Virtual GPU Software R550 for VMware vSphere

Release Notes

RN-07347-001 _v17.0 through 17.3 | July 2024

Table of Contents

Chapter 1. Release Notes

These *Release Notes* summarize current status, information on validated platforms, and known issues with NVIDIA vGPU software and associated hardware on VMware vSphere.

Note: The most current version of the documentation for this release of NVIDIA vGPU software can be found online at NVIDIA Virtual GPU Software Documentation.

1.1. NVIDIA vGPU Software Driver Versions

Each release in this release family of NVIDIA vGPU software includes a specific version of the NVIDIA Virtual GPU Manager, NVIDIA Windows driver, and NVIDIA Linux driver.

NVIDIA vGPU Software Version	NVIDIA Virtual GPU Manager Version	NVIDIA Windows Driver Version	NVIDIA Linux Driver Version
17.3	550.90.05	552.74	550.90.07
17.2	550.90.05	552.55	550.90.07
17.1	550.54.16	551.78	550.54.15
17.0	550.54.10	551.61	550.54.14

For details of which VMware vSphere releases are supported, see Hypervisor Software Releases.

1.2. Compatibility Requirements for the NVIDIA vGPU Manager and Guest VM Driver

The releases of the NVIDIA vGPU Manager and guest VM drivers that you install must be compatible. If you install an incompatible guest VM driver release for the release of the vGPU Manager that you are using, the NVIDIA vGPU fails to load.

See VM running an incompatible NVIDIA vGPU guest driver fails to initialize vGPU when booted.

Note: You must use NVIDIA License System with every release in this release family of NVIDIA vGPU software. All releases in this release family of NVIDIA vGPU software are **incompatible** with all releases of the NVIDIA vGPU software license server.

Compatible NVIDIA vGPU Manager and Guest VM Driver Releases

The following combinations of NVIDIA vGPU Manager and guest VM driver releases are compatible with each other.

- NVIDIA vGPU Manager with guest VM drivers from the same release
- NVIDIA vGPU Manager with guest VM drivers from different releases within the same major release branch
- NVIDIA vGPU Manager from a later major release branch with guest VM drivers from the previous branch

Note:

When NVIDIA vGPU Manager is used with guest VM drivers from a different release within the same branch or from the previous branch, the combination supports **only** the features, hardware, and software (including guest OSes) that are supported on both releases.

For example, if vGPU Manager from release 17.3 is used with guest drivers from release 16.4, the combination does **not** support Windows Server 2019 because NVIDIA vGPU software release 17.3 does not support Windows Server 2019.

The following table lists the specific software releases that are compatible with the components in the NVIDIA vGPU software 17 major release branch.

NVIDIA vGPU Software Component	Releases	Compatible Software Releases
NVIDIA vGPU Manager	17.0 through 17.3	▶ Guest VM driver releases 17.0 through 17.3 ▶ All guest VM driver 16.*x* releases
Guest VM drivers	17.0 through 17.3	NVIDIA vGPU Manager releases 17.0 through 17.3

Incompatible NVIDIA vGPU Manager and Guest VM Driver Releases

The following combinations of NVIDIA vGPU Manager and guest VM driver releases are incompatible with each other.

- NVIDIA vGPU Manager from a later major release branch with guest VM drivers from a production branch two or more major releases before the release of the vGPU Manager
- NVIDIA vGPU Manager from an earlier major release branch with guest VM drivers from a later branch

The following table lists the specific software releases that are incompatible with the components in the NVIDIA vGPU software 17 major release branch.

NVIDIA vGPU Software Component	Releases	Incompatible Software Releases
NVIDIA vGPU Manager	17.0 through 17.3	All guest VM driver releases 15.*x* and earlier
Guest VM drivers	17.0 through 17.3	All NVIDIA vGPU Manager releases 16.*x* and earlier

1.3. Updates in Release 17.3

New Features in Release 17.3

- Security updates - see *Security Bulletin: NVIDIA GPU Display Driver - July 2024*, which is posted shortly after the release date of this software and is listed on the NVIDIA Product Security page

Feature Support Withdrawn in Release 17.3

- Guest OSes no longer supported:
  - Red Hat Enterprise Linux 7.9

1.4. Updates in Release 17.2

New Features in Release 17.2

- Support for a mixture of different types of time-sliced vGPUs on the same physical GPU starting with vSphere 8 Update 3
- Security updates - see *Security Bulletin: NVIDIA GPU Display Driver - June 2024*, which is posted shortly after the release date of this software and is listed on the NVIDIA Product Security page
- Miscellaneous bug fixes

Hardware and Software Support Introduced in Release 17.2

- Newly supported graphics cards:
  - NVIDIA L20 liquid cooled
- Newly supported guest OS releases:
  - Debian 12
  - Red Hat Enterprise Linux 9.4 and 8.10

Feature Support Withdrawn in Release 17.2

- Guest OSes no longer supported:
  - Debian 10
  - Red Hat Enterprise Linux 9.3, 9.0, 8.9, and 8.6

1.5. Updates in Release 17.1

New Features in Release 17.1

- Resolution of an issue that affects graphics cards that are supported only by NVIDIA AI Enterprise.

1.6. Updates in Release 17.0

New Features in Release 17.0

- Distribution of a release of the Virtual GPU Manager and the NVIDIA vGPU software graphics driver for Linux that is based on NVIDIA Linux open GPU kernel modules

- Support in the NVML API and the `nvidia-smi` command for getting information about vGPUs with different amounts of frame buffer on the same physical GPU
- An option to configure guest VMs or physical hosts to acquire NVIDIA vGPU software licenses when a user logs in instead of at boot time
- Miscellaneous bug fixes

Newly Supported Hardware and Software in Release 17.0

- Newly supported graphics cards:
  - NVIDIA RTX 5880 Ada
- Newly supported guest OSes:
  - Microsoft Windows 11 23H2
  - Red Hat Enterprise Linux 9.3
- Newly supported remoting solutions:
  - VMware Horizon 2312 (8.12)

Feature Support Withdrawn in Release 17.0

- Hypervisor software features no longer supported:
  - VMware vSGA
- Graphics cards no longer supported:
  - Tesla M6
  - Tesla M60
  - Tesla P4
  - Tesla P6
  - Tesla P40
  - Tesla P100 PCIe 12 GB
  - Tesla P100 PCIe 16 GB
  - Tesla P100 SXM2 16 GB
- Guest OSes no longer supported:
  - Windows Server 2019

Chapter 2. Validated Platforms

This release family of NVIDIA vGPU software provides support for several NVIDIA GPUs on validated server hardware platforms, VMware vSphere hypervisor software versions, and guest operating systems. It also supports the version of NVIDIA CUDA Toolkit that is compatible with R550 drivers.

2.1. Supported NVIDIA GPUs and Validated Server Platforms

This release of NVIDIA vGPU software on VMware vSphere provides support for several NVIDIA GPUs running on validated server hardware platforms.

For a list of validated server platforms, refer to NVIDIA GRID Certified Servers.

The supported products for each type of NVIDIA vGPU software deployment depend on the GPU.

Note: NVIDIA vGPU software does **not** support vSGA.

GPUs Based on the NVIDIA Ada Lovelace Architecture

GPU	SR-IOV	Mixed vGPU Configuration		Supported NVIDIA vGPU Software Products[1,2,3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
NVIDIA L40S	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise,	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps

GPU	SR-IOV	Mixed vGPU Configuration		Supported NVIDIA vGPU Software Products[1, 2, 3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
		later update releases **17.0, 17.1 only:** N/A			
NVIDIA L40	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases **17.0, 17.1 only:** N/A	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
NVIDIA L20	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases **17.0, 17.1 only:** N/A	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
NVIDIA L20 liquid cooled	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps

GPU	SR-IOV	Mixed vGPU Configuration		Supported NVIDIA vGPU Software Products[1, 2, 3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
		17.0, 17.1 only: N/A			
NVIDIA L4	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases **17.0, 17.1 only:** N/A	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
NVIDIA L2	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases **17.0, 17.1 only:** N/A	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
NVIDIA RTX 6000 Ada	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps

GPU	SR-IOV	Mixed vGPU Configuration		Supported NVIDIA vGPU Software Products[1, 2, 3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
		17.0, 17.1 only: N/A			
NVIDIA RTX 5880 Ada	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases **17.0, 17.1 only:** N/A	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
NVIDIA RTX 5000 Ada	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases **17.0, 17.1 only:** N/A	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps

GPUs Based on the NVIDIA Ampere Architecture

GPU	SR-IOV	Mixed vGPU Configuration		Supported NVIDIA vGPU Software Products[1,2,3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
NVIDIA A40[4]	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases **17.0, 17.1 only:** N/A	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
NVIDIA A16	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases **17.0, 17.1 only:** N/A	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
NVIDIA A10	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps

GPU	SR-IOV	Mixed vGPU Configuration		Supported NVIDIA vGPU Software Products[1,2,3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
		17.0, 17.1 only: N/A			
NVIDIA A2	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases **17.0, 17.1 only:** N/A	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
NVIDIA RTX A6000[4]	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases **17.0, 17.1 only:** N/A	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
NVIDIA RTX A5500[4]	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps

GPU	SR-IOV	Mixed vGPU Configuration		Supported NVIDIA vGPU Software Products[1,2,3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
		17.0, 17.1 only: N/A			
NVIDIA RTX A5000[4]	8.0, 7.0	**Since 17.2:** 8.0 Update 3 and, unless explicitly stated otherwise, later update releases **17.0, 17.1 only:** N/A	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps

GPUs Based on the NVIDIA Turing Architecture

Note: SR-IOV and the configuration of vGPUs with different amounts of frame buffer on the same physical GPU (mixed-size mode) are **not** supported on GPUs based on the NVIDIA Turing™ architecture.

GPU	Mixed vGPU Series Configuration	Supported NVIDIA vGPU Software Products[1,2,3]	
		NVIDIA vGPU	GPU Pass Through
Tesla T4	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
Quadro RTX 6000 [4]	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
Quadro RTX 6000 passive[4]	8.0	▶ vWS ▶ vPC	▶ vWS ▶ vApps

GPU	Mixed vGPU Series Configuration	Supported NVIDIA vGPU Software Products[1,2,3]	
		NVIDIA vGPU	GPU Pass Through
		▶ vApps	
Quadro RTX 8000[4]	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
Quadro RTX 8000 passive[4]	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps

GPUs Based on the NVIDIA Volta Architecture

Note: SR-IOV and the configuration of vGPUs with different amounts of frame buffer on the same physical GPU (mixed-size mode) are **not** supported on GPUs based on the NVIDIA Volta architecture.

GPU	Mixed vGPU Series Configuration	Supported NVIDIA vGPU Software Products[1,2,3]	
		NVIDIA vGPU	GPU Pass Through
Tesla V100 SXM2	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
Tesla V100 SXM2 32GB	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
Tesla V100 PCIe	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
Tesla V100 PCIe 32GB	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps

GPU	Mixed vGPU Series Configuration	Supported NVIDIA vGPU Software Products[1,2,3]	
		NVIDIA vGPU	GPU Pass Through
Tesla V100S PCIe 32GB	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps
Tesla V100 FHHL	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps

GPUs Based on the NVIDIA Maxwell Graphic Architecture

Note: SR-IOV and the configuration of vGPUs with different amounts of frame buffer on the same physical GPU (mixed-size mode) are **not** supported on GPUs based on the NVIDIA Maxwell™ graphic architecture.

GPU	Mixed vGPU Series Configuration	Supported NVIDIA vGPU Software Products[1,2,3]	
		NVIDIA vGPU	GPU Pass Through
Tesla M10	8.0	▶ vWS ▶ vPC ▶ vApps	▶ vWS ▶ vApps

[1] The supported products are as follows:

- vWS: NVIDIA RTX Virtual Workstation
- vPC: NVIDIA Virtual PC
- vApps: NVIDIA Virtual Applications

[2] N/A indicates that the deployment is not supported.
[3] vApps is supported only on Windows operating systems.
[4] This GPU is supported only in displayless mode. In displayless mode, local physical display connectors are disabled.

2.1.1. Support for a Mixture of Time-Sliced vGPU Types on the Same GPU

VMware vSphere Hypervisor (ESXi) support for a mixture of time-sliced vGPU types on the same GPU depends on the VMware vSphere Hypervisor (ESXi) release.

VMware vSphere Hypervisor (ESXi) Release	Mixture of Time-Sliced vGPU Support
Since 17.2: 8 Update 3 and, unless explicitly stated otherwise, later update releases	VMware vSphere Hypervisor (ESXi) supports a mixture of different types of time-sliced vGPUs on the same physical GPU. Any combination of A-series, B-series, and Q-series vGPUs with any amount of frame buffer can reside on the same physical GPU simultaneously. The total amount of frame buffer allocated to the vGPUs on a physical GPU must not exceed the amount of frame buffer that the physical GPU has. For example, the following combinations of vGPUs can reside on the same physical GPU simultaneously: ▶ A40-2B and A40-2Q ▶ A40-2Q and A40-4Q ▶ A40-2B and A40-4Q By default, a GPU supports only vGPUs with the same amount of frame buffer and, therefore, is in equal-size mode. To support vGPUs with different amounts of frame buffer, the GPU must be put into mixed-size mode. When a GPU is in mixed-size mode, the maximum number of some types of vGPU allowed on a GPU is less than when the GPU is in equal-size mode. For more information, refer to *Virtual GPU Software User Guide*.
8.0 through 8 Update 2	VMware vSphere Hypervisor (ESXi) supports a mixture of time-sliced vGPUs with the same amount of frame buffer from different virtual GPU series on the same physical GPU. A-series, B-series, and Q-series vGPUs with the same amount of frame buffer, for example, A40-2B and A40-2Q, can reside on the same physical GPU simultaneously. However, vGPUs with different amounts of frame buffer are not supported on the same GPU. For example, the following combinations of vGPUs are **not** supported on the same GPU: ▶ A40-2Q and A40-4Q ▶ A40-2B and A40-4Q
7.0	VMware vSphere Hypervisor (ESXi) does not support a mixture of different types of time-sliced vGPUs on the same GPU. All vGPUs on a single GPU must be of the same type: They must belong to the same vGPU series and be allocated the same amount of frame buffer.

VMware vSphere Hypervisor (ESXi) Release	Mixture of Time-Sliced vGPU Support
	The requirement for all vGPUs to be of the same type doesn't extend across physical GPUs on the same card. Different physical GPUs on the same card may host different types of virtual GPUs at the same time, provided that the vGPUs on any one physical GPU are all of the same type. For example, an NVIDIA A16 card has four physical GPUs, and can support several types of virtual GPU. ▶ A configuration with a mixture of A16-4A vGPUs and A16-4Q vGPUs on GPU0 is valid. ▶ A configuration with A16-16Q vGPUs on GPU 0 and GPU 1, A16-8Q vGPUs on GPU 2, and A16-4Q vGPUs on GPU3 is valid. ▶ A configuration with a mixture of A16-8Q vGPUs and A16-4Q vGPUs on GPU0 is invalid.

2.1.2. Switching the Mode of a GPU that Supports Multiple Display Modes

Some GPUs support display-off and display-enabled modes but must be used in NVIDIA vGPU software deployments in display-off mode.

The GPUs listed in the following table support multiple display modes. As shown in the table, some GPUs are supplied from the factory in display-off mode, but other GPUs are supplied in a display-enabled mode.

GPU	Mode as Supplied from the Factory
NVIDIA A40	Display-off
NVIDIA L40	Display-off
NVIDIA L40S	Display-off
NVIDIA L20	Display-off
NVIDIA L20 liquid cooled	Display-off
NVIDIA RTX 5000 Ada	Display enabled
NVIDIA RTX 6000 Ada	Display enabled
NVIDIA RTX A5000	Display enabled
NVIDIA RTX A5500	Display enabled
NVIDIA RTX A6000	Display enabled

A GPU that is supplied from the factory in display-off mode, such as the NVIDIA A40 GPU, might be in a display-enabled mode if its mode has previously been changed.

To change the mode of a GPU that supports multiple display modes, use the `displaymodeselector` tool, which you can request from the NVIDIA Display Mode Selector Tool page on the NVIDIA Developer website.

Note: Only the GPUs listed in the table support the `displaymodeselector` tool. Other GPUs that support NVIDIA vGPU software do not support the `displaymodeselector` tool and, unless otherwise stated, do not require display mode switching.

2.1.3. Requirements for Using vGPU on GPUs Requiring 64 GB or More of MMIO Space with Large-Memory VMs

Some GPUs require 64 GB or more of MMIO space. When a vGPU on a GPU that requires 64 GB or more of MMIO space is assigned to a VM with 32 GB or more of memory on ESXi , the VM's MMIO space must be increased to the amount of MMIO space that the GPU requires.

