NVIDIA AI Enterprise

Release Notes

RN-10616-001 _v2.0 | June 2022

Table of Contents

Chapter 1. What's New in NVIDIA AI Enterprise 1

Chapter 2. Supported Hardware and Software 2

2.1. NVIDIA AI Enterprise Software Components 3

2.2. Switching the Mode of a GPU that Supports Multiple Display Modes 3

2.3. Requirements for Using C-Series vCS vGPUs 4

2.4. Requirements for Using vGPU on GPUs Requiring 64 GB or More of MMIO Space with Large-Memory VMs 5

2.5. Linux Only: Error Messages for Misconfigured GPUs Requiring Large MMIO Space 5

2.6. NVIDIA CUDA Toolkit Version Support 6

2.7. vGPU Migration Support 6

2.8. Multiple vGPU Support 6

2.9. Peer-to-Peer CUDA Transfers over NVLink Support 8

2.10. Unified Memory Support 10

2.11. NVIDIA GPU Operator Support 11

Chapter 3. NVIDIA AI Enterprise Supported Cloud Services 12

3.1. Amazon Web Services Elastic Compute Cloud (AWS EC2) 12

3.2. Google Cloud Platform (GCP) 13

3.3. Microsoft Azure 13

3.4. NVIDIA GPU Optimized VMI on CSP Marketplace 14

Chapter 4. CPU Only Server Support 15

Chapter 5. Known Product Limitations 16

5.1. Issues occur when the channels allocated to a vGPU are exhausted 16

5.2. Total frame buffer for vGPUs is less than the total frame buffer on the physical GPU 17

5.3. Single vGPU benchmark scores are lower than pass-through GPU 18

5.4. VMs configured with large memory fail to initialize vGPU when booted 20

Chapter 6. Known Issues 22

6.1. Migration of VMs configured with vGPU stops before the migration is complete 22

Chapter 1. What's New in NVIDIA AI Enterprise

Features in this release of NVIDIA AI Enterprise are as follows:

- Support for Red Hat OpenShift 4.9 and later.
- Added support for CPU based Deep Learning Frameworks and Tools.
- Added TAO Toolkit.
- Enhanced Triton Inference Server to support FIL backend.
- Multi-cloud NVIDIA GPU optimized Virtual Machine Instances.
- Added support for NVIDIA A100X, NVIDIA A30X, NVIDIA A2, and NVIDIA RTX A5500 GPUs.

Chapter 2. Supported Hardware and Software

NVIDIA GPUs:

- NVIDIA A100X
- NVIDIA A100 PCIe 40GB
- NVIDIA A100 HGX 40GB
- NVIDIA A100 PCIe 80GB
- NVIDIA A100 HGX 80GB
- NVIDIA A40
- NVIDIA A30X
- NVIDIA A30
- NVIDIA A10
- NVIDIA A16
- NVIDIA A2
- NVIDIA RTX A6000
- NVIDIA RTX A5500
- NVIDIA RTX A5000
- NVIDIA T4
- NVIDIA V100

For NVIDIA AI Enterprise Compatible servers, refer to the NVIDIA-certified systems page.

Multi-node scaling requires an ethernet NIC that supports RoCE. For best performance, NVIDIA recommends using an NVIDIA® Mellanox® ConnectX®-6 Dx and an NVIDIA A100 GPU in each VM used for multi-node scaling. Refer to the Sizing guide and the Multi-Node Training solution guide for further information.

Hypervisor software:

- VMware vSphere Hypervisor (ESXi) Enterprise Plus Edition 7.0 Update 3
- VMware vCenter Server 7.0 Update 3

- VMware vSphere 6.7 (only for NVIDIA T4 and NVIDIA V100 GPUs)

Guest operating systems:

- Ubuntu 20.04 LTS
- Red Hat Enterprise Linux 8.4
- Red Hat OpenShift 4.9 and later

For more information, see the NVIDIA AI Enterprise Product Support Matrix.

2.1. NVIDIA AI Enterprise Software Components

Software Component	NVIDIA Release
NVIDIA vGPU Software	14.0
NVIDIA GPU Operator	1.10.0
NVIDIA Network Operator	1.1.0
TensorFlow 2	22.02-tf2-nvaie-2.0-py3
TensorFlow 1	22.02-tf1-nvaie-2.0-py3
PyTorch	22.02-nvaie-2.0-py3
NVIDIA Triton Inference Server	22.02-nvaie-2.0-py3 and 22.02-nvaie-2.0-py3-sdk
NVIDIA TensorRT	22.02-nvaie-2.0-py3
NVIDIA RAPIDS	22.02-cuda11.4-ubuntu20.04-py3.8
TAO Toolkit for Language Model (Conv AI)	3.21.08-py3
TAO Toolkit for Conv AI	3.22.02-py3
TAO Toolkit for CV	3.21.11-tf1.15.4-py3-nvaie

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

Some GPUs support displayless and display-enabled modes but must be used in NVIDIA AI Enterprise deployments in displayless mode.

