Hewlett Packard Enterprise

# HPE Reference Architecture for virtual workstations on HPE ProLiant DL380 Gen10 servers

Demonstrating the performance of virtualized GPUs for graphics-intensive use cases

# Contents



## Executive summary

Modern industrial design requires high end computer aided design (CAD) software. CAD software is a resource intensive 3D application that requires graphics processing units (GPUs). At first glance, you instantly think of a dedicated workstation explicitly created for this purpose, but business, regulatory, and security requirements, as well as cost factors, can make the choice of a physical workstation not only unappealing, but untenable. Examples of this case include:

- State-supported industrial espionage is a real and growing threat and local data on portable devices is an obvious target. Many companies have lost cutting edge intellectual property to these entities threatening their very existence.
- Regulations such as the General Data Protection Regulation (GDPR) approved by the European Union Parliament on April 14, 2016 and moved into enforcement on May 25, 2018 changed the way data has to be handled to protect the privacy of its content as well as its integrity.
- For many companies, a worldwide workforce that needs to share large datasets not only within the company, but also across suppliers in real-time, is a functional reality.
- A dedicated remote workstation that is used by only a single user may meet the needs of the most demanding applications but may be a costly solution for downstream consumers of visual data.

In these and other environments, a centralized virtual desktop or application with GPU acceleration can meet the requirements of the designer, the corporation, and data regulations. The HPE ProLiant DL380 Gen10 server solution with either NVIDIA® T4 GPUs or NVIDIA Quadro RTX™ 6000 GPUs, allows designers and consumers of graphics data the ability to experience improved performance with their virtual applications, largely due to the newly designed riser - which can support up to seven single-wide GPUs or two double-wide GPUs.

To meet enterprise requirements for performance and manageability, this Reference Architecture demonstrates the following benefits:

- A graphics enabled environment built on HPE ProLiant DL380 Gen10 servers, equipped with the new riser to support up to seven single-wide GPUs or two double-wide GPUs.
- Test results that can be used to help tailor the environment to meet the demands of high-performance virtual workstations.

**Target audience:** This document is targeted at IT decision makers, architects, and technical personnel seeking to implement virtualized graphics solutions. Experience with VMware® Horizon®, NVIDIA GPU, VMware vSphere®, and high-end CAD applications as well as familiarity with networking is assumed.

**Document purpose:** The purpose of this Reference Architecture is to aid IT departments in understanding the requirements and infrastructure for a virtualized workstation solution capable of running high performance workloads on an HPE ProLiant DL380 Gen10 server accelerated by NVIDIA GPUs and NVIDIA virtual GPU (vGPU) software. This document also details the variety of experiences had, when shared resources are carved into smaller segments, while also helping to create a more secure environment for your intellectual property.

This Reference Architecture describes solution testing performed in August 2019.

## Introduction

Industrial verticals have evolved over the years, driving the need for GPU hardware-based acceleration in end-user computing environments in order to deliver superior performance. Intensive 3D applications in several industrial sectors require the resulting high-resolution graphics in order to provide end users with the appropriate experience. IT shops demand infrastructure and solutions that facilitate this experience while also providing simplified lifecycle management and enhanced security. Failing to meet these requirements can mean the difference between timely and effective deployment of graphics accelerated deliverables and costly slippages and increased support costs.

Given the increasing need for user mobility, data center space constraints, and data security, businesses are continuing to virtualize desktops. To significantly improve their user experience, companies can run their engineering, design and business applications remotely by deploying a GPU-accelerated VDI solution. Moreover, IT organizations can allocate centralized infrastructure resources efficiently and can apply all of the necessary software updates in a shared location.

Hewlett Packard Enterprise is addressing these challenges in an effective way with HPE ProLiant DL380 Gen10 servers and NVIDIA virtual GPUs. The HPE ProLiant DL380 is a secure, resilient server that delivers world-class performance and versatility. Its flexible and forward-looking design keeps up with business needs and helps to maximize ROI. The HPE ProLiant DL380 is an ideal platform for high-performance virtualized workloads. It is equipped with industry-leading security and management services that will ensure a secure and easy deployment in the data

center. With the HPE ProLiant DL380 Gen10 servers' newly designed riser card, up to seven single-wide GPUs or up to two double-wide GPUs can be supported in a server. This provides the opportunity to deploy an expansive range of solutions with HPE ProLiant DL380.

## Solution overview

This Reference Architecture demonstrates best practices to create an immersive, high quality user experience for technical professionals running virtualized workloads. The solution presented in this Reference Architecture includes an HPE ProLiant DL380 Gen10 server used as the system under test (target server). It is equipped with the new riser to support up to seven single-wide or two double-wide GPUs.

The target server is virtualized using VMware ESXi. VMware Horizon is used for implementing the VDI solution with either NVIDIA T4 GPUs or NVIDIA Quadro RTX 6000 GPUs. The Horizon Agent is installed on all the guest VMs on the target server, to deliver them as a desktop, as part of the solution. The Horizon Client is installed on the endpoint devices, which allows users to connect to their virtualized desktops. The client/endpoint devices can be Windows, Apple® Mac, Linux® Zero and thin clients.

A master target VM image was made to ensure easy and consistent deployment of VMs on the target server. Detailed instructions to prepare the master target VM image can be found further in this document.

