Virtual GPU Software R535 for Red Hat Enterprise Linux with KVM

Release Notes
# Table of Contents

Chapter 1. Release Notes......................................................................................................................1
  1.1. NVIDIA vGPU Software Driver Versions.................................................................1
  1.2. Compatibility Requirements for the NVIDIA vGPU Manager and Guest VM Driver....2
  1.3. Updates in Release 16.2.........................................................................................3
  1.4. Updates in Release 16.1.........................................................................................4
  1.5. Updates in Release 16.0.........................................................................................4

Chapter 2. Validated Platforms.......................................................................................................... 6
  2.1. Supported NVIDIA GPUs and Validated Server Platforms.................................6
  2.1.1. Support for a Mixture of Time-Sliced vGPU Types on the Same GPU.............11
  2.1.2. Switching the Mode of a GPU that Supports Multiple Display Modes..........11
  2.1.3. Switching the Mode of a Tesla M60 or M6 GPU............................................12
  2.2. Hypervisor Software Releases............................................................................12
  2.3. Guest OS Support............................................................................................... 13
  2.3.1. Windows Guest OS Support.............................................................................14
   2.3.1.1. Windows Guest OS Support in Release 16.2............................................14
   2.3.1.2. Windows Guest OS Support in Release 16.1............................................15
   2.3.1.3. Windows Guest OS Support in Release 16.0............................................15
  2.3.2. Linux Guest OS Support..................................................................................16
   2.3.2.1. Linux Guest OS Support in Release 16.2................................................16
   2.3.2.2. Linux Guest OS Support in Release 16.1................................................17
   2.3.2.3. Linux Guest OS Support in Release 16.0................................................18
  2.4. NVIDIA CUDA Toolkit Version Support...............................................................19
  2.5. Multiple vGPU Support.......................................................................................19
   2.5.1. vGPUs that Support Multiple vGPUs Assigned to a VM.............................19
   2.5.2. Maximum Number of vGPUs Supported per VM.....................................21
   2.5.3. Hypervisor Releases that Support Multiple vGPUs Assigned to a VM........22
  2.6. Peer-to-Peer CUDA Transfers over NVLink Support.............................................22
   2.6.1. vGPUs that Support Peer-to-Peer CUDA Transfers...................................22
   2.6.2. Hypervisor Releases that Support Peer-to-Peer CUDA Transfers..............23
   2.6.3. Guest OS Releases that Support Peer-to-Peer CUDA Transfers................23
   2.6.4. Limitations on Support for Peer-to-Peer CUDA Transfers.......................23
  2.7. Unified Memory Support......................................................................................24
   2.7.1. vGPUs that Support Unified Memory.........................................................24
   2.7.2. Guest OS Releases that Support Unified Memory......................................25
   2.7.3. Limitations on Support for Unified Memory..............................................25
Chapter 3. Known Product Limitations

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

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

3.3. Nested Virtualization Is Not Supported by NVIDIA vGPU

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

3.5. Virtual GPU hot plugging is not supported

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

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

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

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

3.10. NVENC requires at least 1 Gbyte of frame buffer

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

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

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

Chapter 4. Resolved Issues

4.1. Issues Resolved in Release 16.2

4.2. Issues Resolved in Release 16.1

4.3. Issues Resolved in Release 16.0

Chapter 5. Known Issues

5.1. Pixelation occurs on a Windows VM configured with a Tesla T4 vGPU

5.2. NVIDIA Control Panel is not available in multiuser environments

5.3. 16.0, 16.1 Only: NVWMI functions for faking EDID have no effect

5.4. 16.0 Only: Benign No such file or directory message during capture of configuration data for a bug report

5.5. 16.0 Only: The NVIDIA L40 GPU brand is incorrectly identified if GSP firmware is disabled

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

5.7. Remote desktop connection is lost and the NVIDIA vGPU software graphics driver is unloaded

5.8. 16.0 Only: Graphics applications are corrupted on some Windows vGPU VMs

5.9. VM assigned multiple fractional vGPUs from the same GPU hangs
5.10. CUDA profilers cannot gather hardware metrics on NVIDIA vGPU................................. 46
5.11. NVIDIA vGPU software graphics driver for Windows sends a remote call to ngx.download.nvidia.com............................................................................................................................... 46
5.12. On NVIDIA H100, creation of multiple compute instances after deletion of existing compute instances fails......................................................................................................................... 47
5.13. NLS client fails to acquire a license with the error The allowed time to process response has expired........................................................................................................................................... 48
5.14. NVIDIA vGPU software graphics driver fails to load on KVM-based hypervisors...... 49
5.15. With multiple active sessions, NVIDIA Control Panel incorrectly shows that the system is unlicensed........................................................................................................................................... 50
5.16. VP9 and AV1 decoding with web browsers are not supported on Microsoft Windows Server 2019........................................................................................................................................... 51
5.17. nvidia-smi ignores the second NVIDIA vGPU device added to a Microsoft Windows Server 2016 VM........................................................................................................................................... 51
5.18. After an upgrade of the Linux graphics driver from an RPM package in a licensed VM, licensing fails............................................................................................................................................... 53
5.19. The reported NVENC frame rate is double the actual frame rate........................................ 53
5.20. NVENC does not work with Teradici Cloud Access Software on Windows................... 54
5.21. A licensed client might fail to acquire a license if a proxy is set........................................ 54
5.22. Session connection fails with four 4K displays and NVENC enabled on a 2Q, 3Q, or 4Q vGPU.................................................................................................................................................. 55
5.23. NVIDIA A100 HGX 80GB vGPU names shown as Graphics Device by nvidia-smi..... 56
5.24. Idle Teradici Cloud Access Software session disconnects from Linux VM..................... 57
5.25. GPU Operator doesn't support vGPU on GPUs based on architectures before NVIDIA Turing.................................................................................................................................................... 58
5.26. Idle NVIDIA A100, NVIDIA A40, and NVIDIA A10 GPUs show 100% GPU utilization 58
5.27. Guest VM frame buffer listed by nvidia-smi for vGPUs on GPUs that support SRIOV is incorrect.................................................................................................................................................. 60
5.28. VMs fail to boot on RHV 4.4........................................................................................................ 60
5.29. Driver upgrade in a Linux guest VM with multiple vGPUs might fail................................. 61
5.30. NVIDIA Control Panel fails to start if launched too soon from a VM without licensing information.................................................................................................................................................. 62
5.31. On Linux, the frame rate might drop to 1 after several minutes...................................... 63
5.32. DWM crashes randomly occur in Windows VMs................................................................. 64
5.33. ECC memory settings for a vGPU cannot be changed by using NVIDIA X Server Settings.................................................................................................................................................... 64
5.34. Changes to ECC memory settings for a Linux vGPU VM by nvidia-smi might be ignored.................................................................................................................................................. 65
5.35. Host core CPU utilization is higher than expected for moderate workloads................. 66
5.36. Frame capture while the interactive logon message is displayed returns blank screen......................................................................................................................................................................66
5.37. RDS sessions do not use the GPU with some Microsoft Windows Server releases 67
5.38. When the scheduling policy is fixed share, GPU utilization is reported as higher than expected........................................................................................................................................................................................................68
5.39. License is not acquired in Windows VMs.................................................................................................................................69
5.40. nvidia-smi reports that vGPU migration is supported on all hypervisors..................................................................................69
5.41. Hot plugging and unplugging vCPUs causes a blue-screen crash in Windows VMs........................................................................................................................................................................................................70
5.42. Luxmark causes a segmentation fault on an unlicensed Linux client..........................................................70
5.43. A segmentation fault in DBus code causes nvidia-gridd to exit on Red Hat Enterprise Linux and CentOS............................................................................................................................................... 71
5.44. No Manage License option available in NVIDIA X Server Settings by default...............................72
5.45. Licenses remain checked out when VMs are forcibly powered off.................................................................73
5.46. VM bug checks after the guest VM driver for Windows 10 RS2 is installed............................73
5.47. GNOME Display Manager (GDM) fails to start on Red Hat Enterprise Linux 7.2 and CentOS 7.0........................................................................................................................................................................................................74
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.

