NVIDIA Data Center GPU Driver version 470.182.03 (Linux) / 474.30 (Windows)

Release Notes

RN-08625-470 _v8.0 | March 2023

Table of Contents

Chapter 1. Version Highlights

This section provides highlights of the NVIDIA Data Center GPU R470 Driver (version 470.182.03 Linux and 474.30 Windows).

For changes related to the 470 release of the NVIDIA display driver, review the file "NVIDIA_Changelog" available in the .run installer packages.

- Linux driver release date: 3/30/2023
- Windows driver release date: 3/30/2023

1.1. Software Versions

For this release, the software versions are listed below.

- CUDA Toolkit 11: 11.4

  Note that starting with CUDA 11, individual components of the toolkit are versioned independently. For a full list of the individual versioned components (for example, nvcc, CUDA libraries and so on), see the CUDA Toolkit Release Notes
- NVIDIA Data Center GPU Driver: 470.182.03 (Linux) / 474.30 (Windows)
- Fabric Manager: 470.182.03 (Use `nv-fabricmanager -v`)
- GPU VBIOS:
  - HGX A100 PG506
    - 92.00.45.00.03 SKU200 40GB air cooling (lidless)
    - 92.00.45.00.04 SKU202 40GB hybrid cooling (lidded)
    - 92.00.45.00.05 SKU210 80GB air cooling (lidless)
    - 92.00.45.00.06 SKU212 80GB hybrid cooling (lidded)
  - HGX A100 PG510
    - 92.00.81.00.01 SKU200 40GB air cooling (lidless)
    - 92.00.81.00.02 SKU202 40GB hybrid cooling (lidded)
    - 92.00.81.00.04 SKU210 80GB air cooling (lidless)
    - 92.00.81.00.05 SKU212 80GB hybrid cooling (lidded)

- HGX A800 PG506
  - 92.00.A4.00.01 SKU215 80GB air cooling (lidless)
- HGX A800 PG510
  - 92.00.A4.00.05 SKU215 80GB air cooling (lidless)
- A100 PCIe P1001 SKU230
  - 92.00.90.00.04 (NVIDIA A100 PCIe)
- A800 PCIe P1001
  - 92.00.A4.00.0C 40 GB SKU203 PCIe
  - 92.00.A4.00.0D 80 GB SKU235 PCIe
- NVSwitch VBIOS: 92.10.14.00.01
- 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, NVIDIA HGX A100), refer to the Fabric Manager User Guide.

1.2. Fixed Issues

- Resolved a kernel panic on A100 when using both MIG and DCGM. Pascal GPU page faults were hitting a NULL pointer dereference in the UVM driver while there was not enough system memory available to handle the faults.
- Resolved an issue that sometimes caused abnormal BAR1 memory usage.
- In the L1C submodule when the clock is gated, there is a corner case where the BLCG controller was not woken up from sleep state when an external submodule wants to use L1C. This was fixed by Switching the PROD value to disable L1C BLCG will not cause the hang in the chip until some other event wakes up the BLCG FSM.
- `nvidia-smi -q` returns strange values on A100 and A16 GPUs.
- Fixed an issue where the NVIDIA Linux GPU kernel driver was calling the Linux kernel scheduler while holding a lock with preemption disabled during event notification.
- In the code path related to allocating virtual address space, a call to reallocate memory for tracking structures was allocating less memory than needed, resulting in a potential memory trampler. The size of the reallocation is now correctly calculated.
- Fixed the Race condition between cudaFreeAsync() and cudaDeviceSynchronize() which were being hit if device sync is used instead of stream sync in multi threaded app. Now a Lock is being held for the appropriate duration so that a subpool cannot be modified during a very small window which triggers an assert as the subpool should have been empty due to the free code.

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.
- On HGX A800 8-GPU systems, the nvswitch-audit tool will report 12 NVLinks per GPU. This is a switch configuration report and does not reflect the true number of NVLink interfaces available per-GPU, which remains 8.
- 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.
- 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).
- 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.

- If you encounter an error on RHEL7 when installing with `cuda-drivers-fabricmanager` packages, use the following alternate instructions. For example:

  If you are upgrading from a different branch, for example to driver 470.141.03:

```
new_version=470.141.03
sudo yum swap nvidia-driver-latest-dkms nvidia-driver-latest-dkms-${new_version}
sudo yum install nvidia-fabric-manager-${new_version}
```

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.

OpenCL 3.0 Known Issues

- Device-Side-Enqueue related queries may return 0 values, although corresponding built-ins can be safely used by kernel. This is in accordance with conformance requirements described at https://www.khronos.org/registry/OpenCL/specs/3.0-unified/html/OpenCL_API.html#opencl-3.0-backwardscompatibility
- 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.

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, A40, A30, A16, A10
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 470 driver is supported on the following operating systems:

- Windows x86_64 operating systems:
  - Microsoft Windows® Server 2022
  - Microsoft Windows® Server 2019
  - Microsoft Windows® Server 2016
  - Microsoft Windows® 11
  - 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
Rocky Linux 8.y (where y <= 7)	Yes	No	No
Red Hat Enterprise Linux / CentOS 7.y (where y <= 9)	Yes	No	No
SUSE Linux Enterprise Server 15.y (where y <= 4)	Yes	No	Yes

Distribution	x86_64	POWER	Arm64 Server
Ubuntu 20.04.z LTS (where z <= 5)	Yes	No	Yes
Ubuntu 18.04.z LTS (where z <= 6)	Yes	No	No

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

The Release 470 driver is validated with NVIDIA 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)
  - Red Hat Enterprise Linux 7.9 (in 4/8/16-GPU configurations)
  - Rocky 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 470 driver is validated with NVIDIA HGX A100 and HGX A800 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
  - 1/2/4/8-GPU configurations with Windows x86_64 operating systems:
    - Microsoft Windows® Server 2019

- Microsoft Windows® 10

API Support

This release supports the following APIs:

- NVIDIA® CUDA® 11.4 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) 3.0

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 Data Center GPU 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.

Attention: Release 470 will be the last driver branch to support Data Center GPUs based on the NVIDIA Kepler architecture. This includes discontinued support for the following compute capabilities:

- sm_30 (NVIDIA Kepler)
- sm_32 (NVIDIA Kepler)
- sm_35 (NVIDIA Kepler)
- sm_37 (NVIDIA Kepler)

For more information on GPU products and compute capability, see https://developer.nvidia.com/cuda-gpus.

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 / T-Series Products	
Product	**GPU Architecture**
NVIDIA RTX A6000	NVIDIA Ampere architecture
NVIDIA RTX A5000	NVIDIA Ampere architecture
NVIDIA RTX A4000	NVIDIA Ampere architecture
Quadro RTX 8000	NVIDIA Turing
Quadro RTX 6000	NVIDIA Turing
NVIDIA T1000	NVIDIA Turing
NVIDIA T600	NVIDIA Turing
NVIDIA T400	NVIDIA Turing

Data Center A-Series Products	
Product	**GPU Architecture**
NVIDIA A800	NVIDIA Ampere architecture
NVIDIA A100 NVIDIA A100 80 GB PCIe	NVIDIA Ampere architecture
NVIDIA A40	NVIDIA Ampere architecture
NVIDIA A30	NVIDIA Ampere architecture
NVIDA A16	NVIDIA Ampere architecture
NVIDIA A10	NVIDIA Ampere architecture

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

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

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

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

Data Center 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

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.

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
https://www.nvidia.com