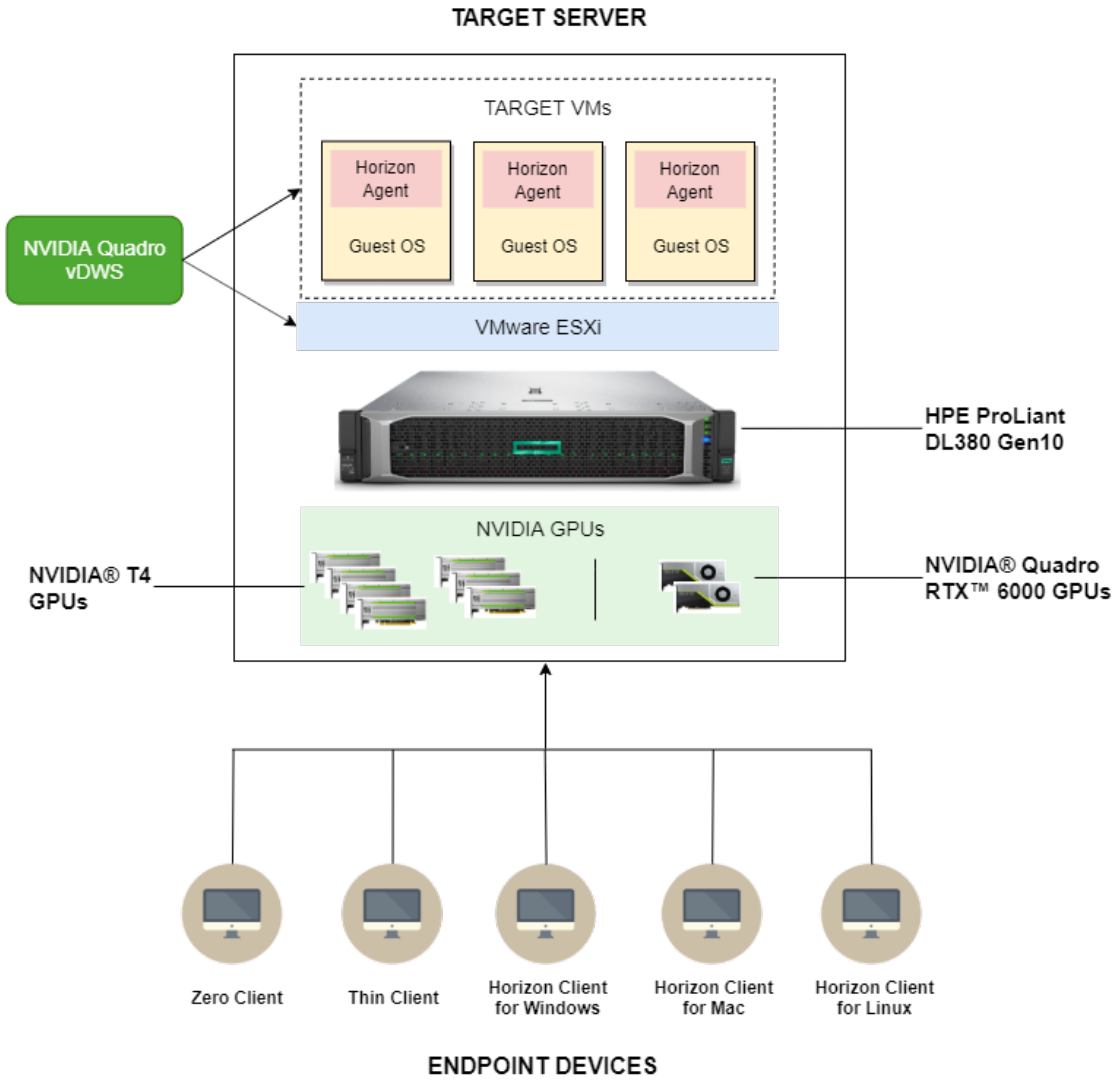


**Figure 1.** Solution setup

For this Reference Architecture, Hewlett Packard Enterprise utilized estimates derived from the SPECviewperf 13 benchmark for Dassault Systèmes CATIA™ workloads in order to analyze the performance and viability of using virtual graphics to support end users with 3D graphics requirements. The HPE ProLiant DL380 Gen10 used in testing for this RA was tested with both two (2) and seven (7) single-wide GPUs configurations and both one (1) and two (2) double-wide GPU configurations.

This Reference Architecture provides perspectives on performance for graphics intensive virtualized workloads using a subset of the SPECviewperf benchmark to generate performance estimates. SPECviewperf, developed by Standard Performance Evaluation Corporation (SPEC), is the worldwide standard for measuring graphics performance based on professional applications. SPECviewperf 13, the latest version of this benchmark, has 3D based viewsets that represent graphics content and behavior from actual applications. For this RA, Hewlett Packard Enterprise used the CATIA viewset to derive the best configuration guidelines to achieve deterministic performance at a given VM density, for end users leveraging GPU accelerated hardware and virtualization technology.

The estimates of the SPECviewperf 13 benchmark for CATIA viewsets are summarized below in Table 1 and will be discussed in greater detail later in this document.

**Table 1.** SPECviewperf 13 estimates for CATIA viewsets workload

| HPE ProLiant DL380 Gen10 | NVIDIA T4 GPUs | Number of Graphics Users | Total Graphics memory | NVIDIA vGPU profile | Estimated Frames per second (FPS) |
|---|---|---|---|---|---|
| 1 | 2 x NVIDIA T4 GPUs | 2 | 32 GB | T4-16Q | 241 |
| 1 | 7 x NVIDIA T4 GPUs | 7 | 112 GB | T4-16Q | 220 |
| 1 | 2 x NVIDIA T4 GPUs | 8 | 32 GB | T4-4Q | 64 |
| 1 | 7 x NVIDIA T4 GPUs | 28 | 112 GB | T4-4Q | 57 |
| 1 | 1 x NVIDIA Quadro RTX 6000 | 1 | 24 GB | RTX6000-24Q | 327 |
| 1 | 2 x NVIDIA Quadro RTX 6000 | 2 | 48 GB | RTX6000-24Q | 330 |
| 1 | 1 x NVIDIA Quadro RTX 6000 | 3 | 24 GB | RTX6000-8Q | 139 |
| 1 | 2 x NVIDIA Quadro RTX 6000 | 6 | 48 GB | RTX6000-8Q | 135 |

## Design principles

In a typical deployment, beside the target server, three more servers are required for management purposes:

- One (1) server for deploying VMware vCenter to manage the virtual infrastructure
- One (1) server for deploying Active Directory and DNS for the infrastructure
- One (1) NVIDIA vGPU License Server to manage NVIDIA licenses

In the testing environment, these servers were run as VMs on a separate management server. HPE ProLiant DL380 Gen10 was used as the management server. It was virtualized with VMware ESXi and the three VMs were deployed on top of it.

For the purpose of testing, clients/endpoint devices were simulated by setting up a separate client server. HPE ProLiant DL380 Gen10 server was used as the client server. It was virtualized with VMware ESXi and multiple VMs were deployed on top of it. Horizon Client was installed on all the VMs on the client server to simulate endpoint devices. Horizon protocol uses H264 encoding and decoding. As a result, the client server requires some GPU processing for H.264 decoding which can be performed by the endpoint devices with built in GPU in a typical deployment scenario. However, the endpoint devices were simulated on an HPE ProLiant DL380 server in the test environment. Two NVIDIA T4 GPUs with T4-1Q profile were enough to perform the decoding for the maximum VM density tested in the environment. The master client VM image was made to ensure easy and consistent deployment of VMs on the client server. Detailed instructions to prepare the master client VM image can be found further in this document.

The target server, management server, and the client server were interconnected using 10GbE networking with an HPE 5900 switch.

# Solution components

## Hardware

### HPE ProLiant DL380 Gen10

The HPE ProLiant DL380 Gen10 platform offers the ultimate flexibility for end-user computing workloads. With a choice of CPUs offering a balance between core counts and core frequencies, very large memory footprints, a broad array of graphics options and a mix of HDD, SSD and NVMe drives, the HPE ProLiant DL380 Gen10 works well for all end-user computing workloads. The HPE ProLiant DL380 Gen10 supports all graphics users from those with simple video needs to workstation class users and does so with support for the various NVIDIA GPUs including the T4 and the Quadro RTX 6000. The HPE ProLiant DL380 Gen10 server supports up to two double-wide or seven single-wide GPUs for workload acceleration.

**Figure 2.** HPE ProLiant DL380 Gen10 server

Table 2 describes the configuration of the HPE ProLiant DL380 Gen10 tested with NVIDIA T4 GPUs for this document.