<table>
<thead>
<tr>
<th>NVIDIA vGPU Software Version</th>
<th>NVIDIA Virtual GPU Manager Version</th>
<th>NVIDIA Windows Driver Version</th>
<th>NVIDIA Linux Driver Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.2</td>
<td>535.129.03</td>
<td>537.70</td>
<td>535.129.03</td>
</tr>
<tr>
<td>16.1</td>
<td>535.104.06</td>
<td>537.13</td>
<td>535.104.05</td>
</tr>
<tr>
<td>16.0</td>
<td>535.54.06</td>
<td>536.25</td>
<td>535.54.03</td>
</tr>
</tbody>
</table>

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 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 16.2 is used with guest drivers from release 13.1, the combination does not support Red Hat Enterprise Linux 8.1 because NVIDIA vGPU software release 16.2 does not support Red Hat Enterprise Linux 8.1.

The following table lists the specific software releases that are compatible with the components in the NVIDIA vGPU software 16 major release branch.
Virtual GPU Software R535 for Red Hat Enterprise Linux
with KVM

Release Notes

NVIDIA vGPU Software Component | Releases | Compatible Software Releases
--- | --- | ---
NVIDIA vGPU Manager | 16.0 through 16.2 | ▶ Guest VM driver releases 16.0 through 16.2  
▶ All guest VM driver 15.x releases  
▶ All guest VM driver 13.x releases
Guest VM drivers | 16.0 through 16.2 | NVIDIA vGPU Manager releases 16.0 through 16.2

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 16 major release branch.

<table>
<thead>
<tr>
<th>NVIDIA vGPU Software Component</th>
<th>Releases</th>
<th>Incompatible Software Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA vGPU Manager</td>
<td>16.0 through 16.2</td>
<td>All guest VM driver releases 14.x and earlier, except 13.x releases</td>
</tr>
<tr>
<td>Guest VM drivers</td>
<td>16.0 through 16.2</td>
<td>All NVIDIA vGPU Manager releases 15.x and earlier</td>
</tr>
</tbody>
</table>

1.3. Updates in Release 16.2

New Features in Release 16.2

▶ Security updates - see Security Bulletin: NVIDIA GPU Display Driver - October 2023, which is posted shortly after the release date of this software and is listed on the NVIDIA Product Security page
▶ Miscellaneous bug fixes
1.4. Updates in Release 16.1

New Features in Release 16.1

‣ New options in the NVML API and the nvidia-smi command for getting the scheduling behavior of time-sliced vGPUs
‣ Miscellaneous bug fixes

Hardware and Software Support Introduced in Release 16.1

‣ Support for the following GPUs:
  ‣ NVIDIA L40S
  ‣ NVIDIA RTX 5000 Ada
‣ Support for Microsoft Windows Server 2022 as a guest OS

1.5. Updates in Release 16.0

New Features in Release 16.0

‣ Support for 4K displays with an aspect ratio of 16:10
‣ Options in the NVML API and the nvidia-smi command for controlling the scheduling behavior of time-sliced vGPUs
‣ Security updates - see Security Bulletin: NVIDIA GPU Display Driver - June 2023, which is posted shortly after the release date of this software and is listed on the NVIDIA Product Security page
‣ Miscellaneous bug fixes

Newly Supported Hardware and Software in Release 16.0

‣ Newly supported hypervisor software:
  ‣ Red Hat Enterprise Linux with KVM 9.2 and 8.8
‣ Newly supported guest OSes:
  ‣ Red Hat Enterprise Linux 9.2 and 8.8

Feature Support Withdrawn in Release 16.0

‣ Product functionality no longer supported:
- NVIDIA Virtual Compute Server (vCS)

**Note:** As a result of this change, C-series vGPU types are no longer available.

Instead, vCS is supported with NVIDIA AI Enterprise.

- Graphics cards no longer supported:
  - Graphics cards that support only C-series vGPUs, namely:
    - NVIDIA H800 PCIe 80GB
    - NVIDIA H100 PCIe 80GB
    - NVIDIA A800 PCIe 80GB
    - NVIDIA A800 PCIe 80GB liquid cooled
    - NVIDIA A800 HGX 80GB
    - NVIDIA A100 PCIe 80GB
    - NVIDIA A100 PCIe 80GB liquid cooled
    - NVIDIA A100X
    - NVIDIA A100 HGX 80GB
    - NVIDIA A100 PCIe 40GB
    - NVIDIA A100 HGX 40GB
    - NVIDIA A30
    - NVIDIA A30X

Instead, these graphics cards are supported with NVIDIA AI Enterprise.

- Hypervisor software no longer supported:
  - Red Hat Enterprise Linux with KVM 9.1, 8.7, and 8.4

- Guest OSes no longer supported:
  - Red Hat Enterprise Linux 9.1, 8.7, and 8.4
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 R535 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 Ada Lovelace Architecture

<table>
<thead>
<tr>
<th>GPU</th>
<th>Supported NVIDIA vGPU Software Products$^{1,2,3}$</th>
<th>GPU Pass Through</th>
</tr>
</thead>
</table>
| Since 16.1: NVIDIA L40S | ◦ vWS  
 ◦ vPC  
 ◦ vApps | ◦ vWS  
 ◦ vApps |
| NVIDIA L40 | ◦ vWS  
 ◦ vPC  
 ◦ vApps | ◦ vWS  
 ◦ vApps |
| NVIDIA L4 | ◦ vWS  
 ◦ vPC  
 ◦ vApps | ◦ vWS  
 ◦ vApps |
### Validated Platforms

<table>
<thead>
<tr>
<th>GPU</th>
<th>Supported NVIDIA vGPU Software Products</th>
<th>GPU Pass Through</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>NVIDIA vGPU</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>vWS</td>
<td>vWS</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td><strong>Since 16.1:</strong> NVIDIA RTX 5000 Ada</td>
<td>vWS</td>
<td>vWS</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
</tbody>
</table>

### GPUs Based on the NVIDIA Ampere Architecture

<table>
<thead>
<tr>
<th>GPU</th>
<th>Supported NVIDIA vGPU Software Products</th>
<th>GPU Pass Through</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>NVIDIA vGPU</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>vWS</td>
<td>vWS</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td>NVIDIA A404</td>
<td>vWS</td>
<td>vWS</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td>NVIDIA A16</td>
<td>vWS</td>
<td>vWS</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td>NVIDIA A10</td>
<td>vWS</td>
<td>vWS</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td>NVIDIA A2</td>
<td>vWS</td>
<td>vWS</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td>NVIDIA RTX A60004</td>
<td>vWS</td>
<td>vWS</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td>NVIDIA RTX A55004</td>
<td>vWS</td>
<td>vWS</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
</tbody>
</table>
## Validated Platforms

### GPU Specifications

#### Supported NVIDIA vGPU Software Products

1. Virtual GPU Software R535 for Red Hat Enterprise Linux with KVM

<table>
<thead>
<tr>
<th>GPU</th>
<th>NVIDIA vGPU</th>
<th>GPU Pass Through</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA RTX A5000</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td>Tesla T4</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td>Quadro RTX 6000</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td>Quadro RTX 6000 passive</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td>Quadro RTX 8000</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td>Quadro RTX 8000 passive</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
</tbody>
</table>

#### GPUs Based on the NVIDIA Turing Architecture

<table>
<thead>
<tr>
<th>GPU</th>
<th>NVIDIA vGPU</th>
<th>GPU Pass Through</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA RTX A5000</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td>Tesla T4</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td>Quadro RTX 6000</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td>Quadro RTX 6000 passive</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td>Quadro RTX 8000</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td>Quadro RTX 8000 passive</td>
<td>vWS, vPC, vApps</td>
<td>vWS, vApps</td>
</tr>
</tbody>
</table>

#### GPUs Based on the NVIDIA Volta Architecture

<table>
<thead>
<tr>
<th>GPU</th>
<th>NVIDIA vGPU</th>
<th>GPU Pass Through</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla V100 SXM2</td>
<td>vWS, vPC</td>
<td>vWS, vApps</td>
</tr>
</tbody>
</table>
**Validated Platforms**

<table>
<thead>
<tr>
<th>GPU</th>
<th>Supported NVIDIA vGPU Software Products</th>
<th>GPU Pass Through</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td><strong>Tesla V100 SXM2 32GB</strong></td>
<td>vWS</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td><strong>Tesla V100 PCIe</strong></td>
<td>vWS</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td><strong>Tesla V100 PCIe 32GB</strong></td>
<td>vWS</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td><strong>Tesla V100S PCIe 32GB</strong></td>
<td>vWS</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td><strong>Tesla V100 FHHL</strong></td>
<td>vWS</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
</tbody>
</table>

