



VIRTUAL GPU MANAGEMENT PACK FOR VMWARE VREALIZE OPERATIONS

DU-08661-001 _v1.0 through 1.1 | May 2020

User Guide



TABLE OF CONTENTS

| | |
|---|-----------|
| Chapter 1. Introduction to the NVIDIA Virtual GPU Management Pack for VMware vRealize Operations..... | 1 |
| Chapter 2. Installing and Configuring the NVIDIA Virtual GPU Management Pack for VMware vRealize Operations..... | 3 |
| 2.1. Installation and Configuration Prerequisites..... | 3 |
| 2.2. Installing or Updating the Management Pack..... | 3 |
| 2.3. Creating an NVIDIA vGPU Adapter Instance..... | 4 |
| 2.4. Assigning Privileges that the NVIDIA vGPU Adapter Requires..... | 6 |
| Chapter 3. Managing Metrics and Analytics for NVIDIA vGPU Software in VMware vRealize Operations..... | 9 |
| 3.1. Viewing Data on NVIDIA Dashboards..... | 9 |
| 3.2. Changing the NVIDIA vGPU Adapter Collection Interval..... | 10 |
| 3.3. Changing the Threshold of a Symptom in an Alert Definition..... | 10 |
| Appendix A. NVIDIA vGPU Alert Definitions..... | 12 |
| A.1. GPU Utilization Is High..... | 12 |
| A.2. vGPU Utilization Is High..... | 13 |
| A.3. vGPU Utilization Is High for Process..... | 13 |
| A.4. GPU Temperature Is High..... | 14 |

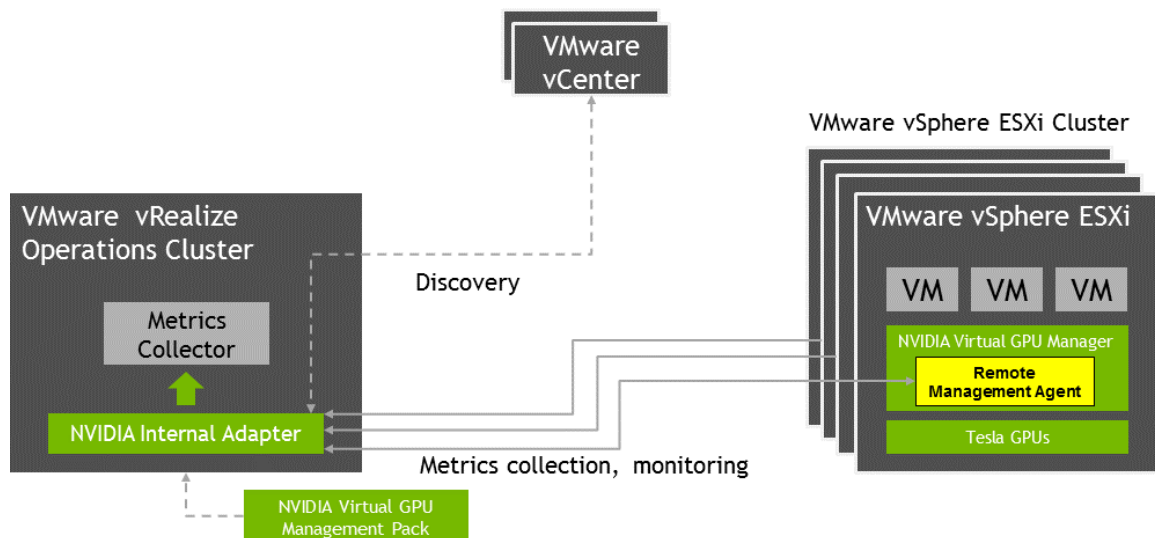
Chapter 1.

INTRODUCTION TO THE NVIDIA VIRTUAL GPU MANAGEMENT PACK FOR VMWARE VREALIZE OPERATIONS

NVIDIA® Virtual GPU Management Pack for VMware vRealize Operations enables you to use a VMware vRealize Operations cluster to monitor the performance of NVIDIA physical GPUs and virtual GPUs.

