# TABLE OF CONTENTS

## Chapter 1. Release Notes

1. Updates in Release 8.0

## Chapter 2. Validated Platforms

2.1. Supported NVIDIA GPUs and Validated Server Platforms
2.2. Hypervisor Software Releases
2.3. Guest OS Support
   2.3.1. Windows Guest OS Support
   2.3.2. Linux Guest OS Support
2.4. NVIDIA CUDA Toolkit Version Support

## Chapter 3. Known Product Limitations

3.1. vGPU profiles with 512 Mbytes or less of frame buffer support only 1 virtual display head on Windows 10
3.2. NVENC requires at least 1 Gbyte of frame buffer
3.3. VM failures or crashes on servers with 1 TB or more of system memory
3.4. VM running older NVIDIA vGPU drivers fails to initialize vGPU when booted
3.5. Virtual GPU fails to start if ECC is enabled
3.6. Single vGPU benchmark scores are lower than pass-through GPU
3.7. VMs configured with large memory fail to initialize vGPU when booted

## Chapter 4. Resolved Issues

## Chapter 5. Security Updates

5.1. Restricting Access to GPU Performance Counters
   5.1.1. Windows: Restricting Access to GPU Performance Counters for One User by Using NVIDIA Control Panel
   5.1.2. Windows: Restricting Access to GPU Performance Counters Across an Enterprise by Using a Registry Key
   5.1.3. Linux Guest VMs: Restricting Access to GPU Performance Counters
   5.1.4. Hypervisor Host: Restricting Access to GPU Performance Counters

## Chapter 6. Known Issues

6.1. Incorrect NVIDIA vGPU software Windows graphics driver version in the installer
6.2. Quadro RTX 8000 and Quadro RTX 6000 GPUs can’t be used with VMware vSphere ESXi 6.5
6.3. Vulkan applications crash in Windows 7 guest VMs configured with NVIDIA vGPU
6.4. Host core CPU utilization is higher than expected for moderate workloads
6.5. H.264 encoder falls back to software encoding on 1Q vGPUs with a 4K display
6.6. H.264 encoder falls back to software encoding on 2Q vGPUs with 3 or more 4K displays
6.7. Frame capture while the interactive logon message is displayed returns blank screen
6.8. RDS sessions do not use the GPU with some Microsoft Windows Server releases
6.9. VMware vMotion fails gracefully under heavy load
6.10. Cloned VMs configured with a vGPU type different than the type in the master image fail to start
6.11. View session freezes intermittently after a Linux VM acquires a license
6.12. Even when the scheduling policy is equal share, unequal GPU utilization is reported...... 28
6.13. When the scheduling policy is fixed share, GPU utilization is reported as higher than expected................................................................................................................................. 29
6.14. nvidia-smi reports that vGPU migration is supported on all hypervisors.................. 30
6.15. VMs resources not available error during VMware instant clone provisioning........... 30
6.16. VMs with 32 GB or more of RAM fail to boot with GPUs requiring 64 GB of MMIO space.. 31
6.17. Module load failed during VIB downgrade from R390 to R384......................................... 32
6.18. Resolution is not updated after a VM acquires a license and is restarted.................. 33
6.19. On Tesla P40, P6, and P4 GPUs, the default ECC setting prevents NVIDIA vGPU from starting............................................................................................................................................................................. 33
6.20. Tesla P40 cannot be used in pass-through mode....................................................... 34
6.21. On Linux, 3D applications run slowly when windows are dragged............................... 34
6.22. A segmentation fault in DBus code causes nvidia-gridd to exit on Red Hat Enterprise Linux and CentOS........................................................................................................ 35
6.23. No Manage License option available in NVIDIA X Server Settings by default............. 36
6.24. Licenses remain checked out when VMs are forcibly powered off............................ 36
6.25. Memory exhaustion can occur with vGPU profiles that have 512 Mbytes or less of frame buffer.................................................................................................................................................................................. 37
6.26. vGPU VM fails to boot in ESXi 6.5 if the graphics type is Shared........................................ 39
6.27. ESXi 6.5 web client shows high memory usage even when VMs are idle....................... 40
6.28. VMs configured with NVIDIA vGPU must not be on a host in a VMware DRS cluster........ 40
6.29. GNOME Display Manager (GDM) fails to start on Red Hat Enterprise Linux 7.2 and CentOS 7.0.................................................................................................................................................................................. 41
6.30. NVIDIA Control Panel fails to start and reports that “you are not currently using a display that is attached to an Nvidia GPU”............................................................ 42
6.31. VM configured with more than one vGPU fails to initialize vGPU when booted............ 43
6.32. A VM configured with both a vGPU and a passthrough GPU fails to start the passthrough GPU.................................................................................................................................................................................. 43
6.33. vGPU allocation policy fails when multiple VMs are started simultaneously................. 44
6.34. Before Horizon agent is installed inside a VM, the Start menu’s sleep option is available 44
6.35. vGPU-enabled VMs fail to start, nvidia-smi fails when VMs are configured with too high a proportion of the server’s memory.................................................................................................................. 45
6.36. On reset or restart VMs fail to start with the error VMIOP: no graphics device is available for vGPU.................................................................................................................................................................................. 46
6.37. nvidia-smi shows high GPU utilization for vGPU VMs with active Horizon sessions........ 46
Chapter 1.
RELEASE NOTES

