Virtual GPU Software R470 for VMware vSphere

Release Notes

RN-07347-001 _v13.0 through 13.4 | July 2022

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
13.4	470.141.05	473.81	470.141.03
13.3	470.129.04	473.47	470.129.06
13.2	470.103.02	472.98	470.103.01
13.1	470.82	472.39	470.82.01
13.0	470.63	471.68	470.63.01

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: This requirement does not apply to the NVIDIA vGPU software license server. All releases in this release family of NVIDIA vGPU software are compatible with **all** releases of the 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
- NVIDIA vGPU Manager from a later long-term support branch with guest VM drivers from the previous long-term support 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 13.4 is used with guest drivers from release 11.2, the combination does **not** support Red Hat Enterprise Linux 7.6 because NVIDIA vGPU software release 13.4 does not support Red Hat Enterprise Linux 7.6.

The following table lists the specific software releases that are compatible with the components in the NVIDIA vGPU software 13 major release branch.

NVIDIA vGPU Software Component	Releases	Compatible Software Releases
NVIDIA vGPU Manager	13.0 through 13.4	▶ Guest VM driver releases 13.0 through 13.4 ▶ All guest VM driver 12.*x* releases ▶ All guest VM driver 11.*x* releases
Guest VM drivers	13.0 through 13.4	NVIDIA vGPU Manager releases 13.0 through 13.4

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 13 major release branch.

NVIDIA vGPU Software Component	Releases	Incompatible Software Releases
NVIDIA vGPU Manager	13.0 through 13.4	All guest VM driver releases 10.*x* and earlier
Guest VM drivers	13.0 through 13.4	All NVIDIA vGPU Manager releases 12.*x* and earlier

1.3. Updates in Release 13.4

New Features in Release 13.4

- Security updates - see *Security Bulletin: NVIDIA GPU Display Driver - August 2022*, 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 13.4

- Support for VMware Horizon 2206 (8.6)

Feature Support Withdrawn in Release 13.4

- VMware vSphere Hypervisor (ESXi) 6.7 and 6.5 are no longer supported.

1.4. Updates in Release 13.3

New Features in Release 13.3

- Security updates - see *Security Bulletin: NVIDIA GPU Display Driver - May 2022*, 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 13.3

- Support for Red Hat Enterprise Linux 8.6 as a guest OS
- Support for VMware Horizon 2203 (8.5)

Feature Support Withdrawn in Release 13.3

- Red Hat Enterprise Linux 8.5 and 8.2 are no longer supported as a guest OS.

1.5. Updates in Release 13.2

New Features in Release 13.2

- Security updates - see *Security Bulletin: NVIDIA GPU Display Driver - February 2022*, 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 13.2

- Support for Red Hat Enterprise Linux 8.5 as a guest OS
- Support for VMware Horizon 2111 (8.4)

Feature Support Withdrawn in Release 13.2

- Red Hat Enterprise Linux 8.1 is no longer supported as a guest OS.
- Red Hat Enterprise Linux 7.8 and 7.7 are no longer supported as a guest OS.

1.6. Updates in Release 13.1

New Features in Release 13.1

- Support for CUDA profilers on vGPUs on the following GPUs:
  - NVIDIA A40
  - NVIDIA A16
  - NVIDIA A10
  - NVIDIA RTX A6000
  - NVIDIA RTX A5000
- NVIDIA Deep Learning Super Sampling (DLSS) support on NVIDIA RTX Virtual Workstation
- Support for vSphere Lifecycle Management (vLCM)
- Support for virtual PCIe topology discovery (requires VMware vSphere ESXi 7.0.3 or later)

  The topology daemon generates virtual device topology mapping in Linux VMs for use by NVIDIA communication libraries.
- Security updates - see *Security Bulletin: NVIDIA GPU Display Driver - October 2021*, which is available on 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 13.1

- Support for Windows 11 21H2 as a guest OS

1.7. Updates in Release 13.0

New Features in Release 13.0

- Support for the following NVIDIA CUDA Toolkit features on NVIDIA vGPU:
  - Development tools such as IDEs, debuggers, profilers, and utilities as listed under *CUDA Toolkit Major Components* in *NVIDIA CUDA Toolkit Release Notes for CUDA 11.4*
  - Tracing and profiling through the CUDA Profiling Tools Interface (CUPTI)
- Compatibility with the guest VM drivers from the previous long-term support branch (11)
- NVIDIA License System support
- Miscellaneous bug fixes

Hardware and Software Support Introduced in Release 13.0

- Support for the following GPUs:
  - NVIDIA A16
- Support for Windows Server 2022 as a guest OS
- Support for VMware Horizon 2106 (8.3)

Feature Support Withdrawn in Release 13.0

- The following GPUs are no longer supported:
  - NVIDIA A100 HGX 80GB
  - NVIDIA A100 PCIe 40GB
  - NVIDIA A100 HGX 40GB

  Instead, these GPUs are supported with NVIDIA AI Enterprise.
- NVIDIA Virtual Compute Server (vCS) is no longer supported and C-series vGPU types are no longer available. Instead, vCS is supported with NVIDIA AI Enterprise.
- The base VMware vSphere 7.0 release and VMware vSphere 7.0 Update 1 are no longer supported.

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 R470 drivers.

2.1. Supported NVIDIA GPUs and Validated Server Platforms

This release of NVIDIA vGPU software provides support for the following NVIDIA GPUs on VMware vSphere, running on validated server hardware platforms:

- GPUs based on the NVIDIA Maxwell™ graphic architecture:
  - Tesla M6 (NVIDIA Virtual Compute Server (vCS) is **not** supported.)
  - Tesla M10 (vCS is **not** supported.)
  - Tesla M60 (vCS is **not** supported.)
- GPUs based on the NVIDIA Pascal™ architecture:
  - Tesla P4
  - Tesla P6
  - Tesla P40
  - Tesla P100 PCIe 16 GB (vSGA, vMotion with vGPU, and suspend-resume with vGPU are **not** supported.)
  - Tesla P100 SXM2 16 GB (vSGA, vMotion with vGPU, and suspend-resume with vGPU are **not** supported.)
  - Tesla P100 PCIe 12GB (vSGA, vMotion with vGPU, and suspend-resume with vGPU are **not** supported.)
- GPUs based on the NVIDIA Volta architecture:
  - Tesla V100 SXM2 (vSGA is **not** supported.)
  - Tesla V100 SXM2 32GB (vSGA is **not** supported.)
  - Tesla V100 PCIe (vSGA is **not** supported.)

