ALIBABA CLOUD

Alibaba Cloud

Elastic Compute Service

Instance

Document Version: 20220713

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.Overview

An Elastic Compute Service (ECS) instance is a virtual server that includes basic components such as vCPUs, memory, an operating system (OS), network configurations, and disks. You can use management tools provided by Alibaba Cloud such as the ECS console and ECS API to create and manage ECS instances. You can manage the status of ECS instances and their deployed applications in the same manner as you would do with local servers. You can also upgrade the capabilities (such as compute and storage capabilities) of your ECS instances as your requirements increase.

Basic instance configurations

The following basic configurations of each ECS instance determine the basic resources that the instance requires:

- Instance type

  Instance types define the basic attributes of ECS instances, such as compute capacity, storage capacity, and networking capacity. Instance types must be used together with images, Elastic Block Storage (EBS) devices, and network resources to create ECS instances that serve different purposes.

  ECS provides a variety of instance families for typical use scenarios. Each instance family consists of multiple instance types that have different compute capabilities to suit different scenarios and different levels of requirements. For information about available instance types, see Instance family. For suggestions about how to select appropriate instance types for different scenarios, see Best practices for instance type selection.

  If you use your own local servers, you may need to purchase new or replacement hardware as your compute needs change. This is costly and inconvenient. If you use ECS instances, you need only to upgrade or downgrade their instance types when your compute needs change. For more information, see Change instance types.

- Image

  Images contain the required information necessary to run ECS instances, such as OSs and initialization data of applications. Alibaba Cloud provides ready-to-use OS images for Windows Server and several mainstream Linux OSs. You can also create or import your own custom images to save time in making repeated configurations. In addition, image providers provide images pre-installed with a variety of runtime environments and software applications in Alibaba Cloud Marketplace. Alibaba Cloud Marketplace images are suitable for specific scenarios such as website building, application development, and visualized management. You can conveniently select Alibaba Cloud Marketplace images based on their purpose.

- Storage

  ECS instances use their attached system disks and data disks for storage. Each instance must have a system disk attached. The first time the instance starts, the OS is installed and instance configurations are initialized based on the image on the system disk.

  Cloud disks can be used as system disks or data disks. Local disks can be used only as data disks and are available only for specific instance types, such as big data instance types and instance types with local SSDs. If you want your instances to have more storage space, you can resize their attached cloud disks or attach more cloud disks after the instances are created. For more information, see Overview and Attach a data disk.

Business data is an important asset. Cloud disks adopt a triplicate mechanism to ensure the durability of data. To ensure that your data remains available, we recommend that you back up your data on a regular basis. You can create snapshots of cloud disks to back up disk data. If you are using local disks, you must implement data redundancy at the application layer to ensure data availability.

In addition to these basic configurations, you can customize network configurations, security groups, OS configurations, and grouping configurations for instances. For more information, see Create an instance by using the wizard.

Instance purchasing options

ECS provides a variety of instance purchasing options. You can select a purchasing option that meets your needs in different scenarios. Examples:

- To obtain stable computing power, you can purchase subscription instances.
- To meet dynamic computing power requirements, you can purchase pay-as-you-go instances.
- To meet the computing power requirements of stateless applications and reduce compute costs, you can purchase preemptible instances.
- To gain flexibility and cost-effectiveness in resource use, you can use pay-as-you-go instances in conjunction with reserved instances or savings plans.
- To gain guaranteed access to compute resources, you can use pay-as-you-go instances in conjunction with resource assurances.

For information about more purchasing options, see Overview.

Usage instructions

- Create instances
- Connect to instances
- Manage instance status
- Manage instance attributes
- Manage instance configurations
- Renew subscription instances
- Change instance configurations
- Limits

Security suggestions

When you use cloud services, we recommend that you follow security suggestions to improve the security of cloud resources. Examples:

- Suggestions for permission control: Use Resource Access Management (RAM) features to control which users can manage resources such as instances and what permissions to grant to the users.
- Suggestions for security features: Use security features such as security hardening and cloud disk encryption to ensure the security of data and runtime environments.
- Suggestions for network security: Use virtual private clouds (VPCs) to isolate services of different security levels. Use security groups to control inbound and outbound traffic for instances and allow instances access to the Internet only when required to minimize the attack surface area of resources.

For more information about how to improve the security of instances, see Best practices for security.

2.Instance lifecycle

An Elastic Compute Service (ECS) instance transitions through different states from the moment it is created to the moment it is released.

Instance states

Instance states are classified into console-based states and API-based states based on where the states can be queried. Console-based states are the instance states that can be queried in the ECS console. API-based states are the instance states that can be queried in ECS API by calling the DescribeInstanceStatus or DescribeInstances operation. An API-based state may correspond to multiple console-based states based on whether a subscription instance has expired or whether a payment is overdue for an instance within your account.

Instance states are classified into stable and transitory states based on their attributes. Transitory states are the states that an instance temporarily enters before it enters a stable state. If an instance remains in a transitory state for an extended period of time, an exception occurs.

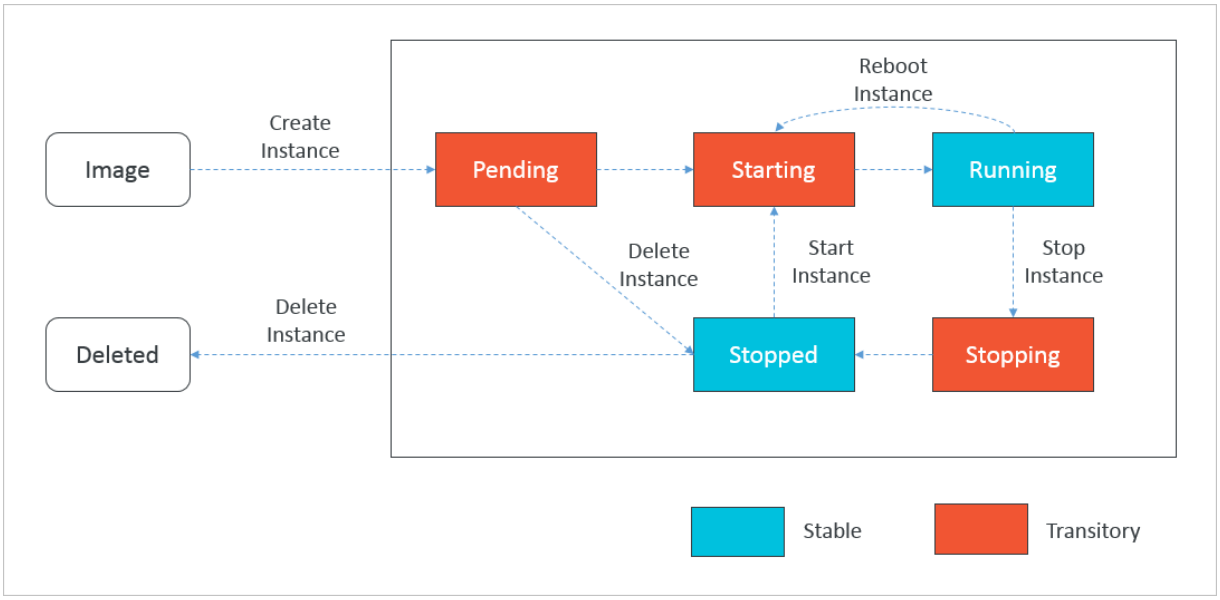
The following table describes the different states that an instance may go through during its lifecycle.

Console-based state	API-based state	State attribute	Description
Pending	Pending	Transitory	After an instance is created, it is in this state before it enters the Starting (Starting) state.
Starting	Starting	Transitory	After an instance is created, started, or restarted, it is in this state before it enters the Running (Running) state.
Running	Running	Stable	When an instance is running normally, it is in this state. **Note** An instance can be externally accessed only when it is in the API-based Running state. Instances in the API-based Running state may be in the console-based Running or Expiring state.
Expiring	Running	Stable	When a subscription instance is about to expire, it enters this state but continues to run normally. We recommend that you renew the instance at your earliest convenience. For more information, see Renewal overview.
Stopping	Stopping	Transitory	After you stop or hibernate an instance, it enters this state before it enters the Stopped (Stopped) state.

Console-based state	API-based state	State attribute	Description
Stopped	Stopped	Stable	When an instance is created but not started or after an instance is stopped or hibernated, it remains in this state. **Note** After you create an instance by using the ECS console or by calling the RunInstances operation, the instance is automatically started.
Expired	Stopped	Stable	When a subscription instance expires or when a pay-as-you-go instance is stopped due to an overdue payment, the instance enters this state and is pending release. For information about whether instance resources are retained, see Changes in resource states after a subscription instance expires and Pay-as-you-go.
Locked	Stopped	Stable	Your instance may enter this state for specific reasons. For example, your instance may enter the Locked state because you have an overdue payment within your account or because security risks are detected within your account. To unlock the instance, you can submit a ticket.
To Be Released	Stopped	Stable	When you apply for a refund for an unexpired subscription instance, the instance enters this state.

Manage the status of instances

The following figure shows the transitions between API-based instance states.



To manage the status of instances, we recommend that you use the management tools provided by Alibaba Cloud, such as the ECS console and ECS API. For example, to enable the economical mode when you are stopping a pay-as-you-go instance or to restart an instance for a new hostname to take effect, you must use the ECS console or call an API operation instead of performing operations within the instance operating system.

To manage the status of instances, you can perform the following operations:

- Create an instance

  The created instance enters the Pending (Pending) state, the Starting (Starting) state, and then the Running (Running) state. You can access the instance when it is in the Running (Running) state. For example, you can connect to the instance to manage its operating system configurations or build websites and use a browser to access the websites.

- Stop an instance

  The instance enters the Stopping (Stopping) state and then the Stopped (Stopped) state. You must stop an instance before you can perform specific operations on the instance, such as replacing the operating system, changing the private IP address, and changing the instance type if the instance is a pay-as-you-go one.

  If you enable the economical mode when you stop a pay-as-you-go instance, the computing resources (vCPUs and memory) and public IP address of the instance are released, and you are no longer charged for them. Other resources of the instance including the disks and the associated elastic IP address (EIP) are retained, and you continue to be charged for them.

- Start an instance

  The instance enters the Starting (Starting) state and then the Running (Running) state.

- Hibernate an instance

  The instance enters the Stopping (Stopping) state and then the Stopped (Stopped) state. When you hibernate an instance, the operating system of the instance saves data stored in memory to the system disk. This data includes running applications and their current states. When you wake the instance, the operating system reads the memory data saved to the system disk. Meanwhile, the operating system resumes the previously running applications and restores them to the states they were in when the instance was hibernated.

  When you hibernate a pay-as-you-go instance for which the No Fees for Hibernated Instances mode is enabled, the computing resources (vCPUs and memory) and public IP address of the instance are released, and you are no longer charged for them. Other resources of the instance including the disks and the associated EIP are retained, and you continue to be charged for them.

- Restart an instance

  The instance enters the Stopping (Stopping) state, the Starting (Starting) state, and then the Running (Running) state. After you perform specific instance operations such as changing the instance type, you must restart the instance for the operations to take effect.

  When an instance is restarted, it may be moved to a new host. If you want your instances to remain on the same host, you can purchase a dedicated host and then associate instances with it.

- Release an instance

  Only instances in the API-based Stopped state can be released. Instances in the API-based Stopped state may be in the console-based Stopped or Expired state. Unexpired subscription instances cannot be released. If you want to release an unexpired subscription instance, you can apply for a refund or change the instance into a pay-as-you-go instance.

When an instance is released, its ID, public IP address, system disk, and data disks for which Release with Instance is enabled are also released and cannot be recovered. If the instance has been associated with an EIP, the EIP is automatically disassociated from the instance and retained. The data disks for which Release with Instance is not enabled are automatically detached from the instance and retained. Proceed with caution when you release instances. To prevent accidental release of instances, we recommend that you enable release protection for the instances.

3.Instance family

An Elastic Compute Service (ECS) instance is a virtual server and the smallest computing service unit in the cloud. An instance type essentially determines the hardware of the host computer used for your instance. Each instance type offers different compute and memory capabilities. This topic describes all ECS instance families available for sale and introduces their features, specifications, and use scenarios.

ECS provides a variety of instance families for different use scenarios or application workloads. Each instance family offers multiple instance types based on their CPU and memory specifications. *ECS instance type* defines the basic properties of an ECS instance, including CPU (CPU model and clock speed) and memory. In addition to the instance type, you must also configure the Elastic Block Storage (EBS) devices, image, and network type when you create an ECS instance.

> **Note** The available instance families and types vary based on regions. You can go to the ECS Instance Types Available for Each Region page to view the available instance types in each region.

Enterprise applications require high business stability. Alibaba Cloud ECS instance families are categorized into enterprise-level and shared instance families based on whether the instance families are suitable for enterprise scenarios. Enterprise-level instance families offer consistent performance and dedicated resources. In enterprise-level instance families, each vCPU corresponds to a hyperthread of an Intel® Xeon® core. For more information about the differences between enterprise-level and shared instance families, see Instance FAQ.

You can upgrade or downgrade instance types within the same instance family or across different instance families. For more information, see Instance families that support instance type changes.

For information about how to choose instance families based on scenarios, see Best practices for instance type selection.

Alibaba Cloud ECS instance families are categorized into the following categories based on their system architecture and use scenarios.

Enterprise-level computing instance families based on the x86 architecture	
Recommended instance families	Other available instance families (If these instance families are sold out, you can use the recommended ones.)
• g7se, storage-enhanced general-purpose instance family • g7a, general-purpose instance family • g7, general-purpose instance family • g7t, security-enhanced general-purpose instance family • g7ne, network-enhanced general-purpose instance family • g6, general-purpose instance family • g6a, general-purpose instance family • g6t, security-enhanced general-purpose instance family • g6e, general-purpose instance family with	

Enterprise-level computing instance families based on the x86 architecture	
enhanced performance • g5, general-purpose instance family • g5ne, network-enhanced general-purpose instance family • c7se, storage-enhanced compute-optimized instance family • c7a, compute-optimized instance family • c7, compute-optimized instance family • c7t, security-enhanced compute-optimized instance family • c6, compute-optimized instance family • c6a, compute-optimized instance family • c6t, security-enhanced compute-optimized instance family • c6e, compute-optimized instance family with enhanced performance • c5, compute-optimized instance family • ic5, compute-intensive instance family • r7se, storage-enhanced memory-optimized instance family • r7a, memory-optimized instance family • r7, memory-optimized instance family • r7t, security-enhanced memory-optimized instance family • r6, memory-optimized instance family • re6p, persistent memory-optimized instance family • r6a, memory-optimized instance family • r6e, memory-optimized instance family with enhanced performance • re6, high-memory instance family • r5, memory-optimized instance family • d3c, compute-intensive big data instance family • d2c, compute-intensive big data instance family • d2s, storage-intensive big data instance family • d1ne, network-enhanced big data instance family • i3g, instance family with local SSDs • i3, instance family with local SSDs • i2, instance family with local SSDs • i2g, instance family with local SSDs • i2ne, instance family with local SSDs • i2gne, instance family with local SSDs • hfc7, compute-optimized instance family with high clock speeds • hfc6, compute-optimized instance family with high clock speeds	• sn2ne, network-enhanced general-purpose instance family • sn1ne, network-enhanced compute-optimized instance family • re4, high-memory instance family • re4e, high-memory instance family • se1ne, network-enhanced memory-optimized instance family • se1, memory-optimized instance family • d1, big data instance family • i1, instance family with local SSDs • hfc5, compute-optimized instance family with high clock speeds • hfg5, general-purpose instance family with high clock speeds

Enterprise-level computing instance families based on the x86 architecture	
• hfg7, general-purpose instance family with high clock speeds • hfg6, general-purpose instance family with high clock speeds • hfr7, memory-optimized instance family with high clock speeds • hfr6, memory-optimized instance family with high clock speeds	

Enterprise-level computing instance families based on the ARM architecture
• g8m, general-purpose instance family

Enterprise-level heterogeneous computing instance families	
Recommended instance families	Other available instance families (If these instance families are sold out, you can use the recommended ones.)
• sgn7i-vws, vGPU-accelerated instance family with shared CPUs • vgn7i-vws, vGPU-accelerated instance family • gn7i, GPU-accelerated compute-optimized instance family • gn7, GPU-accelerated compute-optimized instance family • vgn6i, vGPU-accelerated instance family • gn6i, GPU-accelerated compute-optimized instance family • gn6e, GPU-accelerated compute-optimized instance family • gn6v, GPU-accelerated compute-optimized instance family • f3, FPGA-accelerated compute-optimized instance family	• vgn5i, vGPU-accelerated instance family • gn5, GPU-accelerated compute-optimized instance family • gn5i, GPU-accelerated compute-optimized instance family • f1, FPGA-accelerated compute-optimized instance family

ECS Bare Metal Instance families and Super Computing Cluster (SCC) instance families	
Recommended instance families	Other available instance families (If these instance families are sold out, you can use the recommended ones.)
• ebmgn7e, GPU-accelerated compute-optimized ECS Bare Metal Instance family • ebmgn7i, GPU-accelerated compute-optimized ECS Bare Metal Instance family • ebmgn7, GPU-accelerated compute-optimized ECS Bare Metal Instance family	

ECS Bare Metal Instance families	Super Computing Cluster (SCC) instance families
• ebmgn6ia, GPU-accelerated compute-optimized ECS Bare Metal Instance family • ebmgn6e, GPU-accelerated compute-optimized ECS Bare Metal Instance family • ebmgn6v, GPU-accelerated compute-optimized ECS Bare Metal Instance family • ebmgn6i, GPU-accelerated compute-optimized ECS Bare Metal Instance family • ebmc7, compute-optimized ECS Bare Metal Instance family • ebmc7a, compute-optimized ECS Bare Metal Instance family • ebmc6me, compute-optimized ECS Bare Metal Instance family • ebmc6a, compute-optimized ECS Bare Metal Instance family • ebmc6e, compute-optimized ECS Bare Metal Instance family with enhanced performance • ebmc6, compute-optimized ECS Bare Metal Instance family • ebmg7, general-purpose ECS Bare Metal Instance family • ebmg7a, general-purpose ECS Bare Metal Instance family • ebmg6a, general-purpose ECS Bare Metal Instance family • ebmg6e, general-purpose ECS Bare Metal Instance family with enhanced performance • ebmg6, general-purpose ECS Bare Metal Instance family • ebmr7, memory-optimized ECS Bare Metal Instance family • ebmr7a, memory-optimized ECS Bare Metal Instance family • ebmr6a, memory-optimized ECS Bare Metal Instance family • ebmr6e, memory-optimized ECS Bare Metal Instance family with enhanced performance • ebmr6, memory-optimized ECS Bare Metal Instance family • ebmre6p, persistent memory-optimized ECS Bare Metal Instance family with enhanced performance • ebmre6-6t, memory-optimized ECS Bare Metal Instance family with enhanced performance • ebmhfg7, general-purpose ECS Bare Metal Instance family with high clock speeds • ebmhfc7, compute-optimized ECS Bare Metal Instance family with high clock speeds • ebmhfr7, memory-optimized ECS Bare Metal	• ebmc5s, network-enhanced compute-optimized ECS Bare Metal Instance family • ebmg5s, network-enhanced general-purpose ECS Bare Metal Instance family • ebmr5s, network-enhanced memory-optimized ECS Bare Metal Instance family • ebmg5, general-purpose ECS Bare Metal Instance family • ebmc4, compute-optimized ECS Bare Metal Instance family

ECS Bare Metal Instance families and Super Computing Cluster (SCC) instance families	
Instance family with high clock speeds • ebmhfg6, general-purpose ECS Bare Metal Instance family with high clock speeds • ebmhfc6, compute-optimized ECS Bare Metal Instance family with high clock speeds • ebmhfr6, memory-optimized ECS Bare Metal Instance family with high clock speeds • scchfc6, compute-optimized SCC instance family with high clock speeds • scchfg6, general-purpose SCC instance family with high clock speeds • scchfr6, memory-optimized SCC instance family with high clock speeds • scch5, SCC instance family with high clock speeds • sccg5, general-purpose SCC instance family • sccgn7ex, GPU-accelerated compute-optimized SCC instance family • sccgn6e, GPU-accelerated compute-optimized SCC instance family • sccgn6, GPU-accelerated compute-optimized SCC instance family	

Shared computing instance families based on the x86 architecture	
Recommended instance families	Other available instance families (If these instance families are sold out, you can use the recommended ones.)
• t6, burstable instance family	• t5, burstable instance family • v5, CPU-overprovisioned instance family • xn4, n4, mn4, and e4, previous-generation shared instance families

For information about retired instance families, see Retired instance types.

