Virtual GPU Software R580 for Red Hat Enterprise Linux with KVM

Release Notes

RN-08687-001 _v19.0 | August 2025

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 Red Hat Enterprise Linux with KVM.

> **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
19.0	580.65.05	580.88	580.65.06

For details of which Red Hat Enterprise Linux with KVM releases are supported, see Hypervisor Software Releases.

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

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

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

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

Compatible NVIDIA vGPU Manager and Guest VM Driver Releases

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

- NVIDIA vGPU Manager with guest VM drivers from the same release
- NVIDIA vGPU Manager 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 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 19.0 is used with guest drivers from release 16.4, the combination does **not** support Windows Server 2019 because NVIDIA vGPU software release 19.0 does not support Windows Server 2019.

The following table lists the specific software releases that are compatible with the components in the NVIDIA vGPU software 19 major release branch.

NVIDIA vGPU Software Component	Release	Compatible Software Releases
NVIDIA vGPU Manager	19.0	> Guest VM driver release 19.0 > All guest VM driver 18.*x* releases

NVIDIA vGPU Software Component	Release	Compatible Software Releases
		- All guest VM driver 16.*x* releases
Guest VM drivers	19.0	NVIDIA vGPU Manager release 19.0

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 19 major release branch.

NVIDIA vGPU Software Component	Release	Incompatible Software Releases
NVIDIA vGPU Manager	19.0	All guest VM driver releases 17.*x* and earlier, **except** 16.*x* releases
Guest VM drivers	19.0	All NVIDIA vGPU Manager releases 18.*x* and earlier

1.3. Updates in Release 19.0

New Features in Release 19.0

- Support for Multi-Instance GPU (MIG)-backed vGPUs for graphics on GPUs that support MIG
- Support for time-sliced, MIG-backed vGPUs within a GPU instance on a MIG-enabled GPU
- New B-series vGPU profiles with 3 GB of frame buffer on supported GPUs based on the NVIDIA Ada Lovelace and NVIDIA Blackwell GPU architectures
- Miscellaneous bug fixes

Newly Supported Hardware and Software in Release 19.0

- Newly supported graphics cards:
  - NVIDIA RTX PRO 6000 Blackwell Server Edition

Features Deprecated in Release 19.0

NVIDIA vGPU software 19 is the last release branch to support the following graphics cards:

- Tesla M10
- Tesla V100 SXM2
- Tesla V100 SXM2 32GB
- Tesla V100 PCIe
- Tesla V100 PCIe 32GB
- Tesla V100S PCIe 32GB
- Tesla V100 FHHL
- Quadro RTX 6000
- Quadro RTX 6000 passive
- Quadro RTX 8000
- Quadro RTX 8000 passive

Disabling strict round robin policy is deprecated and NVIDIA vGPU software 19 is the last release branch to support it. Support for this feature is planned to be removed in the next major release of NVIDIA vGPU software.

Chapter 2. Validated Platforms

This release family of NVIDIA vGPU software provides support for several NVIDIA GPUs on validated server hardware platforms, Red Hat Enterprise Linux with KVM hypervisor software versions, and guest operating systems. It also supports the version of NVIDIA CUDA Toolkit that is compatible with R580 drivers.

2.1. Supported NVIDIA GPUs and Validated Server Platforms

This release of NVIDIA vGPU software on Red Hat Enterprise Linux with KVM provides support for several NVIDIA GPUs running on validated server hardware platforms.

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

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

GPUs Based on the NVIDIA Blackwell Architecture

Note: GPUs based on the NVIDIA Blackwell architecture support the manual placement of vGPUs on GPUs in equal-size mode.

GPU	SR-IOV - Red Hat Enterprise Linux with KVM Releases	Mixed vGPU Configuration - Red Hat Enterprise Linux with KVM Releases		Supported NVIDIA vGPU Software Products[1, 2, 3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
NVIDIA RTX PRO 6000 Blackwell Server Edition	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps

GPUs Based on the NVIDIA Ada Lovelace Architecture

Note: GPUs based on the NVIDIA Ada Lovelace architecture support the manual placement of vGPUs on GPUs in equal-size mode.