- Tesla V100 PCIe 32GB (vSGA is **not** supported.)
- Tesla V100S PCIe 32GB (vSGA is **not** supported.)
- Tesla V100 FHHL (vSGA is **not** supported.)

- GPUs based on the NVIDIA Turing™ architecture:
  - Tesla T4 (vSGA is **not** supported.)
  - Quadro RTX 6000 in displayless mode (vSGA is **not** supported.)
  - Quadro RTX 6000 passive in displayless mode (vSGA is **not** supported.)
  - Quadro RTX 8000 in displayless mode (vSGA is **not** supported.)
  - Quadro RTX 8000 passive in displayless mode (vSGA is **not** supported.)

  In displayless mode, local physical display connectors are disabled.
- GPUs based on the NVIDIA Ampere architecture:
  - NVIDIA A40 in displayless mode (vSGA is **not** supported.)
  - NVIDIA A16 (vSGA is **not** supported.)
  - NVIDIA A10 (vSGA is **not** supported.)
  - NVIDIA RTX A6000 in displayless mode (vSGA is **not** supported.)
  - NVIDIA RTX A5000 in displayless mode (vSGA is **not** supported.)

  In displayless mode, local physical display connectors are disabled.

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

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

Some GPUs support displayless and display-enabled modes but must be used in NVIDIA vGPU software deployments in displayless 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 displayless mode, but other GPUs are supplied in a display-enabled mode.

GPU	Mode as Supplied from the Factory
NVIDIA A40	Displayless
NVIDIA RTX A5000	Display enabled
NVIDIA RTX A6000	Display enabled

A GPU that is supplied from the factory in displayless 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 following GPUs support the `displaymodeselector` tool:

- NVIDIA A40
- NVIDIA RTX A5000
- NVIDIA RTX A6000

Other GPUs that support NVIDIA vGPU software do not support the `displaymodeselector` tool and, unless otherwise stated, do not require display mode switching.

2.1.2. Switching the Mode of a Tesla M60 or M6 GPU

Tesla M60 and M6 GPUs support compute mode and graphics mode. NVIDIA vGPU requires GPUs that support both modes to operate in graphics mode.

Recent Tesla M60 GPUs and M6 GPUs are supplied in graphics mode. However, your GPU might be in compute mode if it is an older Tesla M60 GPU or M6 GPU or if its mode has previously been changed.

To configure the mode of Tesla M60 and M6 GPUs, use the `gpumodeswitch` tool provided with NVIDIA vGPU software releases. If you are unsure which mode your GPU is in, use the `gpumodeswitch` tool to find out the mode.

Note:

Only Tesla M60 and M6 GPUs support the `gpumodeswitch` tool. Other GPUs that support NVIDIA vGPU do not support the `gpumodeswitch` tool and, except as stated in Switching the Mode of a GPU that Supports Multiple Display Modes, do not require mode switching.

Even in compute mode, Tesla M60 and M6 GPUs do **not** support NVIDIA Virtual Compute Server vGPU types.

For more information, refer to *gpumodeswitch User Guide*.

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)

- VMs with 32 GB or more of RAM fail to boot with GPUs requiring 64 GB or more of MMIO space

With ESXi 6.7 or later, 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 A6000	128 GB
Quadro RTX 6000 Passive	64 GB
Quadro RTX 8000 Passive	64 GB
Tesla P6	64 GB
Tesla P40	64 GB
Tesla P100 (all variants)	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)
  - Tesla P100 (all variants)
  - Tesla P6
- 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)
  - Tesla P40
- 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.

- The Tesla V100, Tesla P100, and Tesla P6 require ESXi 6.0 Update 1 and later, or ESXi 6.5 and later.
- Because it requires 64 GB of MMIO space, the Tesla P40 requires ESXi 6.0 Update 3 and later, or ESXi 6.5 and later.

  As a result, the VM's MMIO space must be increased to 64 GB as explained in VMware Knowledge Base Article: VMware vSphere VMDirectPath I/O: Requirements for Platforms and Devices (2142307).

2.1.5. 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 Enterprise Plus Edition of VMware vSphere Hypervisor (ESXi). 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) 7.0	7.0 Update 2 and later compatible updates **Note:** The base VMware vSphere Hypervisor (ESXi) 7.0 release and	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
	7.0 Update 1 are **not** supported.	
13.0-13.3 only: VMware vSphere Hypervisor (ESXi) 6.7	6.7 and compatible updates NVIDIA vGPU support requires VMware ESXi 6.7, Patch Release ESXi670-202011002, build 17167734 or later from VMware.	The following GPUs are supported in GPU pass through mode **only**: ▶ NVIDIA RTX A6000 ▶ NVIDIA RTX A5000 ▶ NVIDIA A40 ▶ NVIDIA A16 ▶ NVIDIA A10 Starting with release 6.7 U3, the assignment of multiple vGPUs to a single VM is supported. Starting with release 6.7 U1, vMotion with vGPU and suspend and resume with vGPU are supported on suitable GPUs as listed in Supported NVIDIA GPUs and Validated Server Platforms. Release 6.7 supports only suspend and resume with vGPU. vMotion with vGPU is **not** supported on release 6.7.
13.0-13.3 only: VMware vSphere Hypervisor (ESXi) 6.5	6.5 and compatible updates NVIDIA vGPU support requires VMware ESXi 6.5, Patch Release ESXi650-202102001, build 17477841 or later from VMware.	The following GPUs are supported in GPU pass through mode **only**: ▶ NVIDIA RTX A6000 ▶ NVIDIA RTX A5000 ▶ NVIDIA A40 ▶ NVIDIA A16 ▶ NVIDIA A10 The following features of NVIDIA vGPU software are **not** supported.

