



Virtual GPU Management Pack for VMware Aria Operations

User Guide

Table of Contents

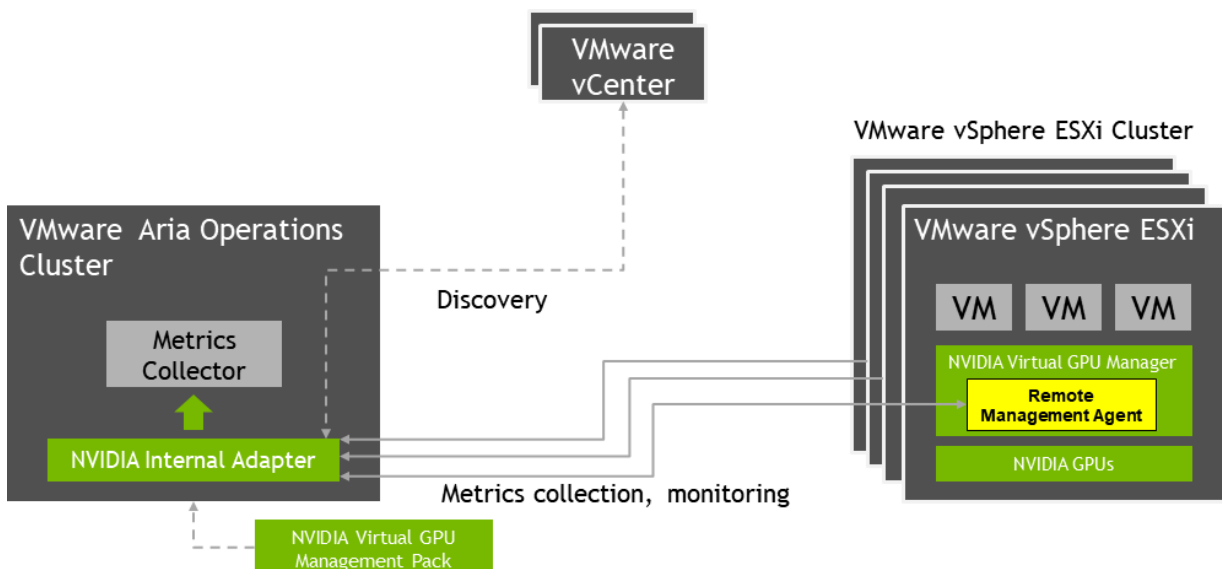
Chapter 1. Introduction to the NVIDIA Virtual GPU Management Pack for VMware Aria Operations.....	1
Chapter 2. Installing and Configuring the NVIDIA Virtual GPU Management Pack for VMware Aria Operations.....	3
2.1. Installation and Configuration Prerequisites.....	3
2.2. Installing or Updating the Management Pack.....	4
2.2.1. Installing or Updating the Management Pack on Premises.....	4
2.2.2. Installing or Updating the Management Pack on VMware Aria Operations Cloud.....	5
2.3. Creating an NVIDIA vGPU Adapter Instance.....	6
2.4. Assigning the CIM Interaction Privileges for the NVIDIA vGPU Adapter.....	8
2.5. Adding a Class to the VMware Aria Operations Logs.....	10
Chapter 3. Managing Metrics and Analytics for NVIDIA vGPU Software in VMware Aria Operations.....	12
3.1. Viewing Data on NVIDIA Dashboards.....	12
3.2. Changing the NVIDIA vGPU Adapter Collection Interval.....	13
3.3. Changing the Threshold of a Symptom in an Alert Definition.....	13
Appendix A. NVIDIA vGPU Alert Definitions.....	15
A.1. GPU Utilization Is High.....	15
A.2. vGPU Utilization Is High.....	16
A.3. vGPU Utilization Is High for Process.....	16
A.4. GPU Temperature Is High.....	17
Appendix B. Metrics Presented by NVIDIA Virtual GPU Management Pack for VMware Aria Operations.....	18
B.1. GPU Metrics.....	18
B.2. NVIDIA vGPU Metrics.....	19
B.3. Process Metrics.....	22

Chapter 1. Introduction to the NVIDIA Virtual GPU Management Pack for VMware Aria Operations

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

VMware Aria 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 Aria Operations documentation](#).