Enterprise-level computing instance families based on the x86 architecture

d3c, compute-intensive big data instance family

> **Note** This instance family is in invitational preview. To use this instance family,.

Features:

- This instance family is equipped with high-capacity and high-throughput local SSDs and can provide maximum bandwidth of 32 Gbit/s between instances.
- Supports online replacement and hot swapping of damaged disks to prevent instance shutdown.

If a local disk fails, you receive a notification about the system event. You can handle the system event by initiating the process of fixing the damaged disk. For more information, see O&M scenarios and system events for instances equipped with local disks.

> **Notice** After you initiate the process of fixing the damaged disk, data in the damaged disk cannot be restored.

- Compute:
  - Uses the third-generation 2.7 GHz Intel® Xeon® Scalable (Ice Lake) processors that deliver an all-core turbo frequency of 3.5 GHz for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports enhanced SSDs (ESSDs), standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Big data computing and storage business scenarios in which services such as Hadoop MapReduce, HDFS, Hive, and HBase are used
  - Scenarios in which EMR JindoFS and Operation Orchestration Service (OOS) are used in combination to separately store hot and cold data and decouple storage from computing
  - Machine learning scenarios such as Spark in-memory computing and MLlib
  - Search and log data processing scenarios in which solutions such as Elasticsearch and Kafka are used

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d3c.3xlarge	14	56.0	1 × 16000	8/burstable up to 10	1,600,000	8	8	30
ecs.d3c.7xlarge	28	112.0	2 × 16000	16/burstable up to 25	2,500,000	16	8	30
ecs.d3c.14xlarge	56	224.0	4 × 16000	32/none	5,000,000	28	8	30
ecs.d3c.16xlarge	64	256.0	4 × 16000	32/none	5,000,000	32	8	30

Note

-
- For more information about these specifications, see Instance family.

d2c, compute-intensive big data instance family

Features:

- This instance family is equipped with high-capacity and high-throughput local SATA HDDs and can provide a maximum bandwidth of 35 Gbit/s between instances.
- Supports online replacement and hot swapping of damaged disks to prevent instance shutdown.

  If a local disk fails, you receive a notification about the system event. You can handle the system event by initiating the process of fixing the damaged disk. For more information, see O&M scenarios and system events for instances equipped with local disks.

  Notice After you initiate the process of fixing the damaged disk, data in the damaged disk cannot be restored.

- Compute:
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Big data computing and storage business scenarios in which services such as Hadoop MapReduce, HDFS, Hive, and HBase are used
  - Scenarios in which EMR JindoFS and OOS are used in combination to separately store hot and cold data and decouple storage from computing
  - Machine learning scenarios such as Spark in-memory computing and MLlib
  - Search and log data processing scenarios in which solutions such as Elasticsearch and Kafka are used

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d2c.6xlarge	24	88.0	3 × 4000	12.0	1,600,000	8	8	20

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d2c.12xlarge	48	176.0	6 × 4000	20.0	2,000,000	16	8	20
ecs.d2c.24xlarge	96	352.0	12 × 4000	35.0	4,500,000	16	8	20

Note

-
- For more information about these specifications, see Instance family.

d2s, storage-intensive big data instance family

Features:

- This instance family is equipped with high-capacity and high-throughput local SATA HDDs and can provide a maximum bandwidth of 35 Gbit/s between instances.
- Supports online replacement and hot swapping of damaged disks to prevent instance shutdown.

  If a local disk fails, you receive a notification about the system event. You can handle the system event by initiating the process of fixing the damaged disk. For more information, see O&M scenarios and system events for instances equipped with local disks.

  Notice After you initiate the process of fixing the damaged disk, data in the damaged disk cannot be restored.

- Compute:
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Big data computing and storage business scenarios in which services such as Hadoop MapReduce, HDFS, Hive, and HBase are used
  - Machine learning scenarios such as Spark in-memory computing and MLlib
  - Search and log data processing scenarios in which solutions such as Elasticsearch and Kafka are used

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d2s.5xlarge	20	88.0	8 × 7300	12.0	1,600,000	8	8	20
ecs.d2s.10xlarge	40	176.0	15 × 7300	20.0	2,000,000	16	8	20
ecs.d2s.20xlarge	80	352.0	30 × 7300	35.0	4,500,000	32	8	20

Note

-
- For more information about these specifications, see Instance family.

d1ne, network-enhanced big data instance family

Features:

- This instance family is equipped with high-capacity and high-throughput local SATA HDDs and can provide a maximum bandwidth of 35 Gbit/s between instances.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4, which is designed for big data scenarios.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios in which services such as Hadoop MapReduce, HDFS, Hive, and HBase are used
  - Machine learning scenarios such as Spark in-memory computing and MLlib
  - Search and log data processing scenarios in which solutions such as Elasticsearch are used

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d1ne.2xlarge	8	32.0	4 × 5500	6.0	1,000,000	4	4	10
ecs.d1ne.4xlarge	16	64.0	8 × 5500	12.0	1,600,000	4	8	20
ecs.d1ne.6xlarge	24	96.0	12 × 5500	16.0	2,000,000	6	8	20
ecs.d1ne-c8d3.8xlarge	32	128.0	12 × 5500	20.0	2,000,000	6	8	20
ecs.d1ne.8xlarge	32	128.0	16 × 5500	20.0	2,500,000	8	8	20
ecs.d1ne-c14d3.14xlarge	56	160.0	12 × 5500	35.0	4,500,000	14	8	20
ecs.d1ne.14xlarge	56	224.0	28 × 5500	35.0	4,500,000	14	8	20

Note

-
- For more information about these specifications, see Instance family.

hfc7, compute-optimized instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses Intel® Xeon® Cooper Lake processors that deliver an all-core turbo frequency of 3.8 GHz and have a minimum clock speed of 3.3 GHz for consistent computing performance.

- Allows you to enable or disable Hyper-Threading.

  Note

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only enhanced SSDs (ESSDs) and provides ultra-high I/O performance.
  - Provides high storage I/O performance based on large compute capacity.

  Note

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance frontend server clusters
  - Frontend servers of massive multiplayer online (MMO) games
  - Data analysis, batch processing, and video encoding
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfc7.large	2	4	1.2/10	900,000	250,000	2	2	6	20,000	1
ecs.hfc7.xlarge	4	8	2/10	1,000,000	250,000	4	3	15	30,000	1.5
ecs.hfc7.2xlarge	8	16	3/10	1,600,000	250,000	8	4	15	45,000	2
ecs.hfc7.3xlarge	12	24	4.5/10	2,000,000	250,000	8	6	15	60,000	2.5

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfc7.4xlarge	16	32	6/10	2,500,000	300,000	8	8	30	75,000	3
ecs.hfc7.6xlarge	24	48	8/10	3,000,000	450,000	12	8	30	90,000	4
ecs.hfc7.8xlarge	32	64	10/none	4,000,000	600,000	16	8	30	105,000	5
ecs.hfc7.12xlarge	48	96	16/none	6,000,000	1,000,000	24	8	30	150,000	8
ecs.hfc7.24xlarge	96	192	32/none	12,000,000	1,800,000	32	15	30	300,000	16

Note

-
- For more information about these specifications, see Instance family.

hfc6, compute-optimized instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2.

- Uses 3.1 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.5 GHz for consistent computing performance.

  > **Note** The processors used by this instance family have a clock speed of 3.1 GHz. However, the Intel System Studio (ISS) feature may cause a lower clock speed to be displayed. Alibaba Cloud is working on this issue. This issue does not affect the actual clock speeds of your instances.
  >
  > You can separately run the following commands to use the turbostat tool to view the actual clock speeds:
  >
  > ```
  > yum install kernel-tools
  > ```
  >
  > ```
  > turbostat
  > ```

- Allows you to enable or disable Hyper-Threading.

  > **Note**

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Provides high storage I/O performance based on large compute capacity.

    > **Note**

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Web frontend servers
  - Frontend servers of MMO games
  - Data analysis, batch processing, and video encoding
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfc6.large	2	4	1/3	300,000	35,000	2	2	6	10,000	1
ecs.hfc6.xlarge	4	8	1.5/5	500,000	70,000	4	3	10	20,000	1.5
ecs.hfc6.2xlarge	8	16	2.5/8	800,000	150,000	8	4	10	25,000	2
ecs.hfc6.3xlarge	12	24	4/10	900,000	220,000	8	6	10	30,000	2.5
ecs.hfc6.4xlarge	16	32	5/10	1,000,000	300,000	8	8	20	40,000	3
ecs.hfc6.6xlarge	24	48	7.5/10	1,500,000	450,000	12	8	20	50,000	4
ecs.hfc6.8xlarge	32	64	10/none	2,000,000	600,000	16	8	20	60,000	5
ecs.hfc6.10xlarge	40	96	12.5/none	3,000,000	1,000,000	32	7	20	100,000	8
ecs.hfc6.16xlarge	64	128	20/none	4,000,000	1,200,000	32	8	20	120,000	10
ecs.hfc6.20xlarge	80	192	25/none	6,000,000	1,800,000	32	15	20	200,000	16

Note

-
- For more information about these specifications, see Instance family.

hfg7, general-purpose instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses Intel® Xeon® Cooper Lake processors that deliver an all-core turbo frequency of 3.8 GHz and have a minimum clock speed of 3.3 GHz for consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs and provides ultra-high I/O performance.
  - Provides high storage I/O performance based on large compute capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Enterprise-grade applications of various types and sizes
  - Game servers
  - Small and medium-sized database systems, caches, and search clusters
  - High-performance scientific computing
  - Video encoding applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfg7.large	2	8	1.2/10	900,000	250,000	2	2	6	20,000	1
ecs.hfg7.xlarge	4	16	2/10	1,000,000	250,000	4	3	15	30,000	1.5
ecs.hfg7.2xlarge	8	32	3/10	1,600,000	250,000	8	4	15	45,000	2
ecs.hfg7.3xlarge	12	48	4.5/10	2,000,000	250,000	8	6	15	60,000	2.5
ecs.hfg7.4xlarge	16	64	6/10	2,500,000	300,000	8	8	30	75,000	3
ecs.hfg7.6xlarge	24	96	8/10	3,000,000	450,000	12	8	30	90,000	4
ecs.hfg7.8xlarge	32	128	10/none	4,000,000	600,000	16	8	30	105,000	5
ecs.hfg7.12xlarge	48	192	16/none	6,000,000	1,000,000	24	8	30	150,000	8
ecs.hfg7.24xlarge	96	384	32/none	12,000,000	1,800,000	32	15	30	300,000	16

Note

-
- For more information about these specifications, see Instance family.

hfg6, general-purpose instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with

the use of SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.

- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 3.1 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.5 GHz for consistent computing performance.

    > **Note** The processors used by this instance family have a clock speed of 3.1 GHz. However, the Intel System Studio (ISS) feature may cause a lower clock speed to be displayed. Alibaba Cloud is working on this issue. This issue does not affect the actual clock speeds of your instances.
    >
    > You can separately run the following commands to use the turbostat tool to view the actual clock speeds:
    >
    > ```
    > yum install kernel-tools
    > ```
    >
    > ```
    > turbostat
    > ```

  - Allows you to enable or disable Hyper-Threading.

    > **Note**

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Provides high storage I/O performance based on large compute capacity.

    > **Note**

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Enterprise-grade applications of various types and sizes
  - Websites and application servers
  - Game servers
  - Small and medium-sized database systems, caches, and search clusters
  - Data analysis and computing
  - Computing clusters and memory-intensive data processing

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfg6.large	2	8	1/3	300,000	35,000	2	2	6	10,000	1
ecs.hfg6.xlarge	4	16	1.5/5	500,000	70,000	4	3	10	20,000	1.5
ecs.hfg6.2xlarge	8	32	2.5/8	800,000	150,000	8	4	10	25,000	2
ecs.hfg6.3xlarge	12	48	4/10	900,000	220,000	8	6	10	30,000	2.5
ecs.hfg6.4xlarge	16	64	5/10	1,000,000	300,000	8	8	20	40,000	3
ecs.hfg6.6xlarge	24	96	7.5/10	1,500,000	450,000	12	8	20	50,000	4
ecs.hfg6.8xlarge	32	128	10/none	2,000,000	600,000	16	8	20	60,000	5
ecs.hfg6.10xlarge	40	192	12.5/none	3,000,000	1,000,000	32	7	20	100,000	8
ecs.hfg6.16xlarge	64	256	20/none	4,000,000	1,200,000	32	8	20	120,000	10
ecs.hfg6.20xlarge	80	384	25/none	6,000,000	1,800,000	32	15	20	200,000	16

Note

-
- For more information about these specifications, see Instance family.

hfr7, memory-optimized instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses Intel® Xeon® Cooper Lake processors that deliver an all-core turbo frequency of 3.8 GHz and have a minimum clock speed of 3.3 GHz for consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs and provides ultra-high I/O performance.
  - Provides high storage I/O performance based on large compute capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance databases and in-memory databases
  - Data analysis, data mining, and distributed memory caching
  - Hadoop clusters, Spark clusters, and other enterprise-level memory-intensive applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfr7.large	2	16	1.2/10	900,000	250,000	2	2	6	20,000	1

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfr7.xlarge	4	32	2/10	1,000,000	250,000	4	3	15	30,000	1.5
ecs.hfr7.2xlarge	8	64	3/10	1,600,000	250,000	8	4	15	45,000	2
ecs.hfr7.3xlarge	12	96	4.5/10	2,000,000	250,000	8	6	15	60,000	2.5
ecs.hfr7.4xlarge	16	128	6/10	2,500,000	300,000	8	8	30	75,000	3
ecs.hfr7.6xlarge	24	192	8/10	3,000,000	450,000	12	8	30	90,000	4
ecs.hfr7.8xlarge	32	256	10/none	4,000,000	600,000	16	8	30	105,000	5
ecs.hfr7.12xlarge	48	384	16/none	6,000,000	1,000,000	24	8	30	150,000	8
ecs.hfr7.24xlarge	96	768	32/none	12,000,000	1,800,000	32	15	30	300,000	16

Note

-
- For more information about these specifications, see Instance family.

hfr6, memory-optimized instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:

- Offers a CPU-to-memory ratio of 1:8.
- Uses 3.1 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.5 GHz for consistent computing performance.

  > **Note** The processors used by this instance family have a clock speed of 3.1 GHz. However, the Intel System Studio (ISS) feature may cause a lower clock speed to be displayed. Alibaba Cloud is working on this issue. This issue does not affect the actual clock speeds of your instances.
  >
  > You can separately run the following commands to use the turbostat tool to view the actual clock speeds:
  >
  > ```
  > yum install kernel-tools
  > ```
  >
  > ```
  > turbostat
  > ```

- Allows you to enable or disable Hyper-Threading.

  > **Note**

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Provides high storage I/O performance based on large compute capacity.

    > **Note**

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance databases and in-memory databases
  - Data analysis, data mining, and distributed memory caching
  - Hadoop clusters, Spark clusters, and other enterprise-level memory-intensive applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfr6.large	2	16	1/3	300,000	35,000	2	2	6	10,000	1
ecs.hfr6.xlarge	4	32	1.5/5	500,000	70,000	4	3	10	20,000	1.5
ecs.hfr6.2xlarge	8	64	2.5/8	800,000	150,000	8	4	10	25,000	2
ecs.hfr6.3xlarge	12	96	4/10	900,000	220,000	8	6	10	30,000	2.5
ecs.hfr6.4xlarge	16	128	5/10	1,000,000	300,000	8	8	20	40,000	3
ecs.hfr6.6xlarge	24	192	7.5/10	1,500,000	450,000	12	8	20	50,000	4
ecs.hfr6.8xlarge	32	256	10/none	2,000,000	600,000	16	8	20	60,000	5
ecs.hfr6.10xlarge	40	384	12.5/none	3,000,000	1,000,000	32	7	20	100,000	8
ecs.hfr6.16xlarge	64	512	20/none	4,000,000	1,200,000	32	8	20	120,000	10
ecs.hfr6.20xlarge	80	768	25/none	6,000,000	1,800,000	32	15	20	200,000	16

Note

-
- For more information about these specifications, see Instance family.

d1, big data instance family

Features:

- This instance family is equipped with high-capacity and high-throughput local SATA HDDs and can provide a maximum bandwidth of 17 Gbit/s between instances.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4, which is designed for big data scenarios.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios in which services such as Hadoop MapReduce, HDFS, Hive, and HBase are used
  - Machine learning scenarios such as Spark in-memory computing and MLlib
  - Scenarios in which customers in industries such as Internet and finance need to compute, store, and analyze big data
  - Search and log data processing scenarios in which solutions such as Elasticsearch are used

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d1.2xlarge	8	32.0	4 × 5500	3.0	300,000	1	4	10
ecs.d1.3xlarge	12	48.0	6 × 5500	4.0	400,000	1	6	10
ecs.d1.4xlarge	16	64.0	8 × 5500	6.0	600,000	2	8	20
ecs.d1.6xlarge	24	96.0	12 × 5500	8.0	800,000	2	8	20
ecs.d1-c8d3.8xlarge	32	128.0	12 × 5500	10.0	1,000,000	4	8	20
ecs.d1.8xlarge	32	128.0	16 × 5500	10.0	1,000,000	4	8	20

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d1-c14d3.14xlarge	56	160.0	12 × 5500	17.0	1,800,000	6	8	20
ecs.d1.14xlarge	56	224.0	28 × 5500	17.0	1,800,000	6	8	20

Note

-
- For more information about these specifications, see Instance family.

hfc5, compute-optimized instance family with high clock speeds

Features

- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses 3.1 GHz Intel® Xeon® Gold 6149 (Skylake) processors.
  - Offers consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - High-performance web frontend servers
  - High-performance scientific and engineering applications
  - MMO gaming and video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.hfc5.large	2	4	1	300,000	2	2	6

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.hfc5.xlarge	4	8	1.5	500,000	2	3	10
ecs.hfc5.2xlarge	8	16	2	1,000,000	2	4	10
ecs.hfc5.3xlarge	12	24	2.5	1,300,000	4	6	10
ecs.hfc5.4xlarge	16	32	3	1,600,000	4	8	20
ecs.hfc5.6xlarge	24	48	4.5	2,000,000	6	8	20
ecs.hfc5.8xlarge	32	64	6	2,500,000	8	8	20

Note

-
- For more information about these specifications, see Instance family.

hfg5, general-purpose instance family with high clock speeds

Features

- Compute:
  - Offers a CPU-to-memory ratio of 1:4 (excluding the instance type with 56 vCPUs).
  - Uses 3.1 GHz Intel® Xeon® Gold 6149 (Skylake) processors.
  - Offers consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - High-performance web frontend servers
  - High-performance scientific and engineering applications
  - MMO gaming and video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.hfg5.large	2	8	1	300,000	2	2	6
ecs.hfg5.xlarge	4	16	1.5	500,000	2	3	10
ecs.hfg5.2xlarge	8	32	2	1,000,000	2	4	10
ecs.hfg5.3xlarge	12	48	2.5	1,300,000	4	6	10
ecs.hfg5.4xlarge	16	64	3	1,600,000	4	8	20
ecs.hfg5.6xlarge	24	96	4.5	2,000,000	6	8	20
ecs.hfg5.8xlarge	32	128	6	2,500,000	8	8	20
ecs.hfg5.14xlarge	56	160	10	4,000,000	14	8	20

Note

-
- For more information about these specifications, see Instance family.