Software	Release Supported	Notes
		▶ Assignment of multiple vGPUs to a single VM ▶ Suspend-resume with vGPU ▶ vMotion with vGPU ▶ Live VMware snapshots with vGPU

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	**Since 13.4:** 2206 (8.6) and compatible updates **Since 13.3:** 2203 (8.5) and compatible updates **Since 13.2:** 2111 (8.4) and compatible updates 2106 (8.3) and compatible updates 2103 (8.2) and compatible updates 2012 (8.1) and compatible updates 2006 (8.0) and compatible updates 7.13 and compatible 7.13.*x* updates 7.12 and compatible 7.12.*x* updates 7.11 and compatible 7.11.*x* updates 7.10 and compatible 7.10.*x* updates 7.9 and compatible 7.9.*x* updates 7.8 and compatible 7.8.*x* updates 7.7 and compatible 7.7.*x* updates 7.6 and compatible 7.6.*x* updates 7.5 and compatible 7.5.*x* updates 7.4 and compatible 7.4.*x* updates

Software	Releases Supported
	7.3 and compatible 7.3.*x* updates 7.2 and compatible 7.2.*x* updates 7.1 and compatible 7.1.*x* updates 7.0 and compatible 7.0.*x* updates
VMware vCenter Server	7.0 Update 2 and later compatible updates 6.7 and compatible updates 6.5 and compatible updates

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 in the table 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

Guest OS	NVIDIA vGPU - VMware vSphere Releases	Pass-Through GPU - VMware vSphere Releases
Windows Server 2022	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7 update 3	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7 update 3

Guest OS	NVIDIA vGPU - VMware vSphere Releases	Pass-Through GPU - VMware vSphere Releases
Windows Server 2019	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5 update 2, 6.5 update 1	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5 update 2, 6.5 update 1
Windows Server 2016 1709, 1607	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5
Windows Server 2012 R2 (not supported on GPUs based on architectures after the NVIDIA Turing™ architecture)	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5
Since 13.1: Windows 11 21H2	7.0	7.0
Since 13.1: Windows 10 November 2021 Update (21H2) and all Windows 10 releases supported by Microsoft up to and including this release See Note (1)	**Since 13.4:** 7.0 **13.1-13.3 only:** 7.0, 6.7, 6.5	**Since 13.4:** 7.0 **13.1-13.3 only:** 7.0, 6.7, 6.5
13.0 only: Windows 10 May 2021 Update (21H1) and all Windows 10 releases supported by Microsoft up to and including this release See Note (1)	7.0, 6.7, 6.5	7.0, 6.7, 6.5

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 in the table 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

Guest OS	NVIDIA vGPU - VMware vSphere Releases	Pass-Through GPU - VMware vSphere Releases
Red Hat CoreOS 4.7	7.0 update 1	7.0 update 1
Since 13.3: Red Hat Enterprise Linux 8.6	**Since 13.4:** 7.0 **13.3 only:** 7.0, 6.7, 6.5 update 3	**Since 13.4:** 7.0 **13.3 only:** 7.0, 6.7, 6.5 update 3
13.2 only: Red Hat Enterprise Linux 8.5	7.0, 6.7, 6.5 update 3	7.0, 6.7, 6.5 update 3
Red Hat Enterprise Linux 8.4	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5 update 3	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5 update 3
13.0-13.2 only: Red Hat Enterprise Linux 8.2	7.0, 6.7, 6.5 update 3	7.0, 6.7, 6.5 update 3
13.0, 13.1 only: Red Hat Enterprise Linux 8.1	7.0, 6.7, 6.5 update 3	7.0, 6.7, 6.5 update 3
Since 13.2: CentOS Linux 8 (2111)	**Since 13.4:** 7.0 **13.2, 13.3 only:** 7.0, 6.7, 6.5 update 3	**Since 13.4:** 7.0 **13.2, 13.3 only:** 7.0, 6.7, 6.5 update 3
CentOS Linux 8 (2105)	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5 update 3	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5 update 3
13.0-13.2 only: CentOS Linux 8 (2004)	7.0, 6.7, 6.5 update 3	7.0, 6.7, 6.5 update 3
13.0, 13.1 only: CentOS Linux 8 (1911)	7.0, 6.7, 6.5 update 3	7.0, 6.7, 6.5 update 3
Since 13.2: Red Hat Enterprise Linux 7.9 and later compatible 7.x versions	**Since 13.4:** 7.0 **13.2, 13.3 only:** 7.0, 6.7, 6.5 update 3	**Since 13.4:** 7.0 **13.2, 13.3 only:** 7.0, 6.7, 6.5 update 3
13.0, 13.1 only: Red Hat Enterprise Linux 7.7-7.9 and later compatible 7.x versions	7.0, 6.7, 6.5	7.0, 6.7, 6.5
CentOS 7.6-7.8 and later compatible 7.x versions	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5
Ubuntu 20.04 LTS	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7 update 1	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7 update 1
Ubuntu 18.04 LTS	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5
Ubuntu 16.04 LTS	**Since 13.4:** 7.0	**Since 13.4:** 7.0

Guest OS	NVIDIA vGPU - VMware vSphere Releases	Pass-Through GPU - VMware vSphere Releases
	13.0-13.3 only: 7.0, 6.7, 6.5	**13.0-13.3 only:** 7.0, 6.7, 6.5
Ubuntu 14.04 LTS	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5
SUSE Linux Enterprise Server 15 SP2	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5
SUSE Linux Enterprise Server 12 SP3	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5	**Since 13.4:** 7.0 **13.0-13.3 only:** 7.0, 6.7, 6.5

2.4. NVIDIA CUDA Toolkit Version Support

The releases in this release family of NVIDIA vGPU software support NVIDIA CUDA Toolkit 11.4.

For more information about NVIDIA CUDA Toolkit, see CUDA Toolkit 11.4 Documentation.

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 only on a subset of supported GPUs, VMware vSphere Hypervisor (ESXi) releases, and guest operating systems.