NVIDIA Virtual GPU Management Pack for VMware Aria 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 Aria Operations cluster, where they are displayed in custom NVIDIA dashboards.



Chapter 2. Installing and Configuring the NVIDIA Virtual GPU Management Pack for VMware Aria Operations

You can install or update the NVIDIA Virtual GPU Management Pack for VMware Aria Operations on an on-premises installation of VMware Aria Operations Manager or on VMware Aria Operations Cloud. After installing the NVIDIA Virtual GPU Management Pack for VMware Aria 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 Aria Operations, ensure that supported versions of the required software are available and configured as follows:

- ▶ VMware Aria Operations Manager is installed or you have a VMware Aria Operations Cloud account.
- ▶ The NVIDIA vGPU software driver package is installed and configured on the hosts in your VMware vSphere ESXi cluster.

If the NVIDIA Virtual GPU Manager in the NVIDIA vGPU software driver package is based on the VMware Daemon SDK (DSDK), **both** components of the NVIDIA Virtual GPU Manager must be installed and configured, namely:

- ▶ NVIDIA vGPU hypervisor host driver
- ▶ NVIDIA GPU Management daemon

For information about how to install and configure the NVIDIA vGPU software driver package, refer to [Virtual GPU Software User Guide](#).

For details about which releases of the required software are supported, refer to [Virtual GPU Management Pack for VMware Aria Operations Release Notes](#).

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

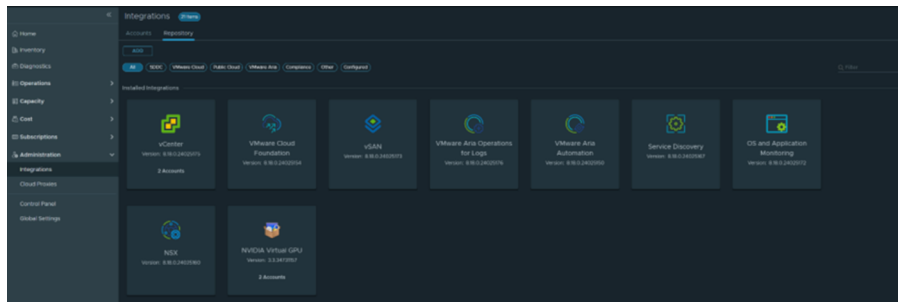
2.2. Installing or Updating the Management Pack

You can install or update the NVIDIA Virtual GPU Management Pack for VMware Aria Operations on an on-premises installation of VMware Aria Operations Manager or on VMware Aria Operations Cloud.

2.2.1. Installing or Updating the Management Pack on Premises

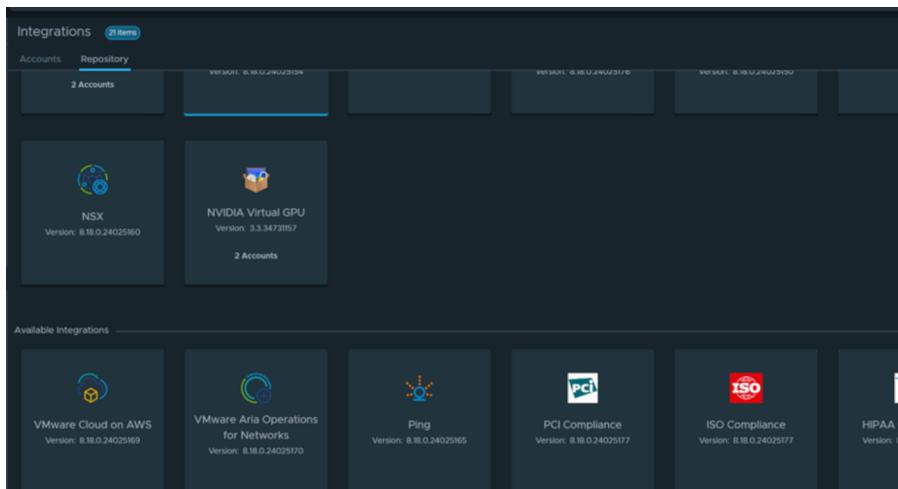
For an installation on VMware Aria Operations Manager on premises, the NVIDIA Virtual GPU Management Pack for VMware Aria Operations is distributed as a ZIP archive that contains a PAK (.pak) file.