**GPUs Based on the NVIDIA Pascal™ Architecture**

<table>
<thead>
<tr>
<th>GPU</th>
<th>Supported NVIDIA vGPU Software Products</th>
<th>GPU Pass Through</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td><strong>Tesla P4</strong></td>
<td>vWS</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td><strong>Tesla P6</strong></td>
<td>vWS</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
<tr>
<td><strong>Tesla P40</strong></td>
<td>vWS</td>
<td>vWS, vApps</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td></td>
</tr>
</tbody>
</table>
## Validated Platforms

### Supported NVIDIA vGPU Software Products

<table>
<thead>
<tr>
<th>GPU</th>
<th>NVIDIA vGPU</th>
<th>GPU Pass Through</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>vWS</td>
<td>vWS</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td>vApps</td>
</tr>
</tbody>
</table>

### NVIDIA vGPU | GPU Pass Through

- **Tesla P100 PCIe 16 GB**
  - vWS
  - vPC
  - vApps

- **Tesla P100 SXM2 16 GB**
  - vWS
  - vPC
  - vApps

- **Tesla P100 PCIe 12 GB**
  - vWS
  - vPC
  - vApps

### GPUs Based on the NVIDIA Maxwell Graphic Architecture

<table>
<thead>
<tr>
<th>GPU</th>
<th>NVIDIA vGPU</th>
<th>GPU Pass Through</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>vWS</td>
<td>vWS</td>
</tr>
<tr>
<td></td>
<td>vPC</td>
<td>vApps</td>
</tr>
<tr>
<td></td>
<td>vApps</td>
<td>vApps</td>
</tr>
</tbody>
</table>

### NVIDIA vGPU | GPU Pass Through

- **Tesla M6**
  - vWS
  - vPC
  - vApps

- **Tesla M10**
  - vWS
  - vPC
  - vApps

- **Tesla M60**
  - vWS
  - vPC
  - 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.
2.1.1. Support for a Mixture of Time-Sliced vGPU Types on the Same GPU

Red Hat Enterprise Linux with KVM supports time-sliced vGPUs with the same amount of frame buffer from different virtual GPU series on the same physical GPU. A-series, B-series, and Q-series vGPUs with the same amount of frame buffer, for example A40-2B and A40-2Q, can reside on the same physical GPU simultaneously.

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.

<table>
<thead>
<tr>
<th>GPU</th>
<th>Mode as Supplied from the Factory</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA A40</td>
<td>Display-off</td>
</tr>
<tr>
<td>NVIDIA L40</td>
<td>Display-off</td>
</tr>
<tr>
<td>Since 16.1: NVIDIA L40S</td>
<td>Display-off</td>
</tr>
<tr>
<td>Since 16.1: NVIDIA RTX 5000 Ada</td>
<td>Display enabled</td>
</tr>
<tr>
<td>NVIDIA RTX 6000 Ada</td>
<td>Display enabled</td>
</tr>
<tr>
<td>NVIDIA RTX A5000</td>
<td>Display enabled</td>
</tr>
<tr>
<td>NVIDIA RTX A5500</td>
<td>Display enabled</td>
</tr>
<tr>
<td>NVIDIA RTX A6000</td>
<td>Display enabled</td>
</tr>
</tbody>
</table>

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.

---

4 This GPU is supported only in displayless mode. In displayless mode, local physical display connectors are disabled.
2.1.3. 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**.

2.2. 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.

<table>
<thead>
<tr>
<th>Software</th>
<th>Releases Supported</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Hat Enterprise Linux with KVM</td>
<td>9.2</td>
<td>All NVIDIA GPUs that NVIDIA vGPU software supports are supported with vGPU and in pass-through mode.</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux with KVM</td>
<td>9.0</td>
<td>All NVIDIA GPUs that NVIDIA vGPU software supports are supported with vGPU and in pass-through mode.</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux with KVM</td>
<td>8.8</td>
<td>All NVIDIA GPUs that NVIDIA vGPU software supports are supported with vGPU and in pass-through mode.</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux with KVM</td>
<td>8.6</td>
<td>All NVIDIA GPUs that NVIDIA vGPU software supports are supported with vGPU and in pass-through mode.</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux with KVM</td>
<td>7.9</td>
<td>The following GPUs are supported in GPU pass through mode only:</td>
</tr>
</tbody>
</table>
### Software

<table>
<thead>
<tr>
<th>Software</th>
<th>Releases Supported</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>▶ NVIDIA A40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>▶ NVIDIA A16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>▶ NVIDIA A10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>▶ NVIDIA A2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>▶ NVIDIA RTX A6000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>▶ NVIDIA RTX A5500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>▶ NVIDIA RTX A5000</td>
</tr>
<tr>
<td><img src="image.jpg" alt="Image" /></td>
<td></td>
<td>All other NVIDIA GPUs that NVIDIA vGPU software supports are</td>
</tr>
<tr>
<td><img src="image.jpg" alt="Image" /></td>
<td></td>
<td>supported with vGPU and in pass-through mode.</td>
</tr>
</tbody>
</table>

| Red Hat Virtualization (RHV)    | 4.4               | All NVIDIA GPUs that NVIDIA vGPU software supports are supported    |
|                                 |                   | with vGPU and in pass-through mode.                                  |

| Red Hat Virtualization (RHV)    | 4.3, 4.2          | Not supported on the following GPUs:                                 |
|                                 |                   | ▶ NVIDIA RTX A6000                                                   |
|                                 |                   | ▶ NVIDIA RTX A5500                                                   |
|                                 |                   | ▶ NVIDIA RTX A5000                                                   |
|                                 |                   | ▶ NVIDIA A40                                                         |
|                                 |                   | ▶ NVIDIA A10                                                         |
|                                 |                   | ▶ NVIDIA A16                                                         |
|                                 |                   | ▶ NVIDIA A2                                                          |
| ![Image](image.jpg)             |                   | All other NVIDIA GPUs that NVIDIA vGPU software supports are        |
| ![Image](image.jpg)             |                   | supported with vGPU and in pass-through mode.                        |

### 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.

2.3.1.1. Windows Guest OS Support in Release 16.2

<table>
<thead>
<tr>
<th>Guest OS</th>
<th>NVIDIA vGPU - Red Hat Enterprise Linux with KVM Releases</th>
<th>Pass-Through GPU - Red Hat Enterprise Linux with KVM Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows Server 2022</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td></td>
<td>RHV 4.4, 4.3, 4.2</td>
<td>RHV 4.4, 4.3, 4.2</td>
</tr>
<tr>
<td>Windows Server 2019</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td></td>
<td>RHV 4.4, 4.3, 4.2</td>
<td>RHV 4.4, 4.3, 4.2</td>
</tr>
<tr>
<td>Windows 11 22H2 and all Windows 11 releases supported by Microsoft up to and including this release</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
</tr>
<tr>
<td></td>
<td>RHV 4.4</td>
<td>RHV 4.4</td>
</tr>
<tr>
<td>Windows 10 2022 Update (22H2) and all Windows 10 releases supported by Microsoft up to and including this release</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
</tr>
<tr>
<td>See Note (1)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note:
1. The hardware-accelerated GPU scheduling feature introduced in Windows 10 May 2020 Update (2004) is not supported on GPUs based on the Maxwell architecture and is supported only in pass-through mode on GPUs based on later architectures.
### 2.3.1.2. Windows Guest OS Support in Release 16.1

<table>
<thead>
<tr>
<th>Guest OS</th>
<th>NVIDIA vGPU - Red Hat Enterprise Linux with KVM Releases</th>
<th>Pass-Through GPU - Red Hat Enterprise Linux with KVM Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows Server 2022</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td></td>
<td>RHEL KVM 4.4, 4.3, 4.2</td>
<td>RHEL KVM 4.4, 4.3, 4.2</td>
</tr>
<tr>
<td>Windows Server 2019</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td></td>
<td>RHEL KVM 4.4, 4.3, 4.2</td>
<td>RHEL KVM 4.4, 4.3, 4.2</td>
</tr>
<tr>
<td>Windows 11 22H2 and all Windows 11 releases supported by Microsoft up to and including this release</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
</tr>
<tr>
<td></td>
<td>RHEL KVM 4.4</td>
<td>RHEL KVM 4.4</td>
</tr>
<tr>
<td>Windows 10 2022 Update (22H2) and all Windows 10 releases supported by Microsoft up to and including this release</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
</tr>
<tr>
<td></td>
<td>RHEL KVM 4.4</td>
<td>RHEL KVM 4.4</td>
</tr>
</tbody>
</table>

