ALIBABA CLOUD

Alibaba Cloud

Elastic GPU Service

User Guide

Document Version: 20210122

Alibaba Cloud

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.

2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company or individual in any form or by any means without the prior written consent of Alibaba Cloud.

3. The content of this document may be changed because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.

4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides this document based on the "status quo", "being defective", and "existing functions" of its products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, special, or punitive damages, including lost profits arising from the use or trust in this document (even if Alibaba Cloud has been notified of the possibility of such a loss).

5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.

6. Please directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
Danger	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	**Danger:** Resetting will result in the loss of user configuration data.
Warning	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	**Warning:** Restarting will cause business interruption. About 10 minutes are required to restart an instance.
Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	**Notice:** If the weight is set to 0, the server no longer receives new requests.
Note	A note indicates supplemental instructions, best practices, tips, and other content.	**Note:** You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

1.Quick reference

This topic provides a quick reference guide for common operations of ECS instances. This topic also introduces common operations on instance resources.

This guide offers solutions for scenarios such as how to connect to ECS instances, change operating systems, resize cloud disks, upgrade or downgrade configurations, and use snapshots or images.

Limits

- For more information about considerations of ECS instances, see Usage notes.
- For more information about limits of ECS resources, see Limits and View quotas (old version).
- To apply for ICP filings for websites that are deployed on your ECS instance, make sure that the instance meets ICP filing requirements. You can apply for a limited number of ICP filing service numbers for each ECS instance. For more information, see Prepare and check the instance and access information.

Create and manage ECS instances

- You can perform the following steps to manage the lifecycle of an ECS instance:
  i. Create an instance by using the wizard
  ii. Connect to an ECS instance
  iii. Stop an instance
  iv. Release an instance
- If the current instance type or network configuration is unsuitable for your business, you can change the instance type, IP address, and peak Internet bandwidth:
  - Subscription instances:
    - Upgrade the instance types of subscription instances
    - Downgrade the configurations of an instance during renewal
  - Pay-as-you-go instances:
    - Change the instance type of a pay-as-you-go instance
    - Modify the bandwidth configurations of pay-as-you-go instances
  - IP addresses of ECS instances:
    - Change the public IP address of an ECS instance
    - Convert the public IP address of a VPC-type instance to an Elastic IP address
- If the current operating system is unsuitable for your business, you can change the operating system. For more information, see Change the operating system.
- You can use the following features to control and manage ECS instances in a fine-grained manner:
  - User data
  - Metadata
  - Instance identity
  - Instance RAM roles

Manage the billing method

- Subscription instances:

  You can use one of the following methods to renew subscription instances:

  - Manually renew an instance
  - Enable auto-renewal for an instance
  - Downgrade the configurations of an instance during renewal

- Pay-as-you-go instances:

  You can enable the No Fees for Stopped Instances (VPC-Connected) feature for pay-as-you-go instances. For more information, see No Fees for Stopped Instances (VPC-Connected).

- Change the billing method of ECS instances:
  - Change the billing method of an instance from pay-as-you-go to subscription
  - Change the billing method of an instance from subscription to pay-as-you-go

Improve cost-effectiveness

- You can purchase preemptible instances to reduce costs and achieve automatic scaling by combining with auto provisioning. For more information, see Create an auto provisioning group and Create a preemptible instance.
- You can purchase reserved instances to improve the flexibility of paying for instances and reduce costs. For more information, see Purchase reserved instances.

Create and manage cloud disks

If you want to use a cloud disk as a data disk, you can perform the following steps:

1. Create a pay-as-you-go disk.
2. Attach a cloud disk.
3. Format a data disk for a Linux instance or Format a data disk for a Windows ECS instance.
4. Create a snapshot to back up data. For more information, see Create a snapshot.
5. If the capacity of the existing system disk or data disk cannot meet your requirements, you can resize the system or data disk. For more information, see Resize disks online for Linux instances and Resize disks offline for Linux instances. If you want to resize the data disk, perform one of the following operations based on your operating system:
   - 在线扩容云盘（Windows系统）
   - Resize partitions and file systems of Linux system disks
   - Resize partitions and file systems of Linux data disks
6. If data error occurs on a cloud disk, you can use a snapshot from a specified point in time to roll back the cloud disk. For more information, see Roll back a disk by using a snapshot.
7. If you want to restore a cloud disk to its initial status, you can reinitialize the disk. For more information, see Re-initialize a data disk.
8. Detach a cloud disk.
9. Release a cloud disk.

Create and manage snapshots

You can perform the following steps to use a snapshot:

1. Create a snapshot by using one of the following methods:
   - Create a normal snapshot.
   - Use an automatic snapshot policy to create snapshots automatically on a regular basis. For more information, see Apply or disable an automatic snapshot policy.
2. View the snapshot size.
3. Delete unnecessary snapshots to save storage space. For more information, see Reduce snapshot fees.

The common application scenarios for snapshots are as follows:

- To copy or back up data: You can use a snapshot to create or roll back a cloud disk. For more information, see Create a disk from a snapshot and Roll back a disk by using a snapshot.
- To ease environment deployment: You can use a system disk snapshot to create a custom image and use the custom image to create instances. For more information, see Create a custom image from a snapshot and Create an ECS instance by using a custom image.

Create and manage custom images

Only custom images can be managed in the ECS console. You can use a custom image to quickly deploy a business environment. You can use one of the following methods to obtain a custom image.

- Create a custom image from a snapshot.
- Create a custom image from an instance.
- Create a custom image by using Packer.
- Copy custom images across regions. For more information, see Copy custom images.
- Share custom images across accounts. For more information, see Share custom images.
- Import custom images.
- Create and import on-premises images by using Packer.

You can export custom images to back up environments. For more information, see Export custom images.

Create and manage security groups

You can perform the following steps to create and manage a security group.

1. Create a security group.
2. Add security group rules.
3. Add an ECS instance to a security group.
4. Delete a security group rule.
5. Delete security groups.

You can clone a security group across regions and network types to simplify business deployment. For more information, see Clone a security group.

If new security group rules disrupt your online business, you can perform a complete or partial restoration of the security group rules. For more information, see Restore security group rules.

Create and bind instance RAM roles

You can perform the following steps to create and bind an instance RAM role.

1. Optional. Authorize a RAM user to use an instance RAM role. For more information, see Authorize a RAM user to manage an instance RAM role.
2. Create and bind an instance RAM role. For more information, see Bind an instance RAM role.
3. Replace the instance RAM role based on your needs. For more information, see Replace an instance RAM role.

Create and manage SSH key pairs

You can perform the following steps to create and manage an SSH key pair:

1. Create an SSH key pair or Import an SSH key pair.
2. Bind an SSH key pair to an instance.
3. Connect to a Linux instance by using an SSH key pair.
4. Unbind an SSH key pair.
5. Delete an SSH key pair.

Create and manage ENIs

You can perform the following steps to create and manage an ENI.

1. Create an ENI.
2. Attach an ENI to an instance or Attach an ENI when you create an instance.
3. Optional. Configure an ENI.
4. Assign secondary private IP addresses.
5. Unbind an ENI.
6. 删除弹性网卡.

Use tags

You can use tags to manage resources to enhance efficiency. You can perform the following steps to use tags:

1. Create or bind a tag.
2. Search for resources by tag.
3. Delete or unbind a tag.

Create and manage launch templates

Launch templates help you create ECS instances that have the same configurations. You can perform the following steps to create a launch template:

1. Create a launch template.
2. 创建实例启动模板的新版本.
3. 删除实例启动模板和版本.

Create and manage deployment sets

Deployment sets help you implement high availability for underlying applications. You can perform the following steps to create and manage a deployment set:

1. Create a deployment set.
2. Create an ECS instance in a deployment set.
3. Change the deployment set of an instance.
4. Delete a deployment set.

Use Cloud Assistant

Cloud Assistant allows you to send remote commands to ECS instances without the need to configure jump servers. You can perform the following steps to use Cloud Assistant:

1. Optional. Manually install and configure the Cloud Assistant client on some ECS instances. For more information, see Install the Cloud Assistant client.
2. Create a command.
3. Run a command.
4. Query execution results and fix common problems.

2.Create a NVIDIA GPU-accelerated instance

This topic describes how to create a NVIDIA GPU-accelerated instance and install a GPU driver to use GPUs.

Prerequisites

You must complete the following preparations to create an ECS instance:

1. Create an account and complete the account information.
   - Create an Alibaba Cloud account. For more information, see Sign up with Alibaba Cloud.
   - Bind your credit card or PayPal account. For more information, see Add a payment method.
   - To purchase ECS instances in mainland China regions, you must complete real-name verification. For more information, see Real-name registration FAQ.
2. Alibaba Cloud provides a default VPC in each region. If you do not want to use the default VPC, you can create a VPC and a VSwitch in the region in which to create the instance. For more information, see Create an IPv4 VPC network.
3. Alibaba Cloud provides a default security group in each region. If you do not want to use the default security group, you can create a security group in the region in which to create the instance. For more information, see Create a security group.

If you need other extended features, you must complete corresponding preparations:

- To specify an SSH key pair when you create a Linux instance, you must create the SSH key pair in the corresponding region. For more information, see Create an SSH key pair.
- To add user data for the instance, you must first prepare user data. For more information about how to prepare user data, see Prepare user data.
- To associate an ECS instance with an instance RAM role, you must create the RAM role, attach permission policies to the role, and then bind the role to the instance. For more information, see Bind an instance RAM role.

Procedure

This topic focuses on the configurations of which you must take note when you create a NVIDIA GPU-accelerated instance in the ECS console. For other general configurations, see Create an instance by using the wizard.

> **Note** If you call the RunInstances operation to create an instance, you can upload the automatic installation script only by setting the UserData parameter. For information about how to prepare the automatic installation script, see the Automatic installation script section in this topic.

1. Go to the Custom Launch tab in the ECS console.
2. Configure the settings in the Basic Configurations step.

   > **Note** GPU-accelerated instance types are available only in specific regions and zones. For more information, see ECS Instance Types Available for Each Region. Select a billing method and enter an instance type name to search for the instance type.

The following table describes the parameters of which you must take note.

Parameter	Description
Instance Type	Set Architecture to **Heterogeneous Computing** and then set Category to **Virtualization Compute Optimized Type with GPU** or **Compute Optimized Type with GPU**. Then, select an instance type. The selected instance type affects the types of drivers that can be installed on the instance. The instances of vGPU-accelerated instance families such as vgn6i and vgn5i are generated from a full GPU virtualization solution with mediated pass-through. You can install only GRID drivers on these instances. However, you can install GPU drivers and GRID drivers on GPU-accelerated compute optimized instances. ◦ GPU drivers: used to drive physical GPUs. ◦ GRID drivers: used to provide instances with graphics acceleration capabilities.
Image	The selected image affects how the GPU driver and GRID driver are installed. For more information, see Installation methods of the drivers.