1. Download the NVIDIA Virtual GPU Management Pack for VMware Aria Operations ZIP archive from the **VMware Solution Exchange** website and extract the PAK (.pak) file. Ensure that the extracted 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. Start the **Add Solution** wizard.
 - a). On the **vRealize Operations Manager Home** page, expand **Administration** and select **Integrations**.
 - b). On the **Integrations** page that opens, click the **Repository** tab.
 - c). On the **Repository** tab, click **ADD**.



4. In the **Add Solution** wizard that opens, click **Browse**, navigate to your copy of the PAK file and select it, and click **Open**.

5. If you have previously installed the NVIDIA Virtual GPU Management Pack for VMware Aria Operations, select these options:
 - ▶ **Install the PAK file even if it is already installed**
 - ▶ **Reset Default Content**
6. Click **Upload**.
7. After VMware Aria Operations has uploaded NVIDIA Virtual GPU Management Pack for VMware Aria Operations, click **Next**.
The End User License Agreement (EULA) appears.
8. Accept the EULA for the NVIDIA Virtual GPU Management Pack for VMware Aria Operations.
9. Click **Next** to start the installation process.
10. When the installation is complete, click **Finish**.
This last page displays progress details for the installation.
11. To confirm that the installation succeeded, on the **Repository** tab of the **Integrations** page, scroll through the **Installed Integrations** section until you see NVIDIA Virtual GPU Management Pack for VMware Aria Operations.



2.2.2. Installing or Updating the Management Pack on VMware Aria Operations Cloud

For an installation on VMware Aria Operations Cloud, the NVIDIA Virtual GPU Management Pack for VMware Aria Operations is available on the **Repository** tab of the **Integrations** page.

1. Log in to **VMware vRealize Operations Cloud** as an administrator user.
2. Navigate to the **Repository** tab of the **Integrations** page.
 - a). On the **VMware vRealize Operations Cloud** page, expand **Data Sources** and select **Integrations**.
 - b). On the **Integrations** page that opens, click the **Repository** tab.

3. On the **Repository** tab, perform the step for installing or updating NVIDIA Virtual GPU Management Pack for VMware Aria Operations.

Action	Step
Install NVIDIA Virtual GPU Management Pack for VMware Aria Operations if it is not already installed.	In the Available Integrations area, locate NVIDIA Virtual GPU Management Pack for VMware Aria Operations and click GET .
Update an existing installation of NVIDIA Virtual GPU Management Pack for VMware Aria Operations.	In the Installed Integrations area, locate NVIDIA Virtual GPU Management Pack for VMware Aria Operations and click UPGRADE .

After you perform this step, VMware Aria Operations Cloud installs or updates the NVIDIA Virtual GPU Management Pack for VMware Aria Operations without prompting you for any information.

2.3. Creating an NVIDIA vGPU Adapter Instance

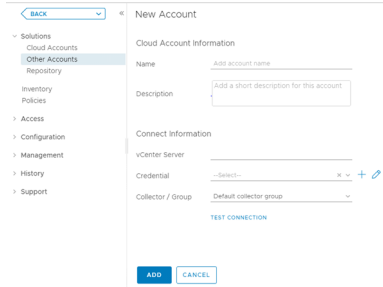
After installing the NVIDIA Virtual GPU Management Pack for VMware Aria Operations, you must configure it by creating an NVIDIA vGPU adapter instance.