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

GPU	Mode as Supplied from the Factory
NVIDIA A40	Displayless
NVIDIA RTX A5000	Display enabled
NVIDIA RTX A5500	Display enabled

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

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

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

Note:

Only the following GPUs support the `displaymodeselector` tool:

- NVIDIA A40
- NVIDIA RTX A5000
- NVIDIA RTX A5500
- NVIDIA RTX A6000

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

2.3. Requirements for Using C-Series vCS vGPUs

Because C-Series vCS vGPUs have large BAR memory settings, using these vGPUs 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 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).
- Because the VM's MMIO space must be increased to 64 GB, vCS requires ESXi 6.0 Update 3 and later, or ESXi 6.5 and later.

2.4. Requirements for Using vGPU on GPUs Requiring 64 GB or More of MMIO Space with Large-Memory VMs

Some GPUs require 64 GB or more of MMIO space. When a vGPU on a GPU that requires 64 GB or more of MMIO space is assigned to a VM with 32 GB or more of memory on ESXi , the VM's MMIO space must be increased to the amount of MMIO space that the GPU requires.

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

No extra configuration is needed.

The following table lists the GPUs that require 64 GB or more of MMIO space and the amount of MMIO space that each GPU requires.

GPU	MMIO Space Required
NVIDIA A10	64 GB
NVIDIA A30	64 GB
NVIDIA A40	128 GB
NVIDIA A100 40GB (all variants)	128 GB
NVIDIA A100 80GB (all variants)	256 GB
NVIDIA RTX A5000	64 GB
NVIDIA RTX A5500	64 GB
NVIDIA RTX A6000	128 GB
Tesla P100 (all variants)	64 GB

2.5. Linux Only: Error Messages for Misconfigured GPUs Requiring Large MMIO Space

In a Linux VM, if the requirements for using C-Series vCS vGPUs or GPUs requiring large MMIO space in pass-through mode are not met, the following error messages are written to the VM's `dmesg` log during installation of the NVIDIA AI Enterprise graphics driver:

```
NVRM: BAR1 is 0M @ 0x0 (PCI:0000:02:02.0)
[   90.823015] NVRM: The system BIOS may have misconfigured your GPU.
[   90.823019] nvidia: probe of 0000:02:02.0 failed with error -1
[   90.823031] NVRM: The NVIDIA probe routine failed for 1 device(s).
```

2.6. NVIDIA CUDA Toolkit Version Support

The releases in this release family of NVIDIA AI Enterprise support NVIDIA CUDA Toolkit 11.4.

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

2.7. vGPU Migration Support

vGPU migration, which includes vMotion and suspend-resume, is supported for both time-sliced and MIG-backed vGPUs on all supported GPUs, hypervisor software releases, and guest operating systems.

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

- Unified memory
- Debuggers
- Profilers

Known Issues with vGPU Migration Support

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

2.8. Multiple vGPU Support

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

Supported vGPUs

Only Q-series and C-series time-sliced vGPUs that are allocated all of the physical GPU's frame buffer are supported. MIG-backed vGPUs are **not** supported.

GPU Architecture	Board	vGPU
Ampere (compute workloads only)	NVIDIA A100 PCIe 80GB NVIDIA A100X	A100D-80C See Note (1).
	NVIDIA A100 HGX 80GB	A100DX-80C See Note (1).

GPU Architecture	Board	vGPU
	NVIDIA A100 PCIe 40GB	A100-40C See Note (1).
	NVIDIA A100 HGX 40GB	A100X-40C See Note (1).
	NVIDIA A30 NVIDIA A30X	A30-24C See Note (1).
Ampere (compute and graphics workloads)	NVIDIA A40	A40-48Q See Note (1).
		A40-48C See Note (1).
	NVIDIA A16	A16-16Q See Note (1).
		A16-16C See Note (1).
	NVIDIA A10	A10-24Q See Note (1).
		A10-24C See Note (1).
	NVIDIA A2	A2-16Q See Note (1).
		A2-16C See Note (1).
	NVIDIA RTX A6000	A6000-48Q See Note (1).
		A6000-48C See Note (1).
	NVIDIA RTX A5500	A5500-24Q See Note (1).
		A5500-24C See Note (1).
	NVIDIA RTX A5000	A5000-24Q See Note (1).
		A5000-24C See Note (1).
Turing	Tesla T4	T4-16Q
		T4-16C
Volta	Tesla V100 SXM2 32GB	V100DX-32Q
		V100D-32C
	Tesla V100 PCIe 32GB	V100D-32Q
		V100D-32C
	Tesla V100S PCIe 32GB	V100S-32Q
		V100S-32C
	Tesla V100 SXM2	V100X-16Q
		V100X-16C
	Tesla V100 PCIe	V100-16Q
		V100-16C
	Tesla V100 FHHL	V100L-16Q

