NVIDIA RTX Enterprise Release 550 Drivers, Version 552.22

For NVDIA RTX, Quadro, NVS, and Data Center GPUs
Windows Server 2022

Release Notes

RN-07956-552.22_v01 | April 16, 2024

Table of Contents

List of Tables

Chapter 1. Introduction to Release Notes

This edition of **Release Notes** describes the Release 550 family of Quadro, NVS, and Data Center drivers for Microsoft® Windows® Server 2022. NVIDIA provides these notes to describe performance improvements and bug fixes in each documented version of the driver.

1.1 Structure of the Document

This document is organized in the following sections:

- Release 550 Driver Changes gives a summary of changes, and fixed and open issues in this version.
- The Release 550 Driver describes the NVIDIA products and languages supported by this driver, the system requirements, and how to install the driver.
- NVIDIA Tesla Compute Cluster Mode describes the Tesla Compute Cluster mode.

1.2 Changes in this Edition

This edition of the Release Notes for Windows Server includes information about NVIDIA graphics driver version 552.22, and lists changes made to the driver since version 551.86.

These changes are discussed beginning with the chapter Release 550 Driver Changes.

Chapter 2. Release 550 Driver Changes

This chapter describes open issues for version 552.22, and resolved issues and driver enhancements for versions of the Release 550 driver up to version 552.22.

The chapter contains these sections:

2.1 Version 552.22 Highlights

This section provides highlights of version 552.22 of the NVIDIA Release 550 Driver for Windows Server 2019/2022.

2.1.1 Existing Support

This release supports the following APIs:

1. OpenCL™ software 3.0
2. OpenGL® 4.6
3. Vulkan® 1.3
4. DirectX 11
5. DirectX 12
6. NVIDIA® CUDA® 12.4

This driver installs NVIDIA RTX Desktop Manager version 204.26.

2.1.2 What's New in Version 552.22

- Refer to What's New in Release 550 for the list of new features introduced since Release 535.
- Support for the following new NVIDIA RTX GPU products:
  - NVIDIA RTX A1000

- NVIDIA RTX A400
- NVIDIA RTX 2000E Ada Generation

> Version 552.22 incorporates the latest bug fixes and driver component enhancements to improve performance.

2.2 Changes in Version 552.22

The following sections list the important changes and the most common issues resolved in this driver version.

2.2.1 Fixed Issues in Version 552.22

> [HTC Vive Pro 2]: System crashes after enabling VR HMD with multi-displays.
> [Houdini]: OptiX applications may crash when SER is combined with user exceptions.
> [Maya]: Maya incorrectly sets a single-thread flag and crashes.

2.2.2 Open Issues in Version 552.22

> [Mosaic]: The dropbox for Refresh rate and resolution in the Mosaic setup of the NVIDIA Control Panel is blank and displays "not supported," preventing users from selecting the desired settings.
> [Mosaic]: SnippingTool.exe application blocks SLI/Mosaic transition.
> [Mosaic]: While enabling Mosaic, the "Horizontal" item occasionally disappears.

2.2.3 What's New in Release 550

The section summarizes the following driver changes in Release 550 (since Release 535):

2.2.3.1 NVIDIA RTX Production Branch Driver

Release 550 is the latest Production Branch release of the NVIDIA RTX Enterprise Driver. Production Branch drivers are recommended for enterprise deployment, certification with professional applications, and users seeking the latest support for NVIDIA Studio features.

For the most stable and fully supported enterprise driver, please use the Production Branch graphics driver downloadable from the NVIDIA driver download page.

2.2.3.2 New Features

> **TensorRT-LLM** — open-source library added that accelerates and optimizes inference performance of latest LLMs on NVIDIA GPUs.
> **Video TrueHDR** — uses AI to enhance SDR to HDR tone mapping with greater color range and brightness levels.
  - Requires Chrome or Edge browser.

- **Bit Depth Expansion** — adds quality improvement and enhanced coding efficiency to video codecs.
- **Execute Indirect Extension NVAPI** — offloads work from the CPU to the GPU.
  - Access provided under NDA.
- **CBL2 Support in OptiX** — improves perfomance by reducing CPU overhead.
- **Video Super Resolution and TrueHDR Support Added to NGX SDK**— support added for app integration of Video Super Resolution and TrueHDR.
  - VSR improves video quality of low-resolution sources.
  - TrueHDR smartly converts SDR content into HDR content.
- **Video Super Resolution support** — adds support for Turing GPU architecture and later.
- **AV1 Vulkan Video support** — support added for Vulkan AV1 decode acceleration through `VK_KHR_video_decode_av1` extension.

2.2.3.3 NVIDIA OpenCL Vulkan Interop

The NVIDIA OpenCL driver has added support for the following new provisional extension specifications released by Khronos. The specifications are for OpenCL external semaphore and external memory.

