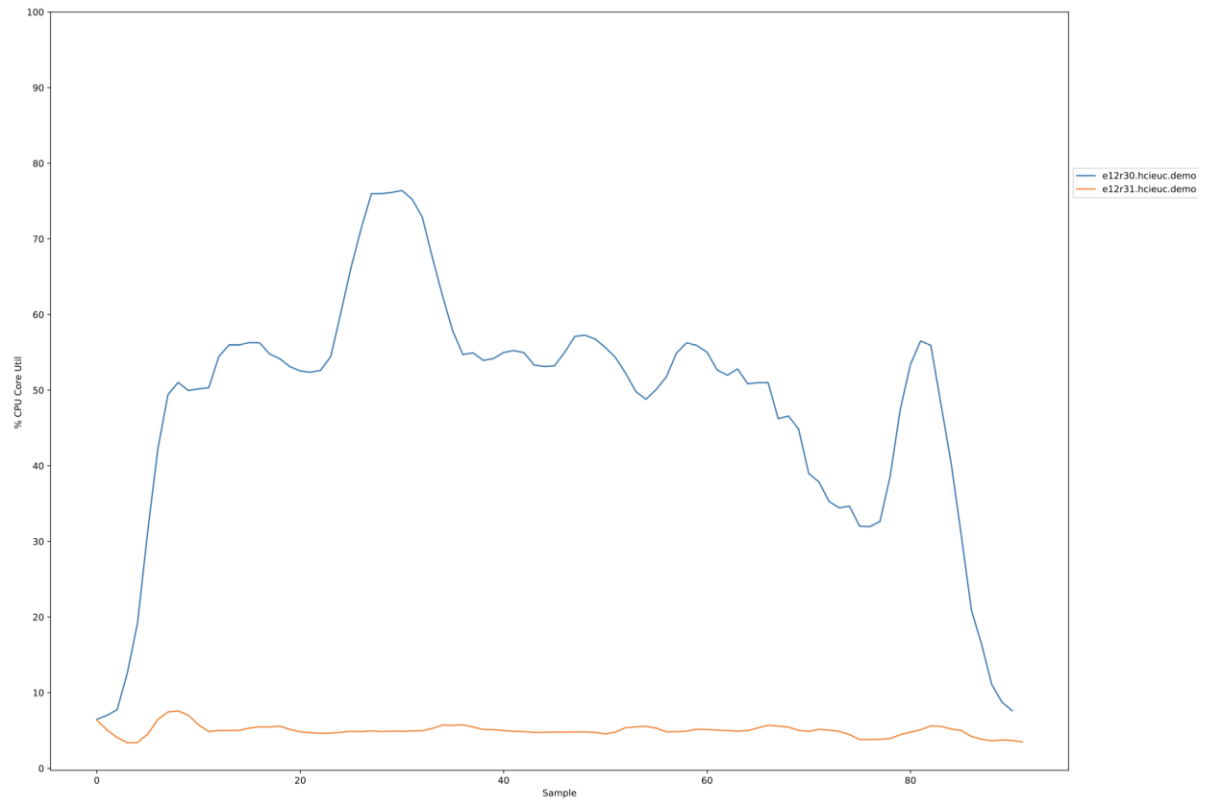


Figure 20) Catia vSphere CPU utilization – 12x4Q.



GPU utilization stayed high for both a single VM ([Figure 21](#)) and for 12 VMs ([Figure 22](#)).

Figure 21) Catia GPU utilization - 1x16Q.

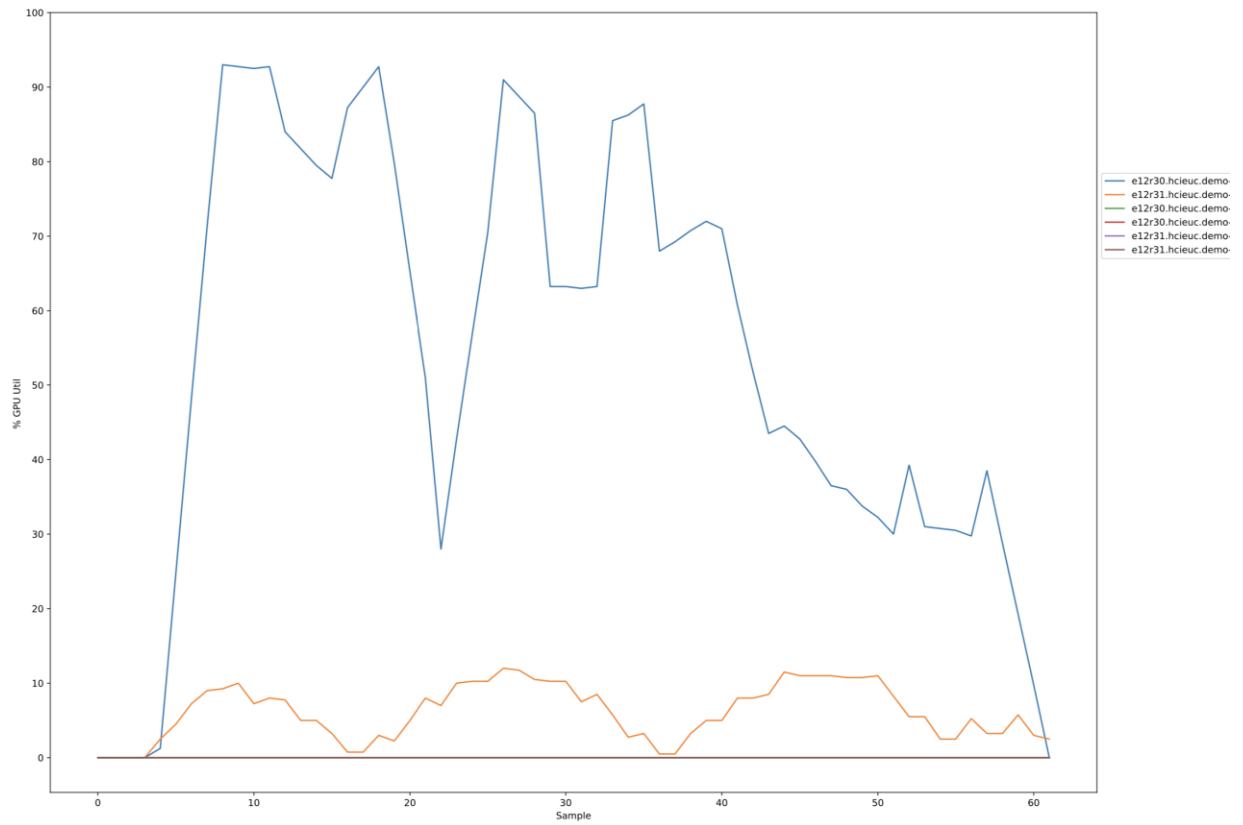


Figure 22) Catia GPU utilization - 12x4Q.

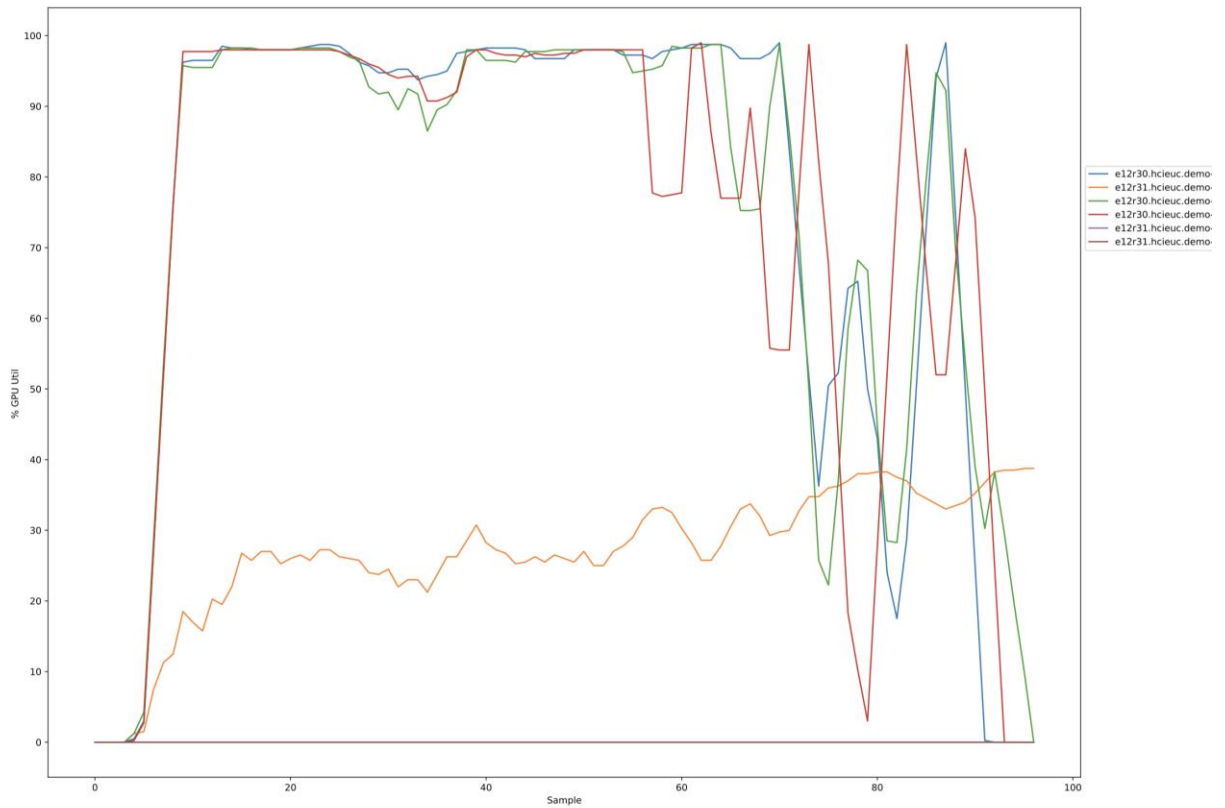
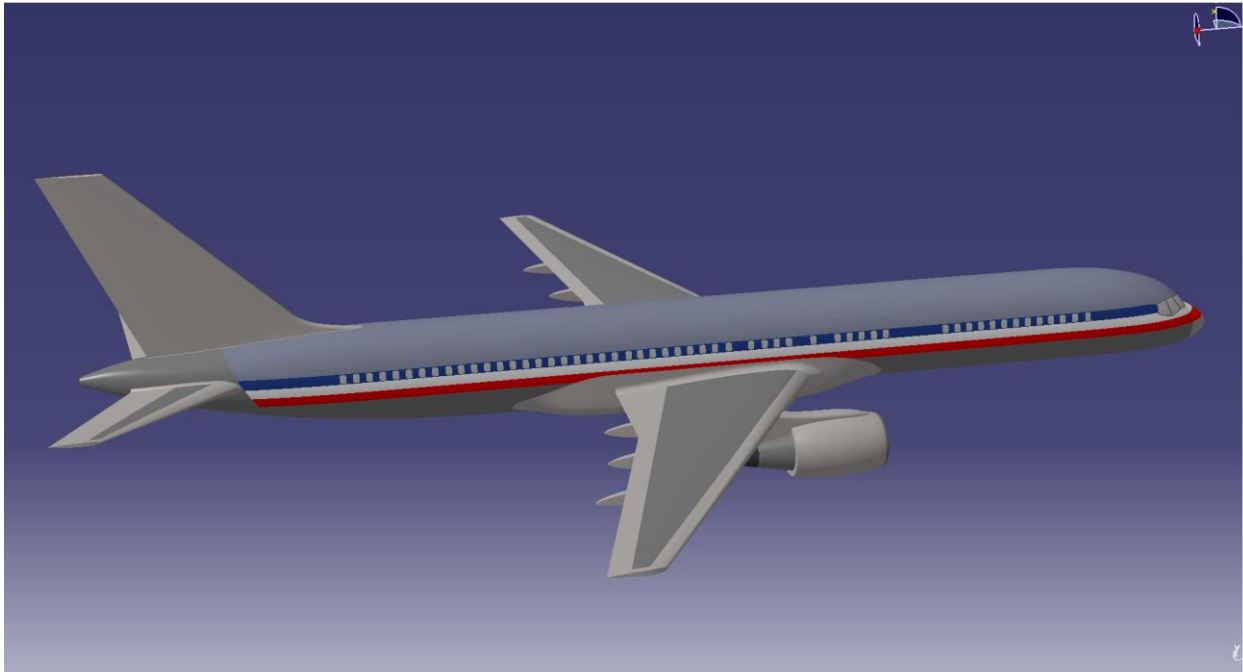


Figure 23 presents a sample screenshot captured during the test.

Figure 23) Catia sample.



6.4 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 ranged from 20 to 48 million vertices.

The viewsets included numerous rendering modes supported by the application.

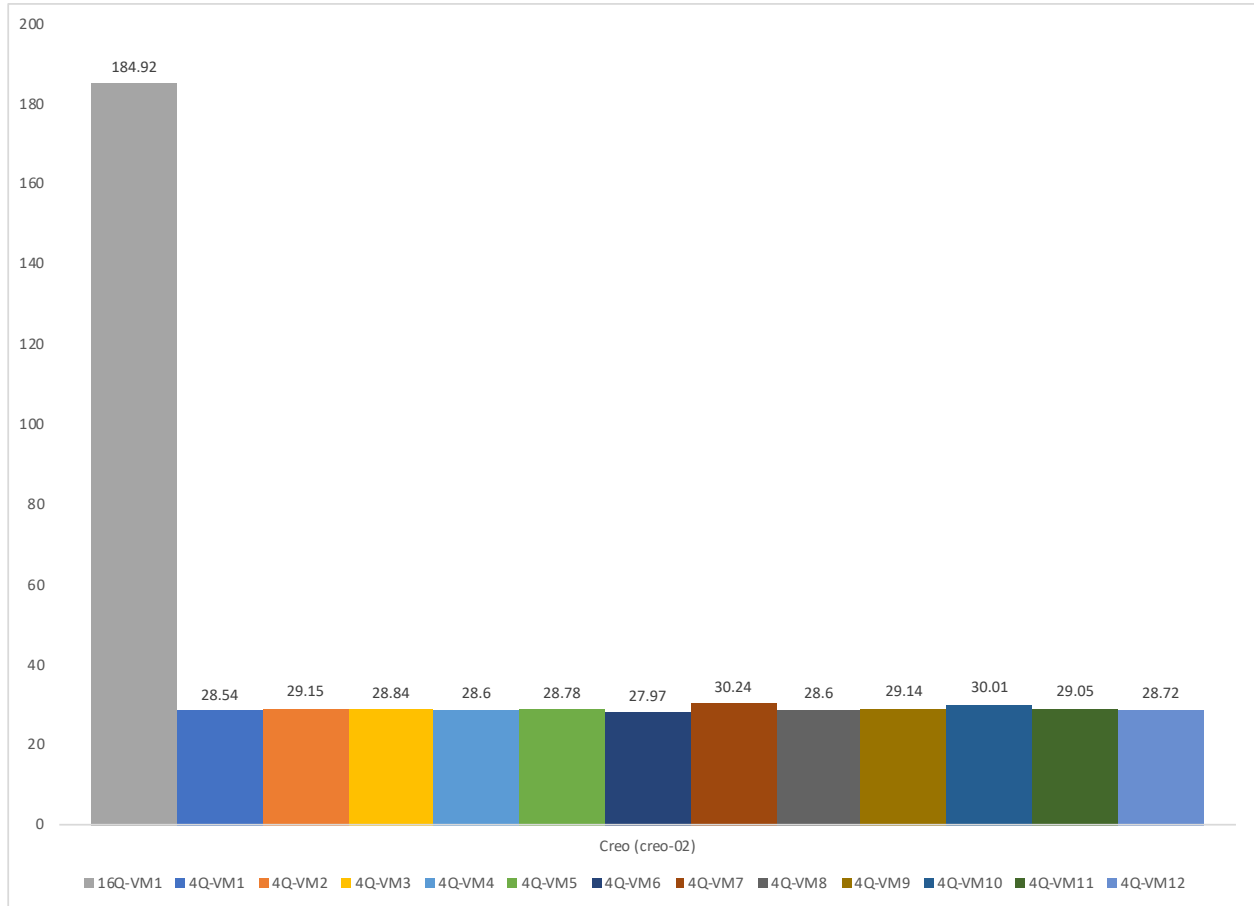
The following viewset tests were run:

- 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 multi-sampled antialiasing.
- Worldcar in shaded mode, with reflections, bump mapping, image background, ambient occlusion, and 8x multi-sampled antialiasing.
- Worldcar in shaded mode
- Engine in shaded mode
- Motorcycle in shaded mode and 4x multi-sampled antialiasing
- Worldcar in shaded-with-edges mode and 4x multi-sampled antialiasing
- Engine in shaded-with-edges mode
- Motorcycle in shaded-with-edges mode
- Four bombers in shaded-with-edges mode and 8x multi-sampled antialiasing (traced from PTC Creo 4)
- Four engines in wireframe mode and 4x multi-sampled antialiasing
- Four bombers in wireframe mode (traced from PTC Creo 4)
- Worldcar in hidden-line mode
- Motorcycle in hidden-line mode and 8x multi-sampled antialiasing

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

The composite scores are showed below for a single VM with the 16Q vGPU profile and 12 VMs with the 4Q vGPU profile.

Figure 24) Creo composite score.



vSphere core utilization stayed below 70% with 12 VMs for the Creo viewset ([Figure 26](#)).

Figure 25) Creo vSphere CPU utilization - 1x16Q.

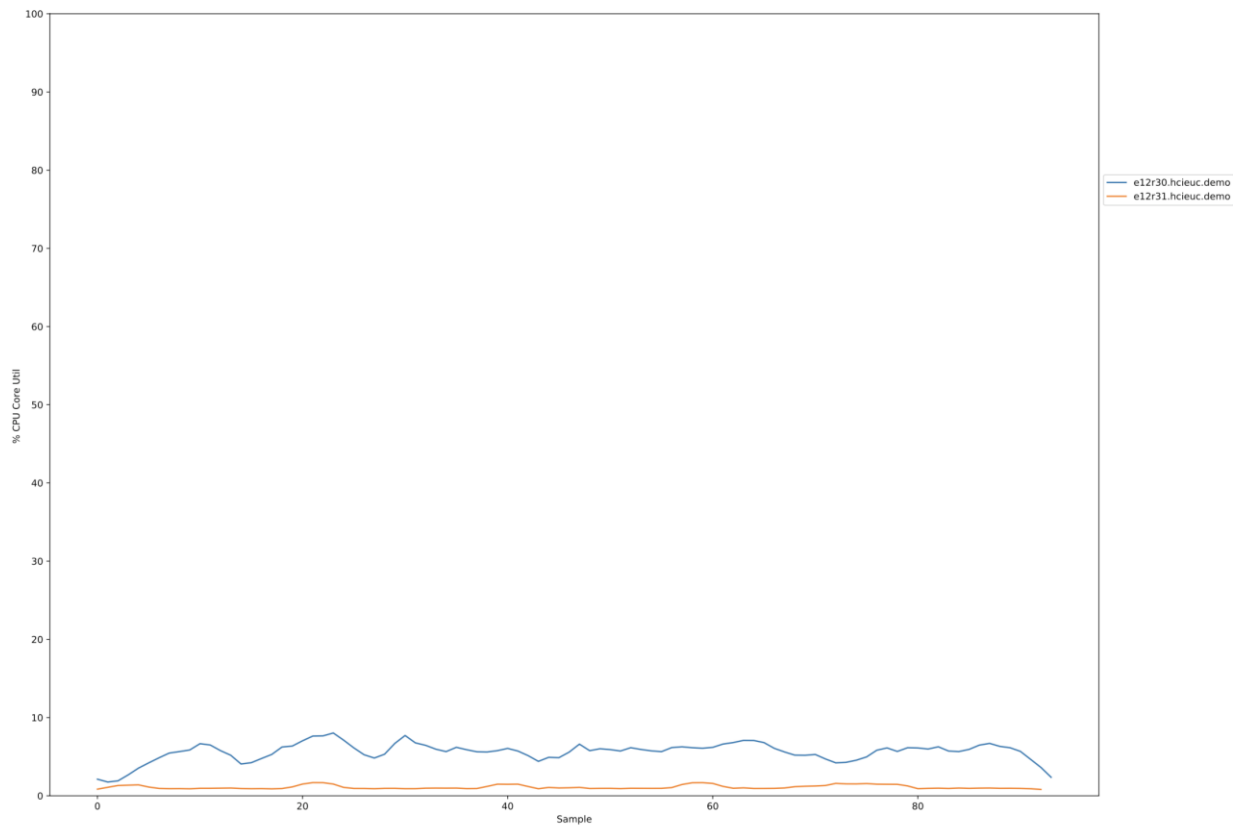
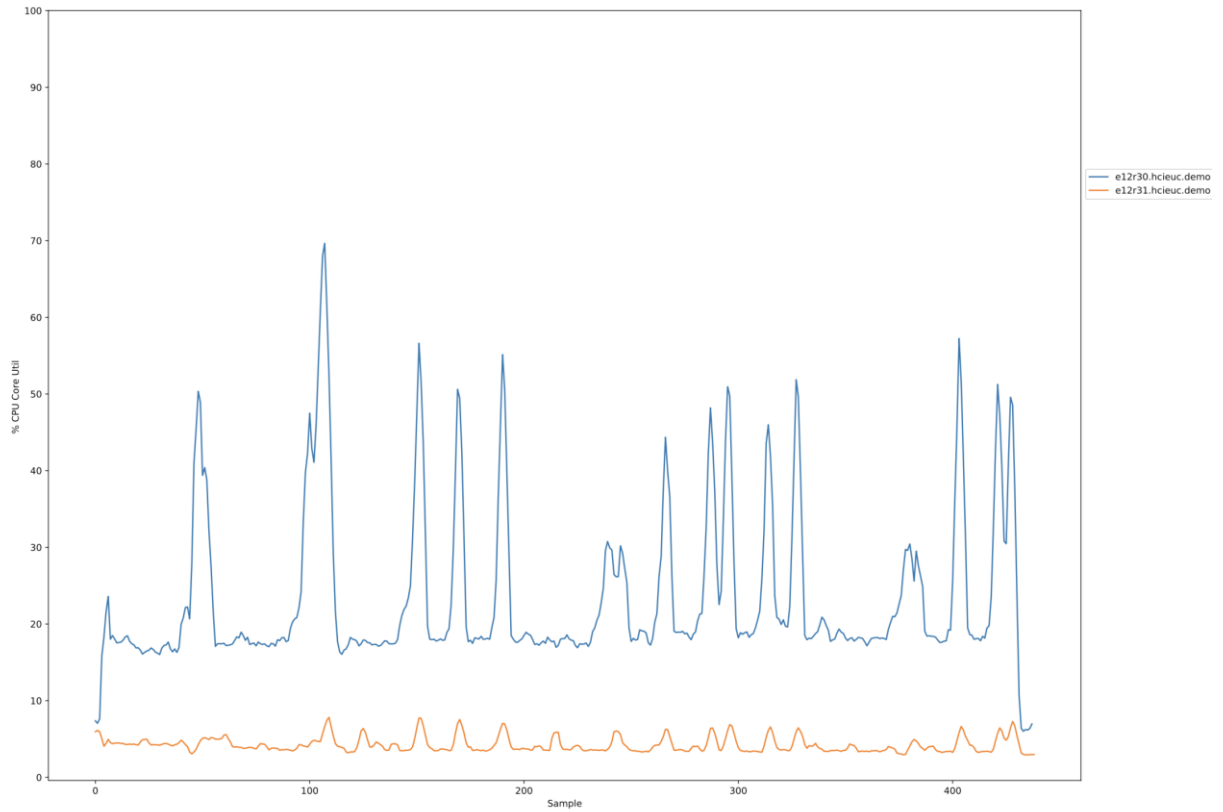


Figure 26) Creo vSphere CPU utilization - 12x4Q.



