ALIBABA CLOUD

# Alibaba Cloud

## Elastic GPU Service

## User Guide

Document Version: 20220704

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.Installation guideline for NVIDIA drivers

If your GPU-accelerated instances are not configured with drivers, you must install NVIDIA drivers to ensure the performance of your instances. The driver types that you can install on the instances may vary based on the scenarios and the instance families. This topic describes how to create GPU-accelerated instances and install NVIDIA drivers on the instances in different scenarios.

## Drivers

You can install the following NVIDIA drivers on Alibaba Cloud GPU-accelerated instances:

- GPU driver: drives physical GPUs.
- GRID driver: accelerates graphics processing.

## Install drivers on vGPU-accelerated instances

The instances of vGPU-accelerated instance families such as vgn6i and vgn5i are configured with vGPUs that are generated from GPU virtualization with mediated pass-through. You can install only GRID drivers on the instances. However, some GPU-related features may be unavailable on the vGPU-accelerated instances because the NVIDIA GRID licenses are not activated for the GPUs that the instances use. In this case, you can use the images in which GRID licenses are activated to create vGPU-accelerated instances. You can also activate GRID licenses to use the GPU-related features. The following information describes how to install the drivers.

| OS | Driver type | Scenario | Installation method |
| --- | --- | --- | --- |
| Windows Server | GRID drivers | Graphics computing scenarios, such as Open Graphics Library (OpenGL) and Direct3D scenarios | • We recommend that you apply for the licenses of GRID drivers, download the installation packages of the drivers, and then install the drivers on vGPU-accelerated instances. To apply for the licenses, submit a ticket. For more information, see Install a GRID driver on a Windows GPU-accelerated instance.<br>• If you have purchased GRID licenses, we recommend that you purchase only vGPU-accelerated instances that are not configured with drivers. For more information, see Create a GPU-accelerated instance that is not configured with a driver.<br>**Note** After you create the vGPU-accelerated instances, you must install GRID drivers on the instances. To install the drivers, contact your license provider. |

| OS | Driver type | Scenario | Installation method |
| --- | --- | --- | --- |
| Linux, such as Alibaba Cloud Linux, CentOS, Ubuntu, or SUSE Linux | GRID drivers | Common computing scenarios, such as deep learning and AI | We recommend that you apply for the licenses of GRID drivers before you install the drivers. To apply for the licenses, submit a ticket. For more information, see Install a GRID driver on a Linux vGPU-accelerated instance. |

## Install drivers on GPU-accelerated compute-optimized instances

GPU-accelerated compute-optimized instance families are classified into the following types:

- GPU-accelerated compute-optimized instance family: gn7i, gn7, gn6i, gn6e, gn6v, gn5i, and gn5
- GPU-accelerated compute-optimized ECS Bare Metal Instance family: ebmgn7i, ebmgn7, ebmgn6e, ebmgn6v, ebmgn6i, ebmgn5, and ebmgn5i

> **Note** The instances of gn7 and ebmgn7 instance families are suitable only for common computing scenarios, such as deep learning, AI, and scientific computing. You must install GPU drivers on the instances.

The following information describes driver types that you can install and how to install the drivers in different scenarios:

- Common computing scenarios such as deep learning, AI, and scientific computing

| OS | Driver type | Installation method |
| --- | --- | --- |
| Linux, such as Alibaba Cloud Linux, CentOS, Ubuntu, or SUSE Linux | GPU drivers | ◦ When you create GPU-accelerated compute-optimized instances, we recommend that you click Public Image and select Auto-install GPU Driver. When the instances are started for the first time, the drivers are installed. For more information, see Create a Linux GPU-accelerated instance configured with a GPU driver.<br>◦ If you cannot find public images of the required OS types or versions, we recommend that you create GPU-accelerated compute-optimized instances that are not configured with drivers and install GPU drivers that you have downloaded from the NVIDIA official website on the instances. For more information about how to install Linux GPU drivers, see the following references:<br>■ Create a GPU-accelerated instance that is not configured with a driver<br>■ Install a GPU driver on a Linux GPU-accelerated compute-optimized instance |

| OS | Driver type | Installation method |
| --- | --- | --- |
| Windows Server | GPU drivers | You cannot configure automatic installation for GPU drivers that run Windows when you create GPU-accelerated compute-optimized instances. We recommend that you create GPU-accelerated compute-optimized instances that are not configured with drivers and install GPU drivers that you have downloaded from the NVIDIA official website on the instances. For more information about how to install Windows Server GPU drivers, see the following references:<br>◦ Create a GPU-accelerated instance that is not configured with a driver<br>◦ Install a Windows GPU driver on a GPU-accelerated compute-optimized instance |

- Graphics computing scenarios such as OpenGL and Direct3D scenarios

| OS | Driver type | Installation method |
| --- | --- | --- |
| Windows Server | GRID drivers | ◦ We recommend that you apply for the licenses of GRID drivers, download the installation packages of the drivers, and then install the drivers on GPU-accelerated compute-optimized instances. To apply for the licenses, submit a ticket. For more information, see Install a GRID driver on a Windows GPU-accelerated instance.<br>◦ If you have purchased GRID licenses, we recommend that you purchase only GPU-accelerated compute-optimized instances that are not configured with drivers. For more information, see Create a GPU-accelerated instance that is not configured with a driver.<br>**Note** After you create the GPU-accelerated compute-optimized instances, you must install GRID drivers on the instances. To install the drivers, contact your license provider. |

| OS | Driver type | Installation method |
| --- | --- | --- |
| Linux, such as Alibaba Cloud Linux, CentOS, Ubuntu, or SUSE Linux | GPU drivers | ◦ When you create GPU-accelerated compute-optimized instances, we recommend that you click Public Image and select Auto-install GPU Driver. When the instances are started for the first time, the drivers are installed. For more information, see Create a Linux GPU-accelerated instance configured with a GPU driver.<br>◦ If you cannot find public images of the required OS types or versions, we recommend that you create GPU-accelerated compute-optimized instances that are not configured with drivers and install GPU drivers that you have downloaded from the NVIDIA official website on the instances. For more information about how to install Linux GPU drivers, see the following references:<br>■ Create a GPU-accelerated instance that is not configured with a driver<br>■ Install a GPU driver on a Linux GPU-accelerated compute-optimized instance |

# 2.Quick reference

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

## 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 instance type or network configuration of your instance is unsuitable for your business, you can change the instance type, IP address, and maximum public 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 operating system of your instance 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 manage ECS instances in a fine-grained manner:
  - User data
  - Instance 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 implement automatic scaling by combining preemptible instances 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 disk.
2. Attach a data 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 for a disk.
5. If the capacity of an existing cloud disk cannot meet your requirements, resize the disk. For more information, see the following topics:
   - Resize disks online for Linux instances
   - Resize disks offline for Linux instances
   - Resize disks online for Windows instances
   - Resize disks offline for Windows instances
6. If a data error occurs on a cloud disk, 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, reinitialize the cloud disk. For more information, see Re-initialize a data disk.
8. Detach a data disk.
9. Release a disk.

## Create and manage snapshots

You can perform the following steps to use a snapshot:

1. Create a snapshot. You can use one of the following methods to manually or automatically create a snapshot:
   - Create a snapshot for a disk.
   - Use an automatic snapshot policy to automatically create snapshots 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 following section describes the common scenarios of snapshots:

- 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 deploy an environment, 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 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 restore all or some 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 manage 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.

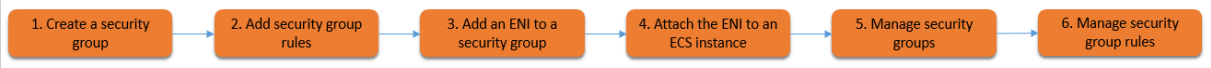
## Create and manage SSH key pairs

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 elastic network interface (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. Delete an ENI.

## Use tags

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

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 use a launch template:

1. Create a launch template.
2. Create a template version.
3. Delete a launch template and a specified template version.

## Create and manage deployment sets

Deployment sets help you implement high availability for underlying applications. You can perform the following steps to use 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.

# 3.Create a GPU-accelerated instance

## 3.1. Create a Linux GPU-accelerated instance configured with a GPU driver

In scenarios in which graphics computing is not required, such as deep learning and AI, we recommend that you use a GPU-accelerated compute-optimized instance configured with a GPU driver. This topic describes how to create a Linux GPU-accelerated compute-optimized instance configured with a GPU driver that supports automatic installation.

### Context

Alibaba Cloud allows you to configure GPU drivers that support automatic installation only when you create specific GPU-accelerated instances and use Linux public images. The instances must belong to GPU-accelerated compute-optimized families, such as the GPU-accelerated compute-optimized instance family and the GPU-accelerated compute-optimized Elastic Compute Service (ECS) Bare Metal Instance family. You cannot configure GPU drivers that support automatic installation in the following scenarios:

- You want to purchase a Windows GPU-accelerated instance and a GPU driver. In this case, you must purchase a Windows GPU-accelerated instance that is not configured with a driver and install a GPU driver. For more information, see Create a GPU-accelerated instance that is not configured with a driver and Install a Windows GPU driver on a GPU-accelerated compute-optimized instance.
- You cannot find a public image of the required Linux distribution or version. In this case, you must create a GPU-accelerated instance that is not configured with a GPU driver and install a GPU driver that you purchased from the NVIDIA official website. For more information, see Create a GPU-accelerated instance that is not configured with a driver and Install a GPU driver on a Linux GPU-accelerated compute-optimized instance.
- You want to use a shared image or custom image when you create a GPU-accelerated instance. In this case, you must install a GPU driver after you create the instance.

The installation methods and driver types may vary based on individual use cases. For more information, see Installation guideline for NVIDIA drivers.

### Preparations

1. Create an Alibaba Cloud account and complete account information.
   - Create an Alibaba Cloud account. For more information, see Sign up with Alibaba Cloud.
   - Complete real-name verification because purchase ECS instances in the Chinese mainland. For more information, see Real-name Registration FAQs.
2. Go to the Custom Launch tab of the instance buy page in the ECS console.

### Procedure

- Step 1: Complete the settings in the Basic Configurations step
- Step 2: Complete the settings in the Networking step
- Step 3: Complete the settings in the System Configurations step

- Step 4: (Optional) Complete the settings in the Grouping step
- Step 5: Complete the settings in the Preview step

## Step 1: Complete the settings in the Basic Configurations step

In the Basic Configurations step, you can configure the basic parameters and resources that are required to purchase an instance. The basic parameters include the billing method, region, and zone. The basic resources include the instance type, image, and storage. After you complete the settings in the Basic Configurations step, click **Next**.

1. Select a billing method.

   The billing method determines how the billing and charging rules are applied to an instance. The billing method also determines how the status of the resources on the instance is changed.

| Billing method | Description | References |
| --- | --- | --- |
| **Subscription** | You pay for resources before you use them. | Subscription |
| **Pay-as-you-go** | You pay for resources after you use them. The billing cycles of pay-as-you-go instances are accurate to the second. You can purchase and release instances based on your business requirements.<br>**Note** We recommend that you use this billing method together with savings plans to reduce costs. | ○ Pay-as-you-go<br>○ Savings plans |
| **Preemptible Instance** | You pay for resources after you use them. The price of a preemptible instance is lower than the price of a pay-as-you-go instance. However, the system may release preemptible instances due to fluctuations in the market price or insufficient resources of instance types. | Preemptible instances |

2. Select a region and a zone.

   Select a region that is close to your geographical location to reduce latency. After an instance is created, the region and the zone of the instance cannot be changed. For more information, see Regions and zones.

3. Select an instance type and configure the relevant settings.

i. Set the Architecture parameter to **Heterogeneous Computing** and set the Category parameter to **Compute Optimized Type with GPU**. Alternatively, set the Architecture parameter to **ECS Bare Metal Instance** and set the Category parameter to **GPU Type**. Then, select an instance type.

> Note
>
> - The available instance types vary based on the selected region. To view the instance types that can be used in each region, go to the ECS Instance Types Available for Each Region page.
> - You may have specific requirements on the settings. For example, you may want to attach multiple elastic network interfaces (ENIs), enhanced SSDs (ESSDs), or local disks to the instance. You must make sure that the selected instance type meets the requirements. For information about the features, scenarios, and specifications of instance types, see Instance family.
> - If you want to purchase an instance for a specific scenario, click the **Scenario-based Selection** tab to view the instance types that are recommended for different scenarios. For example, you can set the Business Scenario parameter to AI Machine Learning to view the recommended GPU-accelerated instance types.

ii. Check whether the value of the **Selected Instance Type** parameter is the same as the selected instance type.

iii. If you set **Billing Method** to **Preemptible Instance**, configure the Use Duration and Maximum Price for Instance Type parameters.

Use Duration specifies the protection period of a preemptible instance. After the protection period ends, the instance may be released due to insufficient resources or a lower bid than the market price. The following table describes the valid values of the Use Duration parameter.

| Value | Description |
| --- | --- |
| **One Hour** | After the preemptible instance is created, it enters a 1-hour protection period during which it cannot be automatically released. |
| **None** | The preemptible instance is created without a protection period. Preemptible instances without a protection period are lower-cost than preemptible instances with a protection period. |

The following table describes the valid values of the Maximum Price for Instance Type parameter.

| Value | Description |
| --- | --- |
| **Use Automatic Bid** | The real-time market price of the instance type is automatically used. The price can be up to but cannot exceed the pay-as-you-go price of the instance type. Automatic bidding can prevent the preemptible instance from being released due to lower bids than the market price, but cannot prevent the instance from being released due to insufficient resources. |
| **Set Maximum Price** | You must specify a maximum price. If the real-time market price exceeds your specified maximum price or if available resources are insufficient, the preemptible instance is released. |

iv. Specify the number of instances to create.

You can create a maximum of 100 instances at a time by using the wizard. In addition, the number of instances within your account cannot exceed your instance quota. The instance quota is displayed on the buy page. For more information, see View and increase instance quotas.

4. Select an image.

i. In the Image section, click Public Image and select the Linux distribution and version that you want to use.

ii. Select **Auto-install GPU Driver**, and determine whether to select **AIACC-Training** and **AIACC-Inference** based on your business requirements. Then, select the versions of the CUDA library, GPU driver, and cuDNN library that you want to use.

> **Note** If you select an instance of the sccgn7ex GPU-accelerated compute-optimized Super Computing Cluster (SCC) instance family, you can determine whether to install a remote direct memory access (RDMA) software stack that supports automatic installation based on your business requirements.

The following information describes GPU drivers, RDMA software stacks, AIACC-Training, and AIACC-Inference:

- GPU drivers are used to drive physical GPUs and can work efficiently when used together with the CUDA and cuDNN libraries. If you select Auto-install GPU Driver, a CUDA library and a cuDNN library are installed when you install the GPU driver. You can select Auto-install GPU Driver only when you use specific Linux public images. The following table lists the image versions and the instance families supported for GPU drivers of different versions.

  > **Note** For a new business system, we recommend that you use the latest versions of the GPU driver, CUDA library, and cuDNN library.

| CUDA library version | GPU driver version | cuDNN library version | Supported Alibaba Cloud public image version | Supported instance family |
| --- | --- | --- | --- | --- |
| 11.4.1 | 470.82.01 | 8.2.4 | ■ Alibaba Cloud Linux 2 and Alibaba Cloud Linux 3<br>■ Ubuntu 20.04, Ubuntu 18.04, and Ubuntu 16.04<br>■ CentOS 8.x and CentOS 7.x<br>■ Debian 10.10<br>**Note** Debian 10.10 is supported only for the sccgn7ex instance family. | ■ gn7i, gn7e, gn6v, gn6i, gn6e, gn5, and gn5i<br>■ ebmgn7, ebmgn7i, ebmgn7e, ebmgn6v, ebmgn6i, ebmgn6e, and ebmgn5i<br>■ sccgn7ex |
| 11.2.2 | 460.91.03 | 8.1.1 | ■ Alibaba Cloud Linux 2 and Alibaba Cloud Linux 3<br>■ Ubuntu 20.04, Ubuntu 18.04, and Ubuntu 16.04<br>■ CentOS 8.x and CentOS 7.x | ■ gn7, gn7i, gn7e, gn6v, gn6i, gn6e, gn5, and gn5i<br>■ ebmgn7, ebmgn7i, ebmgn7e, ebmgn6v, ebmgn6i, ebmgn6e, and ebmgn5i |

| CUDA library version | GPU driver version | cuDNN library version | Supported Alibaba Cloud public image version | Supported instance family |
| --- | --- | --- | --- | --- |
| 11.0.2 | 460.91.03 | ■ 8.1.1<br>■ 8.0.4 | ■ Alibaba Cloud Linux 2<br>■ Ubuntu 20.04, Ubuntu 18.04, and Ubuntu 16.04<br>■ CentOS 8.x and CentOS 7.x | ■ gn7, gn7e, gn6v, gn6i, gn6e, gn5, and gn5i<br>■ ebmgn7, ebmgn7e, ebmgn6v, ebmgn6i, ebmgn6e, and ebmgn5i |
| 10.2.89 | 460.91.03 | ■ 8.1.1<br>■ 8.0.4<br>■ 7.6.5 | ■ Alibaba Cloud Linux 2<br>■ Ubuntu 18.04 and Ubuntu 16.04<br>■ CentOS 8.x and CentOS 7.x | ■ gn6v, gn6i, gn6e, gn5, and gn5i<br>■ ebmgn6v, ebmgn6i, ebmgn6e, and ebmgn5i |
| 10.1.168 | ■ 450.80.02<br>■ 440.64.00 | ■ 8.0.4<br>■ 7.6.5<br>■ 7.5.0 | ■ Ubuntu 18.04 and Ubuntu 16.04<br>■ Centos 7.x | ■ gn6v, gn6i, gn6e, gn5, and gn5i<br>■ ebmgn6v, ebmgn6i, ebmgn6e, and ebmgn5i |
| 10.0.130 | ■ 450.80.02<br>■ 440.64.00 | ■ 7.6.5<br>■ 7.5.0<br>■ 7.4.2<br>■ 7.3.1 | ■ Ubuntu 18.04 and Ubuntu 16.04<br>■ Centos 7.x | ■ gn6v, gn6i, gn6e, gn5, and gn5i<br>■ ebmgn6v, ebmgn6i, ebmgn6e, and ebmgn5i |
| 9.2.148 | ■ 450.80.02<br>■ 440.64.00<br>■ 390.116 | ■ 7.6.5<br>■ 7.5.0<br>■ 7.4.2<br>■ 7.3.1<br>■ 7.1.4 | ■ Ubuntu 16.04<br>■ Centos 7.x | ■ gn6v, gn6e, gn5, and gn5i<br>■ ebmgn6v, ebmgn6e, and ebmgn5i |
| 9.0.176 | ■ 450.80.02<br>■ 440.64.00<br>■ 390.116 | ■ 7.6.5<br>■ 7.5.0<br>■ 7.4.2<br>■ 7.3.1<br>■ 7.1.4<br>■ 7.0.5 | ■ Ubuntu 16.04<br>■ Centos 7.x<br>■ SUSE 12sp2 | ■ gn6v, gn6e, gn5, and gn5i<br>■ ebmgn6v, ebmgn6e, and ebmgn5i |

| CUDA library version | GPU driver version | cuDNN library version | Supported Alibaba Cloud public image version | Supported instance family |
| --- | --- | --- | --- | --- |
| 8.0.61 | ■ 450.80.02<br>■ 440.64.00<br>■ 390.116 | ■ 7.1.3<br>■ 7.0.5 | ■ Ubuntu 16.04<br>■ Centos 7.x | ■ gn5 and gn5i<br>■ ebmgn5i |

> **Note** If you want to change the OS of an instance after the instance is created, make sure that GPU drivers can be automatically installed when you use the selected image.