**Table 2.** HPE ProLiant DL380 Gen10 with NVIDIA T4 GPUs utilized in this Reference Architecture (quantities are per node)

| Hardware | Quantity | Description |
|---|---|---|
| CPU | 2 | Intel® Xeon®-Gold 6254 CPU @3.10GHz, 18-core |
| Memory | 12 | HPE 32GB (1x32GB) Dual Rank x8 DDR4-2666 CAS-19-19-19 Memory Kit |
| GPU | 7 | NVIDIA T4 with NVIDIA Quadro Virtual Data Center Workstation (Quadro vDWS) software |
| Storage Controller | 1 | HPE Smart Array P408i-a SR Gen10 |
| Disks | 6 | HPE 400GB SAS 12G Write Intensive (2.5in) SC 3yr SSD |
| Network Adapter | 1 | HPE Ethernet 10/25Gb 2p 621SFP28 Adapter |

Table 3 describes the configuration of the HPE ProLiant DL380 Gen10 tested with NVIDIA Quadro RTX 6000 GPUs for this document.

**Table 3.** HPE ProLiant DL380 Gen10 with NVIDIA Quadro RTX 6000 GPUs utilized in this Reference Architecture (quantities are per node)

| Hardware | Quantity | Description |
|---|---|---|
| CPU | 2 | Intel Xeon-Gold 6246 CPU @3.3GHz, 12-core |
| Memory | 12 | HPE 32GB (1x32GB) Dual Rank x8 DDR4-2666 CAS-19-19-19 Memory Kit |
| GPU | 2 | NVIDIA Quadro RTX 6000 with NVIDIA Quadro Virtual Data Center Workstation (Quadro vDWS) software |
| Storage Controller | 1 | HPE Smart Array P408i-a SR Gen10 |
| Disks | 6 | HPE 400GB SAS 12G Write Intensive (2.5in) SC 3yr SSD |
| Network Adapter | 1 | HPE Ethernet 10/25Gb 2p 621SFP28 Adapter |

### NVIDIA T4

The NVIDIA T4 GPU is based on the latest NVIDIA Turing architecture which provides support for virtualized workloads with NVIDIA virtual GPU (vGPU) software. It is a single wide card with passive cooling and offers good performance while consuming less power. The NVIDIA Turing architecture includes RT cores for real-time ray tracing acceleration, and batch rendering. It also supports GDDR6 memory which provides improved power efficiency and performance over the previous generation GDDR5 memory. With the help of Turing architecture, T4 is capable of offering the same breakthrough performance and versatility to Virtual Machines (VMs) as achieved on non-virtualized systems. This performance is achieved when T4 is used with NVIDIA vGPU software. Users can achieve a native-PC like experience in a virtualized environment when T4 is combined with NVIDIA vGPU software. The T4 is well suited for various data center workloads including virtual desktops using modern productivity applications, and virtual workstations for creative professionals, engineers and scientists.

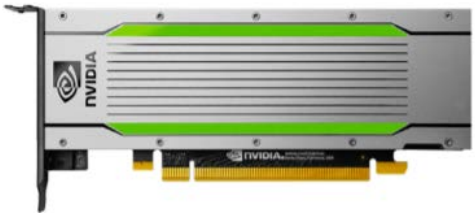


**Figure 3.** NVIDIA T4 GPU

Table 4 describes the specifications of the NVIDIA T4 GPU tested for this document.

**Table 4.** NVIDIA T4 utilized in this Reference Architecture

| NVIDIA T4 | Specifications |
|---|---|
| GPU | 1 NVIDIA Turing GPU |
| CUDA Cores | 2560 |
| Memory Size | 16 GB GDDR6 |
| vGPU Profiles | 1 GB, 2 GB, 4 GB, 8 GB, 16 GB |
| System Interface | x16 PCIe Gen3 |
| Form Factor | Low-Profile PCIe |
| Power | 70 W |
| Thermal Solution | Passive |

### NVIDIA Quadro RTX 6000

NVIDIA Quadro RTX 6000, powered by the NVIDIA Turing™ architecture and the NVIDIA RTX platform, brings the most significant advancement in computer graphics in over a decade to professional workflows. Designers and artists can now wield the power of hardware-accelerated ray tracing, deep learning, and advanced shading to dramatically boost productivity and create amazing content faster than ever before. Equipped with 4608 CUDA cores, 576 Tensor cores, 72 RT Cores and massive 24GB GDDR6 memory, Quadro RTX 6000 can render complex models and scenes with physically accurate shadows, reflections, and refractions to empower users with instant insight. Support for NVIDIA NVLink enables applications to scale memory and performance with multi-GPU configurations.

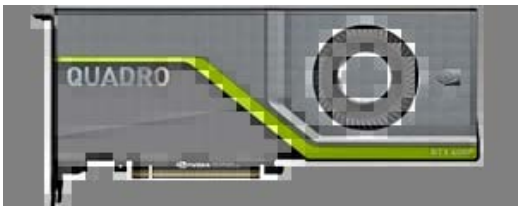


**Figure 4.** NVIDIA Quadro RTX 6000 GPU

Table 5 describes the specifications of the NVIDIA Quadro RTX 6000 GPU tested for this document.

**Table 5.** NVIDIA Quadro RTX 6000 utilized in this Reference Architecture

| NVIDIA Quadro RTX 6000 | Specifications |
|---|---|
| GPU | 1 NVIDIA Turing GPU |
| CUDA Cores | 4608 |
| Memory Size | 24 GB GDDR6 |
| vGPU Profiles | 1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 GB |
| System Interface | x16 PCIe Gen3 |
| Form Factor | Dual Slot, Full Height |
| Power | Total board power: 295 W<br>Total graphics power: 260W |
| Thermal Solution | Active |
| vBIOS version | 90.02.30.00.02 |

## Software

This Reference Architecture uses virtualization infrastructure built on ESXi 6.7 using vCenter Server Appliance (VCSA) 6.7 and VMware Horizon 7. The estimates mentioned in this document were done using the SPECviewperf 13 benchmark.

Table 6 provides the solution software components used in the Reference Architecture.

**Table 6.** Solution software components

| Component | Version |
|---|---|
| **VMware** | |
| • VMware Horizon | 7 |
| • VMware vCenter | 6.7 |
| • VMware ESXi | 6.7 |
| **NVIDIA** | |
| • NVIDIA vGPU Manager | 410.122 (for T4), 430.46 (for Quadro RTX 6000) |
| **SPECviewperf (used for estimating performance)** | |
| • SPECviewperf | 13 |
| **Microsoft®** | |
| • Microsoft Windows® 10 | Enterprise |

### NVIDIA Virtual GPU (vGPU) software

NVIDIA vGPU software is a graphics virtualization platform that provides virtual machines (VMs) access to NVIDIA GPU technology. It enhances the power of a VDI environment to offer a consistent user experience for every virtual workflow. It supports NVIDIA GPUs and is installed in the hypervisor. The software divides the GPU into multiple vGPU instances that each have direct access to the native NVIDIA driver installed in the guest OS. The graphics commands of each virtual machine are passed directly to the GPU, without translation by the hypervisor. This allows the GPU hardware to be allocated for each user to deliver the ultimate experience in shared virtualized graphics performance. With the help of NVIDIA vGPU, multiple NVIDIA GPUs can also be allocated to a single virtual machine to power the most demanding workflows.

