NVIDIA AI Enterprise

Release Notes

RN-10616-001 _v3.1 | October 2024

Table of Contents

Chapter 1. What's New in NVIDIA AI Enterprise

NVIDIA AI Enterprise 3.1 is an update release that introduces some new features and enhancements, and includes bug fixes and security updates.

Changes to Hardware Supported in this Release

- Support for the following GPUs:
  - NVIDIA H800 PCIe 80GB
  - NVIDIA L4
  - NVIDIA L40
  - NVIDIA RTX 6000 Ada

Changes to Virtualization Software in this Release

- New release of NVIDIA vGPU software: 15.2
  - Updates in Release 15.2 for Red Hat Enterprise Linux KVM
  - Updates in Release 15.2 for VMware vSphere

  This release includes security updates for the NVIDIA Virtual GPU Manager and graphics driver for Linux- see *Security Bulletin: NVIDIA GPU Display Driver - March 2023*, which is listed on the NVIDIA Product Security page
- Miscellaneous bug fixes

Changes to Frameworks and Models in this Release

- AI Workflow Reference Solutions for Next Item Prediction as collections
- Addition of NVIDIA RAPIDS Accelerator for Apache Spark 23.02 to NVIDIA AI Enterprise
- New releases of the following software components of NVIDIA AI Enterprise:
  - MONAI (Medical Open Network for Artificial Intelligence) Enterprise 1.0.1
  - NVIDIA Clara Parabricks 4.0.3-1
  - NVIDIA GPU Operator: 23.3.1

- NVIDIA Network Operator: 23.1.0
- NVIDIA RAPIDS: 23.02-runtime-cuda11.8-ubuntu20.04

- New releases of the following NVIDIA deep learning frameworks:

> **Note:** These frameworks support NVIDIA CUDA Toolkit 11.8.0, **not** 12.0 NVIDIA CUDA Toolkit.

  - TensorFlow 2: 23.03-tf2-nvaie-3.1-py3
  - TensorFlow 1: 23.03-tf1-nvaie-3.1-py3
  - PyTorch: 23.03-nvaie-3.1-py3
  - NVIDIA Triton Inference Server: 23.03-nvaie-3.1-py3 and 23.03-nvaie-3.1-py3-sdk
  - NVIDIA TensorRT: 23.03-nvaie-3.1-py3

Changes to Cloud Service Support in this Release

- New Google Cloud Platform (GCP) instances based on the NVIDIA L4 GPU
- New Microsoft Azure NC_A100_v4 virtual machine instances
- Kubernetes management by the following cloud services:
  - Amazon Elastic Kubernetes Service (EKS)
  - Google Kubernetes Engine (GKE)

Chapter 2. Supported Hardware and Software

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

Servers and NVIDIA GPUs Supported

NVIDIA AI Enterprise is supported on NVIDIA® DGX™ servers in bare-metal deployments with the NVIDIA graphics driver for Linux that is included in the DGX OS software.

Note: NVIDIA vGPU software is **not** supported on NVIDIA DGX servers.

NVIDIA AI Enterprise is supported on the following NVIDIA GPUs with the compatible third-party servers that are listed on the NVIDIA-certified systems page.

- NVIDIA A800 PCIe 80GB
- NVIDIA A800 PCIe 80GB liquid cooled
- NVIDIA A800 HGX 80GB
- NVIDIA A100X
- NVIDIA A100 PCIe 40GB
- NVIDIA A100 HGX 40GB
- NVIDIA A100 PCIe 80GB
- NVIDIA A100 PCIe 80GB liquid cooled
- NVIDIA A100 HGX 80GB
- NVIDIA A40
- NVIDIA A30X
- NVIDIA A30
- NVIDIA A10
- NVIDIA A16
- NVIDIA A2
- NVIDIA H100 PCIe 80GB
- NVIDIA H800 PCIe 80GB

- NVIDIA L4
- NVIDIA L40
- NVIDIA RTX A6000
- NVIDIA RTX A5500
- NVIDIA RTX A5000
- NVIDIA RTX 6000 passive
- NVIDIA RTX 8000 passive
- NVIDIA RTX 6000 Ada
- NVIDIA T4
- NVIDIA V100

