NVIDIA Data Center GPU Driver version 450.248.02(Linux) / 454.23 (Windows)

Release Notes

Table of Contents

Chapter 1. Version Highlights

This section provides highlights of the NVIDIA Data Center GPU 450 Driver (version 450.248.02 Linux and 454.23 Windows).

For changes related to the 450 release of the NVIDIA display driver, review the file "NVIDIA_Changelog" available in the .run installer packages.

▶ Linux driver release date: 6/26/2023

▶ Windows driver release date: 6/26/2023

1.1. Software Versions

For this release, the software versions are listed below.

▶ CUDA Toolkit 11: 11.0.3

Note that starting with CUDA 11, individual components of the toolkit are versioned independently. For a full list of the individual versioned components (such as nvcc, CUDA libraries, and so on), see the CUDA Toolkit Release Notes

▶ NVIDIA Data Center GPU Driver: 450.248.02 (Linux) / 454.23 (Windows)

▶ Fabric Manager: 450.248.02 (Use `nv-fabricmanager -v`)

▶ NVFlash: 5.791

Due to a revision lock between the VBIOS and driver, VBIOS versions >= 92.00.18.00.00 must use corresponding drivers >= 450.36.01. Older VBIOS versions will work with newer drivers.

For more information on getting started with the NVIDIA Fabric Manager on NVSwitch-based systems (for example, HGX A100), refer to the Fabric Manager User Guide.

1.2. Fixed Issues

▶ 3954952 - Redfish Row Remapping Failed event was not logged when DRAM uncontained is injected at same address after reset GPU.

▶ 4017428 - Fixed an issue where the activation of Shared NVSwitch partitions fails for GPU partitions of 8x or 16x due to a timeout during GPU opening and initialization.

1.3.   Known Issues

General

▶ The GPU driver build system might not pick the `Module.symvers` file, produced when building the `ofa_kernel` module from MLNX_OFED, from the right subdirectory. Because of that, `nvidia_peermem.ko` does not have the right kernel symbol versions for the APIs exported by the IB core driver, and therefore it does not load correctly. That happens when using MLNX_OFED 5.5 or newer on a Linux Arm64 or ppc64le platform.

To work around this issue, perform the following:

1. Verify that nvidia_peermem.ko does not load correctly.
2. Uninstall old MLNX_OFED if one was installed.
3. Manually remove `/usr/src/ofa_kernel/default` if one exists.
4. Install MLNX_OFED 5.5 or newer.
5. Manually create a soft link:
   `/usr/src/ofa_kernel/default -> /usr/src/ofa_kernel/$(uname -m)/$(uname -r)`
6. Reinstall the GPU driver.

▶ Combining A800 and A100 SXM modules in a single server is not currently supported with this driver version.

▶ Combining A800 and A100 PCIe with NVLink is not fully tested.

▶ There is a known issue with installing the Linux driver where installing the driver interferes with the `ipmitool`.

▶ When installing a driver on SLES15 or openSUSE15 that previously had an R515 driver installed, users need to run the following command afterwards to finalize the installation:

```
sudo zypper install --force nvidia-gfxG05-kmp-default
```

Without doing this, users may see the kernel objects as missing.

▶ By default, Fabric Manager runs as a systemd service. If using `DAEMONIZE=0` in the Fabric Manager configuration file, then the following steps may be required.

1. Disable FM service from auto starting. (`systemctl disable nvidia-fabricmanager`)
2. Once the system is booted, manually start FM process. (`/usr/bin/nv-fabricmanager -c /usr/share/nvidia/nvswitch/fabricmanager.cfg`). Note, since the process is not a daemon, the SSH/Shell prompt will not be returned (use another SSH shell for other activities or run FM as a background task).

▶ There is a known issue with cross-socket GPU to GPU memory consistency that is currently under investigation

GPU Performance Counters

The use of developer tools from NVIDIA that access various performance counters requires administrator privileges. See this note for more details. For example, reading NVLink utilization metrics from `nvidia-smi (nvidia-smi nvlink –g 0)` would require administrator privileges.

NoScanout Mode

NoScanout mode is no longer supported on NVIDIA Data Center GPU products. If NoScanout mode was previously used, then the following line in the "screen" section of /etc/X11/xorg.conf should be removed to ensure that X server starts on data center products:

```
Option          "UseDisplayDevice" "None"
```

NVIDIA Data Center GPU products now support one display of up to 4K resolution.