GPU utilization for the Creo viewset for single VM is shown in [Figure 27](#). Results for 12 VMs is shown in [Figure 28](#).

Figure 27) Creo GPU utilization - 1x16Q.

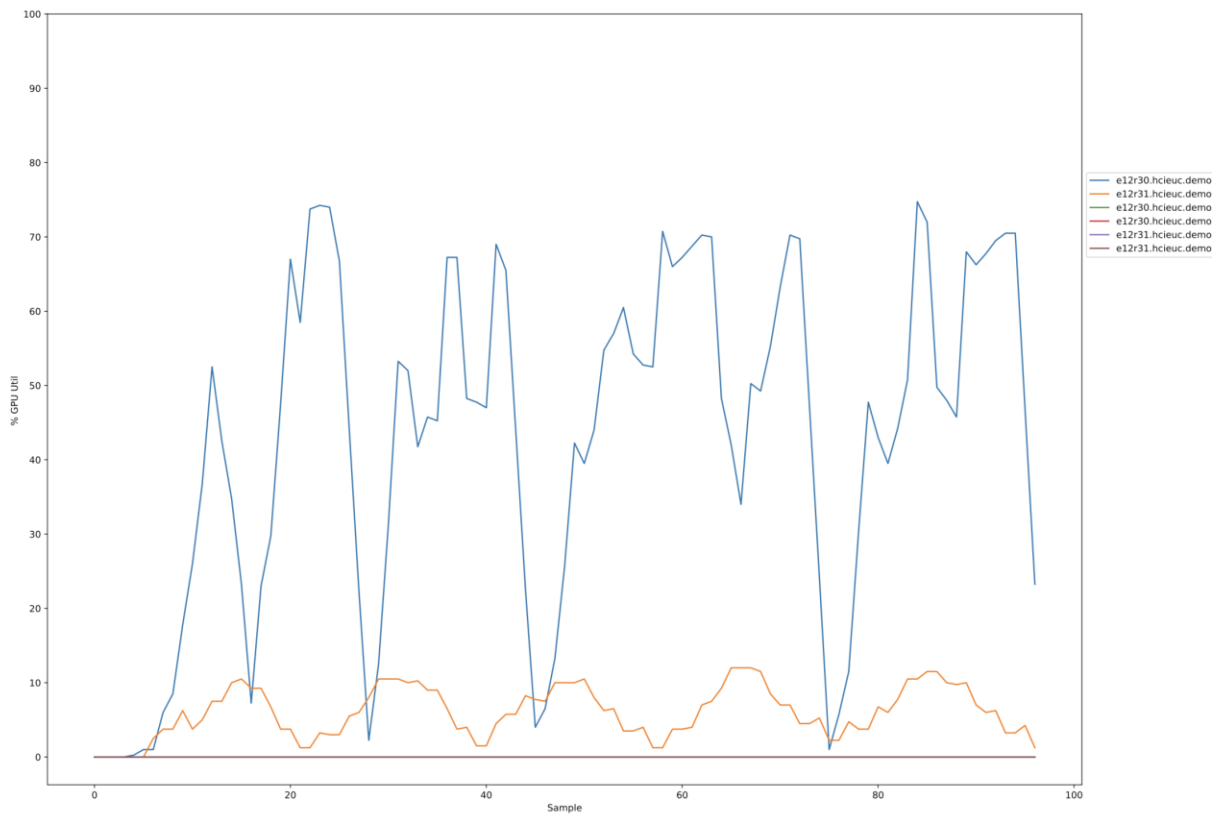


Figure 28) Creo GPU utilization - 12x4Q.

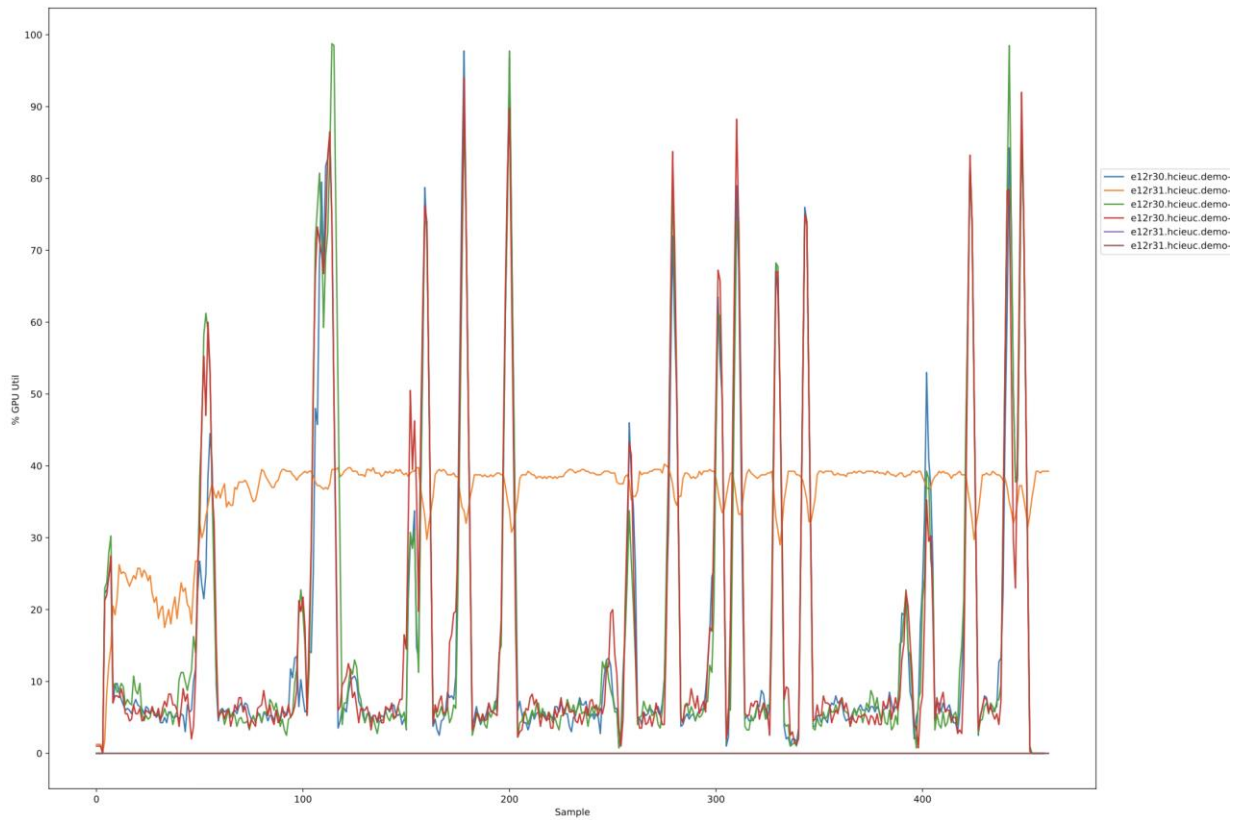
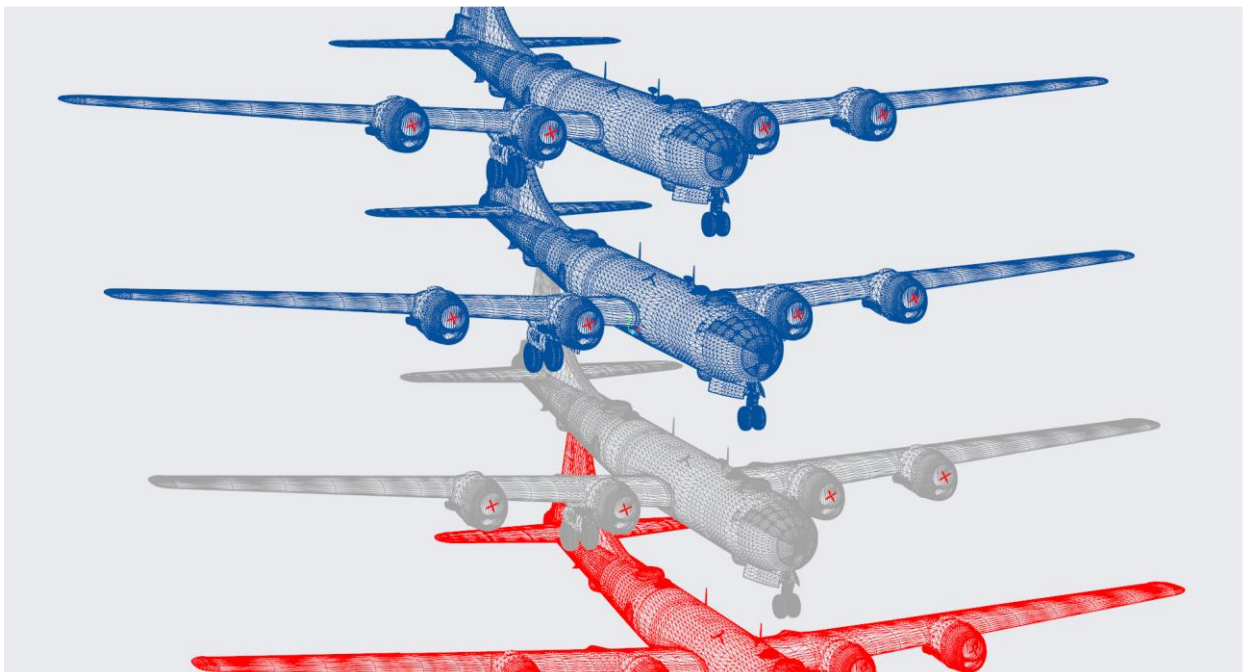


Figure 29 presents a sample screenshot from test run.

Figure 29) Creo Sample.



6.5 Energy (energy-02)

The energy-02 viewset is based on rendering techniques used by the open-source OpendTect seismic visualization application. 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. Transparent lit, color-mapped values are accumulated 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 (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 thus 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 generated by exploration geophysicists and rendered using textured triangle strips.

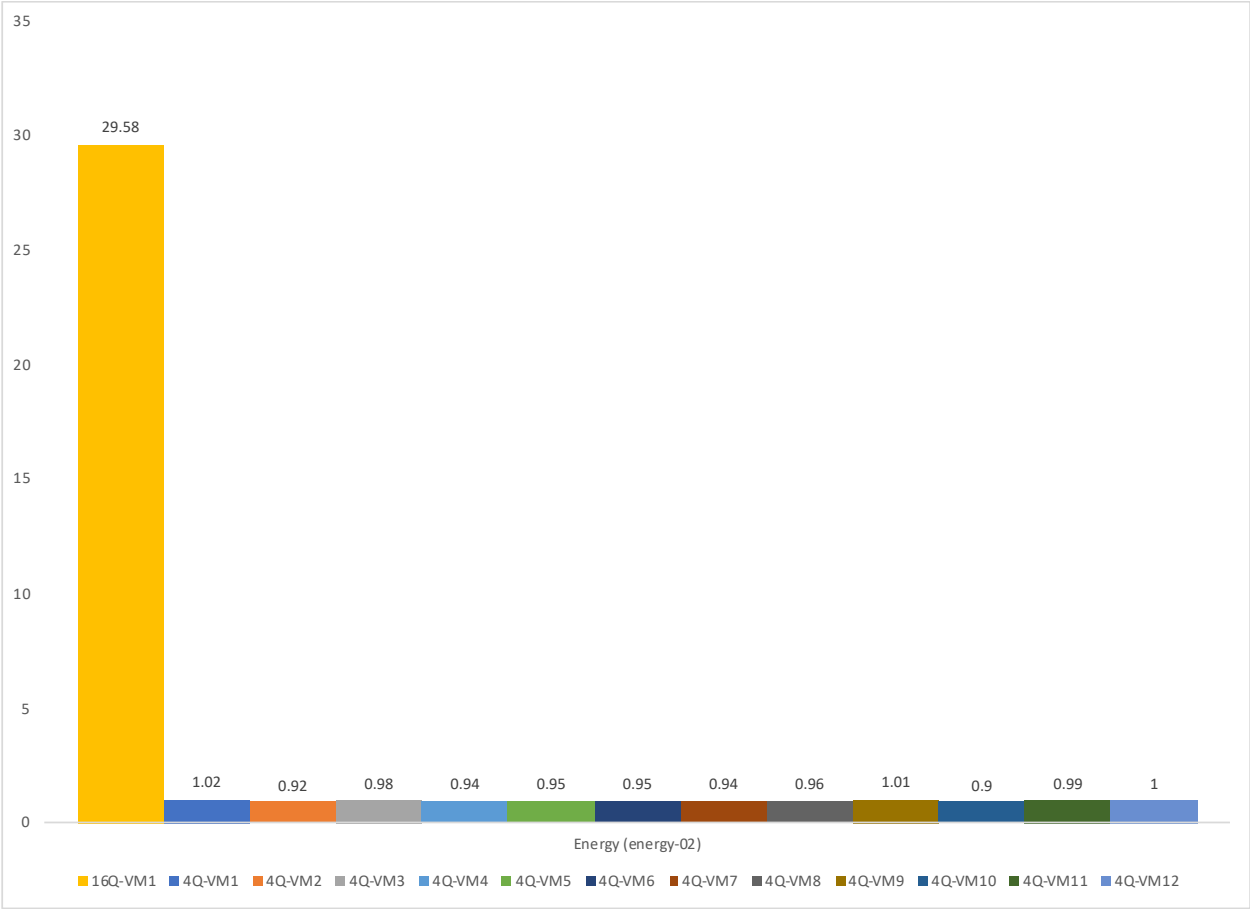
The 3D datasets used in this viewset are real-world seismic datasets found at the [SEG Open Data Wiki page](#). They were translated from their native SEG-Y format and compressed using JPEG-2000.

The following viewset tests were run:

- 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)

The composite score for single 16Q VM and 12 4Q VMs is provided in [Figure 30](#). The frame buffer size affected the composite score significantly for the Energy viewset.

Figure 30) Energy composite score.



vSphere CPU utilization stayed below 80% for 12 VMs with the Energy viewset ([Figure 32](#)).

Figure 31) Energy vSphere CPU utilization - 1x16Q.

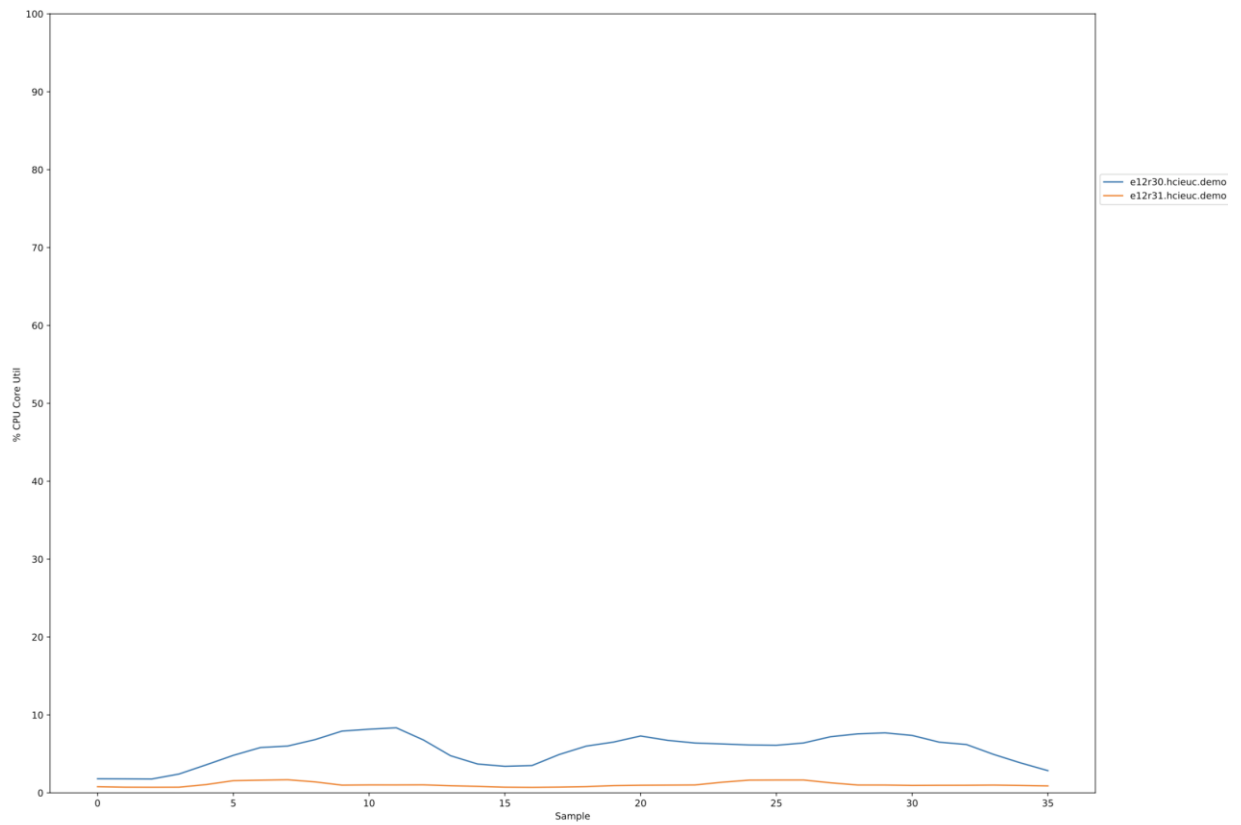


Figure 32) Energy vSphere CPU utilization - 12x4Q.

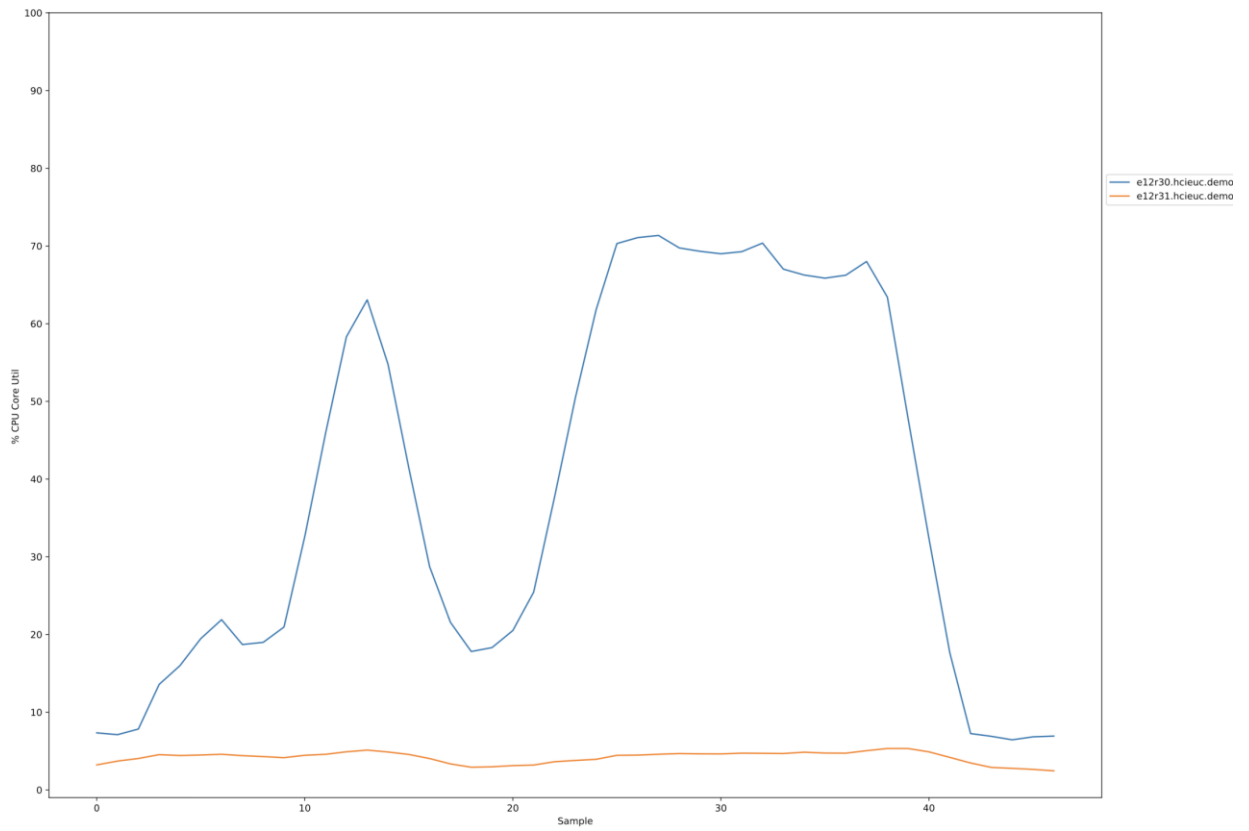


Figure 33) Energy GPU utilization - 1x16Q.

