NVIDIA

VIRTUAL GPU SOFTWARE

QSG-07847-001 _v9.0 through 9.2 | November 2019

Quick Start Guide

TABLE OF CONTENTS

About this Guide.....iii
Chapter 1. Getting Your NVIDIA vGPU Software..... 1
1.1. Before You Begin.....1
1.2. Your Order Confirmation Message.....1
1.3. Creating Your NVIDIA Enterprise Account.....5
1.4. Redeeming Your Product Activation Keys and Downloading Your NVIDIA vGPU Software.....8
Chapter 2. Installing Your NVIDIA vGPU Software License Server and License Files..... 15
2.1. Overview of NVIDIA vGPU software Licensing..... 15
2.2. Platform Configuration Requirements..... 16
2.3. Installing the Java Runtime Environment..... 16
2.4. Installing the License Server Software..... 18
2.5. Obtaining the License Server's MAC Address..... 22
2.6. Registering Your License Server and Getting License Files..... 23
2.7. Installing a License..... 25
Chapter 3. Installing and Configuring NVIDIA vGPU Manager and the Guest Driver..... 27
3.1. Switching the Mode of a Tesla M60 or M6 GPU..... 27
3.2. Installing the NVIDIA Virtual GPU Manager.....28
3.2.1. Installing the NVIDIA Virtual GPU Manager on VMware vSphere..... 28
3.2.2. Installing the NVIDIA Virtual GPU Manager on Citrix Hypervisor.....28
3.3. Disabling and Enabling ECC Memory..... 29
3.3.1. Disabling ECC Memory.....30
3.3.2. Enabling ECC Memory..... 31
3.4. Attaching an NVIDIA vGPU Profile to a VM.....32
3.4.1. Configuring a vSphere VM with NVIDIA vGPU.....32
3.4.2. Configuring a Citrix Hypervisor VM with Virtual GPU..... 34
3.5. Installing the NVIDIA vGPU Software Graphics Driver..... 34
3.6. Licensing an NVIDIA vGPU..... 36

ABOUT THIS GUIDE

Virtual GPU Software Quick Start Guide provides minimal instructions for installing and configuring NVIDIA® virtual GPU software on the Citrix Hypervisor or VMware vSphere hypervisor and for installing and configuring a single-instance NVIDIA virtual GPU License Server on Windows. If you need complete instructions, are using other platforms, or require high availability for the license server, refer to *Virtual GPU Software User Guide* and *Virtual GPU License Server User Guide*.

Chapter 1.
GETTING YOUR NVIDIA VGPU SOFTWARE

After your order for NVIDIA vGPU software is processed, you will receive an order confirmation message from NVIDIA. This message contains information that you need for getting your NVIDIA vGPU software from the NVIDIA Software Licensing Center. To log in to the NVIDIA Software Licensing Center, you must have an NVIDIA Enterprise Account.

1.1. Before You Begin

Before following the procedures in this guide, ensure that the following prerequisites are met:

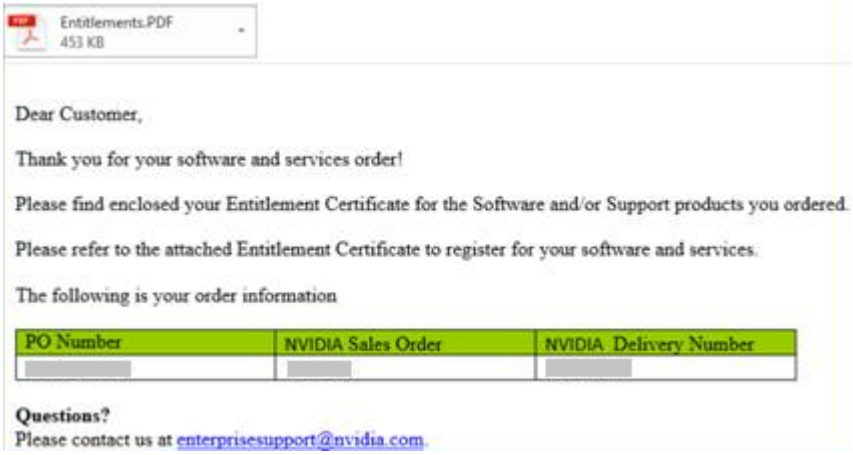
- You have a server platform that is capable of hosting your chosen hypervisor and NVIDIA GPUs that support NVIDIA vGPU software. For a list of validated server platforms, refer to NVIDIA GRID Certified Servers.
- One or more NVIDIA GPUs that support NVIDIA vGPU software is installed in your server platform.
- A supported virtualization software stack is installed according to the instructions in the software vendor's documentation.
- A virtual machine (VM) running a supported Windows guest operating system (OS) is configured in your chosen hypervisor.

For information about supported hardware and software, and any known issues for this release of NVIDIA vGPU software software, refer to the *Release Notes* for your chosen hypervisor:

- *Virtual GPU Software for Citrix Hypervisor Release Notes*
- *Virtual GPU Software for VMware vSphere Release Notes*

1.2. Your Order Confirmation Message

After your order for NVIDIA vGPU software is processed, you will receive an order confirmation message to which your NVIDIA Entitlement Certificate is attached.



Entitlements.PDF
453 KB

Dear Customer,

Thank you for your software and services order!

Please find enclosed your Entitlement Certificate for the Software and/or Support products you ordered.

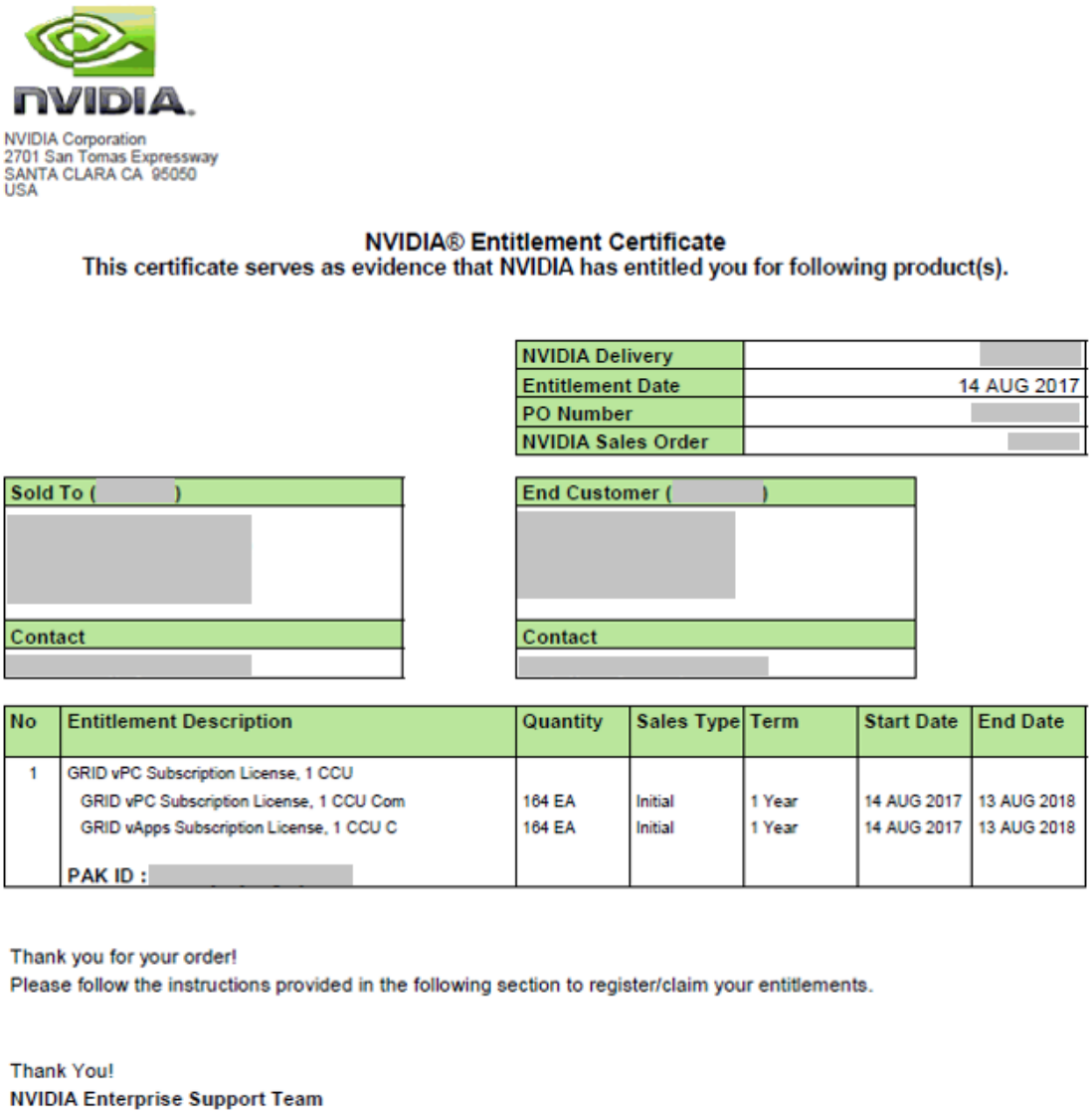
Please refer to the attached Entitlement Certificate to register for your software and services.

The following is your order information

PO Number	NVIDIA Sales Order	NVIDIA Delivery Number

Questions?
Please contact us at enterprisesupport@nvidia.com.

Your NVIDIA Entitlement Certificate contains your product activation keys.



NVIDIA Corporation
2701 San Tomas Expressway
SANTA CLARA CA 95050
USA

NVIDIA® Entitlement Certificate
This certificate serves as evidence that NVIDIA has entitled you for following product(s).

NVIDIA Delivery	
Entitlement Date	14 AUG 2017
PO Number	
NVIDIA Sales Order	

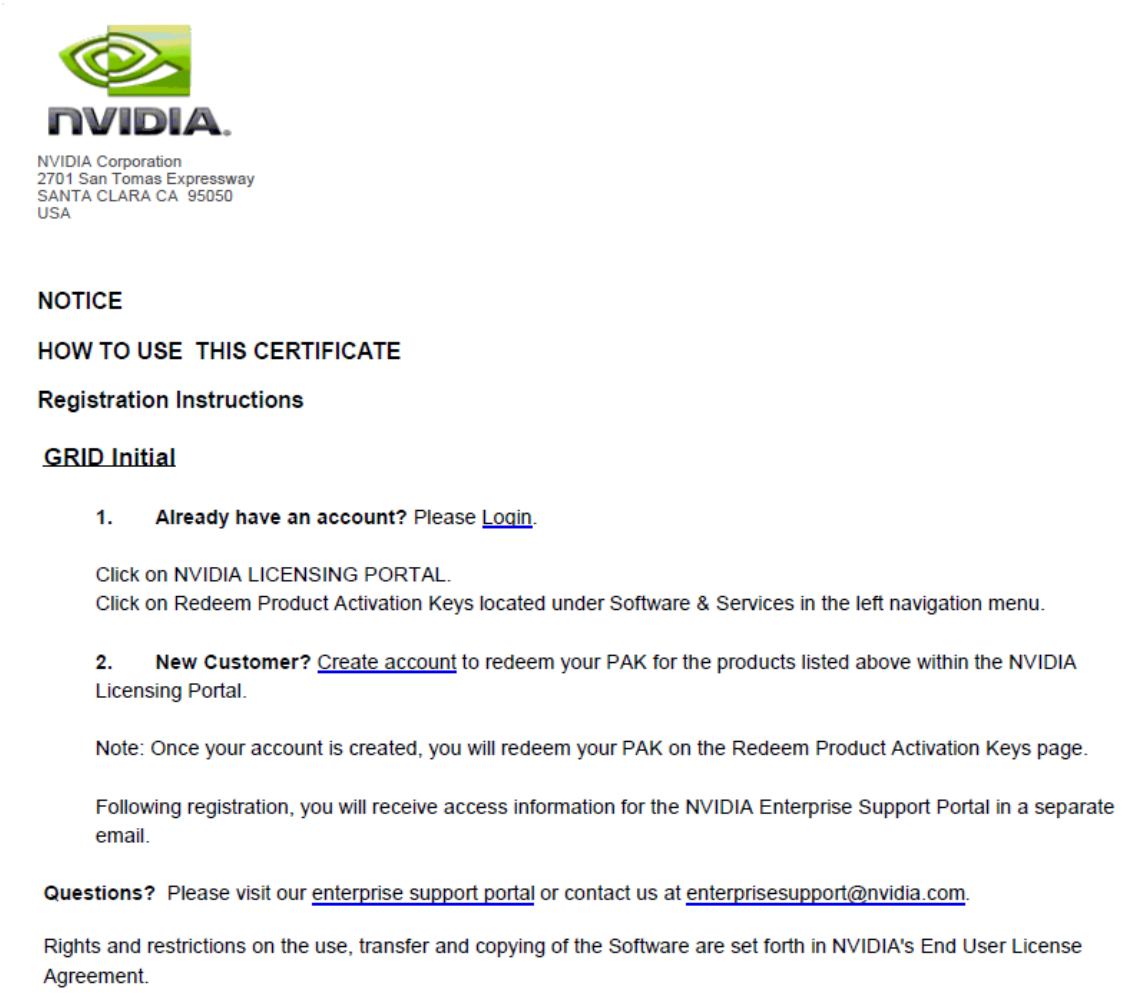
Sold To ( )	End Customer ( )
Contact	Contact

No	Entitlement Description	Quantity	Sales Type	Term	Start Date	End Date
1	GRID vPC Subscription License, 1 CCU					
	GRID vPC Subscription License, 1 CCU Com	164 EA	Initial	1 Year	14 AUG 2017	13 AUG 2018
	GRID vApps Subscription License, 1 CCU C	164 EA	Initial	1 Year	14 AUG 2017	13 AUG 2018
	PAK ID :					

Thank you for your order!
Please follow the instructions provided in the following section to register/claim your entitlements.