1. https://www.khronos.org/registry/OpenCL/specs/3.0-unified/html/OpenCL_Ext.html#cl_khr_semaphore
2. https://www.khronos.org/registry/OpenCL/specs/3.0-unified/html/OpenCL_Ext.html#cl_khr_external_semaphore
3. https://www.khronos.org/registry/OpenCL/specs/3.0-unified/html/OpenCL_Ext.html#cl_khr_external_memory

NVIDIA is seeking developer feedback on this new extension support.

2.2.3.4 OpenCL External semaphore and memory extensions

The set of new External Memory and Semaphore Sharing extensions provides a generic framework that enables OpenCL to import external memory and semaphore handles exported by external APIs—using a methodology that will be familiar to Vulkan developers—and then use those semaphores to synchronize with the external runtime, coordinating the use of shared memory.

The following key features are supported as part of these extensions:

1. Importing memory into buffers using FD, Win32 KMT and NT handles
2. Importing memory into images using FD, Win32 KMT and NT handles
3. Importing binary semaphores using FD, Win32 KMT and NT handles
4. Synchronizing using Wait and Signal on imported semaphores
5. Using buffers and images imported in OpenCL kernels and other APIs such as other regular cl_mem.

2.2.3.5 Limitations of the Current Implementation

1. Support for importing external memory and semaphores using FD handles on Linux and Win32 NT and KMT handles on Windows. No other handle types are currently available.
2. Support for binary semaphores only.
3. No support for exporting semaphore or memory from OpenCL.
4. `clEnqueueAcquireExternalMemObjectsKHR` and `clEnqueueReleaseExternalMemObjectsKHR` APIs are currently not required as execution hand-off can be managed through semaphore wait and signal. But these may be required in the future for correct functionality.

2.2.4 NVIDIA OpenCL Compiler Upgrade

The NVIDIA OpenCL driver used an older OpenCL Just-In-Time (JIT) compiler based on the legacy 3.4 versions of the Clang front-end and NVVM optimizer. NVIDIA has been working on upgrading its OpenCL JIT compiler to use a newer version 7.0 of the Clang front-end and NVVM optimizer component.

NVIDIA introduced this new OpenCL compiler as an opt-in feature in 510 driver release (511.09 on Windows and 510.54 on linux), with the default OpenCL compiler remaining the same. With 520.xx (exact version TBD), NVIDIA OpenCL will use the new Clang + NVVM 7.0 based compiler as the default compiler, replacing the old compiler.

As part of this driver installation, a new compiler library should be visible in the system folder as `libnvidia-nvvm.so/nvvm*.dll` instead of the old `libnvidia-compiler.so*/nvcompiler*.dll`. The driver will pick up the newer compiler library by default.

2.2.4.1 NVVM 7.0 New Compiler Features

The new NVVM 7.0 based compiler takes advantage of years of development in the Clang+LLVM framework. In addition to several minor bug fixes and diagnostic improvements, this compiler introduces the following noteworthy features:

- **16-bit floating point (half) type**

  16-bit floating point types or “half” type is available as a native data type in the new compiler. This type is enabled by the `cl_khr_fp16` feature guard pragma.

  Example:

```
#pragma OPENCL EXTENSION cl_khr_fp16 : enable

half scalar_arith(half n, half k) {
   half w = n + k;
   half x = n - k;
   half y = w * x;
   half z = y / x;
   return -z;
}
 kernel void foo( global int* x) {
   half a = 3.5H, b = 4.5H;
   if (scalar_arith(a, b) == -8.0)
```

```
        *x = 1;
    return;
}
```

Note: While enabling `cl_khr_fp16` pragma/feature macro allows some basic usage of half float data types including basic math operations (add, sub, mul, div) on half floats with the newer compiler, math built-in functions for half floats are currently not supported. Using the same may lead to `clBuildProgram` failing. NVIDIA OpenCL 3.0 drivers do not claim conformance for `cl_khr_fp16`. The device and platform queries (`CL_DEVICE_EXTENSIONS` / `CL_DEVICE_EXTENSIONS_WITH_VERSION` / `CL_PLATFORM_EXTENSIONS` / `CL_PLATFORM_EXTENSIONS_WITH_VERSION`) do not report `cl_khr_fp16` as one of the supported extensions on NVIDIA OpenCL 3.0 drivers even with the compiler upgrade. Please consider this as an experimental feature without any functional, conformance or performance guarantees.

- **128-bit integer type**

  128-bit integer types or “(un)signed long long” is available as a native data type in the new compiler. This type is enabled by default and does not require any macros to be defined.

  Example:

```
typedef unsigned long long ULL;
typedef long long LL;

LL scalar_arith(ULL n, ULL k)  {
    LL w = n + k;
    LL x = n - k;
    LL y = w * x;
    LL z = y / x;
    return -z;
}

__kernel void foo( global int* x)
    ULL a = 0x123456789ABCDEF0ULL;
    ULL b = 0xFEDCBA9876543210ULL;
    if (scalar_arith(a, b) < 0)
        *x = 1;
    return;
}
```

- **Upgraded math libraries**

  The built-in standard math functions (e.g. sin(), cos()) have been upgraded to be on par with CUDA C++. This ensures that your application benefits from high-performance math routines optimized for the latest GPU architectures.