GPU	SR-IOV - Red Hat Enterprise Linux with KVM Releases	Mixed vGPU Configuration - Red Hat Enterprise Linux with KVM Releases		Supported NVIDIA vGPU Software Products[1], [2], [3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
NVIDIA L40S	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	vWS vPC vApps	vWS vApps
NVIDIA L40	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	vWS vPC vApps	vWS vApps
NVIDIA L20	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	vWS vPC vApps	vWS vApps
NVIDIA L20 liquid cooled	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	vWS vPC vApps	vWS vApps
NVIDIA L4	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	vWS vPC vApps	vWS vApps
NVIDIA L2	9.6, 9.4	10.0	10.0	vWS vPC	vWS

GPU	SR-IOV - Red Hat Enterprise Linux with KVM Releases	Mixed vGPU Configuration - Red Hat Enterprise Linux with KVM Releases		Supported NVIDIA vGPU Software Products[1,2,3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
	8.10	9.6, 9.4 8.10	9.6, 9.4 8.10	> vApps	> vApps
NVIDIA RTX 6000 Ada	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
NVIDIA RTX 5880 Ada	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
NVIDIA RTX 5000 Ada	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps

GPUs Based on the NVIDIA Ampere Architecture

Note: The manual placement of vGPUs on GPUs in equal-size mode **is not** supported on GPUs based on the NVIDIA Ampere architecture.

GPU	SR-IOV - Red Hat Enterprise Linux with KVM Releases	Mixed vGPU Configuration - Red Hat Enterprise Linux with KVM Releases		Supported NVIDIA vGPU Software Products[1,2,3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
NVIDIA A40[4]	9.6, 9.4 8.10	10.0 9.6, 9.4	10.0 9.6, 9.4	> vWS > vPC > vApps	> vWS > vApps

GPU	SR-IOV - Red Hat Enterprise Linux with KVM Releases	Mixed vGPU Configuration - Red Hat Enterprise Linux with KVM Releases		Supported NVIDIA vGPU Software Products[1, 2, 3]	
		Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
		8.10	8.10		
NVIDIA A16	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
NVIDIA A10	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
NVIDIA A2	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
NVIDIA RTX A6000[4]	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
NVIDIA RTX A5500[4]	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
NVIDIA RTX A5000[4]	9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps

GPUs Based on the NVIDIA Turing Architecture

Note: SR-IOV and the manual placement of vGPUs on GPUs in equal-size mode are **not** supported on GPUs based on the NVIDIA Turing™ architecture.

	Mixed vGPU Configuration - Red Hat Enterprise Linux with KVM Releases		**Supported NVIDIA vGPU Software Products**[1,2,3]	
GPU	**Frame Buffer Size (Mixed-Size Mode)**	**Series**	**NVIDIA vGPU**	**GPU Pass Through**
Tesla T4	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
Quadro RTX 6000 [4]	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
Quadro RTX 6000 passive[4]	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
Quadro RTX 8000[4]	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
Quadro RTX 8000 passive[4]	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps

GPUs Based on the NVIDIA Volta Architecture

Note: SR-IOV and the manual placement of vGPUs on GPUs in equal-size mode are **not** supported on GPUs based on the NVIDIA Volta architecture.

GPU	Mixed vGPU Configuration - Red Hat Enterprise Linux with KVM Releases		Supported NVIDIA vGPU Software Products[1, 2, 3]	
	Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
Tesla V100 SXM2	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
Tesla V100 SXM2 32GB	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
Tesla V100 PCIe	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
Tesla V100 PCIe 32GB	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
Tesla V100S PCIe 32GB	10.0 9.6, 9.4 8.10	10.0 9.6, 9.4 8.10	> vWS > vPC > vApps	> vWS > vApps
Tesla V100 FHHL	10.0 9.6, 9.4	10.0 9.6, 9.4	> vWS > vPC	> vWS > vApps

GPU	Mixed vGPU Configuration - Red Hat Enterprise Linux with KVM Releases		Supported NVIDIA vGPU Software Products[1, 2, 3]	
	Frame Buffer Size (Mixed-Size Mode)	Series	NVIDIA vGPU	GPU Pass Through
	8.10	8.10	vApps	

GPUs Based on the NVIDIA Maxwell Graphic Architecture

> **Note:**
>
> The following NVIDIA vGPU software features are **not** supported on GPUs based on the NVIDIA NVIDIA Maxwell™ graphic architecture:
>
> - SR-IOV
> - Configuration of vGPUs with different amounts of frame buffer on the same physical GPU (mixed-size mode)
> - Manual placement of vGPUs on GPUs in equal-size mode

GPU	Mixed vGPU Series Configuration - Red Hat Enterprise Linux with KVM Releases	Supported NVIDIA vGPU Software Products[1, 2, 3]	
		NVIDIA vGPU	GPU Pass Through
Tesla M10	10.0 9.6, 9.4 8.10	vWS vPC vApps	vWS vApps

---

[1] The supported products are as follows:

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

[2] N/A indicates that the deployment is not supported.