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

The releases in this release family of NVIDIA vGPU software include the software listed in the following table:

<table>
<thead>
<tr>
<th>Software</th>
<th>8.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA Virtual GPU Manager for the VMware vSphere releases listed in Hypervisor Software Releases</td>
<td>418.66</td>
</tr>
<tr>
<td>NVIDIA Windows driver</td>
<td>425.31</td>
</tr>
<tr>
<td>NVIDIA Linux driver</td>
<td>418.70</td>
</tr>
</tbody>
</table>

**Caution**

If you install the wrong NVIDIA vGPU software packages for the version of VMware vSphere you are using, NVIDIA Virtual GPU Manager will fail to load.

The vGPU Manager and guest VM drivers must be installed together. Older VM drivers will not function correctly with this release of vGPU Manager. Similarly, older releases of vGPU Manager will not function correctly with this release of the guest VM drivers.

See VM running older NVIDIA vGPU drivers fails to initialize vGPU when booted.

This requirement does not apply to the NVIDIA vGPU software license server. All releases of NVIDIA vGPU software are compatible with all releases of the license server.
1.1. Updates in Release 8.0

New Features in Release 8.0

- New -1B4 virtual GPU types
- vGPU migration support on the following GPUs:
  - Tesla V100 SXM2
  - Tesla V100 SXM2 32GB
  - Tesla V100 PCIe
  - Tesla V100 PCIe 32GB
  - Tesla V100 FHHL
  - Quadro RTX 6000
  - Quadro RTX 8000

  For the complete list of GPUs that support vGPU migration, see vGPU Migration Support.

- Security updates
- Miscellaneous bug fixes

Hardware and Software Support Introduced in Release 8.0

- Support for the following GPUs:
  - Quadro RTX 6000
  - Quadro RTX 8000

- Support for the following OS releases as a guest OS:
  - Windows 10 October 2018 Update (1809)
  - Windows Server 2019
Chapter 2.
VALIDATED PLATFORMS

This release family of NVIDIA vGPU software provides support for several NVIDIA GPUs on validated server hardware platforms, VMware vSphere hypervisor software versions, and guest operating systems. It also supports the version of NVIDIA CUDA Toolkit that is compatible with R418 drivers.

2.1. Supported NVIDIA GPUs and Validated Server Platforms

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

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

- Tesla V100 PCIe (vSGA is **not** supported.)
- Tesla V100 PCIe 32GB (vSGA is **not** supported.)
- Tesla V100 FHHL (vSGA is **not** supported.)
- GPUs based on the NVIDIA Turing architecture:
  - Tesla T4
  - Quadro RTX 6000
  - Quadro RTX 8000

For a list of validated server platforms, refer to [NVIDIA GRID Certified Servers](https://www.nvidia.com).

---

**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.

---

**Requirements for Using vGPU on GPUs Requiring 64 GB of MMIO Space with Large-Memory VMs**

Some GPUs require 64 GB of MMIO space. When a vGPU on a GPU that requires 64 GB of MMIO space is assigned to a VM with 32 GB or more of memory on ESXi 6.0 Update 3 and later, or ESXi 6.5 and later updates, the VM's MMIO space must be increased to 64 GB. For more information, see [VMware Knowledge Base Article: VMware vSphere VMDirectPath I/O: Requirements for Platforms and Devices (2142307)](https://kb.vmware.com/s/article/2142307).

With ESXi 6.7, no extra configuration is needed.