For more information, refer to VMware Knowledge Base Article: VMware vSphere VMDirectPath I/O: Requirements for Platforms and Devices (2142307).

No extra configuration is needed.

The following table lists the GPUs that require 64 GB or more of MMIO space and the amount of MMIO space that each GPU requires.

GPU	MMIO Space Required
NVIDIA A10	64 GB
NVIDIA A40	128 GB
NVIDIA RTX A5000	64 GB
NVIDIA RTX A5500	64 GB
NVIDIA RTX A6000	128 GB
Quadro RTX 6000 Passive	64 GB
Quadro RTX 8000 Passive	64 GB
Tesla V100 (all variants)	64 GB

2.1.4. Requirements for Using GPUs Requiring Large MMIO Space in Pass-Through Mode

- The following GPUs require 32 GB of MMIO space in pass-through mode:
  - Tesla V100 (all 16GB variants)
- The following GPUs require 64 GB of MMIO space in pass-through mode.
  - Quadro RTX 8000 passive
  - Quadro RTX 6000 passive

- Tesla V100 (all 32GB variants)

If a GPU that requires more than 32 GB of MMIO space is assigned to a VM, the VM's MMIO space must be increased as explained in VMware Knowledge Base Article: VMware vSphere VMDirectPath I/O: Requirements for Platforms and Devices (2142307).

- Pass through of GPUs with large BAR memory settings has some restrictions on VMware ESXi:
  - The guest OS must be a 64-bit OS.
  - 64-bit MMIO must be enabled for the VM.
  - If the total BAR1 memory exceeds 256 Mbytes, EFI boot must be enabled for the VM.

    Note: To determine the total BAR1 memory, run **`nvidia-smi -q`** on the host.

  - The guest OS must be able to be installed in EFI boot mode.

2.1.5. Requirements for Assigning Multiple GPUs in Pass-Through Mode to a Single VM

If you are assigning multiple GPUs in pass-through mode to a single VM, ensure that you allocate enough MMIO space to the VM for all the GPUs.

1. Calculate the amount of MMIO space that is required for all the GPUs that you want to assign in pass-through mode to the VM.
   a). On the hypervisor host, get the total BAR1 memory usage for each GPU.

```
nvidia-smi -q
==============NVSMI LOG==============

Timestamp                                 : Mon Jul 15 18:36:45 2024
Driver Version                            : 550.90.05
CUDA Version                              :  12.4

Attached GPUs                             : 4
GPU 00000000:01:00.0
...
   BAR1 Memory Usage
        Total                             : 128 GiB
...
```

   In this example, the total BAR1 memory usage for each GPU is 128 GiB.
   b). Multiply the total BAR1 memory usage for each GPU by the number of GPUs that you are assigning in pass-through mode to the VM.
   For example, if you are assigning four GPUs to a VM, the amount of MMIO space that is required for all the GPUs is 4#128 GiB, which equals 512 GiB.
2. Under the VM settings, choose **VM Options** > **Advanced** and set `pciPassthru.use64bitMMIO="TRUE"`.
3. Allocate the required amount of MMIO space to the VM.

```
pciPassthru.64bitMMIOSizeGB = "mmio-space-in-gb"
```

mmio-space-in-gb
The required amount of MMIO space in GiB that you calculated previously. For example, if you are assigning four GPUs to a VM that each use a total of 128 GiB of BAR1 memory, the amount of MMIO space that is required for all the GPUs is 512 GiB.

```
pciPassthru.64bitMMIOSizeGB = "512"
```

2.1.6. Linux Only: Error Messages for Misconfigured GPUs Requiring Large MMIO Space

In a Linux VM, if the requirements for using C-Series vCS vGPUs or GPUs requiring large MMIO space in pass-through mode are not met, the following error messages are written to the VM's `dmesg` log during installation of the NVIDIA vGPU software graphics driver:

```
NVRM: BAR1 is 0M @ 0x0 (PCI:0000:02:02.0)
[   90.823015] NVRM: The system BIOS may have misconfigured your GPU.
[   90.823019] nvidia: probe of 0000:02:02.0 failed with error -1
[   90.823031] NVRM: The NVIDIA probe routine failed for 1 device(s).
```

2.2. Hypervisor Software Releases

Supported VMware vSphere Hypervisor (ESXi) Releases

This release is supported on the VMware vSphere Hypervisor (ESXi) releases listed in the table.

Note:

Support for NVIDIA vGPU software requires the vSphere Foundation edition of VMware vSphere Hypervisor (ESXi) or a vSphere Enterprise Plus license. For details, see VMware vSphere Edition Comparison (PDF).

Updates to a base release of VMware vSphere Hypervisor (ESXi) are compatible with the base release and can also be used with this version of NVIDIA vGPU software unless expressly stated otherwise.

Software	Release Supported	Notes
VMware vSphere Hypervisor (ESXi) 8.0	8.0 and later updates to release 8.0 unless explicitly stated otherwise	This release supports all NVIDIA GPUs with vGPU and in pass-through mode that support NVIDIA vGPU software on VMware vSphere.

Software	Release Supported	Notes
VMware vSphere Hypervisor (ESXi) 7.0	7.0 Update 2 and later updates to release 7.0 unless explicitly stated otherwise **Note:** The base VMware vSphere Hypervisor (ESXi) 7.0 release and 7.0 Update 1 are **not** supported.	This release supports all NVIDIA GPUs with vGPU and in pass-through mode that support NVIDIA vGPU software on VMware vSphere.

Supported Management Software and Virtual Desktop Software Releases

This release supports the management software and virtual desktop software releases listed in the table.

Note: Updates to a base release of VMware Horizon and VMware vCenter Server are compatible with the base release and can also be used with this version of NVIDIA vGPU software unless expressly stated otherwise.

Software	Releases Supported
VMware Horizon	**Note:** All versions supported by earlier NVIDIA vGPU software 17 releases are also supported. 2006 (8.0) through 2312 (8.12) 7.0 through 7.13
VMware vCenter Server	8.0 7.0 Update 2 and later updates to release 7.0 unless explicitly stated otherwise **Note:** The base VMware vCenter Server 7.0 release and 7.0 Update 1 are **not** supported.

2.3. Guest OS Support

NVIDIA vGPU software supports several Windows releases and Linux distributions as a guest OS. The supported guest operating systems depend on the hypervisor software version.

Note:

Use only a guest OS release that is listed as supported by NVIDIA vGPU software with your virtualization software. To be listed as supported, a guest OS release must be supported not only by NVIDIA vGPU software, but also by your virtualization software. NVIDIA **cannot** support guest OS releases that your virtualization software does not support.

NVIDIA vGPU software supports **only** 64-bit guest operating systems. No 32-bit guest operating systems are supported.

2.3.1. Windows Guest OS Support

NVIDIA vGPU software supports **only** the 64-bit Windows releases listed as a guest OS on VMware vSphere. The releases of VMware vSphere for which a Windows release is supported depend on whether NVIDIA vGPU or pass-through GPU is used.

Note:

If a specific release, even an update release, is not listed, it's **not** supported.

VMware vMotion with vGPU and suspend-resume with vGPU are supported on supported Windows guest OS releases.

2.3.1.1. Windows Guest OS Support in Release 17.2

Guest OS	NVIDIA vGPU - VMware vSphere Releases	Pass-Through GPU - VMware vSphere Releases
Windows Server 2022	8.0, 7.0	8.0, 7.0
Windows 11 23H2 and all Windows 11 releases supported by Microsoft up to and including this release	8.0, 7.0	7.0
Windows 10 2022 Update (22H2) and all Windows 10 releases supported by Microsoft up to and including this release See Note (1)	8.0, 7.0	8.0, 7.0

Note:

1. The hardware-accelerated GPU scheduling feature introduced in Windows 10 May 2020 Update (2004) is **not** supported on GPUs based on the Maxwell architecture and is supported only in pass-through mode on GPUs based on later architectures.

2.3.1.2. Windows Guest OS Support in Release 17.1

Guest OS	NVIDIA vGPU - VMware vSphere Releases	Pass-Through GPU - VMware vSphere Releases
Windows Server 2022	8.0, 7.0	8.0, 7.0
Windows 11 23H2 and all Windows 11 releases supported by Microsoft up to and including this release	8.0, 7.0	8.0, 7.0
Windows 10 2022 Update (22H2) and all Windows 10 releases supported by Microsoft up to and including this release See Note (1)	8.0, 7.0	8.0, 7.0

Note:

1. The hardware-accelerated GPU scheduling feature introduced in Windows 10 May 2020 Update (2004) is **not** supported on GPUs based on the Maxwell architecture and is supported only in pass-through mode on GPUs based on later architectures.

2.3.1.3. Windows Guest OS Support in Release 17.0

Guest OS	NVIDIA vGPU - VMware vSphere Releases	Pass-Through GPU - VMware vSphere Releases
Windows Server 2022	8.0, 7.0	8.0, 7.0
Windows 11 23H2 and all Windows 11 releases supported by Microsoft up to and including this release	8.0, 7.0	8.0, 7.0
Windows 10 2022 Update (22H2) and all Windows 10 releases supported by Microsoft up to and including this release See Note (1)	8.0, 7.0	8.0, 7.0

Note:

1. The hardware-accelerated GPU scheduling feature introduced in Windows 10 May 2020 Update (2004) is **not** supported on GPUs based on the Maxwell architecture and is supported only in pass-through mode on GPUs based on later architectures.

2.3.2. Linux Guest OS Support

NVIDIA vGPU software supports **only** the Linux distributions listed as a guest OS on VMware vSphere. The releases of VMware vSphere for which a Linux release is supported depend on whether NVIDIA vGPU or pass-through GPU is used.

Note:

If a specific release, even an update release, is not listed, it's **not** supported.

VMware vMotion with vGPU and suspend-resume with vGPU are supported on supported Linux guest OS releases.

2.3.2.1. Linux Guest OS Support in Release 17.2

Guest OS	NVIDIA vGPU - VMware vSphere Releases	Pass-Through GPU - VMware vSphere Releases
Red Hat CoreOS 4.11	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 9.4	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 9.2	8.0, 7.0	8.0, 7.0
Rocky Linux 9.0	22H2, 23H2 preview, 23H2 8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 8.10	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 8.8	8.0, 7.0	8.0, 7.0
Rocky Linux 8.4	8.0, 7.0	8.0, 7.0
Deprecated: CentOS Linux 8 (2105)	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 7.9 and later compatible 7.x versions	8.0, 7.0	8.0, 7.0
Deprecated: CentOS 7.6-7.8 and later compatible 7.x versions	8.0, 7.0	8.0, 7.0
Ubuntu 24.04 LTS	8.0, 7.0	8.0, 7.0
Ubuntu 22.04 LTS	8.0, 7.0	8.0, 7.0
Ubuntu 20.04 LTS	8.0, 7.0	8.0, 7.0
Debian 12	8.0	8.0
SUSE Linux Enterprise Server 15 SP2	8.0, 7.0	8.0, 7.0
SUSE Linux Enterprise Server 12 SP5	8.0, 7.0	8.0, 7.0

Guest OS	NVIDIA vGPU - VMware vSphere Releases	Pass-Through GPU - VMware vSphere Releases
SUSE Linux Enterprise Server 12 SP3	8.0, 7.0	8.0, 7.0

2.3.2.2. Linux Guest OS Support in Release 17.1

Guest OS	NVIDIA vGPU - VMware vSphere Releases	Pass-Through GPU - VMware vSphere Releases
Red Hat CoreOS 4.11	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 9.3	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 9.2	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 9.0	8.0, 7.0	8.0, 7.0
Rocky Linux 9.0	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 8.9	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 8.8	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 8.6	8.0, 7.0	8.0, 7.0
Rocky Linux 8.4	8.0, 7.0	8.0, 7.0
Deprecated: CentOS Linux 8 (2105)	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 7.9 and later compatible 7.x versions	8.0, 7.0	8.0, 7.0
Deprecated: CentOS 7.6-7.8 and later compatible 7.x versions	8.0, 7.0	8.0, 7.0
Ubuntu 22.04 LTS	8.0, 7.0	8.0, 7.0
Ubuntu 20.04 LTS	8.0, 7.0	8.0, 7.0
Debian 10	8.0, 7.0 Update 3	8.0, 7.0 Update 3
SUSE Linux Enterprise Server 15 SP2	8.0, 7.0	8.0, 7.0
SUSE Linux Enterprise Server 12 SP5	8.0, 7.0	8.0, 7.0
SUSE Linux Enterprise Server 12 SP3	8.0, 7.0	8.0, 7.0

2.3.2.3. Linux Guest OS Support in Release 17.0

Guest OS	NVIDIA vGPU - VMware vSphere Releases	Pass-Through GPU - VMware vSphere Releases
Red Hat CoreOS 4.11	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 9.3	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 9.2	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 9.0	8.0, 7.0	8.0, 7.0
Rocky Linux 9.0	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 8.9	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 8.8	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 8.6	8.0, 7.0	8.0, 7.0
Rocky Linux 8.4	8.0, 7.0	8.0, 7.0
Deprecated: CentOS Linux 8 (2105)	8.0, 7.0	8.0, 7.0
Red Hat Enterprise Linux 7.9 and later compatible 7.x versions	8.0, 7.0	8.0, 7.0
Deprecated: CentOS 7.6-7.8 and later compatible 7.x versions	8.0, 7.0	8.0, 7.0
Ubuntu 22.04 LTS	8.0, 7.0	8.0, 7.0
Ubuntu 20.04 LTS	8.0, 7.0	8.0, 7.0
Debian 10	8.0, 7.0 Update 3	8.0, 7.0 Update 3
SUSE Linux Enterprise Server 15 SP2	8.0, 7.0	8.0, 7.0
SUSE Linux Enterprise Server 12 SP5	8.0, 7.0	8.0, 7.0
SUSE Linux Enterprise Server 12 SP3	8.0, 7.0	8.0, 7.0

2.4. NVIDIA CUDA Toolkit Version Support

The releases in this release family of NVIDIA vGPU software support NVIDIA CUDA Toolkit 12.4.

To build a CUDA application, the system must have the NVIDIA CUDA Toolkit and the libraries required for linking. For details of the components of NVIDIA CUDA Toolkit, refer to *NVIDIA CUDA Toolkit 12.4 Release Notes*.

To run a CUDA application, the system must have a CUDA-enabled GPU and an NVIDIA display driver that is compatible with the NVIDIA CUDA Toolkit release that was used to build the application. If the application relies on dynamic linking for libraries, the system must also have the correct version of these libraries.

For more information about NVIDIA CUDA Toolkit, refer to CUDA Toolkit Documentation 12.4.

Note:

If you are using NVIDIA vGPU software with CUDA on Linux, avoid conflicting installation methods by installing CUDA from a distribution-independent runfile package. Do not install CUDA from a distribution-specific RPM or Deb package.

To ensure that the NVIDIA vGPU software graphics driver is not overwritten when CUDA is installed, deselect the CUDA driver when selecting the CUDA components to install.

For more information, see *NVIDIA CUDA Installation Guide for Linux*.

2.5. vGPU Migration Support

vGPU Migration, which includes vMotion and suspend-resume, is supported on all supported GPUs, but only on a subset of supported VMware vSphere Hypervisor (ESXi) releases and guest operating systems.

Limitations with vGPU Migration Support

vGPU migration is disabled for a VM for which any of the following NVIDIA CUDA Toolkit features is enabled:

- Unified memory
- Debuggers
- Profilers

Supported Hypervisor Software Releases

All supported releases of VMware vSphere

Supported Guest OS Releases

Windows and Linux.

Known Issues with vGPU Migration Support

Use Case	Affected GPUs	Issue
Migration of a guest vGPU VM	All GPUs that support vGPU Migration	17.0, 17.1 Only: Desktop is corrupted with XID errors 13 and 31 after vGPU VM is migrated or suspended and resumed

Use Case	Affected GPUs	Issue
Migration between hosts with different ECC memory configuration	All GPUs that support vGPU Migration	Migration of VMs configured with vGPU stops before the migration is complete

2.6. Multiple vGPU Support

To support applications and workloads that are compute or graphics intensive, multiple vGPUs can be added to a single VM. The assignment of more than one vGPU to a VM is supported only on a subset of vGPUs and hypervisor software releases.