- RDMA software stack

  To further optimize the network performance of GPU-accelerated instances that use the SHENLONG architecture, Alibaba Cloud provides GPU-accelerated compute-optimized SCC instance families, which are named sccgn instance families. sccgn instances provide superior computing power and network communication. RDMA software stacks can be automatically installed for the sccgn7ex instance family. This way, you can use the GPUDirect RDMA feature with ease. For more information, see sccgn instance family.

- AIACC-Training is an AI accelerator that is developed by Alibaba Cloud. AIACC-Training can significantly improve training performance for mainstream AI computing frameworks, such as TensorFlow, PyTorch, MxNet, and Caffe. For more information, see AIACC-Training.

  > **Note** CentOS 6, SUSE Linux, and Alibaba Cloud Linux do not support AIACC-Training.

- AIACC-Inference is an AI accelerator that is developed by Alibaba Cloud. AIACC-Inference can significantly improve inference performance for mainstream AI computing frameworks, such as TensorFlow, and the frameworks that can be converted to the Open Neural Network Exchange (ONNX) format. For more information, see AIACC-Inference.

  > **Note** CentOS 6, SUSE Linux, and Alibaba Cloud Linux do not support AIACC-Inference.

5. Complete the storage and related settings.

   Instances provide storage capabilities based on the system disks, data disks, and Apsara File Storage NAS file systems that are attached to the instances. ECS provides cloud and local disks to meet the storage requirements of different scenarios.

   Cloud disks include ESSDs, standard SSDs, and ultra disks and can be used as system disks or data disks. For more information, see Disks.

   > **Note** The billing method of a cloud disk that is created along with an instance is the same as that of the instance.

Local disks can be used only as data disks. If an instance family (such as instance family with local SSDs and big data instance family) is equipped with local disks, the information of the local disks is displayed. For more information, see Local disks.

> **Note** Local disks cannot be attached to instances on your own.

i. Configure a system disk.

System disks are used to install operating systems. The default capacity of a system disk is 40 GiB. However, the actual minimum capacity is related to the image. The following table describes the capacity ranges of system disks for different images.

| Image | System disk capacity range (GiB) |
| --- | --- |
| Linux (excluding CoreOS and Red Hat) | [max{20, Image size}, 500] |
| FreeBSD | [max {30, Image size}, 500] |
| CoreOS | [max {30, Image size}, 500] |
| Red Hat | [max {40, Image size}, 500] |
| Windows | [max {40, Image size}, 500] |

ii. (Optional)Add data disks.

You can create empty data disks or create data disks from snapshots. A snapshot is a point-in-time backup of a disk. You can import data in a quick manner by creating a disk from a snapshot. When you add a data disk, you can encrypt the disk to meet the requirements of scenarios such as data security and regulatory compliance. For more information about data encryption, see 加密概述.

> **Note** A limited number of data disks can be attached to a single instance. For more information, see the "Elastic Block Storage (EBS) limits" section in Limits.

iii. (Optional)Add NAS file systems.

If you have a large amount of data to share among multiple instances, we recommend that you use a NAS file system to reduce costs in data transmission and synchronization.

Select an existing NAS file system or click **Create a file system** to create a NAS file system in the NAS console. For more information, see 创建文件系统. After a NAS file system is created, go back to the ECS instance creation wizard and click the icon to query the most recent NAS file system list. For more information about how to mount NAS file systems, see Mount NAS file systems when you purchase an ECS instance.

6. (Optional)Configure the snapshot service.

You can use automatic snapshot policies to periodically back up disks to prevent risks such as accidental data deletion.

Select an existing snapshot policy or click **Create Automatic Snapshot Policy** to create an automatic snapshot policy on the Snapshots page. For more information, see Create an automatic snapshot policy. After an automatic snapshot policy is created, go back to the ECS instance creation wizard and click the icon to query the most recent automatic snapshot policy list.

## Step 2: Complete the settings in the Networking step

In the Networking step, you can configure parameters to allow instances to access the Internet and other Alibaba Cloud resources. This ensures the security of your instances. After you complete the settings in the Networking step, click **Next**.

1. Specify parameters in the Network Type and Public IP Address sections.

| Parameter | Description | References |
| --- | --- | --- |
| **Network Type** | Select **VPC**.<br>A virtual private cloud (VPC) is a logically isolated virtual network in Alibaba Cloud. You have full control over VPCs that belong to you. For example, you can specify a CIDR block and configure route tables and gateways for the VPC.<br>If you do not want to use a custom VPC or vSwitch in the specified region when you create an instance, you can skip this operation. Then, the system creates a default VPC and a default vSwitch.<br>**Note** You can skip this operation only if no available VPCs exist in the region where the instance is deployed.<br>Select an existing VPC and vSwitch. You can also click **go to the VPC console** to create a VPC and a vSwitch in the VPC console. After the VPC and the vSwitch are created, go back to the ECS instance creation wizard and click the icon to view the VPC and the vSwitch that you created. | ◦ What is a VPC?<br>◦ Create a VPC<br>◦ Create a vSwitch |

| Parameter | Description | References |
| --- | --- | --- |
| Public IP Address | If you select an image of Windows 2008 R2 or earlier in the **Basic Configurations** step, you can select **Assign Public IPv4 Address**, or you can associate an elastic IP address (EIP) with the instance after the instance is created. This way, you can connect to the instance over other protocols such as the Remote Desktop Protocol (RDP) built into Windows, PC over IP (PCoIP), and XenDesktop HDX 3D. Otherwise, you cannot connect to the instance from a Virtual Network Console (VNC) client after the GPU driver is installed. A persistent black screen or startup interface appears when you attempt to connect to the instance.<br>**Note** RDP does not support some applications such as DirectX and OpenGL applications. If you want to use these applications, you must manually install the VNC service and client.<br>To assign a public IP address, perform the following operations:<br>i. Select **Assign Public IPv4 Address**.<br>ii. Specify the Bandwidth Billing parameter.<br>■ **Pay-By-Bandwidth**: You are charged based on the specified bandwidth. This billing method is suitable for the scenarios that require stable network bandwidth.<br>■ **Pay-By-Traffic**: You are charged based on the traffic that you use. You can configure a peak bandwidth value to avoid excessive fees due to sudden traffic spikes. This billing method is suitable for scenarios that require highly variable bandwidth, such as the scenarios where traffic is low in most cases but spikes occasionally occur.<br>iii. Set Bandwidth or Peak Bandwidth based on your requirements. | What is an EIP? |

2. Select security groups.

   A security group is a virtual firewall that is used to control the inbound and outbound traffic of instances in the security group. For more information, see Overview.

If you do not want to configure security group-related parameters when you create an instance, you can skip the step. The system creates a default security group. The default security group allows inbound traffic over SSH port 22, Remote Desktop Protocol (RDP) port 3389, and Internet Control Message Protocol (ICMP). You can modify the security group configurations after the security group is created.

i. To create a security group, click **create a security group**.

For more information about how to configure a security group, see Create a security group.

ii. Click **Reselect Security Group**.

iii. In the **Select Security Group** dialog box, select one or more security groups and click **Select**.

3. Configure ENIs.

ENIs are classified into primary ENIs and secondary ENIs. Primary ENIs cannot be unbound from instances. They cannot be created or released independently of the instances to which they are bound. Secondary ENIs can be bound to or unbound from instances to allow traffic to be switched between instances. To create a secondary ENI when you create an instance, click the + icon and select a vSwitch to which to connect the secondary ENI.

> **Note** You can bind only one secondary ENI when you create an instance. You can also create secondary ENIs and bind them to an instance after the instance is created. For more information about the number of ENIs that can be bound to an instance of each instance type, see Instance family.

## Step 3: Complete the settings in the System Configurations step

In the System Configurations step, you can configure the parameters to customize what you want to display for the GPU-accelerated instance in the ECS console and in the OS, and how to use the instance. For example, you can configure the Logon Credentials, Host, and User Data parameters. After you complete the settings in the System Configurations step, click **Next**.

1. Configure logon credentials.

We recommend that you set the Logon Credentials parameter to **Key Pair** or **Password**. If you set the Logon Credentials parameter to **Set Later**, you must bind an SSH key pair or reset the password before you connect to the instance from a management terminal. Then, you must restart the instance so that the logon credentials can take effect. If you restart the instance when the GPU driver is being installed, the GPU driver fails to be installed.

2. Specify the instance name that you want to display in the ECS console and the hostname that can be obtained from within the operating system.

If you want to create multiple instances, you can set sequential instance names and hostnames to facilitate management. For more information about how to configure sequential instance names and hostnames, see Batch configure sequential names or hostnames for multiple instances.

3. Configure advanced settings.

i. Select an instance Resource Access Management (RAM) role.

An ECS instance can assume an instance RAM role to obtain the permissions of the role. Then, the instance can securely make API requests to specific Alibaba Cloud services and manage specific Alibaba Cloud resources based on the Security Token Service (STS) temporary credentials of the role.

Select an existing instance RAM role or click **Create Instance RAM Role** to create an instance RAM role in the RAM console. After an instance RAM role is created, go back to the ECS instance creation wizard and click the  icon to query the most recent instance RAM role list.

For more information, see Attach an instance RAM role.

ii. Select an instance metadata access mode.

ECS instance metadata includes instance information in Alibaba Cloud. You can view the metadata of running instances and configure or manage the instances based on their metadata. You can view instance metadata in normal or security hardening mode. For more information, see View instance metadata.

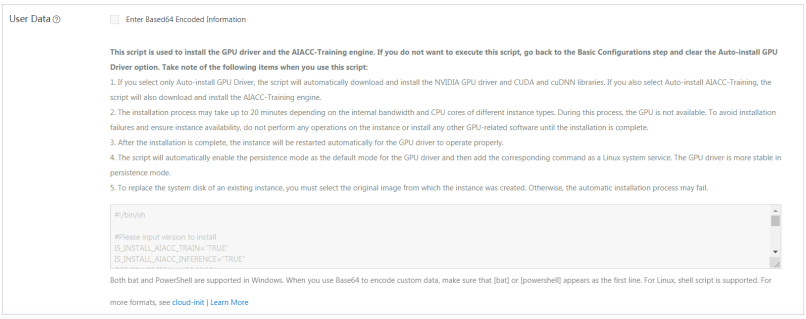
| Instance metadata access mode | Description |
| --- | --- |
| **Normal Mode (Compatible with Security Hardening Mode)** | After the instance is created, you can view its metadata in normal mode or in security hardening mode. |
| **Security Hardening Mode** | After the instance is created, you can view its metadata only in security hardening mode. |

Note

iii. Configure user data.

User data can be run as scripts on instance startup to automate instance configurations, or can be passed to instances as regular data. For more information, see Manage the user data of Linux instances and Manage the user data of Windows instances.

If you select Auto-install GPU Driver, Auto-install RDMA Software Stack, AIACC-Training, and AIACC-Inference in the **Basic Configurations** step, an automatic installation script appears in the lower part of the Advanced section. You can select Auto-install RDMA Software Stack only when you use an instance of the sccgn7ex instance family. The first time the instance is started after the instance is created, cloud-init runs the script.




**Note** You can also customize an automatic installation script and import the script so that a GPU driver, an RDMA software stack, AIACC-Training, and AIACC-Inference can be automatically installed. For more information, see Configure an automatic installation script.

## Step 4: (Optional) Complete the settings in the Grouping (Optional) step

In the Grouping (Optional) step, you can configure parameters such as Tags and Resource Group to batch manage instances. After you complete the settings in the Grouping (Optional) step, click **Next**.

1. Add tags.

   Each tag consists of a key and a value. You can add tags to resources that have identical characteristics, such as resources that belong to the same organization and resources that serve the same purpose. You can use tags to search for and manage resources in an efficient manner. For more information, see Overview.

   Select an existing tag, or enter a key and a value to create a tag.

2. Select a resource group.

   Resource groups allow you to manage resources across regions or across services based on your business requirements and manage the permissions of resource groups. For more information, see Resource groups.

   Select an existing resource group, or click **click here** to create a resource group on the Resource Group page. After a resource group is created, go back to the ECS instance creation wizard and click the icon to query the most recent resource group list. For more information, see Create a resource group.

3. Select a deployment set.

   Deployment sets support the high availability strategy. After you apply the high availability

strategy to a deployment set, all the instances in the deployment set are distributed across different physical servers to ensure business availability and implement underlying disaster recovery.

Select an existing deployment set or click **manage the deployment set** to create a deployment set. After a deployment set is created, go back to the ECS instance creation wizard and click the icon to query the most recent deployment set list. For more information, see Create a deployment set.

4. Select a dedicated host.

   A dedicated host is a cloud host whose physical resources are exclusively reserved for a single tenant. Dedicated hosts meet strict security compliance requirements and support bring your own license (BYOL) when you migrate services to Alibaba Cloud.

   Select an existing dedicated host or click **create a DDH** to create a dedicated host. After the dedicated host is created, go back to the ECS instance creation wizard and click the icon to query the most recent dedicated host list. For more information, see Create a dedicated host.

5. Select a private pool.

   After an elasticity assurance or a capacity reservation is created, the system generates a private pool to reserve resources for a specific number of instances that have specific attributes. During the validity period of the elasticity assurance or capacity reservation, you always have access to the resources reserved in the private pool when you want to create instances. For more information, see Overview.

   > **Note** Only pay-as-you-go instances can be created from the resources reserved by elasticity assurances or capacity reservations.

| Private pool | Description |
| --- | --- |
| **Open** | The capacity in open private pools takes priority over the capacity in the public pool. If no capacity is available in private pools, the system attempts to use the capacity in the public pool. |
| **None** | The capacity in private pools is not used. |
| **Targeted** | The capacity in a specified or open private pool is used to create instances. If no capacity is available in the specified private pool, the instances cannot be created. |

## Step 5: Complete the settings in the Preview step

Before the instance is created, make sure that all selected settings, such as the usage duration, meet your business requirements.

1. Check the selected settings.

   To modify the settings in a step, click the icon to go to the relevant step. You can generate a template based on the selected settings. Then, you can use the template to create instances that have similar settings. The following table describes the buttons that you can use to generate the template.

| Button | Description | References |
| --- | --- | --- |
| **Save as Launch Template** | Saves the settings as a launch template. You can use the launch template to create instances without the need to configure the settings again. | Create an instance by using a launch template |
| **View Open API** | Generates the API workflow and the SDK sample code for your reference. | ○ RunInstances<br>○ Create multiple ECS instances at a time (Java)<br>○ Create multiple ECS instances at a time (Python) |
| **Save as ROS Template** | Saves the settings as a Resource Orchestration Service (ROS) template. You can create stacks from the template to deliver resources in an efficient manner. | Create a stack |

2. Configure the usage duration of the instance.
   - Pay-as-you-go instance: Specify an automatic release time for the instance. You can also manually release the instance or specify an automatic release time for the instance after the instance is created. For more information, see Release an instance.
   - Subscription instance: Specify the usage duration and specify whether to enable auto-renewal. You can also manually renew the instance or enable auto-renewal for the instance after the instance is created. For more information, see Renewal overview.
3. Read *ECS Terms of Service* and *Product Terms of Service*. If you agree to them, select **ECS Terms of Service and Product Terms of Service**.
4. In the lower part of the page, view the total fees of the instance, confirm the order, and then follow on-screen instructions to complete the payment.

   If you select Auto-install GPU Driver, the system installs the GPU driver after the instance is created. The installation duration takes about 10 to 20 minutes and varies based on the internal bandwidth and the number of vCPUs provided by different instance types. You can connect to the instance to view the installation process. You can view the installation logs in the */root/auto_install/auto_install.log* directory after the GPU driver is installed. The following table describes the display effects during the installation process.

| Installation process | Display effect |
| --- | --- |
| In progress | The installation progress bar appears. |
| Installed | The installation result **ALL INSTALL OK** appears. |
| Failed | The installation result **INSTALL FAIL** appears. |

**Notice** When the GPU driver is being installed, the GPU is unavailable. You cannot perform operations or install other GPU-related software on the instance. This prevents an installation failure and ensures instance availability.

## Configure an automatic installation script

You can use the automatic installation script in the following scenarios:

- You do not want to select Auto-install GPU Driver, Auto-install RDMA Software Stack, AIACC-Training, or AIACC-Inference in the **Basic Configurations** step, and you want to enter an automatic installation script in the **System Configurations** step.
- You want to call the RunInstances operation to create a GPU-accelerated instance. In this case, you must upload an automatic installation script by specifying the UserData parameter.

To configure an automatic installation script and use the script to install a GPU driver when you create the instance, perform the following operations:

1. Customize an automatic installation script.

   The automatic installation script contains the following content:

```
#!/bin/sh
#Please input version to install
IS_INSTALL_RDMA=""
IS_INSTALL_AIACC_TRAIN=""
IS_INSTALL_AIACC_INFERENCE=""
DRIVER_VERSION=""
CUDA_VERSION=""
CUDNN_VERSION=""
IS_INSTALL_RAPIDS="FALSE"

INSTALL_DIR="/root/auto_install"

#using .run to install driver and cuda
auto_install_script="auto_install.sh"

script_download_url=$(curl http://100.100.100.200/latest/meta-data/source-address | hea
d -1)"/opsx/ecs/linux/binary/script/${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_AIAC
C_INFERENCE $IS_INSTALL_RDMA $IS_INSTALL_RAPIDS
```

   > **Note** The automatic installation script uses the *.run* installation package to install modules, such as GPU drivers.

   You must add the following parameters to the script based on your business requirements.