Thank You!
NVIDIA Enterprise Support Team

Your NVIDIA Entitlement Certificate also provides instructions for using the certificate.



NVIDIA Corporation
2701 San Tomas Expressway
SANTA CLARA CA 95050
USA

NOTICE

HOW TO USE THIS CERTIFICATE

Registration Instructions

GRID Initial

1. **Already have an account?** Please Login.

   Click on NVIDIA LICENSING PORTAL.
   Click on Redeem Product Activation Keys located under Software & Services in the left navigation menu.

2. **New Customer?** Create account to redeem your PAK for the products listed above within the NVIDIA Licensing Portal.

   Note: Once your account is created, you will redeem your PAK on the Redeem Product Activation Keys page.

   Following registration, you will receive access information for the NVIDIA Enterprise Support Portal in a separate email.

Questions? Please visit our enterprise support portal or contact us at enterprisesupport@nvidia.com.

Rights and restrictions on the use, transfer and copying of the Software are set forth in NVIDIA's End User License Agreement.

To redeem the product activation keys for your NVIDIA vGPU software, you must have an NVIDIA Enterprise Account.

For a Support, Upgrade, and Maintenance Subscription (SUMS) renewal, no action is required. You do not need to redeem the product activation keys for a SUMS renewal.

- If you do not have an account, follow the **Register** link in the instructions for using the certificate to create your account. For details, see Creating Your NVIDIA Enterprise Account.
- If you already have an account, follow the **Login** link in the instructions for using the certificate to log in to the NVIDIA Enterprise Application Hub, go to the NVIDIA Software Licensing Center, and redeem your product activation keys. For details, see Redeeming Your Product Activation Keys and Downloading Your NVIDIA vGPU Software.

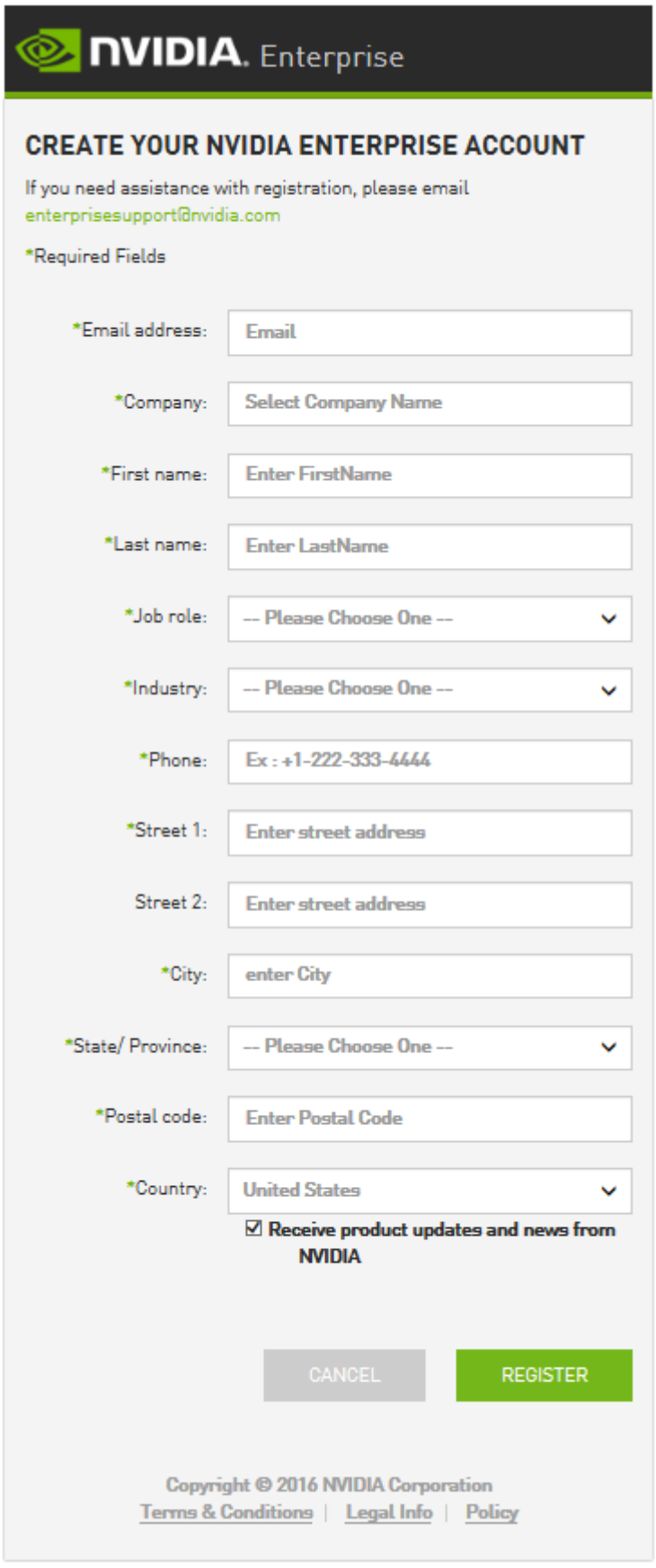
1.3. Creating Your NVIDIA Enterprise Account

If you do not have an NVIDIA Enterprise Account, you must create an account to be able to log in to the NVIDIA Software Licensing Center.

If you already have an account, skip this task and go to Redeeming Your Product Activation Keys and Downloading Your NVIDIA vGPU Software.

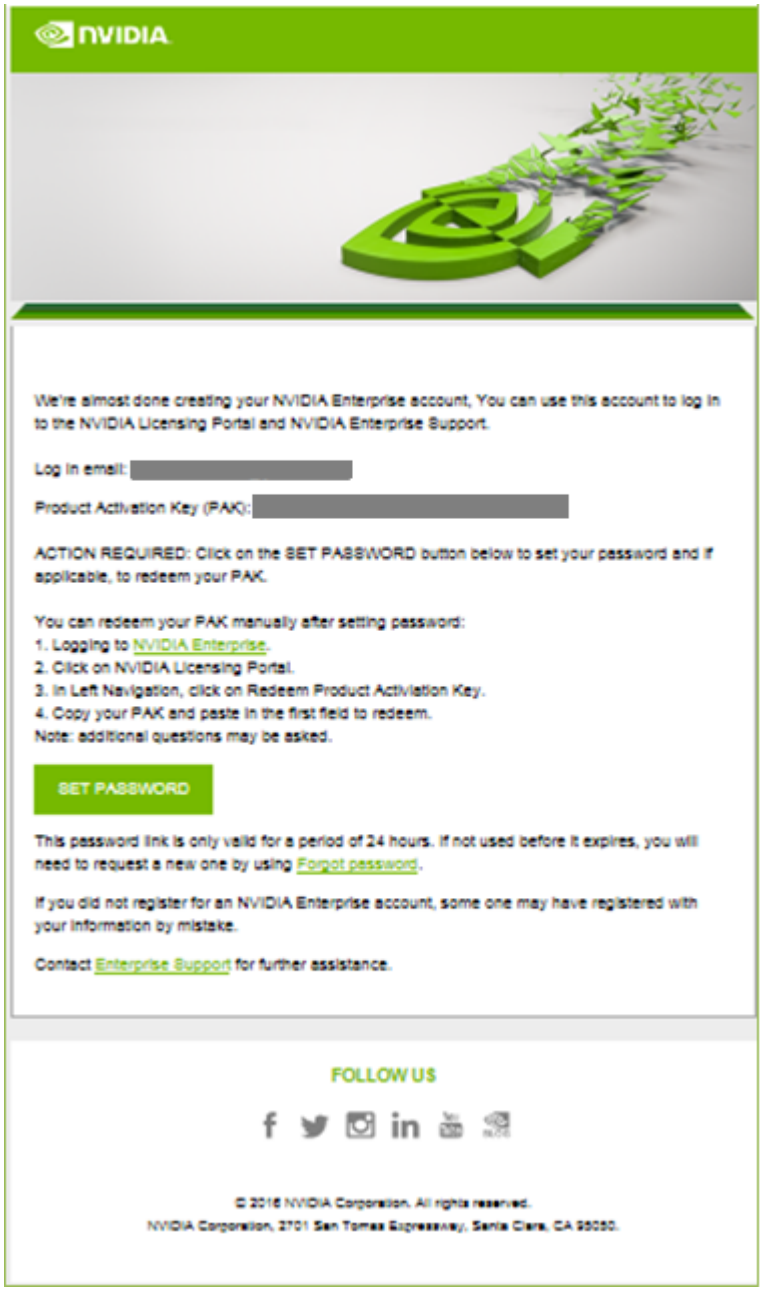
Before you begin, ensure that you have your order confirmation message.

1. In the instructions for using your NVIDIA Entitlement Certificate, follow the **Register** link.
2. Fill out the form on the **Create Your NVIDIA Enterprise Account** page and click **Register**.


NVIDIA. Enterprise
CREATE YOUR NVIDIA ENTERPRISE ACCOUNT
If you need assistance with registration, please email enterprisesupport@nvidia.com
*Required Fields
*Email address:
Email
*Company:
Select Company Name
*First name:
Enter FirstName
*Last name:
Enter LastName
*Job role:
-- Please Choose One --
*Industry:
-- Please Choose One --
*Phone:
Ex : +1-222-333-4444
*Street 1:
Enter street address
Street 2:
Enter street address
*City:
enter City
*State/ Province:
-- Please Choose One --
*Postal code:
Enter Postal Code
*Country:
United States
Receive product updates and news from NVIDIA
CANCEL
REGISTER
Copyright © 2016 NVIDIA Corporation
Terms & Conditions | Legal Info | Policy

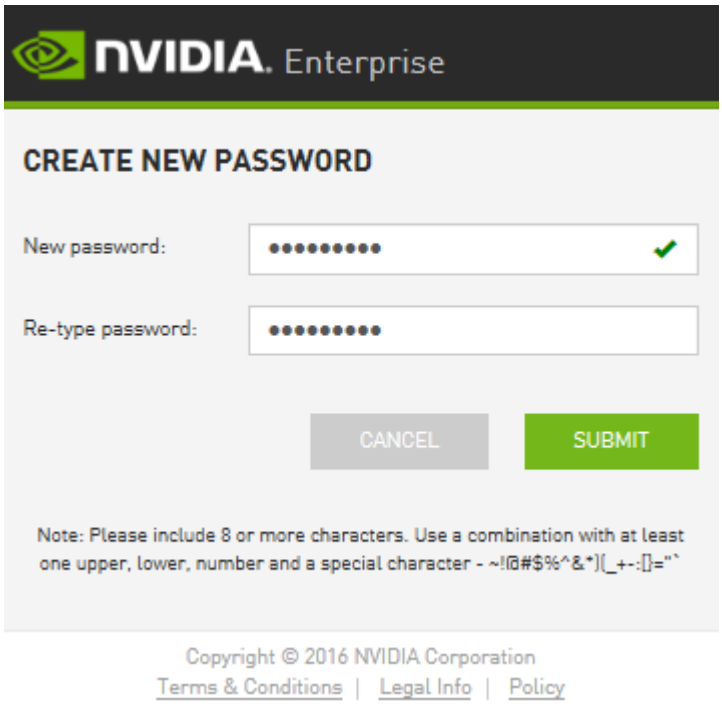
A message confirming that an account has been created appears and an e-mail instructing you to set your NVIDIA password is sent to the e-mail address you provided.