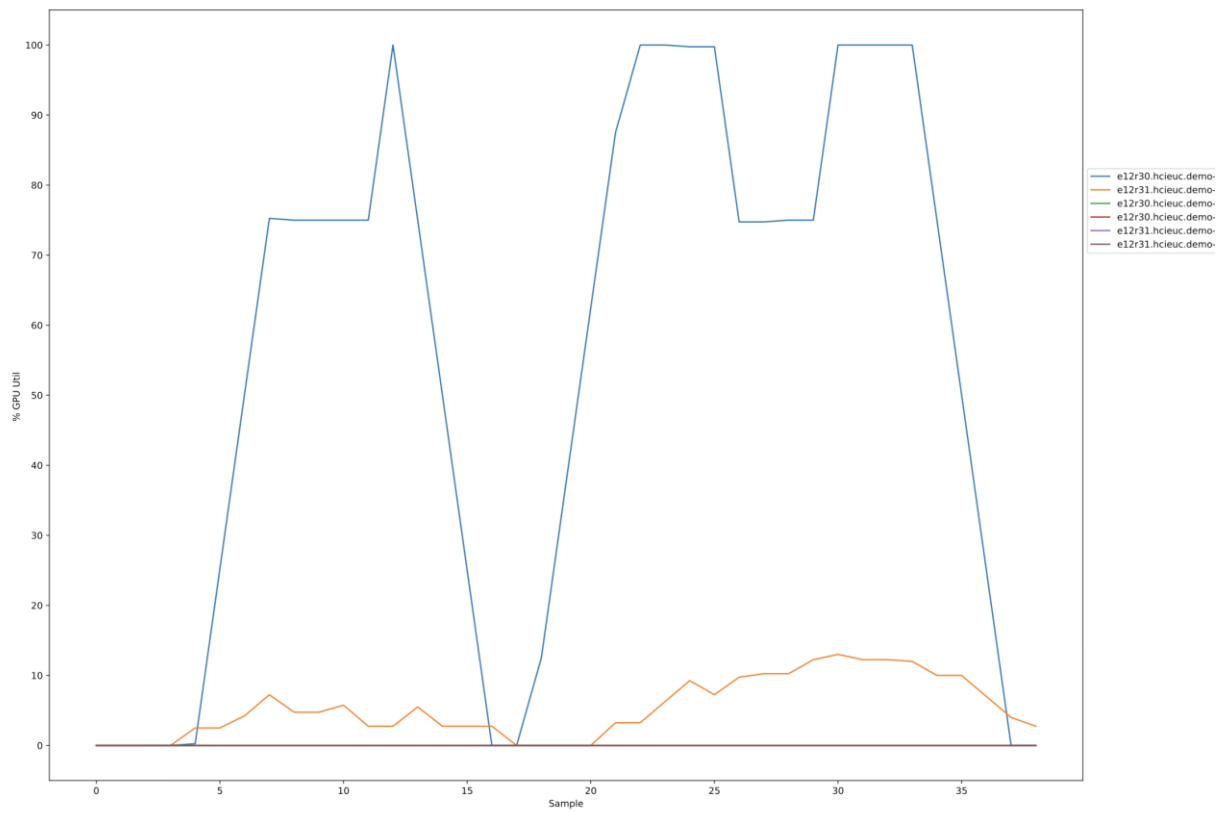
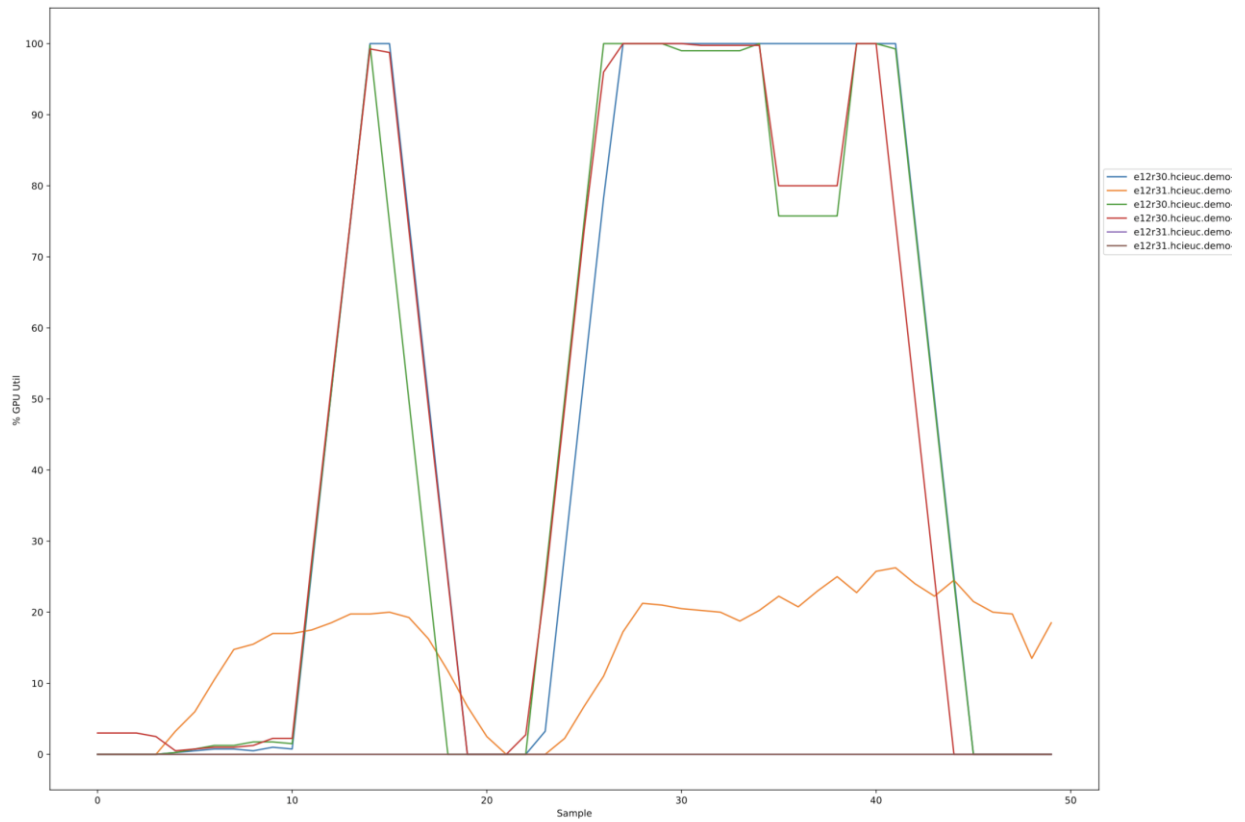
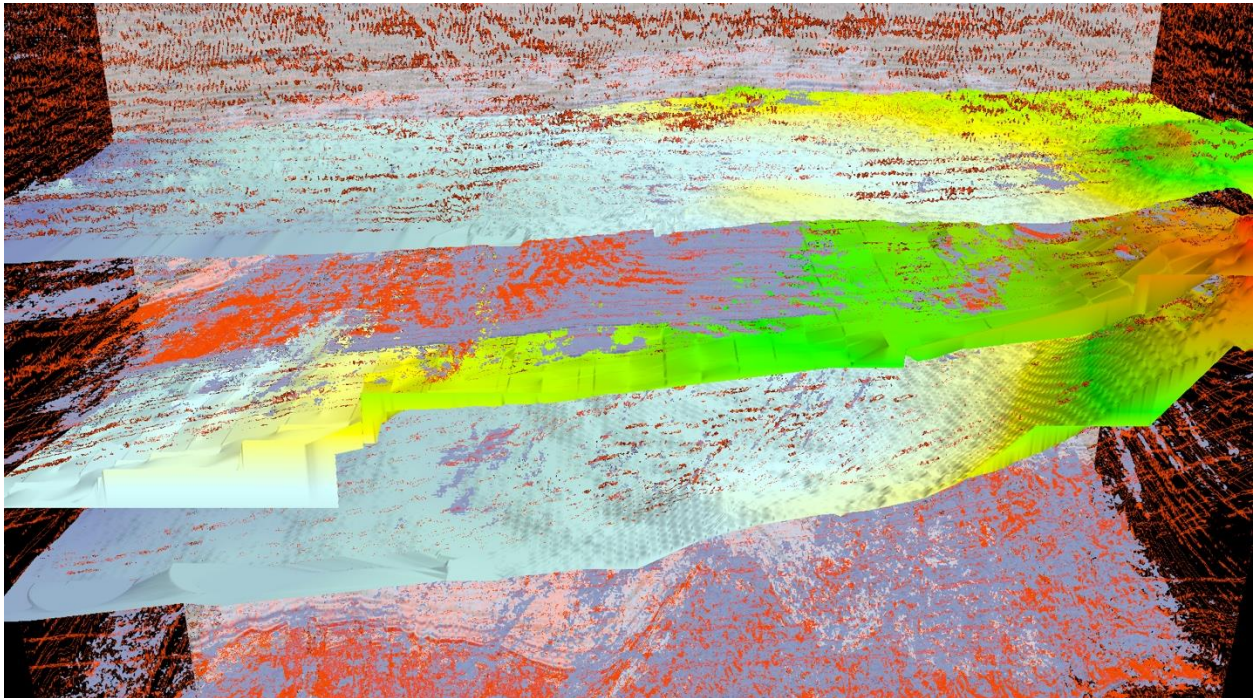


Figure 34) Energy GPU utilization - 12x4Q.



[Figure 35](#) presents a sample screenshot from test run.

Figure 35) Energy Sample.



6.6 Maya (maya-05)

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

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

The following viewset tests were run:

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

The composite scores from the test with a single 16Q VM and 12 4Q VMs are provided in [Figure 36](#).

Figure 36) Maya composite score.

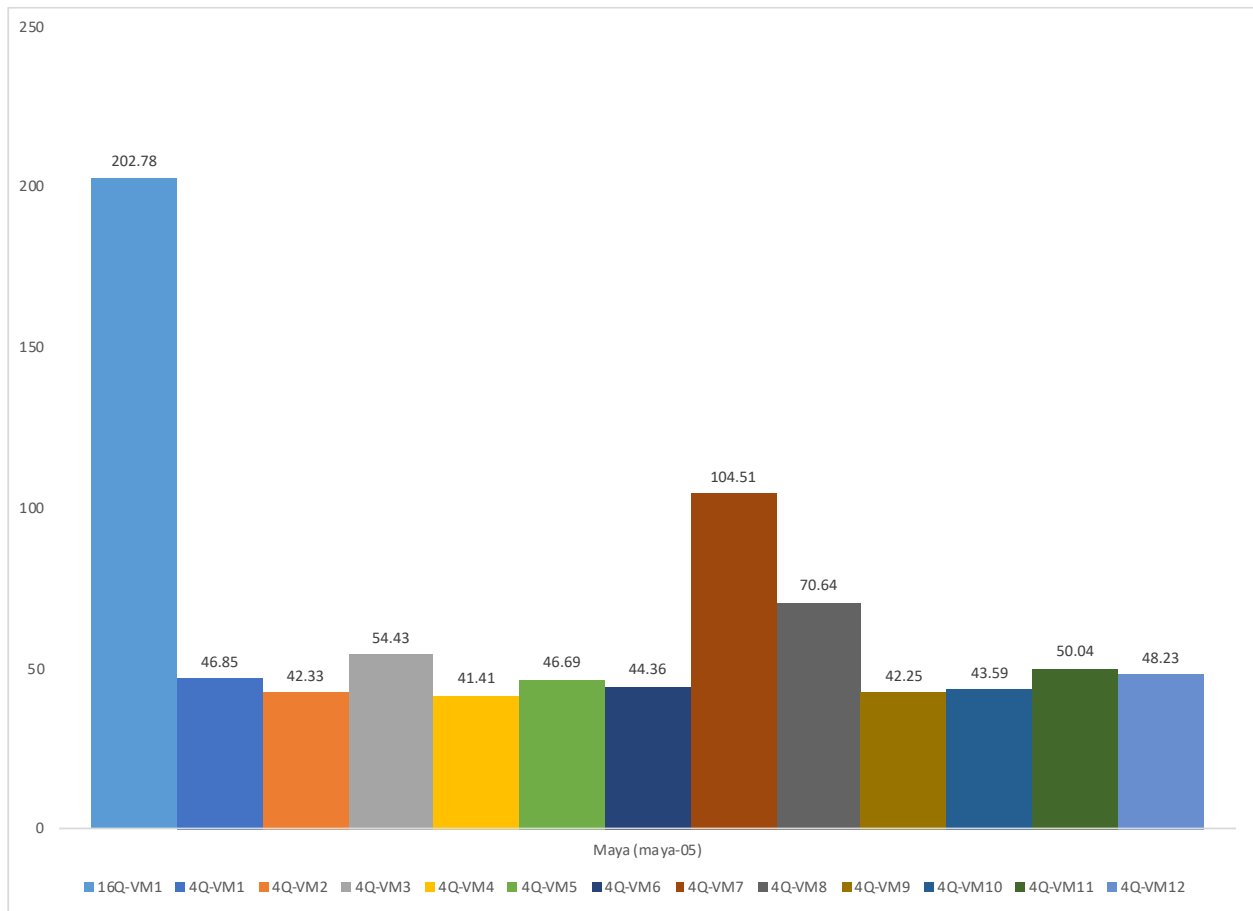


Figure 37) Maya vSphere CPU utilization - 1x16Q.

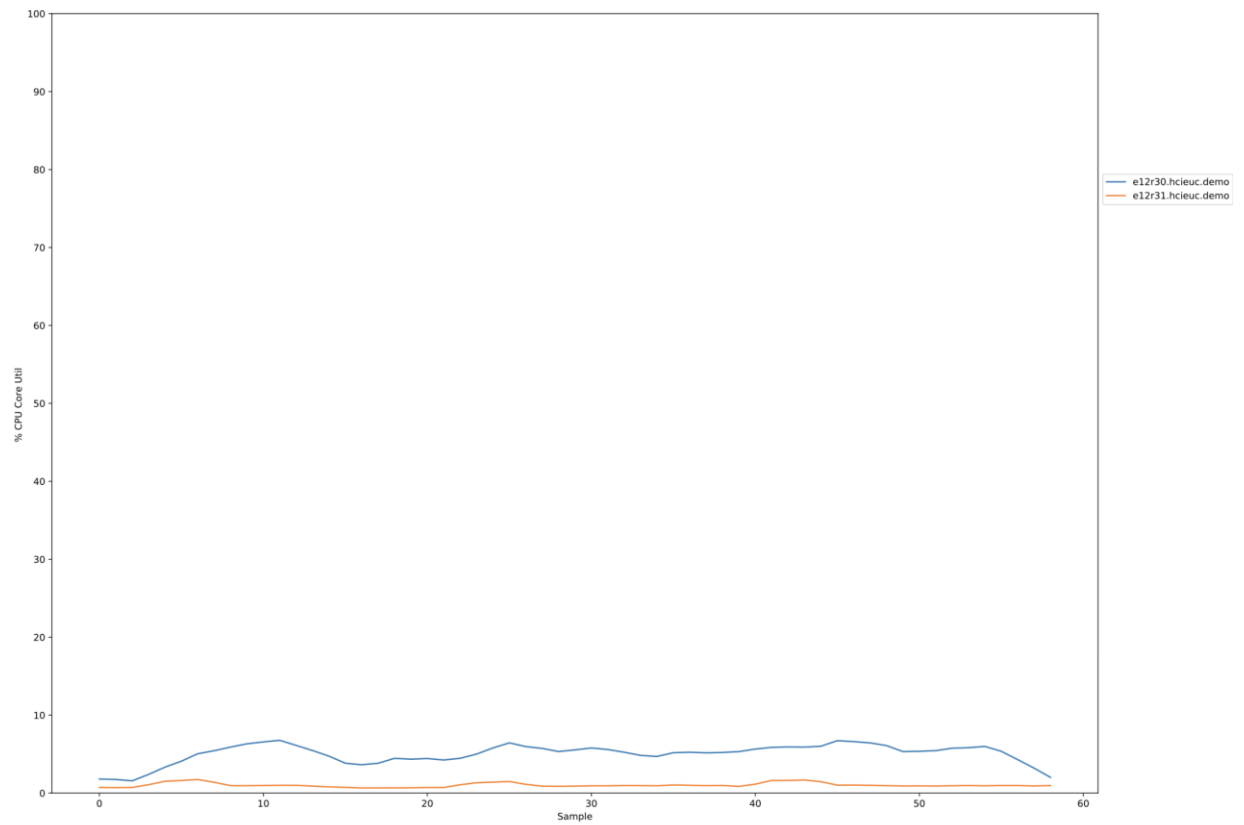


Figure 38) Maya vSphere CPU utilization - 12x4Q.

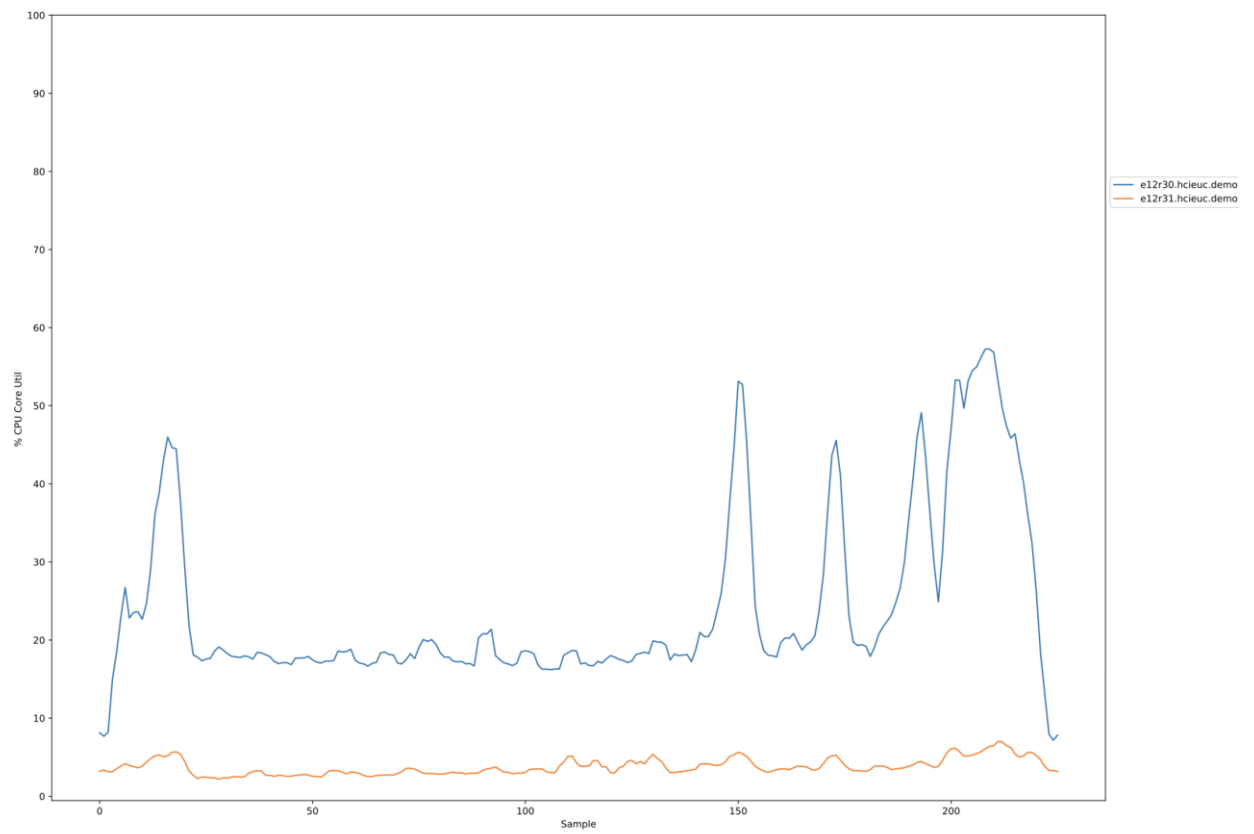


Figure 39) Maya GPU utilization - 1x16Q.

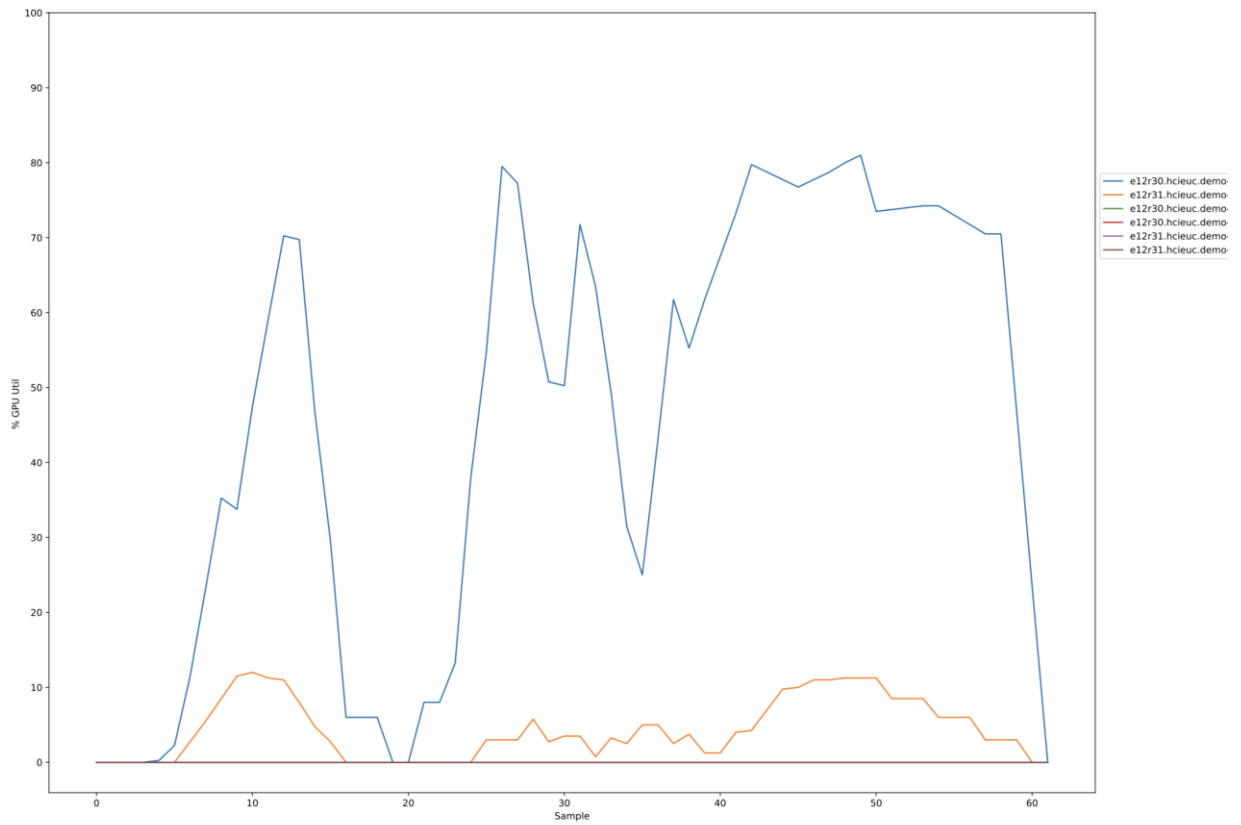


Figure 40) Maya GPU utilization - 12x4Q.

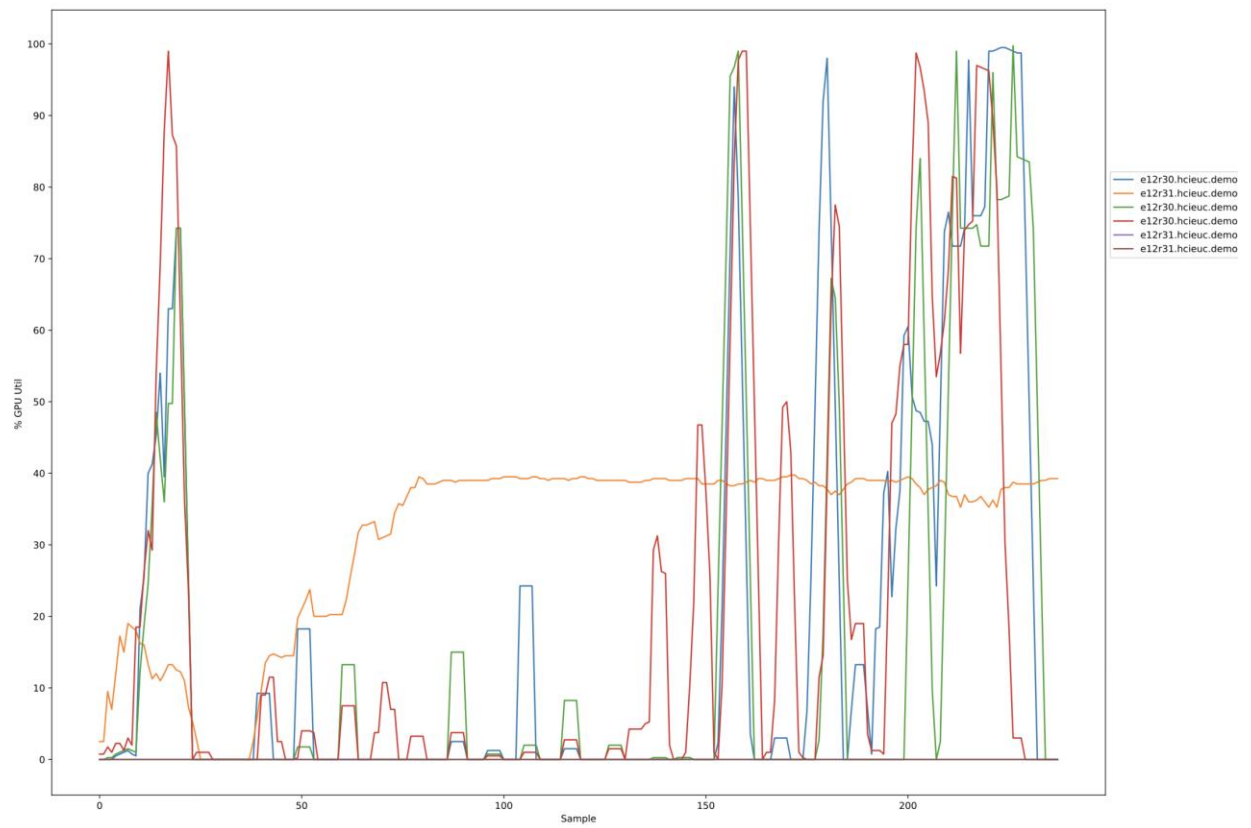


Figure 41) Maya Sample.