The following table describes how the drivers are installed.

Installation methods of the drivers

Instance type	Driver type	Installation method of the driver
vGPU-accelerated instance family such as vgn6i and vgn5i	GRID driver	No images that are pre-installed with GRID drivers are provided. You must purchase a GRID license, and manually install a GRID driver and activate the license after the instance is created.

Instance type	Driver type	Installation method of the driver
GPU-accelerated compute optimized instance families	GPU driver	You can use one of the following methods to install the GPU driver: ○ Select Auto-install GPU Driver. For more information, see Configure the automatic installation script. **Note** Only some Linux public images allow the GPU driver to be automatically installed when you create instances. If you select Shared Image or Custom Image when you create an instance, you can install the GPU driver only after you create the instance. ○ Select an Alibaba Cloud Marketplace image that is pre-installed with a GPU driver and relevant software. Alibaba Cloud Marketplace provides images that have operating systems, application environments, and various software pre-installed. Alibaba Cloud Marketplace images are reviewed by Alibaba Cloud to ensure quality and stability. You can use these images to deploy ECS instances without additional configurations. For example, you can select the **NVIDIA GPU Cloud Virtual Machine Image** deep learning image. The image is pre-installed with a NVIDIA GPU-specific optimized deep learning framework and an optimized environment for HPC application containers. For more information, see Deploy an NGC environment on instances with GPU capabilities. ○ Manually install the GPU driver after you create the instance. For more information, see Manually install a GPU driver.
GPU-accelerated compute optimized instance families	GRID driver	GPU-accelerated compute optimized instances can be installed with a GPU driver. However, no images that are pre-installed with GRID drivers are provided to create instances. You must purchase a GRID license, and manually install a GRID driver and activate the license after the instance is created.

3. Configure the settings in the Networking step. The following table describes the parameters of which you must take note.

Parameter	Description
Network Type	Select **VPC**.

Parameter	Description
Public IP Address	If you select an image of Windows 2008 R2 or an earlier version in the **Basic Configurations** step, you cannot connect to the instance by using a VNC management terminal after the GPU driver is installed. A black screen or the startup interface persists when you attempt to connect to the instance. You can select **Assign Public IPv4 Address** in the Public IP Address section in the Networking step, or associate an elastic IP address (EIP) after you create the instance. This way, you can connect to the instance over other protocols such as Remote Desktop in Windows (RDP), PC over IP (PCoIP), and XenDesktop HDX 3D. **Note** RDP does not support applications such as DirectX and OpenGL. You must install the VNC service and client on your own.

4. Configure the settings in the System Configurations step.The following table describes the parameters of which you must take note.

Parameter	Description
Logon Credentials	We recommend that you select **Key Pair** or **Password**. If you select **Set Later**, you must bind an SSH key pair or set a password by using the password reset feature before you can connect to the instance by using a VNC management terminal. Then, you must restart the instance for the modification to take effect. If you restart the instance while the GPU driver is being installed, the installation fails.
User Data	cloud-init automatically runs the script entered in the **User Data** section when the instance is started for the first time after the instance is created. ◦ If you selected Auto-install GPU Driver, Auto-install AIACC-Training, or Auto-install AIACC-Inference in the **Basic Configurations** step, the automatic installation script is displayed in the User Data section. ◦ If you did not select Auto-install GPU Driver, Auto-install AIACC-Training, or Auto-install AIACC-Inference in the **Basic Configurations** step, you can manually enter the automatic installation script in the User Data section. For information about how to prepare the automatic installation script, see the Automatic installation script section in this topic.

5. Configure the parameters in the Grouping step, confirm the configurations in the Preview step, and then click Create Order or Create Instance.
If you enter the automatic installation script in the **User Data** section, the GPU Driver, AIACC-Training, or AIACC-Inference is automatically installed on the instance after the instance is started. After the GPU driver is installed, the instance is automatically restarted for the GPU driver to run.

Note The GPU driver is more stable in persistence mode. The automatic installation script automatically enables the persistence mode for the GPU driver. Then, the script adds the corresponding commands as a Linux system service to ensure that the persistence mode is automatically enabled for the GPU driver on instance startup.

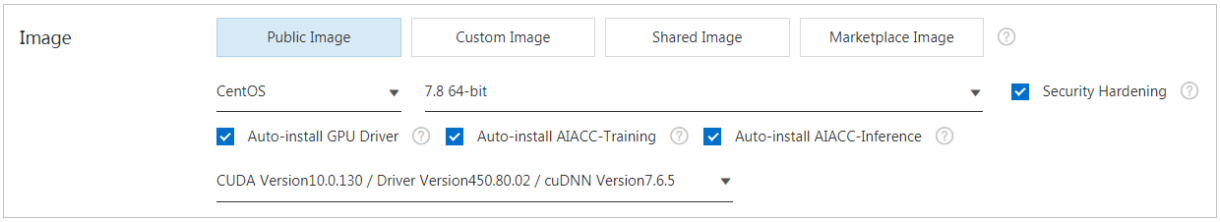
The automatic installation process may take 10 to 20 minutes based on the internal bandwidth and the number of CPU cores of different instance types. You can connect to the instance to view the installation process. You can also view the */root/auto_install/auto_install.log* installation log after the installation is complete. The following table describes the display effects of the installation process.

Installation process	Display effect
The installation is in progress.	The installation progress bar is displayed.
The installation succeeds.	**ALL INSTALL OK** appears as the installation result.
The installation fails.	**INSTALL FAIL** appears as the installation result.

Notice When the installation is in progress, the GPUs are unavailable. To prevent installation failures and keep the instance available, do not perform operations or install other GPU-related software on the instance until the installation is complete.

Configure the automatic installation script

When you create an instance in the ECS console, you can select Auto-install GPU Driver, Auto-install AIACC-Training, or Auto-install AIACC-Inference in the **Image** section of the **Basic Configurations** step. If you select Auto-install GPU Driver, the GPU driver, CUDA, and the NVIDIA CUDA Deep Neural Network library (cuDNN) are installed.



The following section describes the features of GPU drivers, AIACC-Training, and AIACC-Inference, and the available versions of GPU drivers, CUDA, and cuDNN library.

- GPU drivers are used to drive physical GPUs. When used together with CUDA and cuDNN library, GPU drivers can work efficiently. For a new business system, we recommend that you select the latest versions of the GPU driver, CUDA, and cuDNN library. The following table lists the available versions of the GPU driver, CUDA, and cuDNN library.

CUDA	GPU driver	cuDNN	Supported version of the public image (only images supplied and tested by Alibaba Cloud)	Supported instance family
11.0.2	450.80.02	8.0.4	◦ Alibaba Cloud Linux 2 ◦ Ubuntu 20.04, Ubuntu 18.04, and Ubuntu16.04 ◦ CentOS 8.x and CentOS 7.x	◦ gn6v, gn6i, gn6e, gn5, and gn5i ◦ ebmgn6v, ebmgn6i, ebmgn6e, and ebmgn5i

CUDA	GPU driver	cuDNN	Supported version of the public image (only images supplied and tested by Alibaba Cloud)	Supported instance family
10.2.89	◦ 450.80.02 ◦ 440.64.00	◦ 8.0.4 ◦ 7.6.5	◦ Alibaba Cloud Linux 2 ◦ Ubuntu 18.04 and Ubuntu 16.04 ◦ CentOS 8.x, CentOS 7.x, and CentOS 6.x	◦ gn6v, gn6i, gn6e, gn5, and gn5i ◦ ebmgn6v, ebmgn6i, ebmgn6e, and ebmgn5i
10.1.168	◦ 450.80.02 ◦ 440.64.00	◦ 8.0.4 ◦ 7.6.5 ◦ 7.5.0	◦ Ubuntu 18.04 and Ubuntu 16.04 ◦ CentOS 7.x and CentOS 6.x	◦ gn6v, gn6i, gn6e, gn5, and gn5i ◦ ebmgn6v, ebmgn6i, ebmgn6e, and ebmgn5i
10.0.130	◦ 450.80.02 ◦ 440.64.00	◦ 7.6.5 ◦ 7.5.0 ◦ 7.4.2 ◦ 7.3.1	◦ Ubuntu 18.04 and Ubuntu 16.04 ◦ CentOS 7.x and CentOS 6.x	◦ gn6v, gn6i, gn6e, gn5, and gn5i ◦ ebmgn6v, ebmgn6i, ebmgn6e, and ebmgn5i
9.2.148	◦ 450.80.02 ◦ 440.64.00 ◦ 390.116	◦ 7.6.5 ◦ 7.5.0 ◦ 7.4.2 ◦ 7.3.1 ◦ 7.1.4	◦ Ubuntu 16.04 ◦ CentOS 7.x and CentOS 6.x	◦ gn6v, gn6e, gn5, and gn5i ◦ ebmgn6v, ebmgn6e, and ebmgn5i
9.0.176	◦ 450.80.02 ◦ 440.64.00 ◦ 390.116	◦ 7.6.5 ◦ 7.5.0 ◦ 7.4.2 ◦ 7.3.1 ◦ 7.1.4 ◦ 7.0.5	◦ Ubuntu 16.04 ◦ CentOS 7.x and CentOS 6.x ◦ SUSE 12sp2	◦ gn6v, gn6e, gn5, and gn5i ◦ ebmgn6v, ebmgn6e, and ebmgn5i
8.0.61	◦ 450.80.02 ◦ 440.64.00 ◦ 390.116	◦ 7.1.3 ◦ 7.0.5	◦ Ubuntu 16.04 ◦ CentOS 7.x and CentOS 6.x	◦ gn5 and gn5i ◦ ebmgn5i

Note If you replace the operating system after the instance is created, make sure that you use an image that allows GPU drivers to be automatically installed to prevent failures in automatic installation.

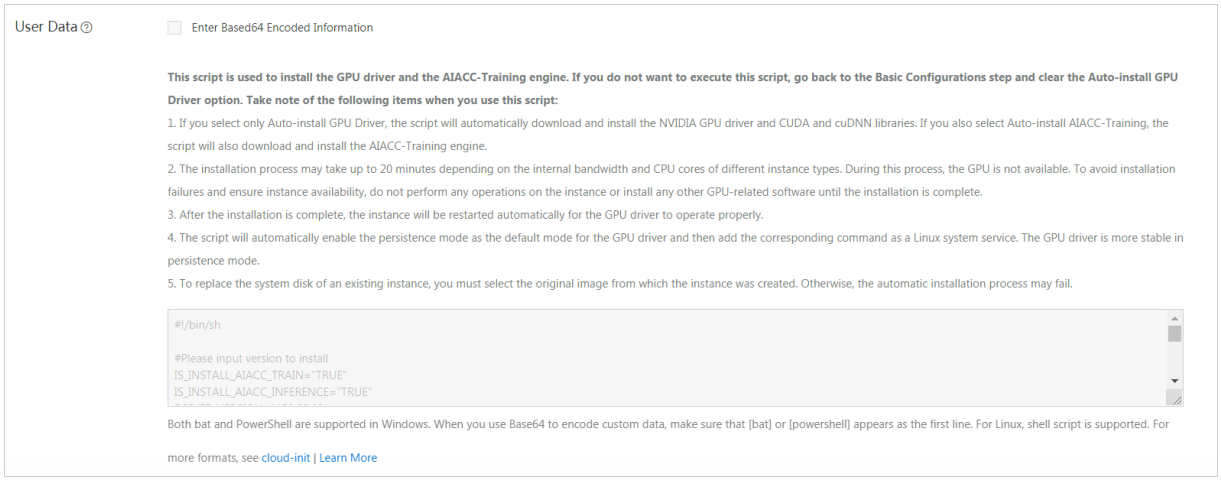
- AIACC-Training is an AI accelerator developed by Alibaba Cloud. AIACC-Training can accelerate major AI computing frameworks such as TensorFlow, PyTorch, MxNet, and Caffe to achieve significant gains in training performance. For more information, see Use AIACC-Training.