   - Specify the versions of the GPU driver, CUDA library, and cuDNN library based on the selected instance family and image version. For more information, see Image versions and instance families supported for GPU drivers. Sample code:

```
DRIVER_VERSION="470.82.01"
CUDA_VERSION="11.4.1"
CUDNN_VERSION="8.2.4"
```

   - Specify whether to install an RDMA software stack.

> **Note** You can install RDMA software stacks only when you use instances that belong to the sccgn7ex instance family.

If you want to install an RDMA software stack, set the IS_INSTALL_RDMA parameter to *TRUE*. If you do not want to install an RDMA software stack, set the IS_INSTALL_RDMA parameter to *FALSE*. Sample code:

```
IS_INSTALL_RDMA="TRUE"
```

- Specify whether to install AIACC-Training and AIACC-Inference.
  - If you want to install AIACC-Training, set the IS_INSTALL_AIACC_TRAIN parameter to *TRUE*. If you do not want to install AIACC-Training, set the IS_INSTALL_AIACC_TRAIN parameter to *FALSE*.
  - If you want to install AIACC-Inference, set the IS_INSTALL_AIACC_INFERENCE parameter to *TRUE*. If you do not want to install AIACC-Inference, set the IS_INSTALL_AIACC_INFERENCE parameter to *FALSE*.

  Sample code:

```
IS_INSTALL_AIACC_TRAIN="TRUE"
IS_INSTALL_AIACC_INFERENCE="FALSE"
```

2. After the script is customized, enter the script in the field below **User Data** in the **Advanced** section in the System Configurations step.

   After the instance is started, the system installs the GPU driver, CUDA library, and cuDNN library. The system also determines whether to install the RDMA software stack, AIACC-Training, and AIACC-Inference based on the script that you entered. After the installation, the system restarts the instance for the GPU driver to run.

   > **Note** The GPU driver in persistence mode is more stable. When you use the automatic installation script, the system enables the persistence mode for the GPU driver in Linux on instance startup. This ensures that the persistence mode is enabled for the GPU driver after the instance is restarted.

# 3.2. Create a GPU-accelerated instance that is not configured with a driver

This topic describes how to create and purchase a GPU-accelerated instance that is not configured with a driver. However, to ensure the performance of your instance, you must manually install a driver for your instance.

## Context

You can create this type of GPU-accelerated instance in the following scenarios:

- You have obtained a GRID license and a GRID driver from a reputable source such as the NVIDIA official website.

- You want to purchase a GPU-accelerated instance that is not configured with a driver from Alibaba Cloud and install a driver that you have purchased from the NVIDIA official website.

The driver installation methods may vary based on individual use cases or the types of drivers that you want to install. For more information, see Installation guideline for NVIDIA drivers.

## Preparations

1. Create an Alibaba Cloud account and complete account information.
   - Create an Alibaba Cloud account. For more information, see Sign up with Alibaba Cloud.
   - Complete real-name verification because purchase ECS instances in the Chinese mainland. For more information, see Real-name Registration FAQs.
2. Go to the Custom Launch tab of the instance buy page in the ECS console.

## Procedure

- Step 1: Complete the settings in the Basic Configurations step
- Step 2: Complete the settings in the Networking step
- Step 3: (Optional) Complete the settings in the System Configurations step
- Step 4: (Optional) Complete the settings in the Grouping step
- Step 5: Complete the settings in the Preview step

## Step 1: Complete the settings in the Basic Configurations step

In the Basic Configurations step, you can configure basic parameters for purchasing an instance, such as the billing method, region, and zone, and basic resources required by the instance, such as the instance type, image, and storage size. After you configure the parameters in the Basic Configurations step, click **Next**.

1. Select a billing method.

   The billing method for an instance determines how the billing and charging rules are applied to the instance. The billing method also determines how the status of the resources that are deployed on the instance is changed at different points of the resource lifecycle.

| Billing method | Description | References |
| --- | --- | --- |
| **Subscription** | A billing method in which you pay for resources before you use them. | Subscription |
| **Pay-As-You-Go** | A billing method in which you use resources first and pay for them afterward. The billing cycles of pay-as-you-go instances are accurate to the second. You can purchase and release instances based on your business requirements.<br>**Note** We recommend that you use this billing method with savings plans to reduce costs. | ○ Pay-as-you-go<br>○ Savings plans |

| Billing method | Description | References |
| --- | --- | --- |
| **Preemptible Instance** | A billing method in which you use resources first and pay for them afterward. The prices of preemptible instances are lower than that of pay-as-you-go instances. However, preemptible instances may be automatically released due to fluctuations in the market price or insufficient resources of instance types. | Preemptible instances |

2. Select a region and a zone.

   Select a region that is close to your geographical location to reduce latency. After an instance is created, the region and the zone of the instance cannot be changed. For more information, see Regions and zones.

3. In the Instance Type section, specify parameters and select an instance type.

   i. Set Architecture to **Heterogeneous Computing**, set Category to **Compute Optimized Type with GPU** or **Visualization Compute Optimized Type with GPU**, and then select an instance type.

   > **Note**
   >
   > - The available instance types vary based on the selected region. To view the instance types that are available in each region, go to the ECS Instance Types Available for Each Region page.
   > - If you have specific configuration requirements for the instance, for example, if you want to bind multiple elastic network interfaces (ENIs), or use enhanced SSDs (ESSDs) or local disks for the instance, make sure that the instance type that you select meets the requirements. For information about the features, scenarios, and specifications of each instance type, see Instance family.
   > - If you want to create an instance that is used for a specific scenario, click the **Scenario-based Selection** tab to view the instance types that are recommended for different scenarios. For example, you can set Business Scenario to AI Machine Learning to view the GPU-accelerated instance types that are available for AI machine learning scenarios.

   ii. Confirm the selected instance type next to **Selected Instance Type**.

iii. If you set **Billing Method** to **Preemptible Instance**, configure the Use Duration and Maximum Price for Instance Type parameters.

Use Duration specifies the protection period of a preemptible instance. After the protection period ends, the instance may be released due to insufficient resources or a lower bid than the market price. The following table describes the valid values of the Use Duration parameter.

| Value | Description |
| --- | --- |
| **One Hour** | After the preemptible instance is created, it enters a 1-hour protection period during which it cannot be automatically released. |
| **None** | The preemptible instance is created without a protection period. Preemptible instances without a protection period are lower-cost than preemptible instances with a protection period. |

The following table describes the valid values of the Maximum Price for Instance Type parameter.

| Value | Description |
| --- | --- |
| **Use Automatic Bid** | The real-time market price of the instance type is automatically used. The price can be up to but cannot exceed the pay-as-you-go price of the instance type. Automatic bidding can prevent the preemptible instance from being released due to lower bids than the market price, but cannot prevent the instance from being released due to insufficient resources. |
| **Set Maximum Price** | You must specify a maximum price. If the real-time market price exceeds your specified maximum price or if available resources are insufficient, the preemptible instance is released. |

iv. Specify the number of instances to create.

You can create a maximum of 100 instances at a time by using the wizard. In addition, the number of instances within your account cannot exceed your instance quota. The instance quota is displayed on the buy page. For more information, see View and increase instance quotas.

4. Select an image.

Images contain the information that is required to run instances. Alibaba Cloud provides a variety of image types for you to access image resources. The following table describes the image types.

| Image type | Description | References |
| --- | --- | --- |
| Public Image | Public images are base images provided by Alibaba Cloud. Public images are licensed and include Windows Server OS images and mainstream Linux OS images. | Overview |

| Image type | Description | References |
| --- | --- | --- |
| Custom Image | You can create or import custom images. Custom images contain the initial system environment, application environment, and software configurations. This eliminates repeated manual configurations. | Overview |
| Shared Image | Shared images are custom images that can be shared across Alibaba Cloud accounts. You can use this type of image to create instances across accounts. | Share or unshare a custom image |
| Marketplace Image | Alibaba Cloud Marketplace provides a wide range of images. These images have been thoroughly reviewed by Alibaba Cloud. You can use these images to create instances for website building and application development in a simplified manner. | Alibaba Cloud Marketplace images |

5. Complete the storage and related settings.

   Instances provide storage capabilities based on the system disks, data disks, and Apsara File Storage NAS file systems that are attached to the instances. ECS provides cloud and local disks to meet the storage requirements of different scenarios.

   Cloud disks include ESSDs, standard SSDs, and ultra disks and can be used as system disks or data disks. For more information, see Disks.

   > **Note** The billing method of a cloud disk that is created along with an instance is the same as that of the instance.

   Local disks can be used only as data disks. If an instance family (such as instance family with local SSDs and big data instance family) is equipped with local disks, the information of the local disks is displayed. For more information, see Local disks.

   > **Note** Local disks cannot be attached to instances on your own.

i. Configure a system disk.

System disks are used to install operating systems. The default capacity of a system disk is 40 GiB. However, the actual minimum capacity is related to the image. The following table describes the capacity ranges of system disks for different images.

| Image | System disk capacity range (GiB) |
| --- | --- |
| Linux (excluding CoreOS and Red Hat) | [max{20, Image size}, 500] |
| FreeBSD | [max {30, Image size}, 500] |
| CoreOS | [max {30, Image size}, 500] |
| Red Hat | [max {40, Image size}, 500] |
| Windows | [max {40, Image size}, 500] |

ii. (Optional)Add data disks.

You can create empty data disks or create data disks from snapshots. A snapshot is a point-in-time backup of a disk. You can import data in a quick manner by creating a disk from a snapshot. When you add a data disk, you can encrypt the disk to meet the requirements of scenarios such as data security and regulatory compliance. For more information about data encryption, see 加密概述.

> **Note** A limited number of data disks can be attached to a single instance. For more information, see the "Elastic Block Storage (EBS) limits" section in Limits.

iii. (Optional)Add NAS file systems.

If you have a large amount of data to share among multiple instances, we recommend that you use a NAS file system to reduce costs in data transmission and synchronization.

Select an existing NAS file system or click **Create a file system** to create a NAS file system in the NAS console. For more information, see 创建文件系统. After a NAS file system is created, go back to the ECS instance creation wizard and click the icon to query the most recent NAS file system list. For more information about how to mount NAS file systems, see Mount NAS file systems when you purchase an ECS instance.

6. (Optional)Configure the snapshot service.

You can use automatic snapshot policies to periodically back up disks to prevent risks such as accidental data deletion.

Select an existing snapshot policy or click **Create Automatic Snapshot Policy** to create an automatic snapshot policy on the Snapshots page. For more information, see Create an automatic snapshot policy. After an automatic snapshot policy is created, go back to the ECS instance creation wizard and click the icon to query the most recent automatic snapshot policy list.

## Step 2: Complete the settings in the Networking step

In the Networking step, you can configure parameters to allow instances to access the Internet and other Alibaba Cloud resources. This ensures the security of your instances. After you complete the settings in the Networking step, click **Next**.

1. Specify parameters in the Network Type and Public IP Address sections.

| Parameter | Description | References |
| --- | --- | --- |
| **Network Type** | Select **VPC**.<br>A virtual private cloud (VPC) is a logically isolated virtual network in Alibaba Cloud. You have full control over VPCs that belong to you. For example, you can specify a CIDR block and configure route tables and gateways for the VPC.<br>If you do not want to use a custom VPC or vSwitch in the specified region when you create an instance, you can skip this operation. Then, the system creates a default VPC and a default vSwitch.<br>**Note** You can skip this operation only if no available VPCs exist in the region where the instance is deployed.<br>Select an existing VPC and vSwitch. You can also click **go to the VPC console** to create a VPC and a vSwitch in the VPC console. After the VPC and the vSwitch are created, go back to the ECS instance creation wizard and click the icon to view the VPC and the vSwitch that you created. | ◦ What is a VPC?<br>◦ Create a VPC<br>◦ Create a vSwitch |

| Parameter | Description | References |
| --- | --- | --- |
| **Public IP Address** | If you select an image of Windows 2008 R2 or earlier in the **Basic Configurations** step, you can select **Assign Public IPv4 Address**, or you can associate an elastic IP address (EIP) with the instance after the instance is created. This way, you can connect to the instance over other protocols such as the Remote Desktop Protocol (RDP) built into Windows, PC over IP (PCoIP), and XenDesktop HDX 3D. Otherwise, you cannot connect to the instance from a Virtual Network Console (VNC) client after the GPU driver is installed. A persistent black screen or startup interface appears when you attempt to connect to the instance.<br><br>**Note** RDP does not support some applications such as DirectX and OpenGL applications. If you want to use these applications, you must manually install the VNC service and client.<br><br>To assign a public IP address, perform the following operations:<br>i. Select **Assign Public IPv4 Address**.<br>ii. Specify the Bandwidth Billing parameter.<br>■ **Pay-By-Bandwidth**: You are charged based on the specified bandwidth. This billing method is suitable for the scenarios that require stable network bandwidth.<br>■ **Pay-By-Traffic**: You are charged based on the traffic that you use. You can configure a peak bandwidth value to avoid excessive fees due to sudden traffic spikes. This billing method is suitable for scenarios that require highly variable bandwidth, such as the scenarios where traffic is low in most cases but spikes occasionally occur.<br>iii. Set Bandwidth or Peak Bandwidth based on your requirements. | What is an EIP? |

2. Select security groups.

   A security group is a virtual firewall that is used to control the inbound and outbound traffic of instances in the security group. For more information, see Overview.

   If you do not want to configure security group-related parameters when you create an instance, you can skip the step. The system creates a default security group. The default security group allows inbound traffic over SSH port 22, Remote Desktop Protocol (RDP) port 3389, and Internet Control Message Protocol (ICMP). You can modify the security group configurations after the security group is created.

   i. To create a security group, click **create a security group**.

      For more information about how to configure a security group, see Create a security group.

   ii. Click **Reselect Security Group**.

   iii. In the **Select Security Group** dialog box, select one or more security groups and click **Select**.

3. Configure ENIs.

   ENIs are classified into primary ENIs and secondary ENIs. Primary ENIs cannot be unbound from instances. They cannot be created or released independently of the instances to which they are bound. Secondary ENIs can be bound to or unbound from instances to allow traffic to be switched between instances. To create a secondary ENI when you create an instance, click the + icon and select a vSwitch to which to connect the secondary ENI.

   > **Note** You can bind only one secondary ENI when you create an instance. You can also create secondary ENIs and bind them to an instance after the instance is created. For more information about the number of ENIs that can be bound to an instance of each instance type, see Instance family.

## Step 3: (Optional) Complete the settings in the System Configurations (Optional) step

In the System Configurations step, you can configure parameters to customize what instance information to display in the ECS console and in the OS or how to use the instance. For example, you can configure the Logon Credentials, Host, and User Data parameters. After you complete the settings in the System Configurations (Optional) step, click **Next**.

1. Configure logon credentials.

   Logon credentials are used to log on to the instance. For more information about how to connect to an instance, see Connection methodsGuidelines on instance connection.

| Logon credential | Description |
| --- | --- |
| **Key Pair** | Select an existing key pair or click **Create Key Pair** to create a key pair. After a key pair is created, go back to the ECS instance creation wizard and click the ↻ icon to query the most recent key pair list.<br>For more information, see Create an SSH key pair.<br>**Note** Key pairs can be used to log on only to Linux instances. |

| Logon credential | Description |
| --- | --- |
| **Password** | Enter and confirm a password. When you log on to an instance by using a username and a password, the default username for Linux is `root` and that for Windows is `administrator`. |
| **Set Later** | After the instance is created, bind the key pair or reset the instance password. For more information, see Bind an SSH key pair to an instance and Reset the logon password of an instance. |

2. Specify the instance name that you want to display in the ECS console and the hostname that can be obtained from within the operating system.

   If you want to create multiple instances, you can set sequential instance names and hostnames to facilitate management. For more information about how to configure sequential instance names and hostnames, see Batch configure sequential names or hostnames for multiple instances.

3. Configure advanced options.

   i. Select an instance Resource Access Management (RAM) role.

   An ECS instance can assume an instance RAM role to obtain the permissions of the role. Then, the instance can securely make API requests to specific Alibaba Cloud services and manage specific Alibaba Cloud resources based on the Security Token Service (STS) temporary credentials of the role.

   Select an existing instance RAM role or click **Create Instance RAM Role** to create an instance RAM role in the RAM console. After an instance RAM role is created, go back to the ECS instance creation wizard and click the icon to query the most recent instance RAM role list.

   For more information, see Attach an instance RAM role.

   ii. Select an instance metadata access mode.

   ECS instance metadata includes instance information in Alibaba Cloud. You can view the metadata of running instances and configure or manage the instances based on their metadata. You can view instance metadata in normal or security hardening mode. For more information, see View instance metadata.

| Instance metadata access mode | Description |
| --- | --- |
| **Normal Mode (Compatible with Security Hardening Mode)** | After the instance is created, you can view its metadata in normal mode or in security hardening mode. |
| **Security Hardening Mode** | After the instance is created, you can view its metadata only in security hardening mode. |

Note

iii. Configure user data.

User data can be run as scripts on instance startup to automate instance configurations, or can be used as common data and passed into instances. For more information, see Manage the user data of Linux instances and Manage the user data of Windows instances.

In the User Data field, enter the user data that you prepared. If the user data is already encoded in Base64, select **Enter Based64 Encoded Information**.

## Step 4: (Optional) Complete the settings in the Grouping (Optional) step

In the Grouping (Optional) step, you can configure parameters such as Tags and Resource Group to batch manage instances. After you complete the settings in the Grouping (Optional) step, click **Next**.

1. Add tags.

   Each tag consists of a key and a value. You can add tags to resources that have identical characteristics, such as resources that belong to the same organization and resources that serve the same purpose. You can use tags to search for and manage resources in an efficient manner. For more information, see Overview.

   Select an existing tag, or enter a key and a value to create a tag.

2. Select a resource group.

   Resource groups allow you to manage resources across regions or across services based on your business requirements and manage the permissions of resource groups. For more information, see Resource groups.

   Select an existing resource group, or click **click here** to create a resource group on the Resource Group page. After a resource group is created, go back to the ECS instance creation wizard and click the icon to query the most recent resource group list. For more information, see Create a resource group.

3. Select a deployment set.

   Deployment sets support the high availability strategy. After you apply the high availability strategy to a deployment set, all the instances in the deployment set are distributed across different physical servers to ensure business availability and implement underlying disaster recovery.

   Select an existing deployment set or click **manage the deployment set** to create a deployment set. After a deployment set is created, go back to the ECS instance creation wizard and click the icon to query the most recent deployment set list. For more information, see Create a deployment set.

4. Select a dedicated host.

   A dedicated host is a cloud host whose physical resources are exclusively reserved for a single tenant. Dedicated hosts meet strict security compliance requirements and support bring your own license (BYOL) when you migrate services to Alibaba Cloud.