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 Supported

- Red Hat Enterprise Linux with KVM hypervisor 9.1, 9.0
- Red Hat Enterprise Linux with KVM hypervisor 8.7, 8.6, and 8.4
- VMware vSphere Hypervisor (ESXi) Enterprise Plus Edition 8.0
- VMware vCenter Server 8.0
- VMware vSphere Hypervisor (ESXi) Enterprise Plus Edition 7.0 Update 3
- VMware vCenter Server 7.0 Update 3

Microsoft Windows Guest Operating Systems Supported

Note:

- NVIDIA AI Enterprise supports **only** the Tesla Compute Cluster (TCC) driver model for Windows guest drivers.
- Windows guest OS support is limited to running applications natively in Windows VMs **without** containers. NVIDIA AI Enterprise features that depend on containerization of applications are not supported on Windows guest operating systems.

Guest OS	Red Hat Enterprise Linux KVM	VMware vSphere
Microsoft Windows Server 2022	9.1, 9.0 8.7, 8.6, 8.4	8.0 7.0 Update 3

Guest OS	Red Hat Enterprise Linux KVM	VMware vSphere
Microsoft Windows Server 2019	9.1, 9.0 8.7, 8.6, 8.4	8.0 7.0 Update 3
Microsoft Windows 11	Not supported	8.0 7.0 Update 3
Microsoft Windows 10	Not supported	8.0 7.0 Update 3

Linux Guest Operating Systems Supported

Note: Red Hat Enterprise Linux guest OS support is limited to running containers by using Docker **without** Kubernetes. NVIDIA AI Enterprise features that depend on Kubernetes, for example, the use of GPU Operator, are not supported on Red Hat Enterprise Linux.

Guest OS	Red Hat Enterprise Linux KVM	VMware vSphere
Red Hat Enterprise Linux 9.1, 9.0	9.1, 9.0 8.7, 8.6, 8.4	8.0 7.0 Update 3
Red Hat Enterprise Linux 8.7, 8.6, 8.4	9.1, 9.0 8.7, 8.6, 8.4	8.0 7.0 Update 3
Red Hat OpenShift 4.9 and later using Red Hat Linux CoreOS (RHCOS)	9.1, 9.0 8.7, 8.6, 8.4	8.0 7.0 Update 3
SUSE Linux Enterprise Server 15 SP2+	Not supported	8.0 7.0 Update 3
Ubuntu 22.04 LTS	Not supported	8.0 7.0 Update 3
Ubuntu 20.04 LTS	Not supported	8.0 7.0 Update 3

2.1. NVIDIA AI Enterprise Software Components

Software Component	NVIDIA Release
NVIDIA vGPU software	15.2: ▶ Virtual GPU Manger: 525.105.14 ▶ Graphics Driver for Windows: 528.89 ▶ Graphics Driver for Linux: 525.105.17
NVIDIA GPU Operator	23.3.1
NVIDIA Network Operator	23.1.0
TensorFlow 2	23.03-tf2-nvaie-3.1-py3
TensorFlow 1	23.03-tf1-nvaie-3.1-py3
PyTorch	23.03-nvaie-3.1-py3
NVIDIA Triton Inference Server	23.03-nvaie-3.1-py3 and 23.03-nvaie-3.1-py3-sdk
NVIDIA TensorRT	23.03-nvaie-3.1-py3
NVIDIA RAPIDS	23.02-runtime-cuda11.8-ubuntu20.04
NVIDIA RAPIDS Accelerator for Apache Spark	23.02
NVIDIA Clara Parabricks	4.0.3-1
NVIDIA DeepStream	6.2.0-triton (PDF)
MONAI (Medical Open Network for Artificial Intelligence) Enterprise	1.0.1
TAO Toolkit for Language Model (Conv AI)	4.0.0-tf2-base
TAO Toolkit for Conv AI	4.0.0-tf2-base
TAO Toolkit for CV	4.0.0-tf1.15.5

Note:

- The NVIDIA deep learning frameworks support NVIDIA CUDA Toolkit11.8.0, **not** 12.0 NVIDIA CUDA Toolkit.
- In this release, TAO Toolkit is supported **only** on the NVIDIA H100 GPU.