VMware vRealize Operations provides integrated performance, capacity, and configuration management capabilities for VMware vSphere, physical and hybrid cloud environments. It provides a management platform that can be extended by adding third-party management packs. For more information, see the [VMware vRealize Operations documentation](#).

NVIDIA Virtual GPU Management Pack for VMware vRealize Operations collects metrics and analytics for NVIDIA vGPU software from virtual GPU manager instances. It then sends these metrics to the metrics collector in a VMware vRealize Operations cluster, where they are displayed in custom NVIDIA dashboards.



Chapter 2.

INSTALLING AND CONFIGURING THE NVIDIA VIRTUAL GPU MANAGEMENT PACK FOR VMWARE VREALIZE OPERATIONS

The NVIDIA Virtual GPU Management Pack for VMware vRealize Operations is distributed as a PAK (.pak) file. After installing the NVIDIA Virtual GPU Management Pack for VMware vRealize Operations, you must configure it by creating an NVIDIA vGPU adapter instance and, if you haven't already done so, by creating a VMware vCenter adapter instance.

2.1. Installation and Configuration Prerequisites

Before installing and configuring the NVIDIA Virtual GPU Management Pack for VMware vRealize Operations, ensure that supported versions of the required software are installed and configured as follows:

- ▶ vRealize Operations Manager is installed.
- ▶ The NVIDIA vGPU software driver package is configured on the hosts in your VMware vSphere ESXi cluster.

For details about which releases of the required software are supported, see *Virtual GPU Management Pack for VMware vRealize Operations Release Notes*.

2.2. Installing or Updating the Management Pack

The NVIDIA Virtual GPU Management Pack for VMware vRealize Operations is distributed as a PAK (.pak) file.

If you have previously installed the NVIDIA Virtual GPU Management Pack for VMware vRealize Operations, back up any customized dashboards before updating the management pack. The update will overwrite any NVIDIA dashboard of the same name.

1. Download the NVIDIA Virtual GPU Management Pack for VMware vRealize Operations PAK (.pak) file.
Ensure that the downloaded file is accessible to the web browser that you are using to manage your **vRealize Operations Manager** instance.
2. Log in to your **vRealize Operations Manager** instance as an administrator user.
3. On the **vRealize Operations Manager Home** page, follow the **Administration** link.
4. Click **Solutions** and click the plus sign in the toolbar.
5. Click **Browse** and navigate to your copy of the PAK file.
6. If you have previously installed the NVIDIA Virtual GPU Management Pack for VMware vRealize Operations, select these options:
 - ▶ **Install the PAK file even if it is already installed**
 - ▶ **Reset Default Content**
7. Select the PAK file and click **Upload**.
8. Accept the EULA for the NVIDIA Virtual GPU Management Pack for VMware vRealize Operations and click **Next**.



Uploading and installing the PAK file may take several minutes. Status information appears in the **Installation Details** text box throughout the installation process.

9. When the installation is complete, click **Finish**.
This last page displays progress details for the installation.

2.3. Creating an NVIDIA vGPU Adapter Instance

After installing the NVIDIA Virtual GPU Management Pack for VMware vRealize Operations, you must configure it by creating an NVIDIA vGPU adapter instance.