Unified Memory Support

CUDA and unified memory is not supported when used with Linux power management states S3/S4.

IMPU FRU for Volta GPUs

The driver does not support the IPMI FRU multi-record information structure for NVLink. See the Design Guide for Tesla P100 and Tesla V100-SXM2 for more information.

Experimental OpenCL Features

Select features in OpenCL 2.0 are available in the driver for evaluation purposes only.

The following are the features as well as a description of known issues with these features in the driver:

Device side enqueue

▶ The current implementation is limited to 64-bit platforms only.

▶ OpenCL 2.0 allows kernels to be enqueued with global_work_size larger than the compute capability of the NVIDIA GPU. The current implementation supports only combinations of global_work_size and local_work_size that are within the compute capability of the NVIDIA GPU. The maximum supported CUDA grid and block size of NVIDIA GPUs is available at http://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#computecapabilities. For a given grid dimension, the global_work_size can be determined by CUDA grid size x CUDA block size.

► For executing kernels (whether from the host or the device), OpenCL 2.0 supports non-uniform ND-ranges where global_work_size does not need to be divisible by the local_work_size. This capability is not yet supported in the NVIDIA driver, and therefore not supported for device side kernel enqueues.

Shared virtual memory

► The current implementation of shared virtual memory is limited to 64-bit platforms only.

Chapter 2. Virtualization

To make use of GPU passthrough with virtual machines running Linux, the hardware platform must support the following features:

▶ A CPU with hardware-assisted instruction set virtualization: Intel VT-x or AMD-V.

▶ Platform support for I/O DMA remapping.

▶ On Intel platforms the DMA remapper technology is called Intel VT-d.

▶ On AMD platforms it is called AMD IOMMU.

Support for these features varies by processor family, product, and system, and should be verified at the manufacturer's website.

The following hypervisors are supported for virtualization:

Hypervisor	Notes
Citrix XenServer	Version 6.0 and later
VMware vSphere (ESX / ESXi)	Version 5.1 and later.
Red Hat KVM	Red Hat Enterprise Linux 7 with KVM
Microsoft Hyper-V	Windows Server 2016 Hyper-V Generation 2 Windows Server 2012 R2 Hyper-V

Tesla products now support one display of up to 4K resolution.

The following GPUs are supported for device passthrough for virtualization:

GPU Family	Boards Supported
NVIDIA Ampere GPU Architecture	NVIDIA A100, A800
Turing	NVIDIA T4
Volta	NVIDIA V100
Pascal	Quadro: P2000, P4000, P5000, P6000, GP100 Tesla: P100, P40, P4

GPU Family	Boards Supported
Maxwell	Quadro: K2200, M2000, M4000, M5000, M6000, M6000 24GB
	Tesla: M60, M40, M6, M4
Kepler	Quadro: K2000, K4000, K4200, K5000, K5200, K6000
	Tesla: K520, K80
	Tesla: K10, K20, K20x, K20Xm, K20c, K20s, K40m, K40c, K40s, K40st, K40t, K80, K520

Chapter 3. Hardware and Software Support

Support for these features varies by processor family, product, and system, and should be verified at the manufacturer's website.

Supported Operating Systems for NVIDIA Data Center GPUs

The Release 450 driver is supported on the following operating systems:

▶ Windows x86_64 operating systems:

  ▶ Microsoft Windows® Server 2019
  ▶ Microsoft Windows® Server 2016
  ▶ Microsoft Windows® 10

▶ The following table summarizes the supported Linux 64-bit distributions. For a complete list of distributions, kernel versions supported, see the CUDA Linux System Requirements documentation.