2.2.4.2 Clang Release Notes

The public LLVM release notes for Clang 3.4 – Clang 7.0 mentioned below summarize the behavioral changes between old and new compiler bases.

- Clang 4: https://releases.llvm.org/4.0.0/tools/clang/docs/ReleaseNotes.html
- Clang 5: https://releases.llvm.org/5.0.0/tools/clang/docs/ReleaseNotes.html

- Clang 6: https://releases.llvm.org/6.0.0/tools/clang/docs/ReleaseNotes.html
- Clang 7: https://releases.llvm.org/7.0.0/tools/clang/docs/ReleaseNotes.html

2.2.4.3 Known Issues with NVVM 7.0-based Compiler

The new Clang/NVVM 7.0 based compiler has stricter error checking compared to the previous compiler. The following use-cases which were allowed with the older compiler may now throw an error.

1. Updating `const` variables after they have been assigned.
2. Using address spaces other than `global` for kernel pointer parameters.
3. Using `variadic` arguments in functions and blocks.

2.2.5 OpenCL 3.0

- Added support for OpenCL 3.0[1], the latest major version of OpenCL maintaining backward compatibility with OpenCL 1.2. NVIDIA OpenCL 3.0 continues to support existing OpenCL 1.2 functionality as well as Khronos and vendor extensions that are already supported with NVIDIA OpenCL 1.2 drivers.

  The following new features beyond existing NVIDIA OpenCL 1.2 features are supported by NVIDIA OpenCL 3.0

  - RGBA vector component naming in OpenCL C kernels
  - pragma_unroll hint
  - opencl_3d_image_writes
  - clCreate*WithProperties APIs which can be used as replacement for existing clCreateBuffer/Image APIs.
  - clSetContextDestructorCallback
  - clCloneKernel from OpenCL 2.1
  - clEnqueueSVMMigrateMem from OpenCL 2.1
- Incorporates the following experimental 2.0 features:
  a. Device side enqueue
     i. The current implementation is limited to 64-bit platforms only.
     ii. 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 https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#compute-capabilities.

     For a given grid dimension, the global_work_size can be determined by CUDA grid size x CUDA block size.

iii. For executing kernels (whether from the host or the device), OpenCL 3.0 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.

1. Khronos has recently released OpenCL 3.0 spec (https://www.khronos.org/registry/OpenCL/specs/3.0-unified/pdf/ OpenCL_API.pdf)

Note: Other OpenCL 2.X entry-points which are now optional and are not supported in NVIDIA OpenCL 3.0 will behave as described at https://www.khronos.org/registry/OpenCL/specs/3.0-unified/html/OpenCL_API.html#opencl-3.0-backwards-compatibility

2.3 Discontinued and Unsupported Features in This Release

2.3.1 End of Support for Windows Server 2019 Operating System

The NVIDIA RTX Driver Release Branch 535 (R535) was the last NVIDIA professional driver to support the Microsoft Windows Server 2019 operating system.

2.3.2 Discontinued Support for Kepler GPUs

Release 470 was the last driver branch to support desktop GPUs from the Kepler architecture generation.

2.3.3 Limitations in This Release

The following features are not currently supported or have limited support in this driver release:

2.3.3.1 External Graphics

- External GPU Surprise Removal

  Not all applications have been designed to address surprise removal of the external GPU; disconnection of the external GPU while applications are running is not advised.
- Mixed GeForce/Quadro Products

  Mixed GeForce/Quadro products are supported (Geforce GPU + Quadro eGPU, or Quadro GPU + Geforce eGPU), but requires installation of the GeForce driver package. The Quadro package does not install GeForce drivers.

2.3.3.2 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-backwards- compatibility.
- Shared virtual memory - the current implementation of shared virtual memory is limited to 64-bit platforms only.

Chapter 3. The Release 550 Driver

This chapter covers the following main topics:

- Driver Security
- Advanced Driver Information
- Known Product Limitations
- Hardware and Software Support
- Driver Installation

3.1 Driver Security

Follow these safe computing practices:

- Only download or execute content and programs from trusted third parties.
- Run your system and programs with the least privilege necessary. Users should run without administrator rights whenever possible.
- When running as administrator, do not elevate UAC privileges for activities or programs that don't need them.

This section describes additional actions to take to mitigate specific known security issues.

3.1.1 Restricting/Enabling Access to GPU Performance Counters

The NVIDIA graphics driver contains a vulnerability (CVE-2018-6260) that may allow access to application data processed on the GPU through a side channel exposed by the GPU performance counters. GPU performance counters are needed by developers to use NVIDIA developer tools such as CUPTI, Nsight Graphics, and Nsight Compute. To address CVE-2018-6260 the driver (starting with version 419.67) automatically disables access for non-admin users. For more information about CVE-2018-6260 visit the NVIDIA Security Bulletin 4772.

