NVIDIA Data Center GPU Driver version 460.32.03 (Linux) / 461.33 (Windows)

Release Notes

RN-08625-460 _v1.1 | February 2021

Table of Contents

Chapter 1. Version Highlights

This section provides highlights of the NVIDIA Data Center GPU R460 Driver (version 460.32.03 Linux and 461.33 Windows).

For changes related to the 460 release of the NVIDIA display driver, review the file "NVIDIA_Changelog" available in the .run installer packages.

- Windows driver release date: 02/08/2021
- Linux driver release date: 01/19/20201

1.1. Software Versions

For this release, the software versions are listed below.

- CUDA Toolkit 11: 11.2.0

  Note that starting with CUDA 11, individual components of the toolkit are versioned independently. For a full list of the individual versioned components (e.g. nvcc, CUDA libraries etc.), see the CUDA Toolkit Release Notes
- NVIDIA Data Center GPU Driver: 460.32.03 (Linux) / 461.33 (Windows)
- Fabric Manager: 460.32.03 (Use `nv-fabricmanager -v`)
- GPU VBIOS:
  - 92.00.19.00.01 (NVIDIA A100 SKU200 with heatsink for HGX A100 8-way and 4-way)
  - 92.00.19.00.02 (NVIDIA A100 SKU202 w/o heatsink for HGX A100 4-way)
- NVSwitch VBIOS: 92.10.14.00.01
- NVFlash: 5.641

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. New Features

General

- Added support for CUDA 11.2. For more information on CUDA 11.2, refer to the CUDA Toolkit 11.2 Release Notes
- Added support for NVIDIA A40.
- Added support for NVIDIA RTX A6000.

1.3. Fixed Issues

- Various security issues were addressed. For additional details on the med-high severity issues, review the NVIDIA Security Bulletin 5142 .

1.4. Known Issues

General

- 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
- On NVSwitch systems with Windows Server 2019 in shared NVSwitch virtualization mode, the host may hang or crash when a GPU is disabled in the guest VM. This issue is 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

Some Unified Memory APIs (for example, CPU page faults) are not supported on Windows in this version of the driver. Review the CUDA Programming Guide on the system requirements for Unified Memory

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 Windows and 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.

Supported Hypervisors

The following hypervisors are supported:

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

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

Supported Graphics Cards

The following GPUs are supported for device passthrough:

GPU Family	Boards Supported
NVIDIA Ampere GPU Architecture	NVIDIA A100, A40
Turing	NVIDIA T4
Volta	NVIDIA V100
Pascal	Tesla: P100, P40, P4

GPU Family	Boards Supported
Maxwell	Tesla: M60, M40, M6, M4
Kepler	Tesla: K520, K80

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 460 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 table below 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 <= 2)	Yes	No	No
Red Hat Enterprise Linux / CentOS 8.y (where y <= 3)	Yes	Yes	Yes
Red Hat Enterprise Linux / CentOS 7.y (where y <= 9)	Yes	No	No
SUSE Linux Enterprise Server 15.x (where y <= 2)	Yes	No	Yes (see note)
Ubuntu 20.04 LTS	Yes	No	No
Ubuntu 18.04.z LTS (where z <= 5)	Yes	Yes	Yes

Distribution	x86_64	POWER	Arm64 Server
Ubuntu 16.04.z LTS (where z <= 7)	Yes	No	No

Note that SUSE Linux Enterprise Server (SLES) 15.1 is provided as a preview for Arm64 server since there are known issues when running some CUDA applications related to dependencies on `glibc 2.27`.

Supported Operating Systems and CPU Configurations for HGX A100

The Release 460 driver is validated with HGX A100 on the following operating systems and CPU configurations:

- Linux 64-bit distributions:
  - Red Hat Enterprise Linux 8.3 (in 4/8/16-GPU configurations)
  - Red Hat Enterprise Linux 7.9 (in 4/8/16-GPU configurations)
  - CentOS Linux 8.3 (in 4/8/16-GPU configurations)
  - CentOS Linux 7.9 (in 4/8/16-GPU configurations)
  - Ubuntu 18.04.5 LTS (in 4/8/16-GPU configurations)
  - SUSE SLES 15.2 (in 4/8/16-GPU configurations)
- CPU Configurations:
  - AMD Rome in PCIe Gen4 mode
  - Intel Skylake/Cascade Lake (4-socket) in PCIe Gen3 mode

Supported Virtualization Configurations

The Release 460 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 18.04.4 LTS
- Shared NVSwitch (guest VMs only have visibility of GPUs and full NVLink bandwidth between GPUs in the same guest VM):
  - 16-GPU configurations with Ubuntu 18.04.4 LTS

API Support

This release supports the following APIs:

- NVIDIA® CUDA® 11.2 for NVIDIA® Kepler™, Maxwell™, Pascal™, Volta™, Turing™ and NVIDIA Ampere architecture GPUs
- OpenGL® 4.6
- Vulkan® 1.2

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

Note that for using graphics APIs on Windows (i.e. OpenGL, Vulkan, DirectX 11 and DirectX 12) or any WDDM 2.0+ based functionality on Tesla 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 A100	A100 and NVSwitch
NVIDIA HGX-2	V100 and NVSwitch

RTX-Series Products	
Product	**GPU Architecture**
NVIDIA RTX A6000	NVIDIA Ampere
Quadro RTX 8000	Turing
Quadro RTX 6000	Turing

A-Series Products	
Product	**GPU Architecture**
NVIDIA A100	NVIDIA Ampere
NVIDIA A40	NVIDIA Ampere

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

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

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

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

Notice

THE INFORMATION IN THIS GUIDE AND ALL OTHER INFORMATION CONTAINED IN NVIDIA DOCUMENTATION REFERENCED IN THIS GUIDE IS PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE INFORMATION FOR THE PRODUCT, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the product described in this guide shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

THE NVIDIA PRODUCT DESCRIBED IN THIS GUIDE IS NOT FAULT TOLERANT AND IS NOT DESIGNED, MANUFACTURED OR INTENDED FOR USE IN CONNECTION WITH THE DESIGN, CONSTRUCTION, MAINTENANCE, AND/OR OPERATION OF ANY SYSTEM WHERE THE USE OR A FAILURE OF SUCH SYSTEM COULD RESULT IN A SITUATION THAT THREATENS THE SAFETY OF HUMAN LIFE OR SEVERE PHYSICAL HARM OR PROPERTY DAMAGE (INCLUDING, FOR EXAMPLE, USE IN CONNECTION WITH ANY NUCLEAR, AVIONICS, LIFE SUPPORT OR OTHER LIFE CRITICAL APPLICATION). NVIDIA EXPRESSLY DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY OF FITNESS FOR SUCH HIGH RISK USES. NVIDIA SHALL NOT BE LIABLE TO CUSTOMER OR ANY THIRD PARTY, IN WHOLE OR IN PART, FOR ANY CLAIMS OR DAMAGES ARISING FROM SUCH HIGH RISK USES.

NVIDIA makes no representation or warranty that the product described in this guide will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do 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 guide. NVIDIA does not accept any 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 guide, or (ii) customer product designs.

Other than the right for customer to use the information in this guide with the product, no other license, either expressed or implied, is hereby granted by NVIDIA under this guide. Reproduction of information in this guide is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

Trademarks

NVIDIA and the NVIDIA logo are trademarks and/or registered trademarks of NVIDIA Corporation in the Unites States 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
www.nvidia.com