NVIDIA Data Center GPU Driver version 550.90.07 (Linux)/ 552.55 (Windows)

Release Notes

RN-08625-550 _v3.0 | June 2024

Table of Contents

Chapter 1. Version Highlights

This section provides highlights of the NVIDIA Data Center GPU R550 Driver (version 550.90.07 Linux and 552.55 Windows).

For changes related to the 550 release of the NVIDIA display driver, review the file "NVIDIA_Changelog" available in the .run installer packages.

- Linux driver release date: 06/06/2024
- Windows driver release date: 06/06/2024

1.1. Software Versions

For this release, the software versions are as follows:

- CUDA Toolkit 12: 12.4

  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: 550.90.07 (Linux) / 552.55 (Windows)
- Fabric Manager: 550.90.07 (Use `nv-fabricmanager -v`)
- NVFlash: 5.791

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

- On Intel Emerald Rapids chipset ID 0x1b81, PCIe P2P sometimes did not work even though the GPU driver reports P2P as supported in `nvidia-smi topo -p2p rw` command. This issue is resolved. 4448950
- Running nvidia-smi caused a kernel panic, which was due to Null pointer dereferencing in the code path. This was fixed by making it a valid pointer. 4644086
- The crashcat buffer was not large enough to contain all crash information, resulting in truncated reports. The buffer size has been increased. 4607635

- The gRPC event poller is now proactively triggered by invoking `waitForConnected()` on the channel with an empirically shorter deadline, which is sufficient to get the channel to READY state. 4605545
- Fixed a null pointer dereference caused by an unchecked return status from the API that allocates the channel notifier action table. 4603505
- RC errors were observed during the channel allocation path, leading to accessing the notifier action table which has not been allocated yet. This issue is resolved. 4589964
- Invalid objects are no longer accessed when a process is killed while profiling with MIG enabled. 4584716
- On some versions of the GPU driver, the `nvidia-smi topo -m` command showed GPU connections as SYS, implying that they are connected across the SMP interconnect between NUMA nodes when they are actually part of the same NUMA node. This issue has been resolved and these connections now correctly show NODE instead of SYS. 4576789
- Resolved an issue that caused driver compilation warnings whenever a `.run` package was installed on systems with kernels built with Clang or newer GCC. 4569283
- Fixed a bug in the compiler's register allocator, where it was previously unable to correctly spill/refill and thereby allocate register for the barrier register class for test cases involving loops with breaks. 4546457
- XID error messages for DRAM and LTCs are no longer limited to 1 per boot. XID error messages for DBEs are now printed when error containment is enabled. 4478617
- The GPU Max Operating Temp now correctly displays the Slowdown temperature. 4507899
- Changed row remapping histogram to only account for uncorrectable row remappings. 4451237
- NVENC session data processing tasks are now scheduled into work-items so processing time won't be propagated up. NVENC NVENC session data are processed as CPU's work-items when no high priority tasks occupy the CPU. 4437306
- Improved the ability of `nvidia-modprobe` to detect whether kernel modules are already loaded. This corrects an issue that prevented `nvidia-persistenced` from setting persistence mode on some systems. 4435513
- The drain and reset flag is now set when uncorrectable row remapping occurs. 4432594
- Fixed an issue with particular types of host memory allocations, for example CMA, where registering that memory buffer on the GPU (via `cuMemHostRegister`) first would later prevent to register the same memory buffer on the RDMA NIC (via `ibv_reg_mr`). 4429264
- Fixed the regression issue where the new hwloc version 2.9.2 (integrated by NVML library) failed to detect two GPUs under a common NUMA node, resulting in the `nvidia-smi topo -m` command showing SYS instead of NODE. 4412560
- When running MPS where the client and server are in different containers, it is possible that they may have access to different versions of the glibc library. There were observed hangs in some combinations of this mismatch due to differing ABI for the glibc semaphore implementation. This version includes a fix for this issue by

making the synchronization used by MPS agnostic to the GLIBC implementation. 4382578

- A clear signal is needed on OOB when the GPU exceeds the SRAM UCE > 4 criteria for RMA. The SRAM Threshold Exceeded flag from nvidia-smi is now reported through OOB. 4315988
- Improved error containment by forwarding poisoned data to target engine instead of handling at GPU granularity. 4060421

1.3. Known Issues

- 4687401 - Some packages may install a different Linux kernel flavor when upgrading the Open-source driver from older branches to 550 and newer.