2.6.1. vGPUs that Support Multiple vGPUs Assigned to a VM

The supported vGPUs depend on the architecture of the GPU on which the vGPUs reside:

- For GPUs based on the NVIDIA Volta architecture and later GPU architectures, the supported vGPUs also depend on the VMware vSphere release:
  - **Since VMware vSphere 8.0: All** Q-series vGPUs are supported.
  - **VMware vSphere 7.x releases:** Only Q-series vGPUs that are allocated all of the physical GPU's frame buffer are supported.
- For GPUs based on the NVIDIA Maxwell™ graphic architecture, only Q-series vGPUs that are allocated all of the physical GPU's frame buffer are supported.

You can assign multiple vGPUs with differing amounts of frame buffer to a single VM, provided the board type and the series of all the vGPUs is the same. For example, you can assign an A40-48Q vGPU and an A40-16Q vGPU to the same VM. However, you cannot assign an A30-8Q vGPU and an A16-8Q vGPU to the same VM.

Multiple vGPU Support on the NVIDIA Ada Lovelace Architecture

Board	vGPU
NVIDIA L40S	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** L40S-48Q
NVIDIA L40	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** L40-48Q
NVIDIA L20 NVIDIA L20 liquid cooled	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** L20-48Q

Board	vGPU
NVIDIA L4	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** L4-24Q
NVIDIA L2	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** L2-24Q
NVIDIA RTX 6000 Ada	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** RTX 6000 Ada-48Q
NVIDIA RTX 5880 Ada	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** RTX 5880 Ada-48Q
NVIDIA RTX 5000 Ada	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** RTX 5000 Ada-32Q

Multiple vGPU Support on the NVIDIA Ampere GPU Architecture

Board	vGPU
NVIDIA A40	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** A40-48Q See Note (1).
NVIDIA A16	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** A16-16Q See Note (1).
NVIDIA A10	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** A10-24Q See Note (1).
NVIDIA A2	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** A2-16Q

Board	vGPU
	See Note (1).
NVIDIA RTX A6000	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** A6000-48Q See Note (1).
NVIDIA RTX A5500	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** A5500-24Q See Note (1).
NVIDIA RTX A5000	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** A5000-24Q See Note (1).

Multiple vGPU Support on the NVIDIA Turing GPU Architecture

Board	vGPU
Tesla T4	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** T4-16Q
Quadro RTX 6000	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** RTX6000-24Q
Quadro RTX 6000 passive	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** RTX6000P-24Q
Quadro RTX 8000	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** RTX8000-48Q
Quadro RTX 8000 passive	**Since VMware vSphere 8.0: All** Q-series vGPUs **VMware vSphere 7.x releases:** RTX8000P-48Q

Multiple vGPU Support on the NVIDIA Maxwell GPU Architecture

Board	vGPU
Tesla M10	M10-8Q

Note:

1. This type of vGPU cannot be assigned with other types of vGPU to the same VM.

2.6.2. Maximum Number of vGPUs Supported per VM

For VMware vSphere, the maximum number of vGPUs per VM supported depends on the hypervisor release:

Hypervisor Release	Maximum Number of vGPUs per VM
Since VMware vSphere 8.0 Update 2	16
VMware vSphere 8.0 and 8.0 Update 1	8
VMware vSphere 7.*x* releases:	4

2.6.3. Hypervisor Releases that Support Multiple vGPUs Assigned to a VM

All hypervisor releases that support NVIDIA vGPU software are supported.

2.7. Peer-to-Peer CUDA Transfers over NVLink Support

Peer-to-peer CUDA transfers enable device memory between vGPUs on different GPUs that are assigned to the same VM to be accessed from within the CUDA kernels. NVLink is a high-bandwidth interconnect that enables fast communication between such vGPUs. Peer-to-Peer CUDA transfers over NVLink are supported only on a subset of vGPUs, VMware vSphere Hypervisor (ESXi) releases, and guest OS releases.

2.7.1. vGPUs that Support Peer-to-Peer CUDA Transfers

Only Q-series vGPUs that are allocated all of the physical GPU's frame buffer on physical GPUs that support NVLink are supported.

Peer-to-Peer CUDA Transfer Support on the NVIDIA Ampere GPU Architecture

Board	vGPU
NVIDIA A40	A40-48Q
NVIDIA A10	A10-24Q
NVIDIA RTX A6000	A6000-48Q
NVIDIA RTX A5500	A5500-24Q
NVIDIA RTX A5000	A5000-24Q

Peer-to-Peer CUDA Transfer Support on the NVIDIA Turing GPU Architecture

Board	vGPU
Quadro RTX 6000	RTX6000-24Q
Quadro RTX 6000 passive	RTX6000P-24Q
Quadro RTX 8000	RTX8000-48Q
Quadro RTX 8000 passive	RTX8000P-48Q

Peer-to-Peer CUDA Transfer Support on the NVIDIA Volta GPU Architecture

Board	vGPU
Tesla V100 SXM2 32GB	V100DX-32Q
Tesla V100 SXM2	V100X-16Q

Note:

1. Supported only on the following hardware:
   - NVIDIA HGX™ A100 4-GPU baseboard with four fully connected GPUs

   Fully connected means that each GPU is connected to every other GPU on the baseboard.

2.7.2. Hypervisor Releases that Support Peer-to-Peer CUDA Transfers

Peer-to-Peer CUDA transfers over NVLink are supported on all hypervisor releases that support the assignment of more than one vGPU to a VM. For details, see Multiple vGPU Support.

2.7.3. Guest OS Releases that Support Peer-to-Peer CUDA Transfers

Linux only. Peer-to-Peer CUDA transfers over NVLink are **not** supported on Windows.

2.7.4. Limitations on Support for Peer-to-Peer CUDA Transfers

- NVSwitch is not supported. Only direct connections are supported.
- Only time-sliced vGPUs are supported. MIG-backed vGPUs are **not** supported.
- PCIe is not supported.
- SLI is not supported.

2.8. Unified Memory Support

Unified memory is a single memory address space that is accessible from any CPU or GPU in a system. It creates a pool of managed memory that is shared between the CPU and GPU to provide a simple way to allocate and access data that can be used by code running on any CPU or GPU in the system. Unified memory is supported only on a subset of vGPUs and guest OS releases.

Note: Unified memory is disabled by default. If used, you must enable unified memory individually for each vGPU that requires it by setting a vGPU plugin parameter. NVIDIA CUDA Toolkit profilers are supported and can be enabled on a VM for which unified memory is enabled.

2.8.1. vGPUs that Support Unified Memory

Only Q-series vGPUs that are allocated all of the physical GPU's frame buffer on physical GPUs that support unified memory are supported.

Unified Memory Support on the NVIDIA Ada Lovelace GPU Architecture

Board	vGPU
NVIDIA L40	L40-48Q
NVIDIA L40S	L40S-48Q
NVIDIA L20 NVIDIA L20 liquid cooled	L20-48Q
NVIDIA L4	L4-24Q

Board	vGPU
NVIDIA L2	L2-24Q
NVIDIA RTX 6000 Ada	RTX 6000 Ada-48Q
NVIDIA RTX 5880 Ada	RTX 5880 Ada-48Q
NVIDIA RTX 5000 Ada	RTX 5000 Ada-32Q

Unified Memory Support on the NVIDIA Ampere GPU Architecture

Board	vGPU
NVIDIA A40	A40-48Q
NVIDIA A16	A16-16Q
NVIDIA A10	A10-24Q
NVIDIA A2	A2-16Q
NVIDIA RTX A6000	A6000-48Q
NVIDIA RTX A5500	A5500-24Q
NVIDIA RTX A5000	A5000-24Q

2.8.2. Guest OS Releases that Support Unified Memory

Linux only. Unified memory is **not** supported on Windows.

2.8.3. Limitations on Support for Unified Memory

- Only time-sliced Q-series and C-series vGPUs that are allocated all of the physical GPU's frame buffer on physical GPUs that support unified memory are supported. Fractional time-sliced vGPUs are **not** supported.
- When unified memory is enabled for a VM, vGPU migration is disabled for the VM.

2.9. NVIDIA GPU Operator Support

NVIDIA GPU Operator simplifies the deployment of NVIDIA vGPU software with software container platforms on immutable operating systems. An immutable operating system does not allow the installation of the NVIDIA vGPU software graphics driver directly on the operating system. NVIDIA GPU Operator is supported only on specific combinations of hypervisor software release, container platform, and guest OS release.

Hypervisor Software Release	Container Platform	Guest OS
VMware vSphere Hypervisor (ESXi) 7.0 Update 2	VMware Tanzu Kubernetes Grid	Ubuntu 20.04 LTS
VMware vSphere Hypervisor (ESXi) 7.0 Update 2	Red Hat Openshift 4.11 with Red Hat Enterprise Linux CoreOS and the CRI-O container runtime	Red Hat CoreOS 4.11
VMware vSphere Hypervisor (ESXi) 7.0 Update 2	Red Hat Openshift 4.8 with Red Hat Enterprise Linux CoreOS and the CRI-O container runtime	Red Hat CoreOS 4.8

2.10. NVIDIA Deep Learning Super Sampling (DLSS) Support

NVIDIA vGPU software supports NVIDIA DLSS on NVIDIA RTX Virtual Workstation.

Supported DLSS versions: 2.0. Version 1.0 is **not** supported.

Supported GPUs:

- NVIDIA L40
- NVIDIA L40S
- NVIDIA L20
- NVIDIA L20 liquid cooled
- NVIDIA L4
- NVIDIA L2
- NVIDIA RTX 6000 Ada
- NVIDIA RTX 5880 Ada
- NVIDIA RTX 5000 Ada
- NVIDIA A40
- NVIDIA A16
- NVIDIA A2
- NVIDIA A10
- NVIDIA RTX A6000
- NVIDIA RTX A5500
- NVIDIA RTX A5000
- Tesla T4
- Quadro RTX 8000

- Quadro RTX 8000 passive
- Quadro RTX 6000
- Quadro RTX 6000 passive

Note: NVIDIA graphics driver components that DLSS requires are installed only if a supported GPU is detected during installation of the driver. Therefore, if the creation of VM templates includes driver installation, the template should be created from a VM that is configured with a supported GPU while the driver is being installed.

Supported applications: only applications that use `nvngx_dlss.dll` version 2.0.18 or newer

2.11. vSphere Lifecycle Management (vLCM) Support

NVIDIA vGPU software supports updating the Virtual GPU Manager for VMware vSphere Hypervisor (ESXi) by using vLCM.

Supported VMware vSphere Hypervisor (ESXi) releases: 7.0 Update 2 and later updates to release 7.0 unless explicitly stated otherwise

Supported VMware vCenter Server releases: 7.0 Update 2 and later updates to release 7.0 unless explicitly stated otherwise

Chapter 3. Known Product Limitations

Known product limitations for this release of NVIDIA vGPU software are described in the following sections.

3.1. 17.0, 17.1 Only: vGPUs of different sizes on the same GPU are not supported

VMware vSphere Hypervisor (ESXi) supports a mixture of time-sliced vGPUs with the same amount of frame buffer from different virtual GPU series on the same physical GPU. A-series, B-series, and Q-series vGPUs with the same amount of frame buffer, for example, A40-2B and A40-2Q, can reside on the same physical GPU simultaneously. However, vGPUs with different amounts of frame buffer are not supported on the same GPU.

VMware vSphere Hypervisor (ESXi) 8 Update 3 and, unless explicitly stated otherwise, later update releases supports a mixture of different types of time-sliced vGPUs on the same physical GPU. Any combination of A-series, B-series, and Q-series vGPUs with any amount of frame buffer can reside on the same physical GPU simultaneously. The total amount of frame buffer allocated to the vGPUs on a physical GPU must not exceed the amount of frame buffer that the physical GPU has.

3.2. NVENC does not support resolutions greater than 4096×4096

Description

The NVIDIA hardware-based H.264 video encoder (NVENC) does not support resolutions greater than 4096×4096. This restriction applies to all NVIDIA GPU architectures and

is imposed by the GPU encoder hardware itself, not by NVIDIA vGPU software. The maximum supported resolution for each encoding scheme is listed in the documentation for NVIDIA Video Codec SDK. This limitation affects any remoting tool where H.264 encoding is used with a resolution greater than 4096×4096. Most supported remoting tools fall back to software encoding in such scenarios.

Workaround

If your GPU is based on a GPU architecture later than the NVIDIA Maxwell® architecture, use H.265 encoding. H.265 is more efficient than H.264 encoding and has a maximum resolution of 8192×8192. On GPUs based on the NVIDIA Maxwell architecture, H.265 has the same maximum resolution as H.264, namely 4096×4096.

Note: Resolutions greater than 4096×4096 are supported only by the H.265 decoder that 64-bit client applications use. The H.265 decoder that 32-bit applications use supports a maximum resolution of 4096×4096.

Because the client-side Workspace App on Windows is a 32-bit application, resolutions greater than 4096×4096 are not supported for Windows clients of Citrix Virtual Apps and Desktops. Therefore, if you are using a Windows client with Citrix Virtual Apps and Desktops, ensure that you are using H.264 hardware encoding with the default Use video codec for compression Citrix graphics policy setting, namely **Actively Changing Regions**. This policy setting encodes only actively changing regions of the screen (for example, a window in which a video is playing). Provided that the number of pixels along any edge of the actively changing region does not exceed 4096, H.264 encoding is offloaded to the NVENC hardware encoder.

3.3. vCS is not supported on VMware vSphere

NVIDIA Virtual Compute Server (vCS) is not supported on VMware vSphere. C-series vGPU types are not available.

Instead, vCS is supported with NVIDIA AI Enterprise. For more information, see NVIDIA AI Enterprise Documentation.

3.4. Nested Virtualization Is Not Supported by NVIDIA vGPU

NVIDIA vGPU deployments do not support nested virtualization, that is, running a hypervisor in a guest VM. For example, enabling the Hyper-V role in a guest VM running the Windows Server OS is **not** supported because it entails enabling nested virtualization.

Similarly, enabling Windows Hypervisor Platform is not supported because it requires the Hyper-V role to be enabled.

3.5. Issues occur when the channels allocated to a vGPU are exhausted

Description

Issues occur when the channels allocated to a vGPU are exhausted and the guest VM to which the vGPU is assigned fails to allocate a channel to the vGPU. A physical GPU has a fixed number of channels and the number of channels allocated to each vGPU is inversely proportional to the maximum number of vGPUs allowed on the physical GPU.

When the channels allocated to a vGPU are exhausted and the guest VM fails to allocate a channel, the following errors are reported on the hypervisor host or in an NVIDIA bug report:

```
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0): Guest attempted to
 allocate channel above its max channel limit 0xfb
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0): VGPU message 6
 failed, result code: 0x1a
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):
 0xc1d004a1, 0xff0e0000, 0xff0400fb, 0xc36f,
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):         0x1,
 0xff1fe314, 0xff1fe038, 0x100b6f000, 0x1000,
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):
 0x80000000, 0xff0e0200, 0x0, 0x0, (Not logged),
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):         0x1, 0x0
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0): , 0x0
```

Workaround

Use a vGPU type with more frame buffer, thereby reducing the maximum number of vGPUs allowed on the physical GPU. As a result, the number of channels allocated to each vGPU is increased.

3.6. Total frame buffer for vGPUs is less than the total frame buffer on the physical GPU

Some of the physical GPU's frame buffer is used by the hypervisor on behalf of the VM for allocations that the guest OS would otherwise have made in its own frame buffer. The frame buffer used by the hypervisor is not available for vGPUs on the physical GPU. In NVIDIA vGPU deployments, frame buffer for the guest OS is reserved in advance, whereas in bare-metal deployments, frame buffer for the guest OS is reserved on the basis of the runtime needs of applications.

If error-correcting code (ECC) memory is enabled on a physical GPU that does not have HBM2 memory, the amount of frame buffer that is usable by vGPUs is further reduced. All types of vGPU are affected, not just vGPUs that support ECC memory.

On all GPUs that support ECC memory and, therefore, dynamic page retirement, additional frame buffer is allocated for dynamic page retirement. The amount that is allocated is inversely proportional to the maximum number of vGPUs per physical GPU. All GPUs that support ECC memory are affected, even GPUs that have HBM2 memory or for which ECC memory is disabled.

The approximate amount of frame buffer that NVIDIA vGPU software reserves can be calculated from the following formula:

max-reserved-fb = *vgpu-profile-size-in-mb*÷16 + 16 + *ecc-adjustments* + *page-retirement-allocation* + *compression-adjustment*

max-reserved-fb
The maximum total amount of reserved frame buffer in Mbytes that is not available for vGPUs.

vgpu-profile-size-in-mb
The amount of frame buffer in Mbytes allocated to a single vGPU. This amount depends on the vGPU type. For example, for the T4-16Q vGPU type, *vgpu-profile-size-in-mb* is 16384.

ecc-adjustments
The amount of frame buffer in Mbytes that is not usable by vGPUs when ECC is enabled on a physical GPU that does not have HBM2 memory.