### NVIDIA vGPU License Server

NVIDIA vGPU License Server acts as a pool of floating licenses to NVIDIA vGPU software licensed products. It is installed in a network accessed by the customer and is configured with licenses obtained from the NVIDIA Software Licensing Center. It helps in keeping track of the usage of licenses. During guest OS boot, all the licensed vGPU functionalities are activated by acquiring a software license from an NVIDIA vGPU License Server. When the guest OS shuts down, the license is returned to the license server.

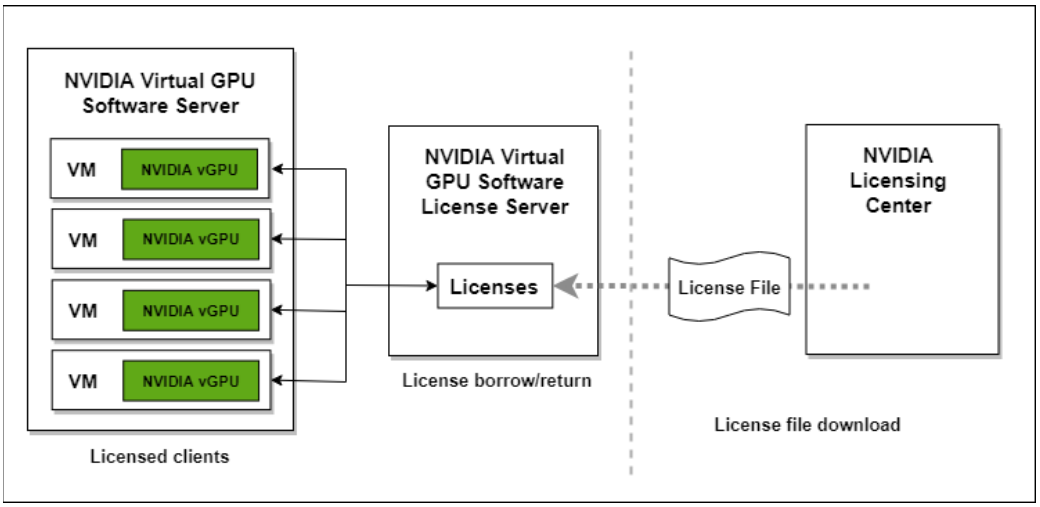


**Figure 5.** NVIDIA vGPU Software Licensing

### VMware Infrastructure

VMware offers a set of software and solutions for virtual desktop infrastructure (VDI). For this reference architecture the following offerings from VMware were used:

### VMware vSphere

VMware vSphere is a suite of software components for virtualization. It uses the power of virtualization to transform data centers into simplified cloud computing infrastructures, enabling IT organizations to deliver flexible and reliable IT services. The two core components of vSphere are VMware ESXi™ and VMware vCenter Server®.

### VMware ESXi

VMware ESXi is the hypervisor on which virtual machines are created and run.

### vCenter Server

vCenter Server is a service that acts as a central administrator for ESXi hosts that are connected on a network. With vCenter Server, we can pool and manage the resources of multiple hosts. vCenter Server allows you to monitor and manage your physical and virtual infrastructure.

### VMware Horizon

VMware Horizon is a centralized desktop virtualization solution used to deliver virtualized desktop services to end users from VMware vSphere servers. The main components of VMware Horizon are Horizon Agent and VMware Horizon Client.

### Horizon Agent

Horizon Agent communicates between Horizon Client and virtual desktops and provides features such as connection monitoring, virtual printing etc.

### VMware Horizon Client

VMware Horizon Client enables the user to access their virtual desktops from a variety of devices such as smartphones, zero clients, thin clients etc. It is installed on every end point. It helps the users to log in to their remote desktops in the data center.

### NVIDIA nVector

NVIDIA's performance engineering teams have developed a methodology and set of benchmarking tools called NVIDIA nVector that simulates, at scale, the end user workflow. It also has the capability to run industry standard benchmarks like SPECViewPerf concurrently on multiple virtual workstations.

NVIDIA nVector aggregates SPECViewPerf scores measured on individual virtual workstations, enabling us to capture benchmark performance at scale. It enables end-to-end automation of performance evaluation of the HPE ProLiant DL380 and provides sizing guidance for high-end graphics virtual workstations. The SPECViewPerf scores, when combined with resource utilization information from the host such as CPU core utilization, virtual machine and network, provide a holistic picture that enables IT to better size their VDI environments for scale, while continuing to ensure a great user experience.

### SPECviewperf 13

The SPECviewperf benchmark is the worldwide standard for measuring graphics performance based on professional applications. The benchmark measures the 3D graphics performance of systems running under the OpenGL and Direct X application programming interfaces. The benchmark's workloads, called viewsets, represent graphics content and behavior from actual applications.

## Configuration guidance for the solution

### VMware vSphere & vCenter server infrastructure

Download the HPE Custom Image for VMware vSphere 6.7 from HPE Support or from VMware Support. The HPE Custom Image has all the necessary management tools for the Agentless Management Service (AMS) and Active Health Service (AHS). It provides drivers, utilities, and tools for device enablement.

VMware vSphere 6.7 should be installed on all HPE ProLiant DL380 Gen10 nodes. SSH is enabled on all the ESXi hosts to allow for further configuration. To manage all the VMware vSphere ESXi hosts, either leverage an existing or create a new vCenter Server in the data center in which the infrastructure resides. All HPE ProLiant nodes associated with this high-performance solution should be housed in a cluster in a VMware data center within the vCenter server.

### NVIDIA vGPU

NVIDIA virtual GPU technology enables multiple virtual machines to access the same physical GPU directly while sharing GPU resources, or multiple physical GPUs can be allocated to a single virtual machine to power the most demanding workloads. These resources are segmented across different virtual machines which can be scaled up or down depending on users needs, which provides an improved TCO. This paper seeks to provide data about performance changes as levels of sharing are increased.

Guest VMs use the NVIDIA vGPUs in a manner similar to accessing a physical GPU that has been passed through by the hypervisor. NVIDIA drivers are loaded in the Guest VMs which enables the guest OS to access the GPU directly. This provides significant fast-path performance. Refer to the NVIDIA vGPU System Architecture shown in Figure 6.