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

- Unified memory
- Debuggers
- Profilers

Supported GPUs

- Tesla M6
- Tesla M10
- Tesla M60
- Tesla P4
- Tesla P6
- Tesla P40
- Tesla V100 SXM2
- Tesla V100 SXM2 32GB
- Tesla V100 PCIe
- Tesla V100 PCIe 32GB
- Tesla V100S PCIe 32GB
- Tesla V100 FHHL
- Tesla T4
- Quadro RTX 6000
- Quadro RTX 6000 passive
- Quadro RTX 8000
- Quadro RTX 8000 passive
- NVIDIA A10
- NVIDIA A16
- NVIDIA A40
- NVIDIA RTX A5000
- NVIDIA RTX A6000

Supported VMware vSphere Hypervisor (ESXi) Releases

- Release 7.0 Update 2 and compatible updates support vMotion with vGPU and suspend-resume with vGPU.
- **13.0-13.3 only:** Release 6.7 U1 and compatible updates support vMotion with vGPU and suspend-resume with vGPU.
- **13.0-13.3 only:** Release 6.7 supports only suspend-resume with vGPU.
- Releases earlier than 6.7 do not support any form of vGPU migration.

Supported Guest OS Releases

Windows and Linux.

Known Issues with vGPU Migration Support

Use Case	Affected GPUs	Issue
Migration to or from a host running an NVIDIA vGPU software 11 release **before 11.6**	Tesla T4	VM hangs after vGPU migration from a host running a vGPU manager 11 release to a host running a vGPU manager 13 release
▶ Migration **from** a host running NVIDIA vGPU software 11.3 **to** a host running a different release ▶ Migration **to** a host running NVIDIA vGPU software 11.3 **from** a host running a different release	Tesla T4	Migrating a VM with a Tesla T4 vGPU between a host running NVIDIA vGPU software 11.3 and a host running a different release fails
Migration from a host that is running a vGPU manager 11 release to a host that is running a vGPU manager 13 release.	▶ Tesla T4 ▶ Tesla V100	Linux VM hangs after vGPU migration to a host running a newer vGPU manager version
Migration to or from a host running an NVIDIA vGPU software 11 release	GPUs based on the NVIDIA Volta™ architecture	Desktop session freezes when a VM is migrated to or from a host running an NVIDIA vGPU software 11 release
Migration to or from a host running an NVIDIA vGPU software 14 release	▶ Tesla T4 ▶ Tesla V100	Desktop session freezes when a VM is migrated to or from a host running an NVIDIA vGPU software 14 release
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 VMware vSphere Hypervisor (ESXi) releases.

Supported vGPUs

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

GPU Architecture	Board	vGPU
Ampere	NVIDIA A40	A40-48Q See Note (1).
	NVIDIA A16	A16-16Q See Note (1).
	NVIDIA A10	A10-24Q See Note (1).
	NVIDIA RTX A6000	A6000-48Q See Note (1).
	NVIDIA RTX A5000	A5000-24Q See Note (1).
Turing	Tesla T4	T4-16Q
	Quadro RTX 6000	RTX6000-24Q
	Quadro RTX 6000 passive	RTX6000P-24Q
	Quadro RTX 8000	RTX8000-48Q
	Quadro RTX 8000 passive	RTX8000P-48Q
Volta	Tesla V100 SXM2 32GB	V100DX-32Q
	Tesla V100 PCIe 32GB	V100D-32Q
	Tesla V100S PCIe 32GB	V100S-32Q
	Tesla V100 SXM2	V100X-16Q
	Tesla V100 PCIe	V100-16Q
	Tesla V100 FHHL	V100L-16Q
Pascal	Tesla P100 SXM2	P100X-16Q
	Tesla P100 PCIe 16GB	P100-16Q
	Tesla P100 PCIe 12GB	P100C-12Q
	Tesla P40	P40-24Q
	Tesla P6	P6-16Q
	Tesla P4	P4-8Q
Maxwell	Tesla M60	M60-8Q
	Tesla M10	M10-8Q

GPU Architecture	Board	vGPU
	Tesla M6	M6-8Q

Note:

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

Maximum vGPUs per VM

NVIDIA vGPU software supports up to a maximum of four vGPUs per VM on VMware vSphere Hypervisor (ESXi).

Supported Hypervisor Releases

VMware vSphere Hypervisor (ESXi) 7.0 and 6.7 U3 and later compatible updates only.

If you upgraded to VMware vSphere 6.7 Update 3 from an earlier version and are using VMs that were created with that version, change the VM compatibility to **vSphere 6.7 Update 2 and later**. For details, see Virtual Machine Compatibility in the VMware documentation.

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.

Supported vGPUs

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

GPU Architecture	Board	vGPU
Ampere	NVIDIA A40	A40-48Q
	NVIDIA A10	A10-24Q
	NVIDIA RTX A6000	A6000-48Q
	NVIDIA RTX A5000	A5000-24Q
Turing	Quadro RTX 6000	RTX6000-24Q
	Quadro RTX 6000 passive	RTX6000P-24Q
	Quadro RTX 8000	RTX8000-48Q

GPU Architecture	Board	vGPU
	Quadro RTX 8000 passive	RTX8000P-48Q
Volta	Tesla V100 SXM2 32GB	V100DX-32Q
	Tesla V100 SXM2	V100X-16Q
Pascal	Tesla P100 SXM2	P100X-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.

Supported Hypervisor Releases

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.

Supported Guest OS Releases

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

Limitations

- Only direct connections are supported. NVSwitch is not 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.

Supported vGPUs

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

GPU Architecture	Board	vGPU
Ampere	NVIDIA A40	A40-48Q
	NVIDIA A16	A16-16Q
	NVIDIA A10	A10-24Q
	NVIDIA RTX A6000	A6000-48Q
	NVIDIA RTX A5000	A5000-24Q

Supported Guest OS Releases

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

Limitations

- When unified memory is enabled for a VM, vGPU migration is disabled for the VM.
- When unified memory is enabled for a VM, NVIDIA CUDA Toolkit profilers are disabled.

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 VMware vSphere Hypervisor (ESXi) release, container platform, and guest OS release.

VMware vSphere Hypervisor (ESXi) 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 4.9 with Red Hat Enterprise Linux CoreOS and the CRI-O container runtime	Red Hat CoreOS 4.9
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. Since 13.1: 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 A40
- NVIDIA A16
- NVIDIA A10
- NVIDIA RTX A6000
- 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. Since 13.1: 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 compatible updates
- 6.7 and compatible updates
- 6.5 and compatible updates

Supported VMware vCenter Server releases: 7.0 Update 2 and later compatible updates

Chapter 3. Known Product Limitations

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

3.1. 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.