   Select an existing dedicated host or click **create a DDH** to create a dedicated host. After the dedicated host is created, go back to the ECS instance creation wizard and click the icon to query the most recent dedicated host list. For more information, see Create a dedicated host.

5. Select a private pool.

After an elasticity assurance or a capacity reservation is created, the system generates a private pool to reserve resources for a specific number of instances that have specific attributes. During the validity period of the elasticity assurance or capacity reservation, you always have access to the resources reserved in the private pool when you want to create instances. For more information, see Overview.

> **Note** Only pay-as-you-go instances can be created from the resources reserved by elasticity assurances or capacity reservations.

| Private pool | Description |
| --- | --- |
| **Open** | The capacity in open private pools takes priority over the capacity in the public pool. If no capacity is available in private pools, the system attempts to use the capacity in the public pool. |
| **None** | The capacity in private pools is not used. |
| **Targeted** | The capacity in a specified or open private pool is used to create instances. If no capacity is available in the specified private pool, the instances cannot be created. |

## Step 5: Confirm the order

Before the instance is created, make sure that the selected configurations such as the use duration meet your requirements.

1. Check the selected configurations.

   To modify the configurations in a step, click the icon to go to the step. You can save the selected configurations as a template. Then, you can use the template to create instances that have similar configurations. The following table describes the buttons that can be used to save the configurations as a template.

| Operation | Description | References |
| --- | --- | --- |
| **Save as Launch Template** | Saves the configurations as a launch template. Then, you can create instances from this launch template without making these configurations again. | Create an instance by using a launch template |
| **View Open API** | Generates the API best-practice workflow and SDK examples for your reference. | ○ RunInstances<br>○ Batch create ECS instances<br>○ Create multiple ECS instances at a time |
| **Save as ROS Template** | Saves the configurations as a Resource Orchestration Service (ROS) template. Then, you can create stacks from this template in the ROS console to deliver resources in a quick manner. | Create a stack |

2. Configure the use duration of the instance.
   - For a pay-as-you-go instance, set an automatic release time for the instance. You can also manually release the instance or set an automatic release time for the instance after it is created. For more information, see Release an instance.
   - For a subscription instance, set Duration and optionally select Enable Auto-renewal. You can also manually renew the instance or enable auto-renewal for the instance after it is created. For more information, see Renewal overview.
3. Read *ECS Terms of Service* and *Product Terms of Service*. If you agree to them, select **ECS Terms of Service and Product Terms of Service**.
4. View the total fees of the instance in the lower part of the page. Confirm the configurations of the instance and complete the payment.

## Related information

### References

- RunInstances
- Install a GPU driver on a Linux GPU-accelerated compute-optimized instance
- Install a Windows GPU driver on a GPU-accelerated compute-optimized instance

# 4.Connect to an instance with GPU capabilities

## 4.1. Overview

GPU-accelerated instances are a type of Elastic Compute Service (ECS) instance and can be connected to in the same manner as common ECS instances. You can use a variety of methods to connect to an ECS instance, including VNC, Workbench and third-party client tools. Select a method to connect to your instance based on the instance operating system, the operating system of your device, and the operations that you want to perform.

### Connection methods

| Operating system of your instance | Operating system of your device | Connection method |
| --- | --- | --- |
| Linux | Windows | • Workbench<br>For information about how to connect to an instance by using a password or a key as the credential, see Connect to a Linux instance by using a password or key.<br>• VNC<br>For more information, see Connect to a Linux instance by using a password.<br>• Client tools such as PuTTY<br>◦ For information about how to connect to an instance by using an SSH key pair as the credential, see Use an SSH key pair to connect to a Linux instance from a Windows device.<br>◦ For information about how to connect to an instance by using a username and password as the credential, see Use a username and password to connect to a Linux instance from a Windows device. |
| | | |

| Operating system of your instance | Operating system of your device | Connection method |
| --- | --- | --- |
| | UNIX-like operating systems such as Linux and macOS | • Workbench<br>For information about how to connect to an instance by using a password or a key as the credential, see Connect to a Linux instance by using a password or key.<br>• VNC<br>For more information, see Connect to a Linux instance by using a password.<br>• SSH commands<br>◦ For information about how to connect to an instance by using an SSH key pair as the credential, see Use an SSH key pair to connect to a Linux instance from a device that supports SSH commands (configure information by using commands).<br>◦ For information about how to connect to an instance by using a username and password as the credential, see Use a username and password to connect to a Linux instance from a Linux or Mac OS X device. |
| | Operating systems of mobile devices, such as iOS and Android | Apps such as SSH Control Lite and JuiceSSH<br>For more information, see Connect to a Linux instance from a mobile device. |
| | Windows | • Workbench<br>For information about how to connect to an instance by using a password or a key as the credential, see Connect to a Windows instance by using a password or key.<br>• VNC<br>For more information, see Connect to a Windows instance by using a password.<br>• Client tools such as Remote Desktop Connection (formerly called MSTSC)<br>For more information, see Connect from a local client that runs a Windows operating system. |

| Operating system of your instance | Operating system of your device | Connection method |
| --- | --- | --- |
| Windows | Linux | • Workbench<br>For information about how to connect to an instance by using a password or a key as the credential, see Connect to a Windows instance by using a password or key.<br>• VNC<br>For more information, see Connect to a Windows instance by using a password.<br>• Client tools such as rdesktop<br>For more information, see Connect from a local client that runs a Linux operating system. |
| | macOS | • Workbench<br>For information about how to connect to an instance by using a password or a key as the credential, see Connect to a Windows instance by using a password or key.<br>• VNC<br>For more information, see Connect to a Windows instance by using a password.<br>• Client tools such as Microsoft Remote Desktop Connection for Mac<br>For more information, see Get started with the macOS client. |
| | Operating systems of mobile devices, such as iOS and Android | Apps such as Microsoft Remote Desktop<br>For more information, see Connect to a Windows instance from a mobile device. |

**Note**

- Except for Workbench and VNC, all connection tools require that instances that you want to connect have public IP addresses or elastic IP addresses (EIPs).
- After a Windows instance is created, it takes 2 to 3 minutes to initialize the operating system. Do not restart the instance while it is being initialized. After a non-I/O optimized Windows instance is created, it takes 10 minutes to initialize the operating system. Do not connect to the instance while it is being initialized.

## Comparison of connection tools

The following table compares the advantages of VNC, Workbench, and other third-party client tools.

| Item | Workbench | VNC | Third-party client tool |
| --- | --- | --- | --- |
| Assignment of a public IP address or an EIP to the instance | Optional.<br>**Note** Workbench cannot be used to troubleshoot network configuration exceptions, such as firewalls being enabled by mistake. | Optional. VNC can be used to troubleshoot network configuration exceptions, such as firewalls being enabled by mistake. | Required. |
| Enabling services such as SSH on the instance | Required. | Optional. VNC can be used to troubleshoot SSH service exceptions, such as SSHD being disabled. | Required. |
| Logons by using the ECS console | Supported. | Supported. | Not supported. The local client must be installed. |
| Independence of the instance operating system | Workbench can be used to connect to both Linux and Windows instances. | VNC can be used to connect to both Linux and Windows instances. | Depends on the client tool. The third-party client tools can be used to connect to Linux or Windows instances. |
| Simultaneous logons by multiple operating system users to a single instance | Supported. | Not supported. | Depends on the client tool. |
| Ease of interaction | Workbench supports copying and pasting text. | VNC does support copying and pasting text. To copy or paste text, use the feature for copying long commands. | Depends on the client tool. |
| Visibility into Linux system file resources | Supported. | Not supported. | Depends on the client tool. |
| Permissions to control and modify hardware | Not supported. | Supported. VNC can be used to manage resources such as BIOS and troubleshoot exceptions such as system startup failures. | Not supported. |

| Item | Workbench | VNC | Third-party client tool |
| --- | --- | --- | --- |
| Terminal configurability | Supported, but depends on the capabilities that Workbench provides. | Not supported. | Supported, but depends on the capabilities that the client tool provides. |

# 4.2. Use Workbench to connect to an instance

## 4.2.1. Connect to a Linux instance by using a password or key

Workbench allows multiple users to connect to a single Elastic Compute Service (ECS) instance at the same time and provides a GUI for users to manage files in Linux instances. Workbench is more efficient and convenient than Virtual Network Console (VNC).

### Prerequisites

- A logon password is set for or a key pair is bound to the Linux instance to which you want to connect.
- The instance is in the **Running** state.
- Security group rules are added to allow the IP addresses related to the Workbench service to access the instance. For more information about the security group rules, see the Add security group rules to allow Workbench access to a Linux instance section.

### Context

By default, a Workbench remote session persists for 6 hours. If you do not perform operations for 6 hours, the remote connection is closed. You must reconnect to the instance.

Workbench can be used to connect to ECS instances over one of the following protocols:

- SSH: By default, Linux instances are connected by using SSH. SSH can also be used to connect to Windows instances on which a GNU-like system such as Cygwin is installed. For information about how to connect to a Linux instance over SSH, see the Connect to a Linux instance over SSH section.
- Remote Desktop Protocol (RDP): By default, Windows instances are connected by using RDP. RDP can also be used to connect to Linux instances on which remote desktop services are enabled. For information about how to connect to a Linux instance over RDP, see the Connect to a Linux instance over RDP section.

  > **Note** If you want to connect to an instance over RDP, make sure that the public bandwidth is at least 5 Mbit/s. If the public bandwidth is less than 5 Mbit/s, the remote desktop freezes.

You can use the GUI provided by Workbench to manage files in your Linux instances in a visual manner. For more information, see Use Workbench to manage files in a Linux instance.

### Connect to a Linux instance over SSH

1.
2.
3.
4. On the **Instances** page, find the instance to which you want to connect, and click **Connect** in the **Actions** column.
5. In the **Connection and Command** dialog box, click **Connect** in the **Workbench Connection** section.
6. In the **Instance Login** dialog box, specify parameters.

   The following table describes the required parameters in the dialog box.

| Parameter | Description |
| --- | --- |
| **Instance** | The information of the current instance is automatically populated. You can also manually enter the IP address or name of another instance. |
| **Connection** | ○ To connect to instances that are located in VPCs, you can use their public or private IP addresses.<br>○ To connect to instances that are located in the classic network, you can use their public or internal IP addresses. |
| **Username**, **Password**, and **Private Key** | Enter a username such as root and select an authentication method. The following authentication methods are supported:<br>○ **Password-based**: Enter the password of your specified username.<br>○ **Certificate-based**: Enter or upload a certificate. If the certificate is encrypted, enter its key passphrase. |

   In the lower part of the dialog box, click **More Options** to show the optional parameters described in the following table.

| Parameter | Description |
| --- | --- |
| **Resource Group** | By default, **All** is selected. You can manually select a resource group from the drop-down list. |
| **Region** | By default, **All** is selected. You can manually select a region from the drop-down list. |
| **Protocol** | By default, **Terminal Connection (SSH)** is selected. |
| **Port** | When Protocol is set to **Terminal Connection (SSH)**, this parameter is automatically set to 22. |
| **Language** | Select your preferred language. The selected language affects the outputs of the instance. We recommend that you select **Default** for Workbench to detect the language settings of the instance and to make configurations accordingly. |

| Parameter | Description |
| --- | --- |
| **Character Set** | Select your preferred character set. The selected character set affects the outputs of the instance. We recommend that you select **Default** for Workbench to detect the character set settings of the instance and to make configurations accordingly. |

7. Click **OK**.

If all of the requirements specified in the prerequisites are met but the instance cannot be connected, perform the following checks on the instance:

- Check whether the sshd service (such as sshd in Linux) is enabled. If not, enable the sshd service.
- Check whether the required terminal connection port (typically port 22) is enabled. If not, enable the port.
- If you log on to the Linux instance as the root user, make sure that `PermitRootLogin yes` is configured in the */etc/ssh/sshd_config* file. For more information, see the Enable root logon over SSH on a Linux instance section.

## Connect to a Linux instance over RDP

1.
2.
3.
4. On the **Instances** page, find the instance to which you want to connect, and click **Connect** in the **Actions** column.
5. In the **Connection and Command** dialog box, click **Connect** in the **Workbench Connection** section.
6. In the **Instance Login** dialog box, specify parameters.
   i. In the lower part of the dialog box, click **More Options**.
   ii. Set **Protocol** to **Remote Desktop (RDP)**.
   iii. In the message that appears, click **OK**.

iv. Specify the parameters described in the following table.

| Parameter | Description |
| --- | --- |
| **Resource Group** | By default, **All** is selected. You can manually select a resource group from the drop-down list. |
| **Region** | By default, **All** is selected. You can manually select a region from the drop-down list. |
| **Instance** | The information of the current instance is automatically populated. You can also manually enter the IP address or name of another instance. |
| **Connection** | ■ To connect to instances that are located in VPCs, you can use their public or private IP addresses.<br>■ To connect to instances that are located in the classic network, you can use their public or internal IP addresses. |
| **Port** | When Protocol is set to **Remote Desktop (RDP)**, this parameter is automatically set to 3389. |
| **Username** and **Password** | Enter a username, such as Administrator, and its password. |

7. Click **OK**.

If all of the requirements specified in the prerequisites are met but the instance cannot be connected, perform the following checks on the instance:

- Check whether a remote desktop service (such as xfreerdp installed on Linux) is enabled. If not, enable a remote desktop service.
- Check whether the required remote desktop port (typically port 3389) is enabled. If not, enable the port.
- If you log on to the Linux instance as the root user, make sure that `PermitRootLogin yes` is configured in the */etc/ssh/sshd_config* file. For more information, see the Enable root logon over SSH on a Linux instance section.

## Enable root logon over SSH on a Linux instance

In some Linux systems, sshd disables root logon by default. If this occurs, when you attempt to connect to an instance as the root user over SSH, you are prompted that your username or password is incorrect. To enable root logon over SSH, perform the following operations.

1. Connect to a Linux instance by using a password with VNC
2. Open the SSH configuration file.

```
vi /etc/ssh/sshd_config
```

3. Change `PermitRootLogin no` to `PermitRootLogin yes`.
4. Press the Esc key and enter **:wq** to save the change.
5. Restart sshd.

```
service sshd restart
```

# Add security group rules to allow Workbench access to a Linux instance

This section describes how to add rules to security groups of different network types in the ECS console to allow Workbench access to a Linux instance.

- If you want to connect to a Linux instance in a VPC, find a security group of the instance, go to the **Security Group Rules** page, and then add a rule on the **Inbound** tab. The following table describes the parameters to be configured for the rule.

| NIC Type | Rule Direction | Action | Protocol Type | Port Range | Priority | Authorization Type | Authorization Object |
|---|---|---|---|---|---|---|---|

| NIC Type | Rule Direction | Action | Protocol Type | Port Range | Priority | Authorization Type | Authorization Object |
|---|---|---|---|---|---|---|---|
| N/A | **Inbound** | **Allow** | ◦ If port 22 is enabled by default on the Linux instance, select **SSH (22)**.<br>◦ If you have manually enabled other ports on the Linux instance, select **Custom TCP**. | ◦ If port 22 is enabled by default on the Linux instance, **22/22** is automatically entered after you select the protocol type.<br>◦ If you have manually enabled other ports on the Linux instance, enter a corresponding port range. | 1 | **IPv4 CIDR Block** | ◦ If you want to connect to the instance by using its public IP address, specify 161.117.90.22/24. The public IP address can be the public IP address that is automatically assigned to the instance or an elastic IP address (EIP) that is associated with the instance.<br>◦ If you want to connect to the instance by using its private IP address, specify 100.104.0.0/16.<br>**Note** You can also specify 0.0.0.0/0 as the authorization object to allow inbound access from all IP addresses. However, this imposes security risks. Proceed with caution. |

- If you want to connect to a Linux instance in the classic network over the Internet, find a security group of the instance, go to the **Security Group Rules** page, and then add a rule on the **Internet Ingress** tab. The following table describes the parameters to be configured for the rule.

| NIC Type | Rule Direction | Action | Protocol Type | Port Range | Priority | Authorization Type | Authorization Object |
|---|---|---|---|---|---|---|---|
| **Public** | **Inbound** | **Allow** | ◦ If port 22 is enabled by default on the Linux instance, select **SSH (22)**.<br>◦ If you have manually enabled other ports on the Linux instance, select **Custom TCP**. | ◦ If port 22 is enabled by default on the Linux instance, **22/22** is automatically entered after you select the protocol type.<br>◦ If you have manually enabled other ports on the Linux instance, enter a corresponding port range. | 1 | **IPv4 CIDR Block** | If you want to connect to the instance by using its public IP address, specify 161.117.90.22/24. The public IP address can be the public IP address that is automatically assigned to the instance or an EIP that is associated with the instance.<br>**Note** You can also specify 0.0.0.0/0 as the authorization object to allow inbound access from all IP addresses. However, this imposes security risks. Proceed with caution. |

- If you want to connect to a Linux instance in the classic network over the internal network, security group of the instance, go to the **Security Group Rules** page, and then add a rule on the **Internal Network Ingress** tab. The following table describes the parameters to be configured for the rule.

| NIC Type | Rule Direction | Action | Protocol Type | Port Range | Priority | Authorization Type | Authorization Object |
|---|---|---|---|---|---|---|---|

| NIC Type | Rule Direction | Action | Protocol Type | Port Range | Priority | Authorization Type | Authorization Object |
|---|---|---|---|---|---|---|---|
| N/A | **Inbound** | **Allow** | ◦ If port 22 is enabled by default on the Linux instance, select **SSH (22)**.<br>◦ If you have manually enabled other ports on the Linux instance, select **Custom TCP**. | ◦ If port 22 is enabled by default on the Linux instance, **22/22** is automatically entered after you select the protocol type.<br>◦ If you have manually enabled other ports on the Linux instance, enter a corresponding port range. | 1 | **IPv4 CIDR Block** | If you want to connect to the instance by using its internal IP address, specify 11.195.184.0/24 and 11.246.55.0/24.<br>**Notice** High security risks may arise if you specify 0.0.0.0/0 as the authorization object. We recommend that you do not specify 0.0.0.0/0. |

# 4.2.2. Connect to a Windows instance by using a password or key

Workbench allows multiple users to connect to a single Elastic Compute Service (ECS) instance at the same time. Workbench is more efficient and convenient than Virtual Network Console (VNC).

## Prerequisites

- A logon password or a key is configured for the Windows instance to which you want to connect.

  **Note** The ECS console cannot be used to bind key pairs to Windows instances. If you want to use a key to log on to a Windows instance, you can enable the sshd service (such as Cygwin SSHD or WinSSHD in Windows) and configure a key on the instance. For more information about how to enable the sshd service in Windows, see Get started with OpenSSH.

- The instance is in the **Running** state.
- Security group rules are added to allow the IP addresses related to the Workbench service to access the instance. For more information, see Add security group rules to allow Workbench access to a Windows instance.