Distribution	x86_64	POWER	Arm64 Server
OpenSUSE Leap 15.x (where y <= 4)	Yes	No	No
Red Hat Enterprise Linux 8.y (where y <= 7)	Yes	Yes	Yes
Red Hat Enterprise Linux / CentOS 7.y (where y <= 9)	Yes	No	No
Rocky Linux 8.x (where x<=7)	Yes	No	No
SUSE Linux Enterprise Server 15.x (where y <= 4)	Yes	No	Yes
Ubuntu 20.04.x LTS (where y <= 6)	Yes	No	No

Distribution	x86_64	POWER	Arm64 Server
Ubuntu 18.04.z LTS (where z <= 6)	Yes	Yes	Yes

Supported Operating Systems and CPU Configurations for NVIDIA HGX A100 and NVIDIA HGX A800

The Release 450 driver is validated with HGX A100 and HGX A800 on the following operating systems and CPU configurations:

▶ Linux 64-bit distributions:

   ▶ Red Hat Enterprise Linux 8.7 (in 4/8/16-GPU configurations)

   ▶ CentOS Linux 7.9 (in 4/8/16-GPU configurations)

   ▶ Ubuntu 18.04.6 LTS (in 4/8/16-GPU configurations)

   ▶ SUSE SLES 15.4 (in 4/8/16-GPU configurations)

▶ Windows 64-bit distributions:

   ▶ Windows Server 2019 (in 1/2/4/8-GPU configurations; 16-GPU configurations are currently not supported)

   Windows is supported only in shared NVSwitch virtualization configurations.

▶ CPU Configurations:

   ▶ AMD Rome in PCIe Gen4 mode

   ▶ Intel Skylake/Cascade Lake (4-socket) in PCIe Gen3 mode

Supported Virtualization Configurations

The Release 450 driver is validated with HGX A100 on the following configurations:

▶ Passthrough (full visibility of GPUs and NVSwitches to guest VMs):

   ▶ 8-GPU configurations with Ubuntu 18.04.6 LTS

▶ Shared NVSwitch (guest VMs only have visibility of GPUs and full NVLink bandwidth between GPUs in the same guest VM):

   ▶ 1/2/4/8/16-GPU configurations with Ubuntu 18.04.6 LTS

   ▶ 1/2/4/8-GPU configurations with Windows Server 2019

API Support

This release supports the following APIs:

▶ NVIDIA® CUDA® 11.0 for NVIDIA® Kepler™, Maxwell™, Pascal™, Volta™, Turing™ and NVIDIA Ampere architecture GPUs

▶ OpenGL® 4.5

- Vulkan® 1.1
- DirectX 11
- DirectX 12 (Windows 10)
- Open Computing Language (OpenCL™ software) 1.2

Note that for using graphics APIs on Windows (such as OpenGL, Vulkan, DirectX 11, and DirectX 12) or any WDDM 2.0+ based functionality on data center GPUs, vGPU is required. See the vGPU documentation for more information.

Supported NVIDIA Data Center GPUs

The NVIDIA Data Center GPU driver package is designed for systems that have one or more Tesla products installed. This release of the driver supports CUDA C/C++ applications and libraries that rely on the CUDA C Runtime and/or CUDA Driver API.

NVIDIA Server Platforms	
Product	**Architecture**
NVIDIA HGX A800	A800 and NVSwitch
NVIDIA HGX A100	A100 and NVSwitch
NVIDIA HGX-2	V100 and NVSwitch

RTX-Series Products	
Product	**GPU Architecture**
Quadro RTX 8000	NVIDIA Turing
Quadro RTX 6000	NVIDIA Turing

A-Series Products	
Product	**GPU Architecture**
NVIDIA A800	NVIDIA Ampere GPU architecture
NVIDIA A100	NVIDIA Ampere GPU architecture

T-Series Products	
Product	**GPU Architecture**
NVIDIA T4	NVIDIA Turing

V-Series Products	
Product	**GPU Architecture**
NVIDIA V100	NVIDIA Volta

Tesla P-Series Products	
Product	**GPU Architecture**
NVIDIA Tesla P100	NVIDIA Pascal
NVIDIA Tesla P40	NVIDIA Pascal
NVIDIA Tesla P4	NVIDIA Pascal

Tesla K-Series Products	
Product	**GPU Architecture**
NVIDIA Tesla K520	NVIDIA Kepler
NVIDIA Tesla K80	NVIDIA Kepler

Tesla M-Class Products	
Product	**GPU Architecture**
NVIDIA Tesla M60	Maxwell
NVIDIA Tesla M40 24 GB	Maxwell
NVIDIA Tesla M40	Maxwell
NVIDIA Tesla M6	Maxwell
NVIDIA Tesla M4	Maxwell

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