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

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

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

GPU	Mode as Supplied from the Factory
NVIDIA A40	Display-off
NVIDIA L40	Display-off
NVIDIA RTX 6000 Ada	Display enabled
NVIDIA RTX A5000	Display enabled
NVIDIA RTX A5500	Display enabled
NVIDIA RTX A6000	Display enabled

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

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

Note:

Only the following GPUs support the `displaymodeselector` tool:

- NVIDIA A40
- NVIDIA L40
- NVIDIA RTX A5000
- NVIDIA RTX 6000 Ada
- 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).

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

GPU	MMIO Space Required
Quadro RTX 6000 Passive	64 GB
Quadro RTX 8000 Passive	64 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 12.0.

To build a CUDA application, the system must have the NVIDIA CUDA Toolkit and the libraries required for linking. For details of the components of NVIDIA CUDA Toolkit, refer to *NVIDIA CUDA Toolkit Release Notes for CUDA 11.4*.

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

For more information about NVIDIA CUDA Toolkit, refer to CUDA Toolkit 12.0 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 hypervisor software releases.

2.8.1. vGPUs that Support Multiple vGPUs Assigned to a VM

The supported vGPUs depend on the hypervisor:

- For Red Hat Enterprise Linux KVM, **all** Q-series and C-series vGPUs are supported. On GPUs that support the Multi-Instance GPU (MIG) feature, both time-sliced and MIG-backed vGPUs are supported.
- For VMware vSphere, the supported vGPUs depend on the hypervisor release:
  - **Since VMware vSphere 8.0: All** Q-series and C-series vGPUs are supported. On GPUs that support the Multi-Instance GPU (MIG) feature, both time-sliced and MIG-backed vGPUs are supported.
  - **VMware vSphere 7.x releases:** Only Q-series vGPUs that are allocated all of the physical GPU's frame buffer are supported.

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

Multiple vGPU Support on the NVIDIA Ada Lovelace Architecture

Board	vGPU
NVIDIA L40	All C-series vGPUs

Board	vGPU
	Since VMware vSphere 8.0: All C-series vGPUs **VMware vSphere 7.x releases:** L40-48C
NVIDIA L4	All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** L4-24C
NVIDIA RTX 6000 Ada	All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** RTX 6000 Ada-48C

Multiple vGPU Support on the NVIDIA Hopper GPU Architecture

Board	vGPU
NVIDIA H800 PCIe 80GB	All C-series vGPUs See Note (1).
NVIDIA H100 PCIe 80GB	All C-series vGPUs See Note (1).

Multiple vGPU Support on the NVIDIA Ampere GPU Architecture

Board	vGPU
NVIDIA A800 PCIe 80GB NVIDIA A800 PCIe 80GB liquid cooled	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** A800D-80C See Note (1).
NVIDIA A800 HGX 80GB	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs

Board	vGPU
	VMware vSphere 7.x releases: A800DX-80C See Note (1).
NVIDIA A100 PCIe 80GB NVIDIA A100 PCIe 80GB liquid cooled	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** A100D-80C See Note (1).
NVIDIA A100 HGX 80GB	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** A100DX-80C See Note (1).
NVIDIA A100 PCIe 40GB	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** A100-40C See Note (1).
NVIDIA A100 HGX 40GB	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** A100X-40C See Note (1).
NVIDIA A40	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs

Board	vGPU
	VMware vSphere 7.x releases: A40-48C See Note (1).
NVIDIA A30	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** A30-24C See Note (1).
NVIDIA A16	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** A16-16C See Note (1).
NVIDIA A10	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** A10-24C See Note (1).
NVIDIA RTX A6000	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** A6000-48C See Note (1).
NVIDIA RTX A5500	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** A5500-24C

Board	vGPU
	See Note (1).
NVIDIA RTX A5000	**Red Hat Enterprise Linux KVM:** All C-series vGPUs **Since VMware vSphere 8.0: All** C-series vGPUs **VMware vSphere 7.x releases:** A5000-24C See Note (1).