3. Open the e-mail instructing you to set your password and click **SET PASSWORD**.

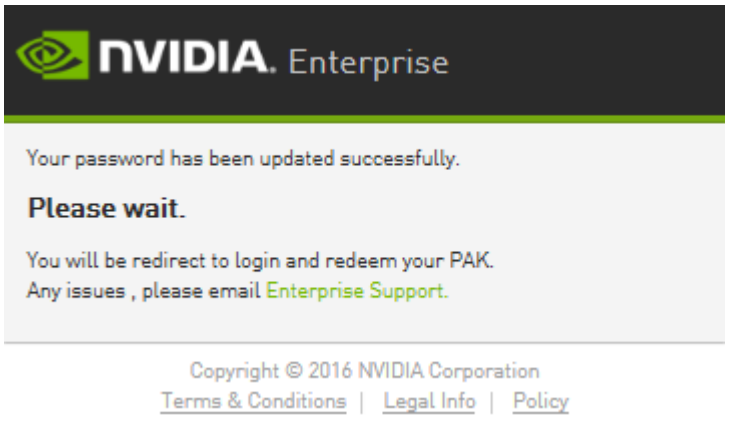


For your account security, the **SET PASSWORD** link in this e-mail is set to expire in 24 hours.

4. Enter and re-enter your new password, and click **SUBMIT**.



A message confirming that your password has been set successfully appears.

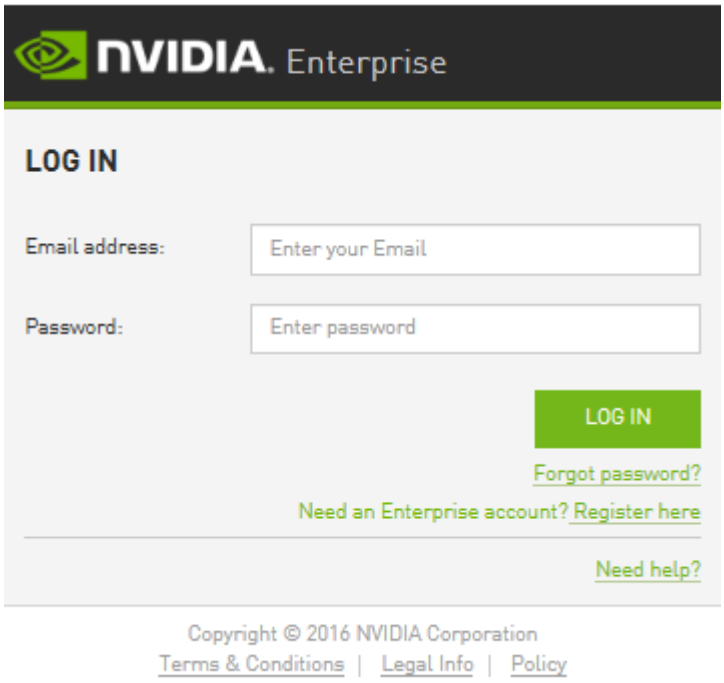


You are then automatically directed to log in to the NVIDIA Software Licensing Center with your new password.

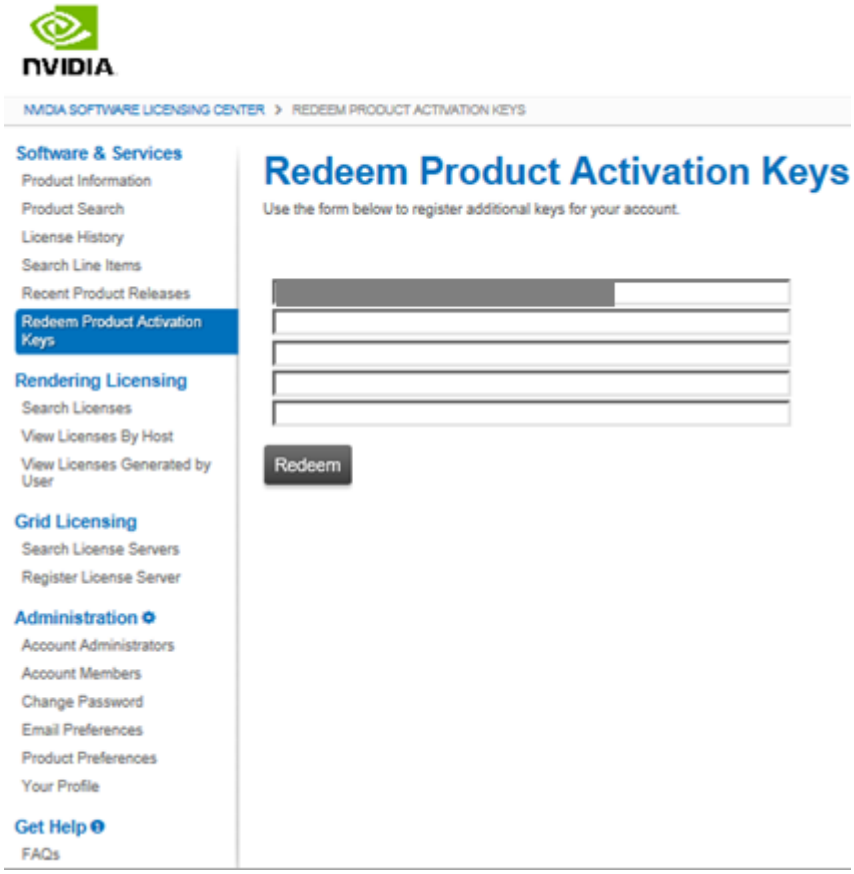
1.4. Redeeming Your Product Activation Keys and Downloading Your NVIDIA vGPU Software

Before you begin, ensure that you have your order confirmation message and have created an NVIDIA Enterprise Account.

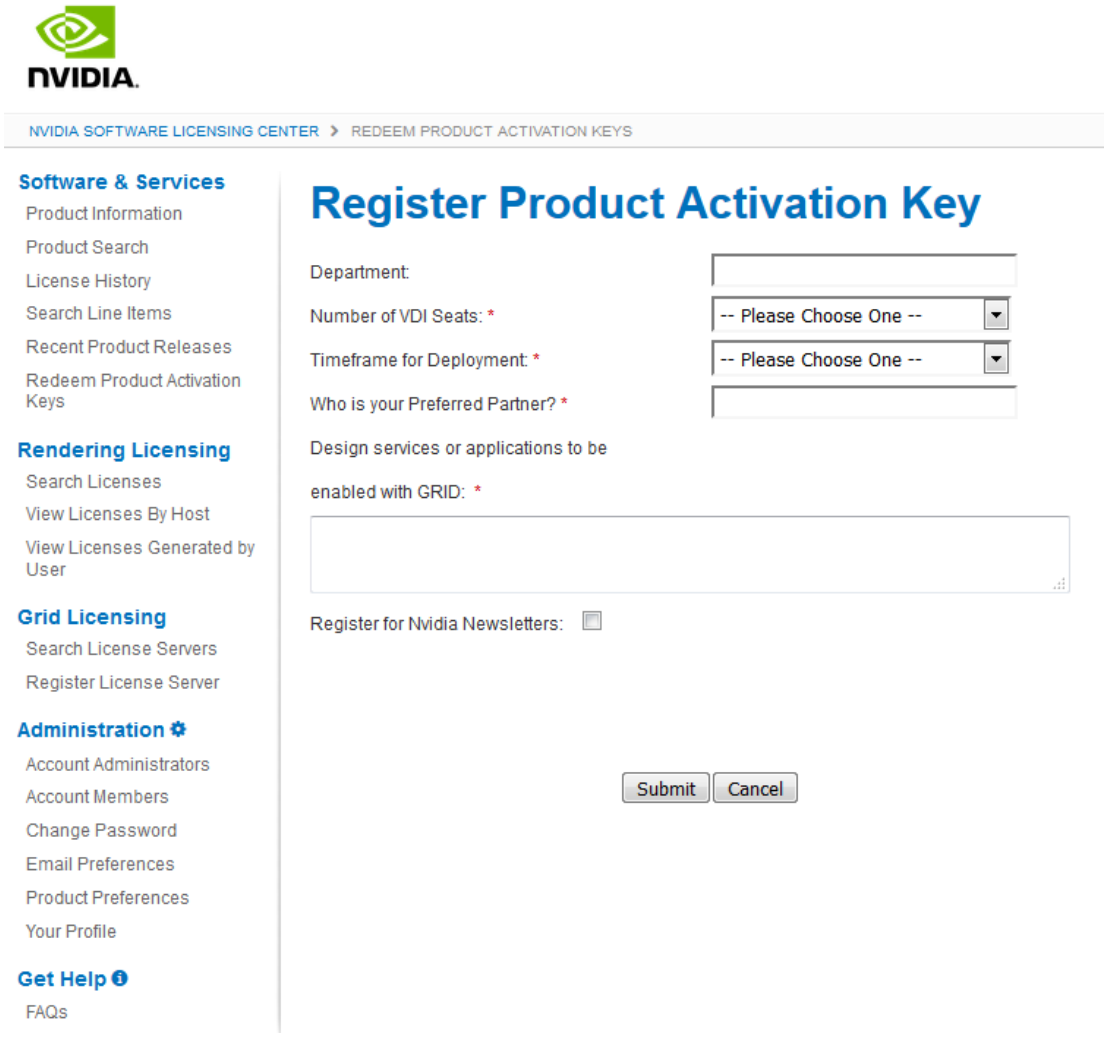
1. Visit the NVIDIA Software Licensing Center by following the **Login** link in the instructions for using your NVIDIA Entitlement Certificate.
2. When prompted, provide your e-mail address and password, and click **LOGIN**.



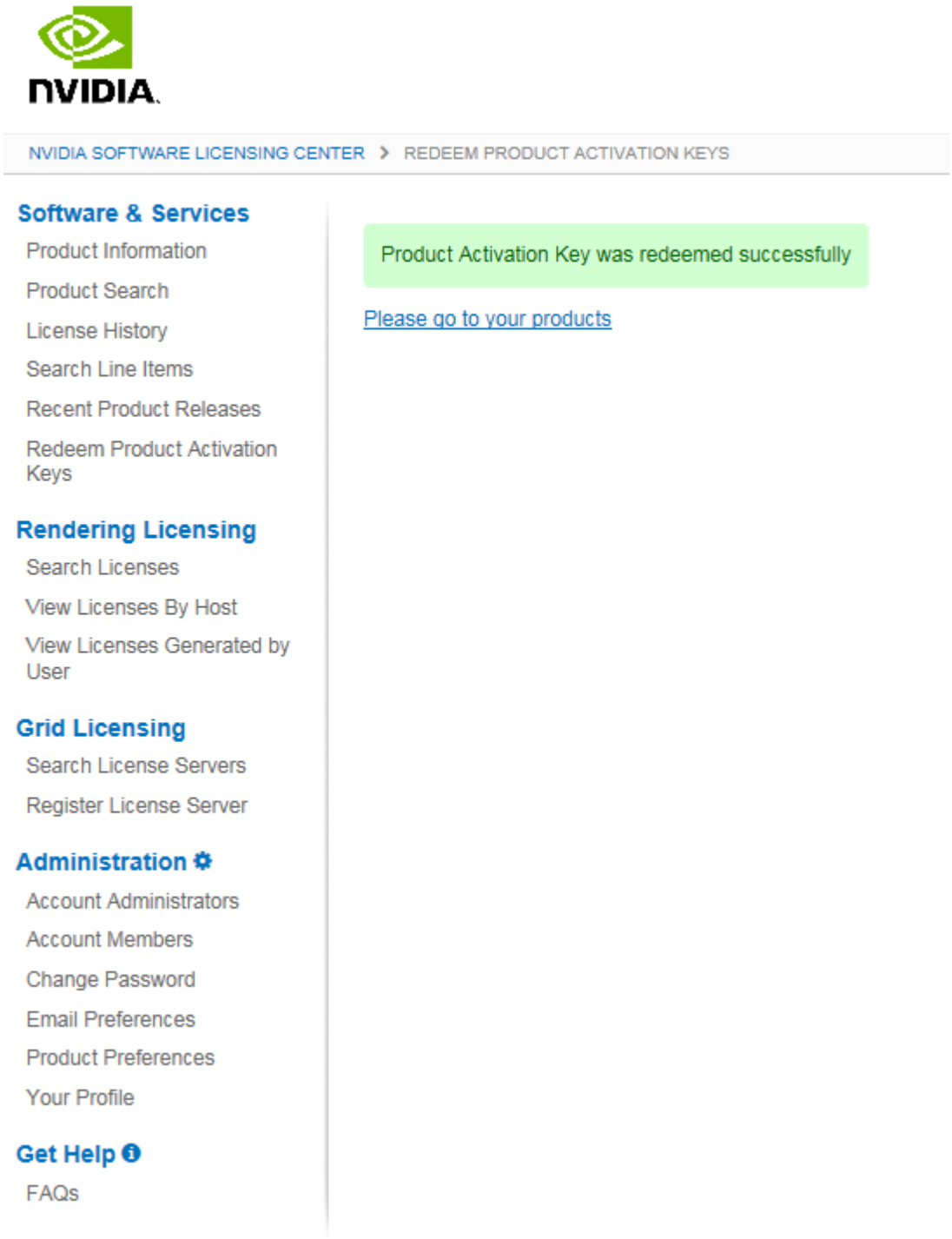
3. On the **Redeem Product Activation Keys** page that opens, do one of the following:
   - If the page is already populated with your product activation keys, click **Redeem**.
   - Otherwise, enter each product activation key in a separate text field and click **Redeem**.