  > **Note** AIACC-Training is not supported in CentOS 8, CentOS 6, SUSE Linux, or Alibaba Cloud Linux.

- AIACC-Inference is an AI accelerator developed by Alibaba Cloud. AIACC-Inference can accelerate the major AI computing framework TensorFlow and exportable frameworks in the Open Neural Network Exchange (ONNX) format to achieve significant gains in inference performance. For more information, see Use AIACC-Inference.

  > **Note** AIACC-Inference is not supported in CentOS 8, CentOS 6, SUSE Linux, or Alibaba Cloud Linux.

If you selected Auto-install GPU Driver, Auto-install AIACC-Training, or Auto-install AIACC-Inference in the **Basic Configurations** step, the automatic installation script is displayed in the **User Data** section of the **System Configurations** step. cloud-init automatically runs the automatic installation script when the instance is started for the first time after the instance is created.



> **Note** If you did not select Auto-install GPU Driver, Auto-install AIACC-Training, or Auto-install AIACC-Inference in the **Basic Configurations** step, you can manually enter the automatic installation script in the **System Configurations** step. For information about how to prepare the automatic installation script, see the Automatic installation script section in this topic.

Automatic installation script

The automatic installation script has been updated to v3.2. The latest version of the automatic installation script has the following benefits:

- Provides the latest versions of the GPU driver, CUDA, and cuDNN Library.
- Shows the installation process after the instance is connected.

The following section lists the content of the automatic installation script:

```
#! /bin/sh
#Please input version to install
IS_INSTALL_AIACC_TRAIN=""
IS_INSTALL_AIACC_INFERENCE=""
DRIVER_VERSION=""
CUDA_VERSION=""
CUDNN_VERSION=""
IS_INSTALL_RAPIDS="FALSE"
INSTALL_DIR="/root/auto_install"
#using .deb to install driver and cuda on ubuntu OS
#using .run to install driver and cuda on ubuntu OS
auto_install_script="auto_install_v3.2.sh"
script_download_url=$(curl http://100.100.100.200/latest/meta-data/source-address | head -1)"/opsx/ecs/linux/binary/script/${auto_install_script}"
echo $script_download_url
mkdir $INSTALL_DIR && cd $INSTALL_DIR
wget -t 10 --timeout=10 $script_download_url && sh ${INSTALL_DIR}/${auto_install_script} $DRIVER_VERSION $CUDA_VERSION $CUDNN_VERSION $IS_INSTALL_AIACC_TRAIN $IS_INSTALL_AIACC_INFERENCE $IS_INSTALL_RAPIDS
```

> **Note** If you use a CentOS, SUSE, or Ubuntu 20.04 image to create the instance, the *.run* installation package is used when you run the automatic installation script. If you use a Ubuntu 18.04 or Ubuntu 16.04 image, the *.deb* installation package is used when you run the automatic installation script.

To use the automatic installation script, you must modify the version parameters of the GPU driver, CUDA, and cuDNN library in the installation script, and specify whether to install AIACC-Training and AIACC-Inference.

- If you want to install AIACC-Training, set IS_INSTALL_AIACC_TRAIN to *TRUE*. Otherwise, set IS_INSTALL_AIACC_TRAIN to *FALSE*.
- If you want to install AIACC-Inference, set IS_INSTALL_AIACC_INFERENCE to *TRUE*. Otherwise, set IS_INSTALL_AIACC_INFERENCE to *FALSE*.

Example:

```
IS_INSTALL_AIACC_TRAIN="FALSE"
IS_INSTALL_AIACC_INFERENCE="FALSE"
DRIVER_VERSION="440.64.00"
CUDA_VERSION="10.2.89"
CUDNN_VERSION="8.0.4"
```

Related information

References

- RunInstances
- Manually install a GPU driver
- Install an NVIDIA GRID driver on a GPU-accelerated Linux instance
- Manually uninstall the GPU driver
- GPU monitoring

3.Connect to an instance with GPU capabilities

3.1. Connect to a Linux instance by using VNC

If you cannot use remote connection software such as PuTTY, Xshell, and SecureCRT to connect to a Linux instance, you can use the **VNC Connection** feature in the ECS console to connect to the Linux instance and view the real-time status of the instance operation interface.

Prerequisites

- An ECS instance is created.
- A logon password is set for the instance. If you have not set a password or if you have forgotten the password, you must reset the password for the instance. For more information, see Reset the logon password of an instance.

Context

The VNC password must be six characters in length and is used to connect to the VNC management terminal in the ECS console, while the instance password is used to log on to the instance.

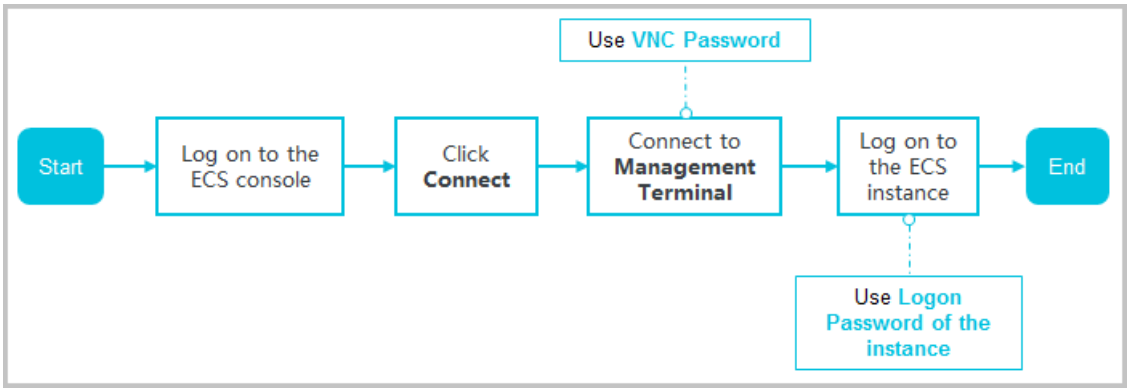
By default, a remote VNC connection session lasts for about 300 seconds. If you do not perform operations within these 300 seconds, the connection to the instance is closed. You must connect to the instance again.

If you cannot use remote connection software to connect to your instance, you can use the **VNC Connection** feature in the ECS console to connect to the instance and view the status of the instance, as described in the following table.

Scenario	Operation
The instance started slowly due to startup self-check.	Check the progress of the self-check.
The firewall of the operating system is enabled by mistake.	Disable the firewall.
The ECS instance is hacked into, which causes a high CPU utilization and high bandwidth usage.	Troubleshoot and terminate abnormal processes.

Procedure

The following figure shows the workflow.



1. Log on to the ECS console.
2. In the left-side navigation pane, choose **Instances & Images > Instances**.
3. In the top navigation bar, select a region.
4. On the **Instances** page, find the instance to be connected and click **Connect** in the **Actions** column.
5. Connect to the VNC management terminal.

   > **Note** The VNC password is used in this step.

   - If you use an Alibaba Cloud account to connect to the VNC management terminal for the first time, perform the following operations:
     a. In the **VNC Password** dialog box, copy the password.

        > **Note** The VNC password is displayed only once when you connect to the VNC management terminal for the first time. Keep the password confidential.

     b. Click **Close**.
     c. In the **Enter VNC Password** dialog box, paste the password and click **OK**.
   - If you forget your password or connect to the VNC management terminal for the first time as a RAM user, perform the following operations:
     a. Modify the VNC password.
     b. In the upper-left corner of the interface, choose **Send Remote Call > Connect VNC**.
     c. In the **Enter VNC Password** dialog box, enter the new password.
     d. Click **OK**.
   - If you connect to the VNC management terminal again by using your Alibaba Cloud account or as a RAM user, perform the following operations:

     In the **Enter VNC Password** dialog box, enter the password and click **OK**.
6. Log on to the instance.

   > **Note** The instance password is used in this step.

   i. Enter the username *root* and press the Enter key.

ii. Enter the instance password and press the Enter key.In the upper-left corner of the interface, choose **Send Remote Call > CTRL+ALT+Fx** (valid values of x: 1 to 10) to switch between different VNC management terminals for connecting to the Linux instance. A persistent black screen indicates that the instance is in the sleep mode. Press any key to wake up the system.

> **Note** The password characters are not displayed when you enter the password. After you enter the password, press the Enter key.

Modify the VNC password

When you connect to the VNC management terminal as a RAM user for the first time, you must modify the VNC password. You can also change the password when you forget or want to update the VNC password.

> **Notice** After you modify the VNC password for a non-I/O optimized instance, you must restart the instance in the ECS console for the new password to take effect. Before you restart the instance, you must stop it. This will lead to service interruption. Proceed with caution.

1. On the **Instances** page, find the instance to be connected and click **Connect** in the **Actions** column.
2. Close the **VNC Password** or **Enter VPC Password** dialog box.
3. In the upper-right corner of the interface, click **Modify VNC Password**.
4. In the **Modify VNC Password** dialog box, enter and confirm the new password, and then click **OK**.
5. (Optional)If the instance is a non-I/O optimized instance, restart the instance.For more information, see Restart an instance.

Copy long commands

If you want to copy a long text item such as a download URL from the local device to the instance, you can use the copy command input feature.

1. On the **Instances** page, find the instance to be connected and click **Connect** in the **Actions** column.
2. Connect to the VNC management terminal.
3. In the upper-right corner of the interface, click **Enter Copy Commands**.
4. In the **Copy and Paste Commands** dialog box, enter the content to be copied and then click **OK**.

3.2. Connect to a Windows instance by using VNC

If you cannot use remote connection software such as RDP and rdesktop to connect to a Windows instance, you can use VNC to establish a remote connection to the instance and view the real-time status of the instance operating interface.

Prerequisites

- An ECS instance is created.
- A logon password is set for the instance. If you have not set a password or if you have forgotten the

password, you must reset the password for the instance. For more information, see Reset the logon password of an instance.