The following GPUs require 64 GB of MMIO space:
- Tesla P6
- Tesla P40

---

**Requirements for Using GPUs Based on the Pascal and Volta Architectures in Pass-Through Mode**

- The Tesla V100, Tesla P100, and Tesla P6 GPUs require 32 GB of MMIO space in pass-through mode.
- The Tesla P40 GPU requires 64 GB of MMIO space in pass-through mode.
- Pass through of GPUs with large BAR memory settings has some restrictions on VMware ESXi:
  - The guest OS must be a 64-bit OS.
• 64-bit MMIO and EFI boot must be enabled for the VM.
• The guest OS must be able to be installed in EFI boot mode.
• The Tesla V100, Tesla P100, and Tesla P6 require ESXi 6.0 Update 1 and later, or ESXi 6.5 and later.
• Because it requires 64 GB of MMIO space, the Tesla P40 requires ESXi 6.0 Update 3 and later, or ESXi 6.5 and later.

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

2.2. Hypervisor Software Releases

Supported VMware vSphere Hypervisor (ESXi) Releases

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

<table>
<thead>
<tr>
<th>Software</th>
<th>Release Supported</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMware vSphere Hypervisor (ESXi) 6.7</td>
<td>6.7 and compatible updates</td>
<td>All NVIDIA GPUs that support NVIDIA vGPU software are supported. Starting with release 6.7 U1, vMotion with vGPU and suspend and resume with vGPU are supported on suitable GPUs as listed in Supported NVIDIA GPUs and Validated Server Platforms. Release 6.7 supports only suspend and resume with vGPU. vMotion with vGPU is not supported on release 6.7.</td>
</tr>
</tbody>
</table>

Support for NVIDIA vGPU software requires the Enterprise Plus Edition of VMware vSphere Hypervisor (ESXi). For details, see Compare VMware vSphere Editions (PDF). Updates to a base release of VMware vSphere Hypervisor (ESXi) are compatible with the base release and can also be used with this version of NVIDIA vGPU software unless expressly stated otherwise.
### Validated Platforms

#### VMware vSphere Hypervisor (ESXi) 6.5
- **Release Supported**: 6.5 and compatible updates
- **Notes**: Requires VMware vSphere Hypervisor (ESXi) 6.5 patch EP3 (ESXi650-201811002, build 10884925) or later from VMware. All NVIDIA GPUs that support NVIDIA vGPU software are supported. Suspend-resume with vGPU and vMotion with vGPU are not supported.

#### VMware vSphere Hypervisor (ESXi) 6.0
- **Release Supported**: 6.0 and compatible updates
- **Notes**: Pending release of a patch from VMware for VMware vSphere Hypervisor (ESXi) 6.0. GPUs based on the Pascal and Volta architectures in pass-through mode require 6.0 Update 3 or later. For vGPU, all NVIDIA GPUs that support NVIDIA vGPU software are supported. Suspend-resume with vGPU and vMotion with vGPU are not supported.

### Supported Management Software and Virtual Desktop Software Releases

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

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

<table>
<thead>
<tr>
<th>Software</th>
<th>Releases Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMware Horizon</td>
<td>7.8 and compatible 7.8.x updates</td>
</tr>
<tr>
<td></td>
<td>7.7 and compatible 7.7.x updates</td>
</tr>
<tr>
<td></td>
<td>7.6 and compatible 7.6.x updates</td>
</tr>
<tr>
<td></td>
<td>7.5 and compatible 7.5.x updates</td>
</tr>
<tr>
<td></td>
<td>7.4 and compatible 7.4.x updates</td>
</tr>
<tr>
<td></td>
<td>7.3 and compatible 7.3.x updates</td>
</tr>
<tr>
<td></td>
<td>7.2 and compatible 7.2.x updates</td>
</tr>
<tr>
<td></td>
<td>7.1 and compatible 7.1.x updates</td>
</tr>
<tr>
<td></td>
<td>7.0 and compatible 7.0.x updates</td>
</tr>
</tbody>
</table>
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.

