# GRID VGPU FOR VMWARE VSPHERE VERSION 367.43/369.17

**Release Notes**

# TABLE OF CONTENTS

# Chapter 1.
# RELEASE NOTES

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

This release includes the following software:

► NVIDIA GRID Virtual GPU Manager version 367.43 for VMware vSphere 6.0 Hypervisor (ESXi)
► NVIDIA Windows drivers for vGPU version 369.17
► NVIDIA Linux drivers for vGPU version 367.43

> 💬 **Caution**
>
> The releases of GRID vGPU Manager and Windows guest VM drivers that you install must be compatible. Older VM drivers will not function correctly with this release of GRID vGPU Manager. However, some older GRID vGPU Managers may function correctly with this release of Windows guest drivers. See VM running older NVIDIA vGPU drivers fails to initialize vGPU when booted.

Updates in this release:

► Support for Tesla M10 cards
► Support for Windows Server 2016 as a guest OS
► Enhanced monitoring of GPU performance:

  ► Monitoring of individual vGPUs
  ► Monitoring from within a guest VM
► Miscellaneous bug fixes

# Chapter 2.
# VALIDATED PLATFORMS

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

- GRID K1
- GRID K2
- Tesla M6
- Tesla M10
- Tesla M60

For a list of validated server platforms, refer to Buy NVIDIA GRID Solutions.

## 2.1. Hypervisor Software Versions

This release has been tested with the following hypervisor software versions:

| Software | Version Tested |
|---|---|
| VMware vSphere Hypervisor (ESXi) | 6.0 RTM build 2494585<br>6.0 update 1<br>6.0 update 2 |
| VMware Horizon | 6.2.1 RTM build 3268071<br>7.0 RTM build 3618085 |
| VMware vCenter Server | 6.0 RTM build 2562643 |

## 2.2. Linux Support

GRID vGPU with Linux guest VMs is supported on Tesla M60, Tesla M10, and Tesla M6, with the following distributions:

- Red Hat Enterprise Linux 6.6
- Red Hat Enterprise Linux 7
- CentOS 6.6

- ▸ CentOS 7
- ▸ Ubuntu 12.04 LTS
- ▸ Ubuntu 14.04 LTS

## 2.3. Hardware Configuration

Tesla M60 and M6 GPUs support compute and graphics modes, which can be configured by using the `gpumodeswitch` tool provided with GRID software releases. GRID vGPU requires that M60 and M6 GPUs are configured in graphics mode.

# Chapter 3.
# KNOWN PRODUCT LIMITATIONS

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

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

**Description**

A VM running older NVIDIA drivers, such as those from a previous vGPU release, will fail to initialize vGPU when booted on a VMware vSphere platform running the current release of GRID Virtual GPU Manager.

In this scenario, the VM boots in standard VGA mode with reduced resolution and color depth. The NVIDIA GRID 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:

```
vthread-10| E105: vmiop_log: VGPU message signature mismatch.
```

**Resolution**

Install the latest NVIDIA vGPU release drivers in the VM.

## 3.2. Virtual GPU fails to start if ECC is enabled

**Description**

GRID K2, Tesla M60, and Tesla M6 support error correcting code (ECC) for improved data integrity. If ECC is enabled, virtual GPU fails to start. The following error is logged in the VMware vSphere VM's log file:

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

Virtual GPU is not currently supported with ECC active. GRID K2 cards and Tesla M60, M6 cards in graphics mode ship with ECC disabled by default, but ECC may subsequently be enabled using `nvidia-smi`.

**Resolution**

Ensure that ECC is disabled on all GPUs.

1. Use `nvidia-smi` to list the status of all GPUs, and check for ECC noted as enabled on GPUs.
2. Change the ECC status to off on each GPU for which ECC is enabled by executing the following command:

   ```
   nvidia-smi -i id -e 0
   ```

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

## 3.3. Single vGPU benchmark scores are lower than passthrough GPU

**Description**

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

Aside from performance differences that may be attributed to a vGPU's smaller framebuffer size, vGPU incorporates a performance balancing feature known as Frame Rate Limiter (FRL), which is enabled on all vGPUs. 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 passthrough GPU.
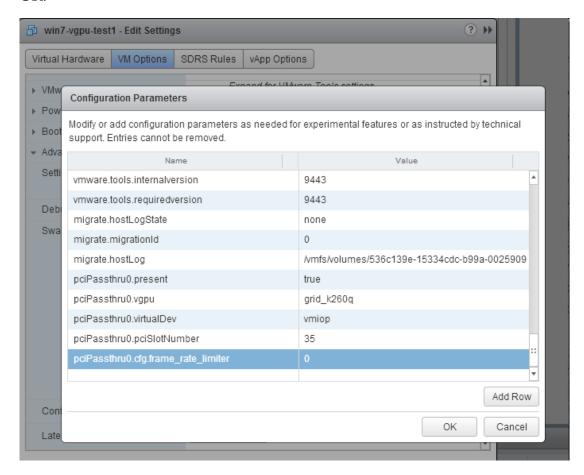
**Resolution**

FRL is controlled by an internal vGPU setting. 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.

# 3.4. GRID K1 and GRID K2 cards do not support monitoring of vGPU engine usage

**Description**

GRID K1 and GRID K2 cards do not support monitoring of vGPU engine usage. All tools and APIs for any vGPU running on GRID K1 or GRID K2 cards report 0 for the following usage statistics:

▸ 3D/Compute
▸ Memory controller bandwidth
▸ Video encoder
▸ Video decoder

# 3.5. 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 GRID GPU is present in **Windows Device Manager** but displays a warning sign, and the following device status:

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

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

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

**Resolution**

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

> 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

| Bug ID | Summary and Description |
|--------|-------------------------|
| 1756897 | **BSOD with 361.40/362.13 drivers on GRID K1 cards with Windows 10**<br><br>If Windows 10 is the guest OS on a server with a GRID K1 card running the 361.40/362.13 drivers, the OS crashes. If the 352.83/354.80 drivers are used, the system drops the connection, but the guest OS continues to function. |

# Chapter 5.
# KNOWN ISSUES

## 5.1. 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 when multiple display heads are used with Citrix XenDesktop or VMware Horizon on a Windows 10 guest VM.

When this error 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
▸ GRID K100, K120Q
▸ GRID K200, K220Q

**Version**

**Workaround**

**Status**

Open

**Ref. #**

200130864

# 5.2. NVIDIA Control Panel is killed during reconnection with a View

**Description**

If NVIDIA Control Panel is running while a View session is disconnected and then reconnected, NVIDIA Control Panel is killed before the View session is reconnected.

**Version**

**Fix**

**Status**

Open

**Ref. #**

200176969

# 5.3. 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!
```

**Version**

**Workaround**

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

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

    ```
    ~]# reboot
    ```

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

**Status**

Not an NVIDIA bug

**Ref. #**

200167868

# 5.4. NVIDIA Control Panel complains 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 complains about not using a display attached to an NVIDIA GPU. This happens because Windows is using VMware's SVGA device instead of NVIDIA vGPU.

**Version**

**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

**Ref. #**

# 5.5. 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)
```

**Version**

**Workaround**

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

**Status**

Open

**Ref. #**

# 5.6. 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 follwoing device status

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

**Version**

**Workaround**

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

**Status**

Open

**Ref. #**

1735002

# 5.7. 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.

**Version**

**Workaround**

Start VMs individually.

**Status**

Not an NVIDIA bug

**Ref. #**

200042690

## 5.8. Prior to installing Horizon agent 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.

**Version**

**Workaround**

Do not use Sleep with vGPU.

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

**Status**

Closed

**Ref. #**

200043405

## 5.9. 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.

**Version**

**Workaround**

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

**Status**

Closed

**Ref. #**

200060499

# 5.10. 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, i.e. maximum number of vGPU VMs the server can support, some VMs might fail to start post a reset or restart operation.

**Version**

**Fix**

Upgrade to ESXi 6.0 Update 1.

**Status**

Closed

**Ref. #**

200097546

## 5.11. `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.

**Version**

**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

## 5.12. Multiple WebGL tabs in Microsoft Internet Explorer may trigger TDR on Windows VMs

**Description**

Running intensive WebGL applications in multiple IE tabs may trigger a TDR on Windows VMs.

**Version**

**Workaround**

Disable hardware acceleration in IE.

To enable software rendering in IE, refer to the Microsoft knowledge base article How to enable or disable software rendering in Internet Explorer.

**Status**

Open

**Ref. #**

200148377