Context

The VNC password must be six characters in length and is used to connect to the VNC management terminal in the ECS console, while the instance password is used to log on to the instance.

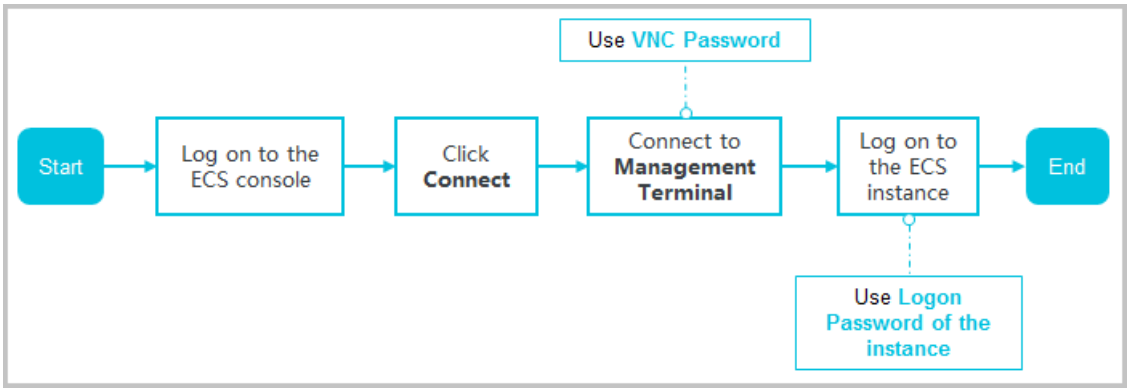
By default, a remote VNC connection session lasts for about 300 seconds. If you do not perform operations within these 300 seconds, the connection to the instance is closed. You must connect to the instance again.

If you cannot use remote connection software to connect to your instance, you can use the **VNC Connection** feature in the ECS console to connect to the instance and view the status of the instance, as described in the following table.

Scenario	Operation
The instance started slowly due to startup self-check.	Check the progress of the self-check.
The firewall of the operating system is enabled by mistake.	Disable the firewall.
The ECS instance is hacked into, which causes a high CPU utilization and high bandwidth usage.	Troubleshoot and terminate abnormal processes.

Procedure

The following figure shows the workflow.



1. Log on to the ECS console.
2. In the left-side navigation pane, choose **Instances & Images > Instances**.
3. In the top navigation bar, select a region.
4. On the **Instances** page, find the instance to be connected and click **Connect** in the **Actions** column.
5. Connect to the VNC management terminal.

   > **Note** The VNC password is used in this step.

   - If you use an Alibaba Cloud account to connect to the VNC management terminal for the first

time, perform the following operations:

a. In the **VNC Password** dialog box, copy the password.

> **Note** The VNC password is displayed only once when you connect to the VNC management terminal for the first time. Keep the password confidential.

b. Click **Close**.

c. In the **Enter VNC Password** dialog box, paste the password and click **OK**.

- If you forget your password or connect to the VNC management terminal for the first time as a RAM user, perform the following operations:

  a. Modify the VNC password.

  b. In the upper-left corner of the interface, choose **Send Remote Call > Connect VNC**.

  c. In the **Enter VNC Password** dialog box, enter the new password.

  d. Click **OK**.

- If you connect to the VNC management terminal again by using your Alibaba Cloud account or as a RAM user, perform the following operations:

  In the **Enter VNC Password** dialog box, enter the password and click **OK**.

6. In the upper-left corner of the **VNC** interface, choose **Send Remote Call > CTRL+ALT+DELETE**.
7. Select an account. Enter the instance password and press the Enter key.By default, the Administrator account is available.

Modify the VNC password

When you connect to the VNC management terminal as a RAM user for the first time, you must modify the VNC password. You can also change the password when you forget or want to update the VNC password.

> **Notice** After you modify the VNC password for a non-I/O optimized instance, you must restart the instance in the ECS console for the new password to take effect. Before you restart the instance, you must stop it. This will lead to service interruption. Proceed with caution.

1.
2. Close the **VNC Password** or **Enter VPC Password** dialog box.
3. In the upper-right corner of the interface, click **Modify VNC Password**.
4. In the **Modify VNC Password** dialog box, enter and confirm the new password, and then click **OK**.
5. (Optional)If the instance is a non-I/O optimized instance, restart the instance.For more information, see Restart an instance.

Copy long commands

If you want to copy a long text item such as a download URL from the local device to the instance, you can use the copy command input feature.

1.
2. Connect to the VNC management terminal.
3. In the upper-right corner of the interface, click **Enter Copy Commands**.

4. In the **Copy and Paste Commands** dialog box, enter the content to be copied and then click **OK**.

4.Manage an instance with GPU capabilities

4.1. Stop an instance

This topic describes how to stop an instance in the ECS console and introduces operations related to the No Fees for Stopped Instances (VPC-Connected) feature.

stop an instance No Charges After Instance Is Stopped

Prerequisites

The instance that you want to stop is in the **Running** state.

Context

If you stop a subscription instance, the billing of the instance is not affected.

If you stop a pay-as-you-go instance, the billing of the instance may be affected based on the network type of the instance and the **No Fees for Stopped Instances (VPC-Connected)** feature setting.

- Pay-as-you-go instances in the classic network do not support the **No Fees for Stopped Instances (VPC-Connected)** feature. A pay-as-you-go instance in the classic network continues to incur fees after the instance is stopped. The billing stops only when the instance is released. For more information, see Release an instance.
- Pay-as-you-go instances in VPCs support the **No Fees for Stopped Instances (VPC-Connected)** feature.
  - If the **No Fees for Stopped Instances (VPC-Connected)** feature is not enabled, the billing of a pay-as-you-go instance continues after the instance is stopped.
  - If the **No Fees for Stopped Instances (VPC-Connected)** feature is enabled, you can use the Stop Mode parameter to configure whether to retain and bill an instance after the instance is stopped. If you set Stop Mode to No Charges After Instance Is Stopped in the Stop Instance dialog box, billing of the vCPUs, memory, and public IP address stops after the instance is stopped. However, you are still charged for other resources. For more information, see No Fees for Stopped Instances (VPC-Connected).

> **Note** Services that are running on an instance will be interrupted if you stop the instance. Exercise caution when you perform this operation.

Stop a subscription instance

To stop a subscription instance, perform the following steps:

1. Log on to the ECS console.
2. In the left-side navigation pane, choose **Instances & Images > Instances**.
3. In the top navigation bar, select a region.
4. Use a suitable method to stop instances.
   - To stop a single instance, find the instance and choose **More > Instance Status > Stop** in the

Actions column.

- To stop multiple instances at a time, select the instances and click **Stop** in the lower part of the Instances page.

5. In the Stop Instance dialog box that appears, set **Stopped By**.
   - **Stop**: stops the instance by shutting it down properly.
   - **Force Stop**: forcibly stops the instance. Forcible stop is equivalent to a physical shutdown, and may cause data loss if instance data has not been written to disks.
6. Click **OK**.

Stop a pay-as-you-go instance

To stop a pay-as-you-go instance, perform the following steps:

> **Note** Services that are running on an instance will be interrupted if you stop the instance. Exercise caution when you perform this operation. The procedure to stop a preemptible instance is the same as that of a pay-as-you-go instance. For more information, see Stop a preemptible instance.

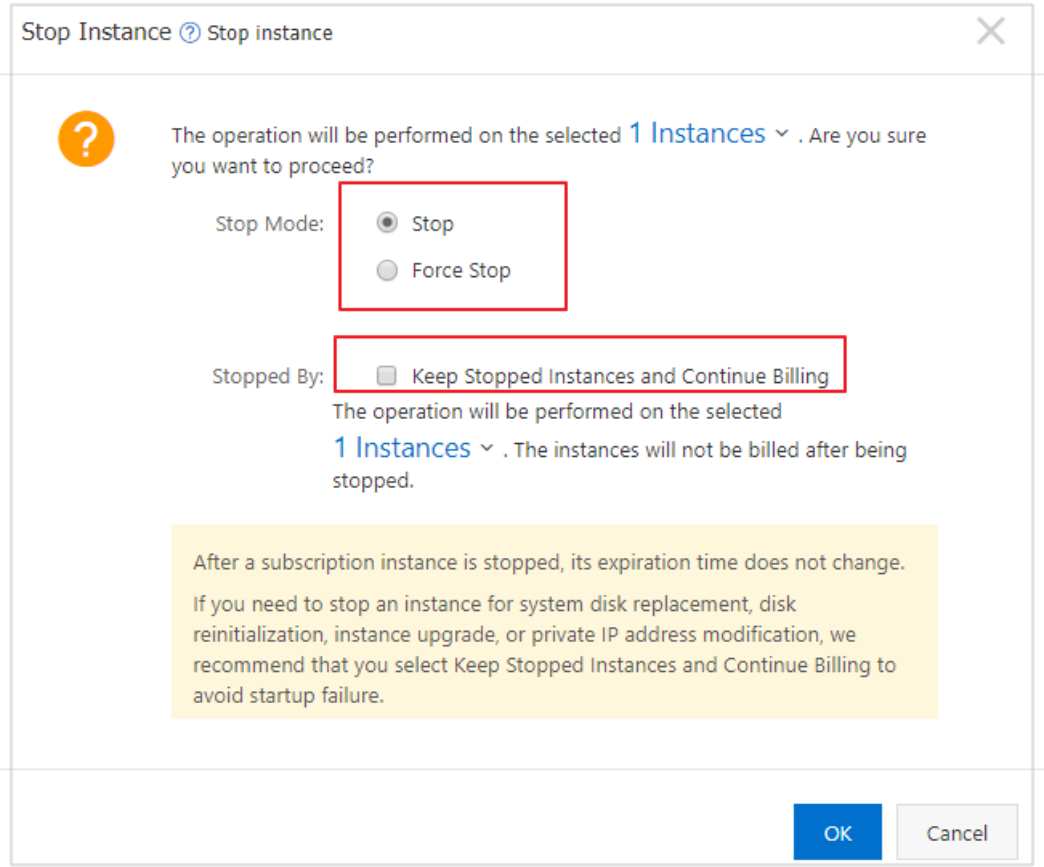
1. Log on to the ECS console.
2. In the left-side navigation pane, choose **Instances & Images > Instances**.
3. In the top navigation bar, select a region.
4. Use a suitable method to stop instances.
   - To stop a single instance, find the instance and choose **More > Instance Status > Stop** in the **Actions** column.
   - To stop multiple instances at a time, select the instances and click **Stop** in the lower part of the Instances page.
5. Configure required parameters based on the instance network type and the **No Fees for Stopped Instances (VPC-Connected)** feature setting.
   - If the network type is classic network or if **No Fees for Stopped Instances (VPC-Connected)** is not enabled:
     a. In the Stop Instance dialog box that appears, set **Stopped By**.
        - **Stop**: stops the instance by shutting it down properly.
        - **Force Stop**: forcibly stops the instance. Forcible stop is equivalent to a physical shutdown, and may cause data loss if instance data has not been written to disks.
     b. Click **OK**.
   - If **No Fees for Stopped Instances (VPC-Connected)** is enabled:
     a. Move the pointer over the icon next to **No Charges After Instance Is Stopped** and read description of the No Fees for Stopped Instances (VPC-Connected) feature.
     b. In the Stop Instance dialog box that appears, set **Stopped By**.
        - **Stop**: stops the instance by shutting it down properly.
        - **Force Stop**: forcibly stops the instance. Forcible stop is equivalent to a physical shutdown, and may cause data loss if instance data has not been written to disks.

c. Set **Stop Mode**.