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

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

2.3.1. Windows Guest OS Support

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

<table>
<thead>
<tr>
<th>Guest OS</th>
<th>NVIDIA vGPU - VMware vSphere Releases</th>
<th>Pass-Through GPU - VMware vSphere Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows Server 2019</td>
<td>6.7, 6.5 update 2, 6.5 update 1</td>
<td>6.7, 6.5 update 2, 6.5 update 1</td>
</tr>
<tr>
<td>Windows Server 2016 1709, 1607</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
<tr>
<td>Windows Server 2012 R2</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
<tr>
<td>Windows Server 2008 R2</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
<tr>
<td>Windows 10:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>‣ October 2018 Update (1809)</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
<tr>
<td>‣ Spring Creators Update (1803)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.3.2. Linux Guest OS Support

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

<table>
<thead>
<tr>
<th>Guest OS</th>
<th>NVIDIA vGPU - VMware vSphere Releases</th>
<th>Pass-Through GPU - VMware vSphere Releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red Hat Enterprise Linux 7.0-7.6 and later compatible 7.x versions</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
<tr>
<td>CentOS 7.0-7.6 and later compatible 7.x versions</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
<tr>
<td>Red Hat Enterprise Linux 6.6 and later compatible 6.x versions</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
<tr>
<td>CentOS 6.6 and later compatible 6.x versions</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
<tr>
<td>Ubuntu 18.04 LTS</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
<tr>
<td>Ubuntu 16.04 LTS</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
<tr>
<td>Ubuntu 14.04 LTS</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
<tr>
<td>SUSE Linux Enterprise Server 12 SP3</td>
<td>6.7, 6.5, 6.0</td>
<td>6.7, 6.5, 6.0</td>
</tr>
</tbody>
</table>

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

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

2.4. NVIDIA CUDA Toolkit Version Support

The releases in this release family of NVIDIA vGPU software support NVIDIA CUDA Toolkit 10.1.
For more information about NVIDIA CUDA Toolkit, see CUDA Toolkit 10.1 Documentation.

### 2.5. vGPU Migration Support

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

**Supported GPUs:**
- Tesla M6
- Tesla M10
- Tesla M60
- Tesla P4
- Tesla P6
- Tesla P40
- Tesla V100 SXM2
- Tesla V100 SXM2 32GB
- Tesla V100 PCIe
- Tesla V100 PCIe 32GB
- Tesla V100 FHHL
- Tesla T4
- Quadro RTX 6000
- Quadro RTX 8000

**Supported VMware vSphere Hypervisor (ESXi) releases:**
- Release 6.7 U1 and compatible updates support vMotion with vGPU and suspend-resume with vGPU.
- Release 6.7 supports only suspend-resume with vGPU.
- Releases earlier than 6.7 do not support any form of vGPU migration.

Supported guest OS releases: Windows and Linux.
Chapter 3. 
KNOWN PRODUCT LIMITATIONS

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

3.1. 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.
3.2. 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.

3.3. VM failures or crashes on servers with 1 TB or more of system memory

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

This limitation applies only to systems with supported GPUs based on the Maxwell architecture: Tesla M6, Tesla M10, and Tesla M60.

Resolution
1. Limit the amount of system memory on the server to 1 TB minus 16 GB by setting `memmapMaxRAMMB` to 1032192, which is equal to 1048576 minus 16384.
2. Reboot the server. 

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

3.4. VM running older NVIDIA vGPU drivers fails to initialize vGPU when booted

**Description**

A VM running a version of the NVIDIA guest VM drivers from a previous main release branch, for example release 4.4, will fail to initialize vGPU when booted on a VMware vSphere platform running the current release of Virtual GPU Manager.

In this scenario, 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)*

Depending on the versions of drivers in use, the VMware vSphere VM's log file reports one of the following errors:

- A version mismatch between guest and host drivers:
  
  ```
  vthread-10| E105: vmiop_log: Guest VGX version(2.0) and Host VGX version(2.1) do not match
  ```

- A signature mismatch:
  
  ```
  ```

**Resolution**

Install the current NVIDIA guest VM driver in the VM.

3.5. Virtual GPU fails to start if ECC is enabled

**Description**

Tesla M60, Tesla M6, and GPUs based on the Pascal GPU architecture, for example Tesla P100 or Tesla P4, support error correcting code (ECC) memory for improved data integrity. Tesla M60 and M6 GPUs in graphics mode are supplied with ECC memory disabled by default, but it may subsequently be enabled using `nvidia-smi`. GPUs based on the Pascal GPU architecture are supplied with ECC memory enabled.

However, NVIDIA vGPU does not support ECC memory. If ECC memory is enabled, NVIDIA vGPU fails to start.
The following error is logged in the VMware vSphere host's log file:

```
vthread10|E105: Initialization: VGX not supported with ECC Enabled.
```

**Resolution**

Ensure that ECC is disabled on all GPUs.

Before you begin, ensure that NVIDIA Virtual GPU Manager is installed on your hypervisor.