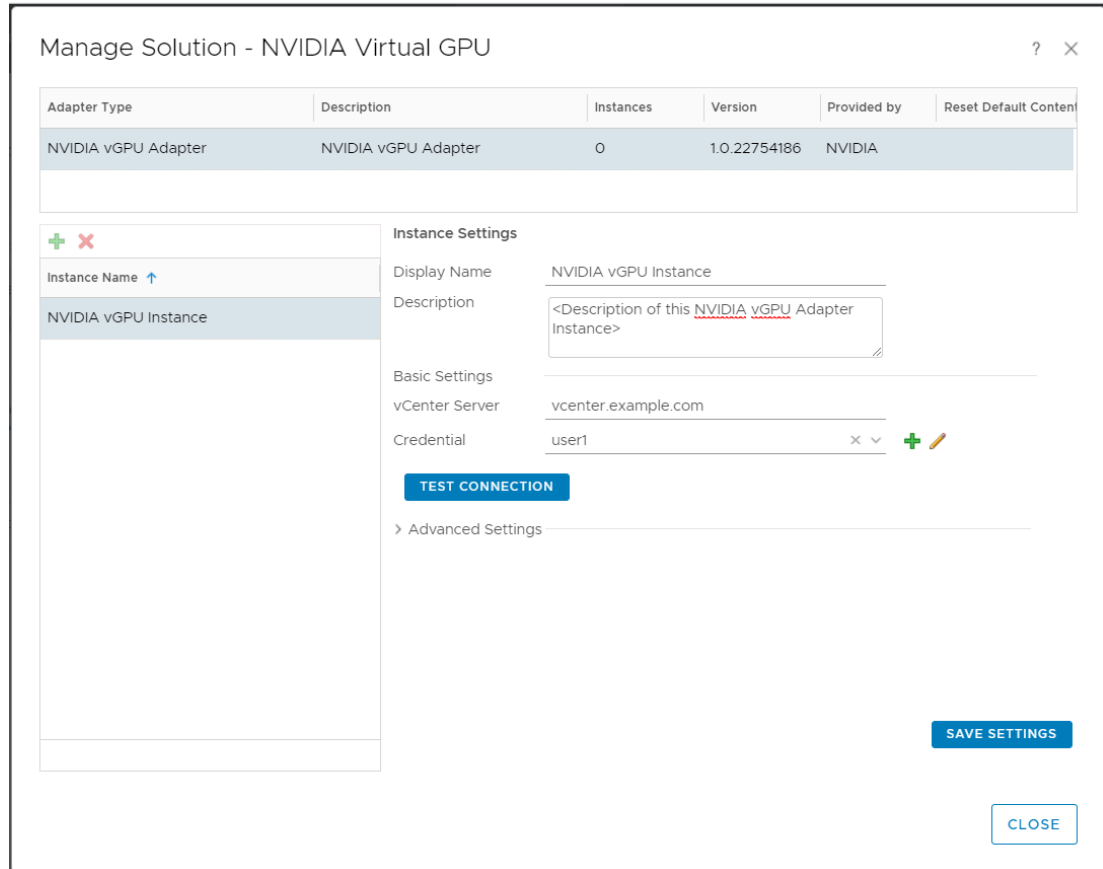


If you haven't already done so, you must also create a VMware vCenter adapter instance.

An NVIDIA vGPU adapter instance connects to a VMware vCenter Server instance and retrieves data from vGPU-enabled hosts in the server instance. You must provide the host name of the VMware vCenter Server instance that the adapter instance will connect to and credentials to be used for connecting to the server instance.

1. If you are not already logged in, log in to your **vRealize Operations Manager** instance as an administrator user.
2. On the **vRealize Operations Manager Home** page, follow the **Administration** link.
3. Click **Solutions**, select **NVIDIA Virtual GPU Management Pack for VMware vRealize Operations**, and click **Configure** on the toolbar.

The **Manage Solution** page opens.



4. From the **Adapter Type** list at the top of page, select **NVIDIA vGPU Adapter**.
5. Click the plus sign.
6. Provide the following information about the adapter instance that you are creating:

Display Name

Enter the name of the instance as you want it to appear in **vRealize Operations Manager**.