- **Retain Instance and Continue Charging After Instance Is Stopped**: After the instance is stopped, resources of the instance are retained and continue to be charged.
- **No Charges After Instance Is Stopped**: After the instance is stopped, computing resources such as vCPUs and memory of the instance are not retained or charged. The cloud disks such as system and data disks, Elastic IP address, and bandwidth continue to be charged. The public IP address is reclaimed and the private IP address is retained.

d. Click **OK**.

> **Note** For information about how to disable **No Fees for Stopped Instances (VPC-Connected)**, see Disable the No Fees for Stopped Instances (VPC-Connected) feature.



Result

The instance enters the **Stopped** state after it is stopped.

Related information

- StopInstance

4.2. Start ECS instances

This topic describes how to start instances in the ECS console.

Prerequisites

The instances that you want to start are in the **Stopped** state.

Procedure

1. Log on to the ECS console.
2. In the left-side navigation pane, choose **Instances & Images > Instances**.
3. Start instances.
   - To start a single instance, choose **More > Instance Status > Start** in the **Actions** column corresponding to the target instance.
   - To start multiple instances at a time, select the instances and click **Start** in the lower-left corner of the Instances page.
4. Verify the information and click **OK**.

Result

After instances are started, they enter the **Running** state.

Related information

- StartInstance

4.3. Restart an instance

This topic describes how to restart an instance by using the ECS console.

Limits

- Only instances in the **Running** state can be restarted.
- Restarting an instance will stop the instance. As a result, services provided by the instance are disrupted.

Procedure

1.
2.
3.
4. Find the target instance, and then choose **More > Instance Status > Restart** in the **Actions** column.

   To restart multiple instances, select all required instances and then click **Restart** at the bottom of the instance list.
5. In the displayed **Restart Instance** dialog box, select a **Restart Mode**, and then click **OK**.

You can also call the RebootInstance API action to complete this task.

4.4. Release an instance

Only pay-as-you-go instances (including preemptible instances) and expired subscription instances can be released. This topic describes how to manually and automatically release a pay-as-you-go instance.

release an instance delete an instance remove an instance Alibaba Cloud ECS deleteinstance release a pay-as-you-go instance

Prerequisites

After an instance is released, its data cannot be recovered. We recommend that you create a snapshot to back up data before releasing an instance. For more information, see Create a snapshot.

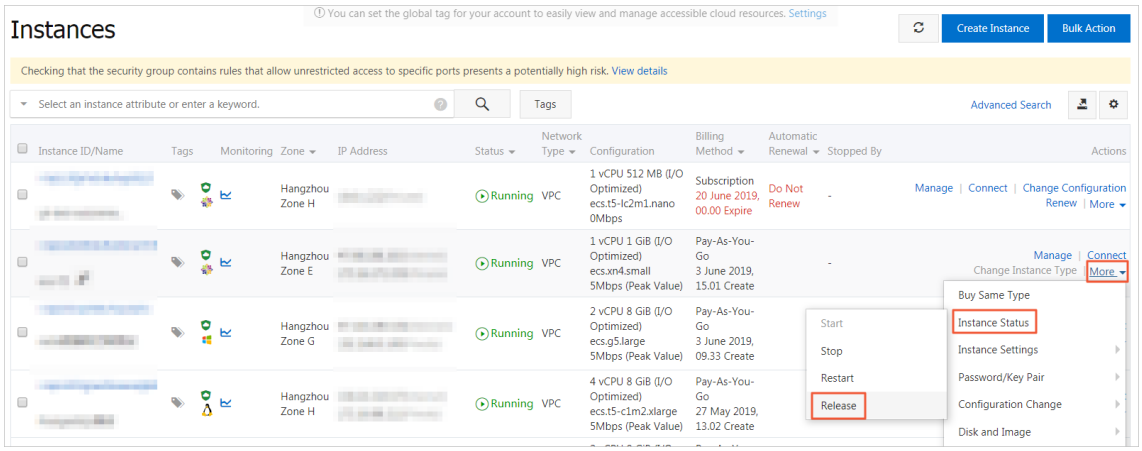
Context

- For a subscription instance, you can manually release the instance after it expires. If you do not renew the instance after it expires, the instance is automatically released.
- For a pay-as-you-go instance, if the No Fees for Stopped Instances (VPC-Connected) feature is not enabled, charges continue to incur until the instance is released.
- You can enable instance release protection for a pay-as-you-go instance to prevent irreversible data loss resulting from accidental or incorrect operations during a manual release. For more information, see Enable and disable instance release protection.
- If the Release Disk with Instance feature is disabled for the disk attached to an instance, the disk is automatically converted to a pay-as-you-go data disk and retained when the instance is released. For more information, see Release a disk.

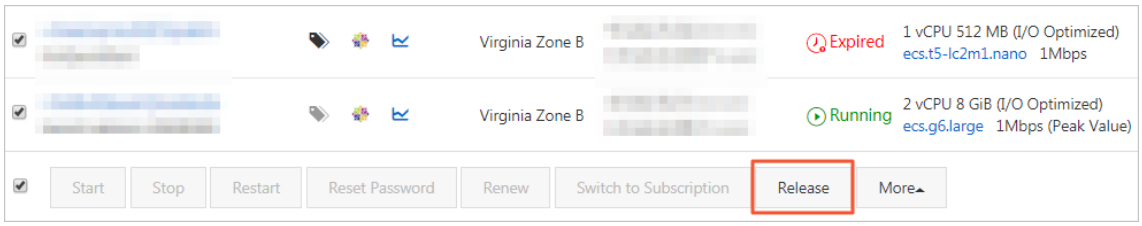
Manually release an instance

You can release a pay-as-you-go instance immediately in the console.

1. Log on to the ECS console.
2. In the left-side navigation pane, choose **Instances & Images > Instances**.
3. In the top navigation bar, select a region.
4. Configure the release.
   - To release only one instance, find the instance that you want to release and choose **More > Instance Status > Release** in the **Actions** column.



   - If you want to release multiple instances, find the pay-as-you-go instances based on the **Billing Method**, select the instances to be released, and click **Release** at the bottom of the list.



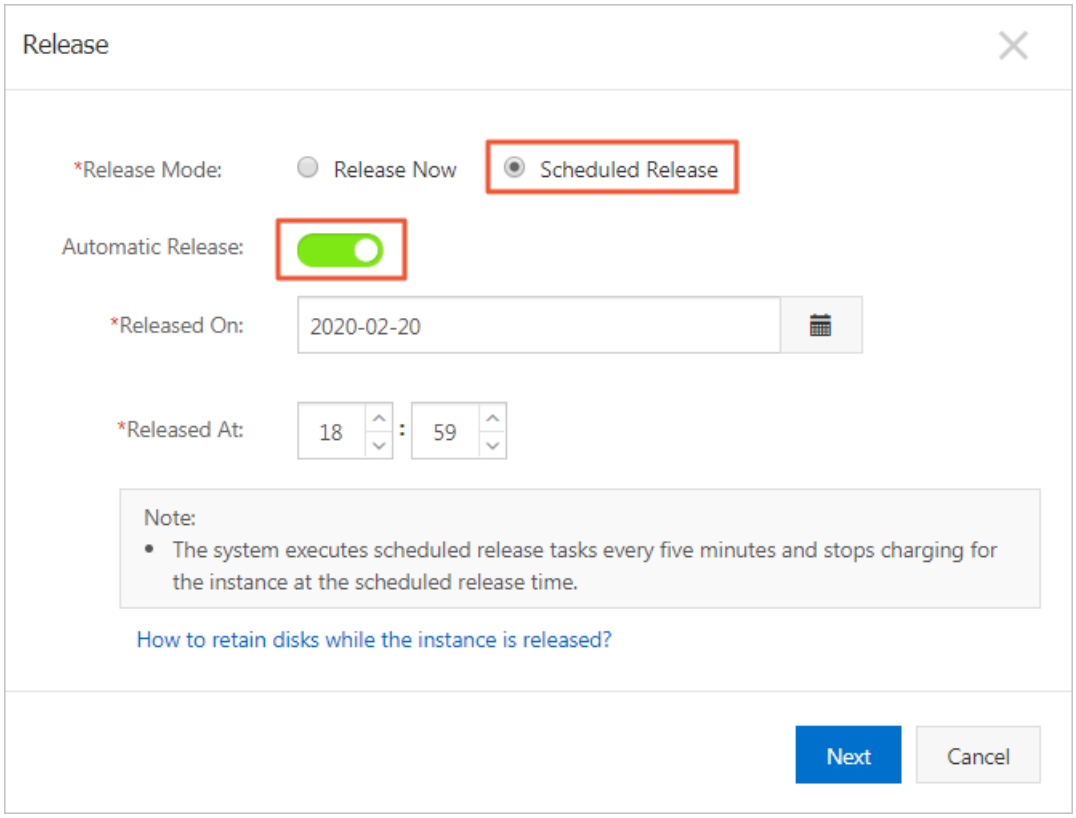
5. In the dialog box that appears, select **Release Now**.
6. Click **Next**, and then click **OK**.

Enable automatic release

You can enable the automatic release function and set a time to automatically release an instance. If you set the automatic release time multiple times, the latest setting works.

1. Log on to the ECS console.
2. In the left-side navigation pane, choose **Instances & Images > Instances**.
3. In the top navigation bar, select a region.
4. Configure the release.
   - To release only one instance, find the instance that you want to release and choose **More > Instance Status > Release** in the **Actions** column.
   - If you want to release multiple instances, find the pay-as-you-go instances based on the **Billing Method**, select the instances to be released, and click **Release** at the bottom of the list.
5. In the dialog box that appears, select **Scheduled Release**.
6. Turn on the automatic release switch, and specify the release date and time.

> **Note** The automatic release time must be at least 30 minutes later than the current time.



7. Click **Next**, and then click **OK**.

Disable automatic release

1. Log on to the ECS console.

2. In the left-side navigation pane, choose **Instances & Images > Instances**.
3. In the top navigation bar, select a region.
4. Release configuration.
   - To disable the automatic release function for only one instance, find the instance that you want to release and choose **More > Instance Status > Release** in the **Actions** column.
   - If you want to disable the automatic release function for multiple instances, find the pay-as-you-go instances based on the **Billing Method**, select the instances for which you want to disable the automatic release function, and click **Release** at the bottom of the list.
5. In the dialog box that appears, select **Scheduled Release**.
6. Turn off the automatic release switch.
7. Click **Next**, and then click **OK**.