- If ECC is enabled on a physical GPU that does not have HBM2 memory *ecc-adjustments* is *fb-without-ecc*/16, which is equivalent to 64 Mbytes for every Gbyte of frame buffer assigned to the vGPU. *fb-without-ecc* is total amount of frame buffer with ECC disabled.
- If ECC is disabled or the GPU has HBM2 memory, *ecc-adjustments* is 0.

page-retirement-allocation
The amount of frame buffer in Mbytes that is reserved for dynamic page retirement.

- On GPUs based on the NVIDIA Maxwell GPU architecture, *page-retirement-allocation* = 4÷*max-vgpus-per-gpu*.
- On GPUs based on NVIDIA GPU architectures **after** the Maxwell architecture, *page-retirement-allocation* = 128÷*max-vgpus-per-gpu*

max-vgpus-per-gpu
The maximum number of vGPUs that can be created simultaneously on a physical GPU. This number varies according to the vGPU type. For example, for the T4-16Q vGPU type, *max-vgpus-per-gpu* is 1.

compression-adjustment

The amount of frame buffer in Mbytes that is reserved for the higher compression overhead in vGPU types with 12 Gbytes or more of frame buffer on GPUs based on the Turing architecture.

compression-adjustment depends on the vGPU type as shown in the following table.

vGPU Type	Compression Adjustment (MB)
T4-16Q T4-16C T4-16A	28
RTX6000-12Q RTX6000-12C RTX6000-12A	32
RTX6000-24Q RTX6000-24C RTX6000-24A	104
RTX6000P-12Q RTX6000P-12C RTX6000P-12A	32
RTX6000P-24Q RTX6000P-24C RTX6000P-24A	104
RTX8000-12Q RTX8000-12C RTX8000-12A	32
RTX8000-16Q RTX8000-16C RTX8000-16A	64
RTX8000-24Q RTX8000-24C RTX8000-24A	96
RTX8000-48Q RTX8000-48C RTX8000-48A	238
RTX8000P-12Q RTX8000P-12C RTX8000P-12A	32
RTX8000P-16Q RTX8000P-16C	64

vGPU Type	Compression Adjustment (MB)
RTX8000P-16A	
RTX8000P-24Q RTX8000P-24C RTX8000P-24A	96
RTX8000P-48Q RTX8000P-48C RTX8000P-48A	238

For all other vGPU types, *compression-adjustment* is 0.

Note: In VMs running Windows Server 2012 R2, which supports Windows Display Driver Model (WDDM) 1.*x*, an additional 48 Mbytes of frame buffer are reserved and not available for vGPUs.

3.7. Issues may occur with graphics-intensive OpenCL applications on vGPU types with limited frame buffer

Description

Issues may occur when graphics-intensive OpenCL applications are used with vGPU types that have limited frame buffer. These issues occur when the applications demand more frame buffer than is allocated to the vGPU.

For example, these issues may occur with the Adobe Photoshop and LuxMark OpenCL Benchmark applications:

- When the image resolution and size are changed in Adobe Photoshop, a program error may occur or Photoshop may display a message about a problem with the graphics hardware and a suggestion to disable OpenCL.
- When the LuxMark OpenCL Benchmark application is run, XID error 31 may occur.

Workaround

For graphics-intensive OpenCL applications, use a vGPU type with more frame buffer.

3.8. In pass through mode, all GPUs connected to each other through NVLink must be assigned to the same VM

Description

In pass through mode, all GPUs connected to each other through NVLink must be assigned to the same VM. If a subset of GPUs connected to each other through NVLink is passed through to a VM, unrecoverable error `XID 74` occurs when the VM is booted. This error corrupts the NVLink state on the physical GPUs and, as a result, the NVLink bridge between the GPUs is unusable.

Workaround

Restore the NVLink state on the physical GPUs by resetting the GPUs or rebooting the hypervisor host.

3.9. vGPU profiles with 512 Mbytes or less of frame buffer support only 1 virtual display head on Windows 10

Description

To reduce the possibility of memory exhaustion, vGPU profiles with 512 Mbytes or less of frame buffer support only 1 virtual display head on a Windows 10 guest OS.

The following vGPU profiles have 512 Mbytes or less of frame buffer:

- Tesla M10-0B
- Tesla M10-0Q

Workaround

Use a profile that supports more than 1 virtual display head and has at least 1 Gbyte of frame buffer.

3.10. NVENC requires at least 1 Gbyte of frame buffer

Description

Using the frame buffer for the NVIDIA hardware-based H.264/HEVC video encoder (NVENC) may cause memory exhaustion with vGPU profiles that have 512 Mbytes or less of frame buffer. To reduce the possibility of memory exhaustion, NVENC is disabled on profiles that have 512 Mbytes or less of frame buffer. Application GPU acceleration remains fully supported and available for all profiles, including profiles with 512 MBytes or less of frame buffer. NVENC support from both Citrix and VMware is a recent feature and, if you are using an older version, you should experience no change in functionality.

The following vGPU profiles have 512 Mbytes or less of frame buffer:

- Tesla M10-0B
- Tesla M10-0Q

Workaround

If you require NVENC to be enabled, use a profile that has at least 1 Gbyte of frame buffer.

3.11. VM failures or crashes on servers with 1 TiB or more of system memory

Description

Support for vGPU is limited to servers with less than 1 TiB of system memory. On servers with 1 TiB or more of system memory, VM failures or crashes may occur. For example, when Citrix Virtual Apps and Desktops is used with a Windows 7 guest OS, a blue screen crash may occur. However, support for vDGA is not affected by this limitation.

Depending on the version of NVIDIA vGPU software that you are using, the log file on the VMware vSphere host might also report the following errors:

```
2016-10-27T04:36:21.128Z cpu74:70210)DMA: 1935: Unable to perform element mapping:
 DMA mapping could not be completed
2016-10-27T04:36:21.128Z cpu74:70210)Failed to DMA map address 0x118d296c000
 (0x4000): Can't meet address mask of the device..
2016-10-27T04:36:21.128Z cpu74:70210)NVRM: VM: nv_alloc_contig_pages: failed to
 allocate memory
```

This limitation applies only to systems with supported GPUs based on the Maxwell architecture, namely, Tesla M10.

Resolution

Limit the amount of system memory on the server to 1 TiB minus 16 GiB.

1. Set `memmapMaxRAMMB` to 1032192, which is equal to 1048576 minus 16384.

   For detailed instructions, see Set Advanced Host Attributes in the VMware vSphere documentation.
2. Reboot the server.

If the problem persists, contact your server vendor for the recommended system memory configuration with NVIDIA GPUs.

3.12. VM running an incompatible NVIDIA vGPU guest driver fails to initialize vGPU when booted

Description

A VM running a version of the NVIDIA guest VM driver that is incompatible with the current release of Virtual GPU Manager will fail to initialize vGPU when booted on a VMware vSphere platform running that release of Virtual GPU Manager.

A guest VM driver is incompatible with the current release of Virtual GPU Manager in either of the following situations:

- The guest driver is from a release in a branch two or more major releases before the current release, for example release 9.4.

  In this situation, the VMware vSphere VM's log file reports the following error:

  ```
  vmiop_log: (0x0): Incompatible Guest/Host drivers: Guest VGX version is older
   than the minimum version supported by the Host. Disabling vGPU.
  ```
- The guest driver is from a later release than the Virtual GPU Manager.

  In this situation, the VMware vSphere VM's log file reports the following error:

  ```
  vmiop_log: (0x0): Incompatible Guest/Host drivers: Guest VGX version is newer
   than the maximum version supported by the Host. Disabling vGPU.
  ```

In either situation, the VM boots in standard VGA mode with reduced resolution and color depth. The NVIDIA virtual GPU is present in **Windows Device Manager** but displays a warning sign, and the following device status:

```
Windows has stopped this device because it has reported problems. (Code 43)
```

Resolution

Install a release of the NVIDIA guest VM driver that is compatible with current release of Virtual GPU Manager.

3.13. Single vGPU benchmark scores are lower than pass-through GPU

Description

A single vGPU configured on a physical GPU produces lower benchmark scores than the physical GPU run in pass-through mode.

Aside from performance differences that may be attributed to a vGPU's smaller frame buffer size, vGPU incorporates a performance balancing feature known as Frame Rate Limiter (FRL). On vGPUs that use the best-effort scheduler, FRL is enabled. On vGPUs that use the fixed share or equal share scheduler, FRL is disabled.

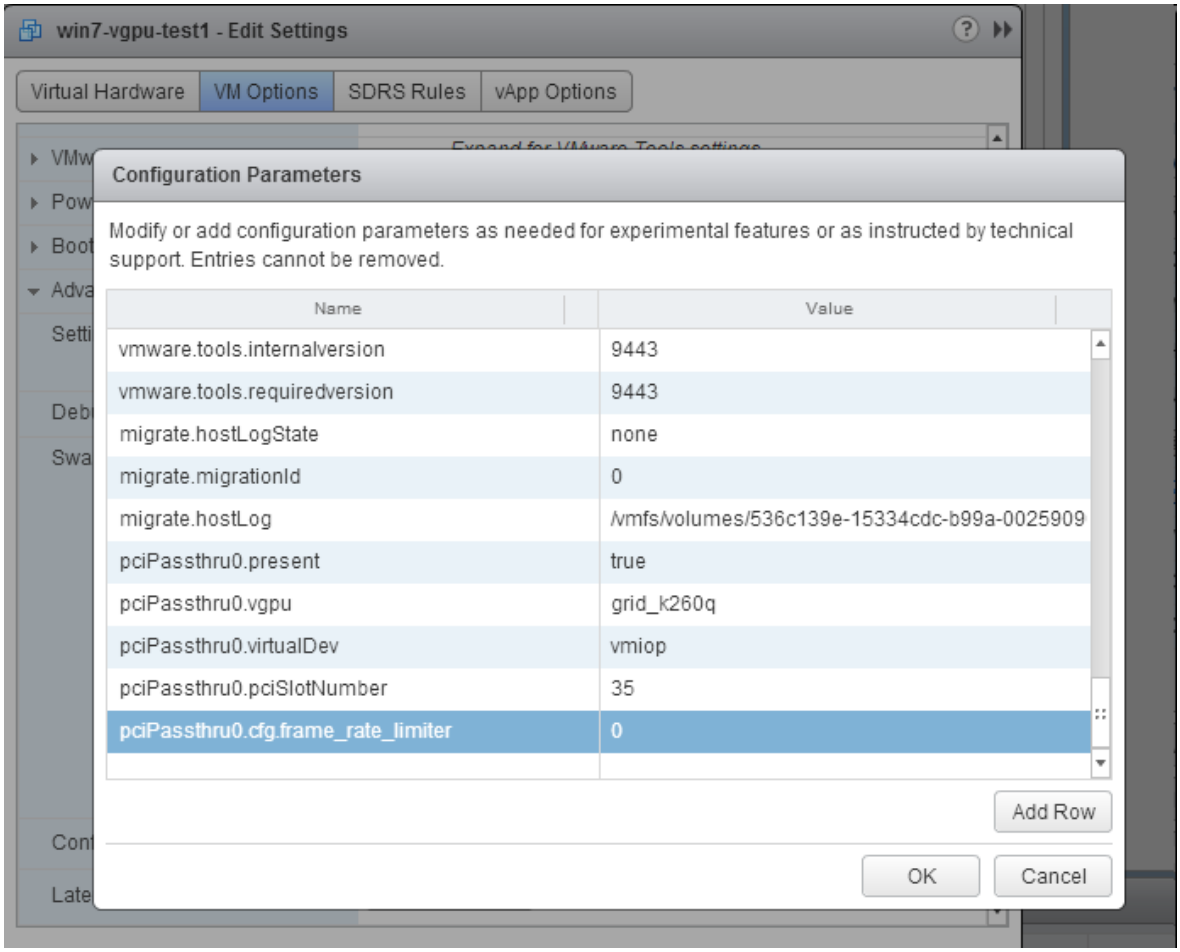
FRL is used to ensure balanced performance across multiple vGPUs that are resident on the same physical GPU. The FRL setting is designed to give good interactive remote graphics experience but may reduce scores in benchmarks that depend on measuring frame rendering rates, as compared to the same benchmarks running on a pass-through GPU.

Resolution

FRL is controlled by an internal vGPU setting. On vGPUs that use the best-effort scheduler, NVIDIA does not validate vGPU with FRL disabled, but for validation of benchmark performance, FRL can be temporarily disabled by adding the configuration parameter `pciPassthru0.cfg.frame_rate_limiter` in the VM's advanced configuration options.

Note: This setting can only be changed when the VM is powered off.

1. Select **Edit Settings**.
2. In **Edit Settings** window, select the **VM Options** tab.
3. From the **Advanced** drop-down list, select **Edit Configuration**.
4. In the **Configuration Parameters** dialog box, click **Add Row**.
5. In the **Name** field, type the parameter name `pciPassthru0.cfg.frame_rate_limiter`, in the **Value** field type 0, and click **OK**.



With this setting in place, the VM's vGPU will run without any frame rate limit. The FRL can be reverted back to its default setting by setting `pciPassthru0.cfg.frame_rate_limiter` to `1` or by removing the parameter from the advanced settings.

3.14. VMs configured with large memory fail to initialize vGPU when booted

Description

When starting multiple VMs configured with large amounts of RAM (typically more than 32GB per VM), a VM may fail to initialize vGPU. In this scenario, the VM boots in VMware SVGA mode and doesn't load the NVIDIA driver. The NVIDIA vGPU software GPU is present in **Windows Device Manager** but displays a warning sign, and the following device status:

```
Windows has stopped this device because it has reported problems. (Code 43)
```

When this error occurs, `VGPU message failed` messages and XID error messages are written to the VMware vSphere VM's log file.

Resolution

vGPU reserves a portion of the VM's framebuffer for use in GPU mapping of VM system memory. The reservation is sufficient to support up to 32GB of system memory, and may be increased to accommodate up to 64GB by adding the configuration parameter `pciPassthru0.cfg.enable_large_sys_mem` in the VM's advanced configuration options

Note: This setting can only be changed when the VM is powered off.

1. Select **Edit Settings**.
2. In **Edit Settings** window, select the **VM Options** tab.
3. From the **Advanced** drop-down list, select **Edit Configuration**.
4. In the **Configuration Parameters** dialog box, click **Add Row**.
5. In the **Name** field, type the parameter name `pciPassthru0.cfg.enable_large_sys_mem`, in the **Value** field type 1, and click **OK**.

With this setting in place, less GPU framebuffer is available to applications running in the VM. To accommodate system memory larger than 64GB, the reservation can be further increased by adding `pciPassthru0.cfg.extra_fb_reservation` in the VM's advanced configuration options, and setting its value to the desired reservation size in megabytes. The default value of 64M is sufficient to support 64 GB of RAM. We recommend adding 2 M of reservation for each additional 1 GB of system memory. For example, to support 96 GB of RAM, set `pciPassthru0.cfg.extra_fb_reservation` to 128.

The reservation can be reverted back to its default setting by setting `pciPassthru0.cfg.enable_large_sys_mem` to 0, or by removing the parameter from the advanced settings.

Chapter 4. Resolved Issues

Only resolved issues that have been previously noted as known issues or had a noticeable user impact are listed. The summary and description for each resolved issue indicate the effect of the issue on NVIDIA vGPU software **before the issue was resolved**.

4.1. Issues Resolved in Release 17.3

No resolved issues are reported in this release for VMware vSphere.

4.2. Issues Resolved in Release 17.2

Bug ID	Summary and Description
4558671	**17.0, 17.1 Only: Disabling or disconnecting a display can cause a TDR on a Windows vGPU VM** Disabling or disconnecting a display can cause a Timeout Detection and Recovery (TDR) error on a Windows VM that is configured with NVIDIA vGPU. The TDR error might cause a VM crash or intermittent black screens with remoting solutions such as VMware Horizon. When this error occurs, TDR error, `XID error 44`, and `XID error 109` messages are written to the log file on the hypervisor host.
3964376	**17.0, 17.1 Only: Desktop is corrupted with XID errors 13 and 31 after vGPU VM is migrated or suspended and resumed** When a vGPU VM is migrated or suspended and resumed, the remote desktop session window is corrupted. XID error 13, XID error 31, or both errors might also occur.
4600308	**17.0, 17.1 Only: XID error 120 causes multiple issues with NVIDIA vGPU on GPUs with a GSP** XID error 120 causes multiple issues with VMs configured with NVIDIA vGPU on a physical GPU that includes a GPU System Processor (GSP), such

Bug ID	Summary and Description
	as GPUs based on the NVIDIA Ada Lovelace GPU architecture. Examples of these issues include: ▶ VMs hang or crash. ▶ VMs fail to power on after hanging or a crashing. ▶ The hypervisor host crashes.
4644559	**17.0, 17.1 Only: XID error 119 causes the hypervisor host to hang or crash when multiple vGPU VMs are shut down** When multiple VMs configured with NVIDIA vGPU are shut down simultaneously, XID error 119 causes the hypervisor host to hang or crash. This issue affects VMs configured with NVIDIA vGPU on a physical GPU that includes a GPU System Processor (GSP), such as GPUs based on the NVIDIA Ada Lovelace GPU architecture.