3.2. 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.

3.3. 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.

3.4. 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

vGPU Type	Compression Adjustment (MB)
RTX8000-48Q RTX8000-48C RTX8000-48A	238
RTX8000P-12Q RTX8000P-12C RTX8000P-12A	32
RTX8000P-16Q RTX8000P-16C RTX8000P-16A	64
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.

3.5. 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.

3.6. 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.

3.7. 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 M6-0B, M6-0Q
- Tesla M10-0B, M10-0Q
- Tesla M60-0B, M60-0Q

Workaround

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

3.8. 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 M6-0B, M6-0Q
- Tesla M10-0B, M10-0Q
- Tesla M60-0B, M60-0Q

Workaround

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

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

Description

Support for vGPU and vSGA 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: Tesla M6, Tesla M10, and Tesla M60.

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.

3.10. 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.

3.11. 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.

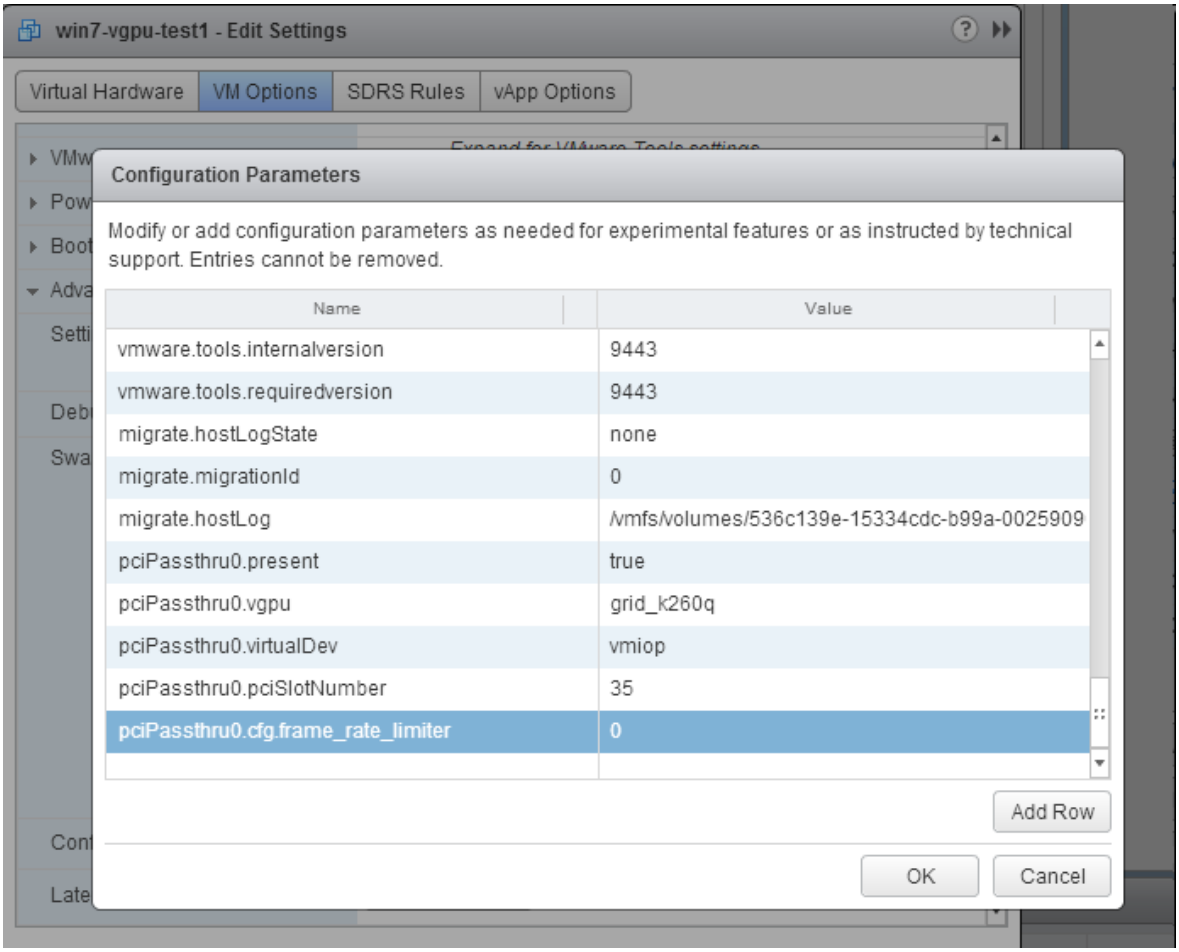
FRL is used to ensure balanced performance across multiple vGPUs that are resident on the same physical GPU. The FRL setting is designed to give good interactive remote graphics experience but may reduce scores in benchmarks that depend on measuring frame rendering rates, as compared to the same benchmarks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

3.12. 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)
```

The VMware vSphere VM's log file contains these error messages:

```
vthread10|E105: NVOS status 0x29
```

```
vthread10|E105: Assertion Failed at 0x7620fd4b:179
vthread10|E105: 8 frames returned by backtrace
 ...
vthread10|E105: VGPU message 12 failed, result code: 0x29
...
vthread10|E105: NVOS status 0x8
vthread10|E105: Assertion Failed at 0x7620c8df:280
vthread10|E105: 8 frames returned by backtrace
...
vthread10|E105: VGPU message 26 failed, result code: 0x8
```

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**.

Issues Resolved in Release 13.4

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

Issues Resolved in Release 13.3

Bug ID	Summary and Description
200756399	**13.0-13.2 Only: Linux VM might fail to return a license after shutdown if the license server is specified by its name** If the license server is specified by its fully qualified domain name, a Linux VM might fail to return its license when the VM is shut down. This issue occurs if the `nvidia-gridd` service cannot resolve the fully qualified domain name of the license server because `systemd-resolved.service` is not available when the service attempts to return the license. When this issue occurs, the `nvidia-gridd` service writes the following message to the `systemd` journal: `General data transfer failure. Couldn't resolve host name`
3465448	**13.0-13.2 Only: Windows vGPU VM sometimes crashes after guest OS upgrade** When a VM that is configured with NVIDIA vGPU is rebooted after an OS upgrade from Windows 10 1909 to Windows 10 20H2, the VM sometimes crashes. This issue is caused by a `NULL` pointer exception in the Virtual GPU Manager plugin (`libnvidia-vgx.so`). This `NULL` pointer exception might also cause the VM to crash in other situations. When this issue occurs, error messages that indicate that the Virtual GPU Manager process crashed are written to the log file `vmware.log` on the hypervisor host.
200724807	**13.0-13.2 Only: Memory leaks in the vGPU manager plugin cause the VM to hang**

Bug ID	Summary and Description
	Applications running in a VM request memory to be allocated and freed by the vGPU manager plugin, which runs on the hypervisor host. When an application requests the vGPU manager plugin to free previously allocated memory, some of the memory is not freed. Some applications request memory more frequently than other applications. If such applications run for a long period of time, for example for two or more days, the failure to free all allocated memory might cause the hypervisor host to run out of memory. As a result, memory allocation for applications running in the VM might fail, causing the applications and, sometimes, the VM to hang.

Issues Resolved in Release 13.2

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

Issues Resolved in Release 13.1

Bug ID	Summary and Description
3392680	**13.0 Only: Windows 2012 R2 licensed clients cannot acquire licenses from a CLS or DLS instance** NVIDIA vGPU software licensed clients running in a Windows 2012 R2 VM cannot acquire licenses from a Cloud License Service (CLS) instance or a Delegated License Service (DLS) instance. During the license acquisition process, the vGPU licensing service compares the size of the message to be sent to the CLS or DLS instance before and after encryption. On Windows 2012 R2, the MSDN API for encrypting the message outputs a shorter encrypted message than the plain text message. As a result, the validation check in the service fails, which causes the client to fail to acquire a license.

Issues Resolved in Release 13.0

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

Chapter 5. Known Issues

5.1. 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. 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

5.2. 13.0-13.2 Only: Linux VM might fail to return a license after shutdown if the license server is specified by its name

Description

If the license server is specified by its fully qualified domain name, a Linux VM might fail to return its license when the VM is shut down. This issue occurs if the `nvidia-gridd` service cannot resolve the fully qualified domain name of the license server because `systemd-resolved.service` is not available when the service attempts to return the license. When this issue occurs, the `nvidia-gridd` service writes the following message to the `systemd` journal:

```
General data transfer failure. Couldn't resolve host name
```

Status

Resolved in NVIDIA vGPU software 13.3

Ref.

200756399

5.3. **NVIDIA Control Panel** is started only for the RDP user that logs on first

Description

On all supported Windows Server guest OS releases, **NVIDIA Control Panel** is started only for the RDP user that logs on first. Other users cannot start **NVIDIA Control Panel**. If more than one RDP user is logged on when **NVIDIA Control Panel** is started, it always opens in the session of the RDP user that logged on first, irrespective of which user started **NVIDIA Control Panel**. Furthermore, on Windows Server 2016, **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.

3334310

5.4. 13.0-13.2 Only: Windows vGPU VM sometimes crashes after guest OS upgrade

Description

When a VM that is configured with NVIDIA vGPU is rebooted after an OS upgrade from Windows 10 1909 to Windows 10 20H2, the VM sometimes crashes. This issue is caused by a `NULL` pointer exception in the Virtual GPU Manager plugin (`libnvidia-vgx.so`). This `NULL` pointer exception might also cause the VM to crash in other situations. When this issue occurs, error messages that indicate that the Virtual GPU Manager process crashed are written to the log file `vmware.log` on the hypervisor host.

Status

Resolved in NVIDIA vGPU software 13.3

Ref.

3465448

5.5. 13.0-13.2 Only: Memory leaks in the vGPU manager plugin cause the VM to hang

Description

Applications running in a VM request memory to be allocated and freed by the vGPU manager plugin, which runs on the hypervisor host. When an application requests the vGPU manager plugin to free previously allocated memory, some of the memory is not freed. Some applications request memory more frequently than other applications. If such applications run for a long period of time, for example for two or more days, the failure to free all allocated

memory might cause the hypervisor host to run out of memory. As a result, memory allocation for applications running in the VM might fail, causing the applications and, sometimes, the VM to hang.

When memory allocation fails, the error messages that are written to the log file on the hypervisor host depend on the hypervisor.

- For VMware vSphere ESXi, the following error messages are written to `vmware.log`:

```
2021-10-05T04:57:35.547Z| vthread-2329002| E110: vmiop_log: Fail to create the
 buffer for translate pte rpc node

2021-06-05T10:48:33.007Z| vcpu-3| E105: PANIC: Unrecoverable memory allocation
 failure
```

- For Citrix Hypervisor and hypervisors based on Linux KVM, the following messages are written to the standard activity log in the `/var/log` directory (`/var/log/messages` or `/var/log/syslog`):

```
Feb 15 09:27:48 bkrzxen1 kernel: [1278743.170072] Out of memory: Kill process
 20464 (vgpu) score 9 or sacrifice child

Feb 15 09:27:48 bkrzxen1 kernel: [1278743.170111] Killed process 20464 (vgpu)
 total-vm:305288kB, anon-rss:56508kB, file-rss:30828kB, shmem-rss:0kB