4. On the **Register Product Activation Key** page that opens, provide the requested information about how you intend to use NVIDIA vGPU software and click **Submit**.

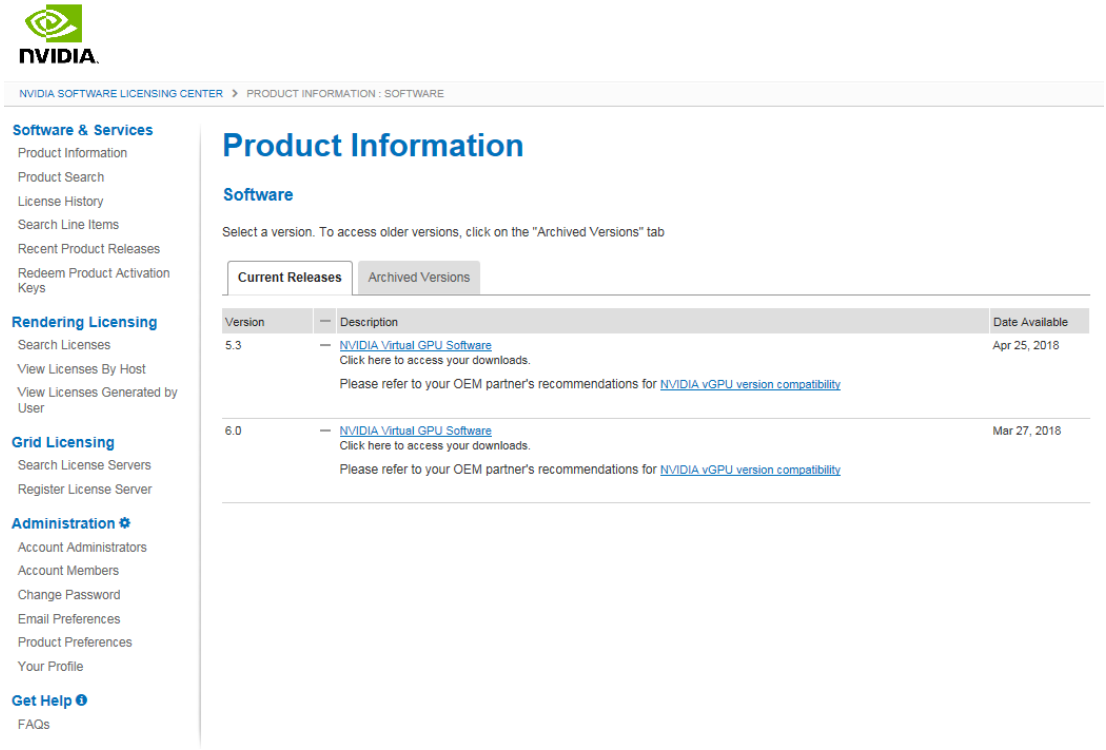


5. On the confirmation page that opens, follow the **Please go to your products** link.



The **Product Information** page opens with the **Current Releases** tab selected. Older versions of NVIDIA vGPU software are available from the **Archived Versions** tab.

6. On the **Product Information** page, follow the **NVIDIA Virtual GPU Software** link.



7. On the **Product Download** page that opens, follow the links to download the software that you require.
   a) Follow the link to the NVIDIA vGPU software download for the brand and version of your chosen hypervisor, for example, **NVIDIA vGPU for vSphere 6.5**.
   b) Follow the link to the NVIDIA vGPU software license server software for your license server host machine's operating system, for example, **License Manager for Windows**.
   c) If you are using an NVIDIA Tesla™ M60 or M6 GPU and think you might need to change its mode, follow the **Mode Change Utility** link.

      For details about when you need to change the mode, see Switching the Mode of a Tesla M60 or M6 GPU.

Product Download

NVIDIA Virtual GPU Software

- Please refer to the NVIDIA vGPU Quick Start Guide for download instructions. More NVIDIA vGPU resources are available here.
- Download appropriate version of the NVIDIA vGPU software below. You can refer to the Product Support Matrix to determine which software package is right for you.

NVIDIA vGPU for vSphere 6.7 / NVIDIA vGPU for vSphere 6.5 / NVIDIA vGPU for vSphere 6.0 / NVIDIA vGPU for XenServer 7.4 / NVIDIA vGPU for XenServer 7.3 / NVIDIA vGPU for XenServer 7.2 / NVIDIA vGPU for XenServer 7.1 / NVIDIA vGPU for XenServer 7.0 / NVIDIA vGPU for Windows / NVIDIA vGPU for RHEL KVM / NVIDIA vGPU for AHV

- Download the GPU mode change utility here: Mode Change Utility
- Download the NVIDIA License Manager here: License Manager for Windows / License Manager for Linux
- Download the NVIDIA management pack for vRealize Operations: NVIDIA Virtual GPU Management Pack for vRealize Operations
- After you install the license server, read the MAC ID displayed on it. Use this MAC ID to map required number of licenses and download license file by clicking on "Register License Server" in the left side panel.

Thank you for your NVIDIA Enterprise software product purchase.

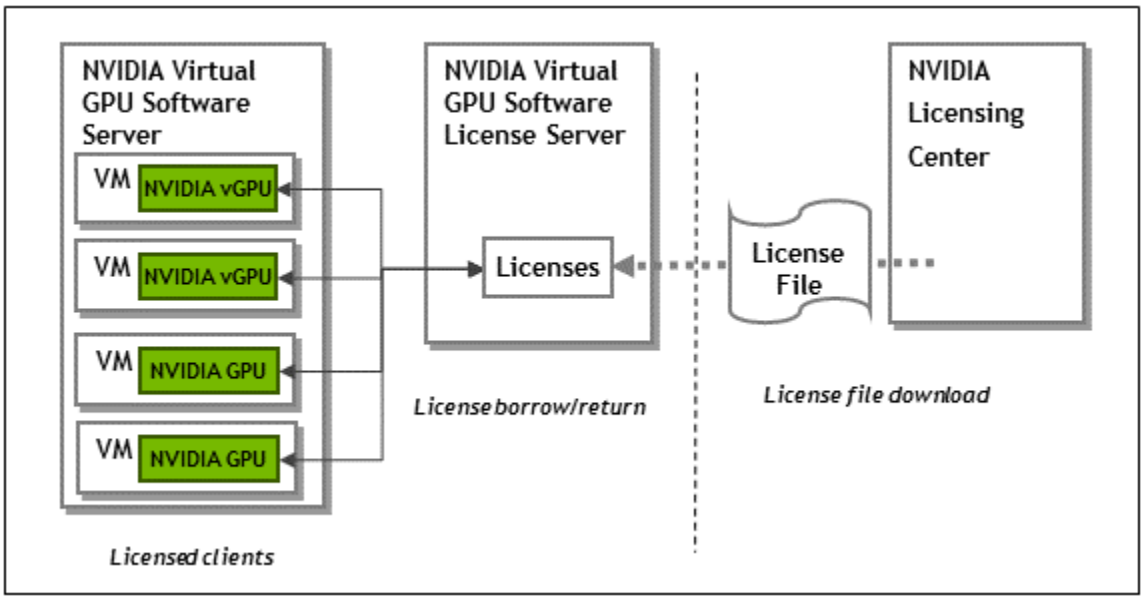
Chapter 2. INSTALLING YOUR NVIDIA VGPU SOFTWARE LICENSE SERVER AND LICENSE FILES

The NVIDIA vGPU software License Server is used to serve a pool of floating licenses to NVIDIA vGPU software licensed products. The license server is designed to be installed at a location that is accessible from a customer's network, and be configured with licenses obtained from the NVIDIA Software Licensing Center.

These instructions cover only the installation and configuration of a single-instance license server. If you require high availability for the license server, refer to *Virtual GPU License Server User Guide* for instrictions for how to configure a primary and secondary license server. If you do not register a secondary server when the primary server is registered, you **cannot** register the secondary server later.

2.1. Overview of NVIDIA vGPU software Licensing

NVIDIA vGPU software is a licensed product. Licensed vGPU functionalities are activated during guest OS boot by the acquisition of a software license served over the network from an NVIDIA vGPU software license server. The license is returned to the license server when the guest OS shuts down.



The licenses served by the NVIDIA vGPU software license server are obtained from the NVIDIA Software Licensing Center as downloadable license files, and installed into the license server through its management interface.

2.2. Platform Configuration Requirements

- The platform must have a fixed (unchanging) IP address. The IP address may be assigned dynamically by DHCP or statically configured, but must be constant.
- The platform must have at least one unchanging Ethernet MAC address, to be used as a unique identifier when registering the server and generating licenses in the NVIDIA Software Licensing Center.
- The platform's date and time must be set accurately. NTP is recommended.

2.3. Installing the Java Runtime Environment

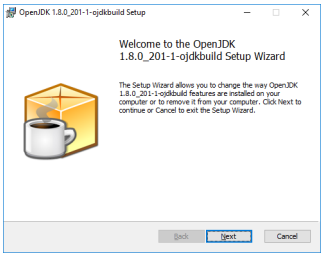
If a suitable Java runtime environment (JRE) version is not already installed on your system, install a supported JRE before running the NVIDIA license server installer.

Install a JRE for the same the processor architecture as the license server that you will install: If you will install a 32-bit license server, install a 32-bit JRE; if you will install a 64-bit license server, install a 64-bit JRE.

1. Download a supported Oracle Java SE JRE or OpenJDK JRE.
   - Download Oracle Java SE JRE from the java.com: Java + You page.
   - Download OpenJDK JRE from the Community builds using source code from OpenJDK project on GitHub.
2. Install the JRE that you downloaded.

- Oracle Java SE JRE installation:

- OpenJDK JRE installation:



3. Set the `JAVA_HOME` system variable to the full path to the `jre` folder of your JRE installation.
   For example:
   - For 64-bit Oracle Java SE JRE: `C:\Program Files\Java\jre1.8.0_191`
   - For 32-bit Oracle Java SE JRE: `C:\Program Files (x86)\Java\jre1.8.0_191`
   - For 64-bit OpenJDK JRE: `C:\Program Files\ojdkbuild\java-1.8.0-openjdk-1.8.0.201-1\jre`

- For 32-bit OpenJDK JRE: `C:\Program Files (x86)\ojdkbuild\java-1.8.0-openjdk-1.8.0.201-1\jre`

Ensure that the path does **not** include any trailing characters, such as a slash or a space.

If you are upgrading to a new version of the JRE, update the value of the `JAVA_HOME` system variable to the full path to the `jre` folder of your **new** JRE version.

4. Ensure that the `Path` system variable contains the path to the `java.exe` executable file.

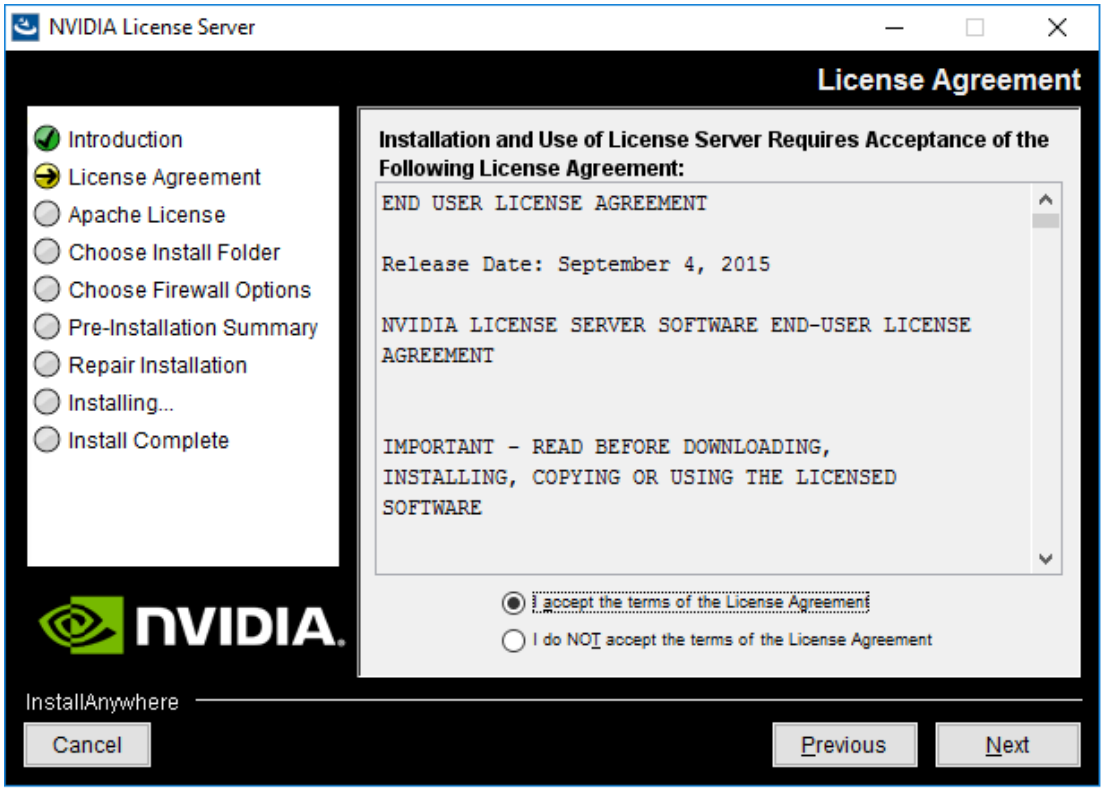
   The installer may have already added this path for you.