Description

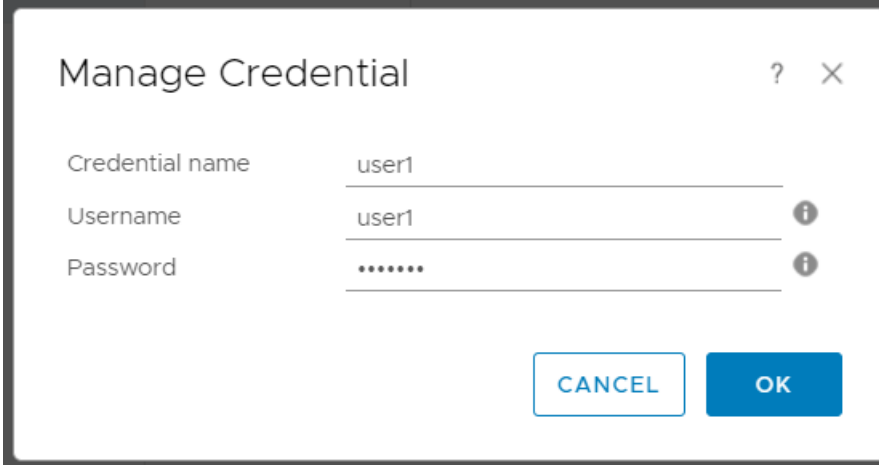
Enter a description that can help distinguish this instance when multiple NVIDIA vGPU adapter instances are configured.

vCenter Server

Enter the IP address of the VMware vCenter Server.

Credential

Click the plus sign and in the **Manage Credential** dialog box that opens, add the credentials for the user that will connect to this vCenter Server instance.



The screenshot shows a 'Manage Credential' dialog box. It has a title bar with a question mark and a close button. The dialog contains three input fields: 'Credential name' with the value 'user1', 'Username' with the value 'user1', and 'Password' with masked characters '.....'. Each input field has an information icon to its right. At the bottom right, there are two buttons: 'CANCEL' and 'OK'.

Credential name

Enter the display name of the user.

Username

Enter the user login name.

Password

Enter the password of the user.

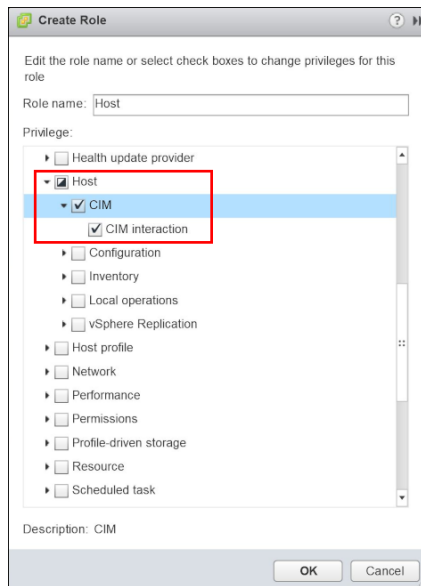
7. Click **Save Settings**.

After installing and configuring NVIDIA Virtual GPU Management Pack for VMware vRealize Operations, verify the installation and configuration as explained in [Viewing Data on NVIDIA Dashboards](#).

2.4. Assigning Privileges that the NVIDIA vGPU Adapter Requires

To collect data from hosts in VMware vCenter that are running NVIDIA GPUs and the NVIDIA vGPU Manager, each user of the NVIDIA vGPU adapter requires the CIM interaction privilege. If this privilege is not assigned, the user cannot use the NVIDIA vGPU adapter to collect data.

1. Log in to vCenter Server by using the vSphere Web Client.
2. Select **Administration** and, in the **Access Control** area, select **Roles**.
3. From the **Roles Provider** list, select your vCenter Server instance.
4. Click the **Create role action (+)** icon.
5. In the **Create Role** window that opens, define the properties of the role.
 - a) In the **Role name** field, type your choice of name for the role.
 - b) Expand the **Host** privilege and then expand the **CIM** privilege under the **Host** privilege.
 - c) Select the **Host**, **CIM**, and **CIM interaction** privileges.
 - d) Click **OK**.



6. In the navigation tree, select **Home > Hosts and Clusters > your-vCenter-Server-instance > Permissions** .
7. In the **Users and Groups** section of the **Add Permission** window, select the users and groups that will use the NVIDIA vGPU adapter.
 - a) Click the **Add permission (+)** icon.
The **Select Users/Groups** window opens.
 - b) For each user or group, select the user or group and click **Add**.
 - c) After adding all the users or groups, click **OK** in the **Select Users/Groups** window.
8. In the **Assigned Role** section of the **Add Permission** window, select the role that you created from the list of roles and select the **Propagate to children** check box.