4.3. Issues Resolved in Release 17.1

No resolved issues are reported in this release for VMware vSphere.

4.4. Issues Resolved in Release 17.0

Bug ID	Summary and Description
3973158	**Pixelation occurs on a Windows VM configured with a Tesla T4 vGPU** Users might experience poor graphics quality on a Windows VM that is configured with a vGPU on a Tesla T4 GPU. This issue can cause random pixelation on the entire screen, or only on some patches of the screen. No errors are reported or written to the log files when this issue occurs.

Chapter 5. Known Issues

5.1. 17.0, 17.1 Only: Disabling or disconnecting a display can cause a TDR on a Windows vGPU VM

Description

Disabling or disconnecting a display can cause a Timeout Detection and Recovery (TDR) error on a Windows VM that is configured with NVIDIA vGPU. The TDR error might cause a VM crash or intermittent black screens with remoting solutions such as VMware Horizon. When this error occurs, TDR error, `XID error 44`, and `XID error 109` messages are written to the log file on the hypervisor host.

Status

Resolved in NVIDIA vGPU software 17.2

After upgrading to a release in which this issue is resolved, you must set a registry key value **in the guest VM** for the vGPU to avoid this issue.

1. Obtain the driver key of the vGPU.
   a). In **Device Manager**, expand **Display Device**, context-click the vGPU and from the menu that pops up, choose **Properties**.
   b). In the **Properties** window that opens, click the **Details** tab and in the **Property** drop-down list, select **Driver key**.
2. Open the **Registry Editor** and navigate to the Windows registry key `Computer\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\Class\`*`driver-key`*.
   driver-key
   The driver key for the vGPU that you got in the previous step, for example, `{4d36e968-e325-11ce-bfc1-08002be10318}\0001`.
3. Add the `R550_4558671 = 1` DWord (`REG_DWORD`) registry value.
4. Restart the guest VM to reload the graphics driver.

Ref.

4558671

5.2. 17.0, 17.1 Only: XID error 120 causes multiple issues with NVIDIA vGPU on GPUs with a GSP

Description

XID error 120 causes multiple issues with VMs configured with NVIDIA vGPU on a physical GPU that includes a GPU System Processor (GSP), such as GPUs based on the NVIDIA Ada Lovelace GPU architecture. Examples of these issues include:

- VMs hang or crash.
- VMs fail to power on after hanging or a crashing.
- The hypervisor host crashes.

Status

Resolved in NVIDIA vGPU software 17.2

Ref.

4600308

5.3. 17.0, 17.1 Only: XID error 119 causes the hypervisor host to hang or crash when multiple vGPU VMs are shut down

Description

When multiple VMs configured with NVIDIA vGPU are shut down simultaneously, XID error 119 causes the hypervisor host to hang or crash. This issue affects VMs configured with NVIDIA vGPU on a physical GPU that includes a GPU System Processor (GSP), such as GPUs based on the NVIDIA Ada Lovelace GPU architecture.

Status

Resolved in NVIDIA vGPU software 17.2

Ref.

4644559

5.4. 17.0, 17.1 Only: Desktop is corrupted with XID errors 13 and 31 after vGPU VM is migrated or suspended and resumed

Description

When a vGPU VM is migrated or suspended and resumed, the remote desktop session window is corrupted. XID error 13, XID error 31, or both errors might also occur.

When this issue occurs, error messages similar to the following examples are written to the log file on the hypervisor host:

```
Apr  8 11:27:28 smc220-0008 kernel: NVRM: Xid (PCI:0000:4b:00): 31, pid=6327,
 name=nvidia-vgpu-mgr, Ch 00000701
...
Apr  8 11:27:29 smc220-0008 nvidia-vgpu-mgr[6327]: error: vmiop_log: (0x0): XID 31
 detected on physical_chid:0x701, guest_chid:0x1
Apr  8 11:27:29 smc220-0008 nvidia-vgpu-mgr[6327]: error: vmiop_log: (0x0): MMU
 Exception data for XID 31: addrLo 0x0, addHi 0x0, faultType 0 engineId 1
```

Status

Resolved in NVIDIA vGPU software 17.2

Ref.

3964376

5.5. vGPU VM fails to boot with error `vmiop-display unable to reserve vgpu`

Description

A VM that is configured with NVIDIA vGPU on any NVIDIA RTX Ada graphics card, such as the NVIDIA RTX 6000 Ada and NVIDIA RTX 5000 Ada, fails to boot. When this issue occurs, the error message `vmiop-display unable to reserve vgpu` is written to the log files for the VM on the hypervisor host. This issue occurs because an issue with VMware vSphere Hypervisor (ESXi) prevents the hypervisor software from parsing the names of the virtual GPU types for these cards.

Version

This issue affects **only** the VMware vSphere Hypervisor (ESXi) 8.0U2 General Availability (GA) release. Other VMware vSphere Hypervisor (ESXi) releases that NVIDIA vGPU software supports are **not** affected.

Status

Not an NVIDIA bug

Ref.

4293546

5.6. **NVIDIA Control Panel** is not available in multiuser environments

Description

After the NVIDIA vGPU software graphics driver for Windows is installed, the **NVIDIA Control Panel** app might be missing from the system. This issue typically occurs in the following situations:

- Multiple users connect to virtual machines by using remote desktop applications such as Microsoft RDP, VMware Horizon, and Citrix Virtual Apps and Desktops.
- VM instances are created by using Citrix Machine Creation Services (MCS) or VMware Instant Clone technology.

- Roaming user desktop profiles are deployed.

This issue occurs because the **NVIDIA Control Panel** app is now distributed through the **Microsoft Store**. The **NVIDIA Control Panel** app might fail to be installed when the NVIDIA vGPU software graphics driver for Windows is installed if the **Microsoft Store** app is disabled, the system is not connected to the Internet, or installation of apps from the **Microsoft Store** is blocked by your system settings.

To determine whether the **NVIDIA Control Panel** app is installed on your system, use the **Windows Settings** app or the `Get-AppxPackage` Windows PowerShell command.

- To use the **Windows Settings** app:
  1. From the Windows **Start** menu, choose **Settings** > **Apps** > **Apps & feautures** .
  2. In the **Apps & features** window, type **`nvidia control panel`** in the search box and confirm that the **NVIDIA Control Panel** app is found.
- To use the `Get-AppxPackage`Windows PowerShell command:
  1. Run **Windows PowerShell** as Administrator.
  2. Determine whether the **NVIDIA Control Panel** app is installed for the current user.

     ```
     PS C:\> Get-AppxPackage -Name NVIDIACorp.NVIDIAControlPanel
     ```
  3. Determine whether the **NVIDIA Control Panel** app is installed for all users.

     ```
     PS C:\> Get-AppxPackage -AllUsers -Name NVIDIACorp.NVIDIAControlPanel
     ```

     This example shows that the **NVIDIA Control Panel** app is installed for the users `Administrator`, `pliny`, and `trajan`.

     ```
     PS C:\> Get-AppxPackage -AllUsers -Name NVIDIACorp.NVIDIAControlPanel

     Name                   : NVIDIACorp.NVIDIAControlPanel
     Publisher              : CN=D6816951-877F-493B-B4EE-41AB9419C326
     Architecture           : X64
     ResourceId             :
     Version                : 8.1.964.0
     PackageFullName        :
      NVIDIACorp.NVIDIAControlPanel_8.1.964.0_x64__56jybvy8sckqj
     InstallLocation        : C:\Program Files\WindowsApps
     \NVIDIACorp.NVIDIAControlPanel_8.1.964.0_x64__56jybvy8sckqj
     IsFramework            : False
     PackageFamilyName      : NVIDIACorp.NVIDIAControlPanel_56jybvy8sckqj
     PublisherId            : 56jybvy8sckqj
     PackageUserInformation :
      {S-1-12-1-530092550-1307989247-1105462437-500 [Administrator]: Installed,

      S-1-12-1-530092550-1307989247-1105462437-1002 [pliny]: Installed,

      S-1-12-1-530092550-1307989247-1105462437-1003 [trajan]: Installed}
     IsResourcePackage      : False
     IsBundle               : False
     IsDevelopmentMode      : False
     NonRemovable           : False
     IsPartiallyStaged      : False
     SignatureKind          : Store
     Status                 : Ok
     ```

Preventing this Issue

If your system does not allow the installation apps from the **Microsoft Store**, download and run the standalone **NVIDIA Control Panel** installer that is available from NVIDIA Licensing Portal. For instructions, refer to *Virtual GPU Software User Guide*.

If your system can allow the installation apps from the **Microsoft Store**, ensure that:

- The Microsoft Store app is enabled.
- Installation of Microsoft Store apps is not blocked by your system settings.
- No local or group policies are set to block Microsoft Store apps.

Workaround

If the **NVIDIA Control Panel** app is missing, install it separately from the graphics driver by downloading and running the standalone **NVIDIA Control Panel** installer that is available from NVIDIA Licensing Portal. For instructions, refer to *Virtual GPU Software User Guide*.

If the issue persists, contact NVIDIA Enterprise Support for further assistance.

Status

Open

Ref.

3999308

5.7. **NVIDIA Control Panel** crashes if a user session is disconnected and reconnected

Description

On all supported Windows Server guest OS releases, **NVIDIA Control Panel** crashes if a user session is disconnected and then reconnected while **NVIDIA Control Panel** is open.

Version

This issue affects all supported Windows Server guest OS releases.

Status

Open

Ref.

4086605

5.8. VM assigned multiple fractional vGPUs from the same GPU hangs

Description

A VM that has been assigned multiple fractional vGPUs from the same physical GPU hangs or becomes inaccessible during installation of the NVIDIA vGPU software graphics driver in the VM. This issue affects only GPUs based on the NVIDIA Turing and NVIDIA Volta GPU architectures. This issue does not occur if the VM has been assigned multiple fractional vGPUs from different physical GPUs.

Version

This issue affects only GPUs based on the NVIDIA Turing and NVIDIA Volta GPU architectures.

Status

Open

Ref.

4020171

5.9. CUDA profilers cannot gather hardware metrics on NVIDIA vGPU

Description

NVIDIA CUDA Toolkit profilers cannot gather hardware metrics on NVIDIA vGPU. This issue affects only traces that gather hardware metrics. Other traces are not affected by this issue and work normally.

Version

This issue affects NVIDIA vGPU software releases starting with 15.2.

Status

Open

Ref.

4041169

5.10. NVIDIA vGPU software graphics driver for Windows sends a remote call to `ngx.download.nvidia.com`

Description

After the NVIDIA vGPU software graphics for windows has been installed in the guest VM, the driver sends a remote call to `ngx.download.nvidia.com` to download and install additional components. Such a remote call might be a security issue.

Workaround

Before running the NVIDIA vGPU software graphics driver installer, disable the remote call to `ngx.download.nvidia.com` by setting the following Windows registry key:

```
[HKEY_LOCAL_MACHINE\SOFTWARE\NVIDIA Corporation\Global\NGXCore]
"EnableOTA"=dword:00000000
```

Note: If this Windows registry key is set to 1 or deleted, the remote call to `ngx.download.nvidia.com` is enabled again.

Status

Open

Ref.

4031840

5.11. Multiple RDP session reconnections on Windows Server 2022 can consume all frame buffer

Description

Multiple RDP session reconnections in a Windows Server 2022 guest VM can consume all the frame buffer of a vGPU or physical GPU. When this issue occurs, users' screens becomes black, their sessions are disconnected but left intact, and they cannot log on again. The following error message is written to the event log on the hypervisor host:

```
The Desktop Window Manager process has exited.
(Process exit code: 0xe0464645, Restart count: 1, Primary display device ID: )
```

Version

This issue affects only the Windows Server 2022 guest OS.

Workaround

Periodically restart the Windows Server 2022 guest VM to prevent all frame buffer from being consumed.

Status

Open

Ref.

3583766

5.12. VM with multiple legacy fractional vGPUs on the same GPU fails to boot

Description

A VM to which multiple legacy fractional vGPUs on the same physical GPU are assigned fails to boot. A fractional vGPU is assigned only a fraction of the physical GPU's frame buffer. A legacy NVIDIA vGPU does not support single root I/O virtualization (SR-IOV). When this issue occurs, error messages similar to the following examples are written to the `vmware.log` file on the hypervisor host:

```
2022-11-23T09:01:06.643Z In(05) vmx - VMIOP: Registered device 0000:da:00.0
...
2022-11-23T09:01:06.715Z In(05) vmx - VMIOP: Failed to register device 0000:da:00.0
 error = Failure
```

Status

Not an NVIDIA bug

Ref.

3879209

5.13. NLS client fails to acquire a license with the error `The allowed time to process response has expired`

Description

A licensed client of NVIDIA License System (NLS) fails to acquire a license with the error `The allowed time to process response has expired`. This error can affect clients of a Cloud License Service (CLS) instance or a Delegated License Service (DLS) instance.

This error occurs when the time difference between the system clocks on the client and the server that hosts the CLS or DLS instance is greater than 10 minutes. A common cause of this error is the failure of either the client or the server to adjust its system clock when daylight savings time begins or ends. The failure to acquire a license is expected to prevent clock windback from causing licensing errors.

Workaround

Ensure that system clock time of the client and any server that hosts a DLS instance match the current time in the time zone where they are located.

To prevent this error from occurring when daylight savings time begins or ends, enable the option to automatically adjust the system clock for daylight savings time:

- **Windows:** Set the **Adjust for daylight saving time automatically** option.
- **Linux:** Use the `hwclock` command.

Status

Not a bug

Ref.

3859889

5.14. With multiple active sessions, **NVIDIA Control Panel** incorrectly shows that the system is unlicensed

Description

In an environment with multiple active desktop sessions, the **Manage License** page of **NVIDIA Control Panel** shows that a licensed system is unlicensed. However, the `nvidia-smi` command and the management interface of the NVIDIA vGPU software license server correctly show that the system is licensed. When an active session is disconnected and reconnected, the **NVIDIA Display Container** service crashes.

The **Manage License** page incorrectly shows that the system is unlicensed because of stale data in **NVIDIA Control Panel** in an environment with multiple sessions. The data is stale because **NVIDIA Control Panel** fails to get and update the settings for remote sessions when multiple sessions or no sessions are active in the VM. The **NVIDIA Display Container** service crashes when a session is reconnected because the session is not active at the moment of reconnection.

Status

Open

Ref.

3761243

5.15. VP9 and AV1 decoding with web browsers are not supported on Microsoft Windows Server 2019

Description

VP9 and AV1 decoding with web browsers are not supported on Microsoft Windows Server 2019 and later supported releases. This issue occurs because starting with Windows Server 2019, the required codecs are not included with the OS and are not available through the **Microsoft Store** app. As a result, hardware decoding is not available for viewing YouTube videos or using collaboration tools such as Google Meet in a web browser.

Version

This issue affects Microsoft Windows Server releases starting with Windows Server 2019.

Status

Not an NVIDIA bug

Ref.

200756564

5.16. `nvidia-smi` ignores the second NVIDIA vGPU device added to a Microsoft Windows Server 2016 VM

Description

After a second NVIDIA vGPU device is added to a Microsoft Windows Server 2016 VM, the device does not appear in the output from the `nvidia-smi` command. This issue occurs only if the VM is already running NVIDIA vGPU software for the existing NVIDIA vGPU device when the second device is added to the VM.

The `nvidia-smi` command cannot retrieve the guest driver version, license status, and accounting mode of the second NVIDIA vGPU device.

```
nvidia-smi vgpu --query
GPU 00000000:37:00.0
    Active vGPUs                      : 1
    vGPU ID                           : 3251695793
        VM ID                         : 3575923
        VM Name                       : SVR-Reg-W(P)-KuIn
        vGPU Name                     : GRID V100D-32Q
        vGPU Type                     : 185
        vGPU UUID                     : 29097249-2359-11b2-8a5b-8e896866496b
        Guest Driver Version          : 551.78
        License Status                : Licensed
        Accounting Mode               : Disabled
...
GPU 00000000:86:00.0
    Active vGPUs                      : 1
    vGPU ID                           : 3251695797
        VM ID                         : 3575923
        VM Name                       : SVR-Reg-W(P)-KuIn
        vGPU Name                     : GRID V100D-32Q
        vGPU Type                     : 185
        vGPU UUID                     : 2926dd83-2359-11b2-8b13-5f22f0f74801
        Guest Driver Version          : Not Available
        License Status                : N/A
        Accounting Mode               : N/A
```

Version

This issue affects only VMs that are running Microsoft Windows Server 2016 as a guest OS.