   For example:

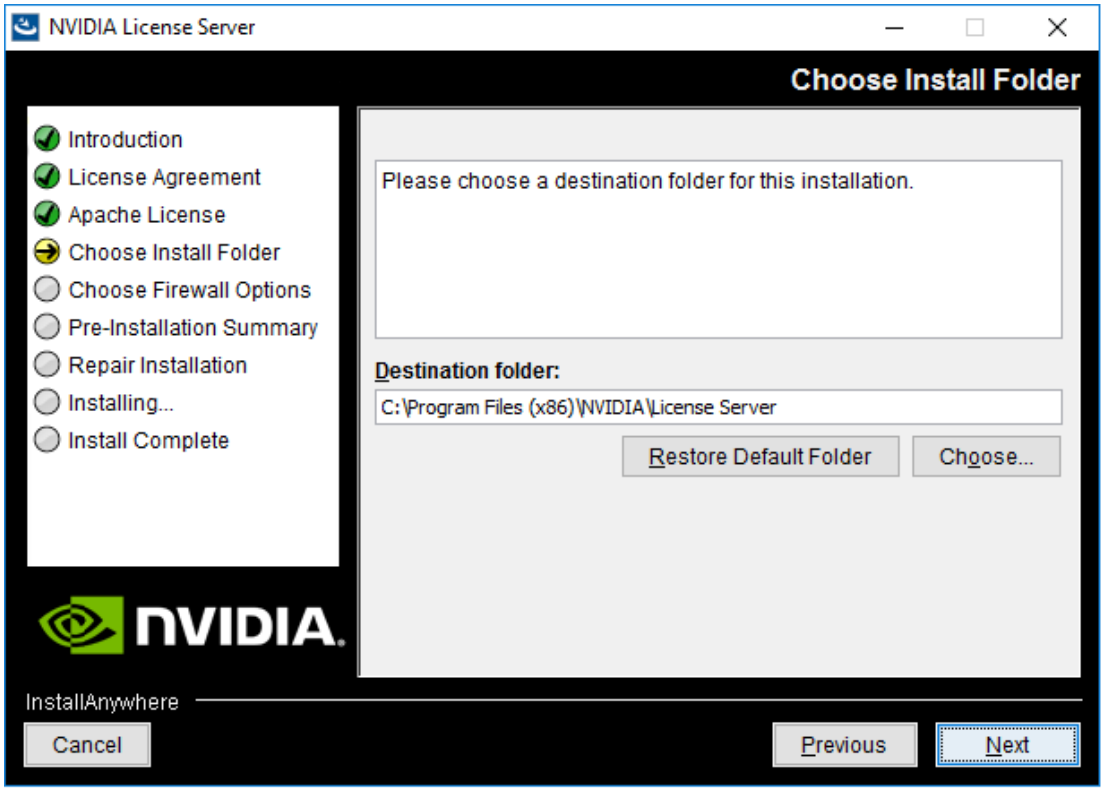
   - For 64-bit Oracle Java SE JRE: `C:\Program Files\Java\jre1.8.0_191\bin`
   - For 32-bit Oracle Java SE JRE: `C:\Program Files (x86)\Java\jre1.8.0_191\bin`
   - For 64-bit OpenJDK JRE: `C:\Program Files\ojdkbuild\java-1.8.0-openjdk-1.8.0.201-1\bin`
   - For 32-bit OpenJDK JRE: `C:\Program Files (x86)\ojdkbuild\java-1.8.0-openjdk-1.8.0.201-1\bin`

2.4. Installing the License Server Software

1. Unzip the license server installer and run `setup.exe`.
2. Accept the EULA for the license server software and the Apache Tomcat software used to support the license server's management interface.



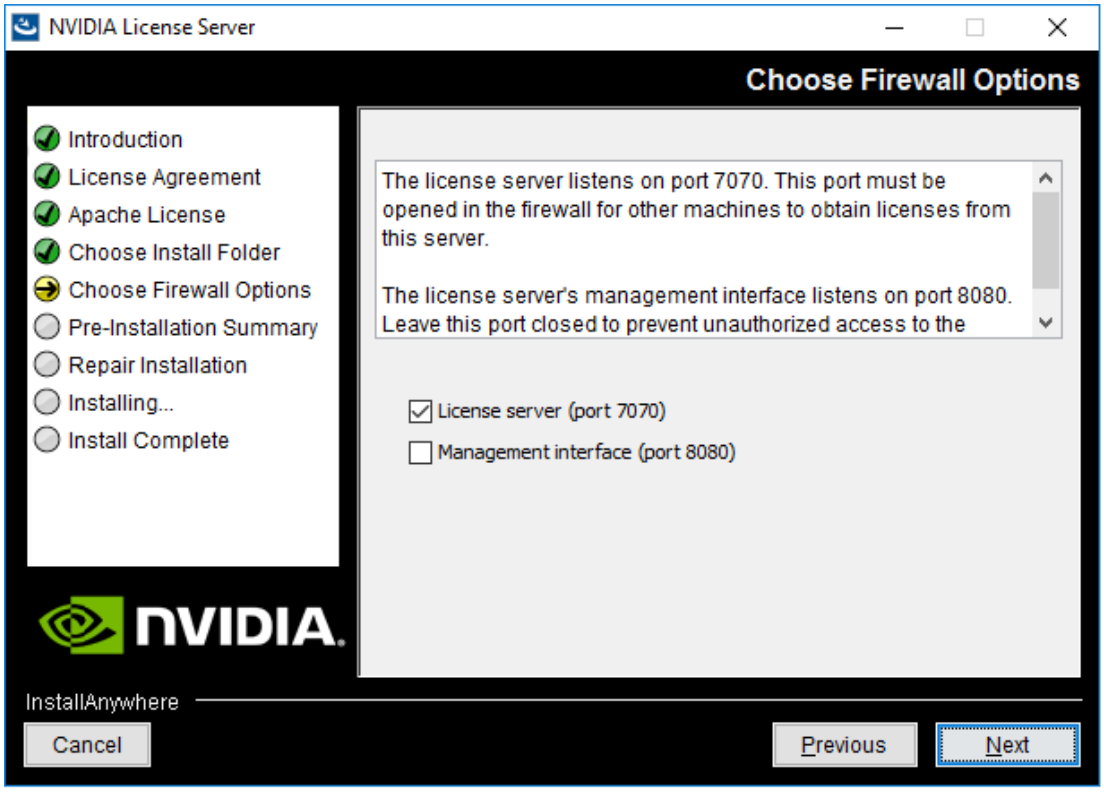
3. Choose the destination folder where you want the license server software to be installed.



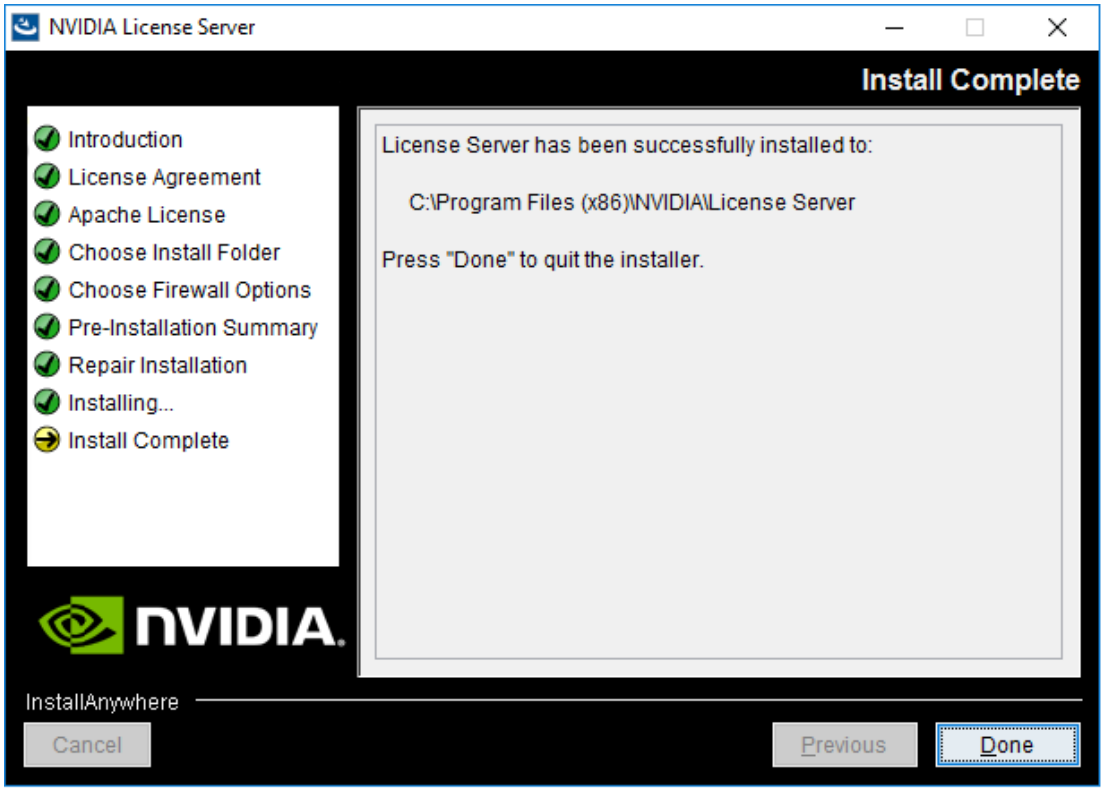
4. In the **Choose Firewall Options** dialog box, select the ports to be opened in the firewall.

   To enable remote clients to access licenses from the server and prevent remote access to the management interface, use the default setting, which sets ports as follows:

   - Port 7070 is open to enable remote clients to access licenses from the server.
   - Port 8080 is closed to ensure that the management interface is available only through a web browser running locally on the license server host.



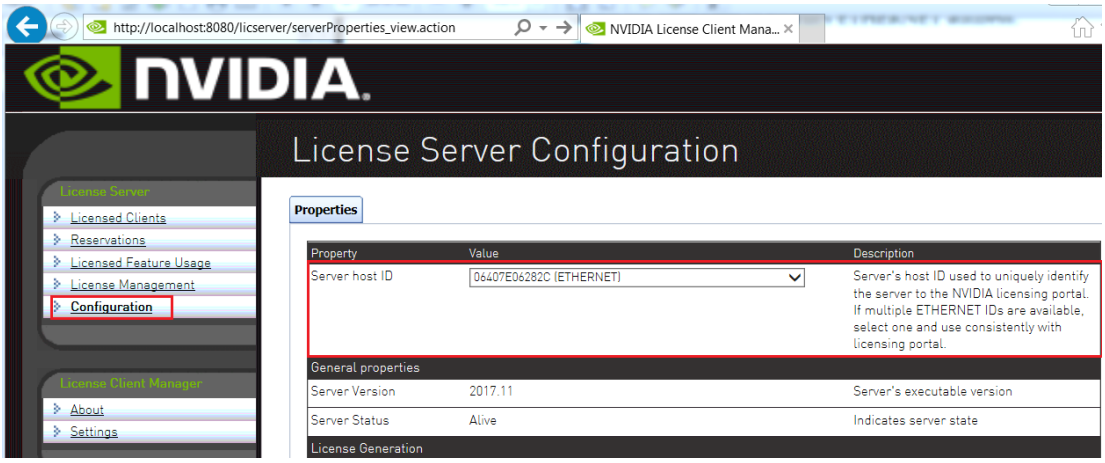
5. After installation has completed successfully, click **Done** to exit the installer.



2.5. Obtaining the License Server's MAC Address

The license server's Ethernet MAC address uniquely identifies your server to the NVIDIA Software Licensing Center. You will need this address to register your license server with the NVIDIA Software Licensing Center to generate license files.