Multiple vGPU Support on the NVIDIA Turing GPU Architecture

Board	vGPU
Tesla T4	**Red Hat Enterprise Linux KVM:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **Since VMware vSphere 8.0:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **VMware vSphere 7.x releases:** ▶ T4-16Q ▶ T4-16C
Quadro RTX 6000 passive	**Red Hat Enterprise Linux KVM:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **Since VMware vSphere 8.0:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **VMware vSphere 7.x releases:** ▶ RTX6000P-24Q ▶ RTX6000P-24C
Quadro RTX 8000 passive	**Red Hat Enterprise Linux KVM:**

Board	vGPU
	▶ All Q-series vGPUs ▶ All C-series vGPUs **Since VMware vSphere 8.0:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **VMware vSphere 7.x releases:** ▶ RTX8000P-48Q ▶ RTX8000P-48C

Multiple vGPU Support on the NVIDIA Volta GPU Architecture

Board	vGPU
Tesla V100 SXM2 32GB	**Red Hat Enterprise Linux KVM:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **Since VMware vSphere 8.0:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **VMware vSphere 7.x releases:** ▶ V100DX-32Q ▶ V100D-32C
Tesla V100 PCIe 32GB	**Red Hat Enterprise Linux KVM:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **Since VMware vSphere 8.0:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **VMware vSphere 7.x releases:** ▶ V100D-32Q ▶ V100D-32C
Tesla V100S PCIe 32GB	**Red Hat Enterprise Linux KVM:**

Board	vGPU
	▶ All Q-series vGPUs ▶ All C-series vGPUs **Since VMware vSphere 8.0:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **VMware vSphere 7.x releases:** ▶ V100S-32Q ▶ V100S-32C
Tesla V100 SXM2	**Red Hat Enterprise Linux KVM:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **Since VMware vSphere 8.0:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **VMware vSphere 7.x releases:** ▶ V100X-16Q ▶ V100X-16C
Tesla V100 PCIe	**Red Hat Enterprise Linux KVM:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **Since VMware vSphere 8.0:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **VMware vSphere 7.x releases:** ▶ V100-16Q ▶ V100-16C
Tesla V100 FHHL	**Red Hat Enterprise Linux KVM:** ▶ All Q-series vGPUs ▶ All C-series vGPUs **Since VMware vSphere 8.0:**

Board	vGPU
	▶ All Q-series vGPUs ▶ All C-series vGPUs **VMware vSphere 7.x releases:** ▶ V100L-16Q ▶ V100L-16C

Note:

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

2.8.2. Maximum Number of vGPUs Supported per VM

For Red Hat Enterprise Linux KVM, NVIDIA AI Enterprise supports up to a maximum of 16 vGPUs per VM.

For VMware vSphere, the maximum number of vGPUs per VM supported depends on the hypervisor release:

- **Since VMware vSphere 8.0:** NVIDIA AI Enterprise supports up to a maximum of eight vGPUs per VM.
- **VMware vSphere 7.x releases:** NVIDIA AI Enterprisesupports up to a maximum of four vGPUs per VM.

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

All hypervisor releases 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.

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

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

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

Board	vGPU
NVIDIA H800 PCIe 80GB	H800-80C
NVIDIA H100 PCIe 80GB	H100-80C

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

Board	vGPU
NVIDIA A800 PCIe 80GB NVIDIA A800 PCIe 80GB liquid cooled	A800D-80C
NVIDIA A800 HGX 80GB	A800DX-80C See Note (1).
NVIDIA A100 PCIe 80GB	A100D-80C
NVIDIA A100 HGX 80GB NVIDIA A100 HGX 80GB liquid cooled	A100DX-80C See Note (1).
NVIDIA A100 PCIe 40GB	A100-40C
NVIDIA A100 HGX 40GB	A100X-40C See Note (1).
NVIDIA A40	A40-48Q A40-48C
NVIDIA A30	A30-24C
NVIDIA A10	A10-24Q

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

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

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

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

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

Note:

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

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

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

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

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

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

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

- NVSwitch is not supported. Only direct connections are supported.
- Only time-sliced vGPUs are supported. MIG-backed vGPUs are **not** supported.
- If unified memory is enabled, peer-to-peer CUDA transfers are not supported on GPUs based on the NVIDIA Ampere GPU architecture that also support MIG-backed vGPUs.
- PCIe is not supported.
- SLI is not supported.

2.10. GPUDirect Technology Support

NVIDIA GPUDirect® Remote Direct Memory Access (RDMA) technology enables network devices to directly access vGPU frame buffer, bypassing CPU host memory altogether. GPUDirect Storage technology enables a direct data path for direct memory access (DMA) transfers between GPU memory and storage. GPUDirect technology is supported only on a subset of vGPUs and guest OS releases.

Supported vGPUs

GPUDirect RDMA and GPUDirect Storage technology are supported on all time-sliced and MIG-backed C-series vGPUs on physical GPUs that support single root I/O virtualization (SR-IOV).