Note: 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. Navigate to the **Account Types** page.
 - a). On the **vRealize Operations Manager Home** page, expand **Administration** and select **Integrations**.
 - b). On the **Integrations** page that opens, click **ADD**.
3. On the **Account Types** page that opens, select **NVIDIA vGPU Adapter**.
The **New Account** page opens.



4. Provide the following information about the adapter instance that you are creating:

Name

Enter the name of the instance as you want it to appear in **vRealize Operations Manager**, for example **vCenter_1**.

Description

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

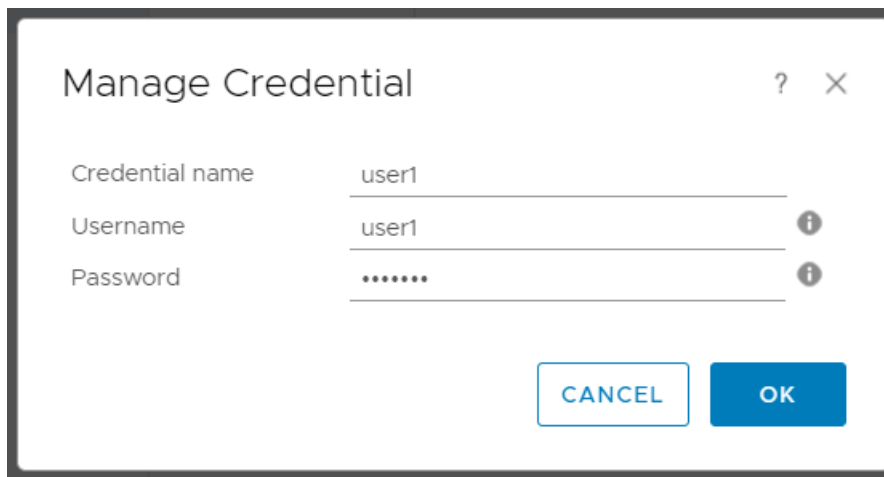
vCenter Server

Enter the fully qualified domain name or IP address of the VMware vCenter Server.

Credential

Enter the credential name for logging in to the VMware vCenter Server instance.

If no suitable credential exists, click the plus sign and in the **Manage Credential** dialog box that opens, add the credentials for the user that will connect to this VMware vCenter Server instance.



Credential Name

Enter your choice of name that uniquely identifies the VMware vCenter Server login credentials.

Username

Enter the login name of the VMware vCenter Server user. The user must have at least CIM privileges on all the hosts in the VMware vCenter Server instance.

Password

Enter the password of the user.

5. Back on the **New Account** page, click **Validate Connection** to test the connection between the new adapter instance and the VMware vCenter Server
6. Click **ADD**.

VMware Aria Operations starts to collect data for NVIDIA virtual GPUs. After approximately ten to fifteen minutes, the Collection State of the NVIDIA vGPU changes to `Collecting` and the Collection Status changes to `Data receiving`.

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

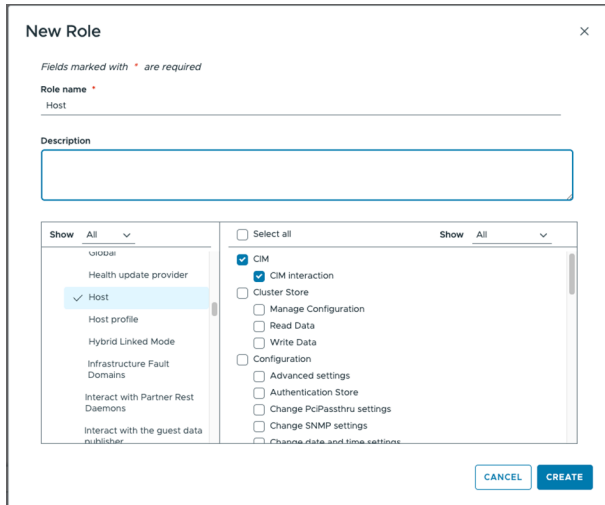
2.4. Assigning the CIM Interaction Privileges for the NVIDIA vGPU Adapter

To collect data from hosts in VMware vCenter that are running NVIDIA GPUs and an NVIDIA GPU Management Daemon that uses CIM Service Ticket-based authentication, which was introduced in NVIDIA vGPU software 20.0, 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.