Enterprise-level computing instance families based on the ARM architecture

Enterprise-level heterogeneous computing instance families

f3, FPGA-accelerated compute optimized instance family

Features

- Uses Xilinx 16nm Virtex UltraScale+ VU9P FPGAs.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.

- Network:
  - Provides high network performance based on large computing capacity.
- Suits the following scenarios:
  - Deep learning and inference
  - Genomics research
  - Database acceleration
  - Image transcoding such as conversion of JPEG images to WebP images
  - Real-time video processing such as H.265 video compression

Instance types

Instance type	vCPUs	Memory (GiB)	FPGAs	Bandwidth (bidirectional), Gbit/s	Packet forwarding rate (bidirectional), Kpps	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.f3-c4f1.xlarge	4	16.0	1 × Xilinx VU9P	1.5	300	2	3	10
ecs.f3-c8f1.2xlarge	8	32.0	1 × Xilinx VU9P	2.5	500	4	4	10
ecs.f3-c16f1.4xlarge	16	64.0	1 × Xilinx VU9P	5.0	1,000	4	8	20
ecs.f3-c16f1.8xlarge	32	128.0	2 × Xilinx VU9P	10.0	2,000	8	8	20
ecs.f3-c16f1.16xlarge	64	256.0	4 × Xilinx VU9P	20.0	2,500	16	8	20
ecs.f3-c22f1.22xlarge	88	336.0	4 × Xilinx VU9P	30.0	4,500	16	8	20

Note

-
- For more information about these specifications, see Instance family.

f1, FPGA-accelerated compute optimized instance family

Features

- Uses Intel® Arria® 10 GX 1150 FPGAs
- Compute:
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
  - Offers a CPU-to-memory ratio of 1:7.5.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Suits the following scenarios:
  - Deep learning and inference
  - Genomics research
  - Financial analysis
  - Image transcoding
  - Computational workloads such as real-time video processing and security management

Instance types

Instance type	vCPUs	Memory (GiB)	FPGAs	Bandwidth (bidirectional), Gbit/s	Packet forwarding rate (bidirectional), Kpps	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.f1-c8f1.2xlarge	8	60.0	Intel ARRIA 10 GX 1150	3.0	400	4	4	10
ecs.f1-c8f1.4xlarge	16	120.0	2 × Intel ARRIA 10 GX 1150	5.0	1,000	4	8	20
ecs.f1-c28f1.7xlarge	28	112.0	Intel ARRIA 10 GX 1150	5.0	2,000	8	8	20
ecs.f1-c28f1.14xlarge	56	224.0	2 × Intel ARRIA 10 GX 1150	10.0	2,000	14	8	20

Note

-
- For more information about these specifications, see Instance family.

ECS Bare Metal Instance families and Super Computing Cluster (SCC) instance families

scchfc6, compute-optimized SCC instance family with high clock speeds

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2.4.
  - Uses 3.1 GHz Intel® Xeon® Platinum 8269 (Cascade Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports enhanced SSDs (ESSDs), standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Supports both RoCE networks and VPCs. RoCE networks are dedicated to RDMA communication.
- Supported scenarios:
  - Large-scale machine learning training
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Physical cores	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.scchfc6.20xlarge	80	40	192.0	30	6,000,000	50	8	32	10

> **Note**
> - ecs.scchfc6.20xlarge provides 80 logical processors on 40 physical cores.
> -
> - For more information about these specifications, see Instance family.

scchfg6, general-purpose SCC instance family with high clock speeds

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.

- Compute:
  - Offers a CPU-to-memory ratio of 1:4.8.
  - Uses 3.1 GHz Intel® Xeon® Platinum 8269 (Cascade Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Supports both RoCE networks and VPCs. RoCE networks are dedicated to RDMA communication.
- Supported scenarios:
  - Large-scale machine learning training
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Physical cores	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.scchfg6.20xlarge	80	40	384.0	30	6,000,000	50	8	32	10

> **Note**
> - ecs.scchfg6.20xlarge provides 80 logical processors on 40 physical cores.
> -
> - For more information about these specifications, see Instance family.

scchfr6, memory-optimized SCC instance family with high clock speeds

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Offers a CPU-to-memory ratio of 1:9.6.
  - Uses 3.1 GHz Intel® Xeon® Platinum 8269 (Cascade Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Storage:
  - Is an instance family in which all instances are I/O optimized.

- Supports ESSDs, standard SSDs, and ultra disks.

- Network:
  - Supports IPv6.
  - Supports both RoCE networks and VPCs. RoCE networks are dedicated to RDMA communication.
- Supported scenarios:
  - Large-scale machine learning training
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Physical cores	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.scchfr6.20xlarge	80	40	768.0	30	6,000,000	50	8	32	10

> **Note**
> - ecs.scchfr6.20xlarge provides 80 logical processors on 40 physical cores.
> -
> - For more information about these specifications, see Instance family.

scch5, SCC instance family with high clock speeds

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Offers a CPU-to-memory ratio of 1:3.
  - Uses 3.1 GHz Intel® Xeon® Gold 6149 (Skylake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports standard SSDs and ultra disks.
- Network:
  - Supports both RoCE networks and VPCs. RoCE networks are dedicated to RDMA communication.
- Supported scenarios:
  - Large-scale machine learning training
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Physical cores	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.scch5.16xlarge	64	32	192.0	10	4,500,000	50	8	32	10

> **Note**
>
> - ecs.scch5.16xlarge provides 64 logical processors on 32 physical cores.
> -
> - For more information about these specifications, see Instance family.

sccg5, general-purpose SCC instance family

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports standard SSDs and ultra disks.
- Network:
  - Supports both RoCE networks and VPCs. RoCE networks are dedicated to RDMA communication.
- Supported scenarios:
  - Large-scale machine learning training
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Physical cores	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.sccg5.24xlarge	96	48	384.0	10	4,500,000	50	8	32	10

Note

- ecs.sccg5.24xlarge provides 96 logical processors on 48 physical cores.
-
- For more information about these specifications, see Instance family.

sccgn6, GPU-accelerated compute-optimized SCC instance family

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Uses NVIDIA V100 GPUs (SXM2-based) that have the following features:
    - Innovative Volta architecture
    - Up to 16 GB HBM2 GPU memory
    - 5,120 CUDA cores
    - 640 Tensor cores
    - GPU memory bandwidth of up to 900 GB/s
    - Support for six NVLink links and a total bandwidth of 300 GB/s (25 GB/s per NVlink link per direction)
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Supports high-performance CPFS.
- Network:
  - Supports IPv6.
  - Supports VPCs.
  - Supports RoCE v2 networks, which are dedicated to low-latency RDMA communication.
- Supported scenarios:
  - Ultra-large-scale training for machine learning on a distributed GPU cluster

- Large-scale high-performance scientific computing and simulation calculation
- Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.sccgn6.24xlarge	96	384.0	NVIDIA V100 × 8	30	4,500,000	50	8	32	10

> **Note**
>
> -
> - For more information about these specifications, see Instance family.

sccgn7ex, GPU-accelerated compute-optimized SCC instance family

Features

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Uses eight NVIDIA A100 GPUs per instance that support NVSwitch and deliver up to 312 TFLOPS of TensorFloat-32 (TF32) computing power.
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses third-generation Intel® Xeon® Scalable (Ice Lake) processors that deliver a base frequency of 2.9 GHz and an all-core turbo frequency of 3.5 GHz and support PCIe 4.0 interfaces.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs. ESSDs at performance level (PL) 3 can deliver a maximum of 1,000,000 IOPS and 4,000 MB/s of throughput, which can meet the cache requirements of training and eliminate the need for local disks.
- Network:
  - Supports IPv6.
  - Supports only virtual private clouds (VPCs).
  - Provides a bandwidth of 800 Gbit/s between sccgn7ex instances (100 GBit/s per port on each of 4 dual-port RDMA network controller) to support GPUDirect. Each GPU is directly connected to a port with the bandwidth of 100 GBit/s on RDMA network interface controllers.
- Supported scenarios: ultra-large-scale training for artificial intelligence.

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues (Primary ENI/Secondary ENI)	ENIs	Private IP addresses per ENI
ecs.sccgn7ex.32xlarge	128	1024	NVIDIA A100 × 8	80 GB × 8	64	24,000,000	800	32/12	32	15

Note

-
- For more information about these specifications, see Instance family.

sccgn6e, GPU-accelerated compute-optimized SCC instance family

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Uses NVIDIA V100 GPUs (SXM2-based) that have the following features:
    - Innovative Volta architecture
    - 32 GB HBM2 GPU memory
    - 5,120 CUDA cores
    - 640 Tensor cores
    - GPU memory bandwidth of up to 900 GB/s
    - Support for six NVLink links and a total bandwidth of 300 GB/s (25 GB/s per NVlink link per direction)
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Supports high-performance Cloud Paralleled File System (CPFS).
- Network:
  - Supports IPv6.
  - Supports VPCs.
  - Supports RoCE v2 networks, which are dedicated to low-latency RDMA communication.

- Supported scenarios:
  - Ultra-large-scale training for machine learning on a distributed GPU cluster
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.scgn6e.24xlarge	96	768.0	NVIDIA V100 × 8	32 GB × 8	32	4,800,000	50	8	32	10

> **Note**
>
> -
> - For more information about these specifications, see Instance family.

Shared computing instance families based on the x86 architecture

t6, burstable instance family

Features:

- Provides baseline CPU performance and is burstable but limited by accrued CPU credits.
- Is more cost-effective when compared with the t5 burstable instance family.
- Compute:
  - Uses 2.5 GHz Intel® Xeon® Cascade Lake processors that deliver a turbo frequency of 3.2 GHz.
  - Uses DDR4 memory.
- Storage:
  - Supports enhanced SSDs (ESSDs), standard SSDs, and ultra disks.

    > **Note** ESSDs at performance level (PL) 2 and 3 cannot provide maximum performance due to the specification limits of burstable instances. We recommend that you use enterprise-level instances or ESSDs that are at lower performance levels.
- Network:
  - Supports IPv6.
  - Supports only virtual private clouds (VPCs).
  - Delivers a bandwidth up to 4 Gbit/s.
- Supported scenarios:

- Web application servers
- Lightweight applications and microservices
- Development and testing environments

Instance types

Instance type	vCPU	Memory (GiB)	Baseline CPU performance	CPU credits per hour	Max CPU credit balance	Base bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.t6-c4m1.large	2	0.5	5%	6	144	0.08	40,000	1	2	2
ecs.t6-c2m1.large	2	1.0	10%	12	288	0.08	60,000	1	2	2
ecs.t6-c1m1.large	2	2.0	20%	24	576	0.08	100,000	1	2	2
ecs.t6-c1m2.large	2	4.0	20%	24	576	0.08	100,000	1	2	2
ecs.t6-c1m4.large	2	8.0	30%	36	864	0.08	100,000	1	2	2
ecs.t6-c1m4.xlarge	4	16.0	40%	96	2304	0.16	200,000	1	2	6
ecs.t6-c1m4.2xlarge	8	32.0	40%	192	4608	0.32	400,000	1	2	6

> **Note**
>
> - Secondary elastic network interfaces (ENIs) cannot be bound to instances of this instance family while the instances are being created and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from instances of the following instance types, the instances must be in the Stopped state: ecs.t6-c1m1.large, ecs.t6-c1m2.large, ecs.t6-c1m4.large, ecs.t6-c2m1.large, and ecs.t6-c4m1.large.
> -
> - For more information about these specifications, see Instance family.

t5, burstable instance family

Features:

- Provides baseline CPU performance and is burstable but limited by accrued CPU credits.
- Offers a balance between compute, memory, and network resources.
- Compute:
  - Offers multiple CPU-to-memory ratios.
  - Uses 2.5 GHz Intel® Xeon® processors.
  - Uses DDR4 memory.
- Network:
  - Supports IPv6.
  - Supports only VPCs.
- Supported scenarios:
  - Web application servers
  - Lightweight applications and microservices
  - Development and testing environments

Instance types

Instance type	vCPU	Memory (GiB)	Baseline CPU performance	CPU credits per hour	Max CPU credit balance	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.t5-lc2m1.nano	1	0.5	20%	12	288	0.1	40,000	1	2	2
ecs.t5-lc1m1.small	1	1.0	20%	12	288	0.2	60,000	1	2	2

Instance type	vCPU	Memory (GiB)	Baseline CPU performance	CPU credits per hour	Max CPU credit balance	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.t5-lc1m2.small	1	2.0	20%	12	288	0.2	60,000	1	2	2
ecs.t5-lc1m2.large	2	4.0	20%	24	576	0.4	100,000	1	2	2
ecs.t5-lc1m4.large	2	8.0	20%	24	576	0.4	100,000	1	2	2
ecs.t5-c1m1.large	2	2.0	25%	30	720	0.5	100,000	1	2	2
ecs.t5-c1m2.large	2	4.0	25%	30	720	0.5	100,000	1	2	2
ecs.t5-c1m4.large	2	8.0	25%	30	720	0.5	100,000	1	2	2
ecs.t5-c1m1.xlarge	4	4.0	25%	60	1440	0.8	200,000	1	2	6
ecs.t5-c1m2.xlarge	4	8.0	25%	60	1440	0.8	200,000	1	2	6
ecs.t5-c1m4.xlarge	4	16.0	25%	60	1440	0.8	200,000	1	2	6

Instance type	vCPU	Memory (GiB)	Baseline CPU performance	CPU credits per hour	Max CPU credit balance	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.t5-c1m1.2xlarge	8	8.0	25%	120	2880	1.2	400,000	1	2	6
ecs.t5-c1m2.2xlarge	8	16.0	25%	120	2880	1.2	400,000	1	2	6
ecs.t5-c1m4.2xlarge	8	32.0	25%	120	2880	1.2	400,000	1	2	6
ecs.t5-c1m1.4xlarge	16	16.0	25%	240	5760	1.2	600,000	1	2	6
ecs.t5-c1m2.4xlarge	16	32.0	25%	240	5760	1.2	600,000	1	2	6

Note

- Secondary ENIs cannot be bound to instances of this instance family while the instances are being created and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from instances of the following instance types, the instances must be in the Stopped state: ecs.t5-lc2m1.nano, ecs.t5-c1m1.large, ecs.t5-c1m2.large, ecs.t5-c1m4.large, ecs.t5-lc1m1.small, ecs.t5-lc1m2.large, ecs.t5-lc1m2.small, and ecs.t5-lc1m4.large.
-
- For more information about these specifications, see Instance family.

Previous-generation shared instance families xn4, n4, mn4, and e4

Features

- Offers multiple CPU-to-memory ratios.

- Uses 2.5 GHz Intel® Xeon® processors.
- Uses DDR4 memory.

Instance family	Description	vCPU-to-memory ratio	Scenario
xn4	Shared compact instance family	1:1	• Frontend web applications • Lightweight applications and microservices • Development and testing environments
n4	Shared compute instance family	1:2	• Websites and web applications • Development environments, servers, code repositories, microservices, and testing and staging environments • Lightweight enterprise applications
mn4	Shared general-purpose instance family	1:4	• Websites and web applications • Lightweight databases and caches • Integrated applications and lightweight enterprise services
e4	Shared memory instance family	1:8	• Applications that require a large memory • Lightweight databases and caches

xn4

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.xn4.small	1	1.0	0.5	50,000	1	2	2

Note

- Secondary elastic network interfaces (ENIs) cannot be bound to instances of this instance family while the instances are being created, and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from an ecs.xn4.small instance, the instance must be in the Stopped state.
-
- For more information about these specifications, see Instance family.

n4

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.n4.small	1	2.0	0.5	50,000	1	2	2
ecs.n4.large	2	4.0	0.5	100,000	1	2	2
ecs.n4.xlarge	4	8.0	0.8	150,000	1	2	6
ecs.n4.2xlarge	8	16.0	1.2	300,000	1	2	6
ecs.n4.4xlarge	16	32.0	2.5	400,000	1	2	6
ecs.n4.8xlarge	32	64.0	5.0	500,000	1	2	6

Note

- Secondary ENIs cannot be bound to instances of this instance family while the instances are being created, and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from an ecs.n4.small or ecs.n4.large instance, the instance must be in the Stopped state.
-
- For more information about these specifications, see Instance family.

mn4

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.mn4.small	1	4.0	0.5	50,000	1	2	2
ecs.mn4.large	2	8.0	0.5	100,000	1	2	2
ecs.mn4.xlarge	4	16.0	0.8	150,000	1	2	6
ecs.mn4.2xlarge	8	32.0	1.2	300,000	1	2	6
ecs.mn4.4xlarge	16	64.0	2.5	400,000	1	2	6
ecs.mn4.8xlarge	32	128.0	5	500,000	2	8	6

Note

- Secondary ENIs cannot be bound to instances of this instance family while the instances are being created, and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from an ecs.mn4.small or ecs.mn4.large instance, the instance must be in the Stopped state.
-
- For more information about these specifications, see Instance family.

e4

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.e4.small	1	8.0	0.5	50,000	1	2	2
ecs.e4.large	2	16.0	0.5	100,000	1	2	2
ecs.e4.xlarge	4	32.0	0.8	150,000	1	2	6
ecs.e4.2xlarge	8	64.0	1.2	300,000	1	3	6

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.e4.4xlarge	16	128.0	2.5	400,000	1	8	6

Note

- Secondary ENIs cannot be bound to instances of this instance family while the instances are being created, and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from an ecs.e4.small or ecs.e4.large instance, the instance must be in the Stopped state.
-
- For more information about these specifications, see Instance family.

Instance type specifications

Specification	Description
Local storage	Local storage (also called local disks or cache disks) refers to the disks attached to the physical servers on which ECS instances are hosted. Local storage provides temporary block storage for instances. Local storage capacity is measured in GiB. When the computing resources (vCPUs and memory) of an instance are released or when an instance is failed over, data stored on its local disks may be lost. For more information, see Local disks.
Bandwidth	The maximum sum of inbound and outbound bandwidth values. For information about how to test the network bandwidth of an instance, see Test the network bandwidth. **Note** Instance type specifications are all verified and obtained within a test environment. In actual scenarios, the performance of an instance may vary based on other factors such as instance load and networking model. We recommend that you perform business stress tests on instances to choose appropriate instance types.
Packet forwarding rate	The maximum sum of inbound and outbound packet forwarding rates. For information about how to test the packet forwarding rate of an instance, see Best practices for testing network performance. **Note** Instance type specifications are all verified and obtained within a test environment. In actual scenarios, the performance of an instance may vary based on other factors such as instance load, image version, and networking model. We recommend that you perform business stress tests on instances to choose appropriate instance types.