- GPUs based on the NVIDIA Ada Lovelace GPU architecture:
  - NVIDIA L40
  - NVIDIA L4
  - NVIDIA RTX 6000 Ada
- GPUs based on the NVIDIA Hopper GPU architecture:

- NVIDIA H800 PCIe 80GB
- NVIDIA H100 PCIe 80GB

- GPUs based on the NVIDIA Ampere GPU architecture:
  - NVIDIA A800 PCIe 80GB
  - NVIDIA A800 PCIe 80GB liquid cooled
  - NVIDIA A800 HGX 80GB
  - NVIDIA A100 PCIe 80GB
  - NVIDIA A100 PCIe 80GB liquid cooled
  - NVIDIA A100 HGX 80GB
  - NVIDIA A100 PCIe 40GB
  - NVIDIA A100 HGX 40GB
  - NVIDIA A100X
  - NVIDIA A30
  - NVIDIA A30X
  - NVIDIA A40
  - NVIDIA A16
  - NVIDIA A10
  - NVIDIA A2
  - NVIDIA RTX A6000
  - NVIDIA RTX A5500
  - NVIDIA RTX A5000

Supported Guest OS Releases

Linux only. GPUDirect technology is **not** supported on Windows.

Supported Network Interface Cards

GPUDirect technology is supported on the following network interface cards:

- NVIDIA ®ConnectX®- 7 SmartNIC
- Mellanox Connect-X 6 SmartNIC
- Mellanox Connect-X 5 Ethernet adapter card

Limitations

GPUDirect Storage technology is supported only on the following guest OS releases:

- Ubuntu 22.04 LTS
- Ubuntu 20.04 LTS

2.11. Unified Memory Support

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

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

2.11.1. vGPUs that Support Unified Memory

On GPUs that support the Multi-Instance GPU (MIG) feature, **all** MIG-backed vGPUs are supported. Only time-sliced vGPUs that are allocated all of the physical GPU's frame buffer on physical GPUs that support unified memory are supported.

Unified Memory Support on the NVIDIA Ada Lovelace GPU Architecture

Board	vGPU
NVIDIA L40	L40-48Q L40-48C
NVIDIA L4	L4-24Q L4-24C
NVIDIA RTX 6000 Ada	RTX 6000 Ada-48Q RTX 6000 Ada-48C

Unified Memory Support on the NVIDIA Hopper GPU Architecture

Board	vGPU
NVIDIA H800 PCIe 80GB	H800-80C **All** MIG-backed vGPUs
NVIDIA H100 PCIe 80GB	H100-80C **All** MIG-backed vGPUs

Unified Memory Support on the NVIDIA Ampere GPU Architecture

Board	vGPU
NVIDIA A800 PCIe 80GB NVIDIA A800 PCIe 80GB liquid cooled	A800D-80C **All** MIG-backed vGPUs
NVIDIA A800 HGX 80GB	A800DX-80C **All** MIG-backed vGPUs
NVIDIA A100 PCIe 80GB NVIDIA A100 PCIe 80GB liquid cooled NVIDIA A100X	A100D-80C **All** MIG-backed vGPUs
NVIDIA A100 HGX 80GB	A100DX-80C **All** MIG-backed vGPUs
NVIDIA A100 PCIe 40GB	A100-40C **All** MIG-backed vGPUs
NVIDIA A100 HGX 40GB	A100X-40C **All** MIG-backed vGPUs
NVIDIA A40	A40-48Q A40-48C
NVIDIA A30	A30-24C **All** MIG-backed vGPUs
NVIDIA A16	A16-16Q A16-16C
NVIDIA A10	A10-24Q A10-24C
NVIDIA RTX A6000	A6000-48Q A6000-48C
NVIDIA RTX A5500	A5500-24Q

Board	vGPU
	A5500-24C
NVIDIA RTX A5000	A5000-24Q A5000-24C

2.11.2. Guest OS Releases that Support Unified Memory

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

2.11.3. Limitations on Support for Unified Memory

- Only time-sliced Q-series and C-series vGPUs that are allocated all of the physical GPU's frame buffer on physical GPUs that support unified memory are supported. Fractional time-sliced vGPUs are **not** supported.
- When unified memory is enabled for a VM, vGPU migration is disabled for the VM.