## Context

By default, a Workbench remote session persists for 6 hours. If you do not perform operations for 6 hours, the remote connection is closed. You must reconnect to the instance.

Workbench can be used to connect to ECS instances over one of the following protocols:

- Remote Desktop Protocol (RDP): By default, Windows instances are connected by using RDP. RDP can also be used to connect to Linux instances on which remote desktop services are enabled. For information about how to connect to a Windows instance over RDP, see the Connect to a Windows instance over RDP section.

  > **Note** If you want to connect to an instance over RDP, make sure that the public bandwidth is at least 5 Mbit/s. If the public bandwidth is less than 5 Mbit/s, the remote desktop freezes.

- SSH: By default, Linux instances are connected by using SSH. SSH can also be used to connect to Windows instances on which a GNU-like system such as Cygwin is installed. For information about how to connect to a Windows instance over RDP, see the Connect to a Windows instance over SSH section.

## Connect to a Windows instance over RDP

1.
2.
3.
4. On the **Instances** page, find the instance to which you want to connect, and click **Connect** in the **Actions** column.
5. In the **Connection and Command** dialog box, click **Connect** in the **Workbench Connection** section.
6. In the **Instance Login** dialog box, specify parameters.

   The following table describes the required parameters in the dialog box.

| Parameter | Description |
| --- | --- |
| **Instance** | The information of the current instance is automatically populated. You can also manually enter the IP address or name of another instance. |
| **Connection** | ○ To connect to instances in virtual private clouds (VPCs), you can use the public or private IP addresses of the instances.<br>○ To connect to instances in the classic network, you can use their public or internal IP addresses. |
| **Username** and **Password** | Enter a username, such as Administrator, and its password. |

   In the lower part of the dialog box, click **More Options** to show the optional parameters described in the following table.

| Parameter | Description |
| --- | --- |
| **Resource Group** | By default, **All** is selected. You can manually select a resource group from the drop-down list. |
| **Region** | By default, **All** is selected. You can manually select a region from the drop-down list. |
| **Protocol** | By default, **Remote Desktop (RDP)** is selected. |
| **Port** | When Protocol is set to **Remote Desktop (RDP)**, this parameter is automatically set to 3389. |

7. Click **OK**.

If all of the requirements specified in the prerequisites are met but the instance cannot be connected, perform the following checks on the instance:

- Check whether a remote desktop service (such as Remote Desktop Services in Windows) is enabled. If not, enable a remote desktop service.
- Check whether the required remote desktop port (typically port 3389) is enabled. If not, enable the port.
- If you log on to the Windows instance as a non-administrator user, the user must belong to the Remote Desktop Users group.

## Connect to a Windows instance over SSH

1.
2.
3.
4. On the **Instances** page, find the instance to which you want to connect, and click **Connect** in the **Actions** column.
5. In the **Connection and Command** dialog box, click **Connect** in the **Workbench Connection** section.
6. In the **Instance Login** dialog box, specify parameters.
   i. In the lower part of the dialog box, click **More Options**
   ii. Set **Protocol** to **Terminal Connection (SSH)**.
   iii. In the Confirm message, click **OK**.

iv. Specify the parameters described in the following table.

| Parameter | Description |
| --- | --- |
| **Resource Group** | By default, **All** is selected. You can manually select a resource group from the drop-down list. |
| **Region** | By default, **All** is selected. You can manually select a region from the drop-down list. |
| **Instance** | The information of the current instance is automatically populated. You can also manually enter the IP address or name of another instance. |
| **Connection** | ■ To connect to instances in VPCs, you can use the public or private IP addresses of the instances.<br>■ To connect to instances in the classic network, you can use their public or internal IP addresses. |
| **Port** | When Protocol is set to **Terminal Connection (SSH)**, this parameter is automatically set to 22. |
| **Username**, **Password**, and **Private Key** | Enter a username such as root and select an authentication method. The following authentication methods are supported:<br>■ **Password-based**: Enter the password of your specified username.<br>■ **Certificate-based**: Enter or upload a certificate. If the certificate is encrypted, enter its key passphrase. |
| **Language** | Select your preferred language. The selected language affects the outputs of the instance. We recommend that you select **Default** for Workbench to detect the language settings of the instance and make configurations accordingly. |
| **Character Set** | Select your preferred character set. The selected character set affects the outputs of the instance. We recommend that you select **Default** for Workbench to detect the character set settings of the instance and make configurations accordingly. |

7. Click **OK**.

If all of the requirements specified in the prerequisites are met but the instance cannot be connected, perform the following checks on the instance:

- Check whether the sshd service (such as Cygwin SSHD or WinSSHD in Windows) is enabled. If not, enable the sshd service.
- Check whether the required terminal connection port (typically port 22) is enabled. If not, enable the port.
- If you log on to the Windows instance as a non-administrator user, the user must belong to the Remote Desktop Users group.

# Add security group rules to allow Workbench access to a Windows instance

This section describes how to add rules to security groups of different network types in the ECS console to allow Workbench access to a Windows instance.

- If you want to connect to a Windows instance in a VPC, find a security group of the instance, go to the **Security Group Rules** page, and then add a rule on the **Inbound** tab. The following table describes the parameters to be configured for the rule.

| NIC Type | Rule Direction | Action | Protocol Type | Port Range | Priority | Authorization Type | Authorization Object |
|---|---|---|---|---|---|---|---|

| NIC Type | Rule Direction | Action | Protocol Type | Port Range | Priority | Authorization Type | Authorization Object |
|---|---|---|---|---|---|---|---|
| N/A | **Inbound** | **Allow** | ◦ If port 3389 is enabled by default on the Windows instance, select **RDP (3389)**.<br>◦ If you have manually enabled other ports on the Windows instance, select **Custom TCP**. | ◦ If port 3389 is enabled by default on the Windows instance, **3389/3389** is automatically entered after you select the protocol type.<br>◦ If you have manually enabled other ports on the Windows instance, enter a corresponding port range. | 1 | **IPv4 CIDR Block** | ◦ If you want to connect to the instance by using its public IP address, specify 161.117.90.22. The public IP address can be the public IP address that is automatically assigned to the instance or an elastic IP address (EIP) that is associated with the instance.<br>◦ If you want to connect to the instance by using its private IP address, specify 100.104.0.0/16.<br>**Note** You can also specify 0.0.0.0/0 as the authorization object to allow inbound access from all IP addresses. However, this imposes security risks. Proceed with caution. |

- If you want to connect to a Windows instance in the classic network over the Internet, find a security group of the instance, go to the **Security Group Rules** page, and then add a rule on the **Internet Ingress** tab. The following table describes the parameters to be configured for the rule.

| NIC Type | Rule Direction | Action | Protocol Type | Port Range | Priority | Authorization Type | Authorization Object |
|---|---|---|---|---|---|---|---|
| **Public** | **Inbound** | **Allow** | ◦ If port 3389 is enabled by default on the Windows instance, select **RDP (3389)**.<br>◦ If you have manually enabled other ports on the Windows instance, select **Custom TCP**. | ◦ If port 3389 is enabled by default on the Windows instance, **3389/3389** is automatically entered after you select the protocol type.<br>◦ If you have manually enabled other ports on the Windows instance, enter a corresponding port range. | 1 | **IPv4 CIDR Block** | If you want to connect to the instance by using its public IP address, specify 161.117.90.22. The public IP address can be the public IP address that is automatically assigned to the instance or an EIP that is associated with the instance.<br>**Note** You can also specify 0.0.0.0/0 as the authorization object to allow inbound access from all IP addresses. However, this imposes security risks. Proceed with caution. |

- If you want to connect to a Windows instance in the classic network over the internal network, find a security group of the instance, go to the **Security Group Rules** page, and then add a rule on the **Internal Network Ingress** tab. The following table describes the parameters to be configured for the rule.

| NIC Type | Rule Direction | Action | Protocol Type | Port Range | Priority | Authorization Type | Authorization Object |
|---|---|---|---|---|---|---|---|

| NIC Type | Rule Direction | Action | Protocol Type | Port Range | Priority | Authorization Type | Authorization Object |
|---|---|---|---|---|---|---|---|
| N/A | **Inbound** | **Allow** | ◦ If port 3389 is enabled by default on the Windows instance, select **RDP (3389)**.<br>◦ If you have manually enabled other ports on the Windows instance, select **Custom TCP**. | ◦ If port 3389 is enabled by default on the Windows instance, **3389/3389** is automatically entered after you select the protocol type.<br>◦ If you have manually enabled other ports on the Windows instance, enter a corresponding port range. | 1 | **IPv4 CIDR Block** | If you want to connect to the instance by using its internal IP address, specify 161.117.90.22.<br>**Notice** High security risks may arise if you specify 0.0.0.0/0 as the authorization object. We recommend that you do not specify 0.0.0.0/0. |

# 4.3. Use VNC to connect to an instance

## 4.3.1. Connect to a Linux instance by using a password

If you cannot use Workbench or connection software such as PuTTY, Xshell, and SecureCRT to connect to an Elastic Compute Service (ECS) 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

A logon password is set for the instance.

**Note** If you have not set a password or forget the password, you can reset the password for the instance. For more information, see Reset the logon password of an instance.

### Context

The following passwords are involved when you use VNC to connect to an instance:

- VNC password: the password of management terminals used to connect to the ECS console.
- Instance logon password: the password used to log on to the instance operating system.

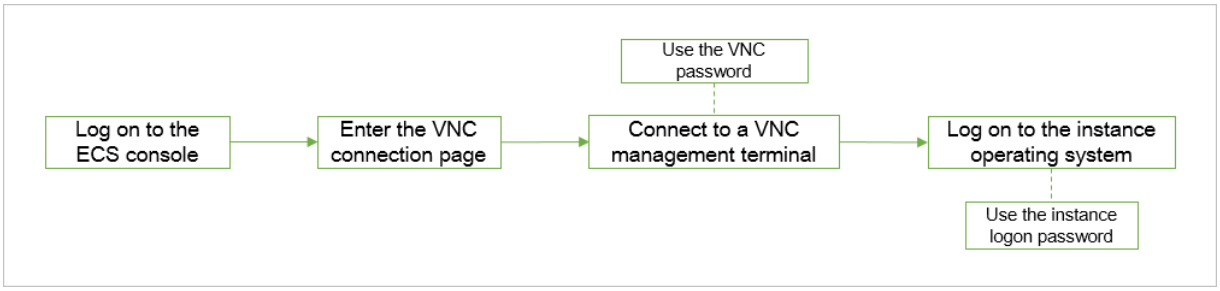
By default, a VNC connection session lasts for about 300 seconds. If you do not perform operations within these 300 seconds, the connection to the instance is automatically closed. You must connect to the instance again.

If you cannot use Workbench or connection software to connect to your instance, you can use the **VNC Connection** feature in the ECS console to connect to the instance. After the instance is connected, you can view the status of the instance and perform operations to resolve issues described in the following table.

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

## Procedure

The following figure shows how to use VNC to connect to an instance.




1.
2.
3.
4. On the **Instances** page, find the instance to be connected and click **Connect** in the **Actions** column.
5. In the **Connection and Command** dialog box, click **Connect** in the **VNC Connection** section.
6. Connect to a VNC management terminal.

   > **Note** In this step, use the VNC password.

   - The first time you connect to a VNC management terminal, perform the following operations:
     a. Change the VNC password. For more information, see the Change the VNC password section in this topic.
     b. In the **Enter VNC Password** dialog box, enter the new password.
     c. Click **OK**.

- If you are not connecting to a VNC management terminal for the first time, perform the following operations:
  a. In the **Enter VNC Password** dialog box, enter the password.
  b. Click **OK**.

7. Log on to the instance operating system.

   > **Note** In this step, use the instance logon password.

   i. Enter the username *root* and press the Enter key.
   ii. Enter the logon password of the instance and press the Enter key.

   > **Note** The characters of the password are hidden when you enter the password. After you enter the password, press the Enter key.

   You can switch between up to 10 different VNC management terminals when you connect to the Linux instance. The default terminal is **CTRL+ALT+F1**. For example, you can choose **Send Remote Call > CTRL+ALT+F2** to switch to **CTRL+ALT+F2**. A persistent black screen indicates that the instance is in sleep mode. Press a key to wake up the instance.

## Change the VNC password

The first time you connect to the VNC management terminal, you must change the VNC password. You can also change the VNC password when you forget the password or when you want to update the password.

> **Notice** After you change 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 can 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. In the **Connection and Command** dialog box, click **Connect** in the **VNC Connection** section.
3. In the **Enter VNC Password** dialog box, click **Reset VNC Password**.
4. In the **Reset 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 your computer to the instance, you can use the command copy feature.

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

# 4.3.2. Connect to a Windows instance by using a password

If you cannot use Workbench or connection software such as Remote Desktop Connection (RDC) and rdesktop to connect to an Elastic Compute Service (ECS) Windows instance, you can use the **VNC Connection** feature in the ECS console to connect to the instance and view the real-time status of the instance operating interface.

## Prerequisites

A logon password is set for the instance.

> **Note** If you have not set a password or forget the password, you can reset the password for the instance. For more information, see Reset the logon password of an instance.

## Context

The following passwords are involved when you use VNC to connect to an instance:

- VNC password: the password of management terminals used to connect to the ECS console.
- Instance logon password: the password used to log on to the instance operating system.

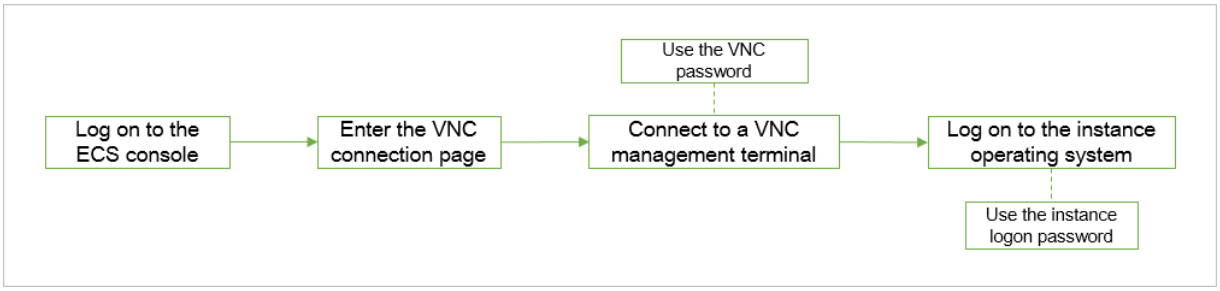
By default, a VNC connection session lasts for about 300 seconds. If you do not perform operations within these 300 seconds, the connection to the instance is automatically closed. You must connect to the instance again.

If you cannot use Workbench or connection software to connect to your instance, you can use the **VNC Connection** feature in the ECS console to connect to the instance. After the instance is connected, you can view the status of the instance and perform operations to resolve issues described in the following table.

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

## Procedure

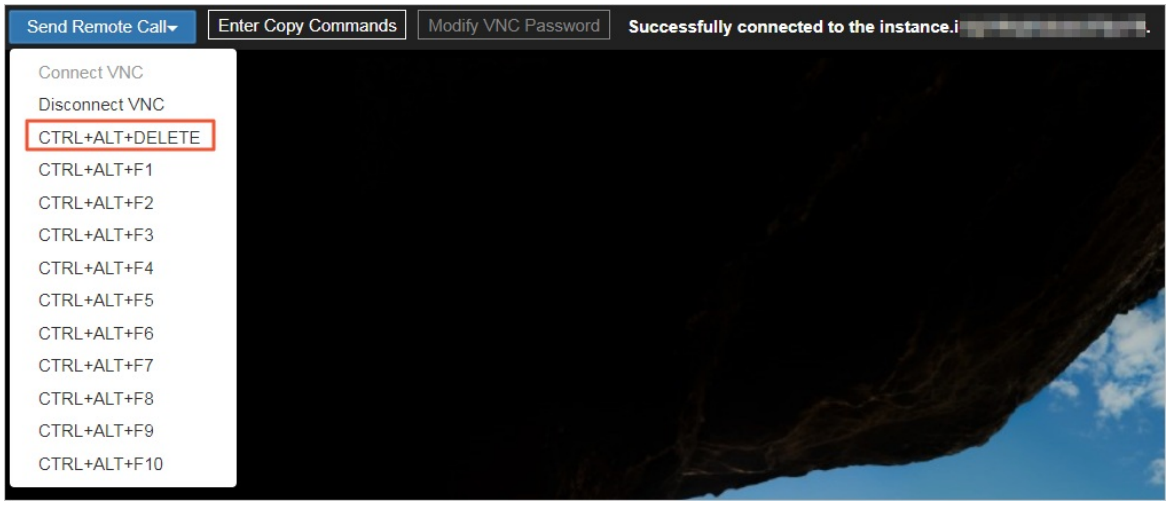
The following figure shows how to use VNC to connect to an instance.




1.
2.
3.
4. On the **Instances** page, find the instance to be connected and click **Connect** in the **Actions** column.
5. Connect to a VNC management terminal.

   > **Note** In this step, use the VNC password.

   - The first time you connect to a VNC management terminal, perform the following operations:
     a. Change the VNC password. For more information, see the Change the VNC password section in this topic.
     b. In the **Enter VNC Password** dialog box, enter the new password.
     c. Click **OK**.
   - If you are not connecting to a VNC management terminal for the first time, perform the following operations:
     a. In the **Enter VNC Password** dialog box, enter the password.
     b. Click **OK**.
6. In the upper-left corner of the **VNC** page, choose **Send Remote Call > CTRL+ALT+DELETE**.




7. Select an account, enter the instance password, and then press the Enter key.

   By default, the Administrator account is available.

## Change the VNC password

The first time you connect to the VNC management terminal, you must change the VNC password. You can also change the VNC password when you forget the password or when you want to update the password.

> **Notice** After you change 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 can 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. In the **Connection and Command** dialog box, click **Connect** in the **VNC Connection** section.
3. In the **Enter VNC Password** dialog box, click **Reset VNC Password**.
4. In the **Reset 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 your computer to the instance, you can use the command copy feature.

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

## FAQ

For more information about how to adjust the resolution of the Windows desktop, see How do I adjust the desktop resolution of a Windows instance?.

# 5.Manage an instance with GPU capabilities

## 5.1. Stop instances

From a resource management standpoint, GPU-accelerated instances are considered as Elastic Compute Service (ECS) instances and are managed in the same way you manage other ECS instances. This topic describes how to stop instances in the ECS console. This topic also describes the operations in ecomomical mode within virtual private clouds (VPCs).

### Prerequisites

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

> **Note** If you stop an instance, services that are running on an instance are interrupted. Proceed with caution when you perform this operation.