Specification	Description
Connections	A connection (also called a session) is the process of connecting a client and a server and transferring data between them. A connection is uniquely defined by the network communication quintuple that consists of a source IP address, a destination IP address, a source port, a destination port, and a protocol. Connections of an ECS instance include TCP, UDP, and Internet Control Message Protocol (ICMP) connections.
NIC queues	The maximum number of network interface controller (NIC) queues supported by the primary NIC of an instance. For the instance types other than ECS Bare Metal Instance types, the maximum number of NIC queues supported by a secondary NIC is the same as that supported by the primary NIC.
ENI	The number of elastic network interfaces (ENIs) per instance that include one primary ENI.

4.Instance type families

4.1. General-purpose instance families

This topic describes the features of general-purpose instance families of Elastic Compute Service (ECS) and lists the instance types of each instance family.

- Recommended instance families
  - g8m, general-purpose instance family
  - g7se, storage-enhanced general-purpose instance family
  - g7a, general-purpose instance family
  - g7, general-purpose instance family
  - g7t, security-enhanced general-purpose instance family
  - g7ne, network-enhanced general-purpose instance family
  - g6a, general-purpose instance family
  - g6t, security-enhanced general-purpose instance family
  - g6e, general-purpose instance family with enhanced performance
  - g6, general-purpose instance family
  - g5, general-purpose instance family
  - g5ne, network-enhanced general-purpose instance family
- Other available instance families (If these instance families are sold out, you can use the recommended ones.)

  sn2ne, network-enhanced general-purpose instance family

g8m, general-purpose instance family

This instance family is in invitational preview. To use this instance family, go to the g8m Instance Free Trial Application Form page to submit an application.

The g8m general-purpose instance family is the first Alibaba Cloud instance family that uses the proprietary Yitian 710 CPU. This instance family is used for general-purpose computing, cloud native, and Android in Cloud scenarios.

Features:

- This instance family uses third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance. This instance family utilizes fast path acceleration on chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.75 GHz Yitian 710 processors to provide consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.

- Supports only enhanced SSDs (ESSDs).
- Provides high storage I/O performance based on large computing capacity.

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides burstable bandwidth capabilities for low-specification instances.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Containers and microservices
  - Scenarios where applications such as DevOps applications are developed and tested
  - Websites and application servers
  - Game servers
  - Other general-purpose enterprise-level applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	Network interface controller (NIC) queues	Elastic network interfaces (ENIs)	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.g8m.small	1	4	1/10	500,000	Up to 250,000	1	2	3	10,000/burstable up to 110,000	1/burstable up to 6
ecs.g8m.large	2	8	2/10	900,000	Up to 250,000	2	3	6	20,000/burstable up to 110,000	1.5/burstable up to 6
ecs.g8m.xlarge	4	16	3/10	1,000,000	Up to 250,000	4	4	15	40,000/burstable up to 110,000	2/burstable up to 6

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	Network interface controller (NIC) queues	Elastic network interfaces (ENIs)	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.g8m.2xlarge	8	32	5/10	1,600,000	Up to 250,000	8	4	15	50,000/burstable up to 110,000	3/burstable up to 6
ecs.g8m.4xlarge	16	64	10/16	3,000,000	300,000	8	8	30	80,000/burstable up to 110,000	5/burstable up to 6
ecs.g8m.8xlarge	32	128	16/none	6,000,000	600,000	16	8	30	150,000	8
ecs.g8m.16xlarge	64	256	32/none	12,000,000	1,200,000	32	8	30	300,000	16
ecs.g8m.32xlarge	128	512	64/none	24,000,000	2,400,000	32	15	30	600,000	32

Note

-
- For more information about these specifications, see Instance family.
- You can change instance types within the g8m instance family or between the g6r and g8m instance families.

g7se, storage-enhanced general-purpose instance family

Features:

- This instance family uses third-generation SHENLONG architecture and Intel Ice Lake processors to improve storage I/O performance.
- This instance family delivers a sequential read/write throughput of up to 64 Gbit/s and a maximum IOPS of 1,000,000 per instance.

- This instance family allows ESSDs to be attached based on the Non-Volatile Memory Express (NVMe) protocol and supports the multi-attach feature to meet the requirements of core enterprise business. For more information, see Enable multi-attach.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses third-generation Intel® Xeon® Scalable (Ice Lake) processors that deliver a base frequency of 2.7 GHz and an all-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs and provides ultra-high I/O performance.
  - Allows a maximum of 64 data disks to be attached to a single instance. You can attach a maximum of 16 data disks to an instance when you create the instance. If the instance requires more data disks, attach more data disks after the instance is created. For more information, see Attach a data disk.
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - I/O-intensive scenarios such as large and medium-sized online transactional processing (OLTP) core databases
  - Large and medium-sized NoSQL databases
  - Search and real-time log analytics
  - Traditional large enterprise-level commercial software such as SAP

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Maximum attached data disks	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.g7se.large	2	8	1.2/burstable up to 3	450,000	Up to 250,000	2	3	6	16	30,000/burstable up to 150,000	3/10
ecs.g7se.xlarge	4	16	2/burstable up to 5	500,000	Up to 250,000	4	4	15	16	60,000/burstable up to 150,000	4/10
ecs.g7se.2xlarge	8	32	3/burstable up to 8	800,000	Up to 250,000	8	4	15	16	100,000/burstable up to 150,000	6/10
ecs.g7se.3xlarge	12	48	4.5/burstable up to 10	1,200,000	Up to 250,000	8	8	15	16	120,000/burstable up to 150,000	8/10
ecs.g7se.4xlarge	16	64	6/burstable up to 10	1,500,000	300,000	8	8	30	24	150,000/none	10/none
ecs.g7se.6xlarge	24	96	8/burstable up to 10	2,250,000	450,000	12	8	30	24	200,000/none	12/none

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Maximum attached data disks	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.g7se.8xlarge	32	128	10/none	3,000,000	600,000	16	8	30	30	300,000/none	16/none
ecs.g7se.16xlarge	64	256	16/none	6,000,000	1,200,000	32	8	30	56	500,000/none	32/none
ecs.g7se.32xlarge	128	512	32/none	12,000,000	2,400,000	32	15	30	64	1,000,000/none	64/none

Note

-
- For more information about these specifications, see Instance family.

g7a, general-purpose instance family

Features:

- This instance family uses third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance. This instance family utilizes fast path acceleration on chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.55 GHz AMD EPYC™ MILAN processors that deliver a maximum single-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides disk burstable IOPS and bandwidth capabilities for low-specification instances.

- Provides high storage I/O performance based on large computing capacity.

  Note

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides burstable bandwidth capabilities for low-specification instances.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Video encoding and decoding
  - Scenarios where large volumes of packets are received and transmitted
  - Websites and application servers
  - Small and medium-sized database systems, caches, and search clusters
  - Game servers
  - Scenarios where applications such as DevOps applications are developed and tested
  - Other general-purpose enterprise-level applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.g7a.large	2	8	1/burstable up to 10	900,000	Up to 250,000	2	3	6	12,500/burstable up to 110,000	1/burstable up to 6
ecs.g7a.xlarge	4	16	1.5/burstable up to 10	1,000,000	Up to 250,000	4	4	15	20,000/burstable up to 110,000	1.5/burstable up to 6

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.g7a.2xlarge	8	32	2.5/burstable up to 10	1,600,000	Up to 250,000	8	4	15	30,000/burstable up to 110,000	2/burstable up to 6
ecs.g7a.4xlarge	16	64	5/burstable up to 10	2,000,000	300,000	8	8	30	60,000/burstable up to 110,000	3/burstable up to 6
ecs.g7a.8xlarge	32	128	8/burstable up to 10	3,000,000	600,000	16	7	30	75,000/burstable up to 110,000	4/burstable up to 6
ecs.g7a.16xlarge	64	256	16/none	6,000,000	1,000,000	32	7	30	150,000/none	8/none
ecs.g7a-nps1.16xlarge	64	256	16/none	6,000,000	1,000,000	32	7	30	150,000/none	8/none
ecs.g7a.32xlarge	128	512	32/none	12,000,000	2,000,000	32	15	30	300,000/none	16/none

> **Note**
>
> -
> - For more information about these specifications, see Instance family.
> - Ubuntu 16 and Debian 9 operating system kernels do not support AMD EPYC™ MILAN processors. Do not use Ubuntu 16 or Debian 9 images to create instances of this instance family. Instances of this instance family created from Ubuntu 16 or Debian 9 images cannot start.

g7, general-purpose instance family

Features:

- This instance family uses third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance. This instance family utilizes fast path acceleration on chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- This instance family supports the virtual Trusted Platform Module (vTPM) feature and implements trusted boot based on Trusted Cryptography Module (TCM) or Trusted Platform Module (TPM) chips to provide ultra-high security capabilities. During a trusted boot, all modules in the boot chain from the underlying server to the ECS instance are measured and verified.
- This instance family supports the Enclave feature and provides a virtualization-based confidential computing environment. For more information, see Build a confidential computing environment by using Enclave.

  > **Note** The Enclave feature is in invitational preview. If you want to use this feature, go to the Enclave product page.

- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses third-generation Intel® Xeon® Scalable (Ice Lake) processors that deliver a base frequency of 2.7 GHz and an all-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides burstable storage I/O performance for low-specification instances.
  - Provides high storage I/O performance based on large computing capacity.
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides burstable network performance for low-specification instances.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:

- Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
- Game servers
- Small and medium-sized database systems, caches, and search clusters
- Enterprise-level applications of various types and sizes
- Websites and application servers
- Data analytics and computing
- Scenarios that require secure and trusted computing
- Blockchain scenarios

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.g7.large	2	8	2/burstable up to 10	900,000	Yes	Up to 250,000	2	3	6	20,000/burstable up to 110,000	1.5/burstable up to 6
ecs.g7.xlarge	4	16	3/burstable up to 10	1,000,000	Yes	Up to 250,000	4	4	15	40,000/burstable up to 110,000	2/burstable up to 6
ecs.g7.2xlarge	8	32	5/burstable up to 10	1,600,000	Yes	Up to 250,000	8	4	15	50,000/burstable up to 110,000	3/burstable up to 6
ecs.g7.3xlarge	12	48	8/burstable up to 10	2,400,000	Yes	Up to 250,000	8	8	15	70,000/burstable up to 110,000	4/burstable up to 6

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.g7.4xlarge	16	64	10/burstable up to 25	3,000,000	Yes	300,000	8	8	30	80,000/burstable up to 110,000	5/burstable up to 6
ecs.g7.6xlarge	24	96	12/burstable up to 25	4,500,000	Yes	450,000	12	8	30	110,000/none	6/none
ecs.g7.8xlarge	32	128	16/burstable up to 25	6,000,000	Yes	600,000	16	8	30	150,000/none	8/none
ecs.g7.16xlarge	64	256	32/none	12,000,000	Yes	1,200,000	32	8	30	300,000/none	16/none
ecs.g7.32xlarge	128	512	64/none	24,000,000	Yes	2,400,000	32	15	30	600,000/none	32/none

Note

-
-
- For more information about these specifications, see Instance family.

g7t, security-enhanced general-purpose instance family

Features:

- This instance family supports up to 256 GiB of encrypted memory and confidential computing based on Intel® Software Guard Extensions (SGX) to protect the confidentiality and integrity of essential code and data from malware attacks.
- This instance family supports Virtual SGX (vSGX) and allows you to select instance types that suit your needs.

  Notice

- This instance family implements trusted boot based on TCM or TPM chips. During a trusted boot, all modules in the boot chain from the underlying server to the guest operating system are measured and verified.
- This instance family offloads a large number of virtualization features to dedicated hardware with the use of third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1: 4. About 50% of memory is encrypted.
  - Uses third-generation Intel® Xeon® Scalable (Ice Lake) processors that deliver a base frequency of 2.7 GHz and an all-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides high storage I/O performance based on large computing capacity.
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios that involve sensitive information such as personal identity information, healthcare information, financial information, and intellectual property data
  - Scenarios where confidential data is shared among multiple parties
  - Blockchain scenarios
  - Confidential machine learning
  - Scenarios that require high security and enhanced trust, such as services for financial organizations, public service sectors, and enterprises
  - Enterprise-level applications of various types and sizes

Instance types

Instance type	vCPUs	Memory (GiB)	Encrypted memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.g7t.large	2	8	4	2/burstable up to 10	900,000	Yes	Up to 250,000	2	3	6	20,000/burstable up to 110,000	1.5/burstable up to 6
ecs.g7t.xlarge	4	16	8	3/burstable up to 10	1,000,000	Yes	Up to 250,000	4	4	15	40,000/burstable up to 110,000	2/burstable up to 6
ecs.g7t.2xlarge	8	32	16	5/burstable up to 10	1,600,000	Yes	Up to 250,000	8	4	15	50,000/burstable up to 110,000	3/burstable up to 6
ecs.g7t.3xlarge	12	48	24	8/burstable up to 10	2,400,000	Yes	Up to 250,000	8	8	15	70,000/burstable up to 110,000	4/burstable up to 6
ecs.g7t.4xlarge	16	64	32	10/burstable up to 25	3,000,000	Yes	300,000	8	8	30	80,000/burstable up to 110,000	5/burstable up to 6

Instance type	vCPUs	Memory (GiB)	Encrypted memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.g7t.6xlarge	24	96	48	12/burstable up to 25	4,500,000	Yes	450,000	12	8	30	110,000/none	6/none
ecs.g7t.8xlarge	32	128	64	16/burstable up to 25	6,000,000	Yes	600,000	16	8	30	150,000/none	8/none
ecs.g7t.16xlarge	64	256	128	32/none	12,000,000	Yes	1,200,000	32	8	30	300,000/none	16/none
ecs.g7t.32xlarge	128	512	256	64/none	24,000,000	Yes	2,400,000	32	15	30	600,000/none	32/none

Note

- For more information about these specifications, see Instance family.

g7ne, network-enhanced general-purpose instance family

Features:

- This instance family significantly improves the network throughput and packet forwarding rate per instance. A single instance can deliver a packet forwarding rate of up to 24,000,000 pps.

- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses Intel® Xeon® Platinum 8369HB (Cooper Lake) or Intel® Xeon® Platinum 8369HC (Cooper Lake) processors that deliver a turbo frequency of 3.8 GHz and a minimum clock speed of 3.3 GHz to provide consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs and provides ultra-high I/O performance.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Network-intensive scenarios such as Network Functions Virtualization (NFV) or Software-defined Wide Area Network (SD-WAN), mobile Internet, on-screen video comments, and telecom data forwarding
  - Small and medium-sized database systems, caches, and search clusters
  - Enterprise-level applications of various types and sizes
  - Big data analysis and machine learning

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.g7ne.large	2	8	1.5/10	900,000	450,000	2	3	10	10,000	0.75
ecs.g7ne.xlarge	4	16	3/10	1,000,000	900,000	4	4	15	20,000	1
ecs.g7ne.2xlarge	8	32	6/15	1,500,000	1,750,000	8	6	15	25,000	1.2
ecs.g7ne.4xlarge	16	64	12/25	3,000,000	3,500,000	16	8	30	40,000	2
ecs.g7ne.8xlarge	32	128	25/none	6,000,000	6,000,000	16	8	30	75,000	5

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.g7ne.12xlarge	48	192	40/none	12,000,000	8,000,000	32	8	30	100,000	8
ecs.g7ne.24xlarge	96	384	80/none	24,000,000	16,000,000	32	15	50	240,000	16

Note

-
- For more information about these specifications, see Instance family.

g6a, general-purpose instance family

Features:

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.6 GHz AMD EPYC™ ROME processors that deliver a turbo frequency of 3.3 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.

- Provides high network performance based on large computing capacity.

- Supported scenarios:
  - Video encoding and decoding
  - Scenarios where large volumes of packets are received and transmitted
  - Websites and application servers
  - Small and medium-sized database systems, caches, and search clusters
  - Game servers
  - Scenarios where applications such as DevOps applications are developed and tested
  - Other general-purpose enterprise-level applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.g6a.large	2	8	1/10	900,000	Up to 250,000	2	2	6	12,500	1
ecs.g6a.xlarge	4	16	1.5/10	1,000,000	Up to 250,000	4	3	15	20,000	1.5
ecs.g6a.2xlarge	8	32	2.5/10	1,600,000	Up to 250,000	8	4	15	30,000	2
ecs.g6a.4xlarge	16	64	5/10	2,000,000	300,000	8	8	30	60,000	3
ecs.g6a.8xlarge	32	128	8/10	3,000,000	600,000	16	7	30	75,000	4
ecs.g6a.16xlarge	64	256	16/none	6,000,000	1,000,000	32	8	30	150,000	8
ecs.g6a.32xlarge	128	512	32/none	12,000,000	2,000,000	32	15	30	300,000	16

Note

-
- For more information about these specifications, see Instance family.

g6t, security-enhanced general-purpose instance family

Features:

- This instance family implements trusted boot based on TCM or TPM chips. During a trusted boot, all modules in the boot chain from the underlying server to the guest operating system are measured and verified.
- This instance family supports vTPMs and delivers a full set of trusted capabilities at the IaaS layer based on integrity monitoring.
- This instance family offloads a large number of virtualization features to dedicated hardware with the use of third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads. This instance family utilizes fast path acceleration on chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269 (Cascade Lake) processors that deliver a turbo frequency of 3.2 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios that require high security and enhanced trust, such as services for financial organizations, public service sectors, and enterprises
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Enterprise-level applications of various types and sizes
  - Websites and application servers

- Game servers
- Small and medium-sized database systems, caches, and search clusters
- Data analytics and computing
- Computing clusters and memory-intensive data processing

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk baseline bandwidth (Gbit/s)
ecs.g6t.large	2	8	1.2/burstable up to 10	900,000	Yes	Up to 250,000	2	3	6	20,000	1
ecs.g6t.xlarge	4	16	2/burstable up to 10	1,000,000	Yes	Up to 250,000	4	4	15	40,000	1.5
ecs.g6t.2xlarge	8	32	3/burstable up to 10	1,600,000	Yes	Up to 250,000	8	4	15	50,000	2
ecs.g6t.4xlarge	16	64	6/burstable up to 10	3,000,000	Yes	300,000	8	8	30	80,000	3
ecs.g6t.8xlarge	32	128	10/none	6,000,000	Yes	600,000	16	8	30	150,000	5
ecs.g6t.13xlarge	52	192	16/none	9,000,000	Yes	900,000	32	7	30	240,000	8
ecs.g6t.26xlarge	104	384	32/none	24,000,000	Yes	1,800,000	32	15	30	480,000	16

Note

-
- For more information about these specifications, see Instance family.
- The results for network capabilities are the maximum values obtained from single-item tests. For example, when network bandwidth is tested, no stress tests are performed on the packet forwarding rate or other network metrics.

g6e, general-purpose instance family with enhanced performance

Features:

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads. This instance family utilizes fast path acceleration on chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269 (Cascade) processors that deliver a turbo frequency of 3.2 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.

    Note
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Enterprise-level applications of various types and sizes
  - Websites and application servers
  - Game servers

- Small and medium-sized database systems, caches, and search clusters
- Data analytics and computing
- Computing clusters and memory-intensive data processing

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.g6e.large	2	8	1.2/burstable up to 10	900,000	Up to 250,000	2	3	6	20,000	1
ecs.g6e.xlarge	4	16	2/burstable up to 10	1,000,000	Up to 250,000	4	4	15	40,000	1.5
ecs.g6e.2xlarge	8	32	3/burstable up to 10	1,600,000	Up to 250,000	8	4	15	50,000	2
ecs.g6e.4xlarge	16	64	6/burstable up to 10	3,000,000	300,000	8	8	30	80,000	3
ecs.g6e.8xlarge	32	128	10/none	6,000,000	600,000	16	8	30	150,000	5
ecs.g6e.13xlarge	52	192	16/none	9,000,000	1,000,000	32	7	30	240,000	8
ecs.g6e.26xlarge	104	384	32/none	24,000,000	1,800,000	32	15	30	480,000	16

Note

-
- For more information about these specifications, see Instance family.
- The results for network capabilities are the maximum values obtained from single-item tests. For example, when network bandwidth is tested, no stress tests are performed on the packet forwarding rate or other network metrics.

g6, general-purpose instance family

Features:

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.2 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.