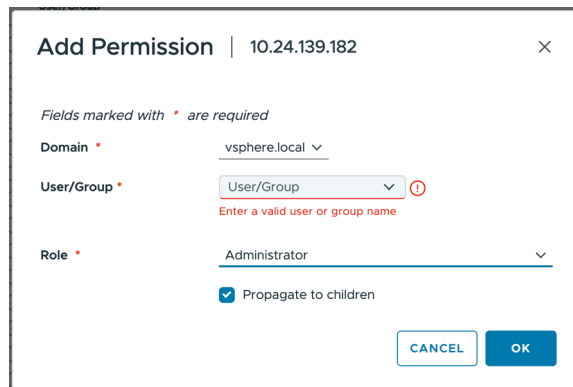


Note: Perform this task only if at least one host in VMware vCenter is running the NVIDIA GPU Management Daemon that uses CIM Service Ticket-based authentication, which was introduced in NVIDIA vGPU software 20.0. If all hosts in VMware vCenter are running an earlier version that does not use CIM Service Ticket-based authentication, omit this task.

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 **New**.
5. In the **New 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). Scroll down and select the **Host** privilege.
 - c). In the right pane of the **New Role** window, click **CIM**.
CIM interaction is automatically selected.
 - d). Click **CREATE**.



6. From the vSphere Web Client home page, go to **Inventory**, select your vCenter Server instance, and click the **Permissions** tab.
7. In the **Users and Groups** section, select the users and groups that will use the NVIDIA vGPU adapter.
 - a). Click **Add**.
 - b). In the **Add Permission** window that opens, select the required users and groups.
 - c). From the **Role** drop-down list, select the role that you created and set the **Propagate to children** option.
 - d). Click **OK**.



See also the following topics in the VMware vSphere documentation:

- ▶ [Using Roles to Assign Privileges](#)
- ▶ [Log In to vCenter Server by Using the vSphere Client](#)
- ▶ [Create a vCenter Server Custom Role](#)
- ▶ [Host CIM Privileges](#)
- ▶ [vSphere Permissions and User Management Tasks](#)

2.5. Adding a Class to the VMware Aria Operations Logs

The NVIDIA vGPU Adapter gathers data only for the classes that are added to the VMware Aria Operations logs. You must set the logging level for a class when you add the class to the VMware Aria Operations logs.

You must add the following classes to the VMware Aria Operations logs:

- ▶ The classes that are required for debugging



Note: In this release of NVIDIA Virtual GPU Management Pack for VMware Aria Operations, **no** required classes for debugging are added to the logs by default. In some earlier releases, required classes for debugging were added to the logs by default with a default logging level.