### Context

The billing of a subscription instance is not affected when you stop the instance.

The billing of a pay-as-you-go instance may be affected when you stop the instance. This depends on whether economical mode is enabled for the instance.

- Pay-as-you-go instances in the classic network do not support economical mode and continue to be billed after they are stopped. Billing stops only when the instances are released. For more information, see Release an instance.
- Pay-as-you-go instances in VPCs support economical mode.
  - If economical mode is disabled for a pay-as-you-go instance in a VPC, the instance continues to be billed after it is stopped.
  - If economical mode is enabled for a pay-as-you-go instance in a VPC, the vCPUs, memory, and public IP address of the instance are no longer billed after the instance is stopped. Other resources continue to be billed. For more information, see Economical mode.

### Stop a subscription instance

1.
2.
3.
4. Use one of the following methods to stop subscription instances:
   - To stop a single instance at a time, 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 Stopped By. Valid values:
   - **Stop**: stops the instance by shutting it down properly.

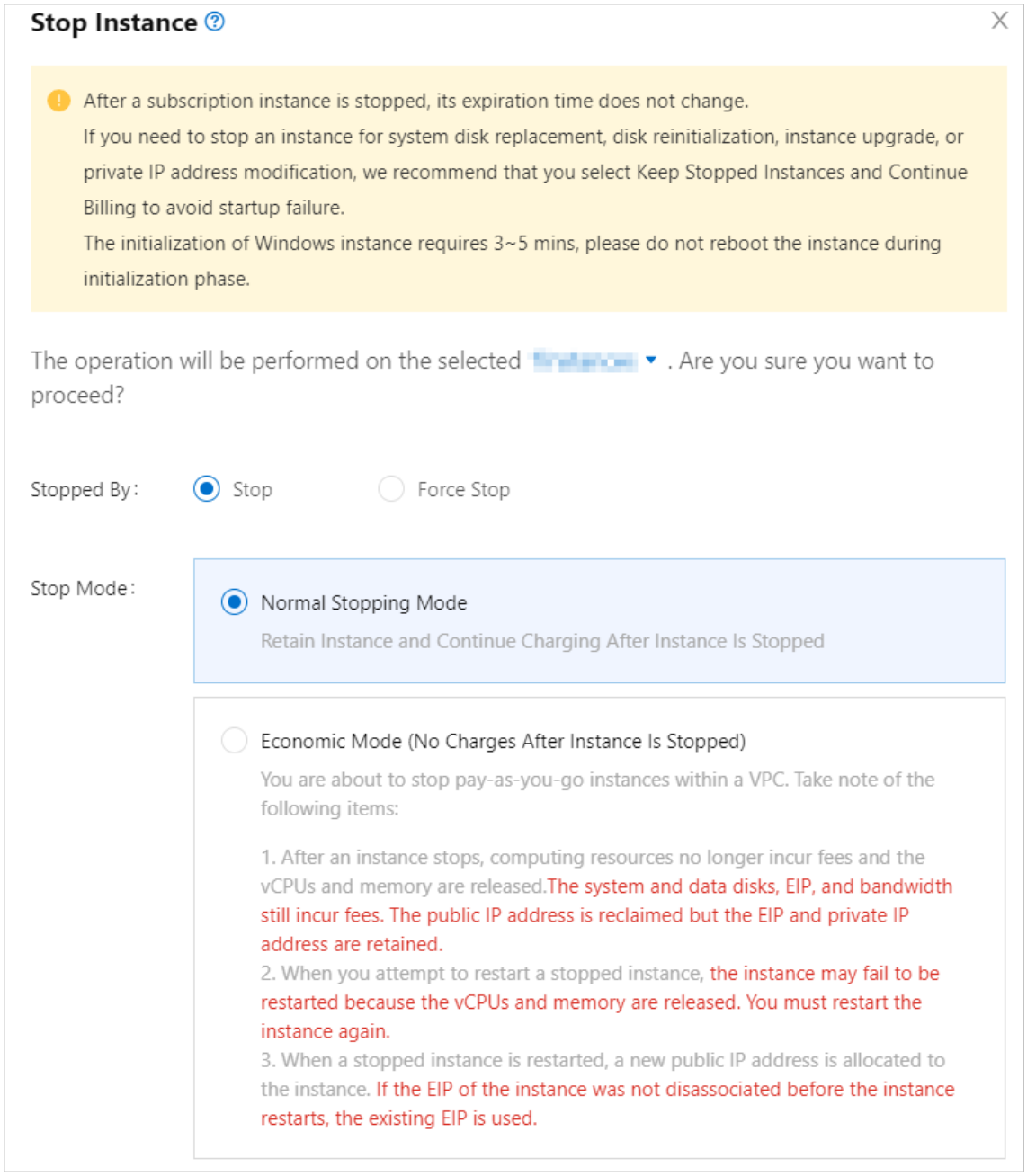
- **Force Stop**: forcibly stops the instance. Forcible stop is equivalent to a physical shutdown and may cause data loss if instance data has not been written to disks.

6. Click **OK**.

## Stop a pay-as-you-go instance

The procedures to stop preemptible instances are the same as those to stop pay-as-you-go instances. However, more factors affect the startup of stopped preemptible instances. For more information, see Stop a preemptible instance.

1.
2.
3.
4. Use one of the following methods to stop pay-as-you-go instances:
   - To stop a single instance at a time, 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 Stopped By and Stop Mode.
   - For a pay-as-you-go instance in the classic network:
     a. Configure Stopped By. Valid values:
        - **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**.
   - For a pay-as-you-go instance in a VPC:
     a. Configure Stopped By. Valid values:
        - **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. Configure Stop Mode. Valid values:
        - **Standard Mode**: The resources of the instance are retained and continue to be billed after the instance is stopped.
        - **Economical Mode (Formerly Known as No Fees for Stopped Instances Mode)**: After the instance is stopped, its computing resources (vCPUs and memory) are released and no longer billed. The cloud disks (including the system disk and data disks), elastic IP addresses (if any), and bandwidth continue to be billed. The public IP address is recycled and the private IP address is retained.
     c. Click **OK**.




## Result

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

## Related information

- StopInstance

# 5.2. Restart instances

GPU-accelerated instances are a type of Elastic Compute Service (ECS) instances and are managed in the same manner as common ECS instances. This article describes how to restart instances in the ECS console.

## Prerequisites

Only instances in the **Running** state can be restarted.

## Context

Restarting an instance will stop the instance. As a result, services provided by the instance are disrupted.

## Procedure

1. Log on to the ECS console.
2. In the left-side navigation pane, choose **Instances & Images > Instances**.
3. Select the target region.
4. Find the target instances.
   - To restart a instance, 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**.

## Related information

- RebootInstance

# 5.3. Release instances

GPU-accelerated instances are a type of Elastic Compute Service (ECS) instances and are managed in the same manner as common ECS instances. Only pay-as-you-go GPU-accelerated instances (including preemptible instances) and expired subscription GPU-accelerated instances can be released. This topic describes how to manually and automatically release pay-as-you-go GPU-accelerated instances.

## Prerequisites

After an instance is released, its data is deleted and cannot be recovered. We recommend that you create snapshots to back up data before you release the instance. For more information, see Create a snapshot for a disk.

> **Note** After an instance is released, snapshots and images that were manually created from the instance are not affected.

## Context

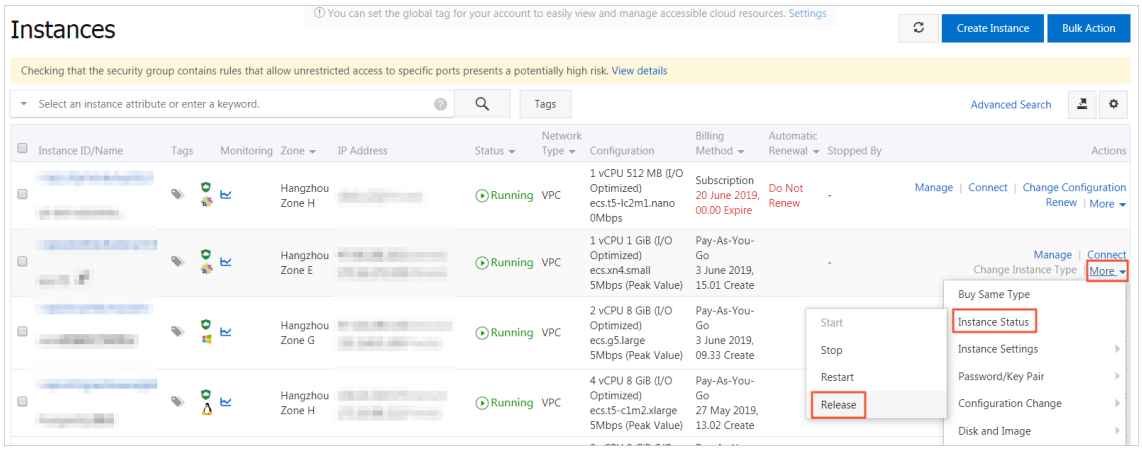
- Subscription instance that have not expired cannot be released. Before a subscription instance that has not expired can be released, you must convert it into a pay-as-you-go instance. For more information, see Change the billing method of an instance from subscription to pay-as-you-go.
- You can manually release expired subscription instances. If you do not renew an expired instance within a specific period of time, the instance is automatically released.
- If economical mode is disabled for a pay-as-you-go instance, you continue to be charged for the instance until it is released.
- You can enable instance release protection for a pay-as-you-go instance to prevent irreversible data loss caused by accidental release operations. For more information, see Enable or disable release protection for ECS instances.

- If the Release Disk with Instance feature is disabled for a disk attached to an instance, the disk is automatically converted into a pay-as-you-go data disk and retained when the instance is released. For more information, see Release a disk.

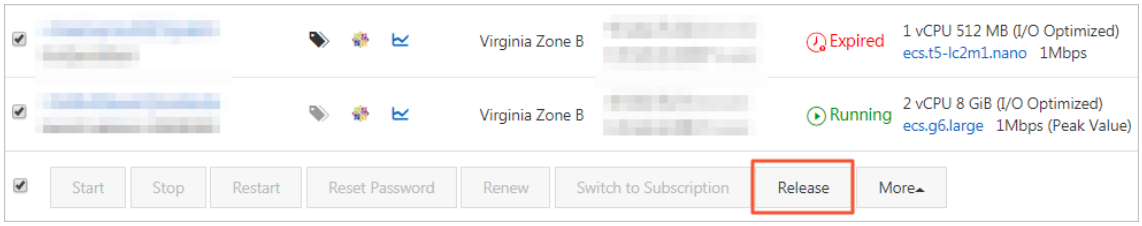
## Manually release instances

You can manually release pay-as-you-go instances in the ECS console.

1.
2.
3.
4. Release one or more pay-as-you-go instances at a time.
   - If you want to release a single pay-as-you-go instance at a time, find the instance that you want to release and choose **More > Instance Status > Release** in the **Actions** column.




   - If you want to release one or more pay-as-you-go instances at a time, click the Filter icon at the top of the **Billing Method** column and select Pay-As-You-Go from the drop-down list. In the displayed list of pay-as-you-go instances, select the instances that you want to release and then click **Release** below the instance list.




5. In the Release dialog box, select **Release Now**.
6. Click **Next**. Then, click **OK**.

## Enable automatic release

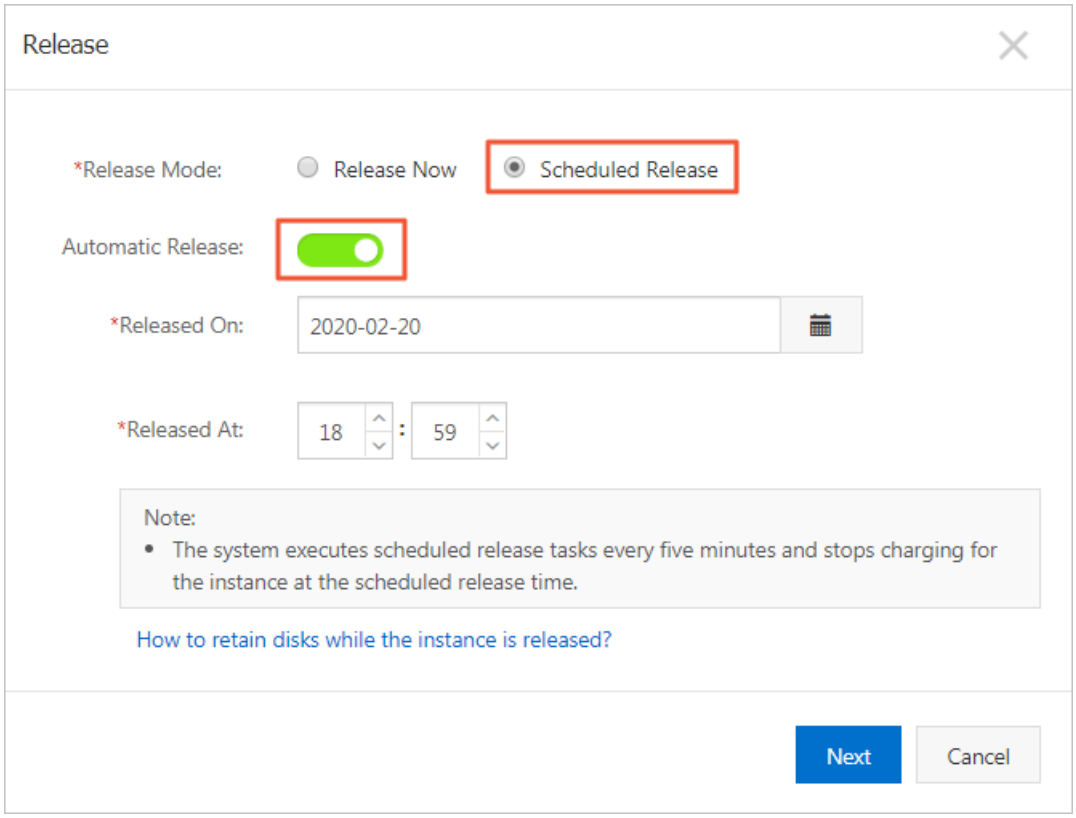
You can enable automatic release for pay-as-you-go instances and set a time to automatically release the instances. If you set the automatic release time more than once, the most recent setting prevails.

1.
2.
3.
4. Configure automatic release for one or more pay-as-you-go instances at a time.

- If you want to have a single pay-as-you-go instance automatically released at a time, find the instance that you want to release and choose **More > Instance Status > Release** in the **Actions** column.
- If you want to have one or more pay-as-you-go instances automatically released at a time, click the Filter icon at the top of the **Billing Method** column and select Pay-As-You-Go from the drop-down list. In the displayed list of pay-as-you-go instances, select the instances that you want to release and then click **Release** below the instance list.

5. In the Release dialog box, select **Scheduled Release**.
6. Turn on **Automatic Release** and specify a date and time to release the selected instances.

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




7. Click **Next**. Then, click **OK**.

## Disable automatic release

1.
2.
3.
4. Disable automatic release for one or more pay-as-you-go instances at a time.
   - If you want to disable automatic release for a single pay-as-you-go instance at a time, find the instance for which you want to disable the automatic release feature and choose **More > Instance Status > Release** in the **Actions** column.
   - If you want to disable automatic release for one or more pay-as-you-go instances at a time, click

the Filter icon at the top of the **Billing Method** column and select Pay-As-You-Go from the drop-down list. In the displayed list of pay-as-you-go instances, select the instances for which you want to disable automatic release and then click **Release** below the instance list.

5. In the Release dialog box, select **Scheduled Release**.
6. Turn off **Automatic Release**.
7. Click **Next**. Then, click **OK**.

## Related information

### References

- DeleteInstance
- ModifyInstanceAutoReleaseTime

# 5.4. GPU monitoring

This topic describes how to view the monitoring data of a GPU-accelerated compute optimized ECS instance in the CloudMonitor console and query the data by calling an API operation.

## Prerequisites

- A GPU-accelerated compute optimized ECS instance is created. The required GPU driver is installed on the instance. For more information, see Create a GPU-accelerated instance that is not configured with a driver.
- The CloudMonitor agent is installed on the ECS instance. For more information, see Install and uninstall the CloudMonitor agent for C++.
- Monitoring charts that include GPU metrics are added. For more information, see Add a monitoring chart.

## GPU metrics

GPU metrics can be viewed based on GPUs, instances, and application groups. The following table lists the GPU metrics.

| Metric | Unit | MetricName | Dimensions |
| --- | --- | --- | --- |
| (Agent)gpu_decoder_utilization | % | gpu_decoder_utilization | userId, instanceId, and gpuId |
| (Agent)gpu_encoder_utilization | % | gpu_encoder_utilization | userId, instanceId, and gpuId |
| (Agent)gpu_gpu_temperature | °C | gpu_gpu_temperature | userId, instanceId, and gpuId |
| (Agent)gpu_gpu_usedutilization | % | gpu_gpu_usedutilization | userId, instanceId, and gpuId |
| (Agent)gpu_memory_freespace | Byte | gpu_memory_freespace | userId, instanceId, and gpuId |

| Metric | Unit | MetricName | Dimensions |
| --- | --- | --- | --- |
| (Agent)gpu_memory_freeutilization | % | gpu_memory_freeutilization | userId, instanceId, and gpuId |
| (Agent)gpu_memory_userdspace | Byte | gpu_memory_userdspace | userId, instanceId, and gpuId |
| (Agent)gpu_memory_usedutilization | % | gpu_memory_usedutilization | userId, instanceId, and gpuId |
| (Agent)gpu_power_readings_power_draw | W | gpu_power_readings_power_draw | userId, instanceId, and gpuId |

## View GPU metric data in the CloudMonitor console

1.
2. In the left-side navigation pane, click **Host Monitoring**.
3. On the **Host Monitoring** page, click the host name or click the icon in the **Actions** column of the host.
4. Click the **GPU Monitoring** tab.

   On the **GPUMonitor** tab, view the monitoring charts for GPU metrics.

## Query GPU metric data by calling an API operation

You can call the DescribeMetricList operation to query the GPU metric data of the ECS instance. For more information, see DescribeMetricList.

> **Note** Set the `Namespace` parameter to `acs_ecs_dashboard` and specify the `MetricName` and `Dimensions` parameters. For more information, see GPU metrics.

# 6.Install NVIDIA drivers

## 6.1. Install a GPU driver on a Linux GPU-accelerated compute-optimized instance

If you do not configure a GPU driver that supports automatic installation or if you cannot find a public image of the required OS type and version when you create a GPU-accelerated compute-optimized instance, you must install a driver on the instance after the instance is created to ensure the performance of the instance. This topic describes how to install a GPU driver on a Linux GPU-accelerated compute-optimized instance after the instance is created.