    Note
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.

    Note
  - Provides high network performance based on large computing capacity.
- Supports changes to instance types in the c6 or r6 instance family.
- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Enterprise-level applications of various types and sizes

- Websites and application servers
- Game servers
- Small and medium-sized database systems, caches, and search clusters
- Data analytics and computing
- Computing clusters and memory-intensive data processing

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.g6.large	2	8	1/3	300,000	Up to 250,000	2	2	6	10,000	1
ecs.g6.xlarge	4	16	1.5/5	500,000	Up to 250,000	4	3	10	20,000	1.5
ecs.g6.2xlarge	8	32	2.5/8	800,000	Up to 250,000	8	4	10	25,000	2
ecs.g6.3xlarge	12	48	4/10	900,000	Up to 250,000	8	6	10	30,000	2.5
ecs.g6.4xlarge	16	64	5/10	1,000,000	300,000	8	8	20	40,000	3
ecs.g6.6xlarge	24	96	7.5/10	1,500,000	450,000	12	8	20	50,000	4
ecs.g6.8xlarge	32	128	10/none	2,000,000	600,000	16	8	20	60,000	5
ecs.g6.13xlarge	52	192	12.5/none	3,000,000	900,000	32	7	20	100,000	8
ecs.g6.26xlarge	104	384	25/none	6,000,000	1,800,000	32	15	20	200,000	16

Note

-
- For more information about these specifications, see Instance family.

g5, general-purpose instance family

Features:

- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) or 8269CY (Cascade Lake) processors to provide consistent computing performance.

    Note Instances of this instance family may be deployed on different server platforms. If your business requires all instances to be deployed on the same server platform, we recommend that you use the g6, g6e, or g7 instance family instead.

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.

    Note

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.

    Note

  - Provides high network performance based on large computing capacity.

- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Enterprise-level applications of various types and sizes
  - Small and medium-sized database systems, caches, and search clusters
  - Data analytics and computing
  - Computing clusters and memory-intensive data processing

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.g5.large	2	8	1	300,000	2	2	6
ecs.g5.xlarge	4	16	1.5	500,000	2	3	10
ecs.g5.2xlarge	8	32	2.5	800,000	4	4	10
ecs.g5.3xlarge	12	48	4	900,000	4	6	10
ecs.g5.4xlarge	16	64	5	1,000,000	4	8	20
ecs.g5.6xlarge	24	96	7.5	1,500,000	6	8	20
ecs.g5.8xlarge	32	128	10	2,000,000	8	8	20
ecs.g5.16xlarge	64	256	20	4,000,000	16	8	20

Note

-
- For more information about these specifications, see Instance family.

g5ne, network-enhanced general-purpose instance family

Features:

- This instance family significantly improves the network throughput and packet forwarding rate per instance. A single instance can deliver a packet forwarding rate of up to 10,000,000 pps.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) or 8269CY (Cascade Lake) processors to provide consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports standard SSDs and ultra disks.
- Network:
  - Supports IPv6.

- Provides high network performance based on large computing capacity.

  > **Note** We recommend that you select instance types in the g5ne instance family to deploy Data Plane Development Kit (DPDK) applications.

- Supported scenarios:
  - DPDK applications
  - Network-intensive scenarios such as NFV or SD-WAN, mobile Internet, on-screen video comments, and telecom data forwarding
  - Small and medium-sized database systems, caches, and search clusters
  - Enterprise-level applications of various types and sizes
  - Big data analysis and machine learning

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.g5ne.large	2	8	1	400,000	450,000	2	3	10	10,000	1
ecs.g5ne.xlarge	4	16	2	750,000	900,000	4	4	15	15,000	1
ecs.g5ne.2xlarge	8	32	3.5	1,500,000	1,750,000	8	6	15	30,000	1
ecs.g5ne.4xlarge	16	64	7	3,000,000	3,500,000	16	8	30	60,000	2
ecs.g5ne.8xlarge	32	128	15	6,000,000	7,000,000	32	8	30	110,000	4
ecs.g5ne.16xlarge	64	256	30	12,000,000	14,000,000	32	8	30	130,000	8
ecs.g5ne.18xlarge	72	288	33	13,500,000	16,000,000	32	15	50	160,000	9

> **Note**
>
> -
> - For more information about these specifications, see Instance family.

sn2ne, network-enhanced general-purpose instance family

Features:

- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) or Platinum 8163 (Skylake) processors to provide consistent computing performance.

    > **Note** Instances of this instance family may be deployed on different server platforms. If your business requires all instances to be deployed on the same server platform, we recommend that you use the g6, g6e, or g7 instance family instead.

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Enterprise-level applications of various types and sizes
  - Small and medium-sized database systems, caches, and search clusters
  - Data analytics and computing
  - Computing clusters and memory-intensive data processing

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.sn2ne.large	2	8	1	300,000	2	2	6
ecs.sn2ne.xlarge	4	16	1.5	500,000	2	3	10

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.sn2ne.2xlarge	8	32	2	1,000,000	4	4	10
ecs.sn2ne.3xlarge	12	48	2.5	1,300,000	4	6	10
ecs.sn2ne.4xlarge	16	64	3	1,600,000	4	8	20
ecs.sn2ne.6xlarge	24	96	4.5	2,000,000	6	8	20
ecs.sn2ne.8xlarge	32	128	6	2,500,000	8	8	20
ecs.sn2ne.14xlarge	56	224	10	4,500,000	14	8	20

Note

-
- For more information about these specifications, see Instance family.

References

- Instance family
- Create an instance by using the wizard

4.2. Compute-optimized instance families

This topic describes the features of compute-optimized instance families of Elastic Compute Service (ECS) and lists the instance types of each family.

- Recommended instance families
  - c7se, storage-enhanced compute-optimized instance family
  - c7a, compute-optimized instance family
  - c7, compute-optimized instance family
  - c7t, security-enhanced compute-optimized instance family
  - c6a, compute-optimized instance family
  - c6t, security-enhanced compute-optimized instance family
  - c6e, compute-optimized instance family with enhanced performance
  - c6, compute-optimized instance family

  - c5, compute-optimized instance family
  - ic5, compute-intensive instance family
- Other available instance families (If these instance families are sold out, you can use the recommended ones.)

  sn1ne, network-enhanced compute-optimized instance family

c7se, storage-enhanced compute-optimized instance family

Features:

- This instance family uses the third-generation SHENLONG architecture and Intel Ice Lake processors to improve storage I/O performance.
- This instance family delivers a sequential read/write throughput of up to 64 Gbit/s and up to 1,000,000 IOPS per instance.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses third-generation Intel® Xeon® Scalable (Ice Lake) processors that deliver a base frequency of 2.7 GHz and an all-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only enhanced SSDs (ESSDs) and provides ultra-high I/O performance.
  - Allows a maximum of 64 data disks to be attached to a single instance. When you create an instance, you can attach a maximum of 16 data disks to the instance. If the instance requires more data disks, you can attach more data disks after the instance is created. For more information, see Attach a data disk.

    Note During the invitational preview of this instance family, you can attach data disks to instances of this instance family only when you create the instances, and a maximum of 16 data disks can be attached to a single instance.
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - I/O-intensive scenarios such as large and medium-sized online transactional processing (OLTP) core databases

- Large and medium-sized NoSQL databases
- Search and real-time log analytics
- Traditional large enterprise-level commercial software such as SAP

Instance types

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	Network interface controller (NIC) queues	Elastic network interfaces (ENIs)	Private IP addresses per ENI	Maximum attached data disks	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.c7se.large	2	4	1.2/burstable up to 3	450,000	Up to 250,000	2	3	6	16	30,000/burstable up to 150,000	3/10
ecs.c7se.xlarge	4	8	2/burstable up to 5	500,000	Up to 250,000	4	4	15	16	60,000/burstable up to 150,000	4/10
ecs.c7se.2xlarge	8	16	3/burstable up to 8	800,000	Up to 250,000	8	4	15	16	100,000/burstable up to 150,000	6/10
ecs.c7se.3xlarge	12	24	4.5/burstable up to 10	1,200,000	Up to 250,000	8	8	15	16	120,000/burstable up to 150,000	8/10
ecs.c7se.4xlarge	16	32	6/burstable up to 10	1,500,000	300,000	8	8	30	24	150,000/none	10/none

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	Network interface controller (NIC) queues	Elastic network interfaces (ENIs)	Private IP addresses per ENI	Maximum attached data disks	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.c7se.6xlarge	24	48	8/burstable up to 10	2,250,000	450,000	12	8	30	24	200,000/none	12/none
ecs.c7se.8xlarge	32	64	10/none	3,000,000	600,000	16	8	30	30	300,000/none	16/none
ecs.c7se.16xlarge	64	128	16/none	6,000,000	1,200,000	32	8	30	56	500,000/none	32/none
ecs.c7se.32xlarge	128	256	32/none	12,000,000	2,400,000	32	15	30	64	1,000,000/none	64/none

Note

-
- For more information about these specifications, see Instance family.

c7a, compute-optimized instance family

Features:

- This instance family uses the third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance. This instance family utilizes the fast path acceleration feature of chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses 2.55 GHz AMD EPYC™ MILAN processors that deliver a maximum single-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides disk burstable IOPS and bandwidth capabilities for low-specification instances.
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides burstable bandwidth capabilities for low-specification instances.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Video encoding and decoding
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Web frontend servers
  - Frontend servers of massively multiplayer online (MMO) games
  - Scenarios where applications such as DevOps applications are developed and tested
  - Data analysis and batch processing
  - High-performance scientific and engineering applications
  - Enterprise-level applications of various types and sizes

Instance types

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.c7a.large	2	4	1/burstable up to 10	900,000	Up to 250,000	2	3	6	12,500/burstable up to 110,000	1/burstable up to 6

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.c7a.xlarge	4	8	1.5/burstable up to 10	1,000,000	Up to 250,000	4	4	15	20,000/burstable up to 110,000	1.5/burstable up to 6
ecs.c7a.2xlarge	8	16	2.5/burstable up to 10	1,600,000	Up to 250,000	8	4	15	30,000/burstable up to 110,000	2/burstable up to 6
ecs.c7a.4xlarge	16	32	5/burstable up to 10	2,000,000	300,000	8	8	30	60,000/burstable up to 110,000	3/burstable up to 6
ecs.c7a.8xlarge	32	64	8/burstable up to 10	3,000,000	600,000	16	7	30	75,000/burstable up to 110,000	4/burstable up to 6
ecs.c7a.16xlarge	64	128	16/none	6,000,000	1,000,000	32	7	30	150,000/none	8/none
ecs.c7a.32xlarge	128	256	32/none	12,000,000	2,000,000	32	15	30	300,000/none	16/none

Note

-
- For more information about these specifications, see Instance family.
- Ubuntu 16 and Debian 9 operating system kernels do not support AMD EPYC™ MILAN processors. Do not use Ubuntu 16 or Debian 9 images to create instances of this instance family. Instances of this instance family created from Ubuntu 16 or Debian 9 images cannot start.

c7, compute-optimized instance family

Features:

- This instance family uses third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance. This instance family utilizes the fast path acceleration feature of chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- This instance family supports the virtual Trusted Platform Module (vTPM) feature and implements trusted boot based on Trusted Cryptography Module (TCM) or Trusted Platform Module (TPM) chips to provide ultra-high security capabilities. During a trusted boot, all modules in the boot chain from the underlying server to the ECS instance are measured and verified.
- This instance family supports the Enclave feature and provides a virtualization-based confidential computing environment. For more information, see Build a confidential computing environment by using Enclave.

  Note The Enclave feature is in invitational preview. If you want to use this feature, go to the Enclave product page.

- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses third-generation Intel® Xeon® Scalable (Ice Lake) processors that deliver a base frequency of 2.7 GHz and an all-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides burstable storage I/O performance for low-specification instances.
  - Provides high storage I/O performance based on large computing capacity.
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides burstable network performance for low-specification instances.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:

- Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
- Frontend servers of MMO games
- Web frontend servers
- Data analytics, batch processing, and video encoding
- High-performance scientific and engineering applications
- Scenarios that require secure and trusted computing
- Enterprise-level applications of various types and sizes
- Blockchain scenarios

Instance types

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.c7.large	2	4	2/burstable up to 10	900,000	Yes	Up to 250,000	2	3	6	20,000/burstable up to 110,000	1.5/burstable up to 6
ecs.c7.xlarge	4	8	3/burstable up to 10	1,000,000	Yes	Up to 250,000	4	4	15	40,000/burstable up to 110,000	2/burstable up to 6
ecs.c7.2xlarge	8	16	5/burstable up to 10	1,600,000	Yes	Up to 250,000	8	4	15	50,000/burstable up to 110,000	3/burstable up to 6
ecs.c7.3xlarge	12	24	8/burstable up to 10	2,400,000	Yes	Up to 250,000	8	8	15	70,000/burstable up to 110,000	4/burstable up to 6

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.c7.4xlarge	16	32	10/burstable up to 25	3,000,000	Yes	300,000	8	8	30	80,000/burstable up to 110,000	5/burstable up to 6
ecs.c7.6xlarge	24	48	12/burstable up to 25	4,500,000	Yes	450,000	12	8	30	110,000/none	6/none
ecs.c7.8xlarge	32	64	16/burstable up to 25	6,000,000	Yes	600,000	16	8	30	150,000/none	8/none
ecs.c7.16xlarge	64	128	32/none	12,000,000	Yes	1,200,000	32	8	30	300,000/none	16/none
ecs.c7.32xlarge	128	256	64/none	24,000,000	Yes	2,400,000	32	15	30	600,000/none	32/none

Note

-
-
- For more information about these specifications, see Instance family.

c7t, security-enhanced compute-optimized instance family

Features:

- This instance family supports up to 128 GiB of encrypted memory and encrypted computing based on Intel® Software Guard Extensions (SGX) to protect the confidentiality and integrity of essential code and data from malware attacks.
- This instance family supports Virtual SGX (vSGX) and allows you to select instance types that suit your needs.

  Notice
- This instance family implements trusted boot based on TCM or TPM chips. During a trusted boot, all modules in the boot chain from the underlying server to the guest OS are measured and verified.
- This instance family offloads a large number of virtualization features to dedicated hardware with the use of the third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2. About 50% of memory is encrypted.
  - Uses third-generation Intel® Xeon® Scalable (Ice Lake) processors that deliver a base frequency of 2.7 GHz and an all-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides high storage I/O performance based on large computing capacity.
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios that involve sensitive information such as personal identity information, healthcare information, financial information, and intellectual property data
  - Scenarios where confidential data is shared among multiple parties
  - Blockchain scenarios
  - Confidential machine learning
  - Scenarios that require high security and enhanced trust, such as services for financial organizations, public service sectors, and enterprises
  - Enterprise-level applications of various types and sizes

Instance types

Instance type	vCPU	Memory (GiB)	Encrypted memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.c7t.large	2	4	2	2/burstable up to 10	900,000	Yes	Up to 250,000	2	3	6	20,000/burstable up to 110,000	1.5/burstable up to 6
ecs.c7t.xlarge	4	8	4	3/burstable up to 10	1,000,000	Yes	Up to 250,000	4	4	15	40,000/burstable up to 110,000	2/burstable up to 6
ecs.c7t.2xlarge	8	16	8	5/burstable up to 10	1,600,000	Yes	Up to 250,000	8	4	15	50,000/burstable up to 110,000	3/burstable up to 6
ecs.c7t.3xlarge	12	24	12	8/burstable up to 10	2,400,000	Yes	Up to 250,000	8	8	15	70,000/burstable up to 110,000	4/burstable up to 6
ecs.c7t.4xlarge	16	32	16	10/burstable up to 25	3,000,000	Yes	300,000	8	8	30	80,000/burstable up to 110,000	5/burstable up to 6

Instance type	vCPU	Memory (GiB)	Encrypted memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.c7t.6xlarge	24	48	24	12/burstable up to 25	4,500,000	Yes	450,000	12	8	30	110,000/none	6/none
ecs.c7t.8xlarge	32	64	32	16/burstable up to 25	6,000,000	Yes	600,000	16	8	30	150,000/none	8/none
ecs.c7t.16xlarge	64	128	64	32/none	12,000,000	Yes	1,200,000	32	8	30	300,000/none	16/none
ecs.c7t.32xlarge	128	256	128	64/none	24,000,000	Yes	2,400,000	32	15	30	600,000/none	32/none

Note

- For more information about these specifications, see Instance family.

c6a, compute-optimized instance family

Features:

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of the SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.

- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses 2.6 GHz AMD EPYC™ ROME processors that deliver a turbo frequency of 3.3 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Video encoding and decoding
  - Scenarios where large volumes of packets are received and transmitted
  - Web frontend servers
  - Frontend servers of MMO games
  - Scenarios where applications such as DevOps applications are developed and tested

Instance types

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.c6a.large	2	4	1/10	900,000	Up to 250,000	2	2	6	12,500	1
ecs.c6a.xlarge	4	8	1.5/10	1,000,000	Up to 250,000	4	3	15	20,000	1.5
ecs.c6a.2xlarge	8	16	2.5/10	1,600,000	Up to 250,000	8	4	15	30,000	2

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.c6a.4xlarge	16	32	5/10	2,000,000	300,000	8	8	30	60,000	3
ecs.c6a.8xlarge	32	64	8/10	3,000,000	600,000	16	7	30	75,000	4,0
ecs.c6a.16xlarge	64	128	16/none	6,000,000	1,000,000	32	8	30	150,000	8
ecs.c6a.32xlarge	128	256	32/none	12,000,000	2,000,000	32	15	30	300,000	16

Note

-
- For more information about these specifications, see Instance family.

c6t, security-enhanced compute-optimized instance family

Features:

- This instance family implements trusted boots based on TPM chips. During a trusted boot, each module in the boot chain from the underlying hardware to the guest OS is measured and verified.
- This instance family supports comprehensive monitoring and provides a full set of trusted capabilities at the IaaS layer.
- This instance family offloads a large number of virtualization features to dedicated hardware with the use of the third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads. This instance family utilizes the fast path acceleration feature of chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269 (Cascade Lake) processors that deliver a turbo frequency of 3.2 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios that require high security and enhanced trust, such as services for financial organizations, public service sectors, and enterprises
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Web frontend servers
  - Frontend servers of MMO games
  - Data analytics, batch processing, and video encoding
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.c6t.large	2	4	1.2/burstable up to 10	900,000	Yes	Up to 250,000	2	3	6	20,000	1
ecs.c6t.xlarge	4	8	2/burstable up to 10	1,000,000	Yes	Up to 250,000	4	4	15	40,000	1.5
ecs.c6t.2xlarge	8	16	3/burstable up to 10	1,600,000	Yes	Up to 250,000	8	4	15	50,000	2

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.c6t.4xlarge	16	32	6/burstable up to 10	3,000,000	Yes	300,000	8	8	30	80,000	3
ecs.c6t.8xlarge	32	64	10/none	6,000,000	Yes	600,000	16	8	30	150,000	5
ecs.c6t.13xlarge	52	96	16/none	9,000,000	Yes	900,000	32	7	30	240,000	8
ecs.c6t.26xlarge	104	192	32/none	24,000,000	Yes	1,800,000	32	15	30	480,000	16

Note

-
- For more information about these specifications, see Instance family.
- The results for network capabilities are the maximum values obtained from single-item tests. For example, when network bandwidth is tested, no stress tests are performed on the packet forwarding rate or other network metrics.

c6e, compute-optimized instance family with enhanced performance

Features:

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of the third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads. This instance family utilizes the fast path acceleration feature of chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269 (Cascade) processors that deliver a turbo frequency of 3.2 GHz to provide consistent computing performance.