- ▶ Any other class for which you want to gather data

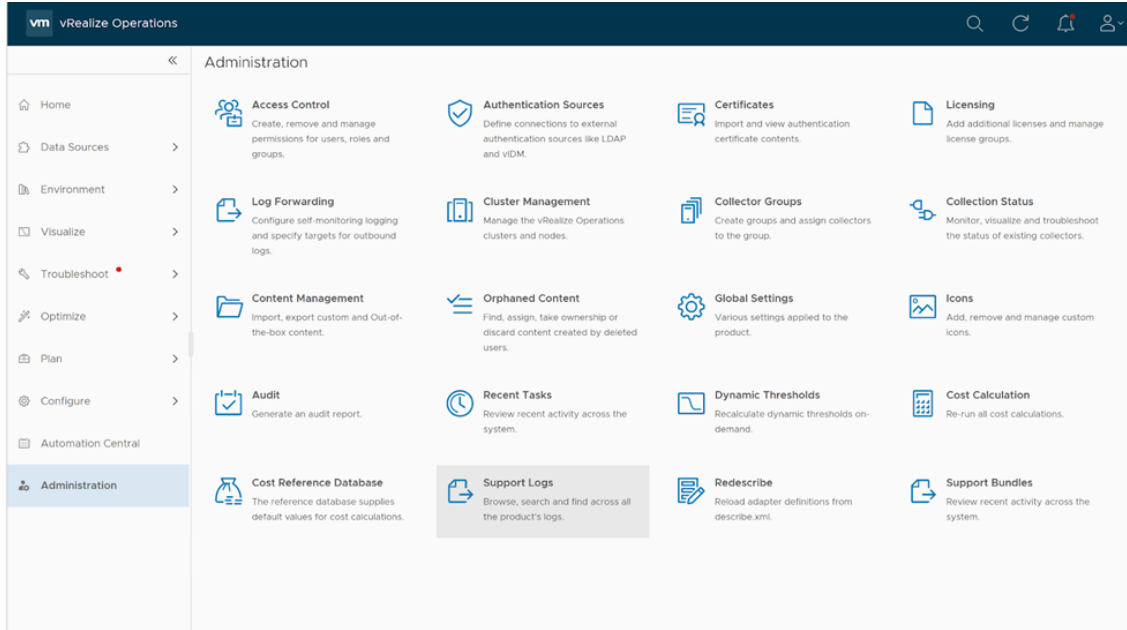
The following classes are required for debugging:

- ▶ `com.nvidia.nvvgpu.adapter.NvVGPUAdapter`
- ▶ `com.nvidia.nvvgpu.adapter.client.VropsInterface`
- ▶ `com.nvidia.nvvgpu.adapter.client.VsphereInterface`
- ▶ `com.nvidia.nvvgpu.adapter.client.Dcgm2xClient`
- ▶ `com.nvidia.nvvgpu.adapter.client.DsdkClient`
- ▶ `com.nvidia.nvvgpu.adapter.client.SrestClient`
- ▶ `com.nvidia.nvvgpu.adapter.util.JsonResponseUtil`

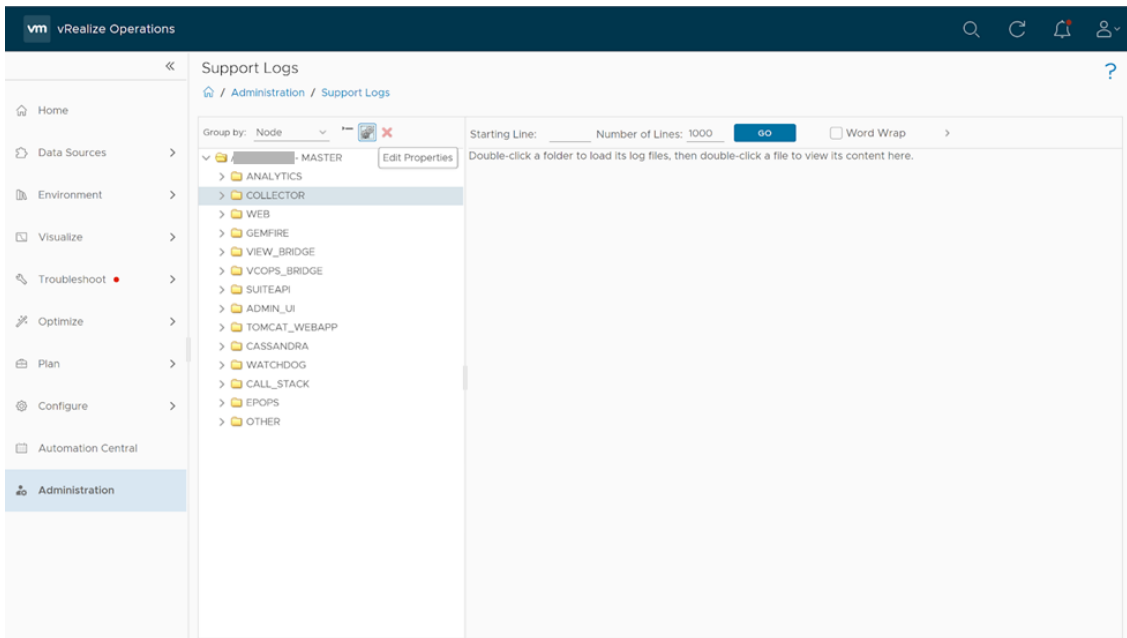
1. If you are not already logged in, log in to your **vRealize Operations Manager** instance as an administrator user.
2. Open the list of VMware Aria Operations log folders.