[3] vApps is supported only on Windows operating systems.

[4] This GPU is supported only in displayless mode. In displayless mode, local physical display connectors are disabled.

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

Red Hat Enterprise Linux with KVM supports a mixture of different types of time-sliced vGPUs on the same physical GPU. Any combination of A-series, B-series, and Q-series vGPUs with any amount of frame buffer can reside on the same physical GPU simultaneously. The total amount of frame buffer allocated to the vGPUs on a physical GPU must not exceed the amount of frame buffer that the physical GPU has.

For example, the following combinations of vGPUs can reside on the same physical GPU simultaneously:

- A40-2B and A40-2Q
- A40-2Q and A40-4Q
- A40-2B and A40-4Q

By default, a GPU supports only vGPUs with the same amount of frame buffer and, therefore, is in equal-size mode. To support vGPUs with different amounts of frame buffer, the GPU must be put into mixed-size mode. When a GPU is in mixed-size mode, the maximum number of some types of vGPU allowed on a GPU is less than when the GPU is in equal-size mode. For more information, refer to *Virtual GPU Software User Guide*.

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

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

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

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

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

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

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

2.2. Hypervisor Software Releases

This release supports **only** the hypervisor software releases listed in the table.

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

Software	Releases Supported	Notes
Red Hat Enterprise Linux with KVM	10.0	All NVIDIA GPUs that NVIDIA vGPU software supports are supported with vGPU and in pass-through mode.
Red Hat Enterprise Linux with KVM	9.6	All NVIDIA GPUs that NVIDIA vGPU software supports are supported with vGPU and in pass-through mode.
Red Hat Enterprise Linux with KVM	9.4	All NVIDIA GPUs that NVIDIA vGPU software supports are supported with vGPU and in pass-through mode.
Red Hat Enterprise Linux with KVM	8.10	All NVIDIA GPUs that NVIDIA vGPU software supports are supported with vGPU and in pass-through mode.
Red Hat OpenShift Virtualization (OSV)	4.16 4.15	Supported on only the following GPUs with vGPU and in pass-through mode: > NVIDIA RTX A6000 > NVIDIA RTX A5500 > NVIDIA RTX A5000 > NVIDIA A40 > NVIDIA A16 > NVIDIA A10 > NVIDIA A2 > NVIDIA RTX 6000 Ada > NVIDIA L40S > NVIDIA L40 > NVIDIA L20 > NVIDIA L4 > NVIDIA L2 Other GPUs that NVIDIA vGPU software supports are **not** supported.

2.3. Guest OS Support

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

> **Note:**
>
> Use only a guest OS release that is listed as supported by NVIDIA vGPU software with your virtualization software. To be listed as supported, a guest OS release must be supported not only by NVIDIA vGPU software, but also by your virtualization software. NVIDIA **cannot** support guest OS releases that your virtualization software does not support.
>
> NVIDIA vGPU software supports **only** 64-bit guest operating systems. No 32-bit guest operating systems are supported.

2.3.1. Windows Guest OS Support

> **Note:** Red Hat Enterprise Linux with KVM and RHV support Windows guest operating systems under specific Red Hat subscription programs. For details, refer to Certified Guest Operating Systems in Red Hat OpenStack Platform, Red Hat Virtualization, Red Hat OpenShift Virtualization and Red Hat Enterprise Linux with KVM.

NVIDIA vGPU software supports **only** the 64-bit Windows releases listed as a guest OS on Red Hat Enterprise Linux with KVM. The releases of Red Hat Enterprise Linux with KVM 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.
>
> Windows Enterprise multi-session is **not** supported.

2.3.1.1. Windows Guest OS Support in Release 19.0

- Windows Server 2022
- Windows 11 24H2 and all Windows 11 releases supported by Microsoft up to and including this release
- Windows 10 2022 Update (22H2) and all Windows 10 releases supported by Microsoft up to and including this release

> **Note:** 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 64-bit Linux distributions listed as a guest OS on Red Hat Enterprise Linux with KVM. The releases of Red Hat Enterprise Linux with KVM 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.
>
> Rocky Linux releases that are compatible with supported Red Hat Enterprise Linux releases are also supported as a guest OS.