### Context

To install a GPU driver on a GPU-accelerated instance, the OS of the GPU driver must be the same as that of the GPU-accelerated instance. This topic describes how to install a GPU driver only on a Linux GPU-accelerated compute-optimized instance. For more information about how to install a GPU driver on a Windows GPU-accelerated compute-optimized instance, see Install a Windows GPU driver on a GPU-accelerated compute-optimized instance.

### Procedure

1. Visit the DOWNLOAD DRIVERS page on the NVIDIA official website.
2. Search for the driver that you want to install.
   i.
   ii. Select a Linux version based on the image of the instance.

   If the Operating System drop-down list does not contain the OS that you want to use, click **Show All Operating Systems** at the bottom of the drop-down list. If you cannot find the Linux version that matches the image of the instance, select **Linux 64-bit**.
   iii. Select a CUDA Toolkit version.
   iv. Select a language.
   v. Click **SEARCH**, find the driver version that you want to download, and then click the driver name.
3. On the **driver details** page, click **DOWNLOAD**. On the **Download** page, right-click **DOWNLOAD** and select **Copy link address** to copy the download address.
4. Connect to the GPU-accelerated compute-optimized instance.

   Use one of the following methods to connect to the instance.

| Connection method | References |
| --- | --- |
| Workbench | Connect to a Linux instance by using a password or key |
| VNC | Connect to a Linux instance by using a password |

5. Paste the download address that you copied in Step 3 to the `wget` command and run the following command to download the installation package. Sample command:

```
wget https://cn.download.nvidia.com/tesla/460.73.01/NVIDIA-Linux-x86_64-460.73.01.run
```

6. Install the GPU driver.
   i. If your instance runs CentOS, run the following command to check whether kernel-devel and kernel-headers packages are installed on the instance. If your instance runs another OS type such as Ubuntu in which the kernel-devel and kernel-headers packages are pre-installed, skip this step.

```
rpm  -qa | grep $(uname -r)
```

   - If the command output contains the following version information about the kernel-devel and kernel-headers packages, the packages are installed.

```
kernel-3.10.0-1062.18.1.el7.x86_64
kernel-devel-3.10.0-1062.18.1.el7.x86_64
kernel-headers-3.10.0-1062.18.1.el7.x86_64
```

   - If the command output does not contain the preceding version information about *kernel-devel-** and *kernel-headers-**, you must download and install the kernel-devel and kernel-headers packages of the kernel version that you want to use.

> **Note** If the kernel-devel version is inconsistent with the kernel version, a compilation error occurs in the driver when you install RPM Package Manager for your driver. Therefore, you must check the version number of *kernel-** in the command output before you download kernel-devel. In the preceding command output, the version number of the kernel is 3.10.0-1062.18.1.el7.x86_64.

   ii. Install the GPU driver.

   In this example, a Linux 64-bit driver package in the .run format is downloaded. Example: NVIDIA-Linux-x86_64-xxxx.run. Run the following commands to authorize and install the GPU driver:

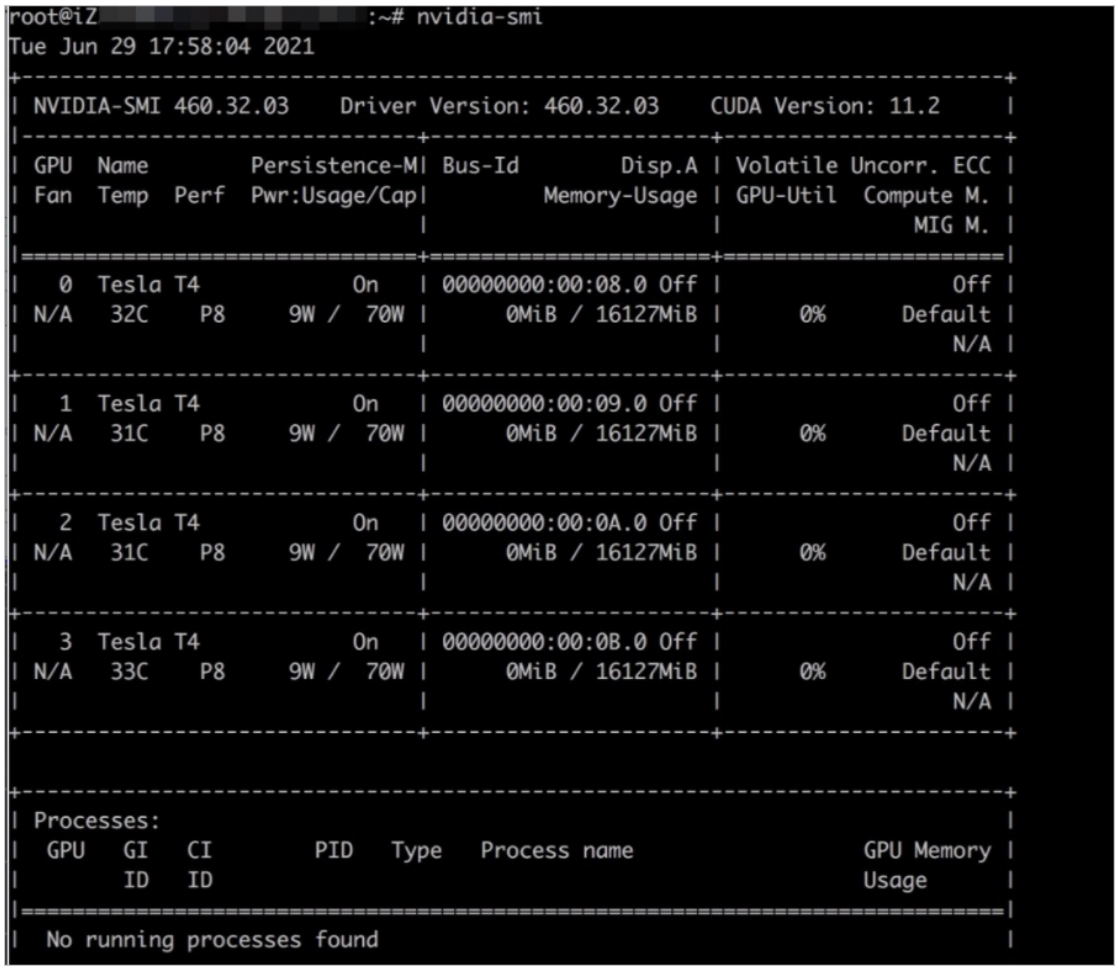
```
chmod +x NVIDIA-Linux-x86_64-xxxx.run
```

```
sh NVIDIA-Linux-x86_64-xxxx.run
```

iii. Run the following command to check whether the driver is installed:

```
nvidia-smi
```

If information similar to the following command output is displayed, the GPU driver is installed.




7. If the GPU-accelerated instance that you created belongs to the instance family ebmgn7, perform the following operations to install NVIDIA Fabric Manager of the version that matches your driver version. Otherwise, you cannot use the instance as expected.

i. Install NVIDIA Fabric Manager.

The commands that you can run to install NVIDIA Fabric Manager vary based on the instance OS. The following information describes how you run the commands.

In this example, a driver of the 460.91.03 version is used. You can change the version number next to `driver_version=` based on your business requirements.

- CentOS 7.x

```
driver_version=460.91.03
yum -y install yum-utils
yum-config-manager --add-repo https://developer.download.nvidia.com/compute/cuda/
repos/rhel7/x86_64/cuda-rhel7.repo
yum install -y nvidia-fabric-manager-${driver_version}-1
```

- CentOS 8.x

```
driver_version=460.91.03
driver_version_main=$(echo $driver_version | awk -F '.' '{print $1}')
distribution=rhel8
ARCH=$( /bin/arch )
dnf config-manager --add-repo http://developer.download.nvidia.com/compute/cuda/r
epos/$distribution/${ARCH}/cuda-$distribution.repo
dnf module enable -y nvidia-driver:${driver_version_main}
dnf install -y nvidia-fabric-manager-0:${driver_version}-1
```

- Ubuntu 16.04 or Ubuntu 18.04

```
driver_version=460.91.03
driver_version_main=$(echo $driver_version | awk -F '.' '{print $1}')
distribution=$(. /etc/os-release;echo $ID$VERSION_ID | sed -e 's/\.//g')
wget https://developer.download.nvidia.com/compute/cuda/repos/$distribution/x86_6
4/cuda-$distribution.pin
mv cuda-$distribution.pin /etc/apt/preferences.d/cuda-repository-pin-600
wget https://developer.download.nvidia.com/compute/cuda/repos/$distribution/x86_6
4/7fa2af80.pub
apt-key add 7fa2af80.pub
rm 7fa2af80.pub
echo "deb http://developer.download.nvidia.com/compute/cuda/repos/$distribution/x
86_64 /" | tee /etc/apt/sources.list.d/cuda.list
apt-get update
apt-get -y install nvidia-fabricmanager-${driver_version_main}=${driver_version}-
*
```

- Ubuntu 20.04

```
driver_version=460.91.03
driver_version_main=$(echo $driver_version | awk -F '.' '{print $1}')
apt-get update
apt-get -y install nvidia-fabricmanager-${driver_version_main}=${driver_version}-
*
```

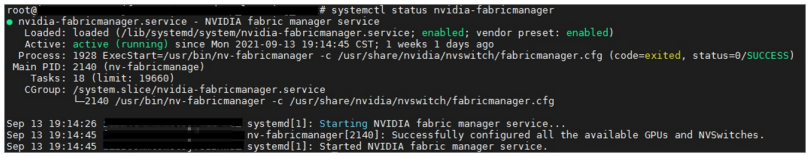
ii. Run the following commands to start NVIDIA Fabric Manager:

```
systemctl enable nvidia-fabricmanager
systemctl start nvidia-fabricmanager
```

iii. Run the following command to view the status of NVIDIA Fabric Manager:

```
systemctl status nvidia-fabricmanager
```

If information similar to the following command output is displayed, NVIDIA Fabric Manager is installed.




# 6.2. Install a Windows GPU driver on a GPU-accelerated compute-optimized instance

Windows GPU-accelerated instances do not support automatic configuration and installation of GPU drivers. To ensure the performance of your GPU-accelerated instance, you can purchase a GPU driver from the NVIDIA official website and install the driver on the instance. This topic describes how to install a Windows GPU driver on a GPU-accelerated compute-optimized instance.

## Context

To install a GPU driver on a GPU-accelerated instance, the OS of the GPU driver must be the same as that of the GPU-accelerated instance. This topic only describes how to install a Windows GPU driver on a GPU-accelerated compute-optimized instance. For more information about how to install a Linux GPU driver on a GPU-accelerated compute-optimized instance, see Install a GPU driver on a Linux GPU-accelerated compute-optimized instance.

## Procedure

1. Connect to the GPU-accelerated compute-optimized instance.

   Use one of the following methods to connect to the instance.

| Connection method | References |
| --- | --- |
| Workbench | Connect to a Windows instance by using a password or key |
| VNC | Connect to a Windows instance by using a password |

2. On your remote desktop, visit the DOWNLOAD DRIVERS page on the NVIDIA official website.
3. Search for the driver that you want to install.

i. From the Product Type, Product Series, and Product drop-down lists, select values based on the GPU with which your GPU-accelerated compute-optimized instance is configured. For more information about how to view instance specifications, see View instance information.

The following table describes the driver specifications that you can select for each instance family.

| Instance family | gn4 | gn5 | gn5i | gn6v | gn6i | gn6e | gn7 | gn7i |
|---|---|---|---|---|---|---|---|---|
| Product Type | Data Center / Tesla | Data Center / Tesla | Data Center / Tesla | Data Center / Tesla | Data Center / Tesla | Data Center / Tesla | Data Center / Tesla | Data Center / Tesla |
| Product Series | M-Class | P-Series | P-Series | V-Series | T-Series | V-Series | A-Series | A-Series |
| Product | M40 | Tesla P100 | Tesla P4 | Tesla V100 | Tesla T4 | Tesla V100 | NVIDIA A100 | NVIDIA A10 |

ii. From the Operating System drop-down list, select a Windows OS based on the image of the instance.

In this example, **Windows 10 64-bit** is used.

iii. From the CUDA Toolkit drop-down list, select a version for CUDA Toolkit.

iv. From the Language drop-down list, select a language.

v. Click **SEARCH**, find the driver version that you want to download, and then click the driver name.

4. On the **driver details** page, click **DOWNLOAD**. On the **Download** page, click **DOWNLOAD**.
5. After the driver is downloaded, open the folder where the driver installation package is stored, double-click the installation package, and then follow on-screen instructions to install the driver.

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

After you install the driver, you can open Device Manager on your computer. In the left-side navigation pane, click Display adapters to check whether the driver is installed. If the driver appears, the driver is installed.

# 6.3. Install a GRID driver on a Windows GPU-accelerated instance

If you need to use Windows GPU-accelerated compute-optimized instances or Windows vGPU-accelerated instances in graphics computing scenarios such as Open Graphics Library (OpenGL) and Direct3D scenarios, you must install a Windows GRID driver. If your GPU-accelerated instance is not configured with a GRID driver, you must install a GRID driver on the instance to ensure the performance of your instance. For example, you must install a GRID driver when you purchase an image that is not configured with a GRID driver from Alibaba Cloud Marketplace and use the image to create a GPU-accelerated instance, or when you cannot find the OS type and version that you want to use for your GRID driver. This topic describes how to install a GRID driver on a Windows GPU-accelerated instance and activate the GRID license.

## Prerequisites

- A GPU-accelerated instance is created. The instance can access the Internet. For more information, see Create a GPU-accelerated instance that is not configured with a driver.

  > **Note** We recommend that you click **Public Image** and select a Windows image in the Image section.

- A remote connection tool is installed on your computer.
- The GRID license and the installation package for the GRID driver are obtained. To obtain the license and the package, .

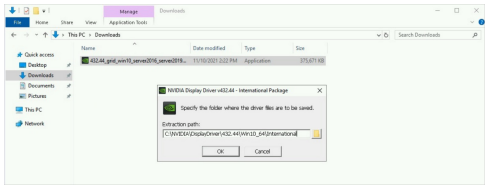
  > **Note** The version of the GRID driver must match the specifications of your GPU-accelerated instance and the region where the instance is deployed. Therefore, you must specify the specifications and the region in the ticket that you submit.

## Context

The operations that you perform to install a GRID driver on a Windows GPU-accelerated compute-optimized instance are similar to those that you perform to install a GRID driver on a Windows vGPU-accelerated instance. This topic describes how to install a GRID driver only on a vGPU-accelerated instance and activate the GRID license. In this example, a vGPU-accelerated instance that belongs to the instance family vgn6i and runs Windows Server 2019 is used.

## Procedure

1. 远程连接实例。
2. Install the GRID driver that you have obtained.
   i. Double-click the installation package. In the dialog box that appears, click **OK**.

   The package is decompressed.

ii. On the **NVIDIA software installation** page, click **AGREE AND CONTINUE**.


iii. Use the default values and click **NEXT**.

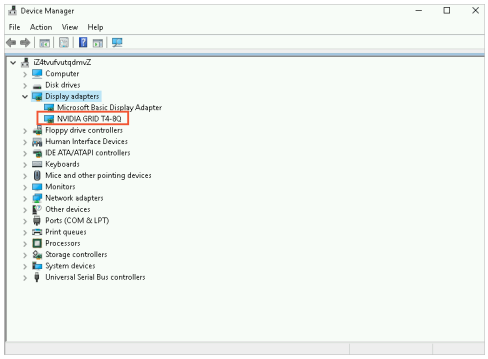

iv. After the driver is installed, click **RESTART NOW**.


3. After the instance restarts, connect to the instance and check whether the GRID driver is installed.

   i. On your Windows desktop, move the pointer over the icon, right-click the icon, and then click **Device Manager**.

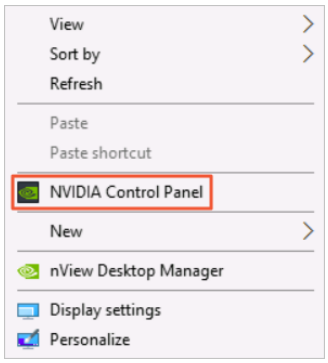
ii. In the **Device Manager** dialog box, click **Display adaptors** and check whether the driver appears. In this example, an **NVIDIA GRID T4-8Q** driver is used.

The following figure shows that the NVIDIA GRID T4-8Q driver is installed.

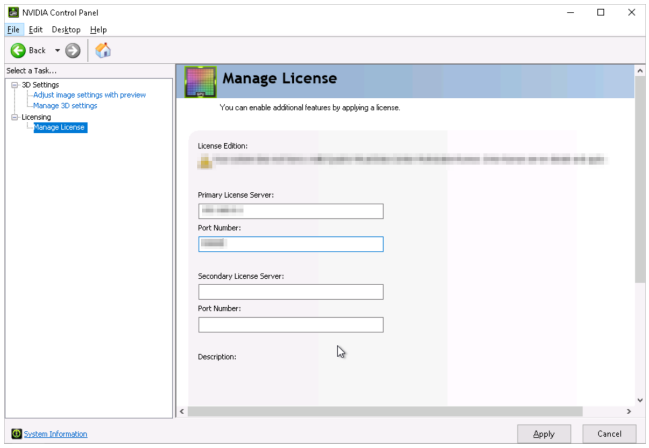



4. Add a license server and activate the license.

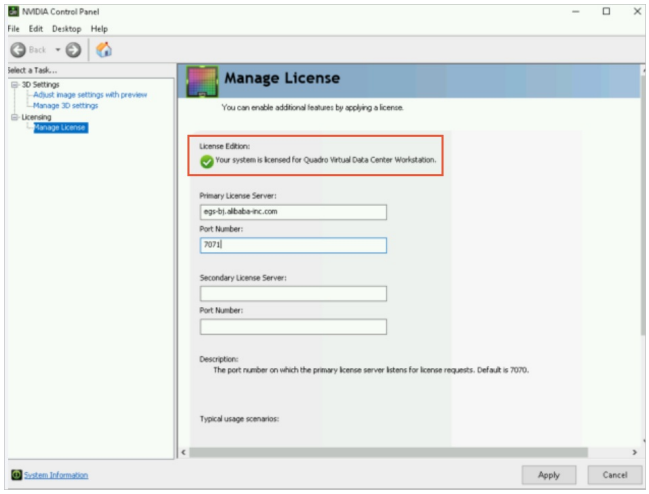
i. On your Windows desktop, right-click anywhere on the screen and select **NVIDIA Control Panel**.




ii. The **NVIDIA Control Panel** dialog box appears. In the left-side navigation pane, choose **Licensing > Manage License**.



iii. On the **Manage License** page, enter the IP address and port number of your license server, and click **Apply**.
Open the **NVIDIA Control Panel** dialog box again. If the message in the following figure appears, the license is activated.