Related information

References

- DeleteInstance
- ModifyInstanceAutoReleaseTime

5.Manage NVIDIA drivers

5.1. Manually install a GPU driver

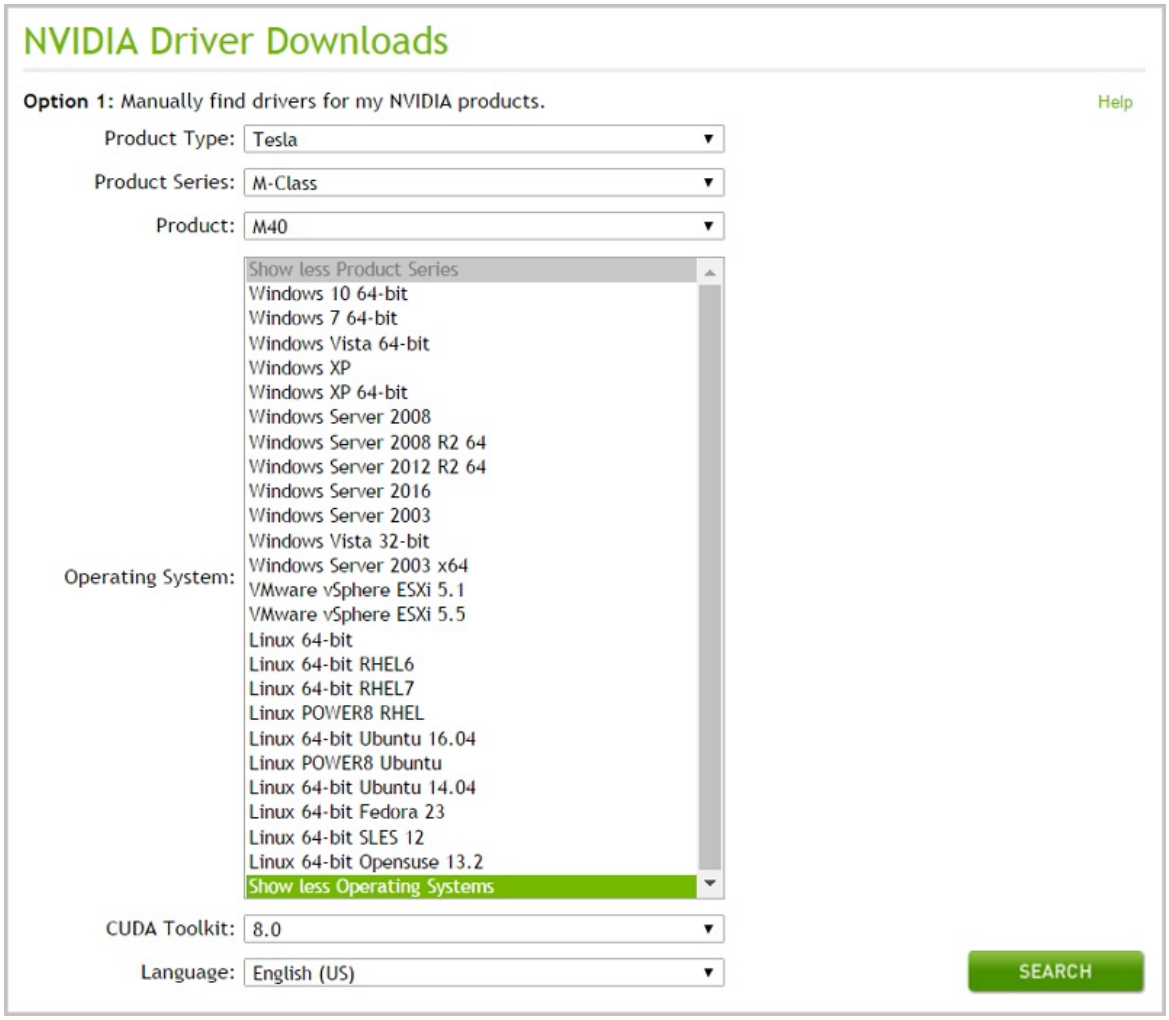
If you do not select to automatically install a GPU driver when you create an instance, you must manually install the GPU driver after the instance is created. This topic describes how to manually install a GPU driver on a GPU-accelerated instance.

Context

You can install only GRID drivers on vgn6i or vgn5i instances. Therefore, the procedure in this topic does not apply to vgn6i or vgn5i instances. For more information about how to install GRID drivers, see Create an NVIDIA GPU-accelerated instance and Install NVIDIA GRID drivers on vgn6i or vgn5i Linux instances.

Procedure

1. Go to the NVIDIA official website.
2. Manually search for the suitable driver.



i. Select a product type, series, and product based on the GPU with which your instance type is equipped. The following table lists the information of GPUs with which different instance types are equipped.

Item	gn4	gn5	gn5i	gn6v	gn6i
Product type	Tesla	Tesla	Tesla	Tesla	Tesla
Product series	M-Class	P-Series	P-Series	V-Series	T-Series
Product	M40	Tesla P100	Tesla P4	Tesla V100	Tesla T4

ii. Select an operating system based on the image of the instance. If your instance runs Debian, select **Linux 64-bit**. If the drop-down list does not contain the operating system that you instance uses, click **Show All Operating Systems** in the lower part of the drop-down list.

iii. Select a CUDA Toolkit version.

iv. Select a language.

v. Click **SEARCH**.

3. Confirm the driver information, and click **DOWNLOAD**. If your instance runs a Linux operating system, do not close the download page. When you install the GPU driver, you may need to refer to the installation steps in the **ADDITIONAL INFORMATION** section.
4. Install the GPU driver.
   - Install the GPU driver on a Linux instance:
     a. Download and install the kernel-devel and kernel-header packages based on your kernel version.

        > **Note** The mismatch between kernel and kernel-devel versions results in a driver compilation error when the driver is installed from the .rpm file. You can run the **rpm -qa | grep kernel** command on an instance to check whether the versions match. Make sure that the versions match and re-install the driver.

     b. Run the **sudo rpm -qa | grep $(uname -r)** command to check whether the kernel-devel and kernel-header packages are downloaded and installed.

        If information similar to the following content is displayed, the kernel-devel and kernel-header packages are installed. CentOS 7.3 is used in this example.

        ```
        kernel-3.10.0-514.26.2.el7.x86_64
        kernel-headers-3.10.0-514.26.2.el7.x86_64
        kernel-tools-libs-3.10.0-514.26.2.el7.x86_64
        python-perf-3.10.0-514.26.2.el7.x86_64
        kernel-tools-3.10.0-514.26.2.el7.x86_64
        ```

c. Perform the following steps in the **ADDITIONAL INFORMATION** section on the download page to install the GPU driver.

The steps in the **ADDITIONAL INFORMATION** section are shown in the following figure. Linux Ubuntu 14.04 64-bit is used in this example.

- Install the GPU driver on a Windows instance:

  Double-click the package and follow the prompts to complete the installation.

> **Note** On Windows instances where the installed GPU drivers have taken effect, Windows Remote Desktop Protocol (RDP) may not support DirectX- and OpenGL-based applications. In this case, you must install the Virtual Network Computing (VNC) service and client or use other protocols that support these applications, such as PC over IP (PCoIP) and XenDesktop HDX 3D.

5.2. Manually uninstall the GPU driver

You can manually uninstall the GPU driver. The uninstall command varies depending on the installation method of the GPU driver and the image type.

Context

The `root` account is used in this topic. If you are a common user, run a `sudo` command to obtain the permissions of a `root` user before you proceed.

If you have installed the GPU driver by using the automatic installation feature, the uninstall method of the GPU driver varies depending on the operating system type:

- The run mode is used for CentOS and SUSE.
- The deb mode is used for Ubuntu.

Uninstall the GPU driver in Ubuntu

If you have used the deb package to install the GPU driver, we recommend that you perform the following operations to uninstall the GPU driver. Driver 410.104, CUDA 10.0.130, and cuDNN 7.5.0 are used in the examples.

1. Run the following command to uninstall the GPU driver:

```
apt-get remove --purge nvidia-*
```

2. Run the following commands to uninstall CUDA and the cuDNN library:

```
apt autoremove --purge cuda-10-0
rm -rf /usr/local/cuda-10.0
```

3. Run the following command to restart the instance:

```
reboot
```

If you have used the run package to install the GPU driver, we recommend that you perform the following operations to uninstall the GPU driver:

1. Run the following command to uninstall the GPU driver:

```
/usr/bin/nvidia-uninstall
```

2. Run the following commands to uninstall CUDA and the cuDNN library:

```
/usr/local/cuda/bin/cuda-uninstaller
rm -rf /usr/local/cuda-10.0
```

> **Note** The uninstall command may differ between CUDA versions. If the *cuda-uninstaller* file does not exist, check whether a file whose name starts with *uninstall_cuda* exists in the */usr/local/cuda/bin/* directory. If yes, replace *cuda-uninstaller* in the command with the file name that starts with uninstall_cuda.

3. Run the following command to restart the instance:

```
reboot
```

Uninstall the GPU driver in CentOS

If you have used the run package to install the GPU driver, we recommend that you perform the following operations to uninstall the GPU driver. Driver 410.104, CUDA 10.0.130, and cuDNN 7.5.0 are used in the examples.

1. Run the following command to uninstall the GPU driver:

```
/usr/bin/nvidia-uninstall
```

2. Run the following commands to uninstall CUDA and the cuDNN library:

```
/usr/local/cuda/bin/cuda-uninstaller
rm -rf /usr/local/cuda-10.0
```

> **Note** The uninstall command may differ between CUDA versions. If the *cuda-uninstaller* file does not exist, check whether a file whose name starts with *uninstall_cuda* exists in the */usr/local/cuda/bin/* directory. If yes, replace *cuda-uninstaller* in the command with the file name that starts with *uninstall_cuda*.

3. Run the following command to restart the instance:

```
reboot
```

If you have used the RPM package to install the GPU driver in CentOS 7, we recommend that you perform the following operations to uninstall the GPU driver:

1. Run the following commands to uninstall the GPU driver:

```
yum remove xorg-x11-drv-nvidia nvidia-kmod cuda-drivers
yum remove nvidia-diag-driver-local-repo-rhel7-410.104
```

2. Run the following commands to uninstall CUDA and the cuDNN library:

```
yum remove /usr/local/cuda-10.0
rm -rf /usr/local/cuda-10.0
```

3. Run the following command to restart the instance:

```
reboot
```

If you have used the RPM package to install the GPU driver in CentOS 6, we recommend that you perform the following operations to uninstall the GPU driver:

1. Run the following commands to uninstall the GPU driver:

```
yum remove xorg-x11-drv-nvidia nvidia-kmod cuda-drivers
yum remove nvidia-diag-driver-local-repo-rhel6-410.104
```

2. Run the following command to uninstall CUDA:

```
yum remove /usr/local/cuda-10.0
```

3. Run the following command to restart the instance:

```
reboot
```

Uninstall the GPU driver in SUSE

If you have used the run package to install the GPU driver, we recommend that you perform the following operations to uninstall the GPU driver. CUDA 9.0.176 is used in the examples.