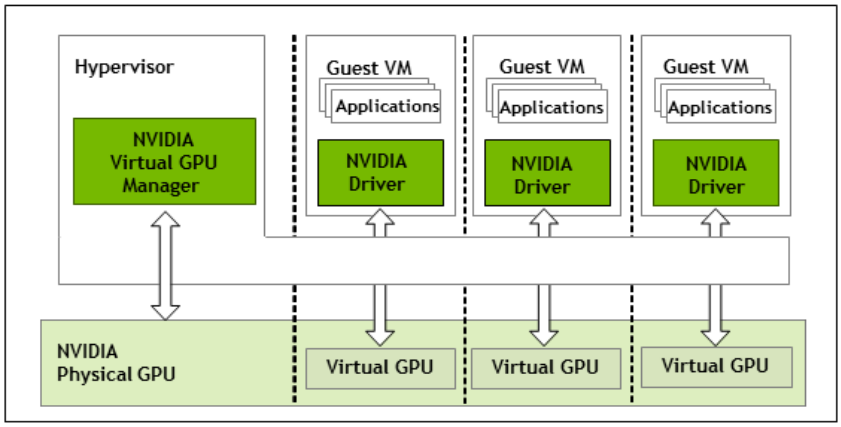


**Figure 6.** NVIDIA vGPU Architecture

## Installing the NVIDIA Virtual GPU Manager Package for VMware vSphere

NVIDIA Virtual GPU Manager is required on each ESXi host. The process of installing the NVIDIA Virtual GPU manager is the same for both NVIDIA T4 and NVIDIA Quadro RTX 6000. To install, follow the steps outlined below:

1. Download the offline VIB bundle and upload it to all the ESXi hosts onto a local vmfs datastore.
2. Ensure that the respective ESXi hosts are in maintenance mode prior to installing the package.
3. Select the ESXi host in the vCenter → right-click on the ESXi host and then select 'Maintenance Mode'.
4. Once the ESXi host goes into maintenance mode, install the NVIDIA vGPU Manager using the following command. Ensure that you fill in the path and VIB in a fashion that reflects your environment.

   `# esxcli software vib install -d /vmfs/volumes/<volume containing the vib>/NVIDIA_ESXi_<vib name>.zip`

5. Once the VIB is successfully installed, reboot the ESXi host and exit the maintenance mode from vCenter.
6. Verify that the NVIDIA kernel driver is installed and loaded in the ESXi host kernel. Search for the NVIDIA kernel module by running the following command:

   `# vmkload_mod -l | grep nvidia`

You should see a line beginning with "nvidia" returned by the command.

To ensure that the NVIDIA kernel driver installed on the ESXi host is communicating to the NVIDIA physical GPU, run the nvidia-smi command to list the GPUs on the HPE ProLiant DL380 Gen10 server. Figure 7 shows the listing of the seven NVIDIA T4 GPUs on a single HPE ProLiant DL380 Gen10 server.

```
[root@target-esxi:~] nvidia-smi
Thu Aug  1 11:22:11 2019
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.122      Driver Version: 410.122      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. |
|===============================+======================+======================|
|   0  Tesla T4            On   | 00000000:12:00.0 Off |                  Off |
| N/A   32C    P8    16W /  70W |  16374MiB / 16383MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla T4            On   | 00000000:13:00.0 Off |                  Off |
| N/A   35C    P8    16W /  70W |  16374MiB / 16383MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla T4            On   | 00000000:37:00.0 Off |                  Off |
| N/A   32C    P8    16W /  70W |  16374MiB / 16383MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla T4            On   | 00000000:86:00.0 Off |                  Off |
| N/A   30C    P8    16W /  70W |  16374MiB / 16383MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   4  Tesla T4            On   | 00000000:AF:00.0 Off |                  Off |
| N/A   29C    P8    16W /  70W |  16374MiB / 16383MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   5  Tesla T4            On   | 00000000:B0:00.0 Off |                  Off |
| N/A   30C    P8    16W /  70W |  16374MiB / 16383MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   6  Tesla T4            On   | 00000000:D8:00.0 Off |                  Off |
| N/A   31C    P8    16W /  70W |  16374MiB / 16383MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
```

**Figure 7.** Verify the NVIDIA package is installed

### Note

For this Reference Architecture, Hewlett Packard Enterprise tested a single HPE ProLiant DL380 Gen10 server scaling from two to seven NVIDIA T4 GPU cards or from one to two NVIDIA Quadro RTX 6000 GPU cards. Different GPU profiles were selected on the virtual machines for scalability and capacity testing to determine the estimated performance as the number of end users accessing the CATIA workload running on the virtual machines increases.

## GPU allocation policy on VMware vSphere on the HPE ProLiant DL380 Gen10 server

For the purpose of testing, the 'Breadth-First' allocation policy was chosen. This choice was based on an analysis on the performance run results. The performance estimates obtained gave consistent performance and minimize sharing of physical GPUs. For the 'Breadth First' allocation policy, ensure that 'Spread VMs across GPUs' is selected. To enable this setting as well as the vGPU on each ESXi host, select 'Shared Direct' followed by 'Spread VMs across GPUs' as shown in Figure 8. When done, select OK.

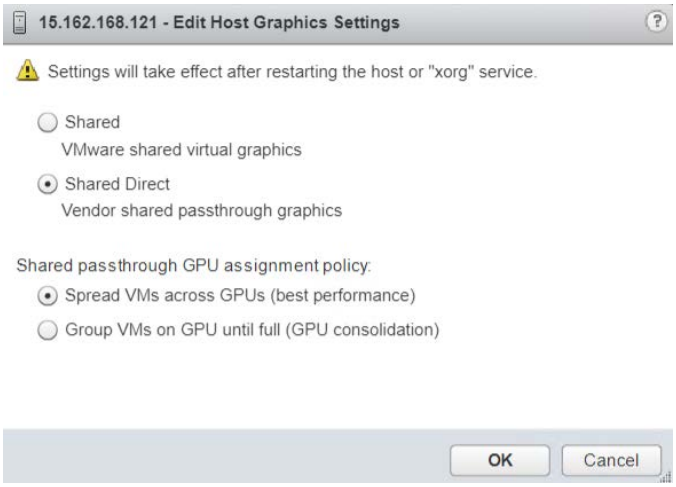


**Figure 8.** Set GPU allocation policy to 'Breadth First' on the HPE ProLiant DL380 Gen10

## Preparing Windows 10 Client VM template/image

A master Windows 10 client template is required. It is used to create more client VMs on the client server. This ensures the same configuration across all client VMs.

The following steps were used to create the master client template/image used in this Reference Architecture:

1. Create a Windows 10 virtual machine with the following specifications:
   - 2 vCPU or more
   - 4 GB vRAM or more
   - 1 VMXNET3 vNIC
   - 40 GB Hard Drive (Thin Provisioned)
   - Windows 10 (1607 LTSB recommended)
   - Reserve all guest memory
   - Shared PCI device set to the appropriate vGPU profile
2. Download the following applications:
   - VMware Horizon client
   - Microsoft Visual C++ 2010 Redistributable Package (x64)
   - NVIDIA vGPU Windows driver corresponding to the VIB version installed in ESXi
3. Once the VM is set up, log in with domain account.
4. Ensure that VMware Tools is installed and up to date.
5. Ensure the power policy in Windows is set to High Performance, and sleep is disabled.

6. Install and configure NVIDIA vGPU drivers to point to the appropriate license server.
7. Install Horizon Client, TightVNC server, Microsoft Visual C++ 2010 Redistributable Package (x64).
8. Change Network settings:
   a. Open inetcpl.cpl
   b. Navigate to the Security tab
   c. Click on the Internet Zone, then click in the Custom level
   d. Find the entry for "Launching applications and unsafe files", and select Enable
   e. Click through and accept the warnings
   f. Repeat for the Trusted Sites Zone
9. Restart the computer.