3.1.2 Enabling Access to GPU Performance Counters Using the NVIDIA Control Panel

Access to GPU performance counters can be enabled for non-admin users who need to use NVIDIA developer tools. Enabling access to GPU performance counters can be

accomplished through the NVIDIA Control Panel -> **Developer** -> **Manage GPU Performance Counters** page. Refer to the **Developer** -> **Manage GPU Performance Counters** section of the NVIDIA Control Panel Help for instructions.

Note: Access to GPU performance counters should be kept disabled for non-admin users who do not need to use NVIDIA developer tools.

3.1.3 Restricting/Enabling Access to GPU Performance Counters Across an Enterprise Using Scripts

Enterprise administrators can use scripts to programmatically apply the settings. The scripts should incorporate the registry key information provided below to automate the deployment.

Caution: These instructions should be performed only be enterprise administrators. Changes to the registry must be made with care. System instability can result if performed incorrectly.

```
[HKLM\SYSTEM\CurrentControlSet\Services\nvlddmkm\Global\NVTweak]
Value: "RmProfilingAdminOnly"
Type: DWORD
Data: 00000001
```

The data value 1 restricts access to admin users, whereas data value 0 allows access to all users.

A system reboot is required for the changes to take effect.

3.2 Advanced Driver Information

This section clarifies instructions for successfully accomplishing the following tasks:

- Turning Off V-Sync to Boost Performance
- NVIDIA Application Configuration Engine (ACE)
- SLI Multi-OS – GPU Assignment in System Virtualization
- Using the WDDM Driver Model with Tesla GPU GOMs

3.2.1 Turning Off V-Sync to Boost Performance

To get the best benchmark and application performance measurements, turn V-Sync off as follows:

1. Open the NVIDIA Control Panel and make sure that *Advanced Settings* is selected from the control panel tool bar.

2. From the *Select a Task* pane, under 3D Settings, click ***Manage 3D Settings***, then click the ***Global Settings*** tab.
3. From the Global presets pull-down menu, select ***Base profile***.
4. From the Settings list box, select ***Vertical sync*** and change its value to ***Force off***, then click ***Apply***.
5. From the Global presets pull-down menu, select ***3D App - Default Global Settings*** (the driver's default profile) or use the application profile that matches the application you are testing, then click ***Apply***.

> **Important:** Be sure to close the NVIDIA Control Panel completely. Leaving it open will affect benchmark and application performance.

3.2.2 NVIDIA Application Configuration Engine (ACE)

This driver includes the NVIDIA Application Configuration Engine (ACE), which automatically detects the workstation application and configures the appropriate profile settings in the NVIDIA Control Panel.

3.2.3 SLI Multi-OS – GPU Assignment in System Virtualization

On systems with two or more graphics cards installed, this driver supports a hypervisor's ability to directly assign GPUs to guest virtual machines (VMs). This direct assignment allows each guest VM to run on their own operating system with their own GPU and driver. The assignment allows full GPU performance and functionality in the guest VM.

3.2.3.1 Hardware Platform Requirements

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.

3.2.3.2 Supported Hypervisors

Please reference the Virtual GPU Software Supported Products page for the latest supported configurations.

3.2.3.3 Supported Graphics Cards

The following GPUs are supported for device passthrough:

GPU Family	Boards supported
NVIDIA Ada Lovelace	NVIDIA Data Center: L40, L20, L2 NVIDIA RTX: 6000 Ada Generation, 5880 Ada Generation, 5000 Ada Generation, 4500 Ada Generation, 4000 Ada Generation, 4000 SFF Ada Generation 2000 Ada Generation
NVIDIA Ampere	NVIDIA Data Center: AX800, A100, A40, A30, A16, A10, A2 NVIDIA RTX: A6000, A5500, A5000, A4500, A4000H, A4000, A2000 12GB, A2000, A1000, A400
Turing	Quadro: RTX 8000, RTX 6000, RTX 5000, RTX 4000 NVIDIA Data Center: T4
Volta	Quadro: GV100 NVIDIA Data Center: V100
Pascal	Quadro: P2200, P2000, P4000, P5000, P6000, GP100 Tesla: P100, P40, P4
Maxwell	Quadro: K2200, M2000, M4000, M5000, M6000, M6000 24GB Tesla: M60, M10, M6

3.2.4 Using the WDDM Driver Model with Tesla GPU GOMs

3.2.4.1 Tesla GPU Operation Modes

The ability to specify the GPU operation mode using NV-SMI/NVML, previously offered by legacy Tesla GPU Accelerators, is still available (refer to: https://developer.nvidia.com/nvidia-management-library-nvml).

By setting the GPU operation mode, developers can selectively turn off certain features in the GPU to get the best performance per watt for certain workloads.

The following are the supported GOMs:

- **Compute-Only**: For running compute tasks only.

  By default, the data center GPUs ship in this mode.
- **Low-Double Precision**: For graphics applications that don't require high bandwidth double precision.

  This is recommended for workloads that are not sensitive to double precision but at the same time need graphics capabilities.
- **All On**: This is recommended only when the workload needs full double precision as well as graphics capabilities.