1. Use `nvidia-smi` to list the status of all GPUs, and check for ECC noted as enabled on GPUs.

   ```
   # nvidia-smi -q
   ================NVSMI LOG===============
   Timestamp : Tue Dec 19 18:36:45 2017
   Driver Version : 384.99
   Attached GPUs : 1
   GPU 0000:02:00.0
   
   Ecc Mode
   Current : Enabled
   Pending : Enabled
   [...]
   
   [...]
   ```

2. Change the ECC status to off on each GPU for which ECC is enabled.

   - If you want to change the ECC status to off for all GPUs on your host machine, run this command:
     ```
     # nvidia-smi -e 0
     ```
   - If you want to change the ECC status to off for a specific GPU, run this command:
     ```
     # nvidia-smi -i id -e 0
     ```
     `id` is the index of the GPU as reported by `nvidia-smi`.
     This example disables ECC for the GPU with index 0000:02:00.0.
     ```
     # nvidia-smi -i 0000:02:00.0 -e 0
     ```

3. Reboot the host.
4. Confirm that ECC is now disabled for the GPU.

   ```
   # nvidia-smi -q
   ================NVSMI LOG===============
   Timestamp : Tue Dec 19 18:37:53 2017
   Driver Version : 384.99
   Attached GPUs : 1
   ```
If you later need to enable ECC on your GPUs, run one of the following commands:

- If you want to change the ECC status to on for all GPUs on your host machine, run this command:

  ```bash
  # nvidia-smi -e 1
  ```

- If you want to change the ECC status to on for a specific GPU, run this command:

  ```bash
  # nvidia-smi -i id -e 1
  ```

  *id* is the index of the GPU as reported by *nvidia-smi*.

  This example enables ECC for the GPU with index `0000:02:00.0`.

  ```bash
  # nvidia-smi -i 0000:02:00.0 -e 1
  ```

After changing the ECC status to on, reboot the host.

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

**Description**

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

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

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

**Resolution**

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

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

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

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

Description

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

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

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

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

Resolution

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

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

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

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

The reservation can be reverted back to its default setting by setting `pciPassthru0.cfg.enable_large_sys_mem` to 0, or by removing the parameter from the advanced settings.
Chapter 4.
RESOLVED ISSUES

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

Issues Resolved in Release 8.0

<table>
<thead>
<tr>
<th>Bug ID</th>
<th>Summary and Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1971698</td>
<td>NVIDIA vGPU encoder and process utilization counters don’t work with Windows Performance Counters. GPU encoder and process utilization counter groups are listed in Windows Performance Counters, but no instances of the counters are available. The counters are disabled by default and must be enabled.</td>
</tr>
</tbody>
</table>
5.1. Restricting Access to GPU Performance Counters

The NVIDIA graphics driver contains a vulnerability (CVE-2018-6260) that may allow access to application data processed on the GPU through a side channel exposed by the GPU performance counters. To address this vulnerability, update the driver and restrict access to GPU performance counters to allow access only by administrator users and users who need to use CUDA profiling tools.

The GPU performance counters that are affected by this vulnerability are the hardware performance monitors used by the CUDA profiling tools such as CUPTI, Nsight Graphics, and Nsight Compute. These performance counters are exposed on the hypervisor host and in guest VMs only as follows:

- On the hypervisor host, they are always exposed. However, the Virtual GPU Manager does not access these performance counters and, therefore, is not affected.
- In Windows and Linux guest VMs, they are exposed only in VMs configured for GPU pass through. They are not exposed in VMs configured for NVIDIA vGPU.

5.1.1. Windows: Restricting Access to GPU Performance Counters for One User by Using NVIDIA Control Panel

Perform this task from the guest VM to which the GPU is passed through.

Ensure that you are running NVIDIA Control Panel version 8.1.950.

1. Open NVIDIA Control Panel:
   - Right-click on the Windows desktop and select NVIDIA Control Panel from the menu.
   - Open Windows Control Panel and double-click the NVIDIA Control Panel icon.
2. In NVIDIA Control Panel, select the Manage GPU Performance Counters task in the Developer section of the navigation pane.

3. Complete the task by following the instructions in the Manage GPU Performance Counters > Developer topic in the NVIDIA Control Panel help.

5.1.2. Windows: Restricting Access to GPU Performance Counters Across an Enterprise by Using a Registry Key

You can use a registry key to restrict access to GPU Performance Counters for all users who log in to a Windows guest VM. By incorporating the registry key information into a script, you can automate the setting of this registry for all Windows guest VMs across your enterprise.

Perform this task from the guest VM to which the GPU is passed through.