GPU Architecture	Board	vGPU
		V100L-16C

Note:

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

Maximum vGPUs per VM

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

Supported Hypervisor Releases

All hypervisors that support NVIDIA AI Enterprise are supported.

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

Peer-to-peer CUDA transfers enable device memory between vGPUs on different GPUs that are assigned to the same VM to be accessed from within the CUDA kernels. NVLink is a high-bandwidth interconnect that enables fast communication between such vGPUs. Peer-to-Peer CUDA transfers over NVLink are supported only on a subset of vGPUs, VMware vSphere Hypervisor (ESXi) releases, and guest OS releases.

Supported vGPUs

Only Q-series 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.

GPU Architecture	Board	vGPU
Ampere (compute workloads only)	NVIDIA A100 PCIe 80GB NVIDIA A100X	A100D-80C
	NVIDIA A100 HGX 80GB	A100DX-80C See Note (1).
	NVIDIA A100 PCIe 40GB	A100-40C
	NVIDIA A100 HGX 40GB	A100X-40C See Note (1).
	NVIDIA A30 NVIDIA A30X	A30-24C
Ampere (compute and graphics workloads)	NVIDIA A40	A40-48Q
		A40-48C

GPU Architecture	Board	vGPU
	NVIDIA A10	A10-24Q
		A10-24C
	NVIDIA RTX A6000	A6000-48Q
		A6000-48C
	NVIDIA RTX A5500	A5500-24Q
		A5500-24C
	NVIDIA RTX A5000	A5000-24Q
		A5000-24C
Volta	Tesla V100 SXM2 32GB	V100DX-32Q
		V100DX-32C
	Tesla V100 SXM2	V100X-16Q
		V100X-16C
NVIDIA RTX A5000	A5500-24Q	

Note:

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

Supported Hypervisor Releases

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

Supported Guest OS Releases

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

Limitations

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

2.10. 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.

Supported vGPUs

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

GPU Architecture	Board	vGPU
Ampere	NVIDIA A40	A40-48C
	NVIDIA A16	A16-16C
	NVIDIA A10	A10-24C
	NVIDIA A2	A2-16C
	NVIDIA RTX A6000	A6000-48C
	NVIDIA RTX A5500	A5500-24C
	NVIDIA RTX A5000	A5000-24C

Supported Guest OS Releases

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

Limitations

- When unified memory is enabled for a VM, vGPU migration is disabled for the VM.

2.11. NVIDIA GPU Operator Support

NVIDIA GPU Operator simplifies the deployment of NVIDIA AI Enterprise with software container platforms. NVIDIA GPU Operator is supported only on specific combinations of VMware vSphere Hypervisor (ESXi) release, container platform, and guest OS release.

VMware vSphere Hypervisor (ESXi) Release	Container Platform	Guest OS
VMware vSphere Hypervisor (ESXi) 7.0 Update 2	Kubernetes 1.21 or later compatible versions	Ubuntu 22.04 LTS
VMware vSphere Hypervisor (ESXi) 7.0 Update 2	Kubernetes 1.21 or later compatible versions	Ubuntu 20.04 LTS

Chapter 3. NVIDIA AI Enterprise Supported Cloud Services

NVIDIA AI Enterprise is supported on several cloud services with bring-your-own-license (BYOL) licensing.