- Allows you to enable or disable Hyper-Threading.

  Note

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides high storage I/O performance based on large computing capacity.

    Note

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.

    Note

  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Web frontend servers
  - Frontend servers of MMO games
  - Data analytics, batch processing, and video encoding
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.c6e.large	2	4	1.2/burstable up to 10	900,000	Up to 250,000	2	3	6	20,000	1
ecs.c6e.xlarge	4	8	2/burstable up to 10	1,000,000	Up to 250,000	4	4	15	40,000	1.5

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.c6e.2xlarge	8	16	3/burstable up to 10	1,600,000	Up to 250,000	8	4	15	50,000	2
ecs.c6e.4xlarge	16	32	6/burstable up to 10	3,000,000	300,000	8	8	30	80,000	3
ecs.c6e.8xlarge	32	64	10/none	6,000,000	600,000	16	8	30	150,000	5
ecs.c6e.13xlarge	52	96	16/none	9,000,000	1,000,000	32	7	30	240,000	8
ecs.c6e.26xlarge	104	192	32/none	24,000,000	1,800,000	32	15	30	480,000	16

Note

-
- For more information about these specifications, see Instance family.
- The results for network capabilities are the maximum values obtained from single-item tests. For example, when network bandwidth is tested, no stress tests are performed on the packet forwarding rate or other network metrics.

c6, compute-optimized instance family

Features:

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of the SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.2 GHz to provide consistent computing performance.

- Allows you to enable or disable Hyper-Threading.

  Note

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.

    Note

  - Provides high storage I/O performance based on large computing capacity.

    Note

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supports changes to instance types in the g6 or r6 instance family.
- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Web frontend servers
  - Frontend servers of MMO games
  - Data analytics, batch processing, and video encoding
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.c6.large	2	4	1/3	300,000	Up to 250,000	2	2	6	10,000	1
ecs.c6.xlarge	4	8	1.5/5	500,000	Up to 250,000	4	3	10	20,000	1.5
ecs.c6.2xlarge	8	16	2.5/8	800,000	Up to 250,000	8	4	10	25,000	2

Instance type	vCPU	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.c6.3xlarge	12	24	4/10	900,000	Up to 250,000	8	6	10	30,000	2.5
ecs.c6.4xlarge	16	32	5/10	1,000,000	300,000	8	8	20	40,000	3
ecs.c6.6xlarge	24	48	7.5/10	1,500,000	450,000	12	8	20	50,000	4
ecs.c6.8xlarge	32	64	10/none	2,000,000	600,000	16	8	20	60,000	5
ecs.c6.13xlarge	52	96	12.5/none	3,000,000	900,000	32	7	20	100,000	8
ecs.c6.26xlarge	104	192	25/none	6,000,000	1,800,000	32	15	20	200,000	16

> **Note**
>
> -
> - For more information about these specifications, see Instance family.

c5, compute-optimized instance family

Features:

- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) or 8269CY (Cascade Lake) processors to provide consistent computing performance.

    > **Note** Instances of this instance family may be deployed on different server platforms. If your business requires all instances to be deployed on the same server platform, we recommend that you use the c6, c6e, or c7 instance family instead.

- Storage:
  - Is an instance family in which all instances are I/O optimized.

- Supports ESSDs, standard SSDs, and ultra disks.

  Note

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Web frontend servers
  - Frontend servers of MMO games
  - Data analytics, batch processing, and video encoding
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPU	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.c5.large	2	4	1	300,000	2	2	6
ecs.c5.xlarge	4	8	1.5	500,000	2	3	10
ecs.c5.2xlarge	8	16	2.5	800,000	4	4	10
ecs.c5.3xlarge	12	24	4	900,000	4	6	10
ecs.c5.4xlarge	16	32	5	1,000,000	4	8	20
ecs.c5.6xlarge	24	48	7.5	1,500,000	6	8	20
ecs.c5.8xlarge	32	64	10	2,000,000	8	8	20
ecs.c5.16xlarge	64	128	20	4,000,000	16	8	20

Note

-
- For more information about these specifications, see Instance family.

ic5, compute-intensive instance family

Features:

- Compute:
  - Offers a CPU-to-memory ratio of 1:1.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) or 8269CY (Cascade Lake) processors to provide consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Web frontend servers
  - Data analytics, batch processing, and video encoding
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Frontend servers of MMO games

Instance types

Instance type	vCPU	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.ic5.large	2	2	1	300,000	2	2	6
ecs.ic5.xlarge	4	4	1.5	500,000	2	3	10
ecs.ic5.2xlarge	8	8	2.5	800,000	2	4	10
ecs.ic5.3xlarge	12	12	4	900,000	4	6	10
ecs.ic5.4xlarge	16	16	5	1,000,000	4	8	20

Note

-
- For more information about these specifications, see Instance family.

sn1ne, network-enhanced compute-optimized instance family

Features:

- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) or Platinum 8163 (Skylake) processors to provide consistent computing performance.

    Note Instances of this instance family may be deployed on different server platforms. If your business requires all instances to be deployed on the same server platform, we recommend that you use the c6, c6e, or c7 instance family instead.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Web frontend servers
  - Frontend servers of MMO games
  - Data analytics, batch processing, and video encoding
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPU	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.sn1ne.large	2	4	1	300,000	2	2	6
ecs.sn1ne.xlarge	4	8	1.5	500,000	2	3	10

Instance type	vCPU	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.sn1ne.2xlarge	8	16	2	1,000,000	4	4	10
ecs.sn1ne.3xlarge	12	24	2.5	1,300,000	4	6	10
ecs.sn1ne.4xlarge	16	32	3	1,600,000	4	8	20
ecs.sn1ne.6xlarge	24	48	4.5	2,000,000	6	8	20
ecs.sn1ne.8xlarge	32	64	6	2,500,000	8	8	20

Note

-
- For more information about these specifications, see Instance family.

4.3. Memory optimized instance families

4.3.1. Memory-optimized instance families

This topic describes the features of memory-optimized instance families of Elastic Compute Service (ECS) and lists the instance types of each instance family.

- Recommended instance families
  - r7se, storage-enhanced memory-optimized instance family
  - r7a, memory-optimized instance family
  - r7, memory-optimized instance family
  - r7t, security-enhanced memory-optimized instance family
  - re6p, persistent memory-optimized instance family
  - r6a, memory-optimized instance family
  - r6e, memory-optimized instance family with enhanced performance
  - r6, memory-optimized instance family
  - re6, high-memory instance family
  - r5, memory-optimized instance family
- Other available instance families (If these instance families are sold out, you can use the recommended ones.)

- re4, high-memory instance family
- re4e, high-memory instance family
- se1ne, network-enhanced memory-optimized instance family
- se1, memory-optimized instance family

r7se, storage-enhanced memory-optimized instance family

Features:

- This instance family uses third-generation SHENLONG architecture and Intel Ice Lake processors to improve storage I/O performance.
- This instance family delivers a sequential read/write throughput of up to 64 Gbit/s and up to 1,000,000 IOPS per instance.
- Compute:
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses third-generation Intel® Xeon® Scalable (Ice Lake) processors that deliver a base frequency of 2.7 GHz and an all-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only enhanced SSDs (ESSDs) and provides ultra-high I/O performance.
  - Allows a maximum of 64 data disks to be attached to a single instance. When you create an instance, you can attach a maximum of 16 data disks to the instance . If the instance requires even more data disks, you can attach more data disks after the instance is created. For more information, see Attach a data disk.

    Note During the invitational preview of this instance family, you can attach data disks to instances of this instance family only when you create the instances, and a maximum of 16 data disks can be attached to a single instance.
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - I/O-intensive scenarios such as large and medium-sized online transactional processing (OLTP) core databases
  - Large and medium-sized NoSQL databases

- Search and real-time log analytics
- Traditional large enterprise-level commercial software such as SAP
- High-density deployment of containers

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IPv4 addresses per ENI	IPv6 addresses per ENI	Maximum attached data disks	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.r7se.large	2	16	1.2/burstable up to 3	450,000	Up to 250,000	2	3	6	6	16	30,000/burstable up to 150,000	3/10
ecs.r7se.xlarge	4	32	2/burstable up to 5	500,000	Up to 250,000	4	4	15	15	16	60,000/burstable up to 150,000	4/10
ecs.r7se.2xlarge	8	64	3/burstable up to 8	800,000	Up to 250,000	8	4	15	15	16	100,000/burstable up to 150,000	6/10
ecs.r7se.3xlarge	12	96	4.5/burstable up to 10	1,200,000	Up to 250,000	8	8	15	15	16	120,000/burstable up to 150,000	8/10

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IPv4 addresses per ENI	IPv6 addresses per ENI	Maximum attached data disks	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.r7se.4xlarge	16	128	6/burstable up to 10	1,500,000	300,000	8	8	30	30	24	150,000/none	10/none
ecs.r7se.6xlarge	24	192	8/burstable up to 10	2,250,000	450,000	12	8	30	30	24	200,000/none	12/none
ecs.r7se.8xlarge	32	256	10/none	3,000,000	600,000	16	8	30	30	30	300,000/none	16/none
ecs.r7se.16xlarge	64	512	16/none	6,000,000	1,200,000	32	8	30	30	56	500,000/none	32/none
ecs.r7se.32xlarge	128	1024	32/none	12,000,000	2,400,000	32	15	30	30	64	1,000,000/none	64/none

Note

-
- For more information about these specifications, see Instance family.

r7a, memory-optimized instance family

Features:

- This instance family uses third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance. This instance family utilizes fast path acceleration on chips to improve storage performance, network performance, and computing stability by an order of magnitude.

- Compute:
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses 2.55 GHz AMD EPYC™ MILAN processors that deliver a maximum single-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides disk burstable IOPS and bandwidth capabilities for low-specification instances.
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides burstable bandwidth capabilities for low-specification instances.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - High-performance databases and in-memory databases
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters
  - Blockchain applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	IPv6 addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	IPv6 addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.r7a.large	2	16	1/burstable up to 10	900,000	Up to 250,000	2	3	6	6	12,500/burstable up to 110,000	1/burstable up to 6
ecs.r7a.xlarge	4	32	1.5/burstable up to 10	1,000,000	Up to 250,000	4	4	15	15	20,000/burstable up to 110,000	1.5/burstable up to 6
ecs.r7a.2xlarge	8	64	2.5/burstable up to 10	1,600,000	Up to 250,000	8	4	15	15	30,000/burstable up to 110,000	2/burstable up to 6
ecs.r7a.4xlarge	16	128	5/burstable up to 10	2,000,000	300,000	8	8	30	30	60,000/burstable up to 110,000	3/burstable up to 6
ecs.r7a.8xlarge	32	256	8/burstable up to 10	3,000,000	600,000	16	7	30	30	75,000/burstable up to 110,000	4/burstable up to 6
ecs.r7a.16xlarge	64	512	16/none	6,000,000	1,000,000	32	7	30	30	150,000/none	8/none

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	IPv6 addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.r7a.32xlarge	128	1024	32/none	12,000,000	2,000,000	32	15	30	30	300,000/none	16/none

> **Note**
>
> -
> - For more information about these specifications, see Instance family.
> - Ubuntu 16 and Debian 9 operating system kernels do not support AMD EPYC™ MILAN processors. Do not use Ubuntu 16 or Debian 9 images to create instances of this instance family. Instances of this instance family created from Ubuntu 16 or Debian 9 images cannot start.

r7, memory-optimized instance family

Features:

- This instance family uses the third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance. This instance family utilizes the fast path acceleration feature of chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- This instance family supports the virtual Trusted Platform Module (vTPM) feature and implements trusted boot based on Trusted Cryptography Module (TCM) or Trusted Platform Module (TPM) chips to provide ultra-high security capabilities. During a trusted boot, all modules in the boot chain from the underlying server to the ECS instance are measured and verified.
- This instance family supports the Enclave feature and provides a virtualization-based confidential computing environment. For more information, see Build a confidential computing environment by using Enclave.

  > **Note** The Enclave feature is in invitational preview. If you want to use this feature, go to the Enclave product page.

- Compute:
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses third-generation Intel® Xeon® Scalable (Ice Lake) processors that deliver a base frequency of 2.7 GHz and an all-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.
- Storage:

- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs.
- Provides burstable storage I/O performance for low-specification instances.
- Provides high storage I/O performance based on large computing capacity.

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides burstable network performance for low-specification instances.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - High-performance databases and in-memory databases
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters
  - Scenarios that require secure and trusted computing

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IPv4 addresses per ENI	IPv6 addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.r7.large	2	16	2/burstable up to 10	900,000	Yes	Up to 250,000	2	3	6	6	20,000/burstable up to 110,000	1.5/burstable up to 6
ecs.r7.xlarge	4	32	3/burstable up to 10	1,000,000	Yes	Up to 250,000	4	4	15	15	40,000/burstable up to 110,000	2/burstable up to 6

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IPv4 addresses per ENI	IPv6 addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.r7.2xlarge	8	64	5/burstable up to 10	1,600,000	Yes	Up to 250,000	8	4	15	15	50,000/burstable up to 110,000	3/burstable up to 6
ecs.r7.3xlarge	12	96	8/burstable up to 10	2,400,000	Yes	Up to 250,000	8	8	15	15	70,000/burstable up to 110,000	4/burstable up to 6
ecs.r7.4xlarge	16	128	10/burstable up to 25	3,000,000	Yes	300,000	8	8	30	30	80,000/burstable up to 110,000	5/burstable up to 6
ecs.r7.6xlarge	24	192	12/burstable up to 25	4,500,000	Yes	450,000	12	8	30	30	110,000/none	6/none
ecs.r7.8xlarge	32	256	16/burstable up to 25	6,000,000	Yes	600,000	16	8	30	30	150,000/none	8/none

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IPv4 addresses per ENI	IPv6 addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.r7.16xlarge	64	512	32/none	12,000,000	Yes	1,200,000	32	8	30	30	300,000/none	16/none
ecs.r7.32xlarge	128	1024	64/none	24,000,000	Yes	2,400,000	32	15	30	30	600,000/none	32/none

Note

-
-
- For more information about these specifications, see Instance family.

r7t, security-enhanced memory-optimized instance family

Features:

- This instance family supports up to 512 GiB of encrypted memory and encrypted computing based on Intel® Software Guard Extensions (SGX) to protect the confidentiality and integrity of your code and data from malware attacks.
- This instance family supports Virtual SGX (vSGX) and allows you to select instance types that suit your needs.

  Notice

- This instance family implements trusted boot based on TCM or TPM chips. During a trusted boot, all modules in the boot chain from the underlying server to the guest OS are measured and verified.
- This instance family offloads a large number of virtualization features to dedicated hardware with the use of the third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:8. About 50% of memory is encrypted.
  - Uses third-generation Intel® Xeon® Scalable (Ice Lake) processors that deliver a base frequency of 2.7 GHz and an all-core turbo frequency of 3.5 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides high storage I/O performance based on large computing capacity.
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Encrypted computing applications for databases
  - Scenarios that involve sensitive information such as personal identity information, healthcare information, financial information, and intellectual property data
  - Scenarios where confidential data is shared among multiple parties
  - Blockchain scenarios
  - Confidential machine learning
  - Scenarios that require high security and enhanced trust, such as services for financial organizations, public service sectors, and enterprises
  - Enterprise-level applications of various types and sizes

Instance types

Instance type	vCPUs	Memory (GiB)	Encrypted memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IPv4 addresses per ENI	IPv6 addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.r7t.large	2	16	8	2/burstable up to 10	900,000	Yes	Up to 250,000	2	3	6	6	20,000/burstable up to 110,000	1.5/burstable up to 6

Instance type	vCPUs	Memory (GiB)	Encrypted memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IPv4 addresses per ENI	IPv6 addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.r7t.xlarge	4	32	16	3/burstable up to 10	1,000,000	Yes	Up to 250,000	4	4	15	15	40,000/burstable up to 110,000	2/burstable up to 6
ecs.r7t.2xlarge	8	64	32	5/burstable up to 10	1,600,000	Yes	Up to 250,000	8	4	15	15	50,000/burstable up to 110,000	3/burstable up to 6
ecs.r7t.3xlarge	12	96	48	8/burstable up to 10	2,400,000	Yes	Up to 250,000	8	8	15	15	70,000/burstable up to 110,000	4/burstable up to 6
ecs.r7t.4xlarge	16	128	64	10/burstable up to 25	3,000,000	Yes	300,000	8	8	30	30	80,000/burstable up to 110,000	5/burstable up to 6

Instance type	vCPUs	Memory (GiB)	Encrypted memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Support for vTPM	Connections	NIC queues	ENIs	Private IPv4 addresses per ENI	IPv6 addresses per ENI	Disk baseline/burst IOPS	Disk baseline/burst bandwidth (Gbit/s)
ecs.r7t.6xlarge	24	192	96	12/burstable up to 25	4,500,000	Yes	450,000	12	8	30	30	110,000/none	6/none
ecs.r7t.8xlarge	32	256	128	16/burstable up to 25	6,000,000	Yes	600,000	16	8	30	30	150,000/none	8/none
ecs.r7t.16xlarge	64	512	256	32/none	12,000,000	Yes	1,200,000	32	8	30	30	300,000/none	16/none
ecs.r7t.32xlarge	128	1024	512	64/none	24,000,000	Yes	2,400,000	32	15	30	30	600,000/none	32/none

Note

- For more information about these specifications, see Instance family.

re6p, persistent memory-optimized instance family

For information about frequently asked questions about persistent memory-optimized instances, see Instance FAQ.

Features:

- This instance family uses Intel® Optane™ persistent memory.

  > **Notice** The reliability of data stored in persistent memory depends on the reliability of persistent memory devices and the physical servers to which these devices are attached. This increases the risks of single points of failure (SPOFs). To ensure the reliability of application data, we recommend that you implement data redundancy at the application layer and use cloud disks for long-term data storage.

- This instance family allows persistent memory to be used as memory or as local SSDs on instances of some instance types.

  > **Note** For more information, see Configure persistent memory usage.

- This instance family provides the ecs.re6p-redis.<nx>large instance types for Redis applications.

  > **Note** ecs.re6p-redis.<nx>large instance types are exclusively provided for Redis applications. Persistent memory on instances of these instance types is used as memory by default and cannot be re-configured as local SSDs. For more information about how to deploy a Redis application, see Deploy Redis applications on persistent memory-optimized instances.