2.12. 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 hypervisor software release, container platform, and guest OS release.

Hypervisor Software Release	Container Platform	Guest OS
Red Hat Enterprise Linux KVM 9.1, 9.0	Red Hat OpenShift 4.9 and later using Red Hat Linux CoreOS (RHCOS) and the CRI-O container runtime	Red Hat OpenShift 4.9 and later using RHCOS
Red Hat Enterprise Linux KVM 8.7, 8.6, 8.4	Red Hat OpenShift 4.9 and later using RHCOS and the CRI-O container runtime	Red Hat OpenShift 4.9 and later using RHCOS
VMware vSphere Hypervisor (ESXi) 8.0	Upstream Kubernetes 1.21 through 1.25	Ubuntu 22.04 LTS
		Ubuntu 20.04 LTS
	VMware vSphere with Tanzu 7.0 U3c	Ubuntu 22.04 LTS
		Ubuntu 22.04 LTS
	HPE Ezmeral Runtime Enterprise 5.5	Red Hat Enterprise Linux 8.4
VMware vSphere Hypervisor (ESXi) 7.0 Update 2, Update 3	Upstream Kubernetes 1.21 through 1.25	Ubuntu 22.04 LTS
		Ubuntu 20.04 LTS

Hypervisor Software Release	Container Platform	Guest OS
	VMware vSphere with Tanzu 7.0 U3c	Ubuntu 22.04 LTS
		Ubuntu 22.04 LTS
	HPE Ezmeral Runtime Enterprise 5.5	Red Hat Enterprise Linux 8.4

2.13. NVIDIA RAPIDS Accelerator for Apache Spark Support

NVIDIA RAPIDS Accelerator for Apache Spark is a software component of NVIDIA AI Enterprise. It uses NVIDIA GPUs to accelerate Spark data frame workloads transparently, that is, without code changes.

NVIDIA AI Enterprise supports RAPIDS Accelerator for Apache Spark on the following platforms:

- Google Cloud Dataproc
- Databricks on the following cloud services:
  - Amazon Web Services (AWS)
  - Microsoft Azure
- Amazon EMR (formerly “Amazon Elastic MapReduce”)

Chapter 3. NVIDIA AI Enterprise Supported Cloud Services

NVIDIA AI Enterprise is supported on several cloud services with bring-your-own-license (BYOL) licensing. Pay-as-you-go licensing is also available with some cloud services.

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

Note: Red Hat Enterprise Linux guest OS support is limited to running containers by using Docker **without** Kubernetes. NVIDIA AI Enterprise features that depend on Kubernetes, for example, the use of GPU Operator, are not supported on Red Hat Enterprise Linux.

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 Red Hat Enterprise Linux 7.9 Red Hat OpenShift 4.10 using Red Hat Linux CoreOS (RHCOS) Red Hat OpenShift 4.9 using Red Hat Linux CoreOS (RHCOS) Ubuntu 22.04 Ubuntu 20.04
NVIDIA V100	P3.2xlarge P3.8xlarge P3.16xlarge	
NVIDIA A10G	g5.xlarge	

GPU	Supported AWS EC2 Instances	Supported Guest Operating Systems
	g5.2xlarge g5.4xlarge g5.8xlarge g5.12xlarge g5.16xlarge g5.24xlarge g5.48xlarge	
NVIDIA A100	p4d.24xlarge	

3.2. Google Cloud Platform (GCP)

Note: Pay-as-you-go licensing is also available for all supported GCP instances.

GPU	Supported GCP Instances	Supported Guest Operating Systems
NVIDIA A100	a2-highgpu-1g a2-highgpu-2g a2-highgpu-4g a2-highgpu-8g a2-megagpu-16g	Red Hat Enterprise Linux 8.4 Red Hat Enterprise Linux 7.9 Red Hat OpenShift 4.10 using Red Hat Linux CoreOS (RHCOS) Red Hat OpenShift 4.9 using Red Hat Linux CoreOS (RHCOS) Ubuntu 22.04 Ubuntu 20.04
NVIDIA L4	g2-standard-4 g2-standard-8 g2-standard-12 g2-standard-16 g2-standard-24 g2-standard-32 g2-standard-48 g2-standard-96	
NVIDIA T4	Any predefined machine type.	
NVIDIA V100	Any custom machine type that can be created in a zone.	