- Amazon Web Services Elastic Compute Cloud (AWS EC2)
- Google Cloud Platform (GCP)
- Microsoft Azure

3.1. Amazon Web Services Elastic Compute Cloud (AWS EC2)

GPU	Supported AWS EC2 Instances	Supported Guest Operating Systems
NVIDIA T4	g4dn.xlarge g4dn.2xlarge g4dn.4xlarge g4dn.8xlarge g4dn.12xlarge g4dn.16xlarge	Red Hat Enterprise Linux 8.4 Ubuntu 20.04
NVIDIA V100	P3.2xlarge P3.8xlarge P3.16xlarge	
NVIDIA A10G	g5.xlarge g5.2xlarge g5.4xlarge g5.8xlarge	

GPU	Supported AWS EC2 Instances	Supported Guest Operating Systems
	g5.12xlarge g5.16xlarge g5.24xlarge g5.48xlarge	
NVIDIA A100	p4d.24xlarge	

3.2. Google Cloud Platform (GCP)

GPU	Supported GCP Instances	Supported Guest Operating Systems
Tesla T4	Any predefined machine type. Any custom machine type that can be created in a zone.	Red Hat Enterprise Linux 8.4 Ubuntu 20.04
Tesla V100		
Tesla A100	a2-highgpu-1g a2-highgpu-2g a2-highgpu-4g a2-highgpu-8g a2-megagpu-16g	

3.3. Microsoft Azure

GPU	Supported Azure Instances	Supported Guest Operating Systems
NVIDIA V100	NC6s_v3 NC12s_v3 NC24s_v3 NC24rs_v3 ND40rs_v2	Red Hat Enterprise Linux 8.4 Ubuntu 20.04
NVIDIA T4	NC4asT4_v3 NC8asT4_v3 NC16asT4_v3 NC64asT4_v3	

GPU	Supported Azure Instances	Supported Guest Operating Systems
NVIDIA A100	ND96asr_v4 ND96amsr_A100_v4	

3.4. NVIDIA GPU Optimized VMI on CSP Marketplace

For ease of use in the cloud, NVIDIA will also provide compute optimized and validated base Virtual Machine Instances (VMI) through CSP marketplaces. The VMI includes key technologies and software from NVIDIA for rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud.

The VMI will have the following software pre-installed:

- Ubuntu Server 20.04
- NVIDIA driver 470TRD - 470.103.01
- Docker-ce 20.10.12
- NVIDIA Container Toolkit 1.8.1
- NVIDIA Container Runtime 3.8.1

Chapter 4. CPU Only Server Support

NVIDIA AI Enterprise supports deployments on CPU only servers that are part of the NVIDIA Certfied Systems list. Customers can deploy both GPU and CPU Only systems with VMware vSphere or Red Hat Enterprise Linux.

NVIDIA AI Enterprise will support the following CPU enabled frameworks:

- TensorFlow
- PyTorch
- Triton Inference Server with FIL backend
- NVIDIA RAPIDS with XGBoost and Dask

Chapter 5. Known Product Limitations

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

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

Description

Issues occur when the channels allocated to a vGPU are exhausted and the guest VM to which the vGPU is assigned fails to allocate a channel to the vGPU. A physical GPU has a fixed number of channels and the number of channels allocated to each vGPU is inversely proportional to the maximum number of vGPUs allowed on the physical GPU.

When the channels allocated to a vGPU are exhausted and the guest VM fails to allocate a channel, the following errors are reported on the hypervisor host or in an NVIDIA bug report:

```
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0): Guest attempted to
 allocate channel above its max channel limit 0xfb
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0): VGPU message 6
 failed, result code: 0x1a
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):
 0xc1d004a1, 0xff0e0000, 0xff0400fb, 0xc36f,
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):         0x1,
 0xff1fe314, 0xff1fe038, 0x100b6f000, 0x1000,
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):
 0x80000000, 0xff0e0200, 0x0, 0x0, (Not logged),
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0):         0x1, 0x0
Jun 26 08:01:25 srvxen06f vgpu-3[14276]: error: vmiop_log: (0x0): , 0x0
```

Workaround

Use a vGPU type with more frame buffer, thereby reducing the maximum number of vGPUs allowed on the physical GPU. As a result, the number of channels allocated to each vGPU is increased.

5.2. 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 AI Enterprise reserves can be calculated from the following formula:

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

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

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

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

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

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

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

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

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

compression-adjustment

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

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

vGPU Type	Compression Adjustment (MB)
T4-16Q T4-16C T4-16A	28

For all other vGPU types, *compression-adjustment* is 0.