Caution Only enterprise administrators should perform this task. Changes to the Windows registry must be made with care and system instability can result if registry keys are incorrectly set.

1. Set the RmProfilingAdminOnly Windows registry key to 1.
   
   `[HKLM\SYSTEM\CurrentControlSet\Services\nvlddmkm\Global\NVTweak]`  
   Value: "RmProfilingAdminOnly"  
   Type: DWORD  
   Data: 00000001

   The data value 1 restricts access, and the data value 0 allows access, to application data processed on the GPU through a side channel exposed by the GPU performance counters.

2. Restart the VM.

5.1.3. Linux Guest VMs: Restricting Access to GPU Performance Counters

On systems where unprivileged users don't need to use GPU performance counters, restrict access to these counters to system administrators, namely users with the CAP_SYS_ADMIN capability set. By default, the GPU performance counters are not restricted to users with the CAP_SYS_ADMIN capability.

Perform this task from the guest VM to which the GPU is passed through.

This task requires sudo privileges.

1. Log in to the guest VM.

2. Set the kernel module parameter NVreg_RestrictProfilingToAdminUsers to 1 by adding this parameter to the `/etc/modprobe.d/nvidia.conf` file.

   ▶ If you are setting only this parameter, add an entry for it to the `/etc/modprobe.d/nvidia.conf` file as follows:

   ```
   options nvidia  
   NVreg_RegistryDwords="NVreg_RestrictProfilingToAdminUsers=1"
   ```
If you are setting multiple parameters, set them in a single entry as in the following example:

```bash
options nvidia NVreg_RegistryDwords="RmPVMRL=0x0"  
"NVreg_RestrictProfilingToAdminUsers=1"
```

If the `/etc/modprobe.d/nvidia.conf` file does not already exist, create it.

3. Restart the VM.

### 5.1.4. Hypervisor Host: Restricting Access to GPU Performance Counters

On systems where unprivileged users don’t need to use GPU performance counters, restrict access to these counters to system administrators. By default, the GPU performance counters are not restricted to system administrators.

Perform this task from your hypervisor host machine.

1. Open a command shell as the root user on your hypervisor host machine.
2. Set the kernel module parameter `NVreg_RestrictProfilingToAdminUsers` to 1 by using the `esxcli` set command.
   - If you are setting only this parameter, set it as follows:
     ```bash
     # esxcli system module parameters set -m nvidia -p "NVreg_RestrictProfilingToAdminUsers=1"
     ```
   - If you are setting multiple parameters, set them in a single command as in the following example:
     ```bash
     # esxcli system module parameters set -m nvidia -p "NVreg_RegistryDwords=RmPVMRL=0x0 NVreg_RestrictProfilingToAdminUsers=1"
     ```
3. Reboot your hypervisor host machine.
Chapter 6.
KNOWN ISSUES

6.1. Incorrect NVIDIA vGPU software Windows graphics driver version in the installer

Description
The NVIDIA vGPU software Windows graphics driver version in the installer is incorrect. The driver version incorrectly appears as 325.31 instead of 425.31 in the Extraction path field and the title of the NVIDIA Graphics Driver (325.31) Package Window.

Version
NVIDIA vGPU software Windows graphics driver version 425.31 in NVIDIA vGPU software release 8.0

Workaround
To simplify future administration of your system, you can correct the driver version in the folder name when the installer prompts you for the extraction path. However, if you do not change the name of the folder in the extraction path, the installation succeeds and the driver functions correctly.

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

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

**Version**
VMware vSphere ESXi 6.5

**Status**
Open

**Ref. #**
200491080

6.3. Vulkan applications crash in Windows 7 guest VMs configured with NVIDIA vGPU

**Description**
In Windows 7 guest VMs configured with NVIDIA vGPU, applications developed with Vulkan APIs crash or throw errors when they are launched. Vulkan APIs require sparse texture support, but in Windows 7 guest VMs configured with NVIDIA vGPU, sparse textures are not enabled.

In Windows 10 guest VMs configured with NVIDIA vGPU, sparse textures are enabled and applications developed with Vulkan APIs run correctly in these VMs.

**Status**
Open

**Ref. #**
200381348
6.4. 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

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

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

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

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

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

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

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

Status
Open

Ref. #
200457177

6.7. 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.
6.8. 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 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.

6.9. VMware vMotion fails gracefully under heavy load

**Description**

Migrating a VM configured with vGPU fails gracefully if the VM is running an intensive workload.
The error stack in the task details on the vSphere web client contains the following error message:

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

**Workaround**

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

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

**Status**

Not an NVIDIA bug

**Ref. #**

200416700

### 6.10. Cloned VMs configured with a vGPU type different than the type in the master image fail to start

**Description**

Cloned VMs configured with a vGPU type different than the type in the master image fail to start.