3.2.4.2 WDDM and TCC Driver Models

Along with the GPU operation mode, the developer needs to select the compatible driver model for GPU accelerators for servers.

- Tesla Compute Cluster (TCC): Optimized for running compute workloads.
- Windows Device Driver Model (WDDM): Designed for graphics application and not recommended for compute workloads.

3.2.4.3 Compatibility Between GOM and Driver Models

Table 1 shows which GPU operation modes are compatible with which driver models.

Table 1. GOM and Driver Model Compatibility

GOM	TCC Driver Model	WDDM Driver Model	Use Case Support
All On	YES	YES	All use cases are supported.
Compute-Only	YES	NO	The following are unsupported: X11 and those that require X11 (GLInterop, OCL conformance and VIPER) 32-bit Windows OS
Low Double Precision	YES	YES	All use cases supported.

The compute-only GOM is supported only on the TCC driver model, while the WDDM driver model supports only GOM modes that enable graphics.

The compute-only GOM and WDDM are incompatible and should not be used simultaneously.

The Tesla K20 Active Accelerators for workstations ship in “compute-only” mode and cannot be modified. Therefore, use only the TCC driver model with these products.

3.2.4.4 Using the WDDM Driver Model

To use the WDDM driver model with GPU Accelerators for servers, first switch the GOM mode from compute-only to All On, then switch from TCC to WDDM.

Do not attempt to specify the driver model by editing the registry. Doing so can result in compute-only GOM and WDDM being configured simultaneously, which might require a clean installation of the driver to fix.

Always use NVIDIA-provided tools to specify a processing mode or to switch between driver models.

Such tools include nvidia-smi or the NVIDIA Control Panel -> Manage Maximus Settings page. These tools provide warnings in case of a conflict.

3.3 Known Product Limitations

This section describes problems that will not be fixed. Usually, the source of the problem is beyond the control of NVIDIA. Following is the list of problems and where they are discussed in this document:

- Some APIs Do Not Report Total Available Graphics Memory Correctly
- Using HDMI/DisplayPort Audio with Displays that have a High Native Resolution
- Using HDMI/DisplayPort Audio in Dualview or Clone Mode Configurations
- GPU Runs at a High Performance Level in Multi-display Modes
- Applying Workstation Application Profiles
- Image Sharpening Control Not Available

3.3.1 Some APIs Do Not Report Total Available Graphics Memory Correctly

3.3.1.1 Background–TAG Memory

In the Windows Display Driver Model (WDDM), Total Available Graphics (TAG) memory is reported as the sum of the following:

- Dedicated Video Memory (video memory dedicated for graphics use),
- Dedicated System Memory (system memory dedicated for graphics use), and
- Shared System Memory (system memory shared between the graphics subsystem and the CPU).

The values for each of these components are computed according to WDDM guidelines when the NVIDIA Display Driver is loaded.

3.3.1.2 Issue

NVIDIA has found that some TAG-reporting APIs represent video memory using 32-bits instead of 64-bits, and consequently do not properly report available graphics memory when the TAG would otherwise exceed 4 gigabytes (GB). This results in under reporting of available memory and potentially undesirable behavior of applications that rely on these APIs to report available memory.

The reported memory can be severely reduced. For example, 6 GB might be reported as 454 MB, and 8 GB might be reported as 1259 MB.

3.3.1.3 NVIDIA Action for Some GeForce-based Systems

For GeForce GPUs with 2.75 GB or less of video memory, the NVIDIA display driver constrains TAG memory to just below 4 GB[1]. In this scenario, the Shared System Memory component of TAG is limited first, before limiting Dedicated Video Memory.

This is a policy decision within the driver, and results in reliable reporting of sub-4 GB TAG memory.

---

1. The WDDM guidelines dictate minimum and maximum values for the components, but the display driver may further constrain the values that are reported (within the allowed minimum and maximum).

3.3.1.4 When TAG Reporting Would Not Be Limited

For GeForce-based GPUs with more than 2.75 GB of video memory, as well as all Quadro and Tesla GPUs, the NVIDIA display driver does not constrain TAG memory reporting.

The disadvantage of constraining TAG on systems with larger amounts of video and system memory is that memory which otherwise would be available for graphics use is no longer available. Since shared system memory is limited first, driver components and algorithms utilizing shared system memory may suffer performance degradation when TAG is constrained.

Since these and similar scenarios are prevalent in many Workstation applications, the NVIDIA driver avoids constraining TAG on all Quadro and Tesla-based systems. Likewise, the driver does not constrain TAG for GeForce-based systems with more than 2.75 GB of video memory.

3.3.2 Using HDMI/DisplayPort Audio with Displays That Have a High Native Resolution

To use HDMI/DisplayPort audio with some displays that have a native resolution higher than 1920x1080, you must set the display to a lower HD resolution.

Some HDMI TVs have a native resolution that exceeds the maximum supported HD mode. For example, TVs with a native resolution of 1920x1200 exceed the maximum supported HD mode of 1920x1080.