# 6.4. Install a GRID driver on a Linux vGPU-accelerated instance

If you create vGPU-accelerated instances, you must install NVIDIA GRID drivers on the instances. If NVIDIA GRID licenses are not activated for the GPUs that are used by GPU-accelerated instances, GPU-related features may not take effect on the instances. You must obtain and activate the GRID licenses before you use these features. This topic describes how to install a GRID driver on a Linux vGPU-accelerated instance and activate the GRID license for the instance. This topic also describes how to test the acceleration effect on graphics. In this example, a vGPU-accelerated instance that belongs to the vgn6i or vgn5i instance family and runs Ubuntu 16.04 64-bit is used.

## Prerequisites

- A vGPU-accelerated instance that belongs to the vgn6i or vgn5i instance family and can access the Internet is created. When you create the instance, we recommend that you click **Public Image** to use a public image.

  > **Note** This topic describes how to install a GRID driver on a Linux vGPU-accelerated instance. For a Windows vGPU-accelerated instance, you can select an image that contains a pre-installed GRID driver and use the image to create the instance.

- A remote connection tool, such as VNC Viewer, is installed on your computer.
- The GRID license and the installation package for the GRID driver are obtained. To obtain the license and the package,.

  > **Note** The version of the GRID driver must match the specifications of your vGPU-accelerated instance and the region in which the instance is deployed. Therefore, you must specify the specifications and the region in the ticket that you submit.

## Context

This topic describes how to install a GRID driver on a Linux vGPU-accelerated instance. If you use a Linux GPU-accelerated compute-optimized instance, we recommend that you install a GPU driver on the instance. For more information, see Install a GPU driver on a Linux GPU-accelerated compute-optimized instance.

## Install a GRID driver

1. Disable Nouveau.

   Nouveau is an open source driver that is pre-installed in specific Linux OSs. If you use a Linux OS that has a Nouveau driver pre-installed, you must disable the driver before you install another driver.

   i. Connect to the vGPU-accelerated instance.

   Use one of the following methods to connect to the instance.

   | Connection method | References |
   | --- | --- |
   | Workbench | Connect to a Linux instance by using a password or key |
   | VNC | Connect to a Linux instance by using a password |

   ii. Check whether the *blacklist-nouveau.conf* file exists.

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

   If the file does not exist, run the following command to create the file:

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

   iii. Add the following information to the *blacklist-nouveau.conf* file to disable Nouveau:

   > **Note** If the following information is added to the file, skip this step.

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

   iv. Create kernel initramfs.

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

   v. Reboot the instance.

   ```
   reboot
   ```

2. Install the obtained GRID driver on the instance.

   - vgn5i

   ```
   chmod +x NVIDIA-Linux-x86_64-418.226.00-grid.run
   ./NVIDIA-Linux-x86_64-418.226.00-grid.run
   ```

   - vgn6i

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

- vgn6i-vws or vgn7i-vws

```
chmod +x NVIDIA-Linux-x86_64-460.91.03-grid.run
./NVIDIA-Linux-x86_64-460.91.03-grid.run
```

3. Test whether the GRID driver is installed.

```
nvidia-smi
```

If the returned information is similar to the following command output, the GRID driver is installed.

```
root@i                    :~# nvidia-smi
Thu Aug 13 11:12:58 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 430.63       Driver Version: 430.63       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  GRID T4-8Q          On   | 00000000:00:07.0 Off |                  N/A |
| N/A   N/A    P0    N/A /  N/A |    528MiB /  8128MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

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

In the returned information, `Driver Version` indicates the version of the GRID driver, and `CUDA Version` indicates the latest Compute Unified Device Architecture (CUDA) version supported by the GRID driver version. The following table describes other available GRID driver versions and the latest CUDA versions supported by these driver versions.

| GRID driver version | Latest CUDA versions supported |
| --- | --- |
| 430.99 | 10.1 |
| 460.91.03 | 11.2 |
| 470.82.01 | 11.4 |

4. 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 following information about the license server to the *gridd.conf* file:

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

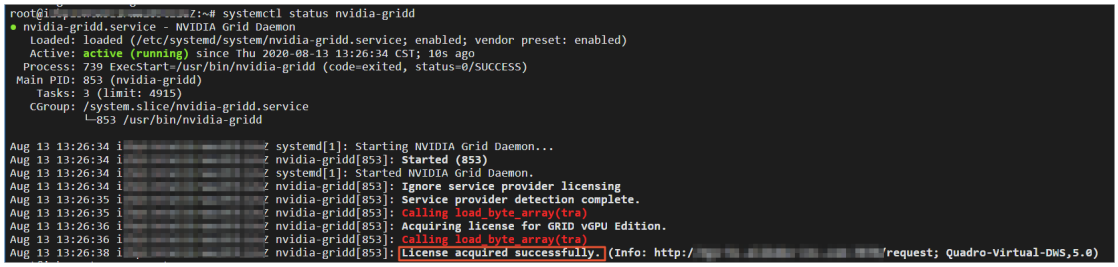
5. Reboot the instance for the configurations of the license server to take effect.

```
reboot
```

6. Check whether the license is activated.
   i. Connect to the instance. For more information, see Guidelines on instance connection.
   ii. Check the status of the license.

```
systemctl status nvidia-gridd
```

If **License acquired successfully** is contained in the command output, the license is activated.




## Test the acceleration effect on graphics

The following section describes how to test the acceleration effect on graphics for the GRID driver that is installed in an OpenGL Extension to the X Window System (GLX) application on the vGPU-accelerated instance. In this example, a vGPU-accelerated instance that runs Ubuntu 16.04 64-bit is used.

1. Prepare an environment to test the acceleration effect on graphics.
   i. Run the following command to install x11vnc:

```
apt-get install x11vnc
```

   ii. Run the `lspci | grep NVIDIA` command to obtain the value of the BusID parameter for the GPU that is used by the instance.

   In this example, the value of the BusID parameter is `00:07.0`.

iii. Configure the X Server environment and reboot the system.

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

b. Add the obtained value of the BusID parameter to `Section "Device"` in the */etc/X11/xorg.conf* file. In this example, `BusID "PCI:0:7:0"` is used.

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

c. Run the `reboot` command to reboot the system.

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

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

3. Run the `startx` command to start X Server.

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

```
root@iZ                Z:~# startx
hostname: Name or service not known
xauth: (stdin):1:  bad display name "iZ                Z:1" in "add" command
```

4. Start an SSH session on a client and run the following command to start x11vnc:

```
x11vnc -display :1
```

If the returned information is similar to the following command output, x11vnc is started. In this case, you can connect to the instance by using a VNC application, such as VNC Viewer.

```
The VNC desktop is:      localhost:0
PORT=5900
```

5. Log on to the ECS console and configure a rule for a security group to which the instance belongs. The rule allows inbound traffic on TCP port 5900. For more information, see Add a security group rule.

6. On your computer, start a VNC application such as VNC Viewer, enter `<Public IP address of the instance>:5900` to connect to the instance, and then go to K Desktop Environment (KDE).

7. Run the `glxinfo` command to view the configurations that are supported by the GRID driver.

i. Start another SSH session on a client.

ii. Run the `export DISPLAY=:1` command.

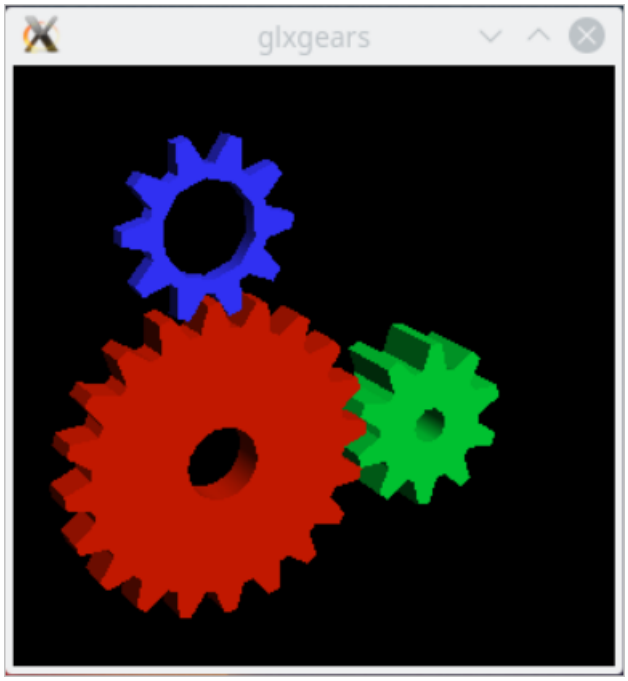
iii. Run the `glxinfo -t` command to obtain the configurations that are supported by the GRID driver.

8. Run the `glxgears` command to test the GRID driver.

i. On KDE, right-click the desktop and select **Run Command**.

i. On RDP, right-click the desktop and select **Run Command**.

ii. Run the `glxgears` command to start the GLX application.
If a window that is similar to the following figure is displayed, the GRID driver works as expected.



# 7.Uninstall the NVIDIA driver

## 7.1. Uninstall a GPU driver

This topic describes how to manually uninstall a GPU driver from a GPU-accelerated instance. The commands that you run to uninstall the GPU driver vary based on the installation method of the driver and the OS that the instance runs.

### Context

You can use GPU-accelerated instances only if relevant drivers are configured for the instances. Before you uninstall a driver from your instance, install another driver that matches your instance family and OS to ensure the performance of your instance. For more information, see Upgrade NVIDIA drivers.

If you select Auto-install GPU Driver when you create GPU-accelerated instances, the operations that you perform to uninstall the drivers vary based on the OSs that the instances run:

- Ubuntu: Uninstall the .deb installation package.
- CentOS or SUSE Linux: Uninstall the .run installation package.

To uninstall GPU drivers from different OSs, see the following references:

- Windows

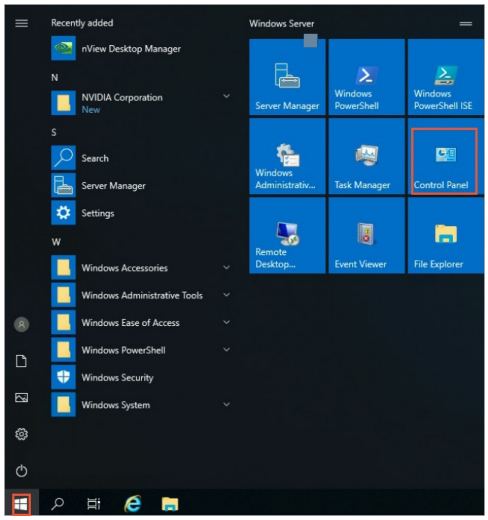
  Uninstall a GPU driver from Windows
- Linux
  - Uninstall a GPU driver from Ubuntu
  - Uninstall a GPU driver from CentOS
  - Uninstall a GPU driver from SUSE Linux

> **Note**
>
> To uninstall a GPU driver from a GPU-accelerated instance that runs Linux, you must log on to the instance as the `root` user before you proceed. If you are a regular user, run the `sudo` command to switch to the `root` user before you proceed.

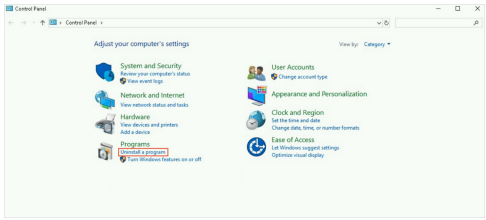
### Uninstall a GPU driver from Windows

In this example, an instance that belongs to the GPU-accelerated compute-optimized instance family gn6i and runs Windows Server 2019 is used.

1. Connect to the instance. For more information, see Connect to a Windows instance by using a password or key.
2. In the lower-left corner of a Windows desktop, click the  icon and click **Control Panel**.




3. In the **Control Panel** dialog box, click **Uninstall a program**.




4. Find the GPU driver that you want to uninstall, right-click the driver, and then click **Uninstall/Change**.
5. In the **NVIDIA Uninstaller** dialog box, click **UNINSTALL**.
6. After you uninstall the driver, click **RESTART NOW**.

## Uninstall a GPU driver from Ubuntu

If you use the .deb installation package to install a GPU driver in Ubuntu, we recommend that you perform the following operations to uninstall the GPU driver. In this example, NVIDIA driver 410.104, CUDA 10.0.130, and cuDNN 7.5.0 are used.

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

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

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

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

3. Run the following command to reboot the instance:

```
reboot
```

If you use the .run installation package to install a GPU driver in Ubuntu, 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 the CUDA and cuDNN libraries:

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

> **Note** The commands that you run to uninstall a driver may vary based on the CUDA version. If you cannot find the *cuda-uninstaller* file, check whether a file whose name contains the *uninstall_cuda* prefix exists in the */usr/local/cuda/bin/* directory. If the file whose name contains the uninstall_cuda prefix exists, replace *cuda-uninstaller* in the preceding command with the name of this file.

3. Run the following command to reboot the instance:

```
reboot
```

## Uninstall a GPU driver from CentOS

If you use the .run installation package to install a GPU driver in CentOS, we recommend that you perform the following operations to uninstall the GPU driver. In this example, NVIDIA driver 410.104, CUDA 10.0.130, and cuDNN 7.5.0 are used.

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

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

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

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

> **Note** The commands that you run to uninstall a driver may vary based on the CUDA version. If you cannot find the *cuda-uninstaller* file, check whether a file whose name contains the *uninstall_cuda* prefix exists in the */usr/local/cuda/bin/* directory. If the file whose name contains the *uninstall_cuda* prefix exists, replace *cuda-uninstaller* in the preceding command with the name of this file.

3. Run the following command to reboot the instance:

```
reboot
```

If you use the .rpm installation package to install a GPU driver in CentOS 7, we recommend that you perform the following operations to uninstall the GPU driver.

1. Run the following command 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 the CUDA and cuDNN libraries:

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

3. Run the following command to reboot the instance:

```
reboot
```

If you use the .rpm installation package to install a GPU driver in CentOS 6, we recommend that you perform the following operations to uninstall the GPU driver.

1. Run the following command 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 the CUDA library:

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

3. Run the following command to reboot the instance:

```
reboot
```

## Uninstall a GPU driver from SUSE Linux

If you use the .run installation package to install a GPU driver in SUSE Linux, we recommend that you perform the following operations to uninstall the GPU driver. In this example, CUDA 9.0.176 is used.

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

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

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

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

3. Run the following command to reboot the instance:

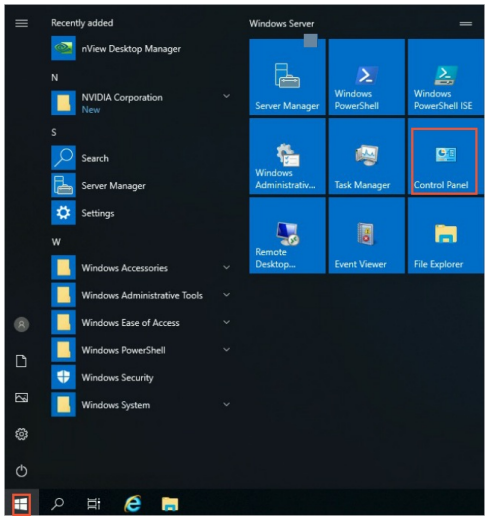
```
reboot
```

# 7.2.

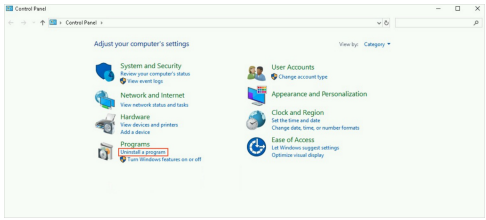
## Context

Upgrade NVIDIA drivers

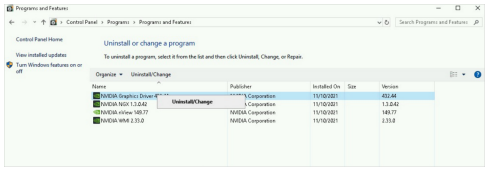
1. Connect from a local client that runs a Windows operating system
2. 



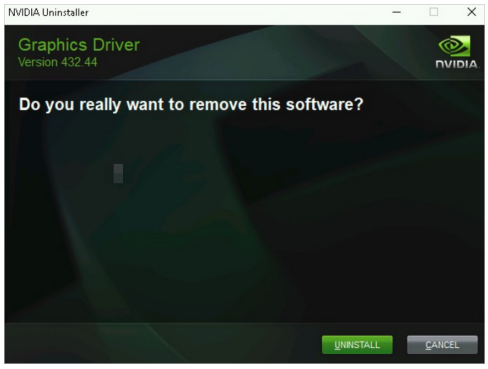

3.




4.




5.




6.

2021-09-16_15-16-43

1.

| | |
|---|---|
| VNC | Connect to a Linux instance by using a password |

```
nvidia-uninstall
```

ii.

Note


If you plan to no longer use the NVIDIA driver, you should make sure that no X screens are configured to use the NVIDIA X driver in your X configuration file. If you used nvidia-xconfig to configure X, it may have created a backup of your original configuration. Would you like to run `nvidia-xconfig --restore-original-backup` to attempt restoration of the original X configuration file?
Yes
No

iii.


Uninstallation of existing driver: NVIDIA Accelerated Graphics Driver for Linux-x86_64 (430.99) is complete.
OK

# 8.Upgrade NVIDIA drivers

If the version of your NVIDIA driver is no longer suitable for your scenarios, or if you install an NVIDIA driver whose type or version is invalid for your GPU-accelerated instance, you can uninstall the existing driver and install a new driver to upgrade the driver on the instance. This topic describes how to uninstall the existing driver and install a new driver on your GPU-accelerated instance. The methods that you can use to uninstall the existing driver and install the new driver vary based on the driver type and the OS that the drivers run.

## Uninstall the existing driver

Use one of the following methods based on the type of your driver:

- Uninstall a GPU driver
-

## Install a new driver

Use one of the following methods based on the type of your driver:

- If you need to install a GPU driver, use one of the following methods based on the OS that the driver runs:
  - Install a Windows GPU driver on a GPU-accelerated compute-optimized instance
  - Install a GPU driver on a Linux GPU-accelerated compute-optimized instance
- If you need to install a GRID driver, use one of the following methods based on the OS that the driver runs:
  - Install a GRID driver on a Windows GPU-accelerated instance

> **Notice** The GRID driver that you can install on a vGPU-accelerated instance varies based on the instance family to which the vGPU-accelerated instance belongs. If you install a GRID driver that does not match your instance family, you cannot use the instance. Therefore, if you need to install a new GRID driver on a vGPU-accelerated instance, install a GRID driver of the version that matches the instance family to which the instance belongs.