When a Windows 10 VM is booted, the VM becomes stuck in a loop and alternately displays **Getting devices ready: 50%** and **Preparation in progress**.

**Workaround**

Create one master image for each vGPU type that you want to use. Do not attempt to configure a cloned VM with a vGPU type different than the type in the master image.

**Status**

Open
6.11. View session freezes intermittently after a Linux VM acquires a license

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

**Workaround**
Resize the view session.

**Status**
Open

Ref. #
2285306

6.12. Even when the scheduling policy is equal share, unequal GPU utilization is reported

**Description**
When the scheduling policy is equal share, unequal GPU engine utilization can be reported for the vGPUs on the same physical GPU.

For example, GPU engine usage for three P40-8Q vGPUs on a Tesla P40 GPU might be reported as follows:

```
[root@localhost:-] nvidia-smi vgpu
Wed Jun 27 10:33:18 2018
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 390.59                Driver Version: 390.59                    |
|-------------------------------+--------------------------------+------------+
| GPU  Name                     | Bus-Id                         | GPU-Util   |
|      vGPU ID    Name          | VM ID    VM Name               | vGPU-Util  |
|===============================+================================+============|
|   0  Tesla P40                | 00000000:81:00.0               |  52%       |
|      2122661    GRID P40-8Q   | 2122682  centos7.4-xmpl-211... |     19%    |
|      2122663    GRID P40-8Q   | 2122692  centos7.4-xmpl-211... |      0%    |
|      2122659    GRID P40-8Q   | 2122664  centos7.4-xmpl-211... |     25%    |
+-------------------------------+--------------------------------+------------+
|   1  Tesla P40                | 00000000:85:00.0               |  58%       |
|      2122662    GRID P40-8Q   | 2122689  centos7.4-xmpl-211... |      0%    |
|      2122658    GRID P40-8Q   | 2122667  centos7.4-xmpl-211... |     59%    |
```
The vGPU utilization of the vGPU 2122658 is reported as 59%. However, the expected vGPU utilization should not exceed 33%.

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

Status
Open

Ref. #
2175888

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

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

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

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

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

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

Status
Open

Ref. #
2227591

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

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

Status
Closed

Ref. #
200407230

6.15. GPU resources not available error during VMware instant clone provisioning

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

This error occurs when options for VMware Virtual Shared Graphics Acceleration (vSGA) are set for a VM that is configured with NVIDIA vGPU. VMware vSGA is a feature of VMware vSphere that enables multiple virtual machines to share the physical GPUs on ESXi hosts and can be used as an alternative to NVIDIA vGPU.

Depending on the combination of options set, one of the following error messages is seen when the VM is powered on:
Module ‘MKS’ power on failed.

This message is seen when the following options are set:

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

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

This message is seen when the following options are set:

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

Resolution

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

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

Status

Not a bug

Ref. #

2369683

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

Description

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

The following GPUs require 64 GB of MMIO space:
Known Issues

- Tesla P6
- Tesla P40

Version

This issue affects the following versions of VMware vSphere ESXi:

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

Workaround

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

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

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

Status

Not an NVIDIA bug

Resolved in VMware vSphere ESXi 6.7

Ref. #

2043171

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

Description

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

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

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

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

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

**Status**

Not an NVIDIA bug

**Ref. #**

200366884

---

**6.18. Resolution is not updated after a VM acquires a license and is restarted**

**Description**

In a Red Enterprise Linux 7.3 guest VM, an increase in resolution from 1024×768 to 2560×1600 is not applied after a license is acquired and the `gridd` service is restarted. This issue occurs if the multimonitor parameter is added to the `xorg.conf` file.

**Version**

Red Enterprise Linux 7.3

**Status**

Open

**Ref. #**

200275925

---

**6.19. On Tesla P40, P6, and P4 GPUs, the default ECC setting prevents NVIDIA vGPU from starting**

**Description**

On Tesla P40, Tesla P6, and Tesla P4 GPUs, the default error-correcting code (ECC) memory setting prevents NVIDIA vGPU from starting. By default, ECC memory is enabled on these GPUs, but NVIDIA vGPU does not support ECC memory.
Known Issues

**Workaround**

Before running NVIDIA vGPU, disable ECC memory as explained in [Virtual GPU Software User Guide](https://www.nvidia.com).