2.3.2.1. Linux Guest OS Support in Release 19.0

- **Deprecated:** CentOS Linux 8 (2105)
- Red Hat CoreOS 4.11
- Red Hat Enterprise Linux 10.0

  Supported only in console and command-line interface (CLI) mode because NVIDIA vGPU software does not support the Wayland display server protocol

  Not supported on Red Hat Enterprise Linux with KVM hypervisor 8 and 9 releases
- Red Hat Enterprise Linux 9.6

  Not supported on Red Hat Enterprise Linux with KVM hypervisor 8 releases
- Red Hat Enterprise Linux 9.4

  Not supported on Red Hat Enterprise Linux with KVM hypervisor 8 releases
- Red Hat Enterprise Linux 8.10

2.4. NVIDIA CUDA Toolkit Version Support

The releases in this release family of NVIDIA vGPU software support NVIDIA CUDA Toolkit 13.0.

To build a CUDA application, the system must have the NVIDIA CUDA Toolkit and the libraries required for linking. For details of the components of NVIDIA CUDA Toolkit, refer to *NVIDIA CUDA Toolkit 12.8 Release Notes*.

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

For more information about NVIDIA CUDA Toolkit, refer to CUDA Toolkit Documentation 13.0.

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 is supported on all supported GPUs, but only on a subset of supported Red Hat Enterprise Linux with KVM releases and guest operating systems.

Limitations with vGPU Migration Support

For a host that is running Red Hat Enterprise Linux with KVM 9.4, the following migrations are **not** supported:

- Migration between hosts that are running different versions of the NVIDIA Virtual GPU Manager driver, even within the same NVIDIA Virtual GPU Manager driver branch
- Migration to or from a host that is running a version of Red Hat Enterprise Linux with KVM that uses of a vendor-specific VFIO framework

Unless explicitly stated otherwise, these restrictions do not apply to a host that is running Red Hat Enterprise Linux with KVM since 9.6. For example, the following migrations are supported:

- Migration between hosts that are running different versions of the NVIDIA Virtual GPU Manager driver
- Migration between a host that is running Red Hat Enterprise Linux with KVM 9.6 and a host that is running Red Hat Enterprise Linux with KVM 10.0

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

- Unified memory
- Debuggers
- Profilers

Supported Hypervisor Software Releases

Since Red Hat Enterprise Linux with KVM 9.4

Supported Guest OS Releases

Windows and Linux.

Known Issues with vGPU Migration Support

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

2.6. Multiple vGPU Support

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

2.6.1. vGPUs that Support Multiple vGPUs Assigned to a VM

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

- For GPUs based on the NVIDIA Volta architecture and later GPU architectures, **all** Q-series vGPUs are supported. On GPUs that support the Multi-Instance GPU (MIG) feature, both time-sliced and MIG-backed vGPUs are supported.
- For GPUs based on the NVIDIA Pascal™ architecture, only Q-series vGPUs that are allocated all of the physical GPU's frame buffer are supported.
- For GPUs based on the NVIDIA NVIDIA Maxwell™ graphic architecture, only Q-series vGPUs that are allocated all of the physical GPU's frame buffer are supported.

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

Multiple vGPU Support on the NVIDIA Blackwell Architecture

Board	vGPU
NVIDIA RTX PRO 6000 Blackwell Server Edition	All Q-series vGPUs

Multiple vGPU Support on the NVIDIA Ada Lovelace Architecture

Board	vGPU
NVIDIA L40S	All Q-series vGPUs

Board	vGPU
NVIDIA L40	All Q-series vGPUs
NVIDIA L20 NVIDIA L20 liquid cooled	All Q-series vGPUs
NVIDIA L4	All Q-series vGPUs
NVIDIA L2	All Q-series vGPUs
NVIDIA RTX 6000 Ada	All Q-series vGPUs
NVIDIA RTX 5880 Ada	All Q-series vGPUs
NVIDIA RTX 5000 Ada	All Q-series vGPUs

Multiple vGPU Support on the NVIDIA Ampere GPU Architecture

Board	vGPU
NVIDIA A40	All Q-series vGPUs See Note (1).
NVIDIA A16	All Q-series vGPUs See Note (1).
NVIDIA A10	All Q-series vGPUs See Note (1).
NVIDIA A2	All Q-series vGPUs See Note (1).
NVIDIA RTX A6000	All Q-series vGPUs See Note (1).
NVIDIA RTX A5500	All Q-series vGPUs See Note (1).
NVIDIA RTX A5000	All Q-series vGPUs See Note (1).