- Compute:
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.2 GHz to provide consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Supports only virtual private clouds (VPCs).
- Supported scenarios:
  - Redis and other NoSQL databases such as Cassandra and MongoDB
  - Structured databases such as MySQL
  - I/O-intensive applications such as e-commerce, online games, and media applications
  - Search scenarios that use solutions such as Elasticsearch
  - Live video streaming, instant messaging, and room-based online games that require persistent connections
  - High-performance relational databases and OLTP systems

Instance types

Instance type	vCPUs	Memory (GiB)	Persistent memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.re6p.large	2	8	31.5	1/3	300,000	Up to 250,000	2	2	10,000	1
ecs.re6p.xlarge	4	16	63	1.5/5	500,000	Up to 250,000	4	3	20,000	1.5
ecs.re6p.2xlarge	8	32	126	2.5/10	800,000	Up to 250,000	8	4	25,000	2
ecs.re6p.13xlarge	52	192	756	12.5/none	3,000,000	900,000	32	7	100,000	8
ecs.re6p.26xlarge	104	384	1512	25/none	6,000,000	1,800,000	32	15	200,000	16,0
ecs.re6p-redis.large	2	8	31.5	1/3	300,000	Up to 250,000	2	2	10,000	1
ecs.re6p-redis.xlarge	4	16	63	1.5/5	500,000	Up to 250,000	4	3	20,000	1.5
ecs.re6p-redis.2xlarge	8	32	126	2.5/10	800,000	Up to 250,000	8	4	25,000	2
ecs.re6p-redis.13xlarge	52	192	756	12.5/none	3,000,000	900,000	32	7	100,000	8

Note

-
- For more information about these specifications, see Instance family.

r6a, memory-optimized instance family

Features:

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of the SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses 2.6 GHz AMD EPYC™ ROME processors that deliver a turbo frequency of 3.3 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Provides high storage I/O performance based on large computing capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Video encoding and decoding
  - Scenarios where large volumes of packets are received and transmitted
  - In-memory databases
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters
  - Scenarios where applications are developed and tested, such as DevOps scenarios

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.r6a.large	2	16	1/10	900,000	Up to 250,000	2	2	6	12,500	1
ecs.r6a.xlarge	4	32	1.5/10	1,000,000	Up to 250,000	5	3	15	20,000	1.5
ecs.r6a.2xlarge	8	64	2.5/10	1,600,000	Up to 250,000	8	4	15	30,000	2
ecs.r6a.4xlarge	16	128	5/10	2,000,000	300,000	8	8	30	60,000	3
ecs.r6a.8xlarge	32	256	8/10	3,000,000	600,000	16	7	30	75,000	4,0
ecs.r6a.16xlarge	64	512	16/none	6,000,000	1,000,000	32	8	30	150,000	8

Note

-
- For more information about these specifications, see Instance family.

r6e, memory-optimized instance family with enhanced performance

Features:

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of the third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads. This instance family utilizes fast path acceleration on chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- Compute:
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269 processors that deliver a turbo frequency of 3.2 GHz to provide consistent computing performance.

- Allows you to enable or disable Hyper-Threading.

  Note

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs.
  - Provides high network and storage I/O performance based on large computing capacity.

    Note

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.

    Note

- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance databases and in-memory databases
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.r6e.large	2	16	1.2/burstable up to 10	900,000	Up to 250,000	2	3	6	20,000	1
ecs.r6e.xlarge	4	32	2/burstable up to 10	1,000,000	Up to 250,000	4	4	15	40,000	1.5
ecs.r6e.2xlarge	8	64	3/burstable up to 10	1,600,000	Up to 250,000	8	4	15	50,000	2

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.r6e.4xlarge	16	128	6/burstable up to 10	3,000,000	300,000	8	8	30	80,000	3
ecs.r6e.8xlarge	32	256	10/none	6,000,000	600,000	16	8	30	150,000	5
ecs.r6e.13xlarge	52	384	16/none	9,000,000	1,000,000	32	7	30	240,000	8
ecs.r6e.26xlarge	104	768	32/none	24,000,000	1,800,000	32	15	30	480,000	16

Note

-
- For more information about these specifications, see Instance family.
- The results for network capabilities are the maximum values obtained from single-item tests. For example, when network bandwidth is tested, no stress tests are performed on the packet forwarding rate or other network metrics.

r6, memory-optimized instance family

Features:

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of the SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.2 GHz to provide consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.

- Supports ESSDs, standard SSDs, and ultra disks.

  Note

- Provides high storage I/O performance based on large computing capacity.

  Note

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supports changes to instance types in the g6 or c6 instance family.
- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance databases and in-memory databases
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.r6.large	2	16	1/3	300,000	Up to 250,000	2	2	6	10,000	1
ecs.r6.xlarge	4	32	1.5/5	500,000	Up to 250,000	4	3	10	20,000	1.5
ecs.r6.2xlarge	8	64	2.5/8	800,000	Up to 250,000	8	4	10	25,000	2
ecs.r6.3xlarge	12	96	4/10	900,000	Up to 250,000	8	6	10	30,000	2.5
ecs.r6.4xlarge	16	128	5/10	1,000,000	300,000	8	8	20	40,000	3

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.r6.6xlarge	24	192	7.5/10	1,500,000	450,000	12	8	20	50,000	4
ecs.r6.8xlarge	32	256	10/none	2,000,000	600,000	16	8	20	60,000	5
ecs.r6.13xlarge	52	384	12.5/none	3,000,000	900,000	32	7	20	100,000	8
ecs.r6.26xlarge	104	768	25/none	6,000,000	1,800,000	32	15	20	200,000	16

Note

-
- For more information about these specifications, see Instance family.

re6, high-memory instance family

Features:

- This instance family is optimized for high-performance databases, in-memory databases, and enterprise-level memory-intensive applications.
- Compute:
  - Offers a CPU-to-memory ratio of 1:15 and up to 3 TiB of memory.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.2 GHz to provide consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
- Supported scenarios:
  - High-performance databases and in-memory databases such as SAP HANA
  - Memory-intensive applications

- Big data processing engines such as Apache Spark and Presto

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.re6.4xlarge	16	256	5	900,000	8	7	20	25,000	2
ecs.re6.8xlarge	32	512	10	1,800,000	16	7	20	50,000	4
ecs.re6.13xlarge	52	768	10	1,800,000	16	7	20	50,000	4
ecs.re6.16xlarge	64	1024	16	3,000,000	32	7	20	100,000	8
ecs.re6.26xlarge	104	1536	16	3,000,000	32	7	20	100,000	8
ecs.re6.32xlarge	128	2048	32	6,000,000	32	15	20	200,000	16
ecs.re6.52xlarge	208	3072	32	6,000,000	32	15	20	200,000	16

Note

-
- For more information about these specifications, see Instance family.

r5, memory-optimized instance family

Features:

- Compute:
  - Offers a CPU-to-memory ratio of 1:8.

- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) or Intel® Xeon® Platinum 8269CY (Cascade Lake) processors to provide consistent computing performance.

  > **Note** Instances of this instance family may be deployed on different server platforms. If your business requires all instances to be deployed on the same server platform, we recommend that you use the r6, r6e, or r7 instance family instead.

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.

  > **Note**

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance databases and in-memory databases
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.r5.large	2	16	1	300,000	2	2	6
ecs.r5.xlarge	4	32	1.5	500,000	2	3	10
ecs.r5.2xlarge	8	64	2.5	800,000	4	4	10
ecs.r5.3xlarge	12	96	4	900,000	4	6	10
ecs.r5.4xlarge	16	128	5	1,000,000	4	8	20
ecs.r5.6xlarge	24	192	7.5	1,500,000	6	8	20

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.r5.8xlarge	32	256	10	2,000,000	8	8	20
ecs.r5.16xlarge	64	512	20	4,000,000	16	8	20

Note

-
- For more information about these specifications, see Instance family.

re4, high-memory instance family

Features:

- This instance family is optimized for high-performance databases, in-memory databases, and enterprise-level memory-intensive applications.
- The ecs.re4.20xlarge and ecs.re4.40xlarge instance types are SAP HANA-certified.
- Compute:
  - Offers a CPU-to-memory ratio of 1:12 and up to 1,920 GiB of memory.
  - Uses 2.2 GHz Intel® Xeon® E7 8880 v4 (Broadwell) processors that deliver a maximum turbo frequency of 2.4 GHz to provide consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
- Supported scenarios:
  - High-performance databases and in-memory databases such as SAP HANA
  - Memory-intensive applications
  - Big data processing engines such as Apache Spark and Presto

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.re4.10xlarge	40	480	8	1,000,000	8	4	10

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.re4.20xlarge	80	960	15	2,000,000	16	8	20
ecs.re4.40xlarge	160	1920	30	4,500,000	16	8	20

Note

-
- For more information about these specifications, see Instance family.

re4e, high-memory instance family

Features:

- This instance family is optimized for high-performance databases, in-memory databases, and enterprise-level memory-intensive applications.
- Compute:
  - Offers a CPU-to-memory ratio of 1:24 and up to 3,840 GiB of memory.
  - Uses 2.2 GHz Intel® Xeon® E7 8880 v4 (Broadwell) processors that deliver a maximum turbo frequency of 2.4 GHz to provide consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
- Supported scenarios:
  - High-performance databases and in-memory databases such as SAP HANA
  - Memory-intensive applications
  - Big data processing engines such as Apache Spark and Presto

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.re4e.40xlarge	160	3840	30	4,500,000	16	15	20

Note

-
- For more information about these specifications, see Instance family.

se1ne, network-enhanced memory-optimized instance family

Features:

- Compute:
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) or Intel® Xeon® Platinum 8163 (Skylake) processors to provide consistent computing performance.

    Note Instances of this instance family may be deployed on different server platforms. If your business requires all instances to be deployed on the same server platform, we recommend that you use the r6, r6e, or r7 instance family instead.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance databases and in-memory databases
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.se1ne.large	2	16	1	300,000	2	2	6
ecs.se1ne.xlarge	4	32	1.5	500,000	2	3	10
ecs.se1ne.2xlarge	8	64	2	1,000,000	4	4	10

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.se1ne.3xlarge	12	96	2.5	1,300,000	4	6	10
ecs.se1ne.4xlarge	16	128	3	1,600,000	4	8	20
ecs.se1ne.6xlarge	24	192	4.5	2,000,000	6	8	20
ecs.se1ne.8xlarge	32	256	6	2,500,000	8	8	20
ecs.se1ne.14xlarge	56	480	10	4,500,000	14	8	20

Note

-
- For more information about these specifications, see Instance family.

se1, memory-optimized instance family

Features:

- Compute:
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors to provide consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - High-performance databases and in-memory databases
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.se1.large	2	16	0.5	100,000	1	2	6
ecs.se1.xlarge	4	32	0.8	200,000	1	3	10
ecs.se1.2xlarge	8	64	1.5	400,000	1	4	10
ecs.se1.4xlarge	16	128	3	500,000	2	8	20
ecs.se1.8xlarge	32	256	6	800,000	3	8	20
ecs.se1.14xlarge	56	480	10	1,200,000	4	8	20

Note

-
- For more information about these specifications, see Instance family.

4.3.2. Configure persistent memory usage

Persistent memory on Elastic Compute Service (ECS) instances can be used as memory or local disks based on the instance type. This topic describes how to configure the usage mode of persistent memory.

Prerequisites

The instance uses an image of one of the following versions:

- Alibaba Cloud Linux 2
- CentOS 7.6 or later
- Ubuntu 18.10 or later
- SUSE Linux 12 SP4 or later

Context

The access latency of persistent memory is lower than that of regular memory. When an instance is stopped or restarted, data in its persistent memory is retained. Persistent memory can be used as memory or local disks.

- When persistent memory is used as memory, you can move data such as non-hot data that does not require high-speed storage access from regular memory to persistent memory. Persistent memory offers large capacity at a low price per GiB and can help reduce the total cost of ownership (TCO) per

GiB of memory.

- When persistent memory is used as local disks, it delivers ultra-high I/O performance and a read/write latency as low as 170 nanoseconds. You can use persistent memory for core application databases that require consistent response time. You can also replace Non-Volatile Memory Express (NVMe) SSDs with persistent memory used as local disks to achieve higher IOPS, higher bandwidth, and lower latency and solve performance bottlenecks.

> Note

- On instances of ecs.re6p, ecs.re7p, and ecs.r7p instance types, persistent memory can be used as memory or local disks.
- On instances of ecs.re6p-redis.<nx>large, ecs.re7p, and ecs.r7p instance types, persistent memory can be used only as memory.

For more information about instance types, see Instance family.

In this example, the following configurations are used:

- Instance type: ecs.re6p.2xlarge
- Image: Alibaba Cloud Linux 2.1903 LTS 64-bit

Configure persistent memory as memory

You can use memkind to allocate memory space. For more information about how to use memkind, visit memkind.

1. Log on to the instance.

   For more information, see Connection methods.

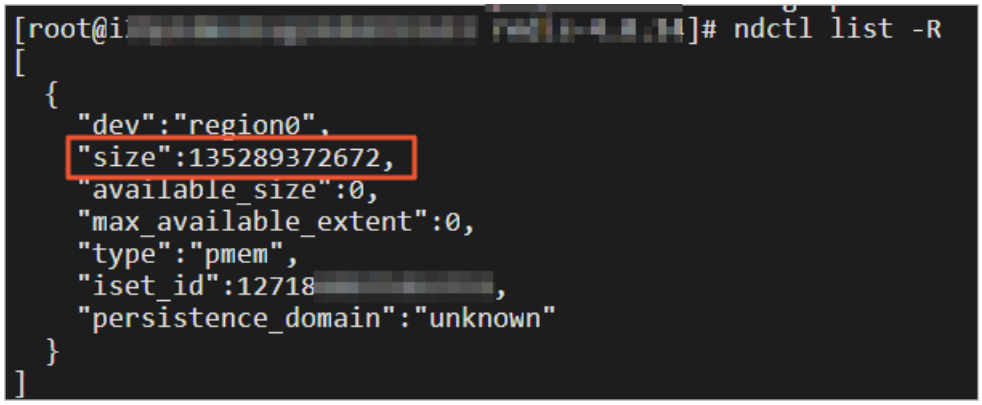
2. Install the persistent memory management tool and set the usage mode to devdax.

   ```
   yum install -y ndctl daxctl && \
   ndctl create-namespace -f -e namespace0.0 --mode=devdax
   ```

3. Check the memory size.

   - Check the size of persistent memory.

     ```
     ndctl list -R
     ```

   - Check the size of regular memory.

     ```
     cat /proc/meminfo
     ```

Configure persistent memory as a local disk

1. Log on to the instance.

   For more information, see Connection methods.

2. Install the persistent memory management tool and set the usage mode to fsdax.

```
yum install -y ndctl daxctl && \
ipmctl create -goal PersistentMemoryType=AppDirectNotInterleaved
ndctl create-namespace --region region0 --mode fsdax
```

3. Format the persistent memory to be used as a local disk and mount the disk.

```
mkfs -t ext4 /dev/pmem0 && \
mkdir /mnt/sdb && \
mount -o dax,noatime /dev/pmem0 /mnt/sdb
```

4. View the mounted disk.

```
df -h
```

```
[root@i                    ~]# df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/vda1        40G  1.9G   36G   5% /
devtmpfs         16G     0   16G   0% /dev
tmpfs            16G     0   16G   0% /dev/shm
tmpfs            16G  564K   16G   1% /run
tmpfs            16G     0   16G   0% /sys/fs/cgroup
tmpfs           3.1G     0  3.1G   0% /run/user/0
/dev/pmem0      122G   61M  116G   1% /mnt/pmem0
```

After the disks are mounted, you can use disk performance test tools to test their performance.

The following table describes the performance comparison between local NVMe SSDs, enhanced SSDs (ESSDs), and persistent memory that is used as local SSDs.

> **Note** The performance data in the following table is for reference only. Data in the results of your own tests prevails.

Metric	Persistent memory of 128 GiB	NVMe SSD of 1,788 GiB	ESSD of 800 GiB at performance level 1 (PL1)
Read bandwidth	8 to 10 GB/s	2 to 3 GB/s	0.2 to 0.3 GB/s
Read/write bandwidth	8 to 10 GB/s	1 to 2 GB/s	0.2 to 0.3 GB/s
Write bandwidth	2 to 3 GB/s	1 to 2 GB/s	0.2 to 0.3 GB/s
Read IOPS	1,000,000	500,000	20,000 to 30,000
Read/write IOPS	1,000,000	300,000	20,000 to 30,000
Write IOPS	1,000,000	300,000	20,000 to 30,000

Metric	Persistent memory of 128 GiB	NVMe SSD of 1,788 GiB	ESSD of 800 GiB at performance level 1 (PL1)
Read latency	300 to 400 nanoseconds	100,000 nanoseconds	250,000 nanoseconds
Write latency	300 to 400 nanoseconds	20,000 nanoseconds	150,000 nanoseconds

4.3.3. Deploy Redis applications on persistent memory-optimized instances

Persistent memory-optimized instances, such as re6p instances, have high CPU-to-memory ratios and can run Redis applications with much lower costs per GiB of memory. This topic describes how to quickly deploy Redis applications on persistent memory-optimized instances. In the examples, some operating systems are used.

Context

The procedures described in this topic are applicable only to the following instance types and image versions:

- Instance types: ecs.re6p-redis.large, ecs.re6p-redis.xlarge, ecs.re6p-redis.2xlarge, ecs.re6p-redis.4xlarge, and ecs.re6p-redis.13xlarge

  > **Note** To use ecs.re6p-redis.4xlarge, submit a ticket.

- Images:
  - Alibaba Cloud Linux 2
  - CentOS 7.6 or later
  - Ubuntu 18.10 or later
  - SUSE Linux 12 SP4 or later

Deploy a Redis application on an instance that runs Alibaba Cloud Linux 2

Alibaba Cloud Linux 2 is tuned for Redis applications. Redis applications deployed on Alibaba Cloud Linux 2 outperform those deployed on community-supported Linux operating systems in overall performance by more than 20%.

The YUM repositories of Redis 6.0.5 and Redis 3.2.12 are built in to Alibaba Cloud Linux 2. You can run the **yum install** command to deploy Redis 6.0.5 and Redis 3.2.12. You can also manually deploy Redis applications of other versions. For more information, see Deploy a Redis application on an instance that runs CentOS and Deploy a Redis application on an instance that runs Ubuntu.

In this example, the following configurations are used:

- Instance type: ecs.re6p-redis.2xlarge
- Image: Alibaba Cloud Linux 2.1903 LTS 64-bit

1. Purchase a persistent memory-optimized instance.

   For more information, see Create an instance by using the wizard. Take note of the following configurations:

   - **Instance Type**: Set Architecture to **x86-Architecture**, set Category to **Memory Optimized**, and then select the ecs.re6p-redis.2xlarge instance type.
   - **Image**: Select Alibaba Cloud Linux 2.1903 LTS 64-bit.

2. Log on to the instance.

   For more information, see Connection methods.

3. Deploy Redis 6.0.5 or Redis 3.2.12.

   - Run the following command to deploy Redis 6.0.5:

     ```
     yum install -y alinux-release-experimentals && \
     yum install -y redis-6.0.5
     ```

   - Run the following command to deploy Redis 3.2.12:

     ```
     yum install -y alinux-release-experimentals && \
     yum install -y redis-3.2.12
     ```

4. Configure network interface controller (NIC) multi-queue.

   NIC multi-queue helps improve the performance of Redis applications.

   ```
   wget https://ecs-image-tools.oss-cn-hangzhou.aliyuncs.com/ecs_mq/ecs_mq_latest.tgz && \
   tar -xzf ecs_mq_latest.tgz && \
   cd ecs_mq && \
   bash install.sh aliyun 2 && \
   systemctl start ecs_mq
   ```

   Note Run a command in the format of `bash install.sh <Operating system name> <Major version of the operating system>`. For example, run the `bash install.sh aliyun 2` command for an Alibaba Cloud Linux 2.1903 image. You must adjust the command parameters based on your image version.