**Status**

Closed.

**Ref. #**

200269717

### 6.20. Tesla P40 cannot be used in pass-through mode

**Description**

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

**Workaround**

Ensure that your GPUs are configured as described in [Requirements for Using GPUs Based on the Pascal and Volta Architectures in Pass-Through Mode](https://www.nvidia.com).

**Status**

Not a bug

**Ref. #**

1944539

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

**Description**

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

This issue does not affect 2D applications.
6.22. 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
6.23. No Manage License option available in NVIDIA X Server Settings by default

Description

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

Workaround

This workaround requires sudo privileges.

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.

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

6.24. 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

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

Description

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

This issue typically occurs in the following situations:

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

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

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

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

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

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

**Workaround**

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

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

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

**Status**

Open

**Ref. #**

- 200130864
- 1803861
6.26. vGPU VM fails to boot in ESXi 6.5 if the graphics type is Shared

Description

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

On VMware vSphere Hypervisor (ESXi) 6.5, after vGPU is configured, VMs to which a vGPU is assigned may fail to start and the following error message may be displayed:

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

The vGPU Manager VIB provides vSGA and vGPU functionality in a single VIB. After this VIB is installed, the default graphics type is Shared, which provides vSGA functionality. To enable vGPU support for VMs in VMware vSphere 6.5, you must change the default graphics type to Shared Direct. If you do not change the default graphics type you will encounter this issue.

Version

VMware vSphere Hypervisor (ESXi) 6.5

Workaround

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

Status

Open

Ref. #

200256224
6.27. ESXi 6.5 web client shows high memory usage even when VMs are idle

**Description**

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

**Version**

VMware vSphere Hypervisor (ESXi) 6.5

**Workaround**

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

**Status**

Not an NVIDIA bug

**Ref. #**

200191065

6.28. VMs configured with NVIDIA vGPU must not be on a host in a VMware DRS cluster

**Description**

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

```
NVIDIA Installer cannot continue
This graphics driver could not find compatible graphics hardware.
```
Furthermore, you cannot overcome this limitation by configuring a VM with NVIDIA vGPU and installing the driver on the VM on a host outside a DRS cluster and moving the host into the DRS cluster after configuring it.

**Workaround**

Move each VM configured with NVIDIA vGPU to a host outside the DRS cluster.

1. Remove NVIDIA Virtual GPU Manager from the host in the DRS cluster.
2. Create a cluster of VMware ESXi hosts outside the DRS domain.
3. Install the NVIDIA Virtual GPU Manager on an ESXi host in the cluster that you created in the previous step.
4. Create a vSphere VM for use with NVIDIA vGPU.
5. Configure the vSphere VM with NVIDIA vGPU.
6. Boot the vSphere VM and install the NVIDIA driver for NVIDIA vGPU.

For instructions for performing these tasks, refer to *Virtual GPU Software User Guide*.

**Status**

Open

**Ref. #**

1933449

### 6.29. 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
Known Issues

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

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

Description

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

Fix

Make NVIDIA vGPU the primary display adapter.

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

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

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

Status

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

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

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

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

Status
Open

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

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

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

Workaround
Do not assign vGPU and passthrough GPUs to a VM simultaneously.
6.33. vGPU allocation policy fails when multiple VMs are started simultaneously

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

Workaround
Start VMs individually.

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

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

Workaround
Do not use Sleep with vGPU.
Installing the VMware Horizon agent will disable the Sleep option.
**Status**
Closed

**Ref. #**
200043405

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

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

- One or more of the VMs may fail to start with the following error:
  ```
  The available Memory resources in the parent resource pool are insufficient for the operation
  ```

- When run in the host shell, the `nvidia-smi` utility returns this error:
  ```
  -sh: can't fork
  ```

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

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

**Status**
Closed

**Ref. #**
200060499
6.36. On reset or restart VMs fail to start with the error VMIOP: no graphics device is available for vGPU...

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

Fix
Upgrade to ESXi 6.0 Update 1.

Status
Closed

Ref. #
200097546

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

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

Workaround
None

Status
Open
Partially resolved for Horizon 7.0.1:
- For Blast connections, GPU utilization is no longer high.
- For PCoIP connections, utilization remains high.
Ref. #
1735009
Notice

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.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication of otherwise under any patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all other information previously supplied. NVIDIA Corporation products are not authorized as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

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, vGPU, Pascal, 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

© 2013-2019 NVIDIA Corporation. All rights reserved.