6.7 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 CT or MRI).

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 and then sent to the graphics hardware for texturing and further calculations. Calculations include 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 is known as bricking and 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.

There are four datasets in this viewset:

- A 4D heart dataset comprising 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 are 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. Tuvok includes a Hilbert Curve implementation, which is copyrighted (1998, Rice University). Tuvok also includes LZ4, which is licensed under the BSD 2-Clause license.

Figure 42) Medical composite score.

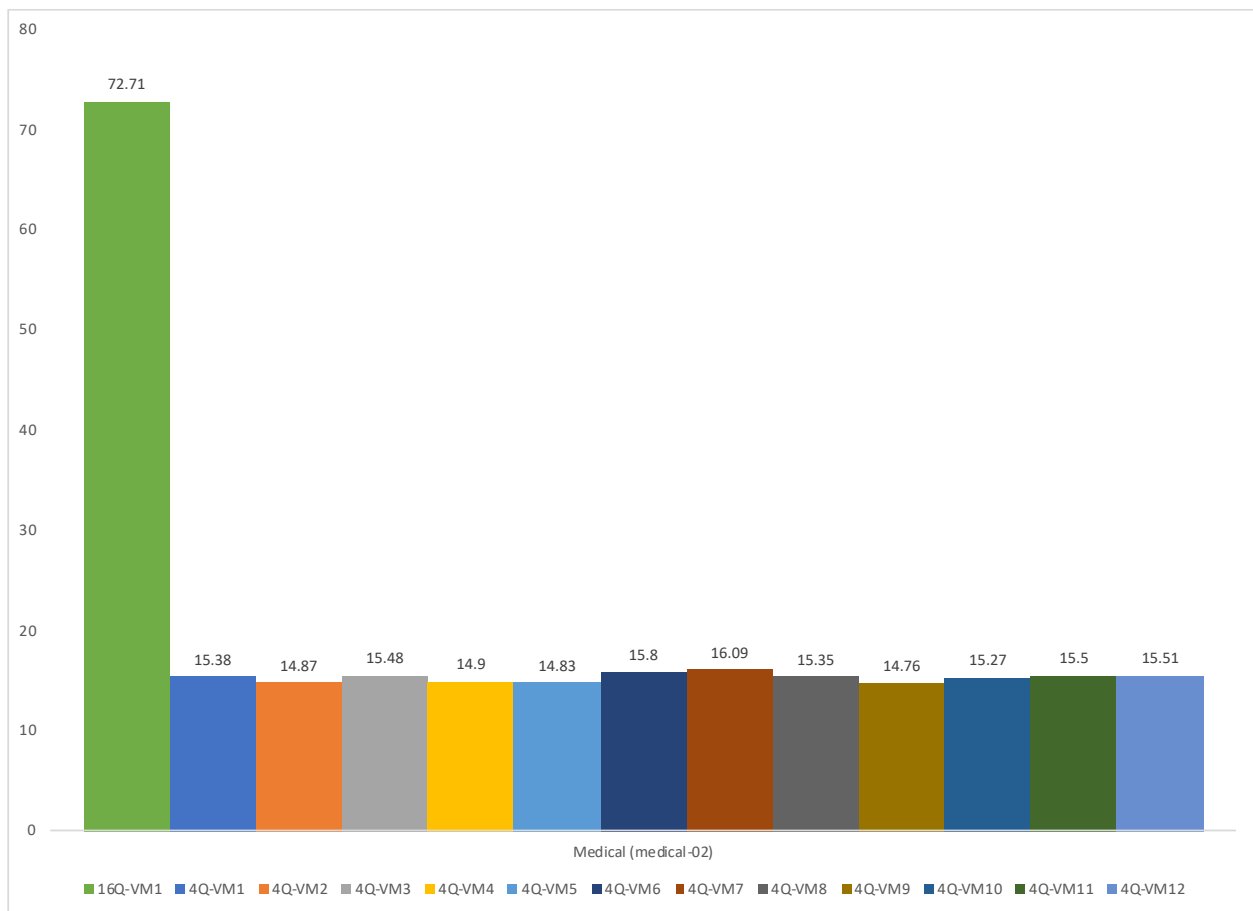


Figure 43) Medical vSphere CPU utilization - 1x16Q.

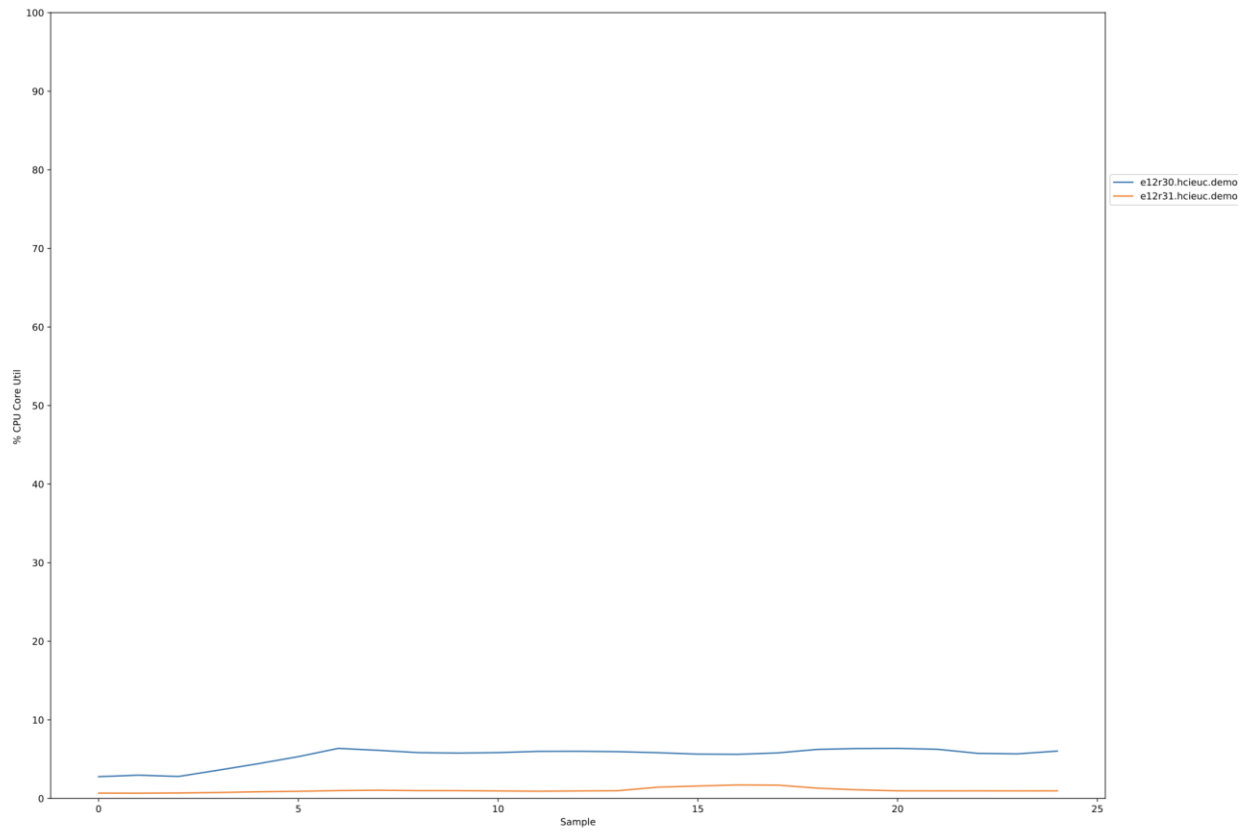


Figure 44) Medical vSphere CPU utilization - 12x4Q.

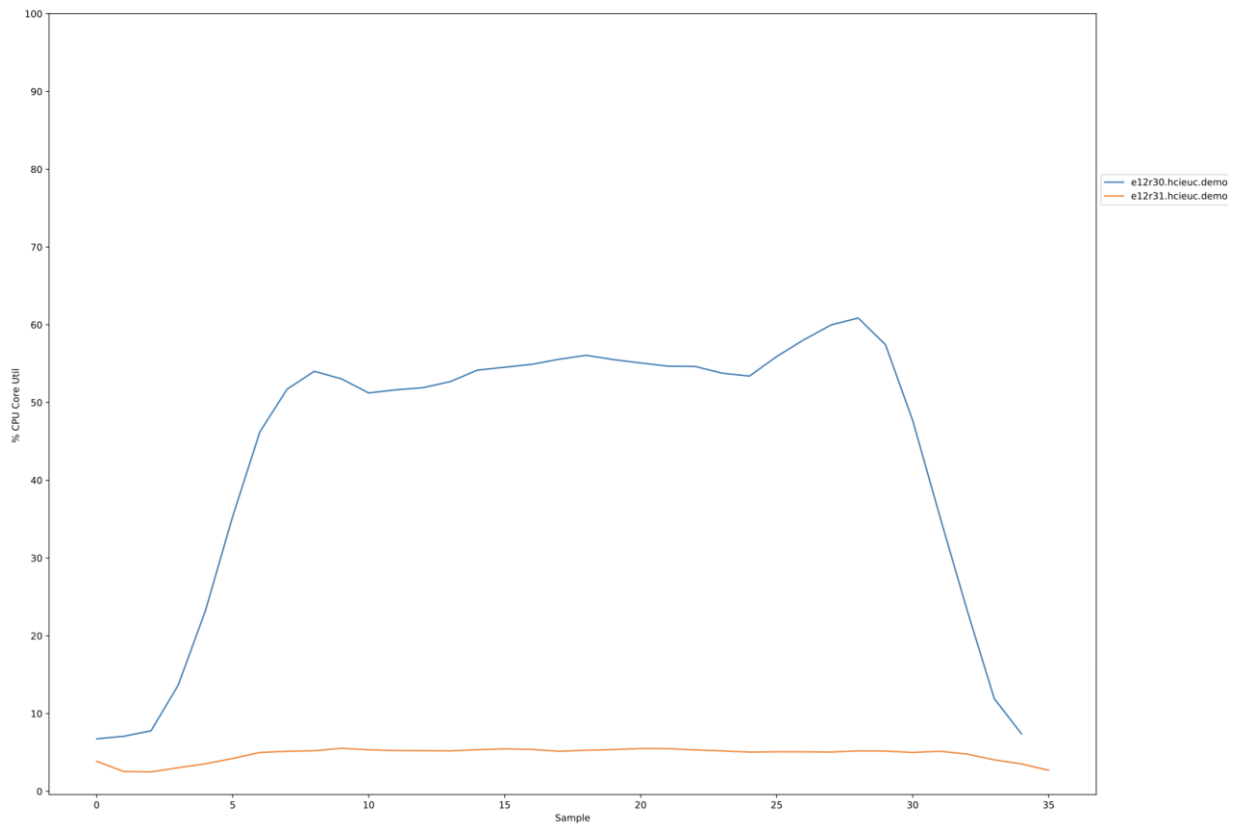


Figure 45) Medical GPU utilization - 1x16Q.

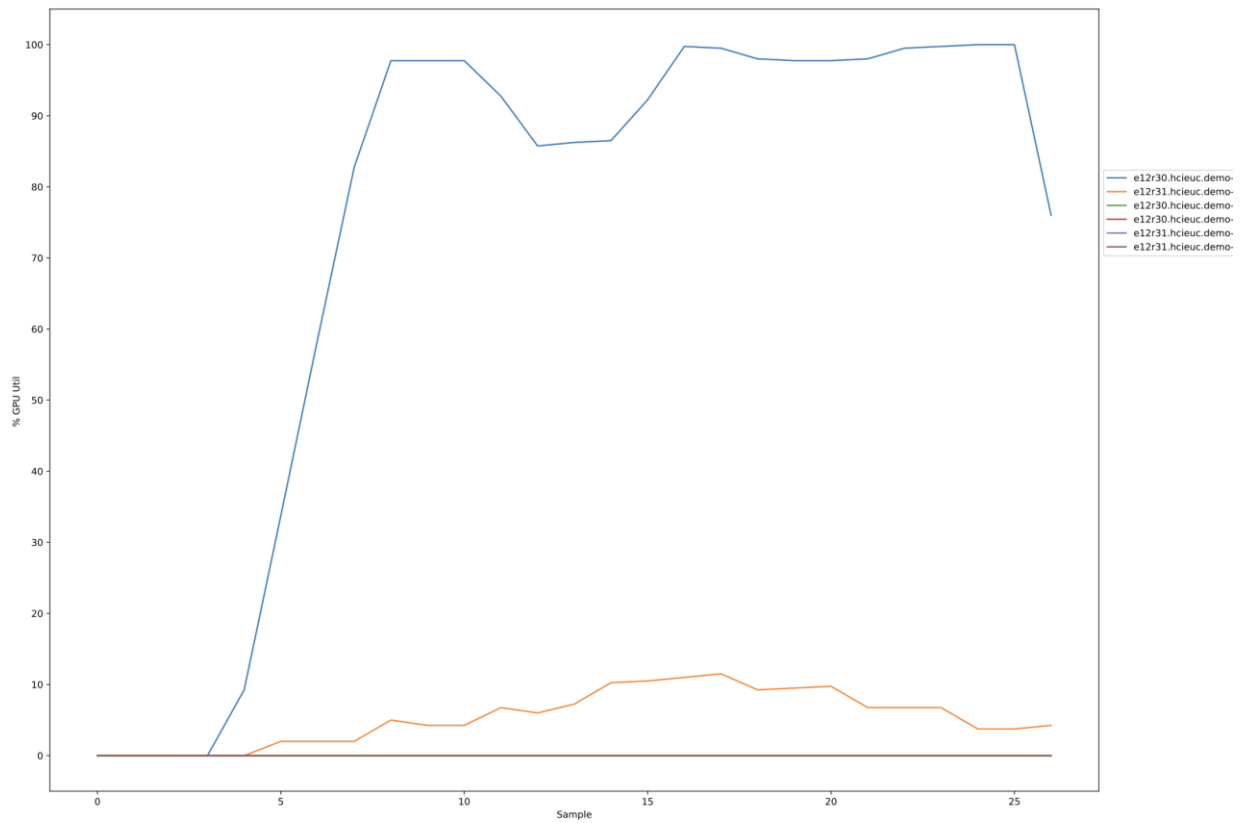


Figure 46) Medical GPU utilization - 12x4Q.

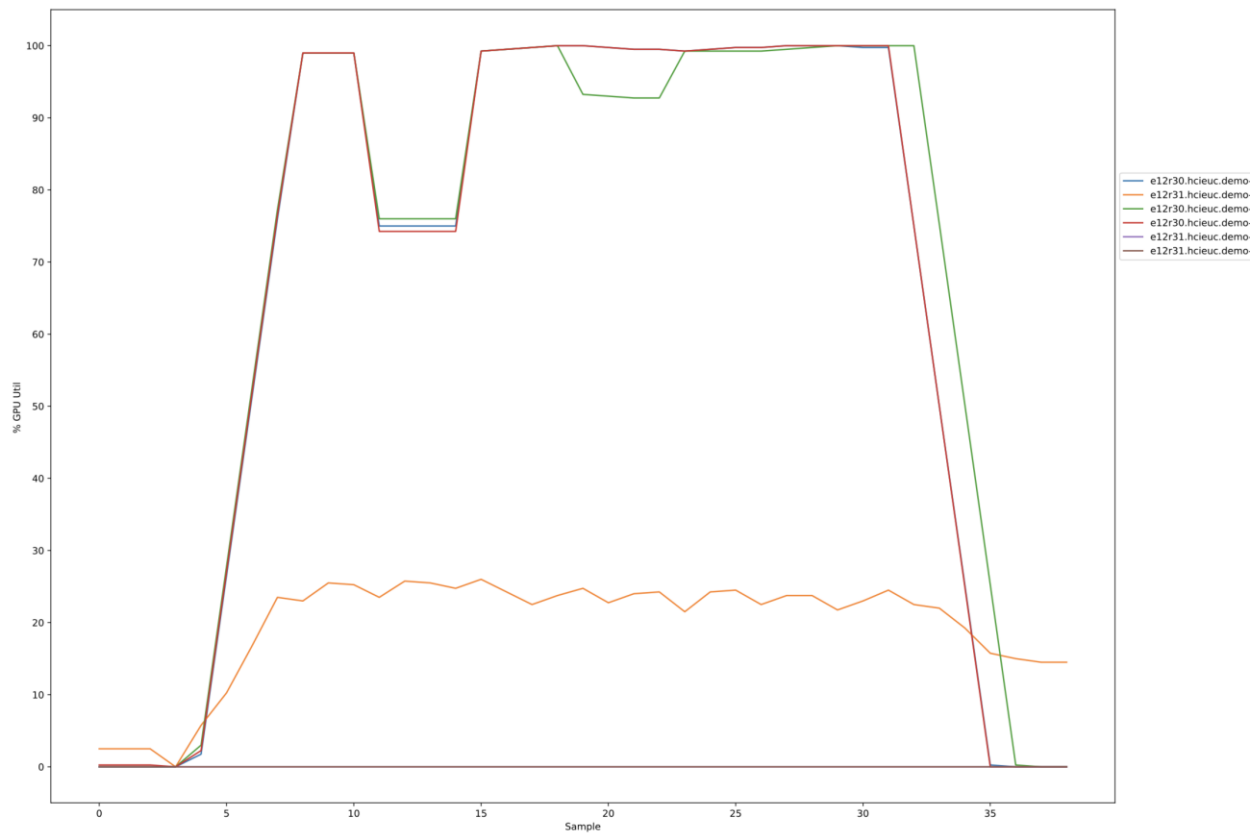
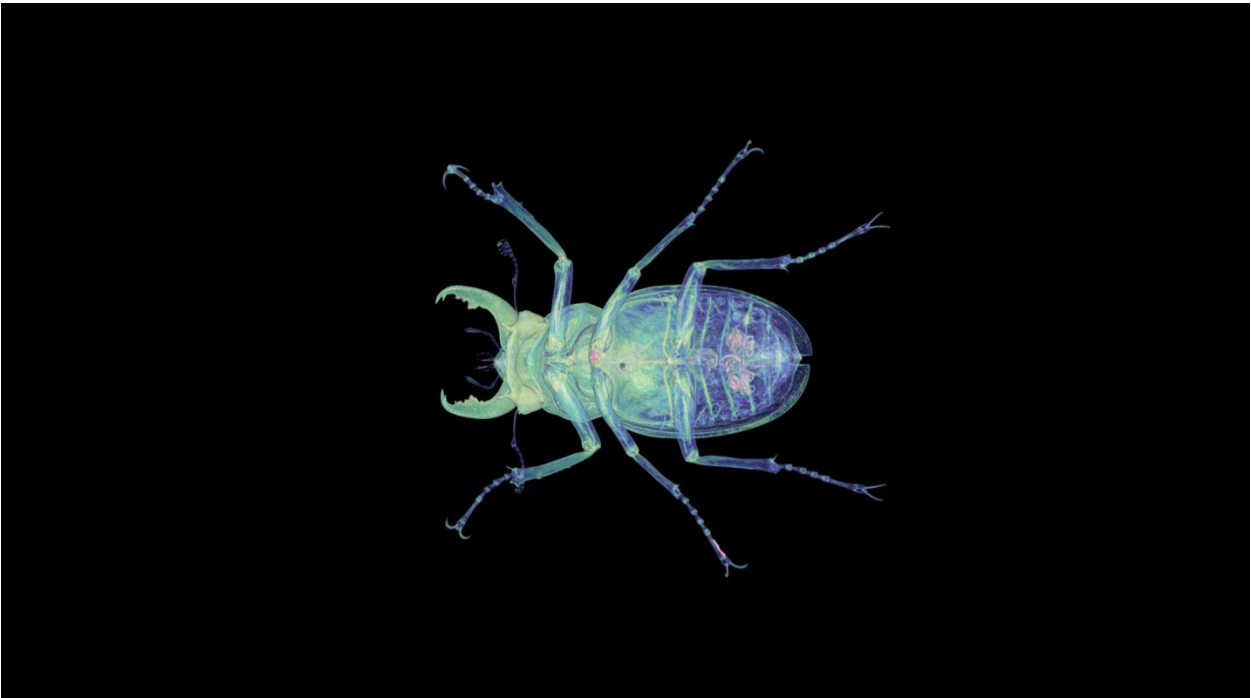


Figure 47) Medical sample.



6.8 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 are included in the viewset:

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

Figure 48) Showcase composite score.

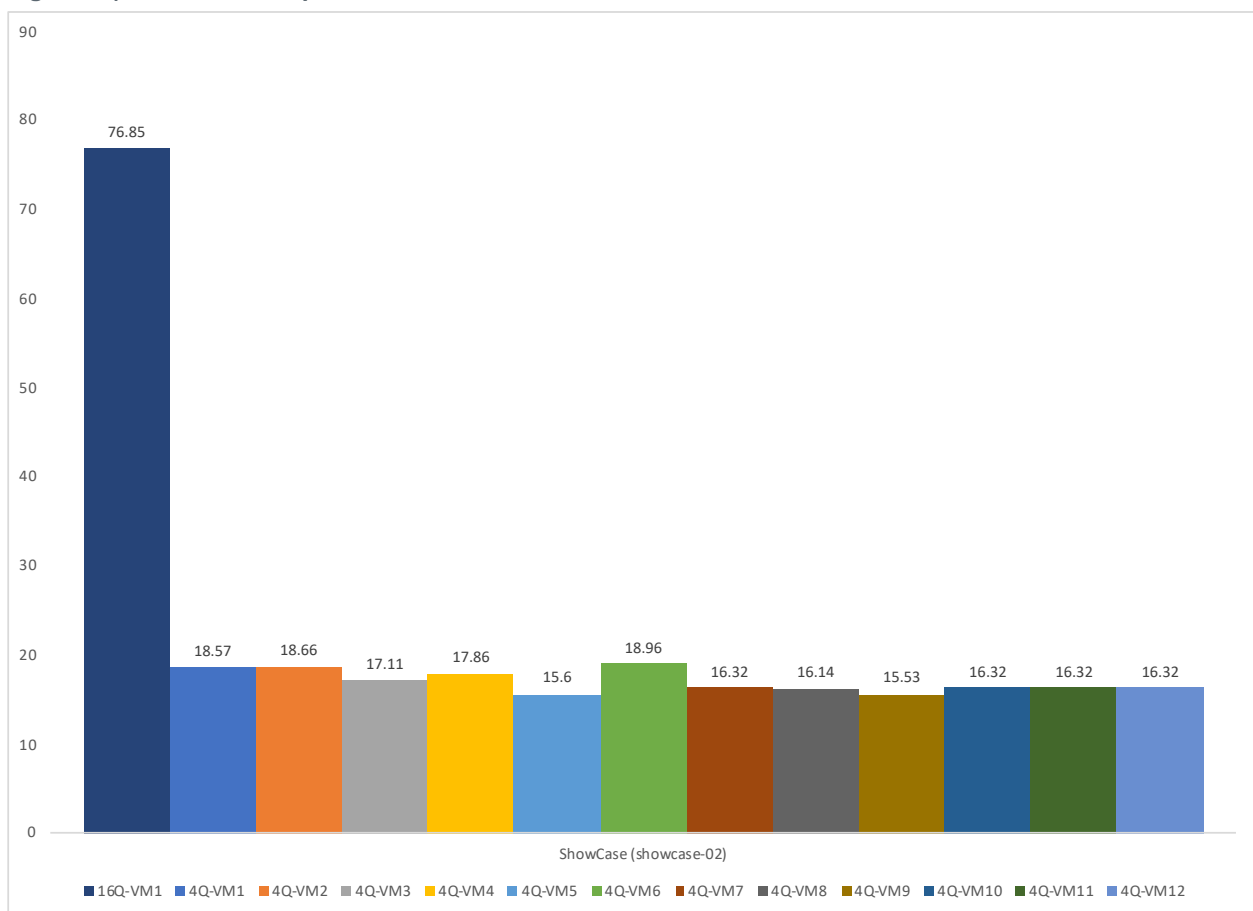


Figure 49) Showcase vSphere CPU utilization - 1x16Q.