1. Open a web browser on the license server host and connect to the URL `http://localhost:8080/licserver`.
2. In the license server management interface, select **Configuration**.
3. On the **License Server Configuration** page that opens, in the **Server host ID** drop-down list, select the platform's ETHERNET address.



2.6. Registering Your License Server and Getting License Files

To be able to generate and download licenses from the NVIDIA Software Licensing Center, you must register your license server. Before downloading a license file, you must generate the file by mapping to it the add-ons that correspond to your licenses. The license file is empty by default and does not contain any licenses it until you map add-ons to it.

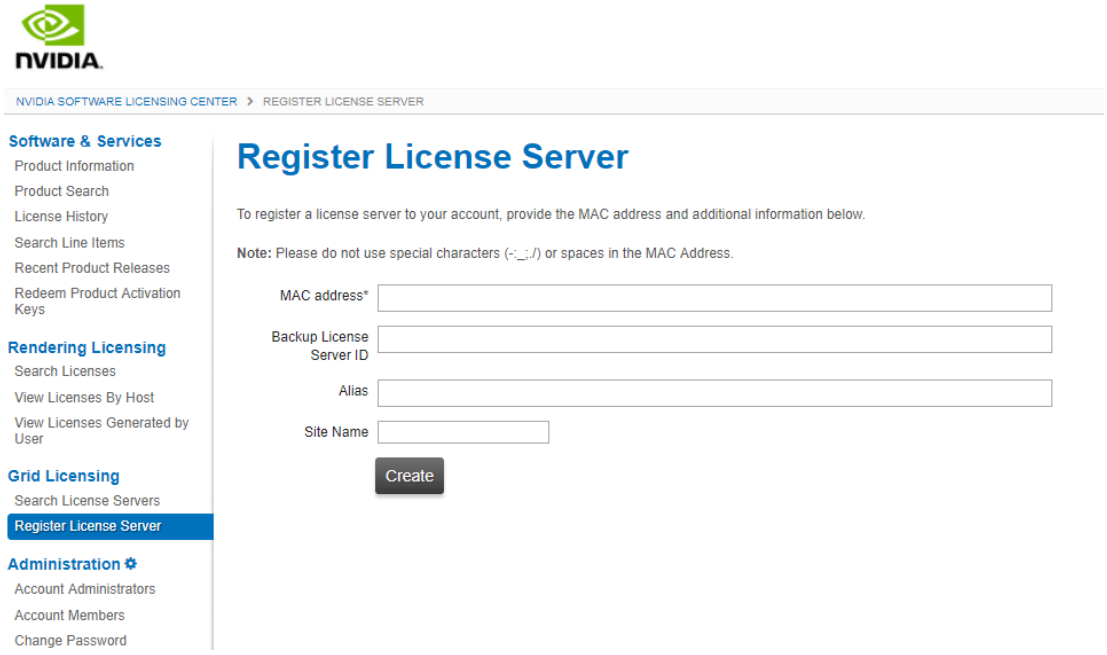
If you require high availability for the license server, you must register a secondary server when the primary server is registered. If you do not register a secondary server when the primary server is registered, you **cannot** register the secondary server later. For information about high availability for the license server, refer to *Virtual GPU License Server User Guide*.

1. If you are not already logged in, log in to the NVIDIA Enterprise Application Hub and click **NVIDIA LICENSING PORTAL** to go to the NVIDIA Software Licensing Center.
2. In the navigation pane of the **NVIDIA Software Licensing Center** page, under **GRID Licensing**, select **Register License Server**.
3. On the **Register License Server** page that opens, provide the details of your license server.
   a) In the **MAC Address** field, enter the MAC address of your license server.
      Enter **only** the numbers and letters in the address. Do **not** enter any spaces or punctuation.
   b) Leave the **Backup License Server ID** field unset.
   c) Optional: Provide an alias and site name for your license server.

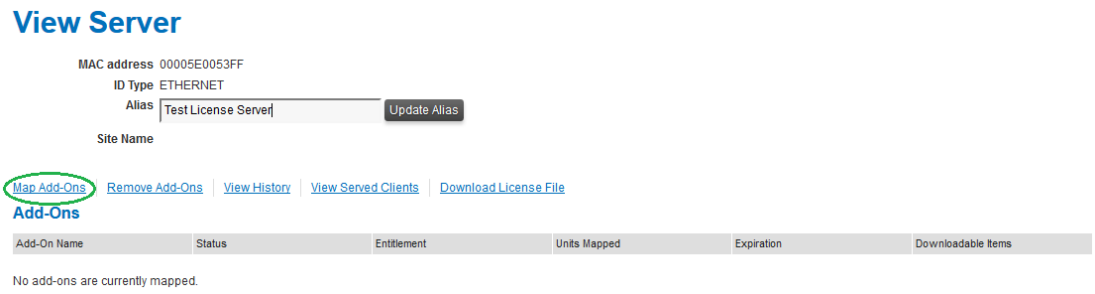
If you do not provide an alias, it is set to the MAC address of your license server. You can edit the alias after your license server is registered.

If you do not provide a site name, your license server is registered without a site name. You **cannot** add a site name after your license server is registered.

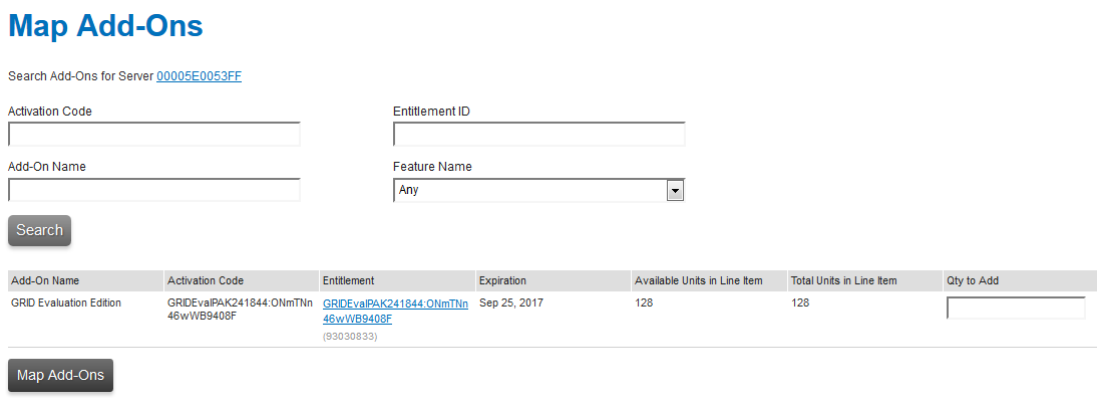
d) Click **Create**.



4. On the **View Server** page that opens, follow the **Map Add-Ons** link.



5. On the **Map Add-Ons** page that opens, in the **Qty to add** field, enter the number of licenses you want to add to your license file and click **Map Add-Ons**.



You can also configure and register multiple servers and distribute your NVIDIA vGPU software licenses across them by mapping add-ons to servers as necessary.

After clicking **Map Add-Ons**, you are returned to the **View Server** page.

6. Back on the **View Server** page, follow the **Download License File** link and save the `.bin` license file to your license server for installation.

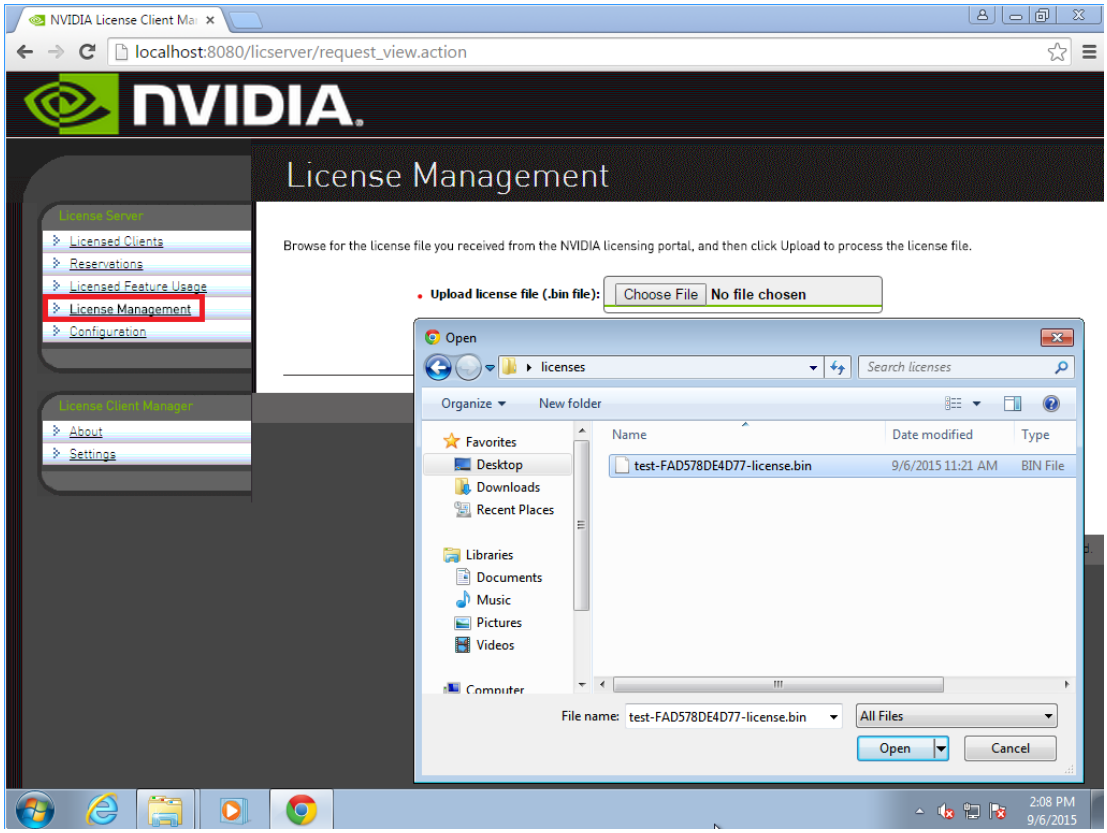
The `.bin` license file is valid for 72 hours. If you do not install the file within 72 hours of downloading it, you must download it again.

2.7. Installing a License

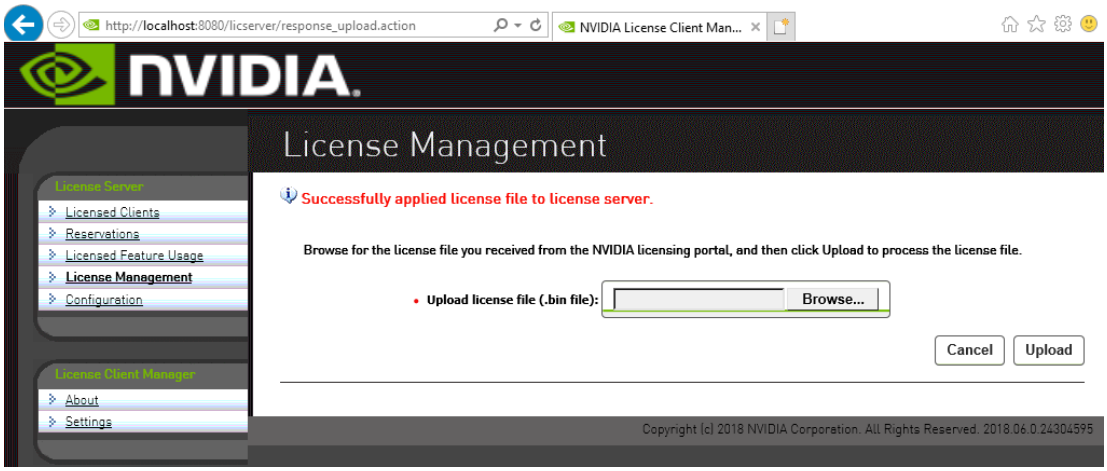
NVIDIA vGPU software licenses are distributed as `.bin` files for download from the NVIDIA Software Licensing Center.

Before installing a license, ensure that you have downloaded the license file from the NVIDIA Software Licensing Center.

1. In the license server management interface, select **License Management**.
2. On the **License Management** page that opens, click **Choose File**.



3. In the file browser that opens, select the `.bin` file and click **Open**.
4. Back on the **License Management** page, click **Upload** to install the license file on the license server.
   The license server should confirm successful installation of the license file.



Chapter 3. INSTALLING AND CONFIGURING NVIDIA VGPU MANAGER AND THE GUEST DRIVER

Before installing and configuring NVIDIA vGPU Manager and the Guest Driver, ensure that a VM running a supported Windows guest OS is configured in your chosen hypervisor.

The factory settings of some supported GPU boards are incompatible with NVIDIA vGPU software. Before configuring NVIDIA vGPU software on these GPU boards, you must configure the boards to change these settings.