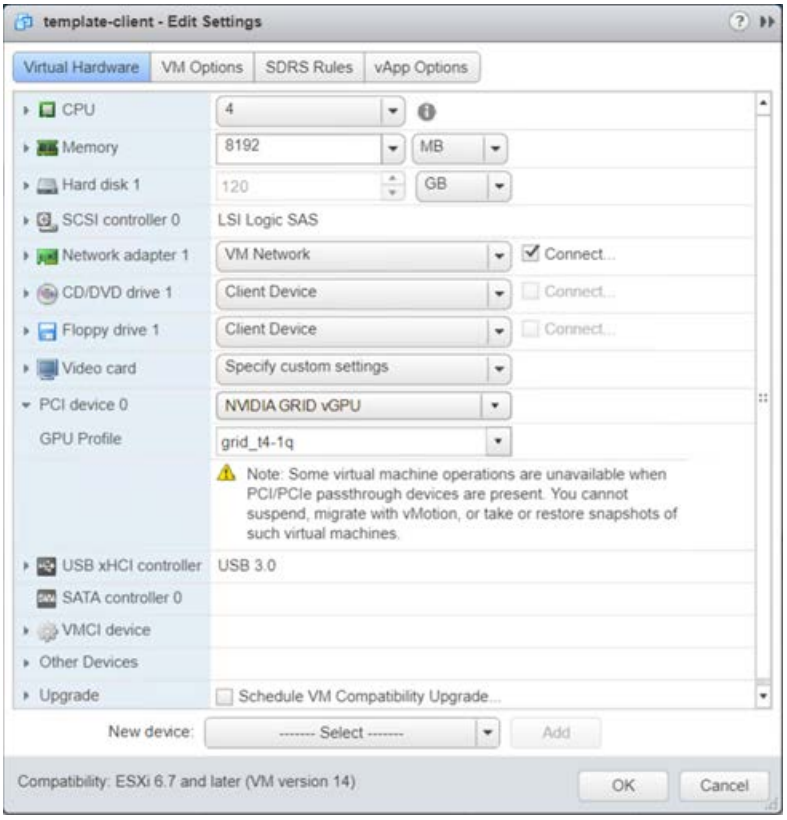


**Figure 9.** Master client VM template configuration

## Preparing Windows 10 Target VM template/image

Following steps were used to create the master target template/image used in this Reference Architecture:

1. Create a Windows 10 virtual machine with the following specifications:
   - 2 vCPUs or more
   - 4 GB vRAM or more
   - 1 VMXNET3 vNIC
   - 40 GB hard drive (thin provisioned)
   - Windows 10 (1607 LTSB recommended)
   - Reserve all guest memory
   - Shared PCI device set to appropriate vGPU profile
2. Download the following applications:
   - VMware Horizon Agent
   - VMware Horizon Direct Connect Agent
   - NVIDIA vGPU Windows driver corresponding to the VIB version installed in ESXi
   - SPECviewperf 13
3. Once the VM is set up, log in with domain account.
4. Ensure that VMware Tools is installed and up to date.
5. Ensure the power policy in Windows is set to High Performance, and sleep is disabled.
6. Install and configure Horizon Agent and Horizon Direct Connect Agent.
7. Install and configure NVIDIA vGPU drivers to point to the appropriate license server.
8. Install Microsoft Office, Google Chrome and SPECviewperf 13.
9. Change Network settings:
   a. Open inetcpl.cpl
   b. Navigate to the Security tab
   c. Click on the Internet Zone, then click in the Custom level
   d. Find the entry for “Launching applications and unsafe files”, and select Enable
   e. Click through and accept the warnings
   f. Repeat for the Trusted Sites Zone

10. Restart the computer.

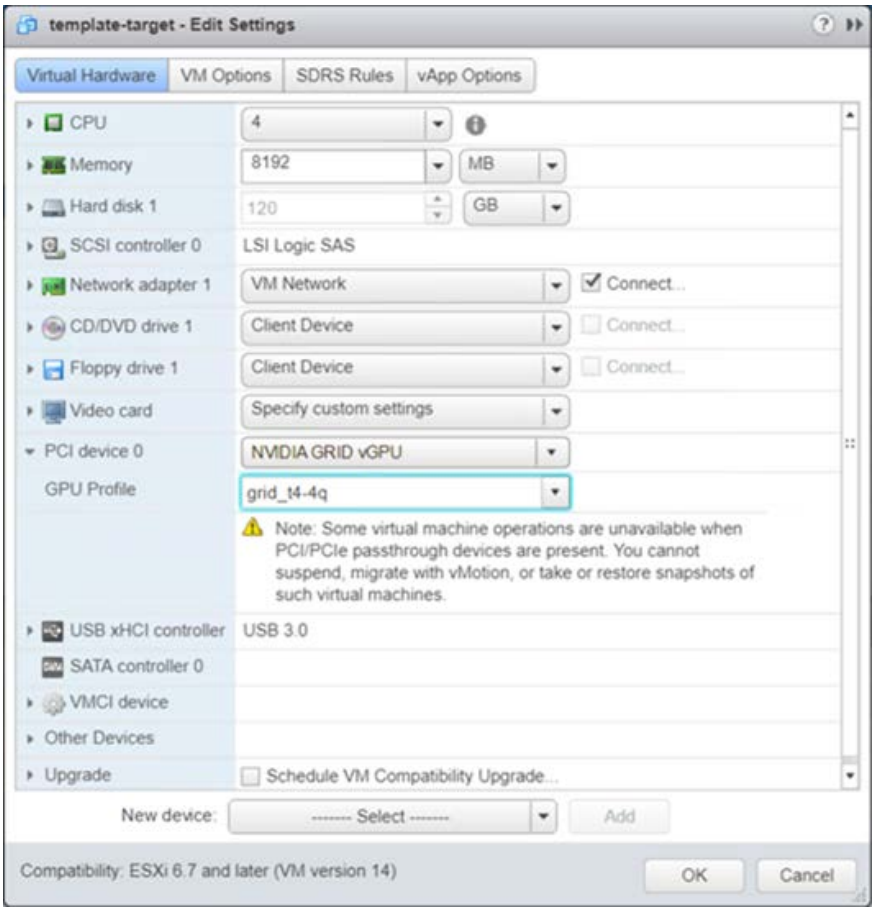


**Figure 10. Master** target VM template configuration

## Capacity and sizing

To publish compliant results that adhere to the SPECviewperf 13 guidelines, the full set of workloads available in the benchmark suite would have needed to be run. As the intent of the document is to provide best practices for implementing, and guidance around the effects of GPU density in the manufacturing space, Hewlett Packard Enterprise chose to only run the SPECviewperf 13 CATIA-05 workload, an application that is commonly used in manufacturing industries. While Hewlett Packard Enterprise ran the workload as it would with any other benchmark, the lack of results for all available workloads prevents these workloads from being publishable as compliant benchmark numbers. As such, the results shared in this paper follow the fair use policy mandated by the SPEC consortium and should be considered as estimates according to the SPEC definition at https://www.spec.org/fairuse.html#DefineEstimate.