1. Run the following command to uninstall the GPU driver:

```
/usr/bin/nvidia-uninstall
```

2. Run the following commands to uninstall CUDA and the cuDNN library:

```
/usr/local/cuda/bin/uninstall_cuda_9.0.pl
rm -rf /usr/local/cuda-9.0
```

3. Run the following command to restart the instance:

```
reboot
```

5.3. Install an NVIDIA GRID driver on a GPU-accelerated Linux instance

This topic describes how to install an NVIDIA GRID driver and build a desktop environment on a GPU-accelerated Linux instance.

Prerequisites

- A GPU-accelerated instance that can access the Internet is created.

  > **Note** This topic describes how to install NVIDIA GRID drivers on GPU-accelerated Linux instances. For GPU-accelerated Windows instances, you can select paid images that have NVIDIA GRID drivers pre-installed when you create the instances. For more information, see Create an NVIDIA GPU-accelerated instance.

  When you create an instance, we recommend that you select an image from the **Public Image** drop-down list. If you select an image pre-installed with the NVIDIA GRID driver in the **Image Marketplace** dialog box, you must disable the nouveau driver after you create the instance.

  nouveau is an open source driver. It must be disabled before you install another driver. You can create a *nouveau.conf* file in the */etc/modprobe.d* directory and add `blacklist nouveau` to the file to disable nouveau.

- A VNC application is installed. VNC Viewer is used in this example.
- A GRID license is obtained from NVIDIA. You must build a license server. You can purchase an ECS instance and build a license server by following the tutorial on the NVIDIA official website.

Context

You must install an NVIDIA GRID driver if your GPU-accelerated instances need to support Open Graphics Library (OpenGL). By default, the NVIDIA GRID license that is granted to NVIDIA GPUs such as P100, P4, and V100 is not activated. You can activate the license by using a trial license to use OpenGL.

> **Note** Only NVIDIA partners can download the driver from the official NVIDIA website. This topic describes how to obtain the installation package of the NVIDIA GRID driver from Alibaba Cloud.

This topic describes how to install NVIDIA GRID drivers on GPU-accelerated instances that are not equipped with vGPUs. For information about how to install NVIDIA GRID drivers on vgn6i or vgn5i GPU-accelerated instances that are equipped with vGPUs, see Install NVIDIA GRID drivers on vgn6i or vgn5i Linux instances.

Procedure

Perform the following operations to install an NVIDIA GRID driver:

- Ubuntu 16.04 64-bit:
  i. Install an NVIDIA GRID driver on a Linux instance that runs Ubuntu 16.04 64-bit
  ii. Test the NVIDIA GRID driver installed on an instance that runs Ubuntu 16.04 64-bit
- CentOS 7.3 64-bit:
  i. Install an NVIDIA GRID driver on a Linux instance that runs CentOS 7.3 64-bit
  ii. Test the NVIDIA GRID driver installed on an instance that runs CentOS 7.3 64-bit

Install an NVIDIA GRID driver on a Linux instance that runs Ubuntu 16.04 64-bit

1. Connect to a Linux instance. For more information, see Connect to a Linux instance by using a username and password.
2. Run the following commands in sequence to upgrade the system and install KDE:

```
apt-get update
apt-get upgrade
apt-get install kubuntu-desktop
```

3. Run the `reboot` command to restart the system.
4. Connect to the Linux instance again. Run the following commands to download and decompress the NVIDIA GRID driver package.

   The NVIDIA GRID driver package contains the drivers for various operating systems. For Linux, select *NVIDIA-Linux-x86_64-410.39-grid.run*.

```
wget http://grid-9-4.oss-cn-hangzhou.aliyuncs.com/NVIDIA-Linux-x86_64-430.99-grid.run
```

5. Run the following commands in sequence and follow the on-screen tips to install the NVIDIA GRID driver:

```
chmod 777 NVIDIA-Linux-x86_64-430.99-grid.run
./NVIDIA-Linux-x86_64-430.99-grid.run
```

6. Run the `nvidia-smi` command to test whether the NVIDIA GRID driver is installed.

If a command output similar to the following one is displayed, the NVIDIA GRID driver is installed.

```
root@i                    :~# nvidia-smi
Thu Nov  5 17:01:38 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 430.99       Driver Version: 430.99       CUDA Version: 10.1     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:00:07.0 Off |                    0 |
| N/A   31C    P0    39W / 300W |      0MiB / 16160MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

7. Add a license server and activate the license.
   i. Run the `cd /etc/nvidia` command to go to the */etc/nvidia* directory.
   ii. Run the `cp gridd.conf.template gridd.conf` command to create a file named *gridd.conf*.
   iii. Add the license server information to the *gridd.conf* file.

```
ServerAddress=<IP address of the license server>
ServerPort=<Port of the license server (default port: 7070)>
FeatureType=2
EnableUI=TRUE
```

8. Run the following command to install x11vnc:

```
apt-get install x11vnc
```

9. Run the `lspci | grep NVIDIA` command to query the GPU BusID.

   In this example, the GPU BusID is `00:07.0`.

10. Configure the X Server environment and restart the system.
    i. Run the `nvidia-xconfig --enable-all-gpus --separate-x-screens` command.
    ii. Add the GPU BusID that you obtained to `Section "Device"` in the */etc/X11/xorg.conf* file. In this example, `BusID "PCI:0:7:0"` is added.

```
Section "Device"
    Identifier     "Device0"
    Driver         "nvidia"
    VendorName     "NVIDIA Corporation"
    BoardName      "Tesla P4"
    BusID          "PCI:0:7:0"
EndSection
```

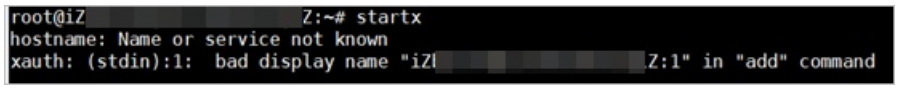
    iii. Run the `reboot` command to restart the system.

Test the NVIDIA GRID driver installed on an instance that runs Ubuntu 16.04 64-bit

1. Run the following command to install the GLX application:

   ```
   apt-get install mesa-utils
   ```

2. Run the `startx` command to start X Server.
   - If the `startx` command is unavailable, run the `apt-get install xinit` command to install the GLX application.
   - If you run the `startx` command, the `hostname: Name or service not known` error may be reported. This error does not affect the startup of X Server. You can run the `hostname` command to query the hostname of the instance. Then, you can modify the */etc/hosts* file by replacing the `hostname` that follows `127.0.0.1` with the actual hostname of your instance.

3. Start a new terminal session of the SSH client and run the following command to start x11vnc:

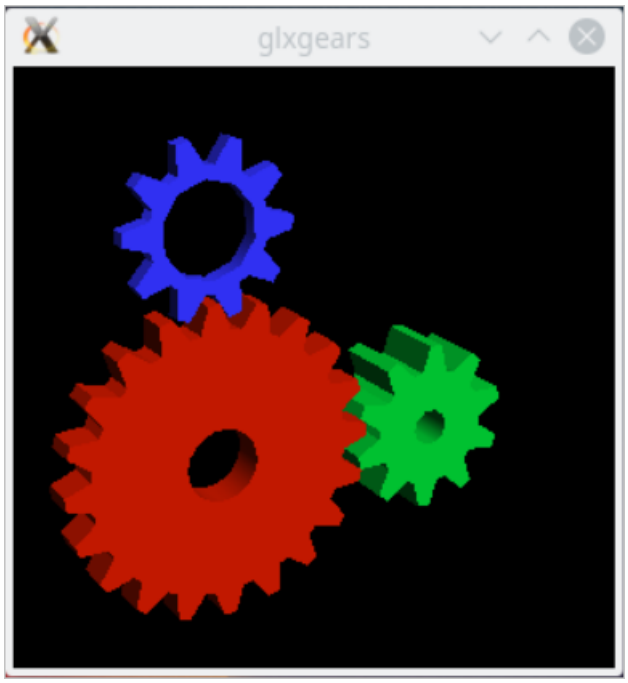
   ```
   x11vnc -display :1
   ```

   If a command output similar to the following one is displayed, x11vnc is started. In this case, you can connect to the instance by using a VNC application. In this example, VNC Viewer is used.

4. Log on to the ECS console and add security group rules to a security group to which the instance is added. The security group rules allow inbound traffic on TCP port 5900. For more information, see Add security group rules.
5. On the local machine, start VNC Viewer and enter `<Public IP address of the instance>:5900` to connect to the instance and go to KDE.
6. Run the `glxinfo` command to view the configurations supported by the current NVIDIA GRID driver.
   i. Start a new terminal session of the SSH client.
   ii. Run the `export DISPLAY=:1` command.
   iii. Run the `glxinfo -t` command to list the configurations supported by the current NVIDIA GRID driver.
7. Run the `glxgears` command to test the NVIDIA GRID driver.
   i. On KDE, right-click the desktop and select **Run Command**.

ii. Run the `glxgears` command to start the testing application.
If a window similar to the following one is displayed, the NVIDIA GRID driver runs normally.



Install an NVIDIA GRID driver on a Linux instance that runs CentOS 7.3 64-bit

1. Connect to a Linux instance. For more information, see Connect to a Linux instance by using a username and password.
2. Run the following commands in sequence to upgrade the system and install KDE:

```
yum update
yum install kernel-devel
yum groupinstall "KDE Plasma Workspaces"
```

3. Run the `reboot` command to restart the system.
4. Connect to the Linux instance again. Run the following command to download and decompress the NVIDIA GRID driver package.

   The NVIDIA GRID driver package contains the drivers for various operating systems. For Linux, select *NVIDIA-Linux-x86_64-430.99-grid.run*.

```
wget http://grid-9-4.oss-cn-hangzhou.aliyuncs.com/NVIDIA-Linux-x86_64-430.99-grid.run
```

5. Disable the nouveau driver:
   i. Run the `vim /etc/modprobe.d/blacklist.conf` command and add `blacklist nouveau` to the file.
   ii. Run the `vim /lib/modprobe.d/dist-blacklist.conf` command and add the following content:

```
blacklist nouveau
options nouveau modeset=0
```

iii. Run the `mv /boot/initramfs-$(uname -r).img /boot/initramfs-$(uname -r)-nouveau.img` command.

iv. Run the `dracut /boot/initramfs-$(uname -r).img $(uname -r)` command.

6. Run the `reboot` command to restart the system.
7. Run the following commands in sequence and follow the on-screen tips to install the NVIDIA GRID driver:

```
chmod 777 NVIDIA-Linux-x86_64-430.99-grid.run
./NVIDIA-Linux-x86_64-430.99-grid.run
```

8. Run the `nvidia-smi` command to test whether the NVIDIA GRID driver is installed.
If a command output similar to the following one is displayed, the NVIDIA GRID driver is installed.