Workaround

To avoid this issue, configure the guest VM with both NVIDIA vGPU devices **before** installing the NVIDIA vGPU software graphics driver.

If you encounter this issue after the VM is configured, use one of the following workarounds:

- Reinstall the NVIDIA vGPU software graphics driver.
- Forcibly uninstall the Microsoft Basic Display Adapter and reboot the VM.
- Upgrade the guest OS on the VM to Microsoft Windows Server 2019.

Status

Not an NVIDIA bug

Ref.

3562801

5.17. After an upgrade of the Linux graphics driver from an RPM package in a licensed VM, licensing fails

Description

After the NVIDIA vGPU software graphics driver for Linux is upgraded from an RPM package in a licensed VM, licensing fails. The `nvidia-smi vgpu -q` command shows the driver version and license status as N/A. Restarting the `nvidia-gridd` service fails with a `Unit not found` error.

Workaround

Perform a clean installation of the NVIDIA vGPU software graphics driver for Linux from an RPM package.

1. Remove the currently installed driver.
2. Install the new version of the driver.

```
$ rpm -iv nvidia-linux-grid-550_550.90.07_amd64.rpm
```

Status

Open

Ref.

3512766

5.18. After an upgrade of the Linux graphics driver from a Debian package, the driver is not loaded into the VM

Description

After the NVIDIA vGPU software graphics driver for Linux is upgraded from a Debian package, the driver is not loaded into the VM.

Workaround

Use one of the following workarounds to load the driver into the VM:

- Reboot the VM.
- Remove the `nvidia` module from the Linux kernel and reinsert it into the kernel.
  1. Remove the `nvidia` module from the Linux kernel.
     ```
     $ sudo rmmod nvidia
     ```
  2. Reinsert the `nvidia` module into the Linux kernel.
     ```
     $ sudo modprobe nvidia
     ```

Status

Not a bug

Ref.

200748806

5.19. Desktop session freezes when a VM is migrated to or from a host running an NVIDIA vGPU software 14 release

Description

When a VM configured with a Tesla V100 or Tesla T4 vGPU is migrated between a host running an NVIDIA vGPU software 14 release and a host running a an NVIDIA vGPU software 13 release, the remote desktop session freezes. After the session freezes, the VM must be rebooted to recover the session. This issue occurs only when the NVIDIA hardware-based H.264/HEVC video encoder (NVENC) is enabled.

Version

The issue affects migrations between a host running an NVIDIA vGPU software 14 release and a host running an NVIDIA vGPU software 13 release.

Workaround

Disable NVENC.

Status

Open

Ref.

3512790

5.20. Application or vGPU VM crashes when multiple application instances are launched

Description

When multiple application instances are launched on a legacy vGPU that is allocated only a fraction of the physical GPU's frame buffer, the application or VM to which the vGPU is assigned crashes. A legacy NVIDIA vGPU does not support single root I/O virtualization (SR-IOV). This issue does **not** affect NVIDIA vGPUs that support SR-IOV.

The symptoms of this issue depend on the release of VMware vSphere Hypervisor (ESXi).

- With VMware vSphere Hypervisor (ESXi) 7.0.3 and later releases, the application crashes but the guest VM remains accessible. When this issue occurs, the following error message is written to the `vmware.log` file:
  ```
  vmiop_log: (0x0): VGPU message 7 failed
  ```
- With VMware vSphere Hypervisor (ESXi) releases before 7.0.3, the guest VMX process crashes. When this issue occurs, the following error message is written to the `vmware.log` file in the host VMFS datastore folder for the VM:
  ```
  E105: PANIC: PhysMem: creating too many Global lookups.
  ```

This issue occurs when the plugin for legacy NVIDIA vGPUs creates more BAR1 mappings than VMware vSphere Hypervisor (ESXi) allows a VM to create. These mappings depend on the number and type of applications running in the VM.

Workaround

A workaround is available for the following GPUs, all of which have a large physical BAR1 memory size:

- Quadro RTX 6000 Passive
- Quadro RTX 8000 Passive

- Tesla V100 (all variants)

Note: This workaround is **not** available for other GPUs that are affected by this issue.

To employ this workaround, set the vGPU plugin parameter `pciPassthru0.cfg.plugin_managed_bar1_va_override` to 1.

Status

Open

Ref.

200680865

5.21. Only one vGPU VM can be powered on with VMware vSphere Hypervisor (ESXi) 7.0.3

Description

Only one VM configured with NVIDIA vGPU can be powered with VMware vSphere Hypervisor (ESXi) 7.0.3. Any attempt to power on a second VM fails with the following error message:

```
Insufficient resources. At least one device (pcipassthru0) required for VM vm-name
 is not available on host. host-name
```

This issue occurs because the release of VMware vCenter Server is incompatible with VMware vSphere Hypervisor (ESXi) 7.0.3. Only VMware vCenter Server 7.0.3 is compatible with VMware vSphere Hypervisor (ESXi) 7.0.3.

Version

VMware vSphere Hypervisor (ESXi) 7.0.3

Workaround

Upgrade VMware vCenter Server to release 7.0.3 to match the release of VMware vSphere Hypervisor (ESXi).

Status

Not an NVIDIA bug

Ref.

3419013

5.22. The reported NVENC frame rate is double the actual frame rate

Description

The frame rate in frames per second (FPS) for the NVIDIA hardware-based H.264/HEVC video encoder (NVENC) reported by the `nvidia-smi encodersessions` command and NVWMI is double the actual frame rate. Only the reported frame rate is incorrect. The actual encoding of frames is **not** affected.

This issue affects only Windows VMs that are configured with NVIDIA vGPU.

Status

Open

Ref.

2997564

5.23. VM fails after a second vGPU is assigned to it

Description

After a second vGPU is added to a VM and the VM is restarted, the VM fails. NVIDIA vGPU software supports up to a maximum of 16 vGPUs per VM on VMware vSphere Hypervisor (ESXi).

When this issue occurs, the following messages are written to the log file on the hypervisor host:

```
2021-09-27T17:11:42.303Z| vthread-2105551| | I005: vmiop_log: (0x0): Start restoring
 vGPU state ...
2021-09-27T17:11:43.465Z| vcpu-0| | E002: vmiop_log: (0x0): Deferred restore for
 RPCs cannot continue, since restore data was not saved
2021-09-27T17:11:43.465Z| vcpu-0| | E002: vmiop_log: (0x0): Deferred call for
 vmiopd_restore_rpc_data failed at un-stun!
2021-09-27T17:11:43.465Z| vcpu-0| | E002: vmiop_log: (0x0): Failed to complete
 restore for deferred functions.
2021-09-27T18:44:27.034Z| vthread-2105550| | E002: vmiop_log: (0x0): VGPU message 1
 failed, guest VGX version is already initialized...
```

```
2021-09-27T18:44:27.034Z| vthread-2105550| | E002: vmiop_log: (0x0): VGPU message 1
 failed, result code: 0x40
...
2021-09-27T18:44:35.359Z| vthread-2105550| | I005: vmiop_log: (0x0): Guest driver
 unloaded!
```

Workaround

To avoid this issue, create your VMs in EFI mode.

If you encounter this issue with a VM that was created in legacy BIOS mode, shut down and restart the VM or power off the VM and power it on again.

Status

Not an NVIDIA bug

Ref.

3386681

5.24. NVENC does not work with Teradici Cloud Access Software on Windows

Description

The NVIDIA hardware-based H.264/HEVC video encoder (NVENC) does not work with Teradici Cloud Access Software on Windows. This issue affects NVIDIA vGPU and GPU pass through deployments.

This issue occurs because the check that Teradici Cloud Access Software performs on the DLL signer name is case sensitive and NVIDIA recently changed the case of the company name in the signature certificate.

Status

Not an NVIDIA bug

This issue is resolved in the latest 21.07 and 21.03 Teradici Cloud Access Software releases.

Ref.

200749065

5.25. When a licensed client deployed by using VMware instant clone technology is destroyed, it does not return the license

Description

When a user logs out of a VM deployed by using VMware Horizon instant clone technology, the VM is deleted and OS is not shut down cleanly. The NVIDIA vGPU software license that was being used by the VM is not returned to the license server, which could cause the license server to run out of licenses.

Workaround

Deploy the instant-clone desktop pool with the following options:

- **Floating** user assignment
- **All Machines Up-Front** provisioning

This configuration will allow the MAC address to be reused on the newly cloned VMs.

For more information, refer to the documentation for the version of VMware Horizon that you are using:

- VMware Horizon 8: Worksheet for Creating an Instant-Clone Desktop Pool in Horizon Console
- VMware Horizon 7: Worksheet for Creating an Instant-Clone Desktop Pool in Horizon Console

Status

Not an NVIDIA bug

Ref.

200744338

5.26. A licensed client might fail to acquire a license if a proxy is set

Description

If a proxy is set with a system environment variable such as `HTTP_PROXY` or `HTTPS_PROXY`, a licensed client might fail to acquire a license.

Workaround

Perform this workaround on each affected licensed client.

1. Add the address of the NVIDIA vGPU software license server to the system environment variable `NO_PROXY`.

   The address must be specified exactly as it is specified in the client's license server settings either as a fully-qualified domain name or an IP address. If the `NO_PROXY` environment variable contains multiple entries, separate the entries with a comma (`,`).

   If high availability is configured for the license server, add the addresses of the primary license server and the secondary license server to the system environment variable `NO_PROXY`.
2. Restart the NVIDIA driver service that runs the core NVIDIA vGPU software logic.
   - On Windows, restart the **NVIDIA Display Container** service.
   - On Linux, restart the `nvidia-gridd` service.

Status

Closed

Ref.

200704733

5.27. Session connection fails with four 4K displays and NVENC enabled on a 2Q, 3Q, or 4Q vGPU

Description

Desktop session connections fail for a 2Q, 3Q, or 4Q vGPU that is configured with four 4K displays and for which the NVIDIA hardware-based H.264/HEVC video encoder (NVENC) is enabled. This issue affects only Teradici Cloud Access Software sessions on Linux guest VMs.

This issue is accompanied by the following error message:

```
This Desktop has no resources available or it has timed out
```

This issue is caused by insufficient frame buffer.

Workaround

Ensure that sufficient frame buffer is available for all the virtual displays that are connected to a vGPU by changing the configuration in one of the following ways:

- Reducing the number of virtual displays. The number of 4K displays supported with NVENC enabled depends on the vGPU.

vGPU	4K Displays Supported with NVENC Enabled
2Q	1
3Q	2
4Q	3

- Disabling NVENC. The number of 4K displays supported with NVENC disabled depends on the vGPU.

vGPU	4K Displays Supported with NVENC Disabled
2Q	2
3Q	2
4Q	4

- Using a vGPU type with more frame buffer. Four 4K displays with NVENC enabled on any Q-series vGPU with at least 6144 MB of frame buffer are supported.

Status

Not an NVIDIA bug

Ref.

200701959

5.28. Disconnected sessions cannot be reconnected or might be reconnected very slowly with NVWMI installed

Description

Disconnected sessions cannot be reconnected or might be reconnected very slowly when the NVIDIA Enterprise Management Toolkit (NVWMI) is installed. This issue affects Citrix Virtual Apps and Desktops and VMware Horizon sessions on Windows guest VMs.

Workaround

Uninstall NVWMI.

Status

Open

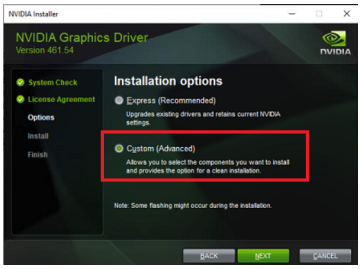
Ref.

3262923

5.29. Windows VM crashes during **Custom (Advanced)** driver upgrade

Description

When the NVIDIA vGPU software graphics driver in a Windows VM is upgraded with the **Custom (Advanced)** option selected, the VM crashes.



Status

Open

Ref.

200700291

5.30. VMs with vGPUs on GPUs based on the NVIDIA Ampere architecture fail to power on

Description

An otherwise correctly configured VMware vSphere ESXi 7.0 Update 2 server fails to boot VMs with vGPUs on GPUs based on the NVIDIA Ampere if the server being managed by a version of VMware vCenter Server older than 7.0.2. This version of VMware vCenter is released with ESXi 7.0 VMware vSphere Update 2.

When this issue occurs, the following error message is seen:

```
Insufficient resources. One or more devices (pciPassthru0) required by VM vm-name
 are not available on host host-name
```

Workaround

Use VMware vCenter Server 7.0.2 or a later compatible update

Status

Open

5.31. Linux VM hangs after vGPU migration to a host running a newer vGPU manager version

Description

When a Linux VM configured with a Tesla V100 or Tesla T4 vGPU is migrated from a host that is running a vGPU manager 11 release before 11.6 to a host that is running a vGPU manager 13 release, the VM hangs. After the migration, the destination host and VM become unstable. When this issue occurs, XID error 31 is written to the log files on the destination hypervisor host.

Version

This issue affects migration from a host that is running a vGPU manager 11 release before 11.6 to a host that is running a vGPU manager 13 release.

Workaround

If the VM is configured with a Tesla T4 vGPU, perform the following sequence of steps before attempting the migration:

1. Upgrade the host that is running a vGPU manager 11 release to release 11.6 or a later vGPU manager 11 release.
2. Disconnect any remoting tool that is using NVENC.

Note: You cannot use this workaround for a VM that is configured with a Tesla V100 vGPU.

Status

Open

Ref.

200691445

5.32. Idle Teradici Cloud Access Software session disconnects from Linux VM

Description

After a Teradici Cloud Access Software session has been idle for a short period of time, the session disconnects from the VM. When this issue occurs, the error messages `NVOS status 0x19` and `vGPU Message 21 failed` are written to the log files on the hypervisor host. This issue affects only Linux guest VMs.

Status

Open

Ref.

200689126

5.33. GPU Operator doesn't support vGPU on GPUs based on architectures before NVIDIA Turing

Description

NVIDIA GPU Operator doesn't support vGPU deployments on GPUs based on architectures before the NVIDIA Turing™ architecture. This issue is caused by the omission of version information for the vGPU manager from the configuration information that GPU Operator requires. Without this information, GPU Operator does not deploy the NVIDIA driver container because the container cannot determine if the driver is compatible with the vGPU manager.

Status

Open

Ref.

3227576

5.34. Idle NVIDIA A100, NVIDIA A40, and NVIDIA A10 GPUs show 100% GPU utilization

Description

The `nvidia-smi` command shows 100% GPU utilization for NVIDIA A100, NVIDIA A40, and NVIDIA A10 GPUs even if no vGPUs have been configured or no VMs are running.

```
[root@host ~]# nvidia-smi
Fri Jul 12 11:45:28 2024
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 550.90.05    Driver Version: 550.90.05   CUDA Version:  12.4     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  A100-PCIE-40GB      On   | 00000000:5E:00.0 Off |                    0 |
| N/A   50C    P0    97W / 250W |      0MiB / 40537MiB |    100%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+

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

Workaround

Boot any VMs that are configured with a vGPU that resides on the GPU.

After this workaround has been completed, the `nvidia-smi` command shows 0% GPU utilization for affected GPUs when they are idle.

```
root@host ~]# nvidia-smi
Fri Jul 12 11:47:38 2024
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 550.90.05    Driver Version: 550.90.05   CUDA Version:  12.4     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  A100-PCIE-40GB      On   | 00000000:5E:00.0 Off |                    0 |
| N/A   50C    P0    97W / 250W |      0MiB / 40537MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
```

```
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Status

Open

Ref.

200605527

5.35. Driver upgrade in a Linux guest VM with multiple vGPUs might fail

Description

Upgrading the NVIDIA vGPU software graphics driver in a Linux guest VM with multiple vGPUs might fail. This issue occurs if the driver is upgraded by overinstalling the new release of the driver on the current release of the driver while the `nvidia-gridd` service is running in the VM.

Workaround

1. Stop the `nvidia-gridd` service.
2. Try again to upgrade the driver.

Status

Open

Ref.