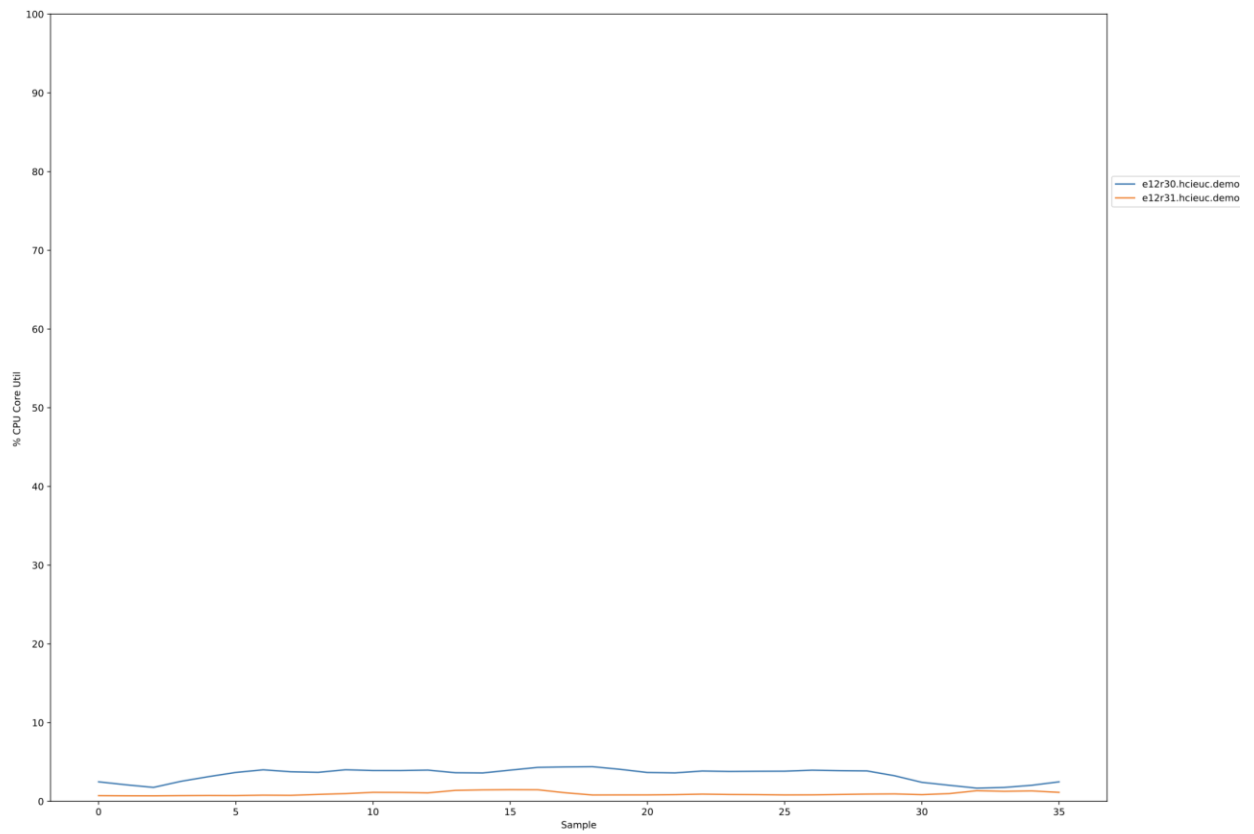


Figure 50) Showcase vSphere CPU utilization - 12x4Q.

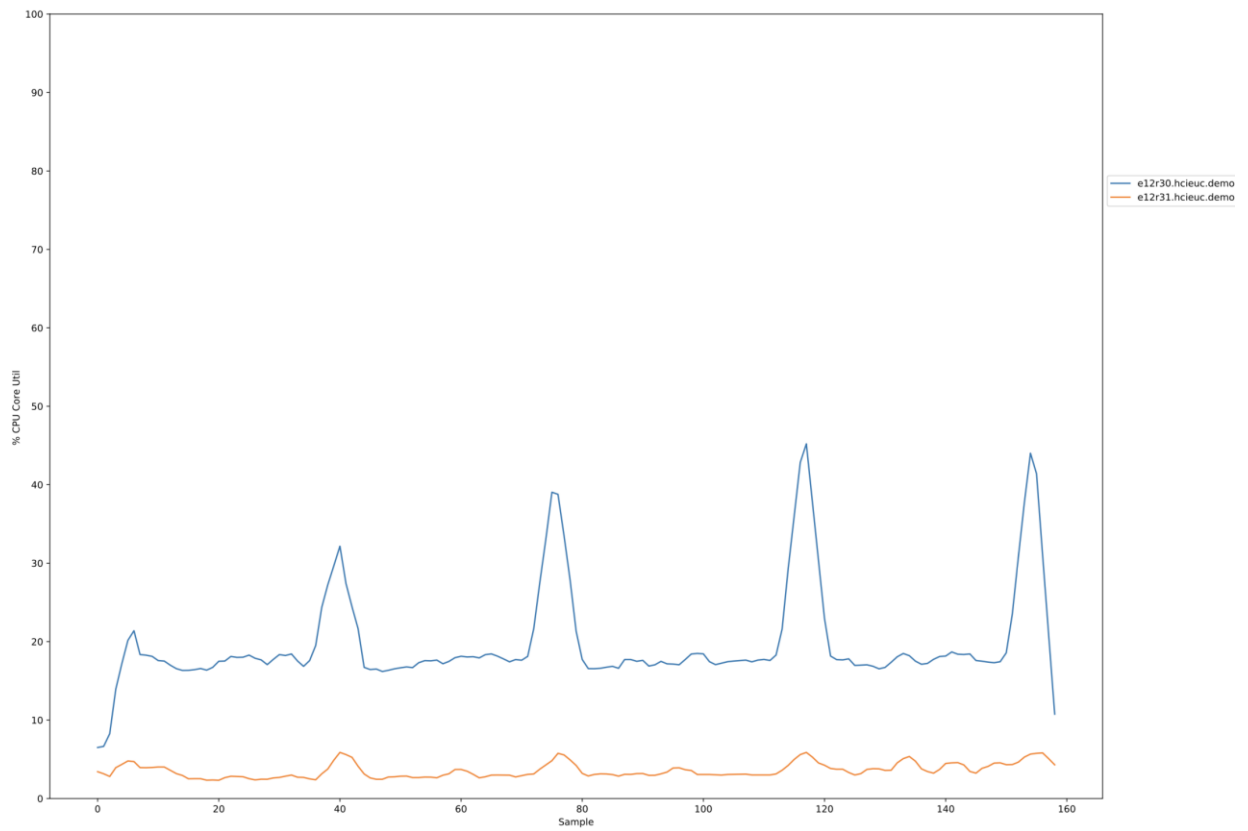


Figure 51) Showcase GPU utilization - 1x16Q.

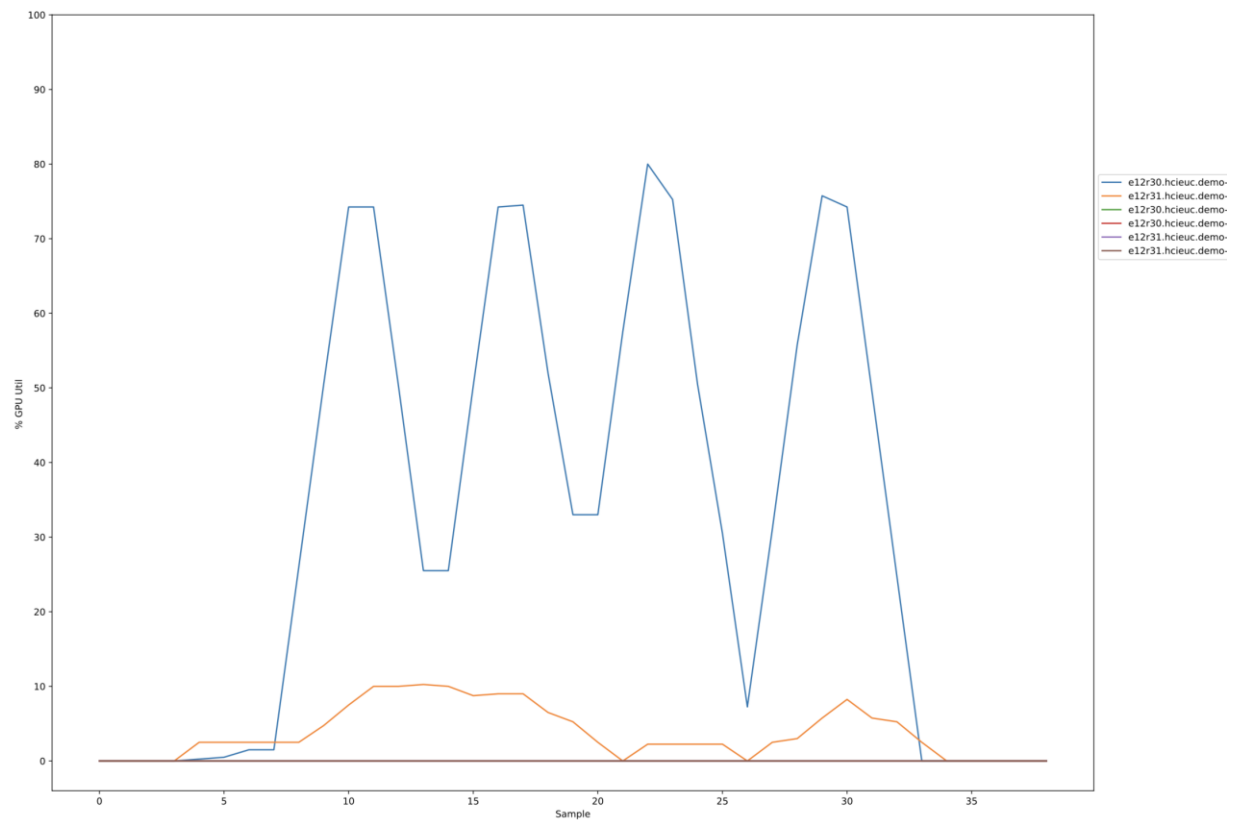


Figure 52) Showcase GPU utilization - 12x4Q.

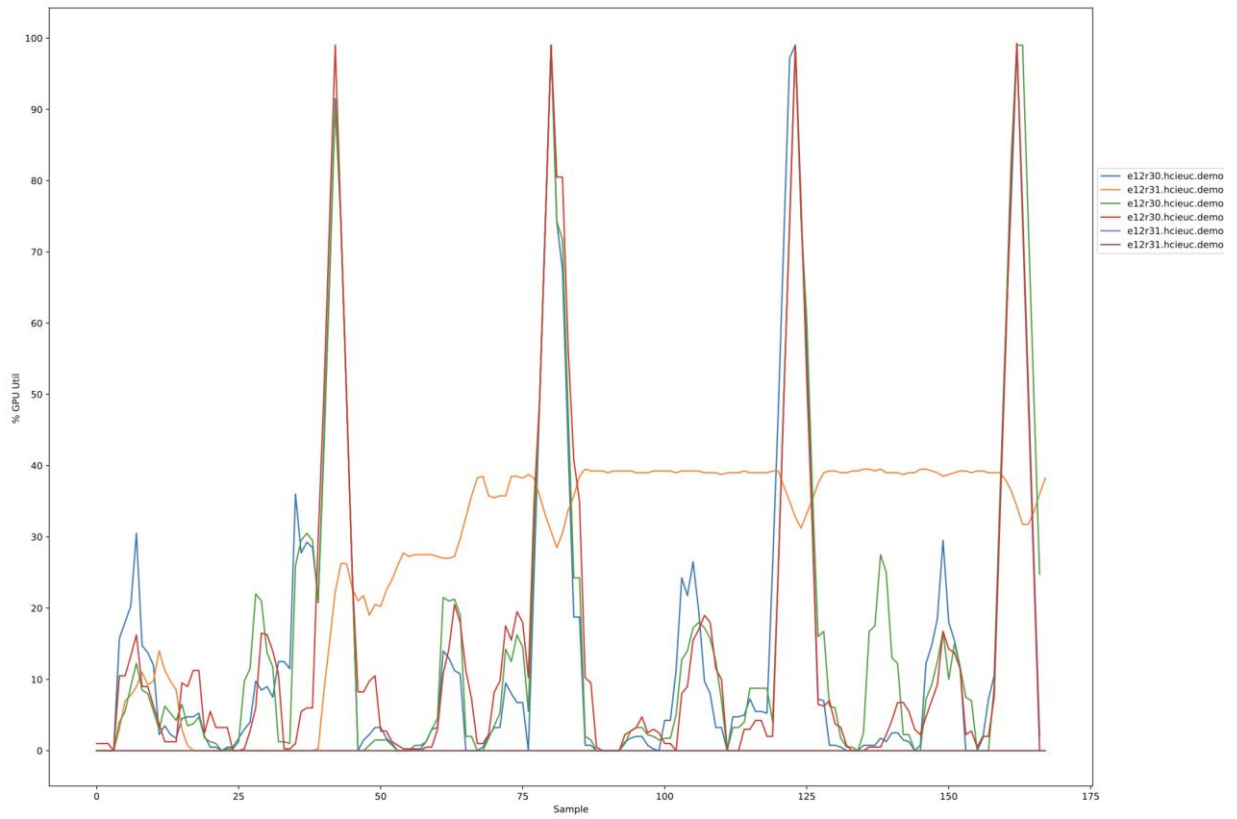
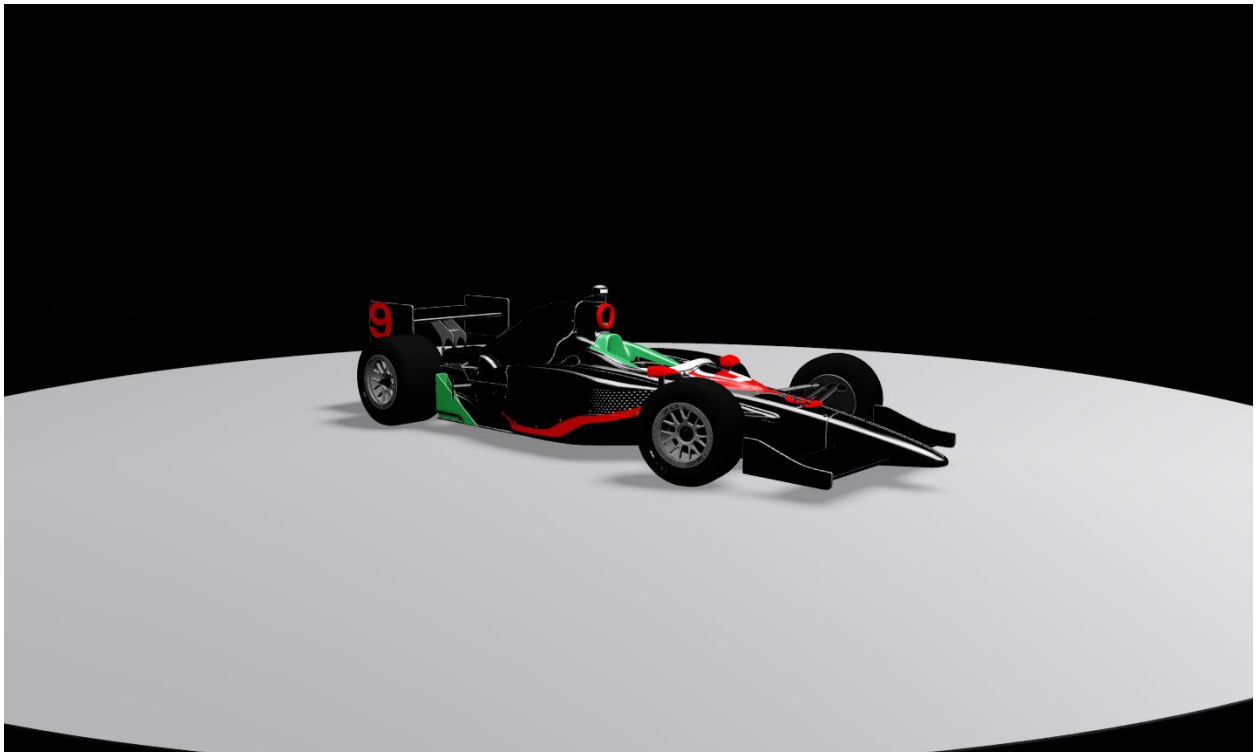


Figure 53) Showcase sample.



6.9 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, anti-aliasing, shaded, shaded with edges, and studio mode.

The following tests are included in the viewset:

The following tests are included within 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

Figure 54) Siemens NX composite score.

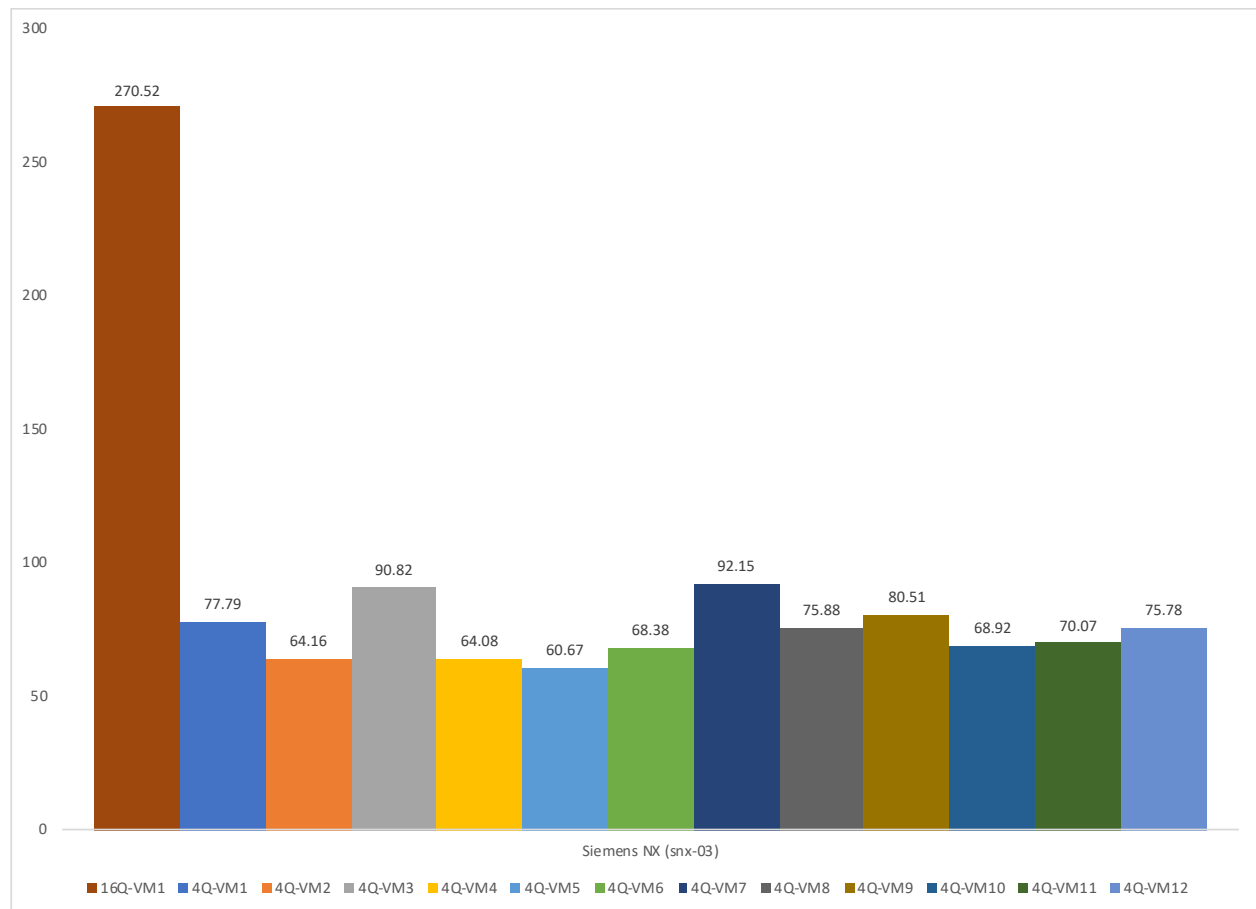


Figure 55) Siemens NX vSphere CPU utilization - 1x16Q.

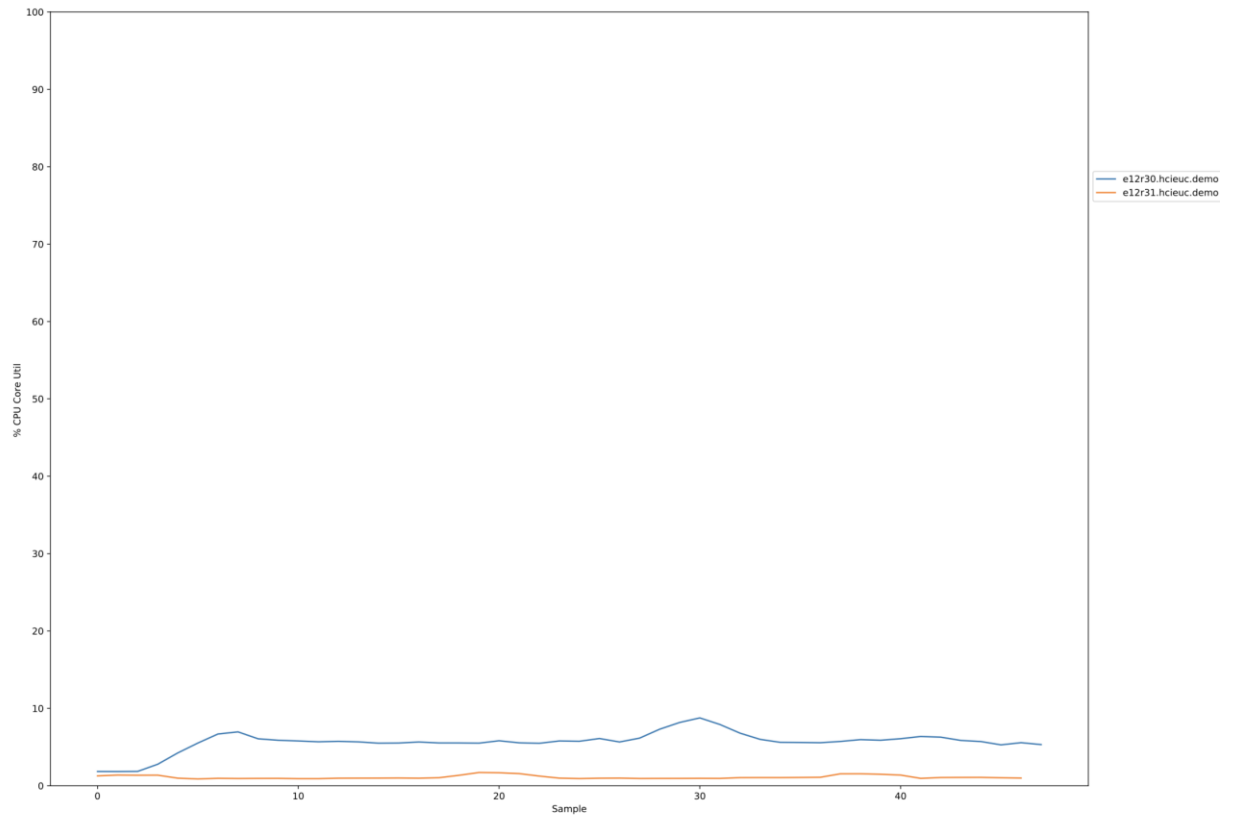


Figure 56) Siemens NX vSphere CPU utilization - 12x4Q.