Chapter 3.

MANAGING METRICS AND ANALYTICS FOR NVIDIA vGPU SOFTWARE IN VMWARE VREALIZE OPERATIONS

Managing metrics and analytics for NVIDIA vGPU software in VMware vRealize Operations involves viewing data on NVIDIA dashboards and changing the settings of the NVIDIA vGPU adapter and NVIDIA vGPU alert definitions.

3.1. Viewing Data on NVIDIA Dashboards

After installing and configuring NVIDIA Virtual GPU Management Pack for VMware vRealize Operations, you can view the data on NVIDIA dashboards to verify the installation and configuration. If you have just completed the installation and configuration, allow the adapter to work for ten to fifteen minutes to collect data to display on the dashboards.

1. On the **vRealize Operations Manager Home** page, click **Dashboards** in the menu bar.
2. In the **All Dashboards** drop-down list, select the **NVIDIA Dashboards** group.

This group contains the following dashboards:

- ▶ **NVIDIA Environment Overview**
- ▶ **NVIDIA Host Summary**
- ▶ **NVIDIA GPU Summary**
- ▶ **NVIDIA vGPU Summary**
- ▶ **NVIDIA Application Summary**

3.2. Changing the NVIDIA vGPU Adapter Collection Interval

If you need to change how frequently the NVIDIA vGPU adapter collects metrics, change the collection interval. The default collection interval is five minutes.

1. If you are not already logged in, log in to your **vRealize Operations Manager** instance as an administrator user.
2. On the **vRealize Operations Manager Home** page, follow the **Administration** link.
3. In the left pane, click **Configuration**.
4. Click **Inventory Explorer** and expand **Adapter Instances** in the center pane.
5. Expand **NVIDIA vGPU Adapter Instance** and select the adapter name.
6. In the right pane, on the **List** tab, select the adapter name and click **Edit Object**.
7. On **Advanced Settings**, enter the new collection interval in the **Collection Interval (Minutes)** field.



The minimum value that you can set is 1 minute.

8. Click **OK**.

3.3. Changing the Threshold of a Symptom in an Alert Definition

An alert definition is a combination of symptoms that identify a problem area and generate alerts for that area. Each symptom in an alert is associated with a metric. For each symptom, a threshold value is defined for its associated metric. If the threshold value is reached, an alert is generated.

For detailed information about the alerts defined for NVIDIA vGPU metrics, including the default threshold values of symptoms in these alerts, see [NVIDIA vGPU Alert Definitions](#).

1. In the menu bar of the **vRealize Operations Manager Home** page, click **Alerts**.
2. In the left pane, click **Alert Settings**.
3. Click **Symptom Definitions**.
4. Click **All Filters**, then click **Object Type**, and type **GPU** or **vGPU**.
The symptom definitions for the object type that you selected are listed.
5. Select the symptom definition that you want to change and click the **Edit** icon.
6. Change the threshold to the new value that you want and click **Save**.

GPU : Utilization|Memory Utilization (%)

Static Threshold

GPU Memory Utilization is moderate| is Warning when metric is greater than or equal 75

► Advanced

Appendix A.

NVIDIA vGPU ALERT DEFINITIONS

The management pack provides alert definitions for the NVIDIA vGPU metrics and analytics that it integrates with VMware vRealize Operations. Each alert definition is a combination of symptoms that identify a problem area and generate alerts for that area.