Feb 15 09:27:48 bkrzxen1 kernel: [1278743.190484] oom_reaper: reaped process
 20464 (vgpu), now anon-rss:0kB, file-rss:27748kB, shmem-rss:4kB".
```

Workaround

If an application or a VM hangs after a long period of usage, restart the VM every couple of days to prevent the hypervisor host from running out of memory.

Status

Resolved in NVIDIA vGPU software 13.3

Ref.

200724807

5.6. `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          : 473.47
        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

5.7. 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

5.8. 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.

Version

Since 13.3: 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 P6
- Tesla P40
- Tesla P100 (all variants)
- Tesla V100 (all variants)

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

Status

Open

Ref.

200680865

5.9. 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

5.10. 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

5.11. VM hangs after vGPU migration from a host running a vGPU manager 11 release to a host running a vGPU manager 13 release

Description

When a VM configured with a 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. 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**, such as 11.4 or 11.5, to a host that is running a vGPU manager 13 release, such as 13.0 or 13.1.

Workaround

Upgrade the host that is running a vGPU manager 11 release to release 11.6 before attempting the migration.

Status

Open

5.12. Windows 2012 R2 licensed clients cannot acquire licenses from a DLS instance

Description

NVIDIA vGPU software licensed clients running in a Windows 2012 R2 VM cannot acquire licenses from a Delegated License Service (DLS) instance. This issue occurs because the TLS handshake between the client VM and DLS instance is failing with schannel error code `0x80090326(SEC_E_ILLEGAL_MESSAGE)`, indicating that the client has encountered an unrecoverable error during the TLS handshake.

Workaround

Use the legacy NVIDIA vGPU software license server instead of NVIDIA License System (NLS).

Status

Open

Ref.

3400123

5.13. 13.0 Only: Windows 2012 R2 licensed clients cannot acquire licenses from a CLS or DLS instance

Description

NVIDIA vGPU software licensed clients running in a Windows 2012 R2 VM cannot acquire licenses from a Cloud License Service (CLS) instance or a Delegated License Service (DLS) instance. During the license acquisition process, the vGPU licensing service compares the size of the message to be sent to the CLS or DLS instance before and after encryption. On Windows 2012 R2, the MSDN API for encrypting the message outputs a shorter encrypted message than the plain text message. As a result, the validation check in the service fails, which causes the client to fail to acquire a license.

Status

Resolved in NVIDIA vGPU software 13.1

Ref.

3392680

5.14. 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 four 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

5.15. Desktop session freezes when a VM is migrated to or from a host running an NVIDIA vGPU software 11 release

Description

The remote desktop session freezes when a VM is migrated to or from a host running an NVIDIA vGPU software 11 release. This issue affects only VMs configured with vGPUs on GPUs based on the NVIDIA Volta™ architecture.

When this issue occurs, the following error messages are written to the VMware vSphere VM's log file on the destination host:

- `XID error 13`
- `XID error 43`

Version

The issue affects migrations to and from a host that is running an NVIDIA vGPU software 11 release.

Status

Open

Ref.

200707632

5.16. 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

5.17. 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

5.18. 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

5.19. 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

5.20. 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

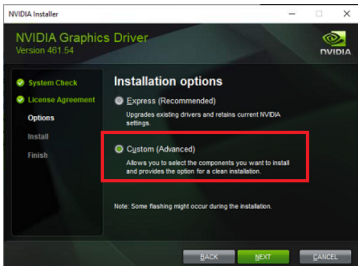
Ref.

3262923

5.21. 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

5.22. 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

5.23. Migrating a VM with a Tesla T4 vGPU between a host running NVIDIA vGPU software 11.3 and a host running a different release fails

Description

Migrating a VM with a Tesla T4 vGPU between hosts where one host is running NVIDIA vGPU software 11.3 and the other host is running a different release fails. After the migration, the destination host and VM become unstable.

When this issue occurs, error message `XID error 38` is written to the VMware vSphere VM's log file `vmware.log` in the guest VM's storage directory. Depending on the host configurations, the following messages might also be written to the log file:

- `XID error 43`
- `VGPU message 58`
- `VGPU message 4`

Version

This issue affects migrations between a host that is running NVIDIA vGPU software release 11.3 and a host that is running a different release. The issue affects migrations to and from the host that is running NVIDIA vGPU software release 11.3.

This issue does **not** affect migrations between two hosts that are both running NVIDIA vGPU software 11.3.

Workaround

If you are migrating a VM between a host that is running NVIDIA vGPU software 11.3 and a host that is running a another NVIDIA vGPU software 11 release, contact NVIDIA Enterprise Support for assistance.

Otherwise, avoid migrating a VM between a host that is running NVIDIA vGPU software 11.3 and a host that is running a different NVIDIA vGPU software release.

Status

Closed

Ref.

200691763

200735219

5.24. NVML fails to initialize with unknown error

Description

After NVIDIA Virtual GPU manager is installed on VMware vSphere Hypervisor (ESXi) 6.7, the `nvidia-smi` command fails with the error `Failed to initialize NVML: Unknown Error`.

Version

VMware vSphere Hypervisor (ESXi) 6.7

Workaround

Apply VMware ESXi 6.7, Patch Release ESXi670-202011002, build 17167734 or later from VMware.

Status

Not an NVIDIA bug

Ref.

3237970

5.25. 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

5.26. 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

5.27. 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

5.28. 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 Aug 12 11:45:28 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 470.141.05   Driver Version: 470.141.05   CUDA Version:  11.4     |
|-------------------------------+----------------------+----------------------+
| 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 Aug 12 11:47:38 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 470.141.05   Driver Version: 470.141.05   CUDA Version:  11.4     |
|-------------------------------+----------------------+----------------------+
| 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

5.29. 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

5.30. **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

5.31. 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

5.32. 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

5.33. 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 13.4 and resuming the VM on a host running the vGPU manager from release 12.4 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

5.34. 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

5.35. 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

5.36. 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

5.37. 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

5.38. 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

5.39. 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

5.40. 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

5.41. 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  470.63.01 -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:  470.63.01 not found
       Error! Bad return status for module build on kernel: 5.3.0-28-generic
(x86_64)