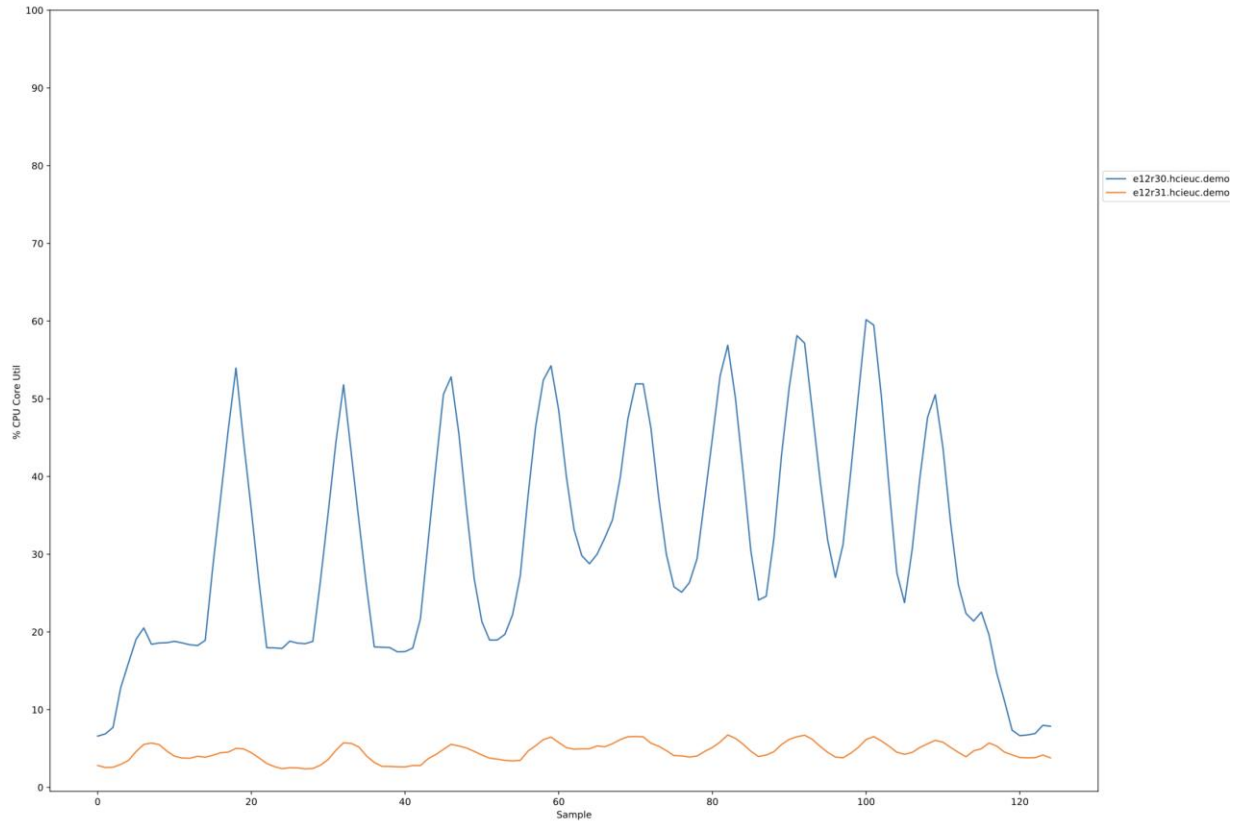


Figure 57) Siemens NX GPU utilization - 1x16Q.

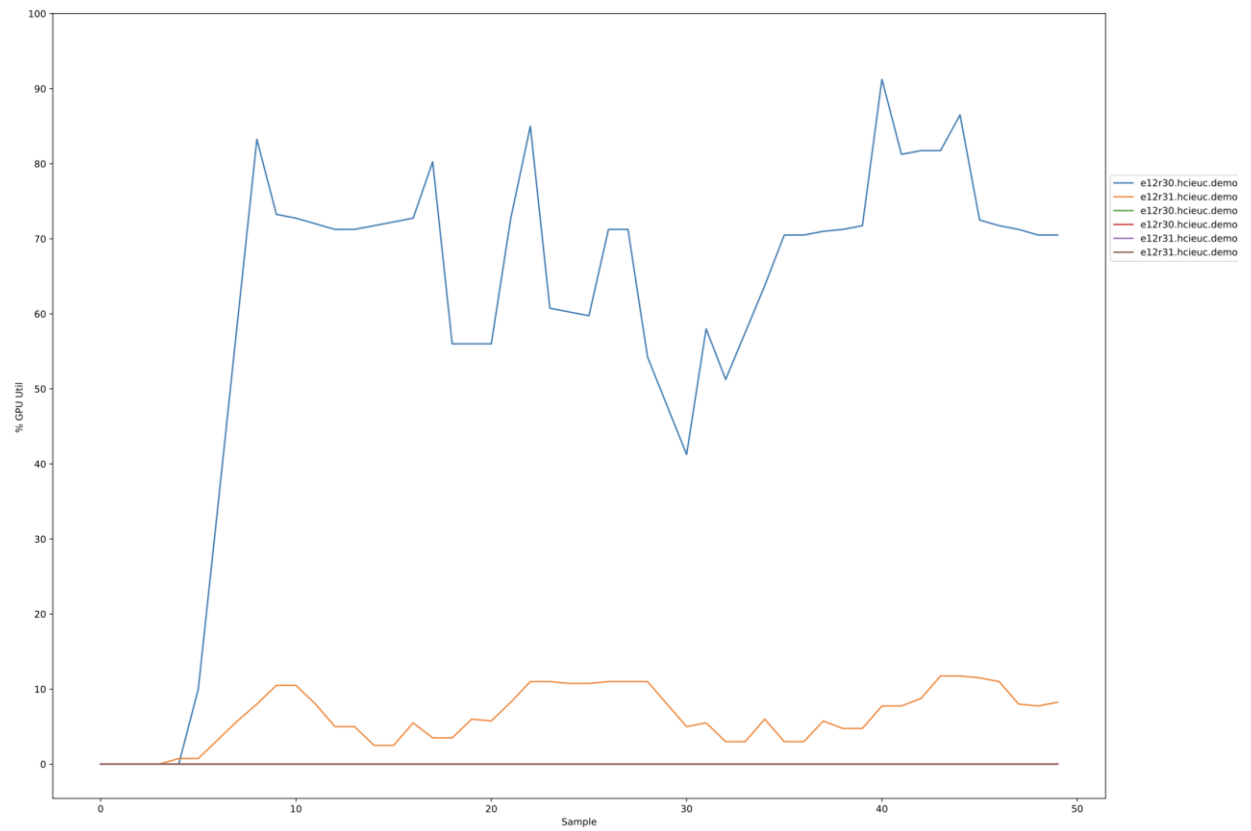


Figure 58) Siemens NX GPU utilization - 12x4Q.

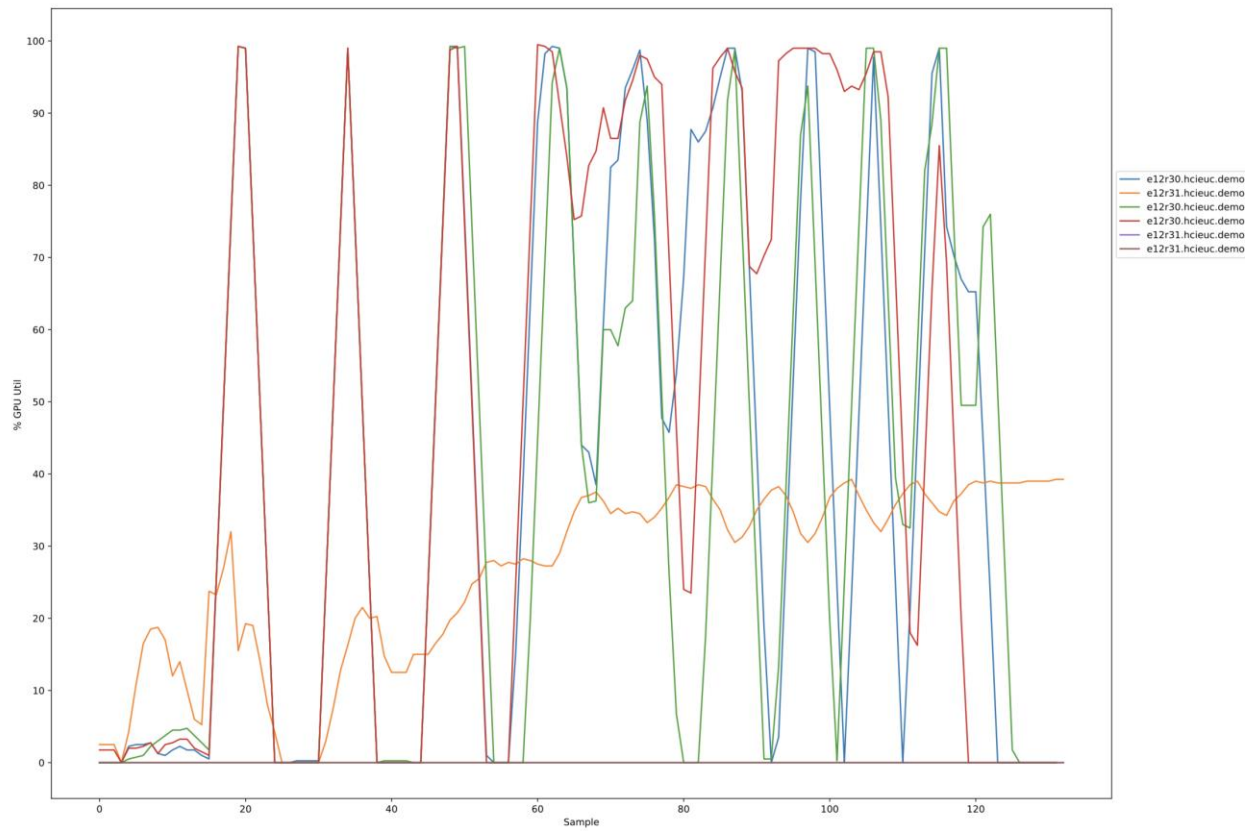
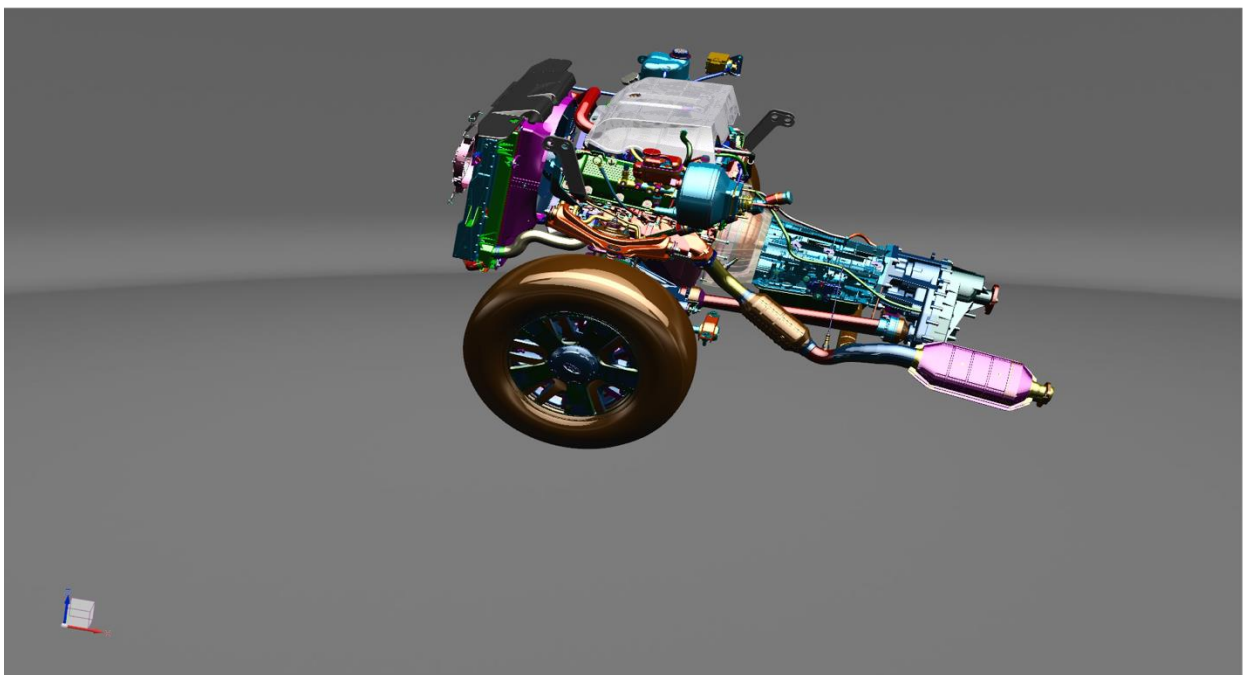


Figure 59) Siemens NX Sample.



6.10 Solidworks (sw-04)

The sw-04 viewset was created from traces of Dassault Systemes' 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 are 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

Figure 60) Solidworks composite score.

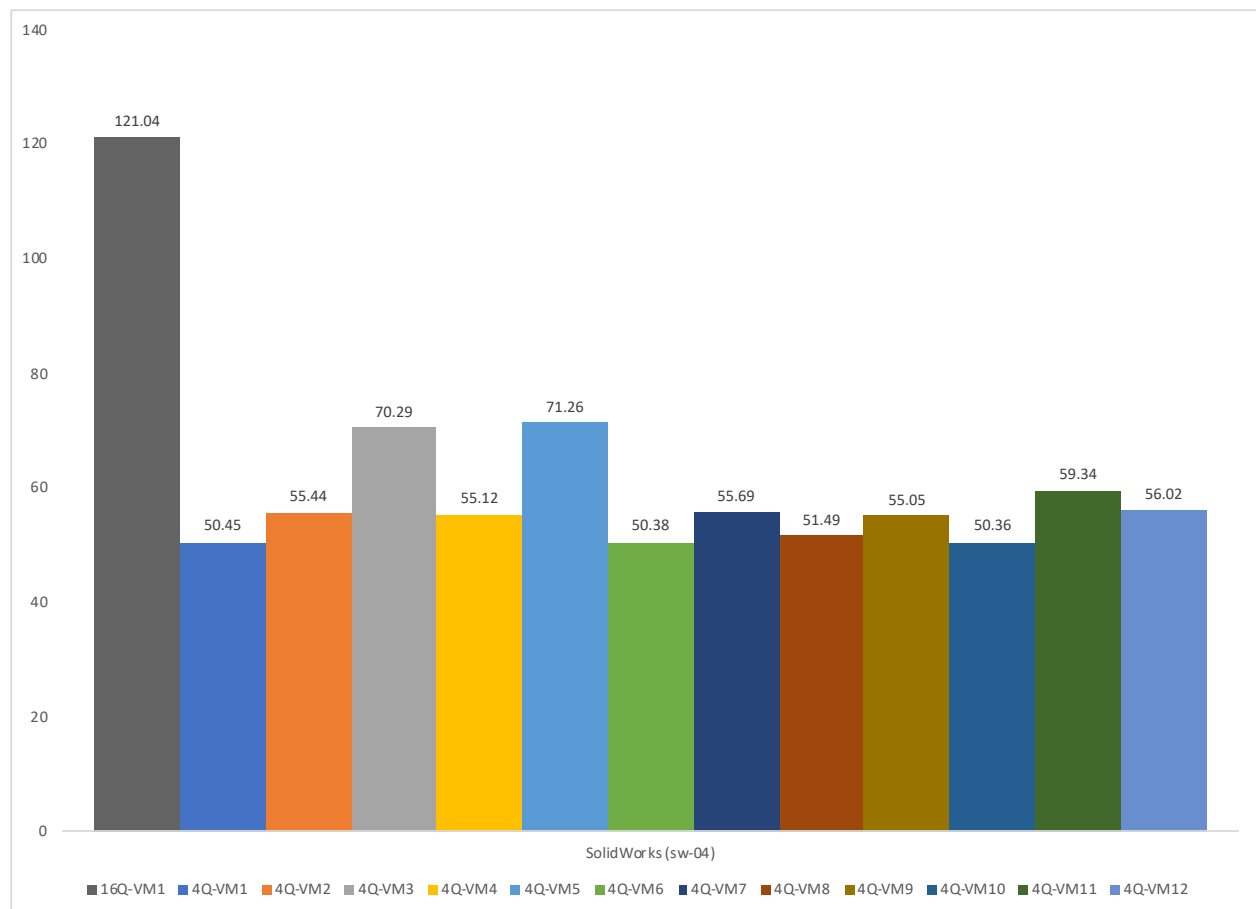


Figure 61) Solidworks vSphere CPU utilization - 1x16Q.

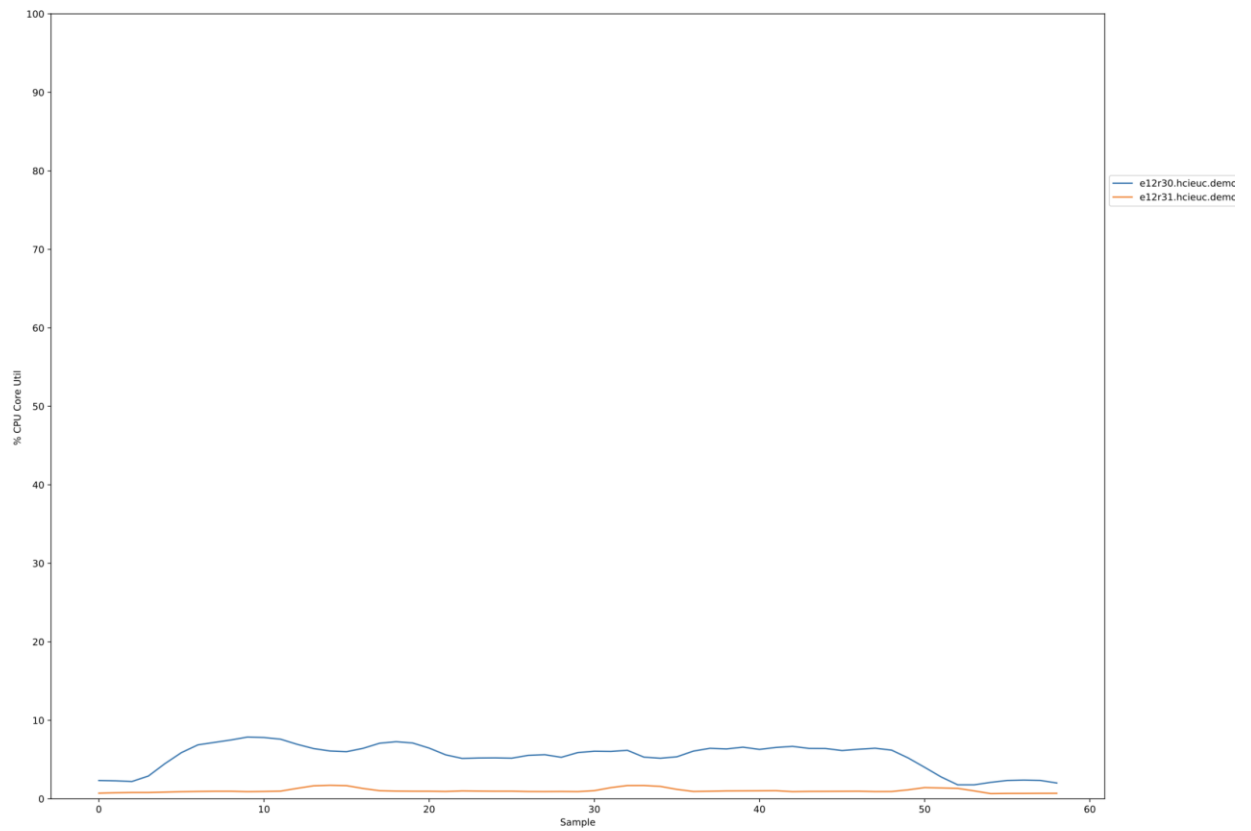


Figure 62) Solidworks vSphere CPU utilization - 12x4Q.

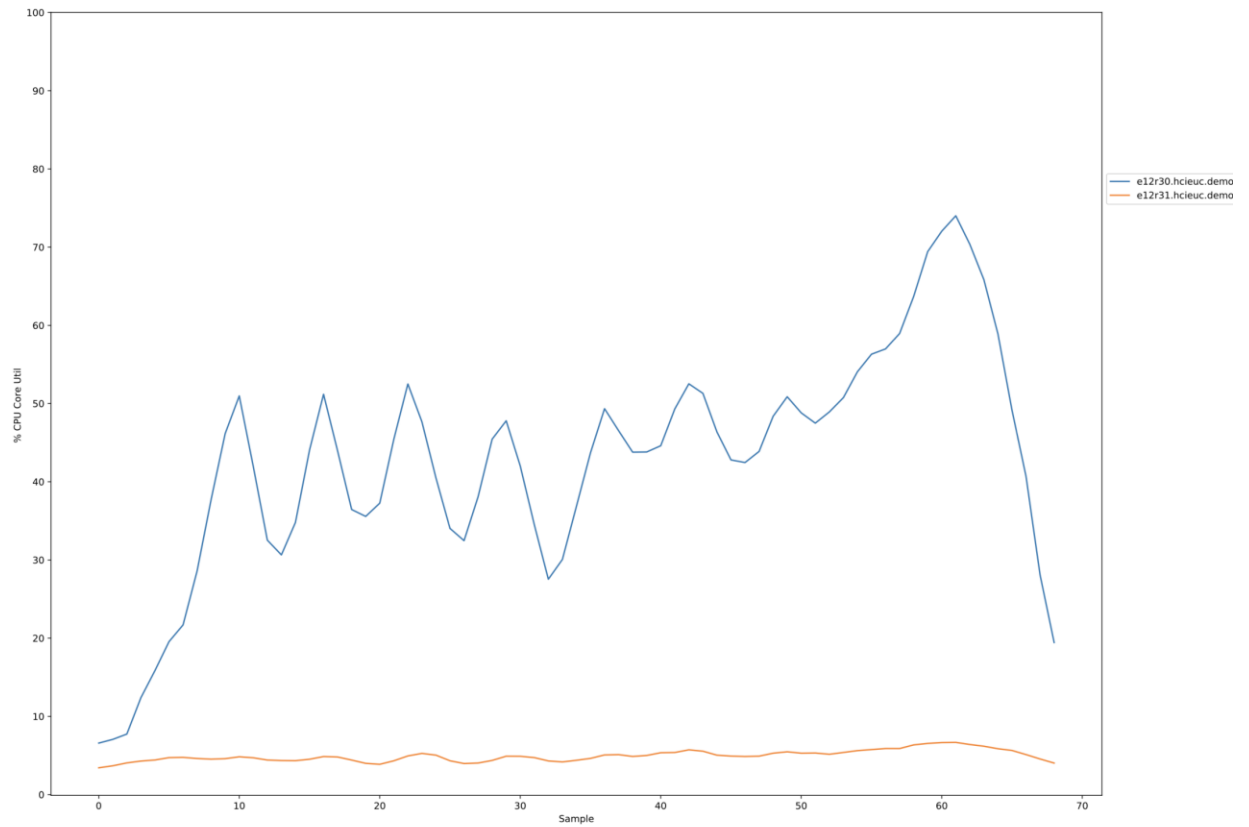


Figure 63) Solidworks GPU utilization - 1x16Q.