Multiple vGPU Support on the NVIDIA Turing GPU Architecture

Board	vGPU
Tesla T4	All Q-series vGPUs
Quadro RTX 6000	All Q-series vGPUs
Quadro RTX 6000 passive	All Q-series vGPUs
Quadro RTX 8000	All Q-series vGPUs
Quadro RTX 8000 passive	All Q-series vGPUs

Multiple vGPU Support on the NVIDIA Volta GPU Architecture

Board	vGPU
Tesla V100 SXM2 32GB	All Q-series vGPUs
Tesla V100 PCIe 32GB	All Q-series vGPUs
Tesla V100S PCIe 32GB	All Q-series vGPUs

Board	vGPU
Tesla V100 SXM2	All Q-series vGPUs
Tesla V100 PCIe	All Q-series vGPUs
Tesla V100 FHHL	All Q-series vGPUs

Multiple vGPU Support on the NVIDIA Maxwell GPU Architecture

Board	vGPU
Tesla M10	M10-8Q

Note:

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

2.6.2. Maximum Number of vGPUs Supported per VM

NVIDIA vGPU software supports up to a maximum of 16 vGPUs per VM.

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

All hypervisor releases that support NVIDIA vGPU software are supported.

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

Peer-to-peer CUDA transfers enable device memory between vGPUs on different GPUs that are assigned to the same VM to be accessed from within the CUDA kernels. NVLink is a high-bandwidth interconnect that enables fast communication between such vGPUs. Peer-to-Peer CUDA transfers over NVLink are supported only on a subset of vGPUs, Red Hat Enterprise Linux with KVM releases, and guest OS releases.

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

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

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

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

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

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

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

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

Note:

1. Supported only on the following hardware:
    - NVIDIA HGX™ A100 4-GPU baseboard with four fully connected GPUs
    - NVIDIA HGX A100 8-GPU baseboards with eight fully connected GPUs

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

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

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

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

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

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

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

2.8. Unified Memory Support

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

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

2.8.1. vGPUs that Support Unified Memory

On GPUs that support the MIG feature and on which this feature is enabled, **only** Q-series MIG-backed vGPUs that occupy an entire GPU instance are supported. All other MIG-backed vGPUs are **not** supported.

On single-instance GPUs, only Q-series vGPUs that are allocated all of the physical GPU's frame buffer on physical GPUs that support unified memory are supported.

Unified Memory Support on the NVIDIA Blackwell GPU Architecture

Board	vGPU
NVIDIA RTX PRO 6000 Blackwell Server Edition	NVIDIA RTX PRO 6000 Blackwell DC-96Q All Q-series MIG-backed vGPUs that occupy an entire GPU instance

Unified Memory Support on the NVIDIA Ada Lovelace GPU Architecture

Board	vGPU
NVIDIA L40	L40-48Q
NVIDIA L40S	L40S-48Q
NVIDIA L20 NVIDIA L20 liquid cooled	L20-48Q
NVIDIA L4	L4-24Q
NVIDIA L2	L2-24Q
NVIDIA RTX 6000 Ada	RTX 6000 Ada-48Q
NVIDIA RTX 5880 Ada	RTX 5880 Ada-48Q
NVIDIA RTX 5000 Ada	RTX 5000 Ada-32Q

Unified Memory Support on the NVIDIA Ampere GPU Architecture

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

2.8.2. Guest OS Releases that Support Unified Memory

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

2.8.3. Limitations on Support for Unified Memory

- Only time-sliced Q-series and C-series vGPUs that are allocated all of the physical GPU's frame buffer on physical GPUs that support unified memory are supported. Fractional time-sliced vGPUs are **not** supported.

2.9. NVIDIA GPU Operator Support

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

For more information, refer to Using NVIDIA vGPU in the NVIDIA GPU Operator documentation.

2.10. NVIDIA Deep Learning Super Sampling (DLSS) Support

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

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

Supported GPUs:

- NVIDIA L40
- NVIDIA L40S
- NVIDIA L20
- NVIDIA L20 liquid cooled
- NVIDIA L4
- NVIDIA L2
- NVIDIA RTX 6000 Ada
- NVIDIA RTX 5880 Ada
- NVIDIA RTX 5000 Ada
- NVIDIA A40
- NVIDIA A16
- NVIDIA A2
- NVIDIA A10
- NVIDIA RTX A6000
- NVIDIA RTX A5500
- NVIDIA RTX A5000
- NVIDIA RTX PRO 6000 Blackwell Server Edition

- 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

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.