```
[root@i                    ~]# nvidia-smi
Thu Nov  5 16:31:22 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 430.99       Driver Version: 430.99       CUDA Version: 10.1     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  On   | 00000000:00:07.0 Off |                    0 |
| N/A   30C    P0    23W / 300W |      0MiB / 16160MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

9. Add a license server and activate the license.

i. Run the `cd /etc/nvidia` command to go to the */etc/nvidia* directory.

ii. Run the `cp gridd.conf.template gridd.conf` command to create a file named *gridd.conf*.

iii. Add the license server information to the *gridd.conf* file.

```
ServerAddress=<IP address of the license server>
ServerPort=<Port of the license server (default port: 7070)>
FeatureType=2
EnableUI=TRUE
```

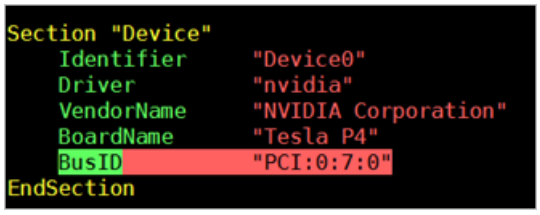
10. Run the following command to install x11vnc:

```
yum install x11vnc
```

11. Run the `lspci | grep NVIDIA` command to query the GPU BusID.In this example, the GPU BusID is `00:07.0`.
12. Configure the X Server environment.

i. Run the `nvidia-xconfig --enable-all-gpus --separate-x-screens` command.

ii. Add your GPU BusID to `Section "Device"` in the */etc/X11/xorg.conf* file. In this example, `BusID "PCI:0:7:0"` is added.



13. Run the `reboot` command to restart the system.

Test the NVIDIA GRID driver installed on an instance that runs CentOS 7.3 64-bit

1. Run the `startx` command to start X Server.
2. Start a new terminal session of the SSH client and run the following command to start x11vnc:

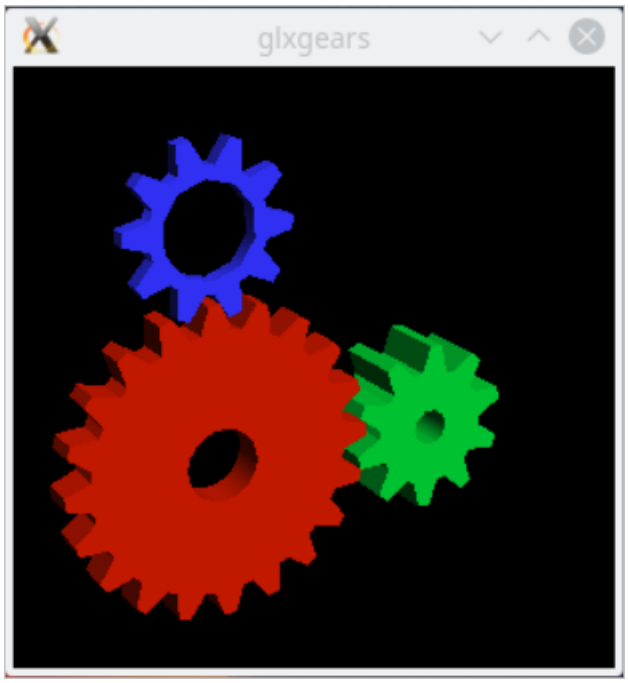
   ```
   x11vnc -display :0
   ```

   If a command output similar to the following one is displayed, x11vnc is started. In this case, you can connect to the instance by using a VNC application. In this example, VNC Viewer is used.

3. Log on to the ECS console and add security group rules to a security group of the instance. The security group rules allow inbound traffic on TCP port 5900. For more information, see Add security group rules.
4. On the local machine, start VNC Viewer and enter `<Public IP address of the instance>:5900` to connect to the instance and go to KDE.
5. Run the `glxinfo` command to view the configurations supported by the current NVIDIA GRID driver.
   i. Start a new terminal session of the SSH client.
   ii. Run the `export DISPLAY=:0` command.
   iii. Run the `glxinfo -t` command to list the configurations supported by the current NVIDIA GRID driver.
6. Run the `glxgears` command to test the NVIDIA GRID driver.
   i. On KDE, right-click the desktop and select **Run Command**.

ii. Run the `glxgears` command to start the testing application.
If a window similar to the following one is displayed, the NVIDIA GRID driver runs normally.



5.4. Install NVIDIA GRID drivers on vgn6i or vgn5i Linux instances

You must install an NVIDIA GRID driver if your GPU-accelerated instances require Open Graphics Library (OpenGL). By default, the NVIDIA GRID license granted to NVIDIA GPUs is not activated. You must purchase and activate the license to use OpenGL. This topic describes how to install an NVIDIA GRID driver and activate the GRID license. vgn6i or vgn5i lightweight GPU-accelerated instances that run the Ubuntu 16.04 64-bit operating system are used in the example.

Prerequisites

- A vgn6i or vgn5i instance that can access the Internet is created. We recommend that you select an image from the **Public Image** tab when you create an instance.

  > **Note** This topic describes how to install GRID drivers on Linux instances. For Windows instances, you can select paid images that have GRID drivers pre-installed when you create the instances. For more information, see Create an NVIDIA GPU-accelerated instance.

- A remote connection tool such as VNC Viewer is installed on your local machine.
- A GRID license is obtained from NVIDIA. You must build a license server. You can purchase an ECS instance and build the license server by following the tutorial on the official NVIDIA website.

Context

This topic describes how to install GRID drivers on vgn6i or vgn5i GPU-accelerated instances that are equipped with vGPUs. For information about how to install GRID drivers on GPU-accelerated instances that are not equipped with vGPUs, see Install an NVIDIA GRID driver on a GPU-accelerated Linux instance.

Procedure

1. Disable nouveau.nouveau is an open source driver. It must be disabled before you can install another driver.
   i. Connect to the Linux instance. For more information, see Overview.
   ii. Check whether the *blacklist-nouveau.conf* file exists.

   ```
   ls /etc/modprobe.d/blacklist-nouveau.conf
   ```

   iii. If the *blacklist-nouveau.conf* file exists and contains the following content, skip this step. If not, run the `vim /etc/modprobe.d/blacklist-nouveau.conf` command to create the file. Then, add the following content to the file to disable nouveau:

   ```
   blacklist nouveau
   blacklist lbm-nouveau
   options nouveau modeset=0
   ```

   iv. Generate kernel initramfs.

   ```
   rmmod nouveau
   update-initramfs -u
   ```

   v. Restart the instance.

   ```
   reboot
   ```

2. Download the NVIDIA GRID driver package.
   i. Connect to the Linux instance. For more information, see Overview.
   ii. Download the NVIDIA GRID driver package.
   - vgn5i GRID guest driver package:

     ```
     wget http://nvidia-418.oss-cn-shenzhen.aliyuncs.com/NVIDIA-Linux-x86_64-418.70-grid.run
     ```

   - vgn6i GRID guest driver package:

     ```
     wget http://grid-9-2.oss-cn-hangzhou.aliyuncs.com/NVIDIA-Linux-x86_64-430.63-grid.run
     ```

3. Install the NVIDIA GRID driver.
   - vgn5i

     ```
     chmod +x NVIDIA-Linux-x86_64-418.70-grid.run
     ./NVIDIA-Linux-x86_64-418.70-grid.run
     ```

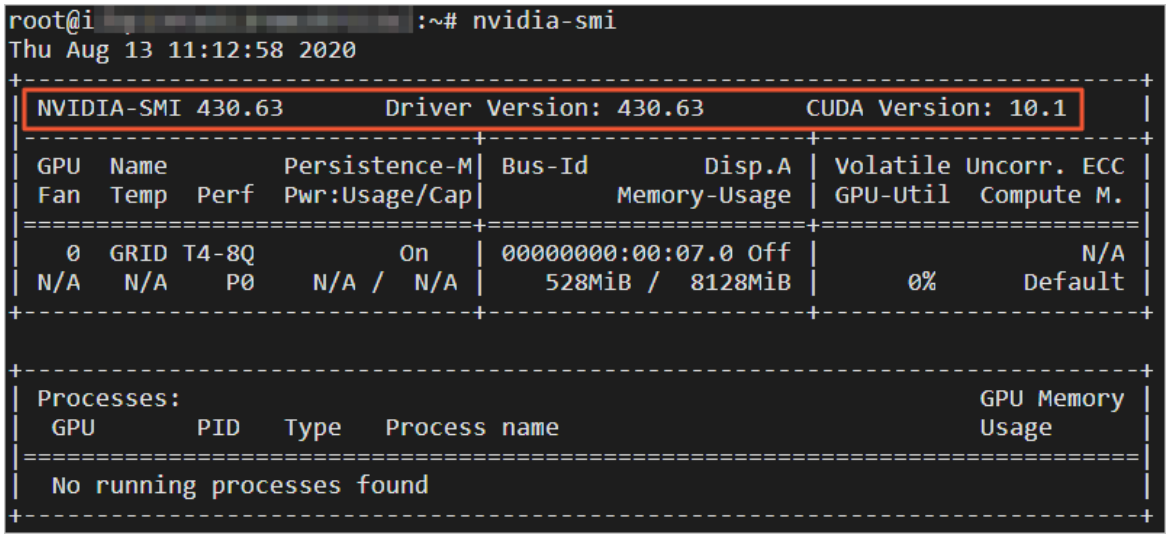
   - vgn6i

     ```
     chmod +x NVIDIA-Linux-x86_64-430.63-grid.run
     ./NVIDIA-Linux-x86_64-430.63-grid.run
     ```

4. Test whether the NVIDIA GRID driver is installed.

```
nvidia-smi
```

If a command output similar to the following one is displayed, the NVIDIA GRID driver is installed.



5. Add a license server.
   i. Go to the */etc/nvidia* directory.

      ```
      cd /etc/nvidia
      ```

   ii. Create a file named *gridd.conf*.

      ```
      cp gridd.conf.template gridd.conf
      ```

   iii. Add the license server information to the *gridd.conf* file.

      ```
      ServerAddress=<IP address of the license server>
      ServerPort=<Port of the license server (default port: 7070)>
      FeatureType=1
      ```

6. Restart the instance for the license server configurations to take effect.

   ```
   reboot
   ```

7. Check whether the license is activated.
   i. Connect to the Linux instance. For more information, see Overview.

ii. Check the license status.

```
systemctl status nvidia-gridd
```

If **License acquired successfully** is displayed, the license is activated.