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

### 2.3.1.3. Windows Guest OS Support in Release 16.0

<table>
<thead>
<tr>
<th>Guest OS</th>
<th>NVIDIA vGPU - Red Hat Enterprise Linux with KVM Releases</th>
<th>Pass-Through GPU - Red Hat Enterprise Linux with KVM Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows Server 2019</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td></td>
<td>RHEL KVM 4.4, 4.3, 4.2</td>
<td>RHEL KVM 4.4, 4.3, 4.2</td>
</tr>
<tr>
<td>Windows 11 22H2 and all Windows 11 releases supported by Microsoft up to and including this release</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
</tr>
<tr>
<td></td>
<td>RHEL KVM 4.4</td>
<td>RHEL KVM 4.4</td>
</tr>
<tr>
<td>Windows 10 2022 Update (22H2) and all Windows 10 releases supported by Microsoft up to and including this release</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
</tr>
<tr>
<td></td>
<td>RHEL KVM 4.4</td>
<td>RHEL KVM 4.4</td>
</tr>
</tbody>
</table>
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.

### 2.3.2.1. Linux Guest OS Support in Release 16.2

<table>
<thead>
<tr>
<th>Guest OS</th>
<th>NVIDIA vGPU - Red Hat Enterprise Linux with KVM Releases</th>
<th>Pass-Through GPU - Red Hat Enterprise Linux with KVM Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Hat CoreOS 4.11</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 9.2</td>
<td>RHEL KVM 9.2, 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
</tr>
<tr>
<td>Rocky Linux 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 8.8</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td>Rocky Linux 8.4</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0</td>
</tr>
<tr>
<td><strong>Deprecated:</strong> CentOS Linux 8 (2105)</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
</tbody>
</table>

---

**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.

---
<table>
<thead>
<tr>
<th>Guest OS</th>
<th>NVIDIA vGPU - Red Hat Enterprise Linux with KVM Releases</th>
<th>Pass-Through GPU - Red Hat Enterprise Linux with KVM Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Hat Enterprise Linux 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td></td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td><strong>Deprecated:</strong> CentOS 7.6-7.8 See Note (1)</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td></td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
</tbody>
</table>

**Note:**
1. CentOS is not a certified guest OS for Red Hat Enterprise Linux with KVM or RHV.

### 2.3.2.2. Linux Guest OS Support in Release 16.1

<table>
<thead>
<tr>
<th>Guest OS</th>
<th>NVIDIA vGPU - Red Hat Enterprise Linux with KVM Releases</th>
<th>Pass-Through GPU - Red Hat Enterprise Linux with KVM Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Hat CoreOS 4.11</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 9.2</td>
<td>RHEL KVM 9.2, 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
</tr>
<tr>
<td>Rocky Linux 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 8.8</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td>Rocky Linux 8.4</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td><strong>Deprecated:</strong> CentOS Linux 8 (2105)</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td><strong>Deprecated:</strong> CentOS 7.6-7.8 See Note (1)</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
</tbody>
</table>
### Validated Platforms

#### Virtual GPU Software R535 for Red Hat Enterprise Linux with KVM

<table>
<thead>
<tr>
<th>Guest OS</th>
<th>NVIDIA vGPU - Red Hat Enterprise Linux with KVM Releases</th>
<th>Pass-Through GPU - Red Hat Enterprise Linux with KVM Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RHV 4.4, 4.3, 4.2</td>
<td>RHV 4.4, 4.3, 4.2</td>
</tr>
</tbody>
</table>

**Note:**
1. CentOS is not a certified guest OS for Red Hat Enterprise Linux with KVM or RHV.

#### 2.3.2.3. Linux Guest OS Support in Release 16.0

<table>
<thead>
<tr>
<th>Guest OS</th>
<th>NVIDIA vGPU - Red Hat Enterprise Linux with KVM Releases</th>
<th>Pass-Through GPU - Red Hat Enterprise Linux with KVM Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Hat CoreOS 4.11</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 9.2</td>
<td>RHEL KVM 9.2, 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
</tr>
<tr>
<td>Rocky Linux 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
<td>RHEL KVM 9.2, 9.0</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 8.8</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 8.6</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td>Rocky Linux 8.4</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td><strong>Deprecated:</strong> CentOS Linux 8 (2105)</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
<tr>
<td><strong>Deprecated:</strong> CentOS 7.6-7.8 See Note (1)</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
<td>RHEL KVM 9.2, 9.0, 8.8, 8.6, 7.9</td>
</tr>
</tbody>
</table>

**Note:**
1. CentOS is not a certified guest OS for Red Hat Enterprise Linux with KVM or RHV.
2.4. NVIDIA CUDA Toolkit Version Support

The releases in this release family of NVIDIA vGPU software support NVIDIA CUDA Toolkit 12.1.

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 Release Notes for CUDA 12.1.0.

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 12.1 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. 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.

2.5.1. 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.
- 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 Ada Lovelace Architecture**

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Since 16.1:</strong> NVIDIA L40S</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>NVIDIA L40</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>NVIDIA L4</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>NVIDIA RTX 6000 Ada</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td><strong>Since 16.1:</strong> NVIDIA RTX 5000 Ada</td>
<td>All Q-series vGPUs</td>
</tr>
</tbody>
</table>

**Multiple vGPU Support on the NVIDIA Ampere GPU Architecture**

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA A40</td>
<td>All Q-series vGPUs See Note (1).</td>
</tr>
<tr>
<td>NVIDIA A16</td>
<td>All Q-series vGPUs See Note (1).</td>
</tr>
<tr>
<td>NVIDIA A10</td>
<td>All Q-series vGPUs See Note (1).</td>
</tr>
<tr>
<td>NVIDIA A2</td>
<td>All Q-series vGPUs See Note (1).</td>
</tr>
<tr>
<td>NVIDIA RTX A6000</td>
<td>All Q-series vGPUs See Note (1).</td>
</tr>
<tr>
<td>NVIDIA RTX A5500</td>
<td>All Q-series vGPUs See Note (1).</td>
</tr>
<tr>
<td>NVIDIA RTX A5000</td>
<td>All Q-series vGPUs See Note (1).</td>
</tr>
</tbody>
</table>

**Multiple vGPU Support on the NVIDIA Turing GPU Architecture**

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla T4</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>Quadro RTX 6000</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>Quadro RTX 6000 passive</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>Quadro RTX 8000</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>Quadro RTX 8000 passive</td>
<td>All Q-series vGPUs</td>
</tr>
</tbody>
</table>
Multiple vGPU Support on the NVIDIA Volta GPU Architecture

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla V100 SXM2 32GB</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>Tesla V100 PCIe 32GB</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>Tesla V100S PCIe 32GB</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>Tesla V100 SXM2</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>Tesla V100 PCIe</td>
<td>All Q-series vGPUs</td>
</tr>
<tr>
<td>Tesla V100 FHHL</td>
<td>All Q-series vGPUs</td>
</tr>
</tbody>
</table>

Multiple vGPU Support on the NVIDIA Pascal GPU Architecture

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla P100 SXM2</td>
<td>P100X-16Q</td>
</tr>
<tr>
<td>Tesla P100 PCIe 16GB</td>
<td>P100-16Q</td>
</tr>
<tr>
<td>Tesla P100 PCIe 12GB</td>
<td>P100C-12Q</td>
</tr>
<tr>
<td>Tesla P40</td>
<td>P40-24Q</td>
</tr>
<tr>
<td>Tesla P6</td>
<td>P6-16Q</td>
</tr>
<tr>
<td>Tesla P4</td>
<td>P4-8Q</td>
</tr>
</tbody>
</table>

Multiple vGPU Support on the NVIDIA Maxwell GPU Architecture

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla M60</td>
<td>M60-8Q</td>
</tr>
<tr>
<td>Tesla M10</td>
<td>M10-8Q</td>
</tr>
<tr>
<td>Tesla M6</td>
<td>M6-8Q</td>
</tr>
</tbody>
</table>

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

2.5.2. Maximum Number of vGPUs Supported per VM

NVIDIA vGPU software supports up to a maximum of 16 vGPUs per VM.
2.5.3. Hypervisor Releases that Support Multiple vGPUs Assigned to a VM

All hypervisor releases that support NVIDIA vGPU software are supported.