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 Red Hat Enterprise Linux 7.9 Red Hat OpenShift 4.10 using Red Hat Linux CoreOS (RHCOS) Red Hat OpenShift 4.9 using Red Hat Linux CoreOS (RHCOS) Ubuntu 22.04 Ubuntu 20.04
NVIDIA T4	NC4asT4_v3 NC8asT4_v3 NC16asT4_v3 NC64asT4_v3	
NVIDIA A100	NC24ads_A100_v4 NC48ads_A100_v4 NC96ads_A100_v4 ND96asr_v4 ND96amsr_A100_v4	
NVIDIA A10	NV6ads_A10_v5 NV12ads_A10_v5 NV18ads_A10_v5 NV36ads_A10_v5 NV36adms_A10_v5 NV72ads_A10_v5	

3.4. Oracle Cloud Infrastructure

GPU	Oracle Cloud Infrastructure Shapes	Supported Guest Operating Systems
NVIDIA P100	VM.GPU2.1 BM.GPU2.2	Linux: ▶ Ubuntu 22.04 ▶ Ubuntu 20.04 Windows:
NVIDIA V100	VM.GPU3.1 VM.GPU3.2 VM.GPU3.4	

GPU	Oracle Cloud Infrastructure Shapes	Supported Guest Operating Systems
	BM.GPU3.8	▶ Microsoft Windows Server 2022 ▶ Microsoft Windows Server 2019
NVIDIA A100	BM.GPU.GM4.8 BM.GPU4.8	
NVIDIA A10	BM.GPU.GU1.4	

3.5. NVIDIA GPU Optimized VMI on CSP Marketplace

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

Each VMI has the following software pre-installed:

- Ubuntu Server 20.04
- NVIDIA driver 525 TRD - 525.60.13
- 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 Certified 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
RTX6000-12Q RTX6000-12C RTX6000-12A	32
RTX6000-24Q RTX6000-24C RTX6000-24A	104
RTX6000P-12Q RTX6000P-12C RTX6000P-12A	32
RTX6000P-24Q RTX6000P-24C RTX6000P-24A	104
RTX8000-12Q RTX8000-12C RTX8000-12A	32
RTX8000-16Q RTX8000-16C RTX8000-16A	64
RTX8000-24Q	96

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

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

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

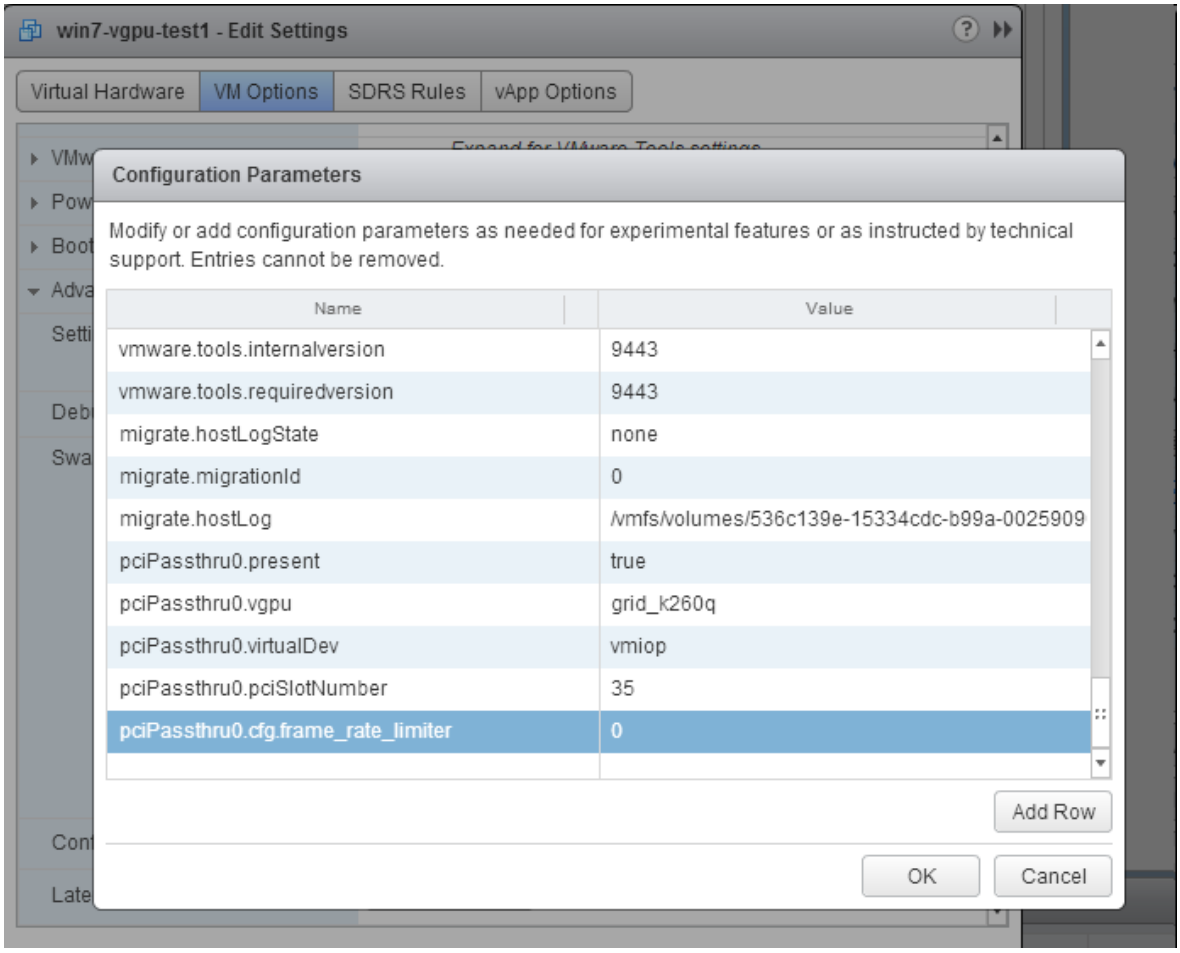
frame rendering rates, as compared to the same benchmarks running on a pass-through GPU.