How to open the list of VMware Aria Operations log folders depends on your VMware Aria Operations release.

- a). On the **vRealize Operations Manager Home** page, click **Administration** and in the **Administration** page that opens, select **Support Logs**.



b). On the **Support Logs** page that opens, expand the list of subfolders under the **vRealize Operations Manager** node.



3. In the expanded list of subfolders, select **COLLECTOR** and click **Edit Properties**.
4. On the **Edit Logger Configuration** page that opens, click the plus sign that represents the **Add Log Class** button.
5. In the **Add Log Class** wizard that starts, specify the class that you want to add and set the logging level.

Chapter 3. Managing Metrics and Analytics for NVIDIA vGPU Software in VMware Aria Operations

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

The metrics are listed in [Metrics Presented by NVIDIA Virtual GPU Management Pack for VMware Aria Operations](#).

3.1. Viewing Data on NVIDIA Dashboards

After installing and configuring NVIDIA Virtual GPU Management Pack for VMware Aria 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 MIG GPU Performance Summary**
- ▶ **NVIDIA MIG vGPU Performance Summary**

► NVIDIA Application Summary



Note: The **NVIDIA MIG GPU Performance Summary** and **NVIDIA MIG vGPU Performance Summary** dashboards provide information for GPUs that are based on the NVIDIA Hopper architecture and later architectures.

3.2. Changing the NVIDIA vGPU Adapter Collection Interval

The collection interval is the length of time between the end of one metrics collection cycle and the start of the next metrics collection cycle. For example, if the collection interval is set to one minute, the NVIDIA vGPU waits for one minute after completing a collection cycle before starting the next collection cycle. The default collection interval is 10 minutes.



Note: The time between successive updates to the collected metrics is longer than the collection interval because the time between updates also includes the time taken to complete a collection cycle. For example, if a collection cycle takes five minutes, and the collection interval is set to one minute, the collected metrics are updated every six 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.



Note: 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**.

The screenshot shows the 'Alert Settings' interface for a symptom named 'GPU : Utilization|Memory Utilization (%)'. At the top, there is a dropdown menu labeled 'Static Threshold'. Below this, the symptom definition is displayed as: 'GPU Memory Utilization is moderate' is 'Warning' when metric 'is greater than or equal' to '75'. There are small up and down arrows next to the '75' value to allow for adjustment. At the bottom left of the form, there is a link labeled 'Advanced' with a right-pointing arrow.

Appendix A. NVIDIA vGPU Alert Definitions

The management pack provides alert definitions for the NVIDIA vGPU metrics and analytics that it integrates with VMware Aria 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

Appendix B. Metrics Presented by NVIDIA Virtual GPU Management Pack for VMware Aria Operations

B.1. GPU Metrics

GPU Information

- ▶ GPU Name
- ▶ GPU UUID
- ▶ GPU Serial Number
- ▶ PCIe Bus Address
- ▶ PCI Device ID
- ▶ PCI Subsystem ID
- ▶ BAR1 Size (GB)
- ▶ Frame Buffer Size (GB)
- ▶ Host Driver Version
- ▶ MIG Mode
- ▶ GPM Support

Temperature

- ▶ Current Temperature (Celsius)
- ▶ Shutdown Temperature (Celsius)
- ▶ Slowdown Temperature (Celsius)
- ▶ GPU Temperature Approaching Shutdown Temperature

Encoder Statistics

- ▶ Active Encoder Session count
- ▶ Trailing Avg Encoder FPS
- ▶ Trailing Avg Encoder Latency (#s)