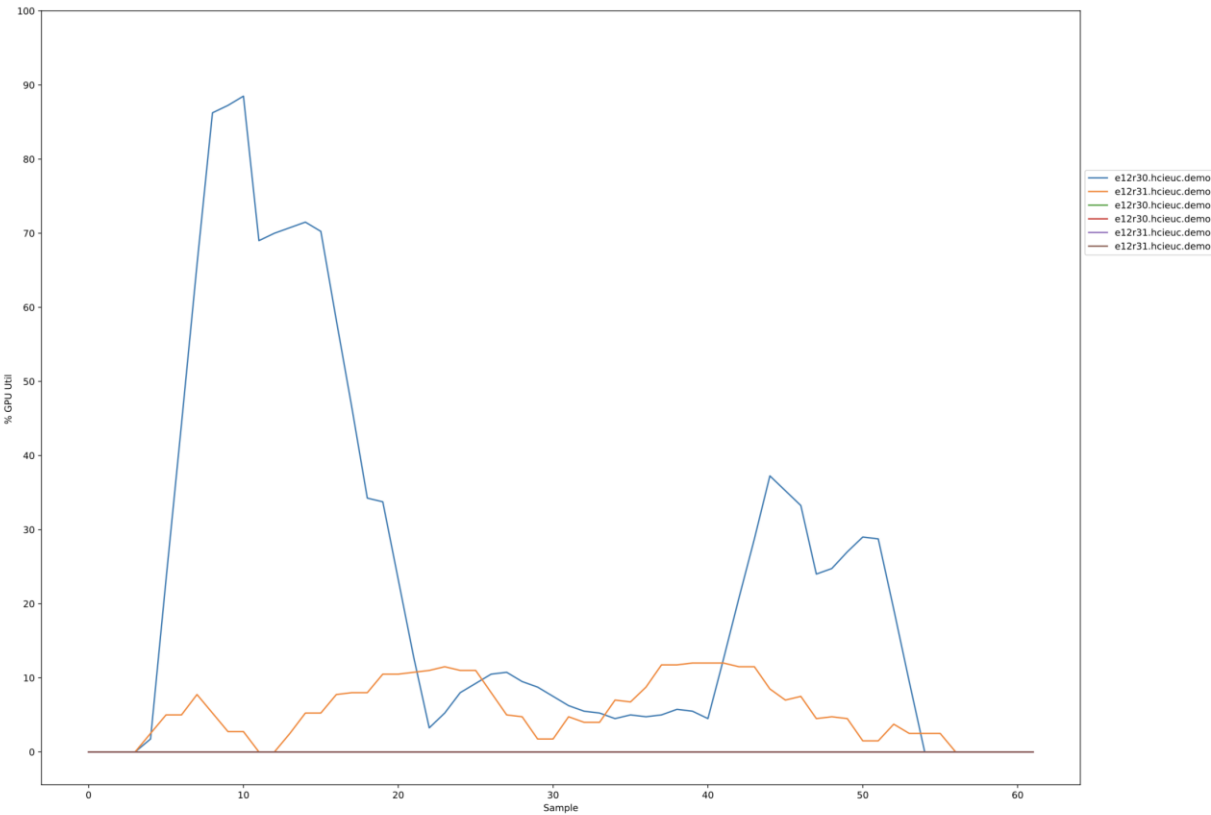


Figure 64) Solidworks GPU utilization - 12x4Q.

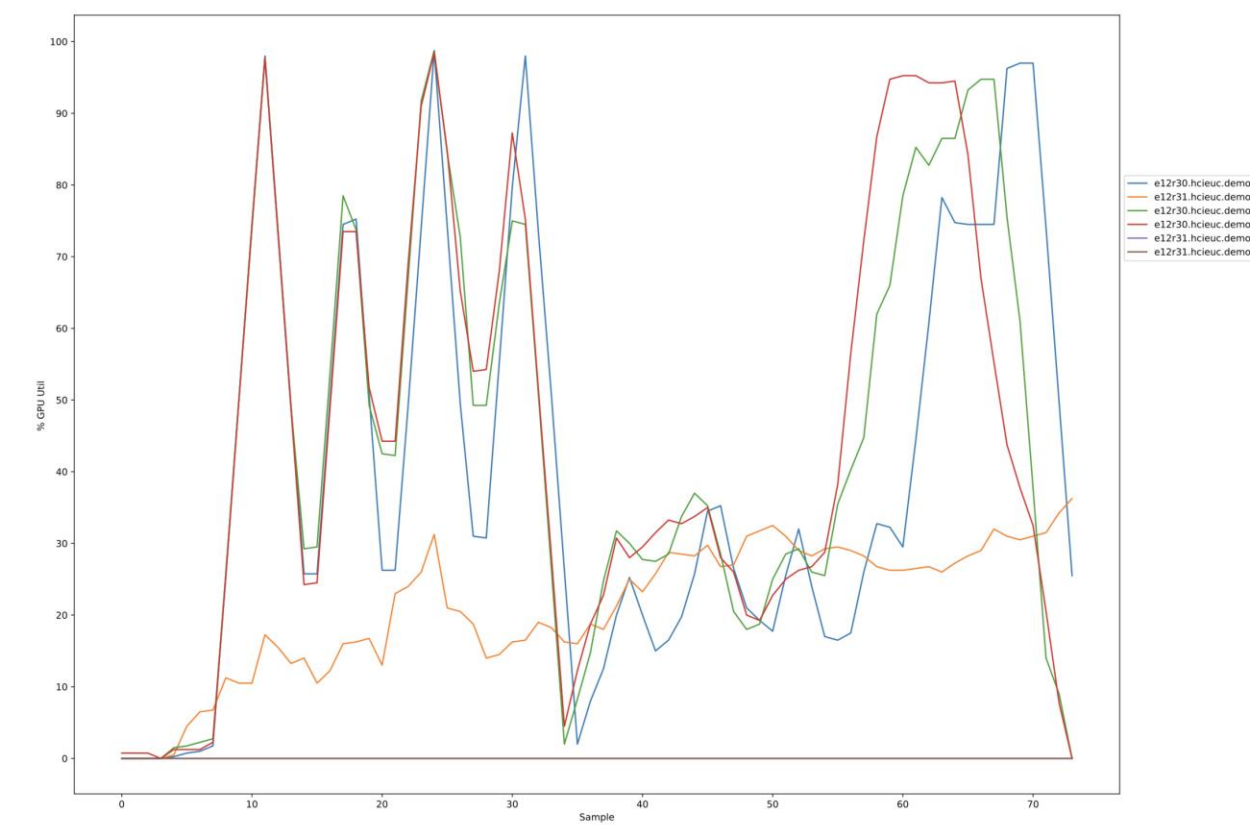
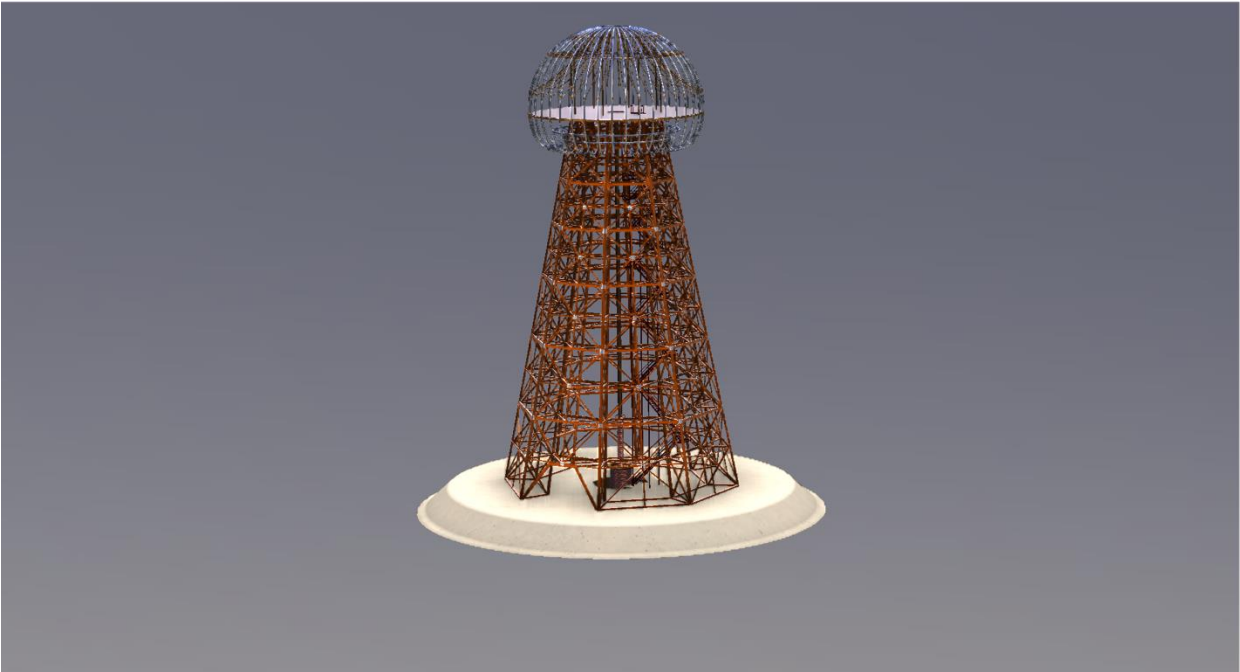


Figure 65) Solidworks sample.



7 Integration with NetApp Private Cloud

Customers looking for the following features can benefit from NetApp Private Cloud:

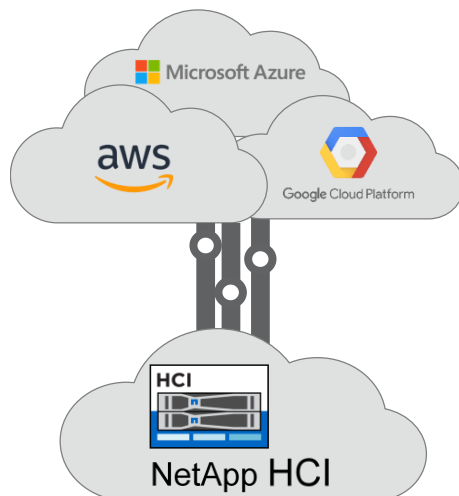
- A self-service portal
- Policy enforcement with features such as placement, access control, and so on
- An extensible framework to integrate with existing components
- Approval workflows
- Machine lifecycle management

If you already have NetApp Private Cloud, you can choose the deployment of desktop pools using blueprints and import into Horizon as a manual pool. VMware vRealize Orchestrator Plug-in for Horizon 7 is available to assist with those automation tasks.

8 Integration with NetApp Kubernetes Service

To maximize the GPU resources on H615C, GPU resources can be used for deep-learning inference tasks. NVIDIA GPU Cloud is a container registry that has ready-to-use containers to accelerate deployment of these tasks. NetApp Kubernetes Service provides a centralized management plane to deploy Kubernetes in either public clouds or on NetApp HCI systems.

Figure 66) The hybrid cloud.



The management node on NetApp HCI provides the Hybrid Cloud Control portal, from which you can also deploy NetApp Kubernetes Service (NKS).

Figure 67) Hybrid Cloud Control.

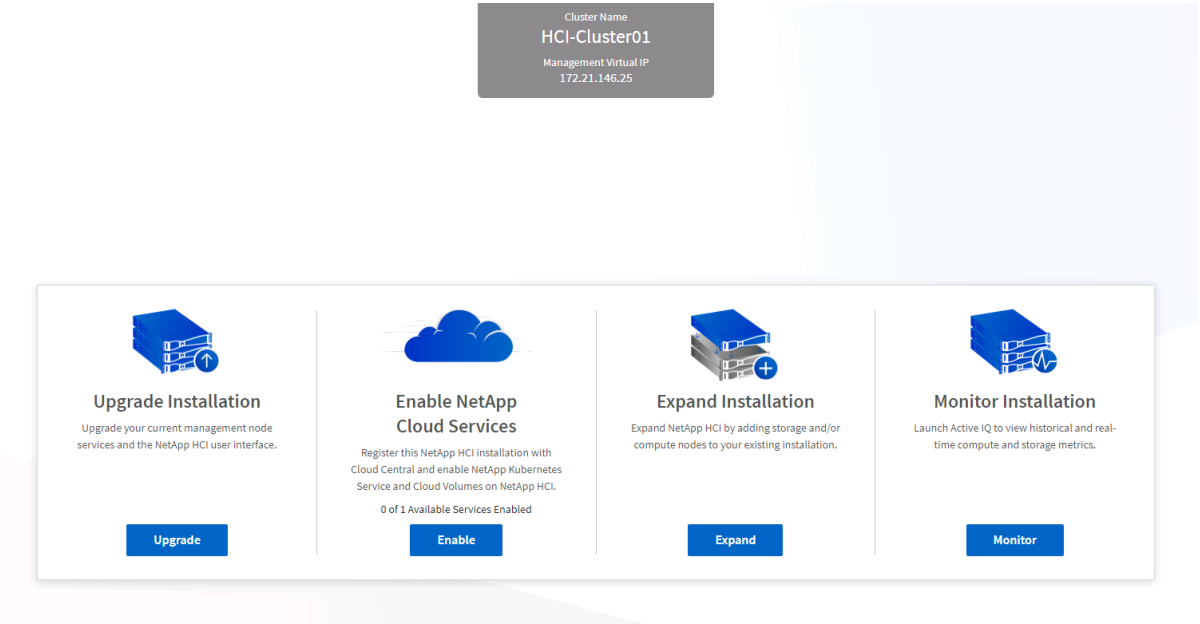
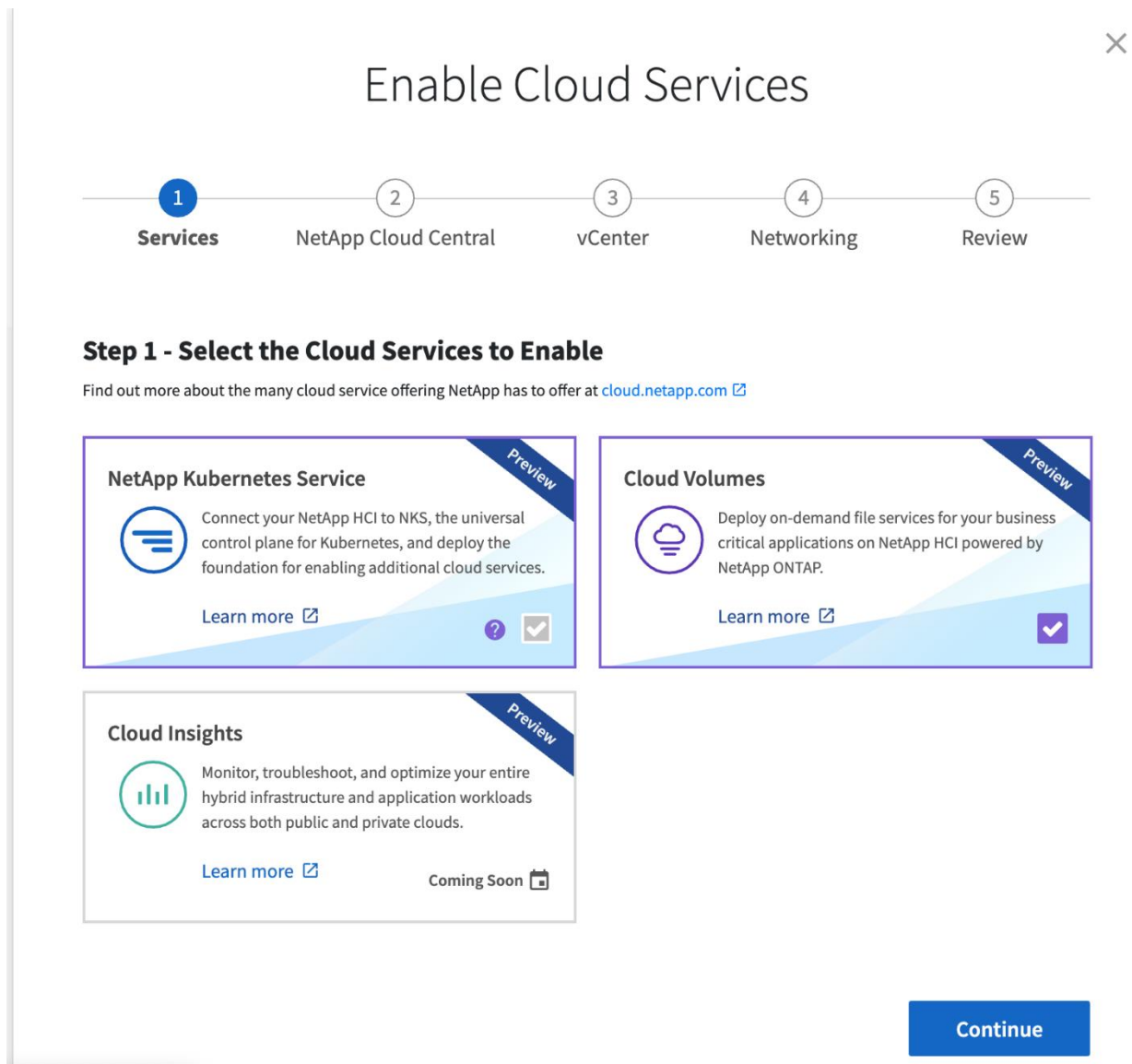
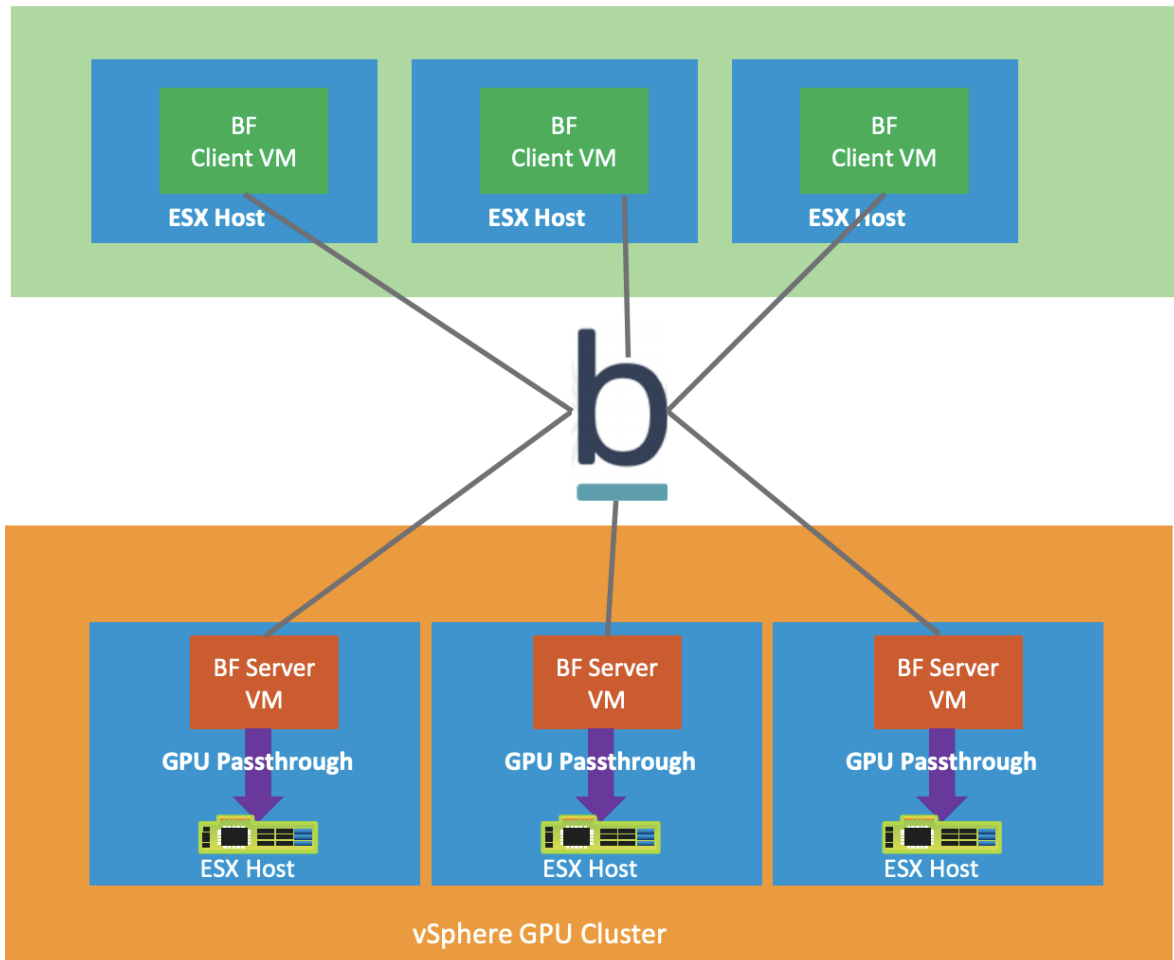


Figure 68) Enable NetApp Cloud Services.



If you would like to run deep-learning inference containers on non-GPU nodes, you should consider FlexDirect from Bitfusion. FlexDirect is a client server architecture in which the server component runs on GPU nodes and the client component can run on different hosts communicating over IP transport.

Figure 69) Bitfusion FlexDirect.



You can view a FlexDirect demonstration video [here](#).

Bitfusion is ideal when a CUDA AI, machine learning, or HPC application is deployed on a server CPU that does not have a GPU. It does not support graphics or virtual client computing (for example, VDI). However, for high performance and workload isolation for 100% CUDA-compatible AI, machine learning, or HPC applications, NetApp recommends that you run apps on a server with NVIDIA GPUs.

9 Summary

The NetApp HCI H615C is well suited for a wide range of data center workloads including the following:

- Virtual desktops for knowledge workers using modern productivity applications
- Virtual workstations for scientists, engineers, and creative professionals
- Deep learning inference computing

The H615C provides 50% more user density for the same rack space compared to the H610C and double the performance for most workloads.

Appendix A: GPU in vSphere

The H615C with vSphere 6.7 Update 1 and a T4 GPU is recognized as a <class> 3D controller.

Figure 70) T4 in vSphere 6.7 Update 1.

The screenshot shows the vSphere 6.7 Update 1 interface. The left sidebar contains a navigation tree with categories like Storage, Networking, Virtual Machines, and System. The main pane is titled 'Graphics Devices' and has two tabs: 'Graphics Devices' (selected) and 'Host Graphics'. Below the tabs is a table with the following data:

Name	Device ID	Vendor	Active Type	Configured Type	Memory
<class> 3D controller	0000:18:00.0	nVidia Corporation	Basic	Shared	0.00 B
<class> 3D controller	0000:18:01.0	nVidia Corporation	Basic	Shared	0.00 B
<class> 3D controller	0000:18:02.0	nVidia Corporation	Basic	Shared	0.00 B
<class> 3D controller	0000:18:03.0	nVidia Corporation	Basic	Shared	0.00 B
<class> 3D controller	0000:18:04.0	nVidia Corporation	Basic	Shared	0.00 B

Below the table, there is a section titled 'VMs associated with the graphics device "<class> 3D controller"'. It contains a table with columns: Name, State, Status, Provisioned Space, Used Space, Host CPU, and Host Mem. The table is empty, and a message at the bottom right says 'No items to display'.

After the H615C is updated to vSphere 6.7 Update 2, it is recognized as a Tesla T4.