       Consult /var/lib/dkms/nvidia/ 470.63.01/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

5.42. 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

5.43. 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

5.44. VMware vCenter shows GPUs with no available GPU memory

Description

VMware vCenter shows some physical GPUs as having 0.0 B of available GPU memory. VMs that have been assigned vGPUs on the affected physical GPUs cannot be booted. The `nvidia-smi` command shows the same physical GPUs as having some GPU memory available.

Workaround

Stop and restart the Xorg service and `nv-hostengine` on the ESXi host.

1. Stop all running VM instances on the host.
2. Stop the Xorg service.
   ```
   [root@esxi:~] /etc/init.d/xorg stop
   ```
3. Stop `nv-hostengine`.
   ```
   [root@esxi:~] nv-hostengine -t
   ```
4. Wait for 1 second to allow `nv-hostengine` to stop.
5. Start `nv-hostengine`.
   ```
   [root@esxi:~] nv-hostengine -d
   ```
6. Start the Xorg service.
   ```
   [root@esxi:~] /etc/init.d/xorg start
   ```

Status

Not an NVIDIA bug

A fix is available from VMware in VMware vSphere ESXi 6.7 U3. For information about the availability of fixes for other releases of VMware vSphere ESXi, contact VMware.

Ref.

2644794

5.45. 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

5.46. 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. After the migration stops, the VM is no longer accessible.

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

Reboot the hypervisor host to recover the VM. 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

A fix is that prevents the VM from becoming inaccessible is available from VMware in VMware vSphere Hypervisor (ESXi) 6.7 Update 3 patch 16075168-04282020. Even with this patch, migration of a VM configured with vGPU requires the ECC memory configuration on both the source and destination hosts to be identical.

Ref.

200520027

5.47. 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

5.48. 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

5.49. 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

5.50. Quadro RTX 8000 and Quadro RTX 6000 GPUs can't be used with VMware vSphere ESXi 6.5

Description

Quadro RTX 8000 and Quadro RTX 6000 GPUs can't be used with VMware vSphere ESXi 6.5. If you attempt to use the Quadro RTX 8000 or Quadro RTX 6000 GPU with VMware vSphere ESXi 6.5, a purple-screen crash occurs after you install the NVIDIA Virtual GPU Manager.

Version

VMware vSphere ESXi 6.5

Workaround

Upgrade VMware vSphere ESXi to patch level ESXi 6.5 P04 (ESXi650-201912002, build 15256549) or later.

VMware resolved this issue in this patch for VMware vSphere ESXi.

Status

Not an NVIDIA bug

Ref.

200491080

5.51. 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

5.52. 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

5.53. 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

5.54. 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

5.55. 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.

5.56. 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

5.57. 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

5.58. 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

5.59. `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

5.60. `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 and can be used as an alternative to NVIDIA vGPU.

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

5.61. VMs with 32 GB or more of RAM fail to boot with GPUs requiring 64 GB or more of MMIO space

Description

VMs with 32 GB or more of RAM fail to boot with GPUs that require 64 GB or more of MMIO space. VMs boot successfully with RAM allocations of less than 32 GB.

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 A6000	128 GB
Quadro RTX 6000 Passive	64 GB
Quadro RTX 8000 Passive	64 GB
Tesla P6	64 GB
Tesla P40	64 GB
Tesla P100 (all variants)	64 GB
Tesla V100 (all variants)	64 GB

Version

This issue affects the following versions of VMware vSphere ESXi:

- 6.0 Update 3 and later updates
- 6.5 and later updates

Workaround

If you want to use a VM with 32 GB or more of RAM with GPUs that require 64 GB or more of MMIO space, use this workaround:

1. Create a VM to which less than 32 GB of RAM is allocated.
2. Choose **VM Options** > **Advanced** and set `pciPassthru.use64bitMMIO="TRUE"`.
3. Allocate the required amount of RAM to the VM.

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

Status

Not an NVIDIA bug

Resolved in VMware vSphere ESXi 6.7

Ref.

2043171

5.62. 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

5.63. Tesla P40 cannot be used in pass-through mode

Description

Pass-through mode on Tesla P40 GPUs and other GPUs based on the Pascal architecture does not work as expected. In some situations, after the VM is powered on, the guest OS crashes or fails to boot.

Workaround

Ensure that your GPUs are configured as described in Requirements for Using GPUs Requiring Large MMIO Space in Pass-Through Mode.

Status

Not a bug

Ref.

1944539

5.64. 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

5.65. 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

5.66. 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

5.67. 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

5.68. 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 M6-0B, M6-0Q
- Tesla M10-0B, M10-0Q
- Tesla M60-0B, M60-0Q

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

5.69. vGPU VM fails to boot in ESXi 6.5 if the graphics type is Shared

Description

Note: If vSGA is being used, this issue shouldn't be encountered and changing the default graphics type is not necessary.

On VMware vSphere Hypervisor (ESXi) 6.5, 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 6.5, you must change the default graphics type to Shared Direct. If you do not change the default graphics type you will encounter this issue.

Version

VMware vSphere Hypervisor (ESXi) 6.5

Workaround

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

Status

Open

Ref.

200256224

5.70. ESXi 6.5 web client shows high memory usage even when VMs are idle

Description

On VMware vSphere Hypervisor (ESXi) 6.5, the web client shows a memory usage alarm with critical severity for VMs to which a vGPU is attached even when the VMs are idle. When memory usage is monitored from inside the VM, no memory usage alarm is shown. The web client does not show a memory usage alarm for the same VMs without an attached vGPU.

Version

VMware vSphere Hypervisor (ESXi) 6.5

Workaround

Avoid using the VMware vSphere Hypervisor (ESXi) 6.5 web client to monitor memory usage for VMs to which a vGPU is attached.

Status

Not an NVIDIA bug

Ref.

200191065

5.71. NVIDIA driver installation may fail for VMs on a host in a VMware DRS cluster

Description

For VMware vSphere releases before 6.7 Update 1, the ESXi host on which VMs configured with NVIDIA vGPU reside must not be a member of an automated VMware Distributed Resource Scheduler (DRS) cluster. The installer for the NVIDIA driver for NVIDIA vGPU software cannot locate the NVIDIA vGPU software GPU card on a host in an automated VMware DRS Cluster. Any attempt to install the driver on a VM on a host in an automated DRS cluster fails with the following error:

```
NVIDIA Installer cannot continue
This graphics driver could not find compatible graphics hardware.
```

Note: This issue does not occur with VMs running VMware vSphere 6.7 Update 1 or later without load balancing support. For these releases, vSphere DRS supports automatic initial placement of VMs configured with NVIDIA vGPU.

Version

VMware vSphere Hypervisor (ESXi) releases **before** 6.7 Update 1.

Workaround

Ensure that the automation level of the DRS cluster is set to **Manual**.

For more information about this setting, see Edit Cluster Settings in the VMware documentation.

Status

Open

Ref.

1933449

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