Alerts defined for GPU utilization can be generated by any of the GPU engines, namely:

- ▶ 3D/Compute
- ▶ Memory controller
- ▶ Video encoder
- ▶ Video decoder

A.1. GPU Utilization Is High

This alert is generated when the utilization of any of the GPU engines is high.

| Symptom | Associated Metric | Criticality | Threshold |
|---|---|-------------|-----------|
| GPU 3D/Compute Utilization is critically high | GPU: Utilization 3D/Compute Utilization | Immediate | 90 |
| GPU 3D/Compute Utilization is moderately high | GPU: Utilization 3D/Compute Utilization | Warning | 75 |
| GPU Memory Utilization is critically high | GPU: Utilization Memory Utilization | Immediate | 90 |
| GPU Memory Utilization is moderately high | GPU: Utilization Memory Utilization | Warning | 75 |
| GPU Encoder Utilization is critically high | GPU: Utilization Encoder Utilization | Immediate | 90 |
| GPU Encoder Utilization is moderately high | GPU: Utilization Encoder Utilization | Warning | 75 |
| GPU Decoder Utilization is critically high | GPU: Utilization Decoder Utilization | Immediate | 90 |
| GPU Decoder Utilization is moderately high | GPU: Utilization Decoder Utilization | Warning | 75 |

A.2. vGPU Utilization Is High

This alert is generated when the utilization of any of the GPU engines is high on any virtual GPU.

| Symptom Name | Associated Metric | Criticality | Threshold |
|--|--|-------------|-----------|
| vGPU 3D/Compute Utilization is critically high | vGPU: Utilization 3D/Compute Utilization | Immediate | 90 |
| vGPU 3D/Compute Utilization is moderately high | vGPU: Utilization 3D/Compute Utilization | Warning | 75 |
| vGPU Memory Utilization is critically high | vGPU: Utilization Memory Utilization | Immediate | 90 |
| vGPU Memory Utilization is moderately high | vGPU: Utilization Memory Utilization | Warning | 75 |
| vGPU Encoder Utilization is critically high | vGPU: Utilization Encoder Utilization | Immediate | 90 |
| vGPU Encoder Utilization is moderately high | vGPU: Utilization Encoder Utilization | Warning | 75 |
| vGPU Decoder Utilization is critically high | vGPU: Utilization Decoder Utilization | Immediate | 90 |
| vGPU Decoder Utilization is moderately high | vGPU: Utilization Decoder Utilization | Warning | 75 |

A.3. vGPU Utilization Is High for Process

This alert is generated when the utilization of any of the GPU engines is high for any process on any virtual GPU.

| Symptom Name | Associated Metric | Criticality | Threshold |
|--|---------------------------------|-------------|-----------|
| vGPU 3D/Compute Utilization is critically high for Process | Process: 3D/Compute Utilization | Immediate | 90 |
| vGPU 3D/Compute Utilization is moderately high for Process | Process: 3D/Compute Utilization | Warning | 75 |
| vGPU Memory Utilization is critically high for Process | Process: Memory Utilization | Immediate | 90 |
| vGPU Memory Utilization is moderately high for Process | Process: Memory Utilization | Warning | 75 |
| vGPU Encoder Utilization is critically high for Process | Process: Encoder Utilization | Immediate | 90 |
| vGPU Encoder Utilization is moderately high for Process | Process: Encoder Utilization | Warning | 75 |
| vGPU Decoder Utilization is critically high for Process | Process: Decoder Utilization | Immediate | 90 |

| Symptom Name | Associated Metric | Criticality | Threshold |
|---|------------------------------|-------------|-----------|
| vGPU Decoder Utilization is moderately high for Process | Process: Decoder Utilization | Warning | 75 |

A.4. GPU Temperature Is High

This alert is generated when the GPU temperature is high enough to force slowdown or shutdown.

| Symptom | Associated Metric | Criticality | Threshold |
|-------------------------------------|--------------------------------------|-------------|------------------------------|
| GPU Temperature is forcing slowdown | GPU: Temperature Current Temperature | Critical | Slowdown Temperature |
| GPU Temperature is forcing shutdown | GPU: Temperature Current Temperature | Immediate | Shutdown Temperature minus 5 |

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.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

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

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA GRID, NVIDIA GRID vGPU, NVIDIA Maxwell, NVIDIA Pascal, NVIDIA Turing, NVIDIA Volta, Quadro, and Tesla are trademarks or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2017-2020 NVIDIA Corporation. All rights reserved.