Figure 71) T4 in vSphere 6.7 Update 2

The screenshot shows the vSphere 6.7 Update 2 interface. The left sidebar is the same as in Figure 70. The main pane is titled 'Graphics Devices' and has two tabs: 'Graphics Devices' (selected) and 'Host Graphics'. Below the tabs is a table with the following data:

Name	Device ID	Vendor	Active Type	Configured Type	Memory
TU104GL [Tesla T4]	0000:18:00.0	NVIDIA Corporation	Basic	Shared	0.00 B
TU104GL [Tesla T4]	0000:18:01.0	NVIDIA Corporation	Basic	Shared	0.00 B
TU104GL [Tesla T4]	0000:18:02.0	NVIDIA Corporation	Basic	Shared	0.00 B
TU104GL [Tesla T4]	0000:18:03.0	NVIDIA Corporation	Basic	Shared	0.00 B
TU104GL [Tesla T4]	0000:18:04.0	NVIDIA Corporation	Basic	Shared	0.00 B

Below the table, there is a section titled 'VMs associated with the graphics device "TU104GL [Tesla T4]"'. It contains a table with columns: Name, State, Status, Provisioned Space, Used Space, Host CPU, and Host Mem. The table is empty, and a message at the bottom right says 'No items to display'.

After an NVIDIA vGPU driver is installed and the type is changed to Shared Direct, the active type is still Shared until it is rebooted.

Figure 72) T4 with NVIDIA vGPU.

Summary

Monitor

Configure

Permissions

VMs

Datastores

Networks

More Objects

Updates

Swap File Location

▼ System

Licensing

Host Profile

Time Configuration

Authentication Servi...

Certificate

Power Management

Advanced System S...

System Resource Re...

Firewall

Services

Security Profile

System Swap

Packages

▼ Hardware

Processors

Memory

PCI Devices

Graphics

Power Management

▼ More

Alarm Definitions

Scheduled Tasks

Graphics Devices

Host Graphics

Graphics Devices

Name	Device ID	Vendor	Active Type	Configured Type	Memory
NVIDIA Tesla T4	0000:18:00.0	NVIDIA Corporation	Shared	Shared Direct	14.98 GB
NVIDIA Tesla T4	0000:18:01.0	NVIDIA Corporation	Basic	Shared Direct	0.00 B
NVIDIA Tesla T4	0000:18:02.0	NVIDIA Corporation	Basic	Shared Direct	0.00 B
NVIDIA Tesla T4	0000:18:03.0	NVIDIA Corporation	Basic	Shared Direct	0.00 B
NVIDIA Tesla T4	0000:18:04.0	NVIDIA Corporation	Basic	Shared Direct	0.00 B

96 Items

VMs associated with the graphics device "NVIDIATesla T4"

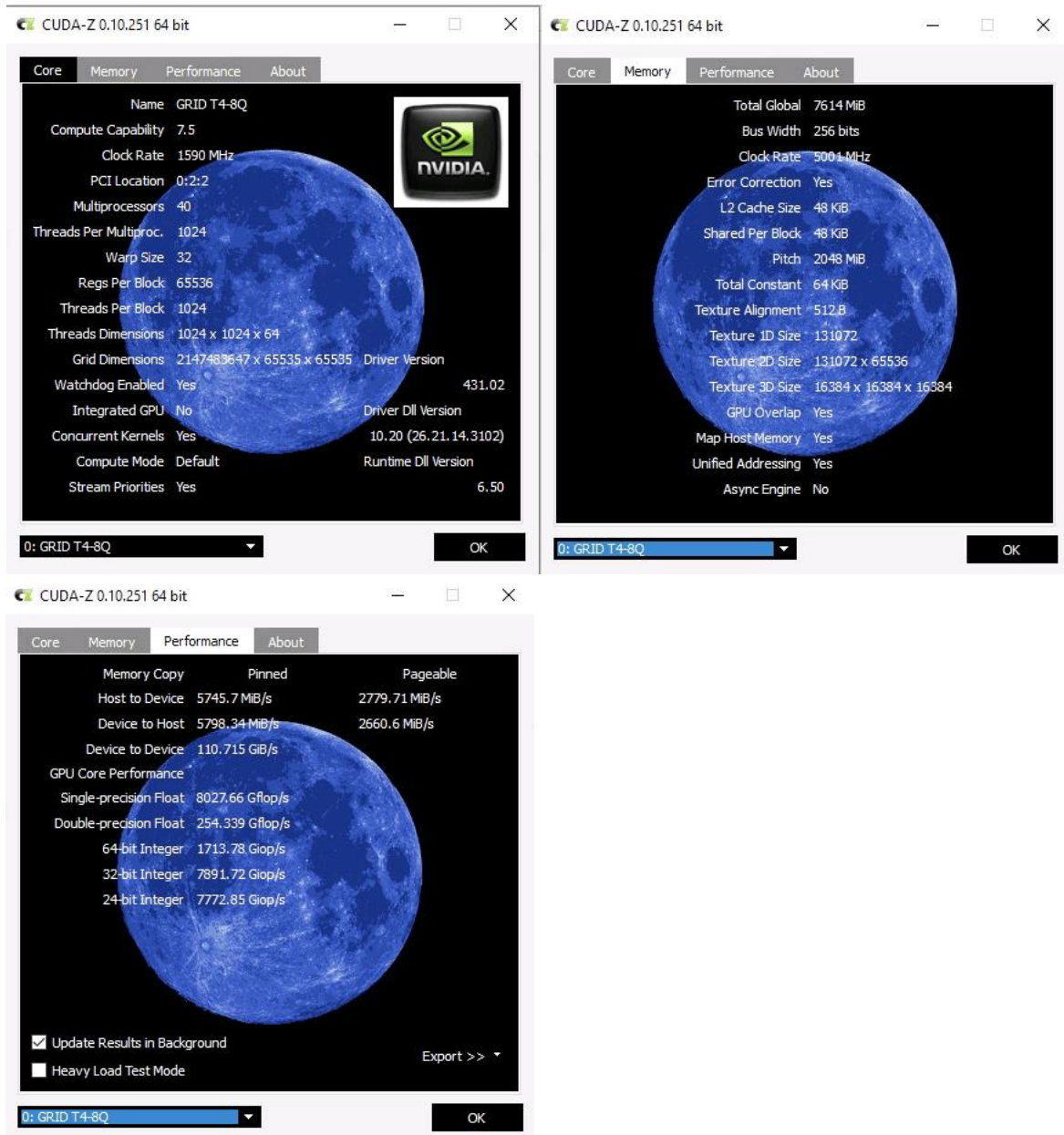
Name	State	Status	Provisioned Space	Used Space	Host CPU	Host Mem
------	-------	--------	-------------------	------------	----------	----------

No Items to display

Appendix B: T4 CUDA-Z Screenshots

CUDA-Z tool can be downloaded from <http://cuda-z.sourceforge.net/>, which provides information about CUDA-enabled GPUs. [Figure 73](#) shows the additional metrics of vGPU.

Figure 73) CUDA-Z screenshots



Appendix C: T4 vGPU Settings

To quickly check the number of displays, the monitor resolution supported for a specific GPU profile, or the framebuffer limit, see the `/usr/share/nvidia/vgpu/vgpuconfig.xml` file. Use this file for reference purposes only. Avoid making changes to it directly.

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE vgpu SYSTEM "http://www.nvidia.com/dtd/vgpuConfig.dtd">
<vgpuconfig>
  <version>1.0</version>
  <globalSettings>
    <homogeneousVgpus>TRUE</homogeneousVgpus>
    <pluginSoName>libnvidia-vgpu</pluginSoName>
  </globalSettings>
</vgpuconfig>
<vgpuType id="230" name="GRID T4-1Q" class="Quadro">
```



```

<devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x130C"/>
<framebuffer>0x38000000</framebuffer>
<numHeads>2</numHeads>
<display width="4096" height="2160"/>
<mappableVideoSize>0x400000</mappableVideoSize>
<fbReservation>0x8000000</fbReservation>
<frlConfig>0x3c</frlConfig>
<cudaEnabled>0x1</cudaEnabled>
<eccSupported>1</eccSupported>
<multiVgpuSupported>0x0</multiVgpuSupported>
<encoderCapacity>0x64</encoderCapacity>
<barlLength>0x100</barlLength>
<frame_rate_limiter>1</frame_rate_limiter>
<license>GRID-Virtual-WS,2.0;Quadro-Virtual-DWS,5.0;GRID-Virtual-WS-Ext,2.0</license>
</vgpuType>
<vgpuType id="231" name="GRID T4-2Q" class="Quadro">
  <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x130D"/>
  <framebuffer>0x74000000</framebuffer>
  <numHeads>4</numHeads>
  <display width="4096" height="2160"/>
  <mappableVideoSize>0x400000</mappableVideoSize>
  <fbReservation>0xC000000</fbReservation>
  <frlConfig>0x3c</frlConfig>
  <cudaEnabled>0x1</cudaEnabled>
  <eccSupported>1</eccSupported>
  <multiVgpuSupported>0x0</multiVgpuSupported>
  <encoderCapacity>0x64</encoderCapacity>
  <barlLength>0x100</barlLength>
  <frame_rate_limiter>1</frame_rate_limiter>
  <license>GRID-Virtual-WS,2.0;Quadro-Virtual-DWS,5.0;GRID-Virtual-WS-Ext,2.0</license>
</vgpuType>
<vgpuType id="232" name="GRID T4-4Q" class="Quadro">
  <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x130E"/>
  <framebuffer>0xEC000000</framebuffer>
  <numHeads>4</numHeads>
  <display width="4096" height="2160"/>
  <mappableVideoSize>0x400000</mappableVideoSize>
  <fbReservation>0x14000000</fbReservation>
  <frlConfig>0x3c</frlConfig>
  <cudaEnabled>0x1</cudaEnabled>
  <eccSupported>1</eccSupported>
  <multiVgpuSupported>0x0</multiVgpuSupported>
  <encoderCapacity>0x64</encoderCapacity>
  <barlLength>0x100</barlLength>
  <frame_rate_limiter>1</frame_rate_limiter>
  <license>GRID-Virtual-WS,2.0;Quadro-Virtual-DWS,5.0;GRID-Virtual-WS-Ext,2.0</license>
</vgpuType>
<vgpuType id="233" name="GRID T4-8Q" class="Quadro">
  <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x130F"/>
  <framebuffer>0x1DC00000</framebuffer>
  <numHeads>4</numHeads>
  <display width="4096" height="2160"/>
  <mappableVideoSize>0x400000</mappableVideoSize>
  <fbReservation>0x24000000</fbReservation>
  <frlConfig>0x3c</frlConfig>
  <cudaEnabled>0x1</cudaEnabled>
  <eccSupported>1</eccSupported>
  <multiVgpuSupported>0x0</multiVgpuSupported>
  <encoderCapacity>0x64</encoderCapacity>
  <barlLength>0x100</barlLength>
  <frame_rate_limiter>1</frame_rate_limiter>
  <license>GRID-Virtual-WS,2.0;Quadro-Virtual-DWS,5.0;GRID-Virtual-WS-Ext,2.0</license>
</vgpuType>
<vgpuType id="234" name="GRID T4-16Q" class="Quadro">
  <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x1310"/>
  <framebuffer>0x3BA40000</framebuffer>
  <numHeads>4</numHeads>
  <display width="4096" height="2160"/>
  <mappableVideoSize>0x400000</mappableVideoSize>
  <fbReservation>0x45C00000</fbReservation>
  <frlConfig>0x3c</frlConfig>

```

```

    <cudaEnabled>0x1</cudaEnabled>
    <eccSupported>1</eccSupported>
    <multiVgpuSupported>0x1</multiVgpuSupported>
    <encoderCapacity>0x64</encoderCapacity>
    <bar1Length>0x100</bar1Length>
    <frame_rate_limiter>1</frame_rate_limiter>
    <license>GRID-Virtual-WS,2.0;Quadro-Virtual-DWS,5.0;GRID-Virtual-WS-Ext,2.0</license>
</vgpuType>
<vgpuType id="319" name="GRID T4-4C" class="Compute">
    <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x139A"/>
    <framebuffer>0xEC000000</framebuffer>
    <numHeads>1</numHeads>
    <display width="4096" height="2160"/>
    <mappableVideoSize>0x400000</mappableVideoSize>
    <fbReservation>0x14000000</fbReservation>
    <frlConfig>0x3c</frlConfig>
    <cudaEnabled>0x1</cudaEnabled>
    <eccSupported>1</eccSupported>
    <multiVgpuSupported>0x0</multiVgpuSupported>
    <encoderCapacity>0x64</encoderCapacity>
    <bar1Length>0x100</bar1Length>
    <frame_rate_limiter>1</frame_rate_limiter>
    <license>NVIDIA-vComputeServer,9.0;Quadro-Virtual-DWS,5.0</license>
</vgpuType>
<vgpuType id="320" name="GRID T4-8C" class="Compute">
    <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x139B"/>
    <framebuffer>0x1DC00000</framebuffer>
    <numHeads>1</numHeads>
    <display width="4096" height="2160"/>
    <mappableVideoSize>0x400000</mappableVideoSize>
    <fbReservation>0x24000000</fbReservation>
    <frlConfig>0x3c</frlConfig>
    <cudaEnabled>0x1</cudaEnabled>
    <eccSupported>1</eccSupported>
    <multiVgpuSupported>0x0</multiVgpuSupported>
    <encoderCapacity>0x64</encoderCapacity>
    <bar1Length>0x100</bar1Length>
    <frame_rate_limiter>1</frame_rate_limiter>
    <license>NVIDIA-vComputeServer,9.0;Quadro-Virtual-DWS,5.0</license>
</vgpuType>
<vgpuType id="321" name="GRID T4-16C" class="Compute">
    <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x1375"/>
    <framebuffer>0x3BA40000</framebuffer>
    <numHeads>1</numHeads>
    <display width="4096" height="2160"/>
    <mappableVideoSize>0x400000</mappableVideoSize>
    <fbReservation>0x45C00000</fbReservation>
    <frlConfig>0x3c</frlConfig>
    <cudaEnabled>0x1</cudaEnabled>
    <eccSupported>1</eccSupported>
    <multiVgpuSupported>0x1</multiVgpuSupported>
    <encoderCapacity>0x64</encoderCapacity>
    <bar1Length>0x100</bar1Length>
    <frame_rate_limiter>1</frame_rate_limiter>
    <license>NVIDIA-vComputeServer,9.0;Quadro-Virtual-DWS,5.0</license>
</vgpuType>
<vgpuType id="225" name="GRID T4-1A" class="NVS">
    <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x1311"/>
    <framebuffer>0x38000000</framebuffer>
    <numHeads>1</numHeads>
    <display width="1280" height="1024"/>
    <mappableVideoSize>0x400000</mappableVideoSize>
    <fbReservation>0x8000000</fbReservation>
    <frlConfig>0x3c</frlConfig>
    <cudaEnabled>0x0</cudaEnabled>
    <eccSupported>0</eccSupported>
    <multiVgpuSupported>0x0</multiVgpuSupported>
    <encoderCapacity>0x64</encoderCapacity>
    <bar1Length>0x100</bar1Length>
    <frame_rate_limiter>1</frame_rate_limiter>
    <license>GRID-Virtual-Apps,3.0</license>

```

```

</vgpuType>
<vgpuType id="226" name="GRID T4-2A" class="NVS">
  <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x1312"/>
  <framebuffer>0x74000000</framebuffer>
  <numHeads>1</numHeads>
  <display width="1280" height="1024"/>
  <mappableVideoSize>0x400000</mappableVideoSize>
  <fbReservation>0xC000000</fbReservation>
  <frlConfig>0x3c</frlConfig>
  <cudaEnabled>0x0</cudaEnabled>
  <eccSupported>0</eccSupported>
  <multiVgpuSupported>0x0</multiVgpuSupported>
  <encoderCapacity>0x64</encoderCapacity>
  <bar1Length>0x100</bar1Length>
  <frame_rate_limiter>1</frame_rate_limiter>
  <license>GRID-Virtual-Apps,3.0</license>
</vgpuType>
<vgpuType id="227" name="GRID T4-4A" class="NVS">
  <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x1313"/>
  <framebuffer>0xEC000000</framebuffer>
  <numHeads>1</numHeads>
  <display width="1280" height="1024"/>
  <mappableVideoSize>0x400000</mappableVideoSize>
  <fbReservation>0x14000000</fbReservation>
  <frlConfig>0x3c</frlConfig>
  <cudaEnabled>0x0</cudaEnabled>
  <eccSupported>0</eccSupported>
  <multiVgpuSupported>0x0</multiVgpuSupported>
  <encoderCapacity>0x64</encoderCapacity>
  <bar1Length>0x100</bar1Length>
  <frame_rate_limiter>1</frame_rate_limiter>
  <license>GRID-Virtual-Apps,3.0</license>
</vgpuType>
<vgpuType id="228" name="GRID T4-8A" class="NVS">
  <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x1314"/>
  <framebuffer>0x1DC000000</framebuffer>
  <numHeads>1</numHeads>
  <display width="1280" height="1024"/>
  <mappableVideoSize>0x400000</mappableVideoSize>
  <fbReservation>0x24000000</fbReservation>
  <frlConfig>0x3c</frlConfig>
  <cudaEnabled>0x0</cudaEnabled>
  <eccSupported>0</eccSupported>
  <multiVgpuSupported>0x0</multiVgpuSupported>
  <encoderCapacity>0x64</encoderCapacity>
  <bar1Length>0x100</bar1Length>
  <frame_rate_limiter>1</frame_rate_limiter>
  <license>GRID-Virtual-Apps,3.0</license>
</vgpuType>
<vgpuType id="229" name="GRID T4-16A" class="NVS">
  <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x1315"/>
  <framebuffer>0x3BA400000</framebuffer>
  <numHeads>1</numHeads>
  <display width="1280" height="1024"/>
  <mappableVideoSize>0x400000</mappableVideoSize>
  <fbReservation>0x45C00000</fbReservation>
  <frlConfig>0x3c</frlConfig>
  <cudaEnabled>0x0</cudaEnabled>
  <eccSupported>0</eccSupported>
  <multiVgpuSupported>0x0</multiVgpuSupported>
  <encoderCapacity>0x64</encoderCapacity>
  <bar1Length>0x100</bar1Length>
  <frame_rate_limiter>1</frame_rate_limiter>
  <license>GRID-Virtual-Apps,3.0</license>
</vgpuType>
<vgpuType id="222" name="GRID T4-1B" class="NVS">
  <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x1309"/>
  <framebuffer>0x38000000</framebuffer>
  <numHeads>4</numHeads>
  <display width="2560" height="1600"/>
  <mappableVideoSize>0x400000</mappableVideoSize>

```

```

        <fbReservation>0x8000000</fbReservation>
        <ftrlConfig>0x2d</ftrlConfig>
        <cudaEnabled>0x0</cudaEnabled>
        <eccSupported>0</eccSupported>
        <multiVgpuSupported>0x0</multiVgpuSupported>
        <encoderCapacity>0x64</encoderCapacity>
        <barlLength>0x100</barlLength>
        <frame_rate_limiter>1</frame_rate_limiter>
        <license>GRID-Virtual-PC,2.0;GRID-Virtual-WS,2.0;Quadro-Virtual-DWS,5.0;GRID-Virtual-WS-
Ext,2.0</license>
    </vgpuType>
    <vgpuType id="252" name="GRID T4-1B4" class="NVS">
        <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x1345"/>
        <framebuffer>0x38000000</framebuffer>
        <numHeads>1</numHeads>
        <display width="4096" height="2160"/>
        <mappableVideoSize>0x400000</mappableVideoSize>
        <fbReservation>0x8000000</fbReservation>
        <ftrlConfig>0x2d</ftrlConfig>
        <cudaEnabled>0x0</cudaEnabled>
        <eccSupported>0</eccSupported>
        <multiVgpuSupported>0x0</multiVgpuSupported>
        <encoderCapacity>0x64</encoderCapacity>
        <barlLength>0x100</barlLength>
        <frame_rate_limiter>1</frame_rate_limiter>
        <license>GRID-Virtual-PC,2.0;GRID-Virtual-WS,2.0;Quadro-Virtual-DWS,5.0;GRID-Virtual-WS-
Ext,2.0</license>
    </vgpuType>
    <vgpuType id="223" name="GRID T4-2B" class="NVS">
        <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x130A"/>
        <framebuffer>0x74000000</framebuffer>
        <numHeads>2</numHeads>
        <display width="4096" height="2160"/>
        <mappableVideoSize>0x400000</mappableVideoSize>
        <fbReservation>0xC000000</fbReservation>
        <ftrlConfig>0x2d</ftrlConfig>
        <cudaEnabled>0x0</cudaEnabled>
        <eccSupported>0</eccSupported>
        <multiVgpuSupported>0x0</multiVgpuSupported>
        <encoderCapacity>0x64</encoderCapacity>
        <barlLength>0x100</barlLength>
        <frame_rate_limiter>1</frame_rate_limiter>
        <license>GRID-Virtual-PC,2.0;GRID-Virtual-WS,2.0;Quadro-Virtual-DWS,5.0;GRID-Virtual-WS-
Ext,2.0</license>
    </vgpuType>
    <vgpuType id="224" name="GRID T4-2B4" class="NVS">
        <devId vendorId="0x10de" deviceId="0x1EB8" subsystemVendorId="0x10de" subsystemId="0x130B"/>
        <framebuffer>0x74000000</framebuffer>
        <numHeads>4</numHeads>
        <display width="2560" height="1600"/>
        <mappableVideoSize>0x400000</mappableVideoSize>
        <fbReservation>0xC000000</fbReservation>
        <ftrlConfig>0x2d</ftrlConfig>
        <cudaEnabled>0x0</cudaEnabled>
        <eccSupported>0</eccSupported>
        <multiVgpuSupported>0x0</multiVgpuSupported>
        <encoderCapacity>0x64</encoderCapacity>
        <barlLength>0x100</barlLength>
        <frame_rate_limiter>1</frame_rate_limiter>
        <license>GRID-Virtual-PC,2.0;GRID-Virtual-WS,2.0;Quadro-Virtual-DWS,5.0;GRID-Virtual-WS-
Ext,2.0</license>
    </vgpuType>

```

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>

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>
- 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 vGPU Deployment Guide for VMware Horizon 7.5 on VMware vSphere 6.7
<https://images.nvidia.com/content/pdf/vgpu/guides/vgpu-deployment-guide-horizon-on-vsphere-final.pdf>
- NVIDIA Management and Monitoring
<https://www.nvidia.com/en-us/data-center/virtualization/it-management/>

VMware

- TechZone
<https://techzone.vmware.com/>
- Blast Extreme Display Protocol in VMware Horizon 7
<https://techzone.vmware.com/resource/blast-extreme-display-protocol-vmware-horizon-7>
- Bitfusion FlexDirect Demo
<https://www.youtube.com/watch?v=qWDVzmoaoBo>

Applications

- Autodesk Maya
<https://www.autodesk.com/products/maya/overview>
- Autodesk 3DSMax
<https://www.autodesk.com/products/3ds-max/overview>
- Autodesk Revit
<https://www.autodesk.com/products/revit/overview>
- Autodesk AutoCAD
<https://www.autodesk.com/products/autocad/overview>
- Dassault Systemes SOLIDWORKS
<https://www.solidworks.com/>
- Dassault Systemes CATIA
<https://www.3ds.com/products-services/catia/>

- PTC Creo
<https://www.ptc.com/en/products/cad/creo>
- Siemens NX
<https://www.plm.automation.siemens.com/global/en/products/nx/>
- Adobe Creative Cloud
<https://www.plm.automation.siemens.com/global/en/products/nx/>

SPEC

- SPECviewperf 13 benchmark
<https://www.spec.org/gwpg/gpc.static/vp13info.html>
- The SPEC Consortium
<https://www.spec.org/consortium/>

Version History

Version	Date	Document Version History
Version 1.0	August 2019	Initial document

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

Copyright © 2019 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

Data contained herein pertains to a commercial item (as defined in FAR 2.101) and is proprietary to NetApp, Inc. The U.S. Government has a non-exclusive, non-transferrable, non-sublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b).

Trademark Information

NETAPP, the NETAPP logo, and the marks listed at <http://www.netapp.com/TM> are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.