2.6. 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.

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

Only Q-series and C-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

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA A800 PCIe 40GB active cooled</td>
<td>A800-40C</td>
</tr>
<tr>
<td>NVIDIA A40</td>
<td>A40-48Q</td>
</tr>
<tr>
<td>NVIDIA A10</td>
<td>A10-24Q</td>
</tr>
<tr>
<td>NVIDIA RTX A6000</td>
<td>A6000-48Q</td>
</tr>
<tr>
<td>NVIDIA RTX A5500</td>
<td>A5500-24Q</td>
</tr>
<tr>
<td>NVIDIA RTX A5000</td>
<td>A5000-24Q</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadro RTX 6000</td>
<td>RTX6000-24Q</td>
</tr>
<tr>
<td>Quadro RTX 6000 passive</td>
<td>RTX6000P-24Q</td>
</tr>
<tr>
<td>Quadro RTX 8000</td>
<td>RTX8000-48Q</td>
</tr>
</tbody>
</table>
**Validated Platforms**

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

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadro RTX 8000 passive</td>
<td>RTX8000P-48Q</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla V100 SXM2 32GB</td>
<td>V100DX-32Q</td>
</tr>
<tr>
<td>Tesla V100 SXM2</td>
<td>V100X-16Q</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla P100 SXM2</td>
<td>P100X-16Q</td>
</tr>
</tbody>
</table>

**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.

2.6.2. 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.

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

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

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

- NVIDIA NVSwitch is supported only on the hardware platforms, vGPUs, and hypervisor software releases listed in `unique_34`. Otherwise, only direct connections are supported.
2.7. 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.

2.7.1. vGPUs that Support Unified Memory

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

Unified Memory Support on the NVIDIA Ada Lovelace GPU Architecture

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA L40</td>
<td>L40-48Q</td>
</tr>
<tr>
<td><strong>Since 16.1:</strong> NVIDIA L40S</td>
<td>L40S-48Q</td>
</tr>
<tr>
<td>NVIDIA L4</td>
<td>L4-24Q</td>
</tr>
<tr>
<td>NVIDIA RTX 6000 Ada</td>
<td>RTX 6000 Ada-48Q</td>
</tr>
<tr>
<td><strong>Since 16.1:</strong> NVIDIA RTX 5000 Ada</td>
<td>RTX 5000 Ada-32Q</td>
</tr>
</tbody>
</table>

Unified Memory Support on the NVIDIA Ampere GPU Architecture

<table>
<thead>
<tr>
<th>Board</th>
<th>vGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA A800 PCIe 40GB active cooled</td>
<td>A800-40C</td>
</tr>
<tr>
<td>All MIG-backed vGPUs</td>
<td></td>
</tr>
<tr>
<td>NVIDIA A40</td>
<td>A40-48Q</td>
</tr>
<tr>
<td>NVIDIA A16</td>
<td>A16-16Q</td>
</tr>
<tr>
<td>NVIDIA A10</td>
<td>A10-24Q</td>
</tr>
<tr>
<td>NVIDIA A2</td>
<td>A2-16Q</td>
</tr>
</tbody>
</table>
2.7.2. Guest OS Releases that Support Unified Memory

Linux only. Unified memory is not supported on Windows.

2.7.3. 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.

2.8. 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.

<table>
<thead>
<tr>
<th>Hypervisor Software Release</th>
<th>Container Platform</th>
<th>Guest OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Hat Enterprise Linux with KVM 8.2</td>
<td>Red Hat Openshift 4.11 with Red Hat Enterprise Linux CoreOS and the CRI-O container runtime</td>
<td>Red Hat CoreOS 4.11</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux with KVM 8.2</td>
<td>Red Hat Openshift 4.8 with Red Hat Enterprise Linux CoreOS and the CRI-O container runtime</td>
<td>Red Hat CoreOS 4.8</td>
</tr>
</tbody>
</table>

2.9. 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
- Since 16.1: NVIDIA L40S
- NVIDIA L4
- NVIDIA RTX 6000 Ada
- Since 16.1: NVIDIA RTX 5000 Ada
- NVIDIA A40
- NVIDIA A16
- NVIDIA A2
- NVIDIA A10
- NVIDIA RTX A6000
- NVIDIA RTX A5500
- NVIDIA RTX A5000
- Tesla T4
- Quadro RTX 8000
- Quadro RTX 8000 passive
- Quadro RTX 6000
- Quadro RTX 6000 passive

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

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

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: 0xl1
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0): 0xc1d004a1, 0xff0e0000, 0xff0400fb, 0xc36f, 0x80000000, 0xff0e0200, 0x0, 0x0, (Not logged),
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0): , 0x0
```

Virtual GPU Software R535 for Red Hat Enterprise Linux with KVM

RN-08687-001_v16.0 through 16.2 | 28
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:

\[
\text{max-reserved-fb} = \frac{\text{vgpu-profile-size-in-mb}}{16} + 16 + \text{ecc-adjustments} + \text{page-retirement-allocation} + \text{compression-adjustment}
\]

**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.

<table>
<thead>
<tr>
<th>vGPU Type</th>
<th>Compression Adjustment (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T4-16Q</td>
<td>28</td>
</tr>
<tr>
<td>T4-16C</td>
<td>28</td>
</tr>
<tr>
<td>T4-16A</td>
<td>28</td>
</tr>
<tr>
<td>RTX6000-12Q</td>
<td>32</td>
</tr>
<tr>
<td>RTX6000-12C</td>
<td>32</td>
</tr>
<tr>
<td>RTX6000-12A</td>
<td>32</td>
</tr>
<tr>
<td>RTX6000-24Q</td>
<td>104</td>
</tr>
<tr>
<td>RTX6000-24C</td>
<td>104</td>
</tr>
<tr>
<td>RTX6000-24A</td>
<td>104</td>
</tr>
<tr>
<td>RTX6000P-12Q</td>
<td>32</td>
</tr>
<tr>
<td>RTX6000P-12C</td>
<td>32</td>
</tr>
<tr>
<td>vGPU Type</td>
<td>Compression Adjustment (MB)</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>RTX6000P-12A</td>
<td></td>
</tr>
<tr>
<td>RTX6000P-24Q</td>
<td>104</td>
</tr>
<tr>
<td>RTX6000P-24C</td>
<td></td>
</tr>
<tr>
<td>RTX6000P-24A</td>
<td></td>
</tr>
<tr>
<td>RTX8000-12Q</td>
<td>32</td>
</tr>
<tr>
<td>RTX8000-12C</td>
<td></td>
</tr>
<tr>
<td>RTX8000-12A</td>
<td></td>
</tr>
<tr>
<td>RTX8000-16Q</td>
<td>64</td>
</tr>
<tr>
<td>RTX8000-16C</td>
<td></td>
</tr>
<tr>
<td>RTX8000-16A</td>
<td></td>
</tr>
<tr>
<td>RTX8000-24Q</td>
<td>96</td>
</tr>
<tr>
<td>RTX8000-24C</td>
<td></td>
</tr>
<tr>
<td>RTX8000-24A</td>
<td></td>
</tr>
<tr>
<td>RTX8000-48Q</td>
<td>238</td>
</tr>
<tr>
<td>RTX8000-48C</td>
<td></td>
</tr>
<tr>
<td>RTX8000-48A</td>
<td></td>
</tr>
<tr>
<td>RTX8000P-12Q</td>
<td>32</td>
</tr>
<tr>
<td>RTX8000P-12C</td>
<td></td>
</tr>
<tr>
<td>RTX8000P-12A</td>
<td></td>
</tr>
<tr>
<td>RTX8000P-16Q</td>
<td>64</td>
</tr>
<tr>
<td>RTX8000P-16C</td>
<td></td>
</tr>
<tr>
<td>RTX8000P-16A</td>
<td></td>
</tr>
<tr>
<td>RTX8000P-24Q</td>
<td>96</td>
</tr>
<tr>
<td>RTX8000P-24C</td>
<td></td>
</tr>
<tr>
<td>RTX8000P-24A</td>
<td></td>
</tr>
<tr>
<td>RTX8000P-48Q</td>
<td>238</td>
</tr>
<tr>
<td>RTX8000P-48C</td>
<td></td>
</tr>
<tr>
<td>RTX8000P-48A</td>
<td></td>
</tr>
</tbody>
</table>

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 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.

3.10. 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 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.

### 3.11. 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 \texttt{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
```

   \textbf{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 \texttt{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:

```bash
# 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.