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.

Resolution

FRL is controlled by an internal vGPU setting. On vGPUs that use the best-effort scheduler, NVIDIA does not validate vGPU with FRL disabled, but for validation of benchmark performance, FRL can be temporarily disabled by setting `frame_rate_limiter=0` in the vGPU configuration file.

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

For example:

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

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

With this setting in place, the VM's vGPU will run without any frame rate limit.

The FRL can be reverted back to its default setting as follows:

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

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

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

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

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

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

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

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. MIG mode cannot be changed on a single NVIDIA H100 or H800 in a multi-GPU system

Description

MIG mode cannot be enabled or disabled on a single NVIDIA H100 or NVIDIA H800 GPU in a multi-GPU system. When this issue occurs, the following error message is displayed:

```
NVML: Unable to get MIG mode: Invalid Argument
```

This issue occurs **only** in response to running the **`nvidia-smi -mig -i gpu-index`** command to change the MIG mode of a single NVIDIA H100 or H800 GPU in a multi-GPU system.

This issue does not occur in any of the following situations:

- The command is run to change the MIG mode of any other GPU that supports the MIG feature, such as any variant of the NVIDIA A100 and NVIDIA A800 GPUs.
- The system contains only one NVIDIA H100 or NVIDIA H800 GPU.
- The **`-i gpu-index`** is omitted from the command to change the MIG mode.

Status

Open

Ref.

4008029

6.2. Virtual GPU Manager upgrade fails on VMware vSphere Hypervisor (ESXi)

Description

Upgrading the Virtual GPU Manager from an earlier NVIDIA AI Enterprise release branch to the current release fails on VMware vSphere Hypervisor (ESXi). The installation result contains the message `Host is not changed`.

Version

This issue affects upgrades of the Virtual GPU Manager from an earlier NVIDIA AI Enterprise release branch to the current release.

Workaround

Uninstall the Virtual GPU Manager from the earlier NVIDIA AI Enterprise release branch before installing the current release of the Virtual GPU Manager.

Status

Open

Ref.

3913505

6.3. The NVIDIA MOFED driver container fails to install the driver if Network Operator is installed

Description

The NVIDIA MOFED driver container fails to install the driver if Network Operator is installed. The installation fails because the container fails to unload the `ib_core` module. The `rdma-core` package is installed as part of the Red Hat CoreOS installation. This package loads the `ib_core` module if the system has Mellanox network interface cards (NICs).

Status

Open

Ref.

3565857

6.4. 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.

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

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

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