  To work around this, end-users on openSUSE and SLES should upgrade their NVIDIA Open GPU kernel modules using the following:

  ```
  sudo zypper -v install $(zypper search -s nvidia-open-driver-G06-kmp-<flavor> | sed 's| ||g' | awk -F '|' '/<branch>/ {print $2"="$4}')
  ```

  ```
  sudo zypper -v install cuda-drivers-<branch>
  ```

  For example:

  ```
  sudo zypper -v install $(zypper search -s nvidia-open-driver-G06-kmp-default | sed 's| ||g' | awk -F '|' '/550/ {print $2"="$4}')
  ```

  ```
  sudo zypper -v install cuda-drivers-550
  ```

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 2019 Hyper-V Generation 2

Data Center products now support one display of up to 2560x1600 resolution.

The following GPUs are supported for device passthrough for virtualization:

GPU Family	Boards Supported
NVIDIA Ada Lovelace	NVIDIA L40, L4
NVIDIA Grace Hopper	NVIDIA GH200
NVIDIA Hopper	NVIDIA H100, NVIDIA H800
NVIDIA Ampere GPU Architecture	NVIDIA A800, A100, A40, A30, A16, A10, A10G, A2, AX800
NVIDIA Turing	NVIDIA T4, NVIDIA T4G
NVIDIA Volta	NVIDIA V100
NVIDIA Pascal	Quadro: P2000, P4000, P5000, P6000, GP100

GPU Family	Boards Supported
NVIDIA Maxwell	Tesla: P100, P40, P4 Quadro: K2200, M2000, M4000, M5000, M6000, M6000 24GB Tesla: M60, M40, M6, M4

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 550 driver is supported on the following operating systems:

- Windows x86_64 operating systems:
  - Microsoft Windows® Server 2022
  - Microsoft Windows® 11 21H2 - SV1
  - Microsoft Windows® 11 22H2 - SV2
  - Microsoft Windows® 11 23H2
  - Microsoft Windows® 10 21H2
  - Microsoft Windows® 10 22H2
- 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	Arm64 Server
Debian 12.x (where x <= 5)	Yes	No
Debian 11.x (where x <= 9)	Yes	No
Debian 10. x (where x <= 13)	Yes	No
OpenSUSE Leap 15.x (where y <= 5)	Yes	No
Fedora 39	Yes	No
Red Hat Enterprise Linux 9.y (where y <= 4)	Yes	Yes
Rocky Linux 9.y (where y <= 4)	Yes	No

Distribution	x86_64	Arm64 Server
Red Hat Enterprise Linux 8.y (where y <= 9)	Yes	Yes
Rocky Linux 8.y (where y <= 9)	Yes	No
Red Hat Enterprise Linux / CentOS 7.y (where y <= 9)	Yes	No
SUSE Linux Enterprise Server 15.y (where y <= 5)	Yes	Yes
Ubuntu 24.04 LTS	Yes	Yes
Ubuntu 22.04.z LTS (where z <= 4)	Yes	Yes
Ubuntu 20.04.z LTS (where z <= 6)	Yes	Yes
KylinOS V10 SP2	Yes	No
CBL-Mariner 2.0*	Yes	No

* CBL-Mariner will be supported by TRD via runfile. CUDA Toolkit will not support this OS as this is a deployment OS.

Supported Operating Systems and CPU Configurations for NVIDIA HGX H20

- Hopper Linux distributions:
  - Red Hat Enterprise Linux 9.4
  - Ubuntu 22.04 with NVIDIA HWE kernel
- Windows 64-bit distributions:
  - Windows Server 2022

Supported Operating Systems and CPU Configurations for NVIDIA HGX GH200

- Grace Hopper Linux distributions:
  - Red Hat Enterprise Linux 9.4
  - SUSE Linux Enterprise Server 15 SP5 QU1
  - Ubuntu 22.04 with NVIDIA HWE kernel

    RHEL and SLES feature parity with NVIDIA HWE Kernels. The latest RHEL 9 and SLES 15 SP5 kernels support bare metal.

Supported Operating Systems and CPU Configurations for NVIDIA HGX H200

The Release 550 driver is validated with NVIDIA HGX H200 on the following operating systems and CPU configurations:

- Linux 64-bit distributions:
  - Red Hat Enterprise Linux 9.4 (in 4/8/16-GPU configurations)
  - Ubuntu 22.04.2 LTS (in 4/8/16-GPU configurations)
- Windows 64-bit distributions:
  - Windows Server 2022
  - Windows is supported only in shared NVSwitch virtualization configurations.