Framebuffer Capture (FBC) Statistics

- ▶ Active FBC Session count
- ▶ Trailing Avg FBC FPS
- ▶ Trailing Avg FBC Latency (#s)

Utilization

- ▶ 3D/Compute Utilization (%)
- ▶ Encoder Utilization (%)
- ▶ Decoder Utilization (%)
- ▶ Frame Buffer Usage (MB)
- ▶ Memory Utilization (%)

vGPU Information

- ▶ Supported vGPU Types
- ▶ Creatable vGPU Types
- ▶ Number of vGPUs active
- ▶ Array of vGPU Instance IDs

B.2. NVIDIA vGPU Metrics

NVIDIA vGPU Summary

- ▶ vGPU Name
- ▶ vGPU Type ID
- ▶ vGPU Instance ID
- ▶ PCIe Bus Address
- ▶ Frame Buffer Size (GB)
- ▶ License Status
- ▶ UUID

Encoder Statistics

- ▶ Active Encoder Session count
- ▶ Trailing Avg Encoder FPS
- ▶ Trailing Avg Encoder Latency (#s)

Frame Buffer Capture (FBC) Statistics

- ▶ Active FBC Session count
- ▶ Trailing Avg FBC FPS
- ▶ Trailing Avg FBC Latency (#s)

Utilization

- ▶ 3D/Compute Utilization (%)
- ▶ Encoder Utilization (%)
- ▶ Decoder Utilization (%)
- ▶ Frame Buffer Usage (MB)
- ▶ Memory Utilization (%)

MIG Information

- ▶ GPU instance information:
 - ▶ GPU Instance ID
 - ▶ GPU Instance Name
- ▶ GPU instance profile information:
 - ▶ GPU Instance Profile ID
 - ▶ Peer-to-Peer Support
 - ▶ GPU Instance Slice Count
 - ▶ Memory Size
 - ▶ SM (Multiprocessor) Count
 - ▶ Copy Engine Count
 - ▶ Decoder Count
 - ▶ Encoder Count
 - ▶ JPEG Count
 - ▶ OFA Count
- ▶ Compute instance information:
 - ▶ Compute Instance ID
 - ▶ Compute Instance Name

- ▶ Compute Instance Profile Information
 - ▶ Multiprocessor count
 - ▶ Shared Copy Engine count
 - ▶ Shared Decoder count
 - ▶ Shared Encoder count
 - ▶ Shared JPEG count
 - ▶ Shared OFA count
 - ▶ Compute Instance slice count
 - ▶ Compute Instance Alive

GPM Information

- ▶ GPU instance information:
 - ▶ Graphics Activity
 - ▶ SM Activity
 - ▶ SM Occupancy
 - ▶ Integer Activity
 - ▶ Tensor Activity
 - ▶ DFMA Tensor Activity
 - ▶ HMMA Tensor Activity
 - ▶ IMMA Tensor Activity
 - ▶ DRAM Activity
 - ▶ FP64 Activity
 - ▶ FP32 Activity
 - ▶ FP16 Activity
- ▶ Compute instance information:
 - ▶ Graphics Activity
 - ▶ SM Activity
 - ▶ SM Occupancy
 - ▶ Integer Activity
 - ▶ Tensor Activity
 - ▶ DFMA Tensor Activity
 - ▶ HMMA Tensor Activity
 - ▶ IMMA Tensor Activity
 - ▶ FP64 Activity
 - ▶ FP32 Activity

- ▶ FP16 Activity
- ▶ Compute Instance Active

B.3. Process Metrics

- ▶ Process ID
- ▶ Process Name
- ▶ 3D/Compute Utilization (%)
- ▶ Encoder Utilization (%)
- ▶ Decoder Utilization (%)
- ▶ Frame Buffer Usage (MB)
- ▶ Memory Utilization (%)

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 NON-INFRINGEMENT, 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

© 2017-2026 NVIDIA Corporation. All rights reserved.