3.2. vCS is not supported on Red Hat Enterprise Linux with KVM

NVIDIA Virtual Compute Server (vCS) is not supported on Red Hat Enterprise Linux with KVM. C-series vGPU types are not available.

3.3. Nested Virtualization Is Not Supported by NVIDIA vGPU

In general, NVIDIA vGPU deployments do not support nested virtualization, that is, running a hypervisor in a guest VM. For example, enabling the Hyper-V role in a guest VM running the Windows Server OS is **not** supported because it entails enabling nested virtualization. Similarly, enabling Windows Hypervisor Platform is not supported because it requires the Hyper-V role to be enabled.

3.4. 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.

3.5. Virtual GPU hot plugging is not supported

NVIDIA vGPU software does not support the addition of virtual function I/O (VFIO) mediated device (`mdev`) devices after the VM has been started by QEMU. All `mdev` devices must be added before the VM is started.

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

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

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

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

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

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

max-reserved-fb

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

vgpu-profile-size-in-mb

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

ecc-adjustments

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

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

page-retirement-allocation

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

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

max-vgpus-per-gpu

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

compression-adjustment

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

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

vGPU Type	Compression Adjustment (MB)
T4-16Q T4-16C T4-16A	28
RTX6000-12Q RTX6000-12C RTX6000-12A	32
RTX6000-24Q RTX6000-24C RTX6000-24A	104
RTX6000P-12Q RTX6000P-12C RTX6000P-12A	32

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

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

Description

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

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

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

Workaround

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

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

Description

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

Workaround

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

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

Description

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

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

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

Workaround

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

3.10. NVENC requires at least 1 Gbyte of frame buffer

Description

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

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

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

Workaround

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

3.11. VM 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 Red Hat Enterprise Linux with KVM 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 Red Hat Enterprise Linux with KVM VM's `/var/log/messages` 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 Red Hat Enterprise Linux with KVM VM's `/var/log/messages` 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.

3.12. 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 setting `frame_rate_limiter=0` in the vGPU configuration file.

```
# echo "frame_rate_limiter=0" > /sys/bus/mdev/devices/vgpu-id/nvidia/vgpu_params
```

For example:

```
# echo "frame_rate_limiter=0" > /sys/bus/mdev/devices/aa618089-8b16-4d01-a136-25a0f3c73123/
nvidia/vgpu_params
```

The setting takes effect the next time any VM using the given vGPU type is started.

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 as follows:

1. Clear all parameter settings in the vGPU configuration file.

   ```
   # echo " " > /sys/bus/mdev/devices/vgpu-id/nvidia/vgpu_params
   ```

   > **Note:** You cannot clear specific parameter settings. If your vGPU configuration file contains other parameter settings that you want to keep, you must reinstate them in the next step.

2. Set `frame_rate_limiter=1` in the vGPU configuration file.

   ```
   # echo "frame_rate_limiter=1" > /sys/bus/mdev/devices/vgpu-id/nvidia/vgpu_params
   ```

If you need to reinstate other parameter settings, include them in the command to set `frame_rate_limiter=1`. For example:

```
# echo "frame_rate_limiter=1 disable_vnc=1" > /sys/bus/mdev/devices/aa618089-8b16-4d01-
a136-25a0f3c73123/nvidia/vgpu_params
```

3.13. `nvidia-smi` fails to operate when all GPUs are assigned to GPU pass-through mode

Description

If all GPUs in the platform are assigned to VMs in pass-through mode, `nvidia-smi` will return an error:

```
[root@vgx-test ~]# nvidia-smi
Failed to initialize NVML: Unknown Error
```

This is because GPUs operating in pass-through mode are not visible to `nvidia-smi` and the NVIDIA kernel driver operating in the Red Hat Enterprise Linux with KVM host.

To confirm that all GPUs are operating in pass-through mode, confirm that the `vfio-pci` kernel driver is handling each device.

```
# lspci -s 05:00.0 -k
05:00.0 VGA compatible controller: NVIDIA Corporation GM204GL [Tesla M60] (rev a1)
             Subsystem: NVIDIA Corporation Device 113a
             Kernel driver in use: vfio-pci
```

Resolution

N/A

Chapter 4. Resolved Issues

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

4.1. Issues Resolved in Release 19.0

No resolved issues are reported in this release for Red Hat Enterprise Linux with KVM.

Chapter 5. Known Issues

Refer to the separate *Known Issues* document

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