```bash
# 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 16.2

<table>
<thead>
<tr>
<th>Bug ID</th>
<th>Summary and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4309888</td>
<td><strong>16.0, 16.1 Only: NVWMI functions for faking EDID have no effect</strong></td>
</tr>
<tr>
<td></td>
<td>The NVIDIA Enterprise Management Toolkit (NVWMI) functions for faking Extended Display Identification Data (EDID),</td>
</tr>
<tr>
<td></td>
<td>namely, fakeEDID, fakeEDIDAll, and fakeEDIDOnPort have no effect. This issue affects only Windows guest VMs and</td>
</tr>
<tr>
<td></td>
<td>can prevent a VM from being enabled with multiple displays. When this issue occurs, unable to fake EDID events</td>
</tr>
<tr>
<td></td>
<td>can be seen in Event Viewer.</td>
</tr>
</tbody>
</table>

4.2. Issues Resolved in Release 16.1

<table>
<thead>
<tr>
<th>Bug ID</th>
<th>Summary and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4142288</td>
<td><strong>16.0 Only: The NVIDIA L40 GPU brand is incorrectly identified if GSP firmware is disabled</strong></td>
</tr>
<tr>
<td></td>
<td>If GPU System Processor (GSP) firmware is disabled, the NVIDIA Virtual GPU Manager incorrectly identifies the brand of the NVIDIA L40 GPU. This</td>
</tr>
<tr>
<td></td>
<td>incorrect identification of the GPU brand might cause performance degradation with some applications that are optimised for features of the NVIDIA</td>
</tr>
<tr>
<td></td>
<td>L40 that are not available in the incorrect brand. However, the output from the nvidia-smi command is not affected.</td>
</tr>
<tr>
<td>4052185</td>
<td><strong>16.0 Only: Benign No such file or directory message during capture of configuration data for a bug report</strong></td>
</tr>
<tr>
<td>Bug ID</td>
<td>Summary and Description</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| 3641947 | **16.0 Only: Graphics applications are corrupted on some Windows vGPU VMs**  
Graphics applications are corrupted on Windows VMs that are configured with one or more vGPUs that are based on the NVIDIA Ampere or NVIDIA Ada Lovelace GPU architecture. |
| 4096848 | **Optical Flow object allocation fails on VMs configured with vGPUs based on the NVIDIA Ampere architecture**  
Optical Flow object allocation fails on VMs configured with vGPUs that reside on GPUs based on the NVIDIA Ampere GPU architecture. This issue has been observed as the failure of the Omniverse Kit container on a VM configured with NVIDIA vGPU. |
| 3936030 | **CUDA applications fail on any VM configured with multiple vGPUs when unified memory is enabled**  
CUDA applications fail on any VM configured with multiple vGPUs based on the NVIDIA Ada Lovelace GPU architecture when unified memory is enabled for the VM. Whenever a CUDA application fails, the following message is observed on the hypervisor host:  
VGPU message 2 failed, result code: 0xff100004 |
| 3334310 | **NVIDIA Control Panel is started only for the RDP user that logs on first**  
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. |
Chapter 5. Known Issues

5.1. Pixelation occurs on a Windows VM configured with a Tesla T4 vGPU

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

Workaround
Contact NVIDIA Enterprise Support for assistance with a workaround for this issue.

Status
Open

Ref. #
3973158

5.2. NVIDIA Control Panel is not available in multiuser environments

Description
After the NVIDIA vGPU software graphics driver for Windows is installed, the NVIDIA Control Panel app might be missing from the system. This issue typically occurs when multiple users connect to virtual machines by using remote desktop applications such as Microsoft RDP, VMware Horizon, and Citrix Virtual Apps and Desktops.
Known Issues

Virtual GPU Software R535 for Red Hat Enterprise Linux with KVM  

RN-08687-001 _v16.0 through 16.2   |   40

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

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

- To use the Windows Settings app:
  1. From the Windows Start menu, choose Settings > Apps > Apps & feautures.
  2. In the Apps & features window, type nvidia control panel in the search box and confirm that the NVIDIA Control Panel app is found.

- To use the Get-AppxPackageWindows PowerShell command:
  1. Run Windows PowerShell as Administrator.
  2. Determine whether the NVIDIA Control Panel app is installed for the current user.
     ```
     PS C:\> Get-AppxPackage -Name NVIDIACorp.NVIDIAControlPanel
     ```
  3. Determine whether the NVIDIA Control Panel app is installed for all users.
     ```
     PS C:\> Get-AppxPackage -AllUsers -Name NVIDIACorp.NVIDIAControlPanel
     ```
     This example shows that the NVIDIA Control Panel app is installed for the users Administrator, pliny, and trajan.
     ```
     Name                      : NVIDIACorp.NVIDIAControlPanel
     Publisher                 : CN=D6816951-877F-493B-B4EE-41AB9419C326
     Architecture             : X64
     ResourceId               :
     Version                  : 8.1.964.0
     PackageFullName          :
     InstallLocation          : C:\Program Files\WindowsApps\NVIDIACorp.NVIDIAControlPanel_8.1.964.0_x64__56jybvy8sckqj
     PackageFamilyName        :
     PublisherId              : 56jybvy8sckqj
     PackageUserInformation   :
     {S-1-12-1-530092550-1307989247-1105462437-500 [Administrator]: Installed,
      S-1-12-1-530092550-1307989247-1105462437-1002 [pliny]: Installed,
      S-1-12-1-530092550-1307989247-1105462437-1003 [trajan]: Installed
     }
     IsResourcePackage        : False
     IsBundle                 : False
     IsDevelopmentMode        : False
     NonRemovable             : False
     IsPartiallyStaged        : False
     SignatureKind            : Store
     Status                   : Ok
     ```

Preventing this Issue

To prevent this issue from occurring, ensure that:
Known Issues

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

Workaround

If the NVIDIA Control Panel app is missing from a system that is running Windows 11 or a modern version of Windows 10, you can install the NVIDIA Control Panel app by using the winget command-line tool of Windows Package Manager.

Note: The winget command-line tool is not available on the Windows Server OS.

Before using the winget command-line tool to install the NVIDIA Control Panel app, ensure that the following prerequisites are met:

- Your system is connected to the Internet.
- The Microsoft Store app is enabled.
- Packages on which winget depends, such as Microsoft.UI.Xaml and Microsoft.VCLibs.x64, are installed.

To use the winget command-line tool to install the NVIDIA Control Panel app, run the following command:

```
PS C:\> winget install "NVIDIA Control Panel" --id 9NF8H0H7WMLT -s msstore
        --accept-package-agreements --accept-source-agreements
```

For information about how to download and use the latest winget version, refer to Use the winget tool to install and manage applications on the Microsoft documentation site.

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

Status
Open

Ref. #
3999308

5.3. 16.0, 16.1 Only: NVWMI functions for faking EDID have no effect

Description

The NVIDIA Enterprise Management Toolkit (NVWMI) functions for faking Extended Display Identification Data (EDID), namely, fakeEDID, fakeEDIDAll, and fakeEDIDOnPort...
have no effect. This issue affects only Windows guest VMs and can prevent a VM from being enabled with multiple displays. When this issue occurs, unable to fake EDID events can be seen in Event Viewer.

Status
Resolved in NVIDIA vGPU software 16.2

Ref. #
4309888

5.4. 16.0 Only: Benign No such file or directory message during capture of configuration data for a bug report

Description
While the nvidia-bug-report.sh script on is running on the hypervisor host to capture configuration data for a bug report, the following error message is displayed:

sysctl: cannot stat /proc/sys/vm/compression_proactiveness: No such file or directory

Workaround
Ignore this message as it is benign. The bug report is generated correctly.

Status
Resolved in NVIDIA vGPU software 16.1

Ref. #
4052185
5.5. 16.0 Only: The NVIDIA L40 GPU brand is incorrectly identified if GSP firmware is disabled

Description
If GPU System Processor (GSP) firmware is disabled, the NVIDIA Virtual GPU Manager incorrectly identifies the brand of the NVIDIA L40 GPU. This incorrect identification of the GPU brand might cause performance degradation with some applications that are optimised for features of the NVIDIA L40 that are not available in the incorrect brand. However, the output from the `nvidia-smi` command is not affected.