The HPE ProLiant DL380 Gen10 target server, with two or seven NVIDIA T4 cards or with one or two NVIDIA Quadro RTX 6000 card(s), was the target for the CATIA workload. Each NVIDIA T4 GPU card has 16 GB memory and 2560 CUDA cores whereas each NVIDIA Quadro RTX 6000 has 24 GB memory and 4608 CUDA cores. This section highlights the estimated frames per second achieved during testing.

### CATIA Workload on HPE ProLiant DL380 Gen10 with two NVIDIA T4 GPU cards

The CATIA workload benchmark performance was measured on multiple VMs ranging from 2 to 16.

Table 7 shows that the estimated frames per second (FPS) achieved is between 32 and 241 FPS.

**Table 7.** Frames per second for different profiles achieved using SPECviewperf 13 with two NVIDIA T4 GPUs

| NVIDIA T4 GPU | Number of graphic users | Graphic Memory per VM | Estimated Frames per second (FPS) | CPU Utilization |
|---|---|---|---|---|
| 2 | 2 | 16 GB | 241 | 11% |
| 2 | 4 | 8 GB | 131 | 20% |
| 2 | 8 | 4 GB | 64 | 40% |
| 2 | 16 | 2 GB | 32 | 65% |

## CATIA Workload on HPE ProLiant DL380 Gen10 with seven NVIDIA T4 GPU cards

The CATIA workload benchmark performance was measured on multiple VMs ranging from 7 to 28. Table 8 shows the estimated frames per second (FPS) achieved is between 57 and 220 FPS.

**Table 8.** Frames per second for different profiles achieved using SPECviewperf 13 with seven NVIDIA T4 GPUs

| NVIDIA T4 GPU | Number of graphic users | Graphic Memory per VM | Estimated Frames per second (FPS) | CPU Utilization |
|---|---|---|---|---|
| 7 | 7 | 16 GB | 220 | 40% |
| 7 | 14 | 8 GB | 124 | 70% |
| 7 | 21 | 4 GB | 70 | 90% |
| 7 | 28 | 4 GB | 57 | 97% |

## CATIA Workload on HPE ProLiant DL380 Gen10 with one NVIDIA Quadro RTX 6000 GPU card

The CATIA workload benchmark performance was measured on multiple VMs ranging from one to three. Table 9 shows that the estimated frames per second (FPS) achieved is between 139 and 327 FPS.

**Table 9.** Frames per second for different profiles achieved using SPECviewperf 13 with one NVIDIA Quadro RTX 6000 GPU

| NVIDIA Quadro RTX 6000 GPU | Number of graphic users | Graphic Memory per VM | Estimated Frames per second (FPS) | CPU Utilization |
|---|---|---|---|---|
| 1 | 1 | 24 GB | 327 | 7% |
| 1 | 2 | 12 GB | 195 | 11% |
| 1 | 3 | 8 GB | 139 | 16% |

## CATIA Workload on HPE ProLiant DL380 Gen10 with two NVIDIA Quadro RTX 6000 GPU card

The CATIA workload benchmark performance was measured on multiple VMs ranging from three to nine. Table 10 shows that the estimated frames per second (FPS) achieved is between 135 and 330 FPS.

**Table 10.** Frames per second for different profiles achieved using SPECviewperf 13 with two NVIDIA Quadro RTX 6000 GPU

| NVIDIA Quadro RTX 6000 GPU | Number of graphic users | Graphic Memory per VM | Estimated Frames per second (FPS) | CPU Utilization |
|---|---|---|---|---|
| 2 | 2 | 24 GB | 330 | 13% |
| 2 | 4 | 12 GB | 192 | 22% |
| 2 | 6 | 8 GB | 135 | 34% |

## Analysis and recommendations

The following is a list of findings and recommendations derived from testing done by Hewlett Packard Enterprise:

- Follow all the settings described in software configuration sections. Ensure the BIOS settings shown below are selected for optimum performance:
  - Workload Profile = Virtualization Max Performance
  - Thermal Configuration = Maximum Cooling
- All the test results shared in this document use "Best Effort" scheduler. The "Best Effort" scheduler allows a vGPU to use GPU processing cycles which are not being used by other vGPUs. When we have only a single VM on the physical GPU, all available CUDA cores are available to that VM. However, the GPU memory is limited by the designated vGPU profile (4 GB for the T4-4Q profile and 8 GB for the RTX6000-8Q profile).
- For the multi-VM test runs, the SPECviewperf13 was launched simultaneously on all the VM using NVIDIA's nVector in order to generate maximum load on the system and analyze system behavior in this stressed condition.
- The FPS achieved is directly related to the number of VMs running on a GPU and the results for the CATIA viewset show an almost linear scaling of performance as we scale the vGPU profiles. For example, for the 2 GPU configuration, a score of 241 FPS on a T4-16Q vGPU profile for a single VM scales down by a factor of 2, as we scale the number of VMs. Similarly, the score of 64 FPS on a T4-4Q vGPU profile for a single VM scales down by a factor of 2 when a second VM is introduced.
- Table 7 shows that even with as little as 2 GB GPU memory, when all the VMs are being simultaneously utilized to their fullest by the test software, the score of 32 FPS on NVIDIA T4 is 33% better than movie quality refresh rate of 24 FPS.
- With 28 VMs on a single HPE ProLiant DL380 Gen10 equipped with seven NVIDIA T4, SPECviewperf 13 reported an average of 57 FPS.
  - This test was run with oversubscribing the host CPU by more than 3 times
  - Total VM memory allocation was less than 60% of the total available system memory capacity
  - Network bandwidth utilization was minimal
  - As the number of VMs increased to 30 or more, bottlenecks with CPU utilization were observed during the testing.
  - Despite the above-mentioned facts, the frames per second (FPS) numbers are close to the most commonly used computer monitors refresh rate of 60 Hz
- As the number of GPUs increased, similar scores were achieved keeping the vGPU profile unchanged. Also, higher CPU utilization was observed due to increased number of VMs running graphics workload simultaneously. For example, 2 VMs on a RTX6000-24Q vGPU profile with 2 NVIDIA Quadro RTX 6000 GPUs obtained a score of 330 FPS while utilizing 13% of CPU resources, whereas 1 VM on same profile with 1 NVIDIA Quadro RTX 6000 GPU obtained a score of 327 FPS while utilizing 7% of CPU resources.
- As shown in Table 8, the T4-4Q vGPU profile with 28 VMs utilized 97% of the CPU. With this high CPU utilization, the testing was restricted to T4-4Q vGPU profile on the seven GPU configuration.
- NVIDIA Quadro RTX 6000 GPUs are best suited for high performance workloads that require 8Q profile or higher. High performance workloads like CATIA can be simultaneously run on up to 6 VMs with RTX6000-8Q profile with an FPS of 135.
- For the purpose of testing, the vBIOS version of NVIDIA Quadro RTX 6000 GPU was upgraded to version 90.02.30.00.02 to restrict the board power of the GPUs.
- The results offer some useful insights into the nature of the graphics workload used (CATIA viewset in this particular case). This workload scales more with GPU compute cycles than with the GPU memory. When assigning vGPU profiles, customers would benefit from characterizing the workloads with some test runs in order to better understand what the dominant factors are influencing performance. This helps drive the ability to choose the right vGPU profiles and scheduler policies for individual workloads.
- While a "Breadth First" GPU allocation policy can offer very good performance under light user load, it should be noted that such performance is more opportunistic. As the hypervisor assigns vGPUs to VMs in a round-robin fashion, we start to see the effects of additional load on performance. Hewlett Packard Enterprise results indicate that the "Breadth First" allocation policy may be best suited to environments where deterministic performance is not a hard requirement and the goal is to maximize application performance under light loads, but the customer

can live with reduced performance at heavy loads. This would be a choice that customers will need to make depending on their workload performance needs.