3.1. Switching the Mode of a Tesla M60 or M6 GPU

Tesla M60 and M6 GPUs support compute mode and graphics mode. NVIDIA vGPU requires GPUs that support both modes to operate in graphics mode.

Only Tesla M60 and M6 GPUs require and support mode switching. Other GPUs that support NVIDIA vGPU do not require or support mode switching.

Even in compute mode, Tesla M60 and M6 GPUs do **not** support NVIDIA vComputeServer vGPU types.

Recent Tesla M60 GPUs and M6 GPUs are supplied in graphics mode. However, your GPU might be in compute mode if it is an older Tesla M60 GPU or M6 GPU, or if its mode has previously been changed.

If your GPU supports both modes but is in compute mode, you must use the `gpumodeswitch` tool to change the mode of the GPU to graphics mode. If you are unsure which mode your GPU is in, use the `gpumodeswitch` tool to find out the mode.

For more information, see *gpumodeswitch User Guide*.

3.2. Installing the NVIDIA Virtual GPU Manager

Before guests enabled for NVIDIA vGPU can be configured, the NVIDIA Virtual GPU Manager must be installed in your chosen hypervisor. The process for installing the NVIDIA Virtual GPU Manager depends on the hypervisor that you are using.

If you need more detailed instructions, refer to the appropriate NVIDIA vGPU installation guide.

3.2.1. Installing the NVIDIA Virtual GPU Manager on VMware vSphere

The NVIDIA Virtual GPU Manager for VMware vSphere ESXi is distributed as a vSphere Installation Bundle (VIB) file.

1. Copy the NVIDIA Virtual GPU Manager VIB file to the ESXi host.
2. Put the ESXi host into maintenance mode.

   ```
   $ esxcli system maintenanceMode set --enable true
   ```
3. Run the `esxcli` command to install the NVIDIA Virtual GPU Manager from the VIB file.

   ```
   $ esxcli software vib install -v directory/NVIDIA**.vib
   ```

   directory is the path to the directory that contains the VIB file.
4. Exit maintenance mode.

   ```
   $ esxcli system maintenanceMode set --enable false
   ```
5. Reboot the ESXi host.

   ```
   $ reboot
   ```
6. Verify that the NVIDIA kernel driver can successfully communicate with the physical GPUs in your system by running the `nvidia-smi` command without any options.

   ```
   $ nvidia-smi
   ```

   If successful, the `nvidia-smi` command lists all the GPUs in your system.

3.2.2. Installing the NVIDIA Virtual GPU Manager on Citrix Hypervisor

The NVIDIA Virtual GPU Manager for Citrix Hypervisor is distributed as an RPM Package Manager (RPM) file. It runs in the Citrix Hypervisor Control Domain (dom0) shell.

1. Copy the NVIDIA Virtual GPU Manager RPM file to the Citrix Hypervisor dom0 shell.

2. Run the `rpm` command to install the package.

```
[root@xenserver ~]# rpm -iv NVIDIA-**.rpm
```

3. Reboot the Citrix Hypervisor platform.

```
[root@xenserver ~]# shutdown -r now
```

4. After the Citrix Hypervisor host has rebooted, verify the installation of the NVIDIA Virtual GPU Manager package for Citrix Hypervisor by checking for the NVIDIA kernel driver in the list of kernel-loaded modules.

```
[root@xenserver ~]# lsmod |grep nvidia
nvidia 8152994 0
i2c_core 20294 2 nvidia,i2c_
```

3.3. Disabling and Enabling ECC Memory

Some GPUs that support NVIDIA vGPU software support error correcting code (ECC) memory with NVIDIA vGPU. ECC memory improves data integrity by detecting and handling double-bit errors. However, not all GPUs, vGPU types, and hypervisor software versions support ECC memory with NVIDIA vGPU.

On GPUs that support ECC memory with NVIDIA vGPU, ECC memory is supported with C-series and Q-series vGPUs, but not with A-series and B-series vGPUs. Although A-series and B-series vGPUs start on physical GPUs on which ECC memory is enabled, enabling ECC with vGPUs that do not support it might incur some costs.

On physical GPUs that do not have HBM2 memory, the amount of frame buffer that is usable by vGPUs is reduced. All types of vGPU are affected, not just vGPUs that support ECC memory.

The effects of enabling ECC memory on a physical GPU are as follows:

- ECC memory is exposed as a feature on all supported vGPUs on the physical GPU.
- In VMs that support ECC memory, ECC memory is enabled, with the option to disable ECC in the VM.
- ECC memory can be enabled or disabled for individual VMs. Enabling or disabling ECC memory in a VM does not affect the amount of frame buffer that is usable by vGPUs.

GPUs based on the Pascal GPU architecture and later GPU architectures support ECC memory with NVIDIA vGPU. These GPUs are supplied with ECC memory enabled.

Tesla M60 and M6 GPUs support ECC memory when used without GPU virtualization, but NVIDIA vGPU does not support ECC memory with these GPUs. In graphics mode, these GPUs are supplied with ECC memory disabled by default.

Some hypervisor software versions do not support ECC memory with NVIDIA vGPU.

If you are using a hypervisor software version or GPU that does not support ECC memory with NVIDIA vGPU and ECC memory is enabled, NVIDIA vGPU fails to start. In this situation, you must ensure that ECC memory is disabled on all GPUs if you are using NVIDIA vGPU.

3.3.1. Disabling ECC Memory

If ECC memory is unsuitable for your workloads but is enabled on your GPUs, disable it. You must also ensure that ECC memory is disabled on all GPUs if you are using NVIDIA vGPU with a hypervisor software version or a GPU that does not support ECC memory with NVIDIA vGPU. If your hypervisor software version or GPU does not support ECC memory and ECC memory is enabled, NVIDIA vGPU fails to start.

Where to perform this task from depends on whether you are changing ECC memory settings for a physical GPU or a vGPU.

- For a physical GPU, perform this task from the hypervisor host.
- For a vGPU, perform this task from the VM to which the vGPU is assigned.

ECC memory must be enabled on the physical GPU on which the vGPUs reside.

Before you begin, ensure that NVIDIA Virtual GPU Manager is installed on your hypervisor. If you are changing ECC memory settings for a vGPU, also ensure that the NVIDIA vGPU software graphics driver is installed in the VM to which the vGPU is assigned.

1. Use `nvidia-smi` to list the status of all physical GPUs or vGPUs, and check for ECC noted as enabled.

```
# nvidia-smi -q

==============NVSMI LOG==============

Timestamp                           : Fri Nov 22 18:36:45 2019
Driver Version                      : 430.67

Attached GPUs                       : 1
GPU 0000:02:00.0

[...]

    Ecc Mode
        Current                     : Enabled
        Pending                     : Enabled

[...]
```

2. Change the ECC status to off for each GPU for which ECC is enabled.
   - If you want to change the ECC status to off for all GPUs on your host machine or vGPUs assigned to the VM, run this command:

```
# nvidia-smi -e 0
```

   - If you want to change the ECC status to off for a specific GPU or vGPU, run this command:

```
# nvidia-smi -i id -e 0
```

     id is the index of the GPU or vGPU as reported by `nvidia-smi`.

This example disables ECC for the GPU with index `0000:02:00.0`.

```
# nvidia-smi -i 0000:02:00.0 -e 0
```

3. Reboot the host or restart the VM.
4. Confirm that ECC is now disabled for the GPU or vGPU.

```
# nvidia–smi –q

==============NVSMI LOG==============

Timestamp                           : Fri Nov 22 18:37:53 2019
Driver Version                      : 430.67

Attached GPUs                       : 1
GPU 0000:02:00.0
[...]

    Ecc Mode
        Current                     : Disabled
        Pending                     : Disabled

[...]
```

3.3.2. Enabling ECC Memory

If ECC memory is suitable for your workloads and is supported by your hypervisor software and GPUs, but is disabled on your GPUs or vGPUs, enable it.

Where to perform this task from depends on whether you are changing ECC memory settings for a physical GPU or a vGPU.

- For a physical GPU, perform this task from the hypervisor host.
- For a vGPU, perform this task from the VM to which the vGPU is assigned.

ECC memory must be enabled on the physical GPU on which the vGPUs reside.

Before you begin, ensure that NVIDIA Virtual GPU Manager is installed on your hypervisor. If you are changing ECC memory settings for a vGPU, also ensure that the NVIDIA vGPU software graphics driver is installed in the VM to which the vGPU is assigned.

1. Use `nvidia-smi` to list the status of all physical GPUs or vGPUs, and check for ECC noted as disabled.

```
# nvidia-smi -q

==============NVSMI LOG==============

Timestamp                           : Fri Nov 22 18:36:45 2019
Driver Version                      : 430.67

Attached GPUs                       : 1
GPU 0000:02:00.0

[...]
```

```
    Ecc Mode
        Current                     : Disabled
        Pending                     : Disabled

[...]
```

2. Change the ECC status to on for each GPU or vGPU for which ECC is enabled.
   - If you want to change the ECC status to on for all GPUs on your host machine or vGPUs assigned to the VM, run this command:

     ```
     # nvidia-smi -e 1
     ```
   - If you want to change the ECC status to on for a specific GPU or vGPU, run this command:

     ```
     # nvidia-smi -i id -e 1
     ```

     id is the index of the GPU or vGPU as reported by `nvidia-smi`.

     This example enables ECC for the GPU with index `0000:02:00.0`.

     ```
     # nvidia-smi -i 0000:02:00.0 -e 1
     ```
3. Reboot the host or restart the VM.
4. Confirm that ECC is now enabled for the GPU or vGPU.

```
# nvidia–smi –q

==============NVSMI LOG==============

Timestamp                           : Fri Nov 22 18:37:53 2019
Driver Version                      : 430.67

Attached GPUs                       : 1
GPU 0000:02:00.0
[...]

    Ecc Mode
        Current                     : Enabled
        Pending                     : Enabled

[...]
```

3.4. Attaching an NVIDIA vGPU Profile to a VM

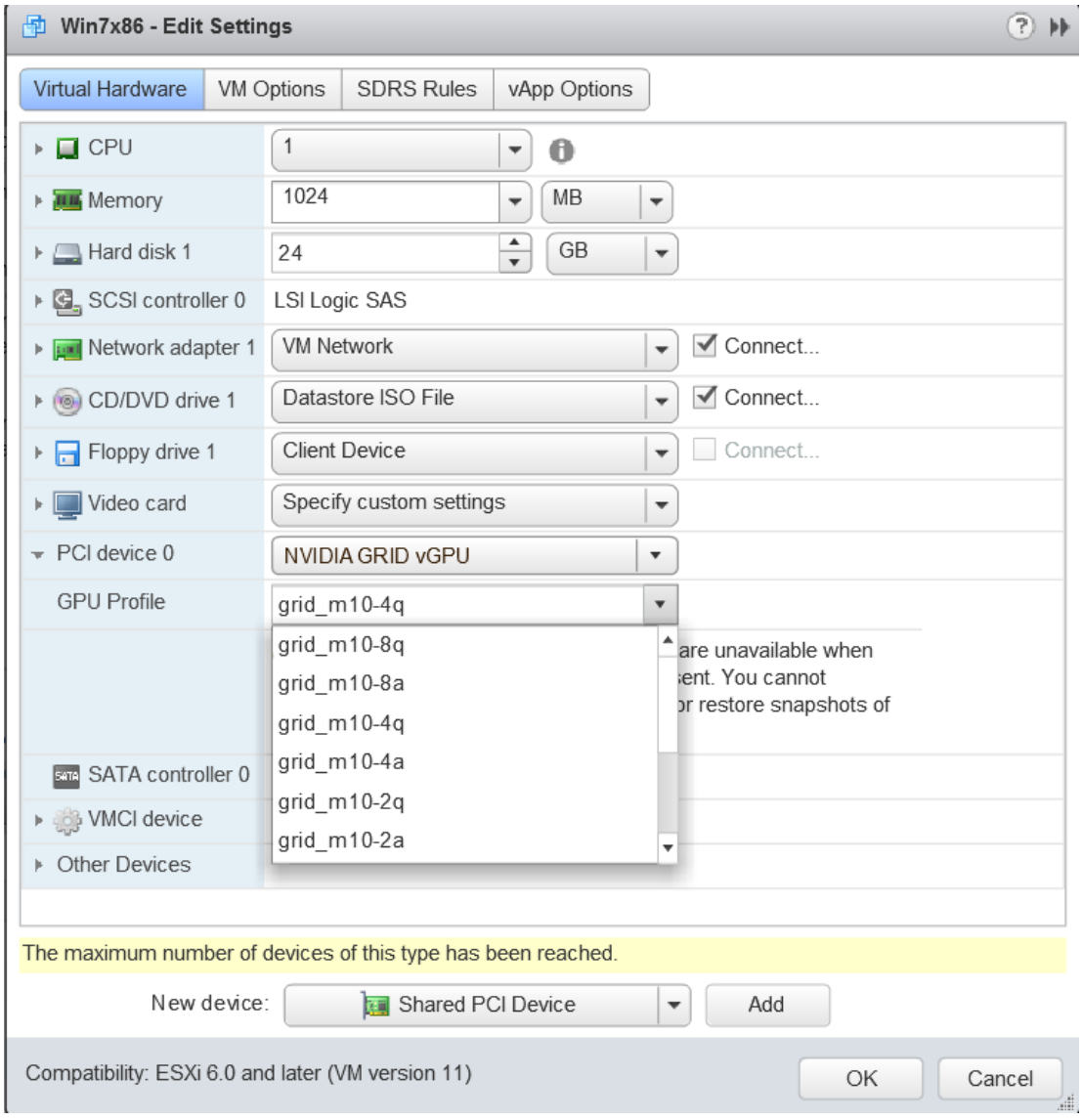
To attach an NVIDIA vGPU profile to a virtual machine (VM), you must configure the VM hardware. The process for attaching an NVIDIA vGPU profile to a VM depends on the hypervisor that you are using.