Applying this native mode results in display overscan which cannot be resized using the NVIDIA Control Panel since the mode is not an HD mode.

To avoid this situation and provide a better user experience, the driver treats certain TVs–such as the Viewsonic VX2835wm and the Westinghouse LVM- 37w3–as a DVI monitor when applying the native mode. Because the driver does not treat the TV as an HDMI in this case, the HDMI audio is not used.

3.3.3 Using HDMI/DisplayPort Audio in Dualview or Clone Mode Configurations

3.3.3.1 Two Audio-Enabled Ports

In a multi-display configuration where both HDMI/DisplayPort audio ports are enabled, only the primary display will provide the audio.

3.3.3.2 One Audio-Enabled Port

In a multi-display configuration where only one audio port is enabled, such as when one display is a DVI display, then the HDMI/DisplayPort display can provide the audio whether is it the primary or secondary display.

3.3.4 GPU Runs at a High-Performance Level in Multi-Display Modes

This is a hardware limitation and not a software bug. Even when no 3D programs are running, the driver will operate the GPU at a high-performance level to efficiently drive multiple displays. In the case of SLI or multi-GPU PCs, the second GPU will always operate with full clock speeds; again, to efficiently drive multiple displays. Today, all hardware from all GPU vendors has this limitation.

3.3.5 Applying Workstation Application Profiles

- Background

  The workstation application profiles are software settings used by the NVIDIA Display Drivers to provide optimum performance when using a selected application. The profile also works around known application issues and bugs.

  If there is an available setting for an application, it should be used, otherwise incorrect behavior or reduced performance is likely to occur.

- Issues

  Configuration changes require that you restart the application.

  Once an application is running, it does not receive notification of configuration changes. Therefore, if you change the configuration while the application is running, you must exit and restart the application for the configuration changes to take effect.

3.3.6 Image Sharpening Control Not Available

The Image sharpening slider on the NVIDIA Control Panel-> Display->Adjust Desktop Color Settings page is grayed out.

This control is intentionally disabled because image sharpening is not supported on current GPUs.

Chapter 4. Hardware and Software Support

This chapter covers the following main topics:

- Supported Operating Systems
- Supported NVIDIA Products
- Supported Languages

4.1 Supported Operating Systems

The Release 550 driver, version 552.22, has been tested with the following Microsoft operating systems:

- Windows Server 2019/2022 (64-bit)

4.2 Supported NVIDIA Products

The following tables list the NVIDIA products supported by the Release 550 driver, version 552.22.

- Supported NVIDIA Workstation Products
- Supported NVIDIA Quadro Sync II Products
- Supported NVIDIA Quadro Sync Products
- Supported NVIDIA Quadro Blade/Embedded Graphics Board Series
- Supported NVIDIA Data Center Products

Table 2. Supported NVIDIA Workstation Products

Product	GPU Architecture
NVIDIA RTX 6000 Ada Generation	NVIDIA Ada Lovelace
NVIDIA RTX 5880 Ada Generation	
NVIDIA RTX 5000 Ada Generation	
NVIDIA RTX 4500 Ada Generation	
NVIDIA RTX 4000 Ada Generation	
NVIDIA RTX 4000 SFF Ada Generation	

Product	GPU Architecture
NVIDIA RTX 2000 Ada Generation	
NVIDIA RTX 2000E Ada Generation	
NVIDIA RTX A6000	NVIDIA Ampere
NVIDIA RTX A5500	
NVIDIA RTX A5000	
NVIDIA RTX A4500	
NVIDIA RTX A4000H	
NVIDIA RTX A4000	
NVIDIA RTX A2000 12GB	
NVIDIA RTX A2000	
NVIDIA RTX A1000	
NVIDIA RTX A400	
NVIDIA Quadro RTX 8000	Turing
NVIDIA Quadro RTX 6000	
NVIDIA Quadro RTX 5000	
NVIDIA Quadro RTX 4000	
NVIDIA T1000 8GB	
NVIDIA T1000	
NVIDIA T600	
NVIDIA T400 4GB	
NVIDIA T400	
NVIDIA T400E	
NVIDIA Quadro GV100	Volta
NVIDIA Quadro GP100	Pascal
NVIDIA Quadro P6000	
NVIDIA Quadro P5000	
NVIDIA Quadro P4000	
NVIDIA Quadro P2200	
NVIDIA Quadro P2000	
NVIDIA Quadro P1000	
NVIDIA Quadro P600	
NVIDIA Quadro P400	
NVIDIA Quadro M6000 24 GB	Maxwell
NVIDIA Quadro M6000	
NVIDIA Quadro M5000	
NVIDIA Quadro M4000	
NVIDIA Quadro M2000	
NVIDIA Quadro K2200	