- The "Breadth First" allocation policy tends to cause higher power draw as it needs to spread out the VMs across all available GPUs. This tends to have a negative effect on energy efficiency. This is because the active GPUs will draw a lot more power than idle GPUs (limited only by the designated TDP of the GPU). In addition to the power drawn by the GPUs, thermal requirements may drive fans at higher speeds which adds to the overall platform power consumption. On the other hand, a "Depth First" policy tries to pack VMs onto the physical GPU until the GPU is fully loaded. For vGPU profiles smaller than the available GPU memory, the load will be contained to just those physical GPUs that have a running workload. The resulting energy efficiency can offer improved cost savings, and thereby drive TCO benefits. This would be an important consideration for customers as they strive to balance performance demands with power budgets and energy efficiency requirements.

## Summary

The HPE ProLiant DL380 Gen10 equipped with the newly designed riser card (with support for seven single-wide GPUs or two double-wide GPUs) and combined with either NVIDIA T4 GPUs or NVIDIA Quadro RTX 6000 GPUs, has been tested with SPECviewperf 13 for high performance virtualized workloads, with the goal of helping IT departments understand what changes in user density mean for designers, scientists, engineers, CAD users etc. The key takeaways include:

- The newly designed riser card for HPE ProLiant DL380 Gen10 can support up to seven single-wide GPUs or up to two double-wide GPUs without compromising performance for high performance virtualized workloads.
- This solution describes the implementation of a graphic based workload on HPE ProLiant DL380 Gen10 leveraging either NVIDIA T4 GPUs or NVIDIA Quadro RTX 6000 GPUs and VMware Horizon capabilities for end users and high-performance applications.
- Performance testing performed using CATIA viewsets with SPECviewperf 13 delivered impressive frames per second (FPS) for every graphic user.

## Appendix A: Bill of materials

The following table shows the bill of materials (BOM) for the target server in this solution.

**Note**

Part numbers are at time of publication/testing and subject to change. The bill of materials does not include complete support options or other rack and power requirements. If you have questions regarding ordering, please consult with your HPE Reseller or HPE Sales Representative for more details. hpe.com/us/en/services/consulting.html

**Table 11.** Bill of materials

| Quantity | Part number | Description |
|---|---|---|
| 1 | 868703-B21 | HPE ProLiant DL380 Gen10 8SFF Configure-to-order Server |
| 12 | 815100-B21 | HPE 32GB (1x32GB) Dual Rank x4 DDR4-2666 CAS-19-19-19Memory Kit |
| 6 | P09098-B21 | HPE 400GB SAS 12G Write Intensive (2.5in) SC 3yr SSD |
| 1 | 804331-B21 | HPE Smart Array P408i-a SR Gen10 (8 Internal Lanes/2GB Cache) 12G SAS Modular Controller |
| 1 | 867328-B21 | HPE Ethernet 10/25Gb 2p 621SFP28 Adapter |
| 1 | P01366-B21 | HPE 96W Smart Storage Battery (up to 20 Devices) with 145mm Cable Kit |
| 1 | 867810-B21 | HPE DL38X Gen10 High Performance Temperature Fan Kit |
| 2 | 826706-B21 | HPE DL380 Gen10 High Performance Heatsink |
| 1 per user | Q2D79A | NVIDIA Quadro Virtual Data Center Workstation (Quadro vDWS) License |

| Quantity | Part number | Description |
|---|---|---|
| | | **Components specific to 2 NVIDIA T4 GPU configuration** |
| 2 | P02517-L21 | HPE DL380 Gen10 Intel Xeon-Gold 6254 (3.1GHz/18-core/200W) FIO Processor Kit |
| 2 | R0W29A | HPE NVIDIA T4 16GB Computational Accelerator |
| 1 | 826694-B21 | HPE DL38X Gen10 x16/x16 Slot 1/2 Riser |
| 1 | 871674-B21 | HPE DL38X Gen10 x16/x16 Slot 1/2 Riser FIO |
| | | **Components specific to 7 NVIDIA T4 GPU configuration** |
| 2 | P02517-L21 | HPE DL380 Gen10 Intel Xeon-Gold 6254 (3.1GHz/18-core/200W) FIO Processor Kit |
| 7 | R0W29A | HPE NVIDIA T4 16GB Computational Accelerator |
| 1 | P14373-B21 | HPE DL38X Gen10 3x16 Sec GPU FIO Kit |
| 1 | P14374-B21 | HPE DL38X Gen10 3x16 Prim GPU FIO Kit |
| 1 | 826700-B21 | HPE DL38X Gen10 x16 Tertiary Riser |
| | | **Components specific to 1 NVIDIA Quadro RTX 6000 GPU configuration** |
| 2 | P15758-L21 | HPE DL380 Gen10 Intel Xeon-Gold 6246 (3.3GHz/12-core/165W) FIO Processor Kit |
| 1 | R0Z45A | HPE NVIDIA Quadro RTX 6000 Graphics Accelerator |
| 1 | 826694-B21 | HPE DL38X Gen10 x16/x16 Slot 1/2 Riser |
| | | **Components specific to 2 NVIDIA Quadro RTX 6000 GPU configuration** |
| 2 | P15758-L21 | HPE DL380 Gen10 Intel Xeon-Gold 6246 (3.3GHz/12-core/165W) FIO Processor Kit |
| 2 | R0Z45A | HPE NVIDIA Quadro RTX 6000 Graphics Accelerator |
| 1 | 826694-B21 | HPE DL38X Gen10 x16/x16 Slot 1/2 Riser |
| 1 | 871674-B21 | HPE DL38X Gen10 x16/x16 Slot 1/2 Riser FIO |
| 1 | 826700-B21 | HPE DL38X Gen10 x16 Tertiary Riser |

## Resources and additional links

HPE Reference Architectures, https://www.hpe.com/info/ra

HPE ProLiant servers, https://www.hpe.com/servers/proliant

NVIDIA, https://www.nvidia.com/en-us/design-visualization/graphics-cards-for-virtualization

NVIDIA drivers, https://www.nvidia.com/Download/index.aspx

VMware Horizon, https://www.vmware.com/in/products/horizon.html

To help us improve our documents, please provide feedback at https://hpe.com/contact/feedback/



a00049201enw, version 2, October 2019