Supported Operating Systems and CPU Configurations for NVIDIA HGX H100/H800

The Release 550 driver is validated with NVIDIA HGX H100 on the following operating systems and CPU configurations:

- Linux 64-bit distributions:
  - Red Hat Enterprise Linux 8.9 (in 4/8/16-GPU configurations)
  - Red Hat Enterprise Linux 9.4 (in 4/8/16-GPU configurations)
  - SUSE Linux Enterprise Server 15.5 (in 4/8/16-GPU configurations)
  - Ubuntu 24.04 LTS (in 4/8/16-GPU configurations)
  - Ubuntu 22.04.2 LTS (in 4/8/16-GPU configurations)
- Windows 64-bit distributions:
  - Windows Server 2022
  - Windows is supported only in shared NVSwitch virtualization configurations.

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

The Release 550 driver is validated with NVIDIA HGX A100 on the following operating systems and CPU configurations:

- Linux 64-bit distributions:
  - Debian 12.5
  - Debian 11.9
  - Debian 10.13
  - Red Hat Enterprise Linux 8.9 (in 4/8/16-GPU configurations)

- Red Hat Enterprise Linux 7.9 (in 4/8/16-GPU configurations)
- Rocky Linux 8.9 (in 4/8/16-GPU configurations)
- Red Hat Enterprise Linux 9.4 (in 4/8/16-GPU configurations)
- CentOS Linux 7.9 (in 4/8/16-GPU configurations)
- Ubuntu 24.04 LTS (in 4/8/16-GPU configurations)
- Ubuntu 22.04.4 LTS (in 4/8/16-GPU configurations)
- Ubuntu 20.04.6 LTS (in 4/8/16-GPU configurations)
- SUSE SLES 15.5 (in 4/8/16-GPU configurations)
- KylinOS V10 SP2

- Windows 64-bit distributions:
  - Windows Server 2022
  - 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 550 driver is validated with NVIDIA HGX A100, HGX A800, H100, and H800 on the following configurations:

- Passthrough (full visibility of GPUs and NVSwitches to guest VMs):
  - 8-GPU configurations with Ubuntu 20.04.6 and 22.4.2
- 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 20.04.6 LTS

API Support

This release supports the following APIs:

- NVIDIA® CUDA® 12.4 for NVIDIA® Maxwell™, Pascal™, Volta™, Turing™, Hopper™, NVIDIA Ampere architecture, and NVIDIA Ada Lovelace GPU architecture GPUs
- OpenGL® 4.6
- Vulkan® 1.3
- 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 was 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 H20 8-GPU	H20 and NVSwitch
NVIDIA HGX H100 8-GPU	H100 and NVSwitch
NVIDIA HGX H800 8-GPU	H800 and NVSwitch
NVIDIA HGX H100 4-GPU	H100 and NVLink
NVIDIA HGX A800 8-GPU	A800 and NVSwitch
NVIDIA HGX A100 8-GPU	A100 and NVSwitch
NVIDIA HGX A100 4-GPU	A100 and NVLink
NVIDIA HGX-2	V100 and NVSwitch

Data Center L-Series Products	
Product	**GPU Architecture**
NVIDIA L2	NVIDIA Ada Lovelace
NVIDIA L20	NVIDIA Ada Lovelace
NVIDIA L40	NVIDIA Ada Lovelace
NVIDIA L40S	NVIDIA Ada Lovelace
NVIDIA L4	NVIDIA Ada Lovelace

Data Center H-Series Products	
Product	**GPU Architecture**
NVIDIA H100 PCIe	NVIDIA Hopper
NVIDIA H100 NVL	NVIDIA Hopper
NVIDIA H800 PCIe	NVIDIA Hopper
NVIDIA H800 NVL	NVIDIA Hopper

RTX-Series / T-Series Products	
Product	**GPU Architecture**
NVIDIA RTX 6000 Ada Generation	NVIDIA Ada Lovelace
NVIDIA RTX 4000 SFF Ada Generation	NVIDIA Ada Lovelace
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
Quadro RTX 4000	NVIDIA Turing
NVIDIA T1000	NVIDIA Turing
NVIDIA T600	NVIDIA Turing
NVIDIA T400	NVIDIA Turing

Data Center A-Series Products	
Product	**GPU Architecture**
NVIDIA A2	NVIDIA Ampere architecture
NVIDIA A800, AX800	NVIDIA Ampere architecture
NVIDIA A100X	NVIDIA Ampere architecture
NVIDIA A100 NVIDIA A100 80 GB PCIe	NVIDIA Ampere architecture
NVIDIA A40	NVIDIA Ampere architecture
NVIDIA A30. A30X	NVIDIA Ampere architecture
NVIDIA A16	NVIDIA Ampere architecture
NVIDIA A10, A10M, A10G	NVIDIA Ampere architecture

Data Center T-Series Products	
Product	**GPU Architecture**
NVIDIA T4, T4G	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 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