200633548

5.36. **NVIDIA Control Panel** fails to start if launched too soon from a VM without licensing information

Description

If NVIDIA licensing information is not configured on the system, any attempt to start **NVIDIA Control Panel** by right-clicking on the desktop within 30 seconds of the VM being started fails.

Workaround

Restart the VM and wait at least 30 seconds before trying to launch **NVIDIA Control Panel**.

Status

Open

Ref.

200623179

5.37. Citrix Virtual Apps and Desktops session corruption occurs in the form of residual window borders

Description

When a window is dragged across the desktop in a Citrix Virtual Apps and Desktops session, corruption of the session in the form of residual window borders occurs.

Version

This issue affects only Citrix Virtual Apps and Desktops version 7 2003

Workaround

Use Citrix Virtual Apps and Desktops version 7 1912 or 2006.

Status

Not an NVIDIA bug

Ref.

200608675

5.38. VMware Horizon clients cannot connect to a Windows 10 2004 VM with multiple displays

Description

Some VMware Horizon clients cannot connect to a Windows 10 2004 VM with multiple displays. When this issue occurs, the VM becomes unusable and clients cannot connect to the VM even if only a single display is connected to it.

This issue occurs because the desktop capture mechanism for the affected VMware Horizon clients is provided by NVIDIA® Frame Buffer Capture (NVFBC) and NVFBC is deprecated on Windows 10 starting with Windows 10 October 2019 Update. For more information, see NVFBC Windows 10 Support Deprecation Technical Bulletin (PDF).

Version

This issue affects only Windows 10 May 2020 Update (2004) guest VMs.

Workaround

Contact VMware to obtain a version of VMware Horizon for which the desktop capture mechanism is **not** provided by NVFBC.

Status

Not an NVIDIA bug

Ref.

200607827

5.39. Suspend and resume between hosts running different versions of the vGPU manager fails

Description

Suspending a VM configured with vGPU on a host running one version of the vGPU manager and resuming the VM on a host running a version from an older main release branch fails. For example, suspending a VM on a host that is running the

vGPU manager from release 17.3 and resuming the VM on a host running the vGPU manager from release 16.6 fails. When this issue occurs, the error `One or more devices (pciPassthru0) required by VM vm-name are not available on host host-name` is reported on VMware vCenter Server.

Status

Not an NVIDIA bug

Ref.

200602087

5.40. On Linux, a VMware Horizon 7.12 session freezes after a switch to full screen

Description

On a Linux VM configured with a -1Q vGPU, one 4K display, and VMware Horizon 7.12, the VMware Horizon session might become unresponsive after a switch from large screen (windowed) to full screen. When this issue occurs, the VMware vSphere VM's log file contains the error message `Unable to set requested topology`.

Version

This issue affects deployments that use VMware Horizon 7.12.

Workaround

Use VMware Horizon 7.11.

Status

Open

Ref.

200617112

5.41. On Linux, a VMware Horizon 7.12 session with two 4K displays freezes

Description

On a Linux VM configured with a -1Q vGPU, two 4K displays, and VMware Horizon 7.12, the VMware Horizon session might become unresponsive. When this issue occurs, the VMware vSphere VM's log file contains the error message `Failed to setup capture session (error 8). Unable to allocate video memory`.

Version

This issue affects deployments that use VMware Horizon 7.12.

Workaround

Use VMware Horizon 7.11 or a vGPU with more frame buffer.

Status

Open

Ref.

200617081

5.42. On Linux, the frame rate might drop to 1 after several minutes

Description

On Linux, the frame rate might drop to 1 frame per second (FPS) after NVIDIA vGPU software has been running for several minutes. Only some applications are affected, for example, `glxgears`. Other applications, such as Unigine Heaven, are not affected. This behavior occurs because Display Power Management Signaling (DPMS) for the Xorg server is enabled by default and the display is detected to be inactive even when the application is running. When DPMS is enabled, it enables power saving behavior of the display after several minutes of inactivity by setting the frame rate to 1 FPS.

Workaround

1. If necessary, stop the Xorg server.

```
# /etc/init.d/xorg stop
```

2. In a plain text editor, edit the `/etc/X11/xorg.conf file` to set the options to disable DPMS and disable the screen saver.

   a). In the `Monitor` section, set the `DPMS` option to `false`.

```
Option "DPMS" "false"
```

   b). At the end of the file, add a `ServerFlags` section that contains option to disable the screen saver.

```
Section "ServerFlags"
    Option "BlankTime" "0"
  EndSection
```

   c). Save your changes to `/etc/X11/xorg.conf file` and quit the editor.

3. Start the Xorg server.

```
# etc/init.d/xorg start
```

Status

Open

Ref.

200605900

5.43. Frame buffer consumption grows with VMware Horizon over Blast Extreme

Description

When VMware Horizon is used with the Blast Extreme display protocol, frame buffer consumption increases over time after multiple disconnections from and reconnections to a VM. This issue occurs even if the VM is in an idle state and no graphics applications are running.

Workaround

Reboot the VM.

Status

Not an NVIDIA bug

Ref.

200602520

5.44. DWM crashes randomly occur in Windows VMs

Description

Desktop Windows Manager (DWM) crashes randomly occur in Windows VMs, causing a blue-screen crash and the bug check `CRITICAL_PROCESS_DIED`. Computer Management shows problems with the primary display device.

Version

This issue affects Windows 10 1809, 1903 and 1909 VMs.

Status

Not an NVIDIA bug

Ref.

2730037

5.45. Remote desktop session freezes with assertion failure and XID error 43 after migration

Description

After multiple VMs configured with vGPU on a single hypervisor host are migrated simultaneously, the remote desktop session freezes with an assertion failure and XID error 43. This issue affects only GPUs that are based on the Volta GPU architecture. It does not occur if only a single VM is migrated.

When this error occurs, the following error messages are logged to the VMware vSphere Hypervisor (ESXi) log file:

```
Jan  3 14:35:48 ch81-m1 vgpu-12[8050]: error: vmiop_log: NVOS status 0x1f
Jan  3 14:35:48 ch81-m1 vgpu-12[8050]: error: vmiop_log: Assertion Failed at
 0x4b8cacf6:286
...
```

```
Jan  3 14:35:59 ch81-m1 vgpu-12[8050]: error: vmiop_log: (0x0): XID 43 detected on
 physical_chid:0x174, guest_chid:0x14
```

Status

Open

Ref.

200581703

5.46. Citrix Virtual Apps and Desktops session freezes when the desktop is unlocked

Description

When a Citrix Virtual Apps and Desktops session that is locked is unlocked by pressing **Ctrl**+**Alt**+**Del**, the session freezes. This issue affects only VMs that are running Microsoft Windows 10 1809 as a guest OS.

Version

Microsoft Windows 10 1809 guest OS

Workaround

Restart the VM.

Status

Not an NVIDIA bug

Ref.

2767012

5.47. NVIDIA vGPU software graphics driver fails after Linux kernel upgrade with DKMS enabled

Description

After the Linux kernel is upgraded (for example by running `sudo apt full-upgrade`) with Dynamic Kernel Module Support (DKMS) enabled, the `nvidia-smi` command fails to run. If DKMS is enabled, an upgrade to the Linux kernel triggers a rebuild of the NVIDIA vGPU software graphics driver. The rebuild of the driver fails because the compiler version is incorrect. Any attempt to reinstall the driver fails because the kernel fails to build.

When the failure occurs, the following messages are displayed:

```
-> Installing DKMS kernel module:
       ERROR: Failed to run `/usr/sbin/dkms build -m nvidia -v  550.54.14 -k
 5.3.0-28-generic`:
       Kernel preparation unnecessary for this kernel. Skipping...
       Building module:
       cleaning build area...
       'make' -j8 NV_EXCLUDE_BUILD_MODULES='' KERNEL_UNAME=5.3.0-28-generic
 IGNORE_CC_MISMATCH='' modules...(bad exit status: 2)
       ERROR (dkms apport): binary package for nvidia:  550.54.14 not found
       Error! Bad return status for module build on kernel: 5.3.0-28-generic
 (x86_64)
       Consult /var/lib/dkms/nvidia/ 550.54.14/build/make.log for more information.
       -> error.
       ERROR: Failed to install the kernel module through DKMS. No kernel module
 was installed;
       please try installing again without DKMS, or check the DKMS logs for more
 information.
       ERROR: Installation has failed. Please see the file '/var/log/nvidia-
installer.log' for details.
       You may find suggestions on fixing installation problems in the README
 available on the Linux driver download page at www.nvidia.com.
```

Workaround

When installing the NVIDIA vGPU software graphics driver with DKMS enabled, use one of the following workarounds:

- Before running the driver installer, install the `dkms` package, then run the driver installer with the `-dkms` option.
- Run the driver installer with the `--no-cc-version-check` option.

Status

Not a bug.

Ref.

2836271

5.48. Red Hat Enterprise Linux and CentOS 6 VMs hang during driver installation

Description

During installation of the NVIDIA vGPU software graphics driver in a Red Hat Enterprise Linux or CentOS 6 guest VM, a kernel panic occurs, and the VM hangs and cannot be rebooted. This issue is observed on older Linux kernels when the NVIDIA device is using message-signaled interrupts (MSIs).

Version

This issue affects the following guest OS releases:

- Red Hat Enterprise Linux 6.6 and later compatible 6.*x* versions
- CentOS 6.6 and later compatible 6.*x* versions

Workaround

1. Disable MSI in the guest VM to fall back to INTx interrupts by adding the following line to the file `/etc/modprobe.d/nvidia.conf`:

   ```
   options nvidia NVreg_EnableMSI=0
   ```

   If the file `/etc/modprobe.d/nvidia.conf` does not exist, create it.
2. Install the NVIDIA vGPU Software graphics driver in the guest VM.

Status

Closed

Ref.

200556896

5.49. Tesla T4 is enumerated as 32 separate GPUs by VMware vSphere ESXi

Description

Some servers, for example, the Dell R740, do not configure SR-IOV capability if the SR-IOV SBIOS setting is disabled on the server. If the SR-IOV SBIOS setting is disabled on such a server that is being used with the Tesla T4 GPU, VMware vSphere ESXi enumerates the Tesla T4 as 32 separate GPUs. In this state, you cannot use the GPU to configure a VM with NVIDIA vGPU or for GPU pass through.

Workaround

Ensure that the SR-IOV SBIOS setting is enabled on the server.

Status

Not an NVIDIA bug

A fix is available from VMware in VMware vSphere ESXi 7.0 Update 2.

Ref.

2697051

5.50. Users' sessions may freeze during vMotion migration of VMs configured with vGPU

Description

When vMotion is used to migrate a VM configured with vGPU to another host, users' sessions may freeze for up to several seconds during the migration.

These factors may increase the length of time for which a session freezes:

- Continuous use of the frame buffer by the workload, which typically occurs with workloads such as video streaming
- A large amount of vGPU frame buffer
- A large amount of system memory

- Limited network bandwidth

Workaround

Administrators can mitigate the effects on end users by avoiding migration of VMs configured with vGPU during business hours or warning end users that migration is about to start and that they may experience session freezes.

End users experiencing this issue must wait for their sessions to resume when the migration is complete.

Status

Open

Ref.

2569578

5.51. Migration of VMs configured with vGPU stops before the migration is complete

Description

When a VM configured with vGPU is migrated to another host, the migration stops before it is complete.

This issue occurs if the ECC memory configuration (enabled or disabled) on the source and destination hosts are different. The ECC memory configuration on both the source and destination hosts must be identical.

Workaround

Before attempting to migrate the VM again, ensure that the ECC memory configuration on both the source and destination hosts are identical.

Status

Not an NVIDIA bug

Ref.

200520027

5.52. ECC memory settings for a vGPU cannot be changed by using **NVIDIA X Server Settings**

Description

The ECC memory settings for a vGPU cannot be changed from a Linux guest VM by using **NVIDIA X Server Settings**. After the ECC memory state has been changed on the **ECC Settings** page and the VM has been rebooted, the ECC memory state remains unchanged.

Workaround

Use the `nvidia-smi` command in the guest VM to enable or disable ECC memory for the vGPU as explained in *Virtual GPU Software User Guide*.

If the ECC memory state remains unchanged even after you use the `nvidia-smi` command to change it, use the workaround in Changes to ECC memory settings for a Linux vGPU VM by nvidia-smi might be ignored.

Status

Open

Ref.

200523086

5.53. Changes to ECC memory settings for a Linux vGPU VM by `nvidia-smi` might be ignored

Description

After the ECC memory state for a Linux vGPU VM has been changed by using the `nvidia-smi` command and the VM has been rebooted, the ECC memory state might remain unchanged.

This issue occurs when multiple NVIDIA configuration files in the system cause the kernel module option for setting the ECC memory state `RMGuestECCState` in `/etc/modprobe.d/nvidia.conf` to be ignored.

When the `nvidia-smi` command is used to enable ECC memory, the file `/etc/modprobe.d/nvidia.conf` is created or updated to set the kernel module option `RMGuestECCState`. Another configuration file in `/etc/modprobe.d/` that contains the keyword `NVreg_RegistryDwordsPerDevice` might cause the kernel module option `RMGuestECCState` to be ignored.

Workaround

This workaround requires administrator privileges.

1. Move the entry containing the keyword `NVreg_RegistryDwordsPerDevice` from the other configuration file to `/etc/modprobe.d/nvidia.conf`.
2. Reboot the VM.

Status

Open

Ref.

200505777

5.54. Black screens observed when a VMware Horizon session is connected to four displays

Description

When a VMware Horizon session with Windows 7 is connected to four displays, a black screen is observed on one or more displays.

This issue occurs because a VMware Horizon session does not support connections to four 4K displays with Windows 7.

Status

Not an NVIDIA bug

Ref.

200503538

5.55. Host core CPU utilization is higher than expected for moderate workloads

Description

When GPU performance is being monitored, host core CPU utilization is higher than expected for moderate workloads. For example, host CPU utilization when only a small number of VMs are running is as high as when several times as many VMs are running.

Workaround

Disable monitoring of the following GPU performance statistics:

- vGPU engine usage by applications across multiple vGPUs
- Encoder session statistics
- Frame buffer capture (FBC) session statistics
- Statistics gathered by performance counters in guest VMs

Status

Open

Ref.

2414897

5.56. H.264 encoder falls back to software encoding on 1Q vGPUs with a 4K display

Description

On 1Q vGPUs with a 4K display, a shortage of frame buffer causes the H.264 encoder to fall back to software encoding.

Workaround

Use a 2Q or larger virtual GPU type to provide more frame buffer for each vGPU.

Status

Open

Ref.

2422580

5.57. H.264 encoder falls back to software encoding on 2Q vGPUs with 3 or more 4K displays

Description

On 2Q vGPUs with three or more 4K displays, a shortage of frame buffer causes the H.264 encoder to fall back to software encoding.

This issue affects only vGPUs assigned to VMs that are running a Linux guest OS.

Workaround

Use a 4Q or larger virtual GPU type to provide more frame buffer for each vGPU.

Status

Open

Ref.

200457177

5.58. Frame capture while the interactive logon message is displayed returns blank screen

Description

Because of a known limitation with NvFBC, a frame capture while the interactive logon message is displayed returns a blank screen.

An NvFBC session can capture screen updates that occur after the session is created. Before the logon message appears, there is no screen update after the message is shown

and, therefore, a black screen is returned instead. If the NvFBC session is created after this update has occurred, NvFBC cannot get a frame to capture.

Workaround

Press **Enter** or wait for the screen to update for NvFBC to capture the frame.

Status

Not a bug

Ref.

2115733

5.59. RDS sessions do not use the GPU with some Microsoft Windows Server releases

Description

When some releases of Windows Server are used as a guest OS, Remote Desktop Services (RDS) sessions do not use the GPU. With these releases, the RDS sessions by default use the Microsoft Basic Render Driver instead of the GPU. This default setting enables 2D DirectX applications such as Microsoft Office to use software rendering, which can be more efficient than using the GPU for rendering. However, as a result, 3D applications that use DirectX are prevented from using the GPU.

Version

- Windows Server 2019
- Windows Server 2016
- Windows Server 2012

Solution

Change the local computer policy to use the hardware graphics adapter for all RDS sessions.

1. Choose **Local Computer Policy** > **Computer Configuration** > **Administrative Templates** > **Windows Components** > **Remote Desktop Services** > **Remote Desktop Session Host** > **Remote Session Environment** .
2. Set the **Use the hardware default graphics adapter for all Remote Desktop Services sessions** option.

5.60. VMware vMotion fails gracefully under heavy load

Description

Migrating a VM configured with vGPU fails gracefully if the VM is running an intensive workload.