This issue occurs only if GPU System Processor (GSP) firmware is disabled. It does not occur if GSP firm is enabled.

Status
Resolved in NVIDIA vGPU software 16.1

Ref. #
4142288

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

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

Version
This issue affects all supported Windows Server guest OS releases.

Status
Open
5.7. Remote desktop connection is lost and the NVIDIA vGPU software graphics driver is unloaded

Description

The remote desktop connection is lost and the NVIDIA vGPU software graphics driver is unloaded after an attempt to access a VM over RDP and VMware Horizon agent direct connect. After an attempt to log in again, a black screen is displayed.

When this issue occurs, the following errors are written to the log files on the guest VM:

- A timeout detection and recovery (TDR) error:
  
  ```
  vmiop_log: (0x0): Timeout occurred, reset initiated.
  vmiop_log: (0x0): TDR_DUMP:0x52445456 0x006907d0 0x000001cc 0x00000001
  ```

- XID error 43:
  
  ```
  vmiop_log: (0x0): XID 43 detected on physical_chid
  ```

- vGPU error 22:
  
  ```
  vmiop_log: (0x0): VGPU message 22 failed
  ```

- Guest driver unloaded error:
  
  ```
  vmiop_log: (0x0): Guest driver unloaded!
  ```

Workaround

To recover from this issue, reboot the VM.

To prevent this issue from occurring, disable translation lookaside buffer (TLB) invalidation by setting the vGPU plugin parameter `tlb_invalidate_enabled` to 0.

Status

Open

Ref. #

4086605

3596327
5.8. 16.0 Only: Graphics applications are corrupted on some Windows vGPU VMs

Description
Graphics applications are corrupted on Windows VMs that are configured with one or more vGPUs that are based on the NVIDIA Ampere or NVIDIA Ada Lovelace GPU architecture.

Status
Resolved in NVIDIA vGPU software 16.1

Ref. #
3641947

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

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

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

Status
Open

Ref. #
4020171
5.10. **CUDA profilers cannot gather hardware metrics on NVIDIA vGPU**

Description

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

Version

This issue affects NVIDIA vGPU software releases starting with 15.2.

Status

Open

Ref. #

4041169

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

Description

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

Workaround

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

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

Note: If this Windows registry key is set to 1 or deleted, the remote call to ngx.download.nvidia.com is enabled again.
5.12. On NVIDIA H100, creation of multiple compute instances after deletion of existing compute instances fails

Description

After compute instances are created and deleted on an NVIDIA H100 GPU, creation of multiple instances in a single `nvidia-smi` command fails. For example, the command `nvidia-smi mig -cci 0,1,2` fails with the following error message:

```
Unable to create a compute instance on GPU  0 GPU instance ID  0 using profile 0:
  Invalid Argument
Failed to create compute instances: Invalid Argument
```

Workaround

Create each compute instance in a separate `nvidia-smi` command, for example:

```
$ nvidia-smi mig -cci 0
$ nvidia-smi mig -cci 1
$ nvidia-smi mig cci 2
```
5.13. **NLS client fails to acquire a license with the error** The allowed time to process response has expired

**Description**

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

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

**Workaround**

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

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

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

**Status**

Not a bug

**Ref. #**

3859889
5.14. **NVIDIA vGPU software graphics driver fails to load on KVM-based hypervisors**

**Description**

The NVIDIA vGPU software graphics driver fails to load on hypervisors based on Linux with KVM. This issue affects UEFI VMs configured with a vGPU or pass-through GPU that requires a large BAR address space. This issue does not affect VMs that are booted in legacy BIOS mode. The issue occurs because BAR resources are not mapped into the VM.

On a Windows VM, error code 12 is reported in **Device Manager** for the vGPU or pass-through GPU.

**Workaround**

1. In `virsh`, open for editing the XML document of the VM to which the vGPU or GPU is assigned.
   ```bash
   # virsh edit vm-name
   vm-name
   ```
   The name of the VM to which the vGPU or GPU is assigned.

2. Declare the custom `libvirt` XML namespace that supports command-line pass through of QEMU arguments.

   Declare this namespace by modifying the start tag of the top-level `domain` element in the first line of the XML document.

   ```xml
   <domain type='kvm' xmlns:qemu='http://libvirt.org/schemas/domain/qemu/1.0'>
   ```

3. At the end of the XML document, between the `</devices>` end tag and the `</domain>` end tag, add the highlighted `qemu` elements.

   These elements pass the QEMU arguments for mapping the required BAR resources into the VM.

   ```xml
   </devices>
   <qemu:commandline>
     <qemu:arg value='-fw_cfg'/>
     <qemu:arg value='opt/ovmf/X-PciMmio64Mb,string=262144'/>
   </qemu:commandline>
   </domain>
   ```

4. Start the VM to which the vGPU or GPU is assigned.

   ```bash
   # virsh start vm-name
   vm-name
   ```
   The name of the VM to which the vGPU or GPU is assigned.
5.15. With multiple active sessions, **NVIDIA Control Panel** incorrectly shows that the system is unlicensed

**Description**

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

The Manage License page incorrectly shows that the system is unlicensed because of stale data in **NVIDIA Control Panel** in an environment with multiple sessions. The data is stale because **NVIDIA Control Panel** fails to get and update the settings for remote sessions when multiple sessions or no sessions are active in the VM. The **NVIDIA Display Container** service crashes when a session is reconnected because the session is not active at the moment of reconnection.
5.16. VP9 and AV1 decoding with web browsers are not supported on Microsoft Windows Server 2019

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

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

Status
Not an NVIDIA bug

Ref. #
200756564

5.17. 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 : 537.13
  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
5.18. After an upgrade of the Linux graphics driver from an RPM package in a licensed VM, licensing fails

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

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

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

```bash
$ rpm -iv nvidia-linux-grid-525_535.129.03_amd64.rpm
```

Status
Open

Ref. #
3512766

5.19. 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.
5.20. 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

5.21. 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

### 5.22. 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.
Known Issues

**Virtual GPU Software R535 for Red Hat Enterprise Linux with KVM**

<table>
<thead>
<tr>
<th>vGPU</th>
<th>4K Displays Supported with NVENC Enabled</th>
</tr>
</thead>
<tbody>
<tr>
<td>2Q</td>
<td>1</td>
</tr>
<tr>
<td>3Q</td>
<td>2</td>
</tr>
<tr>
<td>4Q</td>
<td>3</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>vGPU</th>
<th>4K Displays Supported with NVENC Disabled</th>
</tr>
</thead>
<tbody>
<tr>
<td>2Q</td>
<td>2</td>
</tr>
<tr>
<td>3Q</td>
<td>2</td>
</tr>
<tr>
<td>4Q</td>
<td>4</td>
</tr>
</tbody>
</table>

- 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

### 5.23. NVIDIA A100 HGX 80GB vGPU names shown as Graphics Device by nvidia-smi

**Description**

The names of vGPUs that reside on the NVIDIA A100 80GB GPU are incorrectly shown as Graphics Device by the `nvidia-smi` command. The correct names indicate the vGPU type, for example, A100DX-40C.

```bash
$ nvidia-smi
Mon Jan 25 02:52:57 2021
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 460.32.04    Driver Version: 460.32.04    CUDA Version: 11.2     |
|-------------------------------+----------------------+----------------------+
| 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 Graphics Device On       | 00000000:07:00.0 Off|         00%      Default |
| N/A          N/A P0 N/A / N/A| 6053MiB / 81915MiB |              Disabled |
+-------------------------------+----------------------+----------------------+
```

Virtual GPU Software R535 for Red Hat Enterprise Linux with KVM
### Known Issues

**Virtual GPU Software R535 for Red Hat Enterprise Linux with KVM**

#### RN-08687-001 _v16.0 through 16.2 |

<table>
<thead>
<tr>
<th>Graphics Device</th>
<th>On</th>
<th>00000000:08:00.0 Off</th>
<th>0%</th>
<th>Default</th>
<th>Disabled</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A N/A P0 N/A / N/A</td>
<td>6053MiB / 81915MiB</td>
<td>0%</td>
<td>Default</td>
<td>Disabled</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Processes:</th>
<th>GPU</th>
<th>GI</th>
<th>CI</th>
<th>PID</th>
<th>Type</th>
<th>Process name</th>
<th>GPU Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Usage</td>
</tr>
</tbody>
</table>

| No running processes found |

### Status

Open

### Ref. #

200691204

### 5.24. 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
5.25. 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