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

Description

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

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

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

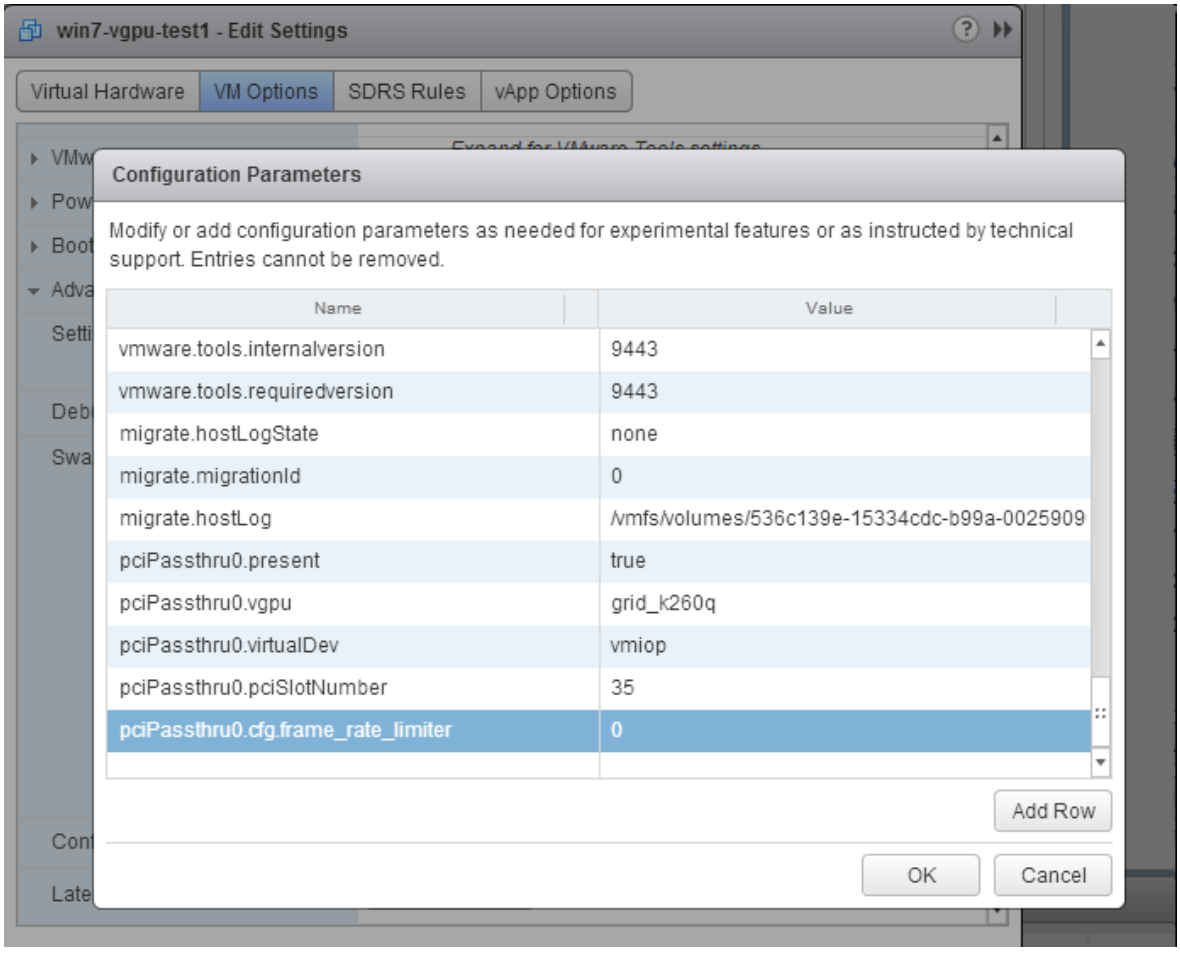
Resolution

FRL is controlled by an internal vGPU setting. On vGPUs that use the best-effort scheduler, NVIDIA does not validate vGPU with FRL disabled, but for validation of benchmark

performance, FRL can be temporarily disabled by adding the configuration parameter `pciPassthru0.cfg.frame_rate_limiter` in the VM's advanced configuration options.

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

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



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

5.4. 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 AI Enterprise GPU is present in **Windows Device Manager** but displays a warning sign, and the following device status:

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

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

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

Resolution

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

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

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

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

reservation for each additional 1 GB of system memory. For example, to support 96 GB of RAM, set `pciPassthru0.cfg.extra_fb_reservation` to 128.

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

Chapter 6. Known Issues

6.1. Migration of VMs configured with vGPU stops before the migration is complete

Description

When a VM configured with vGPU is migrated to another host, the migration stops before it is complete. After the migration stops, the VM is no longer accessible.

This issue occurs if the ECC memory configuration (enabled or disabled) on the source and destination hosts are different. The ECC memory configuration on both the source and destination hosts must be identical.

Workaround

Reboot the hypervisor host to recover the VM. Before attempting to migrate the VM again, ensure that the ECC memory configuration on both the source and destination hosts are identical.

Status

Not an NVIDIA bug

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

Ref.

200520027

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

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

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA Maxwell, NVIDIA Pascal, NVIDIA Turing, NVIDIA Volta, Quadro, and Tesla are trademarks or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com