Product	GPU Architecture
NVIDIA Quadro K1200	
NVIDIA Quadro K620	
NVIDIA NVS 810	

Table 3. Supported NVIDIA Quadro Sync II Products

Product	GPU Architecture
NVIDIA RTX 6000 Ada Generation	NVIDIA Ada Lovelace
NVIDIA RTX 4000 SFF Ada Generation	
NVIDIA RTX A6000	NVIDIA Ampere
NVIDIA RTX A5500	
NVIDIA RTX A5000	
NVIDIA RTX A4500	
NVIDIA RTX A4000H	
NVIDIA RTX A4000	
NVIDIA RTX A2000 12GB	
NVIDIA RTX A2000	
NVIDIA T1000 8GB	Turing
NVIDIA T1000	
NVIDIA T600	
NVIDIA T400 4GB	
NVIDIA T400	
NVIDIA T400E	
NVIDIA Quadro GP100	Pascal
NVIDIA Quadro P6000	
NVIDIA Quadro P5000	
NVIDIA Quadro P4000	

Table 4. Supported NVIDIA Quadro Sync Products

Product	GPU Architecture
NVIDIA Quadro M6000 24 GB	Maxwell
NVIDIA Quadro M6000	
NVIDIA Quadro M5000	
NVIDIA Quadro M4000	

Table 5. Supported NVIDIA Quadro Blade/Embedded Graphics Board Series

Product	GPU Architecture
NVIDIA Quadro RTX 5000	Turing
NVIDIA Quadro RTX 3000	
NVIDIA Quadro T1000	
NVIDIA Quadro P5000	Pascal
NVIDIA Quadro P3000	
NVIDIA Quadro P2000	
NVIDIA Quadro P1000	
NVIDIA Quadro M5000 SE	Maxwell
NVIDIA Quadro M3000 SE	

Table 6. Supported NVIDIA Data Center Products

Product	GPU Architecture
NVIDIA L-Series Products	
NVIDIA L40	NVIDIA Ada Lovelace
NVIDIA L20	
NVIDIA L4	
NVIDIA L2	
NVIDIA H-Series Products	
NVIDIA H100	NVIDIA Hopper
NVIDIA A-Series Products	
NVIDIA AX800	NVIDIA Ampere
NVIDIA A100	
NVIDIA A40	
NVIDIA A30	
NVIDIA A16	
NVIDIA A10	
NVIDIA A2	
NVIDIA T-Series Products	
NVIDIA T4	Turing
NVIDIA V-Series Products	
NVIDIA V100	Volta
Tesla P-Series Products	
NVIDIA Tesla P100	Pascal
NVIDIA Tesla P40	
NVIDIA Tesla P4	
Tesla M-Series Products	
NVIDIA Tesla M60	Maxwell

Product	GPU Architecture
NVIDIA Tesla M6	

4.3 Supported Languages

The Release 550 Graphics Drivers supports the following languages in the main driver Control Panel:

English (USA)	German	Portuguese (Euro/Iberian)
English (UK)	Greek	Russian
Arabic	Hebrew	Slovak
Chinese (Simplified)	Hungarian	Slovenian
Chinese (Traditional)	Italian	Spanish
Czech	Japanese	Spanish (Latin America)
Danish	Korean	Swedish
Dutch	Norwegian	Thai
Finnish	Polish	Turkish
French	Portuguese (Brazil)	

4.4 Driver Installation

4.4.1 Minimum Hard Disk Space

The hard disk space requirement is approximately 1.5x the size of the installation download to accommodate extracted and temporary files.

4.4.2 Before You Begin

If you have previously installed NVIDIA nTune, NVIDIA recommends that you uninstall nTune before installing this driver. After the driver install is complete, you can reinstall NVIDIA nTune.

4.5 Installation Instructions

1. Follow the instructions on the NVIDIA.com website driver download page to locate the appropriate driver to download, based on your hardware and operating system.
2. Click the driver download link.
3. The license agreement dialog box appears.
4. Click Accept if you accept the terms of the agreement, then either open the file or save the file to your PC and open it later.
5. Extract the zip files to a temporary folder on your PC.

6. Open the NVIDIA driver installation .EXE file to run the NVIDIA Package Launcher to extract the driver files.
7. Follow the instructions in the NVIDIA InstallShield Wizard to complete the installation.

Chapter 5. NVIDIA Tesla Compute Cluster Mode

This chapter describes the Tesla Compute Cluster (TCC) mode.

5.1 About Tesla Compute Cluster Mode

5.1.1 TCC Overview

Tesla Compute Cluster (TCC) mode is designed for compute cluster nodes that have one or more Tesla or supported NVIDIA RTX / Quadro products installed.

5.1.1.1 Benefits

- TCC drivers make it possible to use NVIDIA GPUs in nodes with non-NVIDIA integrated graphics.
- NVIDIA GPUs on systems running the TCC drivers will be available via Remote Desktop, both directly and via cluster management systems that rely on Remote Desktop.
- NVIDIA GPUs will be available to applications running as a Windows service (i.e., in Session 0) on systems running the Tesla/Quadro driver in TCC mode.

5.1.1.2 TCC Does not Support Graphics Acceleration

- TCC mode does not provide CUDA–DirectX/OpenGL interoperability.

  It is a “non-display” driver, and NVIDIA GPUs using this driver will not support DirectX or OpenGL hardware acceleration.