5.26. 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. A GPU is affected by this issue only if the sriov-manage script has **not** been run to enable the virtual function for the GPU in the **sysfs** file system.

```
[root@host ~]# nvidia-smi
Fri Nov 10 11:45:28 2023
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 535.129.03   Driver Version: 535.129.03  CUDA Version:  12.1     |
|-------------------------------+----------------------+----------------------+
| 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 |
Workaround

Run the `sriov-manage` script to enable the virtual function for the GPU in the `sysfs` file system as explained in Virtual GPU Software User Guide.

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

```bash
root@host ~]# nvidia-smi
Fri Nov 10 11:47:38 2023

+-----------------------------------------------------------------------------+
| NVIDIA-SMI 535.129.03   Driver Version: 535.129.03  CUDA Version: 12.1     |
|-------------------------------+----------------------+----------------------+
| 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
5.27. Guest VM frame buffer listed by `nvidia-smi` for vGPUs on GPUs that support SRIOV is incorrect

Description
The amount of frame buffer listed in a guest VM by the `nvidia-smi` command for vGPUs on GPUs that support Single Root I/O Virtualization (SR-IOV) is incorrect. Specifically, the amount of frame buffer listed is the amount of frame buffer allocated for the vGPU type minus the size of the VMMU segment (`vmmu_page_size`). Examples of GPUs that support SRIOV are GPUs based on the NVIDIA Ampere architecture, such as NVIDIA A100 PCIe 40GB or NVIDIA A100 HGX 40GB.

For example, frame buffer for -4C and -20C vGPU types is listed as follows:
- For -4C vGPU types, frame buffer is listed as 3963 MB instead of 4096 MB.
- For -20C vGPU types, frame buffer is listed as 20347 MB instead of 20480 MB.

Status
Open

Ref. #
200524749

5.28. VMs fail to boot on RHV 4.4

Description
On RHV 4.4, VMs fail to boot with the error `Host doesn't support passthru of host PCI device`. This issue affects GPU pass through deployments with all supported GPUs and NVIDIA vGPU deployments with GPUs based on the NVIDIA Ampere architecture. This issue occurs because the `intel_iommu` parameter and the `nouveau.modeset` parameter are not set correctly.

Version
This issue affects RHV 4.4.
Workaround

Perform this workaround on the hypervisor host. This workaround requires root user privileges on the hypervisor host.

1. In a plain-text editor, edit the file `/boot/loader/entries/rhvh-4.4.1.1-0.20200722.0+1-4.18.0-193.13.2.e18_2.x86_64.conf` to add the following options to the boot options.
   - `nouveau.modeset=0`
   - `intel_iommu=on`

   **Note:** Line breaks have been added to this example to enhance readability.

   ```
   title rhvh-4.4.1.1-0.20200722.0 (4.18.0-193.13.2.e18_2.x86_64)
   version 4.18.0-193.13.2.e18_2.x86_64
   linux //rhvh-4.4.1.1-0.20200722.0+1/vmlinuz-4.18.0-193.13.2.e18_2.x86_64
   initrd //rhvh-4.4.1.1-0.20200722.0+1/initramfs-4.18.0-193.13.2.e18_2.x86_64.img
   options crashkernel=auto resume=/dev/mapper/rhvh00-swap \
   rd.lvm.lv=rhvh00/rhvh-4.4.1.1-0.20200722.0+1 rd.lvm.lv=rhvh00/swap \
   root=/dev/rhvh00/rhvh-4.4.1.1-0.20200722.0+1 \
   boot=UUID=38ff2175-b761-403d-8a91-d7ec9f7ec2f7 rootflags=discard \
   img.bootid=rhvh-4.4.1.1-0.20200722.0+1 intel_iommu=on nouveaumodeset=0
   grub_users $grub_users
   grub_arg --unrestricted
   grub_class kernel
   ```

2. Reboot the hypervisor host machine.

Status

Not an NVIDIA bug

Ref. #

200653675

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.
Known Issues

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. 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
5.32. 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

5.33. 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.
5.34. 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. #**

200523086

200505777
5.35. 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

5.36. 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

5.37. 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.

2. Set the Use the hardware default graphics adapter for all Remote Desktop Services sessions option.
5.38. 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    win7-xmpl-146048-2    |     39%    |
|                          | 88095    win7-xmpl-146048-3    |     26%    |
|                          | 89170    win7-xmpl-146048-4    |      0%    |
|                          | 90475    win7-xmpl-146048-5    |      0%    |
|                          | 93363    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
5.39. License is not acquired in Windows VMs

Description

When a windows VM configured with a licensed vGPU is started, the VM fails to acquire a license.

Error messages in the following format are written to the NVIDIA service logs:

```
[000000020.860152600 sec] - [Logging.lib] ERROR: [nvGridLicensing.FlexUtility] 353@FlexUtility::LogFneError : Error: Failed to add trusted storage. Server URL : license-server-url - [1,7E2,2,1[7000003F,0,9B00A7]]
```

System machine type does not match expected machine type.

Workaround

This workaround requires administrator privileges.

1. Stop the NVIDIA Display Container LS service.
2. Delete the contents of the folder `%SystemDrive%\Program Files\NVIDIA Corporation\Grid Licensing`.
3. Start the NVIDIA Display Container LS service.

Status

Closed

Ref. #

200407287

5.40. 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.
5.41. Hot plugging and unplugging vCPUs causes a blue-screen crash in Windows VMs

Description

Hot plugging or unplugging vCPUs causes a blue-screen crash in Windows VMs that are running NVIDIA vGPU software graphics drivers.

When the blue-screen crash occurs, one of the following error messages may also be seen:

- SYSTEM_SERVICE_EXCEPTION(nvlddmkm.sys)
- DRIVER_IRQL_NOT_LESS_OR_EQUAL(nvlddmkm.sys)

NVIDIA vGPU software graphics drivers do not support hot plugging and unplugging of vCPUs.

Status

Closed

Ref. #

200407230

5.42. Luxmark causes a segmentation fault on an unlicensed Linux client

Description

If the Luxmark application is run on a Linux guest VM configured with NVIDIA vGPU that is booted without acquiring a license, a segmentation fault occurs and the application core dumps. The fault occurs when the application cannot allocate a CUDA object on
NVIDIA vGPUs where CUDA is disabled. On NVIDIA vGPUs that can support CUDA, CUDA is disabled in unlicensed mode.

**Status**
Not an NVIDIA bug.

**Ref. #**
200330956

### 5.43. 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
5.44. 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](https://example.com).

**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
5.45. 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

5.46. VM bug checks after the guest VM driver for Windows 10 RS2 is installed

**Description**

When the VM is rebooted after the guest VM driver for Windows 10 RS2 is installed, the VM bug checks. When Windows boots, it selects one of the standard supported video modes. If Windows is booted directly with a display that is driven by an NVIDIA driver, for example a vGPU on Citrix Hypervisor, a blue screen crash occurs.
This issue occurs when the screen resolution is switched from VGA mode to a resolution that is higher than 1920×1200.

**Fix**

Download and install [Microsoft Windows Update KB4020102](#) from the Microsoft Update Catalog.

**Workaround**

If you have applied the fix, ignore this workaround.

Otherwise, you can work around this issue until you are able to apply the fix by not using resolutions higher than 1920×1200.

1. Choose a GPU profile in Citrix XenCenter that does not allow resolutions higher than 1920×1200.
2. Before rebooting the VM, set the display resolution to 1920×1200 or lower.

**Status**

Not an NVIDIA bug

**Ref. #**

200310861

### 5.47. 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
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 OR IN ANY WAY RELATING TO THE MATERIALS OR THE USE THEREOF. IN NO EVENT WILL NVIDIA'S AGGREGATE LIABILITY EXCEED THE ACTUAL AMOUNT PAID BY CUSTOMER FOR THE PRODUCT.

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

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 OR IN ANY WAY RELATING TO THE MATERIALS OR THE USE THEREOF. IN NO EVENT WILL NVIDIA'S AGGREGATE LIABILITY EXCEED THE ACTUAL AMOUNT PAID BY CUSTOMER 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, GPDirect, 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
© 2013-2023 NVIDIA Corporation & affiliates. All rights reserved.