3.4.1. Configuring a vSphere VM with NVIDIA vGPU

Caution Output from the VM console in the VMware vSphere Web Client is not available for VMs that are running vGPU. Make sure that you have installed an alternate means of accessing the VM (such as VMware Horizon or a VNC server) before you configure vGPU.

VM console in vSphere Web Client will become active again once the vGPU parameters are removed from the VM's configuration.

1. Open the vCenter Web UI.
2. In the vCenter Web UI, right-click the VM and choose **Edit Settings**.
3. Click the **Virtual Hardware** tab.
4. In the **New device** list, select **Shared PCI Device** and click **Add**.
   The **PCI device** field should be auto-populated with **`NVIDIA GRID vGPU`**.



5. From the **GPU Profile** drop-down menu, choose the type of vGPU you want to configure and click **OK**.
6. Ensure that VMs running vGPU have all their memory reserved:
   a) Select **Edit virtual machine settings** from the vCenter Web UI.

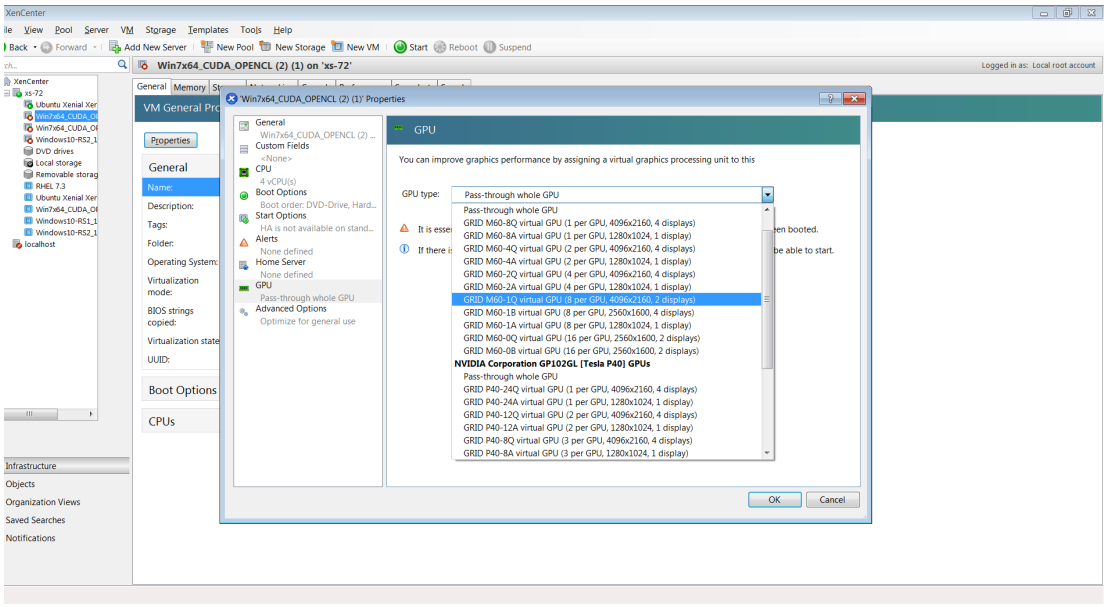
b) Expand the **Memory** section and click **Reserve all guest memory (All locked)**.

After you have configured a vSphere VM with a vGPU, start the VM. VM console in vSphere Web Client is not supported in this vGPU release. Therefore, use VMware Horizon or VNC to access the VM's desktop.

3.4.2. Configuring a Citrix Hypervisor VM with Virtual GPU

1. Ensure the VM is powered off.
2. Right-click the VM in XenCenter, select **Properties** to open the VM's properties, and select the **GPU** property.

   The available GPU types are listed in the GPU type drop-down list:



After you have configured a Citrix Hypervisor VM with a vGPU, start the VM, either from XenCenter or by using `xe vm-start` in a dom0 shell. You can view the VM's console in XenCenter.

3.5. Installing the NVIDIA vGPU Software Graphics Driver

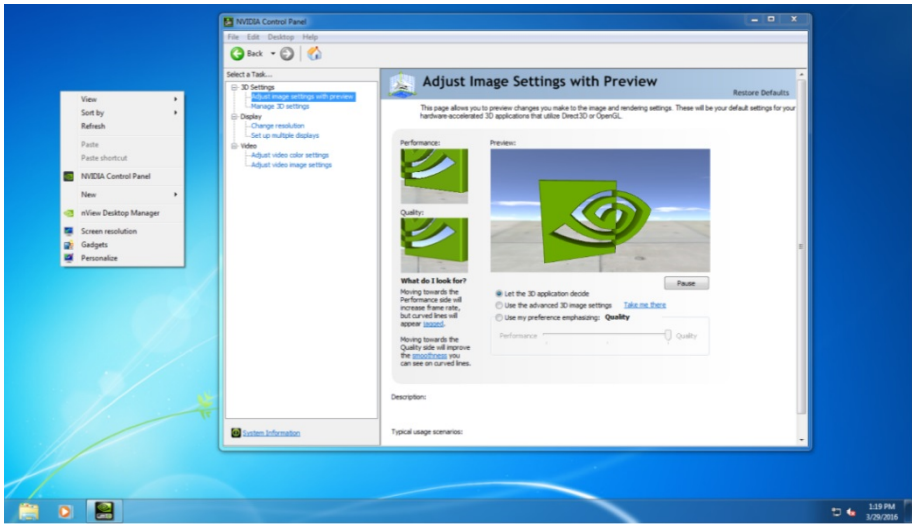
After you create a Windows VM on the hypervisor and boot the VM, the VM should boot to a standard Windows desktop in VGA mode at 800×600 resolution. You can use the Windows screen resolution control panel to increase the resolution to other standard resolutions, but to fully enable GPU operation, the NVIDIA vGPU software graphics driver must be installed.

1. Copy the NVIDIA Windows driver package to the guest VM where you are installing the driver.
2. Execute the package to unpack and run the driver installer.

3. Click through the license agreement.
4. Select **Express Installation** and click **NEXT**.
   After the driver installation is complete, the installer may prompt you to restart the platform.
5. If prompted to restart the platform, do one of the following:
   - Select **Restart Now** to reboot the VM.
   - Exit the installer and reboot the VM when you are ready.

   After the VM restarts, it boots to a Windows desktop.
6. Verify that the NVIDIA driver is running.
   a) Right-click on the desktop.
   b) From the menu that opens, choose **NVIDIA Control Panel**.
   c) In the **NVIDIA Control Panel**, from the **Help** menu, choose **System Information**.

      NVIDIA Control Panel reports the vGPU that is being used, its capabilities, and the NVIDIA driver version that is loaded.



3.6. Licensing an NVIDIA vGPU

Perform this task from the guest VM to which the vGPU is assigned.

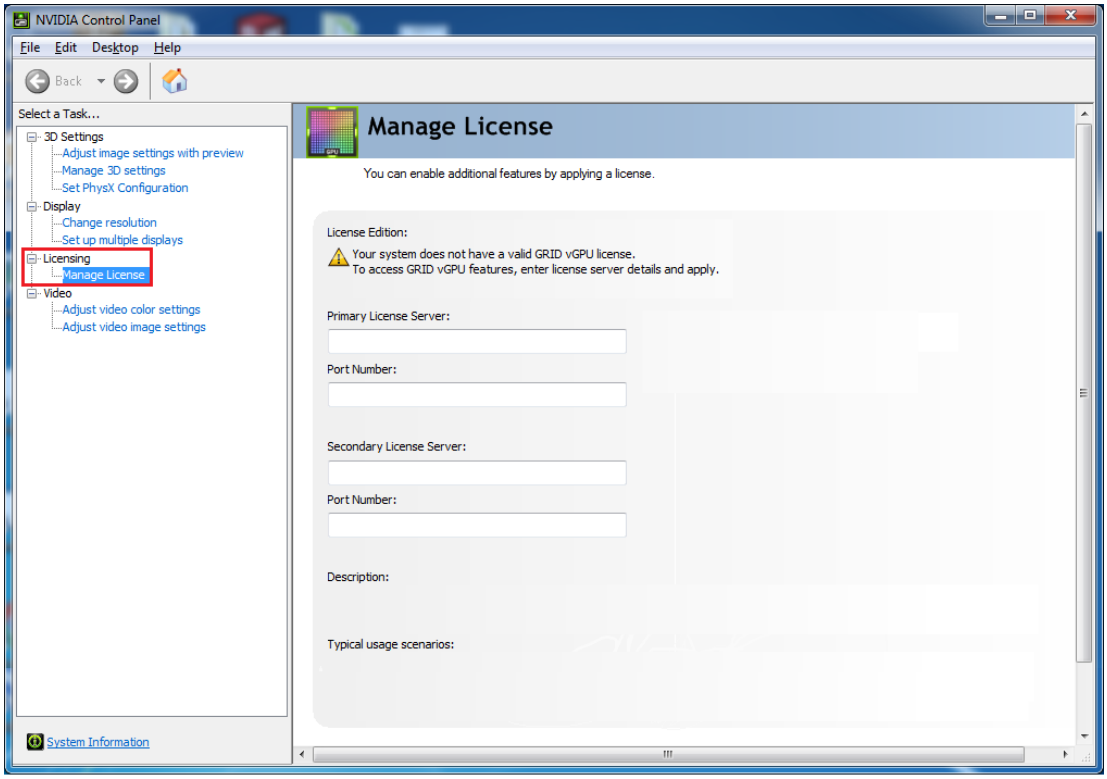
The **NVIDIA Control Panel** tool that you use to perform this task detects that a vGPU is assigned to the VM and, therefore, provides no options for selecting the license type. After you license the vGPU, NVIDIA vGPU software automatically selects the correct type of license based on the vGPU type.

These instructions cover only licensing for a single-instance license server. If you configured a primary and secondary license server, refer to *Virtual GPU Client Licensing User Guide* for instructions for how to specify the address and port number of the primary and secondary license servers.

1. Open **NVIDIA Control Panel**:
   - Right-click on the Windows desktop and select **NVIDIA Control Panel** from the menu.
   - Open **Windows Control Panel** and double-click the **NVIDIA Control Panel** icon.
2. In **NVIDIA Control Panel**, select the **Manage License** task in the **Licensing** section of the navigation pane.

   If the **Licensing** section and **Manage License** task are not displayed in **NVIDIA Control Panel**, the system has been configured to hide licensing controls in **NVIDIA Control Panel**. For information about registry settings, see *Virtual GPU Client Licensing User Guide*.

   The **Manage License** task pane shows that NVIDIA vGPU is currently unlicensed.



3. In the **Primary License Server** field, enter the address of your NVIDIA vGPU software License Server.

   The address can be a fully-qualified domain name such as `gridlicense1.example.com`, or an IP address such as `10.31.20.45`.
4. Leave the **Port Number** field under the **Primary License Server** field unset.
   The port defaults to `7070`, which is the default port number used by NVIDIA vGPU software License Server.
5. Leave the **Secondary License Server** field and the **Port Number** field under it unset.
6. Click **Apply** to assign the settings.
   The system requests the appropriate license for the current vGPU from the configured license server.

The vGPU within the VM should now exhibit full frame rate, resolution, and display output capabilities. The VM is now capable of running the full range of DirectX and OpenGL graphics applications.

You have now completed the process of installing NVIDIA vGPU software and setting up a single VM to use NVIDIA vGPU.

If the system fails to obtain a license, see *Virtual GPU Client Licensing User Guide* for guidance on troubleshooting.

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.

OpenCL

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

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA GRID, vGPU, Pascal, 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

© 2013-2019 NVIDIA Corporation. All rights reserved.