5.1.2 Running CUDA Applications

- This release of the NVIDIA RTX / Quadro driver supports CUDA C/C++ applications and libraries that rely on the CUDA C Runtime and/or CUDA Driver API.
- NVIDIA GPUs running the NVIDIA RTX / Quadro driver in TCC mode will be available for CUDA applications running via services or Remote Desktop.

- In this release, all GPUs will be in compute exclusive mode. As a result, only one CUDA context may exist on a particular device at a time.
- SDK applications that use graphics will not run properly under TCC mode. The following are examples of CUDA SDK applications that are not supported:

bicubicTexture	boxFilter	cudaDecodeD3D9	smokeParticles
cudaDecodeGL	fluidsD3D9	fluidsGL	SobelFilter
imageDenoising	Mandelbrot	marchingCubes	volumeRender
nbody	oceanFFT	particles	
postProcessGL	recursiveGaussian	simpleD3D10	
simpleD3D10Texture	simpleD3D11Texture	simpleD3D9	
simpleD3D9Texture	simpleGL	simpleTexture3D	

5.2 Operating on Systems with non-TCC NVIDIA GPUs

- NVIDIA GPUs running under TCC mode may coexist with other display devices.
- The NVIDIA RTX / Quadro driver is installed over any NVIDIA display driver in the system – the NVIDIA Tesla driver then becomes the only driver for all NVIDIA GPUs in the system.
- If the NVIDIA RTX / Quadro driver is uninstalled later, the previous driver is not restored.
- NVIDIA GPUs that do not support TCC mode will appear as “VGA adapters” in the Windows Device Manager and can be used to drive displays.

  Non-supported NVIDIA GPUs can still function as CUDA devices, but the GPU’s graphics functionality is not available to applications.

Notice

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.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication of otherwise under any patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all other information previously supplied. NVIDIA Corporation products are not authorized as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

Macrovision Compliance Statement

NVIDIA Products that are Macrovision enabled can only be sold or distributed to buyers with a valid and existing authorization from Macrovision to purchase and incorporate the device into buyer's products.

Macrovision copy protection technology is protected by U.S. patent numbers 5,583,936; 6,516,132; 6,836,549; and 7,050,698 and other intellectual property rights. The use of Macrovision's copy protection technology in the device must be authorized by Macrovision and is intended for home and other limited pay-per-view uses only, unless otherwise authorized in writing by Macrovision. Reverse engineering or disassembly is prohibited.

OpenCL Notice

Portions of the NVIDIA system software contain components licensed from third parties under the

following terms: Clang & LLVM:

Copyright (c) 2003-2008 University of Illinois at

Urbana-Champaign. All rights reserved.

Portions of LLVM's System library:

Copyright (C) 2004 eXtensible

Systems, Inc. Developed by:

LLVM Team

University of Illinois at Urbana-

Champaign http://llvm.org

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal with the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimers.

Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimers in the documentation and/or other materials provided with the distribution.

Neither the names of the LLVM Team, University of Illinois at Urbana-Champaign, nor the names of its contributors may be used to endorse or promote products derived from this Software without specific prior written permission.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE CONTRIBUTORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS WITH THE SOFTWARE.

Vulkan Notice

Copyright (c) 2015-2016 The Khronos

Group Inc. Copyright (c) 2015-2016

LunarG, Inc.

Copyright (c) 2015-2016 Valve Corporation

The Vulkan Runtime is comprised of 100% open source components (MIT, and Apache 2.0). The text of such licenses is included below along with the copyrights.

ALL INFORMATION HERE IS PROVIDED "AS IS." LUNARG MAKES NO REPRESENTATIONS OR WARRANTIES, EXPRESS OR IMPLIED, WITH REGARD TO THIS LIST OR ITS ACCURACY OR COMPLETENESS, OR WITH RESPECT TO ANY RESULTS TO BE OBTAINED FROM USE OR

DISTRIBUTION OF THE LIST. BY USING OR DISTRIBUTING THIS LIST, YOU AGREE THAT IN NO EVENT SHALL LUNARG BE HELD LIABLE FOR ANY DAMAGES WHATSOEVER RESULTING FROM ANY USE OR DISTRIBUTION OF THIS LIST, INCLUDING, WITHOUT LIMITATION, ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR OTHER DIRECT OR INDIRECT DAMAGES.

Softfloat Notice

Portions of the driver use the Softfloat floating point emulation library.

Softfloat Release 3e (http://www.jhauser.us/arithmetic/SoftFloat.html) is provided under the following terms:

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions, and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions, and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of the University nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE REGENTS AND CONTRIBUTORS "AS IS", AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL,

SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF

LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

==========Apache 2.0==========

Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on as "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

==========MIT==========

Copyright (c) 2009 Dave Gamble

Copyright (c) 2015-2016 The Khronos Group Inc.
Copyright (c) 2015-2016 Valve Corporation

Copyright (c) 2015-2016 LunarG, Inc.

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Trademarks

Copyright