The error stack in the task details on the vSphere web client contains the following error message:

```
The migration has exceeded the maximum switchover time of 100 second(s).
ESX has preemptively failed the migration to allow the VM to continue running on the
 source.
To avoid this failure, either increase the maximum allowable switchover time or wait
 until
the VM is performing a less intensive workload.
```

Workaround

Increase the maximum switchover time by increasing the `vmotion.maxSwitchoverSeconds` option from the default value of 100 seconds.

For more information, see VMware Knowledge Base Article: vMotion or Storage vMotion of a VM fails with the error: The migration has exceeded the maximum switchover time of 100 second(s) (2141355).

Status

Not an NVIDIA bug

Ref.

200416700

5.61. View session freezes intermittently after a Linux VM acquires a license

Description

In a Linux VM, the view session can sometimes freeze after the VM acquires a license.

Workaround

Resize the view session.

Status

Not an NVIDIA bug

Ref.

200426961

5.62. When the scheduling policy is fixed share, GPU utilization is reported as higher than expected

Description

When the scheduling policy is fixed share, GPU engine utilization can be reported as higher than expected for a vGPU.

For example, GPU engine usage for six P40-4Q vGPUs on a Tesla P40 GPU might be reported as follows:

```
[root@localhost:~] nvidia-smi vgpu
Mon Aug 20 10:33:18 2018
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 390.42                 Driver Version: 390.42                    |
|-------------------------------+----------------------------------+----------+
| GPU  Name                     | Bus-Id                           | GPU-Util |
|      vGPU ID    Name          | VM ID    VM Name                 | vGPU-Util|
|===============================+==================================+==========|
|   0  Tesla P40                | 00000000:81:00.0                 |  99%     |
|      85109      GRID P40-4Q   | 85110    win7-xmpl-146048-1      |     32%  |
|      87195      GRID P40-4Q   | 87196    win7-xmpl-146048-2      |     39%  |
|      88095      GRID P40-4Q   | 88096    win7-xmpl-146048-3      |     26%  |
|      89170      GRID P40-4Q   | 89171    win7-xmpl-146048-4      |      0%  |
|      90475      GRID P40-4Q   | 90476    win7-xmpl-146048-5      |      0%  |
|      93363      GRID P40-4Q   | 93364    win7-xmpl-146048-6      |      0%  |
+-------------------------------+----------------------------------+----------+
|   1  Tesla P40                | 00000000:85:00.0                 |   0%     |
+-------------------------------+----------------------------------+----------+
```

The vGPU utilization of vGPU 85109 is reported as 32%. For vGPU 87195, vGPU utilization is reported as 39%. And for 88095, it is reported as 26%. However, the expected vGPU utilization of any vGPU should not exceed approximately 16.7%.

This behavior is a result of the mechanism that is used to measure GPU engine utilization.

Status

Open

Ref.

2227591

5.63. `nvidia-smi` reports that vGPU migration is supported on all hypervisors

Description

The command `nvidia-smi vgpu -m` shows that vGPU migration is supported on all hypervisors, even hypervisors or hypervisor versions that do not support vGPU migration.

Status

Closed

Ref.

200407230

5.64. `GPU resources not available` error during VMware instant clone provisioning

Description

A `GPU resources not available` error might occur during VMware instant clone provisioning. On Windows VMs, a `Video TDR failure - NVLDDMKM.sys` error causes a blue screen crash.

This error occurs when options for VMware Virtual Shared Graphics Acceleration (vSGA) are set for a VM that is configured with NVIDIA vGPU. VMware vSGA is a feature of VMware vSphere that enables multiple virtual machines to share the physical GPUs on ESXi hosts. NVIDIA vGPU software does **not** support VMware vSGA.

Depending on the combination of options set, one of the following error messages is seen when the VM is powered on:

- `Module 'MKS' power on failed.`

  This message is seen when the following options are set:

- **Enable 3D support** is selected.
- **3D Renderer** is set to **Hardware**
- The graphics type of all GPUs on the ESXi host is Shared Direct.

- `Hardware GPU resources are not available. The virtual machine will use software rendering.`

  This message is seen when the following options are set:

  - **Enable 3D support** is selected.
  - **3D Renderer** is set to **Automatic**.
  - The graphics type of all GPUs on the ESXi host is Shared Direct.

Resolution

If you want to use NVIDIA vGPU, unset any options for VMware vSGA that are set for the VM.

1. Ensure that the VM is powered off.
2. Open the vCenter Web UI.
3. In the vCenter Web UI, right-click the VM and choose **Edit Settings**.
4. Click the **Virtual Hardware** tab.
5. In the device list, expand the **Video card** node and de-select the **Enable 3D support** option.
6. Start the VM.

Status

Not a bug

Ref.

2369683

5.65. Module load failed during VIB downgrade from R390 to R384

Description

Some registry keys are available only with the R390 Virtual GPU Manager, for example, `NVreg_IgnoreMMIOCheck`. If any keys that are available only with the R390 Virtual GPU Manager are set, the NVIDIA module fails to load after a downgrade from R390 to R384.

When `nvidia-smi` is run without any arguments to verify the installation, the following error message is displayed:

```
NVIDIA-SMI has failed because it couldn't communicate with the NVIDIA driver. Make
 sure that the latest NVIDIA driver is installed and running.
```

Workaround

Before uninstalling the R390 VIB, clear all parameters of the `nvidia` module to remove any registry keys that are available only for the R390 Virtual GPU Manager.

```
# esxcli system module parameters set -p "" -m nvidia
```

Status

Not an NVIDIA bug

Ref.

200366884

5.66. On Linux, 3D applications run slowly when windows are dragged

Description

When windows for 3D applications on Linux are dragged, the frame rate drops substantially and the application runs slowly.

This issue does not affect 2D applications.

Status

Open

Ref.

1949482

5.67. A segmentation fault in DBus code causes `nvidia-gridd` to exit on Red Hat Enterprise Linux and CentOS

Description

On Red Hat Enterprise Linux 6.8 and 6.9, and CentOS 6.8 and 6.9, a segmentation fault in DBus code causes the `nvidia-gridd` service to exit.

The `nvidia-gridd` service uses DBus for communication with **NVIDIA X Server Settings** to display licensing information through the **Manage License** page. Disabling the GUI for licensing resolves this issue.

To prevent this issue, the GUI for licensing is disabled by default. You might encounter this issue if you have enabled the GUI for licensing and are using Red Hat Enterprise Linux 6.8 or 6.9, or CentOS 6.8 and 6.9.

Version

Red Hat Enterprise Linux 6.8 and 6.9

CentOS 6.8 and 6.9

Status

Open

Ref.

- 200358191
- 200319854
- 1895945

5.68. No **Manage License** option available in **NVIDIA X Server Settings** by default

Description

By default, the **Manage License** option is not available in **NVIDIA X Server Settings**. This option is missing because the GUI for licensing on Linux is disabled by default to work around the issue that is described in A segmentation fault in DBus code causes nvidia-gridd to exit on Red Hat Enterprise Linux and CentOS.

Workaround

This workaround requires `sudo` privileges.

Note: Do **not** use this workaround with Red Hat Enterprise Linux 6.8 and 6.9 or CentOS 6.8 and 6.9. To prevent a segmentation fault in DBus code from causing the `nvidia-gridd` service from exiting, the GUI for licensing must be disabled with these OS versions.

If you are licensing a physical GPU for vCS, you **must** use the configuration file `/etc/nvidia/gridd.conf`.

1. If **NVIDIA X Server Settings** is running, shut it down.
2. If the `/etc/nvidia/gridd.conf` file does not already exist, create it by copying the supplied template file `/etc/nvidia/gridd.conf.template`.
3. As root, edit the `/etc/nvidia/gridd.conf` file to set the `EnableUI` option to `TRUE`.
4. Start the `nvidia-gridd` service.

```
# sudo service nvidia-gridd start
```

When **NVIDIA X Server Settings** is restarted, the **Manage License** option is now available.

Status

Open

5.69. Licenses remain checked out when VMs are forcibly powered off

Description

NVIDIA vGPU software licenses remain checked out on the license server when non-persistent VMs are forcibly powered off.

The NVIDIA service running in a VM returns checked out licenses when the VM is shut down. In environments where non-persistent licensed VMs are not cleanly shut down, licenses on the license server can become exhausted. For example, this issue can occur in automated test environments where VMs are frequently changing and are not guaranteed to be cleanly shut down. The licenses from such VMs remain checked out against their MAC address for seven days before they time out and become available to other VMs.

Resolution

If VMs are routinely being powered off without clean shutdown in your environment, you can avoid this issue by shortening the license borrow period. To shorten the license borrow period, set the `LicenseInterval` configuration setting in your VM image. For details, refer to *Virtual GPU Client Licensing User Guide*.

Status

Closed

Ref.

1694975

5.70. Memory exhaustion can occur with vGPU profiles that have 512 Mbytes or less of frame buffer

Description

Memory exhaustion can occur with vGPU profiles that have 512 Mbytes or less of frame buffer.

This issue typically occurs in the following situations:

- Full screen 1080p video content is playing in a browser. In this situation, the session hangs and session reconnection fails.
- Multiple display heads are used with Citrix Virtual Apps and Desktops or VMware Horizon on a Windows 10 guest VM.
- Higher resolution monitors are used.
- Applications that are frame-buffer intensive are used.
- NVENC is in use.

To reduce the possibility of memory exhaustion, NVENC is disabled on profiles that have 512 Mbytes or less of frame buffer.

When memory exhaustion occurs, the NVIDIA host driver reports Xid error 31 and Xid error 43 in the VMware vSphere log file `vmware.log` in the guest VM's storage directory.

The following vGPU profiles have 512 Mbytes or less of frame buffer:

- Tesla M10-0B
- Tesla M10-0Q

The root cause is a known issue associated with changes to the way that recent Microsoft operating systems handle and allow access to overprovisioning messages and errors. If your systems are provisioned with enough frame buffer to support your use cases, you should not encounter these issues.

Workaround

- Use an appropriately sized vGPU to ensure that the frame buffer supplied to a VM through the vGPU is adequate for your workloads.
- Monitor your frame buffer usage.
- If you are using Windows 10, consider these workarounds and solutions:
  - Use a profile that has 1 Gbyte of frame buffer.
  - Optimize your Windows 10 resource usage.

    To obtain information about best practices for improved user experience using Windows 10 in virtual environments, complete the NVIDIA GRID vGPU Profile Sizing Guide for Windows 10 download request form.

    Additionally, you can use the VMware OS Optimization Tool to make and apply optimization recommendations for Windows 10 and other operating systems.

Status

Open

Ref.

- 200130864

- 1803861

5.71. vGPU VM fails to boot in ESXi if the graphics type is Shared

Description

On VMware vSphere Hypervisor (ESXi), after vGPU is configured, VMs to which a vGPU is assigned may fail to start and the following error message may be displayed:

```
The amount of graphics resource available in the parent resource pool is
 insufficient for the operation.
```

The vGPU Manager VIB provides vSGA and vGPU functionality in a single VIB. After this VIB is installed, the default graphics type is Shared, which provides vSGA functionality. To enable vGPU support for VMs in VMware vSphere, you must change the default graphics type to Shared Direct. If you do not change the default graphics type you will encounter this issue.

Note: NVIDIA vGPU software does **not** support VMware vSGA.

Workaround

Change the default graphics type to Shared Direct as explained in *Virtual GPU Software User Guide*.

Status

Open

Ref.

200256224

5.72. GNOME Display Manager (GDM) fails to start on Red Hat Enterprise Linux 7.2 and CentOS 7.0

Description

GDM fails to start on Red Hat Enterprise Linux 7.2 and CentOS 7.0 with the following error:

```
Oh no! Something has gone wrong!
```

Workaround

Permanently enable permissive mode for Security Enhanced Linux (SELinux).

1. As root, edit the `/etc/selinux/config` file to set `SELINUX` to `permissive`.
   ```
   SELINUX=permissive
   ```
2. Reboot the system.
   ```
   ~]# reboot
   ```

For more information, see Permissive Mode in *Red Hat Enterprise Linux 7 SELinux User's and Administrator's Guide*.

Status

Not an NVIDIA bug

Ref.

200167868

5.73. NVIDIA Control Panel fails to start and reports that “you are not currently using a display that is attached to an Nvidia GPU”

Description

When you launch NVIDIA Control Panel on a VM configured with vGPU, it fails to start and reports that you are not using a display attached to an NVIDIA GPU. This happens because Windows is using VMware’s SVGA device instead of NVIDIA vGPU.

Fix

Make NVIDIA vGPU the primary display adapter.

Use Windows screen resolution control panel to make the second display, identified as “2” and corresponding to NVIDIA vGPU, to be the active display and select the Show desktop only on 2 option. Click Apply to accept the configuration.

You may need to click on the Detect button for Windows to recognize the display connected to NVIDIA vGPU.

Note: If the VMware Horizon/View agent is installed in the VM, the NVIDIA GPU is automatically selected in preference to the SVGA device.

Status

Open

Ref.

5.74. VM configured with more than one vGPU fails to initialize vGPU when booted

Description

Using the current VMware vCenter user interface, it is possible to configure a VM with more than one vGPU device. When booted, the VM boots in VMware SVGA mode and doesn't load the NVIDIA driver. The additional vGPU devices are present in Windows Device Manager but display a warning sign, and the following device status:

```
Windows has stopped this device because it has reported problems. (Code 43)
```

Workaround

NVIDIA vGPU currently supports a single virtual GPU device per VM. Remove any additional vGPUs from the VM configuration before booting the VM.

Status

Open

Ref.

5.75. A VM configured with both a vGPU and a passthrough GPU fails to start the passthrough GPU

Description

Using the current VMware vCenter user interface, it is possible to configure a VM with a vGPU device and a passthrough (direct path) GPU device. This is not a currently supported configuration for vGPU. The passthrough GPU appears in Windows Device Manager with a warning sign, and the following device status:

```
Windows has stopped this device because it has reported problems. (Code 43)
```

Workaround

Do not assign vGPU and passthrough GPUs to a VM simultaneously.

Status

Open

Ref.

1735002

5.76. vGPU allocation policy fails when multiple VMs are started simultaneously

Description

If multiple VMs are started simultaneously, vSphere may not adhere to the placement policy currently in effect. For example, if the default placement policy (breadth-first) is in effect, and 4 physical GPUs are available with no resident vGPUs, then starting 4 VMs simultaneously should result in one vGPU on each GPU. In practice, more than one vGPU may end up resident on a GPU.

Workaround

Start VMs individually.

Status

Not an NVIDIA bug

Ref.

200042690

5.77. Before Horizon agent is installed inside a VM, the Start menu's sleep option is available

Description

When a VM is configured with a vGPU, the **Sleep** option remains available in the **Windows Start** menu. Sleep is not supported on vGPU and attempts to use it will lead to undefined behavior.

Workaround

Do not use Sleep with vGPU.

Installing the VMware Horizon agent will disable the **Sleep** option.

Status

Closed

Ref.

200043405

5.78. vGPU-enabled VMs fail to start, `nvidia-smi` fails when VMs are configured with too high a proportion of the server's memory.

Description

If vGPU-enabled VMs are assigned too high a proportion of the server's total memory, the following errors occur:

- One or more of the VMs may fail to start with the following error:

```
The available Memory resources in the parent resource pool are insufficient for
 the operation
```

- When run in the host shell, the `nvidia-smi` utility returns this error:

```
-sh: can't fork
```

For example, on a server configured with 256G of memory, these errors may occur if vGPU-enabled VMs are assigned more than 243G of memory.

Workaround

Reduce the total amount of system memory assigned to the VMs.

Status

Closed

Ref.

200060499

5.79. On reset or restart VMs fail to start with the error `VMIOP: no graphics device is available for vGPU...`

Description

On a system running a maximal configuration, that is, with the maximum number of vGPU VMs the server can support, some VMs might fail to start post a reset or restart operation.

Fix

Upgrade to ESXi 6.0 Update 1.

Status

Closed

Ref.

200097546

5.80. `nvidia-smi` shows high GPU utilization for vGPU VMs with active Horizon sessions

Description

vGPU VMs with an active Horizon connection utilize a high percentage of the GPU on the ESXi host. The GPU utilization remains high for the duration of the Horizon session even if there are no active applications running on the VM.

Workaround

None

Status

Open

Partially resolved for Horizon 7.0.1:

- For Blast connections, GPU utilization is no longer high.
- For PCoIP connections, utilization remains high.

Ref.

1735009

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA GRID, NVIDIA GRID vGPU, NVIDIA Maxwell, NVIDIA Pascal, NVIDIA Turing, NVIDIA Volta, GPUDirect, Quadro, and Tesla are trademarks or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com