5. Start Redis and configure the default amounts of regular and persistent memory allocated to Redis.

   Sample commands:

   - Run the following command to set the regular memory-to-persistent memory ratio to a recommended value of 1:4:

     ```
     export MEMKIND_DAX_KMEM_NODES=1 && \
     redis-server /etc/redis.conf --port 8369 --memory-alloc-policy ratio --dram-pmem-ratio 1 4 --hashtable-on-dram yes --daemonize yes --logfile /tmp/redis_8369.log --protected-mode no --bind 0.0.0.0
     ```

   - You can also customize a regular memory-to-persistent memory ratio and leave some regular memory available for other applications. For example, you can run the following command to set the regular memory-to-persistent memory ratio to 1:16 and the total amount of memory allocated to Redis to 34 GiB (including 2 GiB of regular memory and 32 GiB of persistent memory):

```
export MEMKIND_DAX_KMEM_NODES=1 && \
redis-server /etc/redis.conf --port 8369 --memory-alloc-policy ratio --dram-pmem-rati
o 1 16 --maxmemory 34G
```

Deploy a Redis application on an instance that runs CentOS

In this example, the following configurations are used:

- Instance type: ecs.re6p-redis.2xlarge
- Image: CentOS 7.6
- Redis: Redis 4.0.14
- memkind: memkind 1.10.1-rc2

> **Note** Resources required to perform some of the following steps need to be downloaded from GibHub at https://github.com/. Make sure that the required resources are downloaded before you proceed to the steps. If an attempt to download a resource from GitHub fails, repeat the corresponding command until the resource is downloaded.

1. Purchase a persistent memory-optimized instance.

   For more information, see Create an instance by using the wizard. Take note of the following configurations:

   - **Instance Type**: Set Architecture to **x86-Architecture**, set Category to **Memory Optimized**, and then select the ecs.re6p-redis.2xlarge instance type.
   - **Image**: Select CentOS 7.6 64-bit.

2. Log on to the instance.

   For more information, see Connection methods.

3. Prepare the compiling environment.

   ```
   export MEMKIND_DAX_KMEM_NODES=1 && \
   yum -y install numactl-devel.x86_64 && \
   yum -y groupinstall 'Development Tools'
   ```

4. Prepare Redis 4.0.14 source code.

   ```
   wget https://github.com/redis-io/redis/archive/4.0.14.tar.gz && \
   wget https://github.com/redis/redis/compare/4.0.14...tieredmemdb:4.0.14-devel.diff -O r
   edis_4.0.14_diff_tieredmemdb.patch && \
   tar xzvf 4.0.14.tar.gz && \
   cd redis-4.0.14 && \
   git apply --ignore-whitespace ../redis_4.0.14_diff_tieredmemdb.patch
   ```

   > **Note** The patch is used to enable persistent memory and varies per Redis version. For more information, see Download patches that enable Redis applications to use persistent memory.

5. Prepare memkind source code.

   memkind is a memory management tool used to allocate and manage persistent memory.

i. Download memkind source code.

```
wget https://github.com/memkind/memkind/archive/v1.10.1-rc2.tar.gz && \
tar xzvf v1.10.1-rc2.tar.gz && \
mv memkind-1.10.1-rc2/* ./deps/memkind
```

ii. (Optional)Adjust Makefile.

> **Note** You can first run the `ldd --version` command to view the version of glibc. If the version of glibc is 2.17 or later, skip the following operations and compile Redis directly.

```
cd ./deps/memkind && \
wget https://github.com/memKeyDB/memKeyDB/wiki/files/0001-Use-secure_getenv-when-possible.patch && \
git apply --ignore-whitespace 0001-Use-secure_getenv-when-possible.patch && \
cd /root/redis-4.0.14
```

> **Note** If the patch cannot be downloaded, run the command that excludes `cd ./deps/memkind && \`

6. Compile and install Redis.

```
make clean && \
make distclean && \
make MALLOC=memkind -j 4 && \
make install
```

7. Configure NIC multi-queue.

   NIC multi-queue helps improve the performance of Redis applications.

```
wget https://ecs-image-tools.oss-cn-hangzhou.aliyuncs.com/ecs_mq/ecs_mq_latest.tgz && \
tar -xzf ecs_mq_latest.tgz && \
cd ecs_mq && \
bash install.sh centos 7 && \
systemctl start ecs_mq
```

> **Note** Run a command in the format of `bash install.sh <Operating system name> <Major version of the operating system>`. For example, run the `bash install.sh centos 7` command for a CentOS 7.6 image. You must adjust the command parameters based on your image version.

8. Start Redis and configure the default amounts of regular memory and persistent memory allocated to Redis.

   Sample commands:

   - Run the following command to set the regular memory-to-persistent memory ratio to a recommended value of 1:4:

```
redis-server /root/redis-4.0.14/redis.conf --port 8369 --memory-alloc-policy ratio --
dram-pmem-ratio 1 4 --hashtable-on-dram yes --daemonize yes --logfile /tmp/redis_8369
.log --protected-mode no --bind 0.0.0.0
```

- You can also customize a regular memory-to-persistent memory ratio and leave some regular memory available for other applications. For example, you can run the following command to set the regular memory-to-persistent memory ratio to 1:16 and the total amount of memory allocated to Redis to 34 GiB (including 2 GiB of regular memory and 32 GiB of persistent memory):

```
redis-server /root/redis-4.0.14/redis.conf --port 8369 --memory-alloc-policy ratio --
dram-pmem-ratio 1 16 --maxmemory 34G
```

Deploy a Redis application on an instance that runs Ubuntu

In this example, the following configurations are used:

- Instance type: ecs.re6p-redis.2xlarge
- Image: Ubuntu 20.04
- Redis: Redis 6.2.5
- memkind: memkind 1.10.1-rc2

> **Note** Resources required to perform some of the following steps need to be downloaded from GibHub at https://github.com/. Make sure that the required resources are downloaded before you proceed to the steps. If an attempt to download a resource from GitHub fails, repeat the corresponding command until the resource is downloaded.

1. Purchase a persistent memory-optimized instance.

   For more information, see Create an instance by using the wizard. Take note of the following configurations:

   - **Instance Type**: Set Architecture to **x86-Architecture**, set Category to **Memory Optimized**, and then select the ecs.re6p-redis.2xlarge instance type.
   - **Image**: Select Ubuntu 20.04 64-bit.

2. Log on to the instance.

   For more information, see Connection methods.

3. Prepare the compiling environment.

```
export MEMKIND_DAX_KMEM_NODES=1 && \
apt update && \
apt -y install git && \
apt install -y libnuma-dev && \
apt install -y numactl
```

4. Prepare Redis 6.2.5 source code.

```
wget https://download.redis.io/releases/redis-6.2.5.tar.gz && \
wget https://github.com/redis/redis/compare/6.2.5...tieredmemdb:6.2.5-devel.diff -O red
is_6.2.5_diff_tieredmemdb.patch && \
tar xzf redis-6.2.5.tar.gz && \
cd redis-6.2.5 && \
git apply --ignore-whitespace ../redis_6.2.5_diff_tieredmemdb.patch
```

> **Note** The patch is used to enable persistent memory and varies per Redis version. For more information, see Download patches that enable Redis applications to use persistent memory.

5. Prepare memkind source code.

   memkind is a memory management tool used to allocate and manage persistent memory.

   i. Download memkind source code.

   ```
   wget https://github.com/memkind/memkind/archive/v1.10.1-rc2.tar.gz && \
   tar xzvf v1.10.1-rc2.tar.gz && \
   mv memkind-1.10.1-rc2/* ./deps/memkind/
   ```

   ii. Adjust Makefile.

   > **Note** You can first run the `ldd --version` command to view the version of glibc. If the version of glibc is 2.17 or later, skip the following operations and compile Redis directly.

   ```
   cd ./deps/memkind && \
   wget --no-check-certificate https://github.com/memKeyDB/memKeyDB/wiki/files/0001-Use-secure_getenv-when-possible.patch && \
   git apply --ignore-whitespace 0001-Use-secure_getenv-when-possible.patch && \
   cd /root/redis-6.2.5
   ```

   > **Note** If the patch cannot be downloaded, run the command that excludes `cd ./deps/memkind && \`

6. Compile and install Redis.

   ```
   make clean && \
   make distclean && \
   make MALLOC=memkind -j 4 && \
   make install
   ```

7. Configure NIC multi-queue.

   NIC multi-queue helps improve the performance of Redis applications.

   ```
   wget https://ecs-image-tools.oss-cn-hangzhou.aliyuncs.com/ecs_mq/ecs_mq_latest.tgz && \
   tar -xzf ecs_mq_latest.tgz && \
   cd ecs_mq && \
   bash install.sh ubuntu 20 && \
   systemctl start ecs_mq
   ```

   > **Note** Run a command in the format of `bash install.sh <Operating system name> <Major version of the operating system>`. For example, run the `bash install.sh ubuntu 20` command for a Ubuntu 20 image. You must adjust the command parameters based on your image version.

8. Start Redis and configure the default amounts of regular memory and persistent memory allocated

to Redis.

Sample commands:

- Run the following command to set the regular memory-to-persistent memory ratio to a recommended value of 1:4:

```
redis-server /root/redis-6.2.5/redis.conf --port 8369 --memory-alloc-policy ratio --dram-pmem-ratio 1 4 --hashtable-on-dram yes --daemonize yes --logfile /tmp/redis_8369.log --protected-mode no --bind 0.0.0.0
```

- You can also customize a regular memory-to-persistent memory ratio and leave some regular memory available for other applications. For example, you can run the following command to set the regular memory-to-persistent memory ratio to 1:16 and the total amount of memory allocated to Redis to 34 GiB (including 2 GiB of regular memory and 32 GiB of persistent memory):

```
redis-server /root/redis-6.2.5/redis.conf --port 8369 --memory-alloc-policy ratio --dram-pmem-ratio 1 16 --maxmemory 34G
```

Download patches that enable Redis applications to use persistent memory

In the sample command, replace the download URL and the version number that corresponds to the file name. For example, run the following command to download a patch suitable for Redis 6.2.5:

```
wget https://github.com/redis/redis/compare/6.2.5...tieredmemdb:6.2.5-devel.diff -O redis_6.2.5_diff_tieredmemdb.patch
```

The following section lists the download URLs for supported patches:

- Redis 6.0
  - https://github.com/redis/redis/compare/6.0.9...tieredmemdb:6.0.9-devel.diff
  - https://github.com/redis/redis/compare/6.0.5...tieredmemdb:6.0.5-devel.diff
  - https://github.com/redis/redis/compare/6.0.3...tieredmemdb:6.0.3-devel.diff
  - https://github.com/redis/redis/compare/6.0.0...tieredmemdb:6.0.0-devel.diff
- Redis 5.0
  - https://github.com/redis/redis/compare/5.0.9...tieredmemdb:5.0.9-devel.diff
  - https://github.com/redis/redis/compare/5.0.2...tieredmemdb:5.0.2-devel.diff
  - https://github.com/redis/redis/compare/5.0.0...tieredmemdb:5.0.0-devel.diff
- Redis 4.0
  - https://github.com/redis/redis/compare/4.0.14...tieredmemdb:4.0.14-devel.diff
  - https://github.com/redis/redis/compare/4.0.9...tieredmemdb:4.0.9-devel.diff
  - https://github.com/redis/redis/compare/4.0.2...tieredmemdb:4.0.2-devel.diff
  - https://github.com/redis/redis/compare/4.0.0...tieredmemdb:4.0.0-devel.diff
- Redis 3.0
  - https://github.com/redis/redis/compare/3.2.12...tieredmemdb:3.2.diff

> **Note** If you want to use other versions of Redis, submit a ticket.

4.4. Big data instance families

This topic describes the features of big data instance families of Elastic Compute Service (ECS) and lists the instance types of each instance family.

- Recommended instance families
  - d3c, compute-intensive big data instance family
  - d2c, compute-intensive big data instance family
  - d2s, storage-intensive big data instance family
  - d1ne, network-enhanced big data instance family
- Other available instance families (If these instance families are sold out, you can use the recommended ones.)
  - d1, big data instance family

Description

Big data instance families are designed to provide cloud computing and big data storage to support the needs of big data-oriented enterprises. These instance families are suitable for scenarios that require offline computing and big data storage, such as Hadoop distributed computing, extensive log processing, and large-scale data warehousing. Big data instance families are ideal for business that uses distributed networks and has high requirements on storage, capacity, and internal bandwidth.

These instance families are suitable for customers in industries such as Internet and finance that need to compute, store, and analyze big data. Big data instance families use local storage to ensure large amounts of storage space and high storage performance.

Big data instances have the following benefits:

- Enterprise-level computing power ensures efficient and stable data processing.
- Network performance is enhanced with higher maximum internal bandwidth per instance and higher maximum packet forwarding rates to satisfy data transfer demands such as shuffling in Hadoop MapReduce at peak times.
- When an instance is created or started for the first time, its disks must be pre-warmed before they can achieve optimal performance. Each disk can deliver sequential read and write performance of up to 190 MB/s, and each instance can deliver a storage throughput of up to 5 GB/s. This reduces the amount of time required to read data from or write data to Hadoop Distributed File System (HDFS) files.
- The cost of local storage is 97% lower than that of standard SSDs. This significantly reduces the cost to build Hadoop clusters.

When you use big data instances, take note of the following items:

Best practices for mounting a file system to a big data instance

The first time you mount a file system such as ext4, you must initialize the inode table. By default, the lazyinit feature is enabled in Linux kernel v2.6.37 and later, which causes the inode table not to be initialized until file systems are mounted. In addition, local disks consume a large amount of throughput when they are being initialized, such as 600 MB/s for 30 local disks. This may affect service stability. The concurrency of lazyinit in Linux kernel v4.x is improved to resolve this problem. For more information, see index: kernel/git/stable/linux.git. We recommend that you use the following best practices to initialize the inode table at your earliest opportunity:

1. Obtain a list of all local serial advanced technology attachment (SATA) HDDs.
2. Run the following command to initialize each local disk separately.

   In this example, an ext4 file system is created on a local disk whose device name is */dev/vdb*.

   ```
   mkfs.ext4 -E lazy_itable_init=0,lazy_journal_init=0 /dev/vdb &
   ```

3. After all local disks are initialized, run the **iostat -x 5** command until the I/O activities of all local disks are displayed as 0.
4. Batch run the mount command.

d3c, compute-intensive big data instance family

> **Note** This instance family is in invitational preview. To use this instance family,.

Features:

- This instance family is equipped with high-capacity and high-throughput local SSDs and can provide maximum bandwidth of 32 Gbit/s between instances.
- Supports online replacement and hot swapping of damaged disks to prevent instance shutdown.

  If a local disk fails, you receive a notification about the system event. You can handle the system event by initiating the process of fixing the damaged disk. For more information, see O&M scenarios and system events for instances equipped with local disks.

  > **Notice** After you initiate the process of fixing the damaged disk, data in the damaged disk cannot be restored.

- Compute:
  - Uses the third-generation 2.7 GHz Intel® Xeon® Scalable (Ice Lake) processors that deliver an all-core turbo frequency of 3.5 GHz for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports enhanced SSDs (ESSDs), standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Big data computing and storage business scenarios in which services such as Hadoop MapReduce, HDFS, Hive, and HBase are used

- Scenarios in which EMR JindoFS and Operation Orchestration Service (OOS) are used in combination to separately store hot and cold data and decouple storage from computing
- Machine learning scenarios such as Spark in-memory computing and MLlib
- Search and log data processing scenarios in which solutions such as Elasticsearch and Kafka are used

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d3c.3xlarge	14	56.0	1 × 16000	8/burstable up to 10	1,600,000	8	8	30
ecs.d3c.7xlarge	28	112.0	2 × 16000	16/burstable up to 25	2,500,000	16	8	30
ecs.d3c.14xlarge	56	224.0	4 × 16000	32/none	5,000,000	28	8	30
ecs.d3c.16xlarge	64	256.0	4 × 16000	32/none	5,000,000	32	8	30

Note

-
- For more information about these specifications, see Instance family.

d2c, compute-intensive big data instance family

Features:

- This instance family is equipped with high-capacity and high-throughput local SATA HDDs and can provide a maximum bandwidth of 35 Gbit/s between instances.
- Supports online replacement and hot swapping of damaged disks to prevent instance shutdown.

  If a local disk fails, you receive a notification about the system event. You can handle the system event by initiating the process of fixing the damaged disk. For more information, see O&M scenarios and system events for instances equipped with local disks.

  Notice After you initiate the process of fixing the damaged disk, data in the damaged disk cannot be restored.

- Compute:
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors.
- Storage:

- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.

- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Big data computing and storage business scenarios in which services such as Hadoop MapReduce, HDFS, Hive, and HBase are used
  - Scenarios in which EMR JindoFS and OOS are used in combination to separately store hot and cold data and decouple storage from computing
  - Machine learning scenarios such as Spark in-memory computing and MLlib
  - Search and log data processing scenarios in which solutions such as Elasticsearch and Kafka are used

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d2c.6xlarge	24	88.0	3 × 4000	12.0	1,600,000	8	8	20
ecs.d2c.12xlarge	48	176.0	6 × 4000	20.0	2,000,000	16	8	20
ecs.d2c.24xlarge	96	352.0	12 × 4000	35.0	4,500,000	16	8	20

Note

-
- For more information about these specifications, see Instance family.

d2s, storage-intensive big data instance family

Features:

- This instance family is equipped with high-capacity and high-throughput local SATA HDDs and can provide a maximum bandwidth of 35 Gbit/s between instances.
- Supports online replacement and hot swapping of damaged disks to prevent instance shutdown.

  If a local disk fails, you receive a notification about the system event. You can handle the system event by initiating the process of fixing the damaged disk. For more information, see O&M scenarios and system events for instances equipped with local disks.

> **Notice** After you initiate the process of fixing the damaged disk, data in the damaged disk cannot be restored.

- Compute:
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Big data computing and storage business scenarios in which services such as Hadoop MapReduce, HDFS, Hive, and HBase are used
  - Machine learning scenarios such as Spark in-memory computing and MLlib
  - Search and log data processing scenarios in which solutions such as Elasticsearch and Kafka are used

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d2s.5xlarge	20	88.0	8 × 7300	12.0	1,600,000	8	8	20
ecs.d2s.10xlarge	40	176.0	15 × 7300	20.0	2,000,000	16	8	20
ecs.d2s.20xlarge	80	352.0	30 × 7300	35.0	4,500,000	32	8	20

> **Note**
>
> -
> - For more information about these specifications, see Instance family.

d1ne, network-enhanced big data instance family

Features:

- This instance family is equipped with high-capacity and high-throughput local SATA HDDs and can provide a maximum bandwidth of 35 Gbit/s between instances.
- Compute:

- Offers a CPU-to-memory ratio of 1:4, which is designed for big data scenarios.
- Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios in which services such as Hadoop MapReduce, HDFS, Hive, and HBase are used
  - Machine learning scenarios such as Spark in-memory computing and MLlib
  - Search and log data processing scenarios in which solutions such as Elasticsearch are used

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d1ne.2xlarge	8	32.0	4 × 5500	6.0	1,000,000	4	4	10
ecs.d1ne.4xlarge	16	64.0	8 × 5500	12.0	1,600,000	4	8	20
ecs.d1ne.6xlarge	24	96.0	12 × 5500	16.0	2,000,000	6	8	20
ecs.d1ne-c8d3.8xlarge	32	128.0	12 × 5500	20.0	2,000,000	6	8	20
ecs.d1ne.8xlarge	32	128.0	16 × 5500	20.0	2,500,000	8	8	20
ecs.d1ne-c14d3.14xlarge	56	160.0	12 × 5500	35.0	4,500,000	14	8	20

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d1ne.14xlarge	56	224.0	28 × 5500	35.0	4,500,000	14	8	20

Note

-
- For more information about these specifications, see Instance family.

d1, big data instance family

Features:

- This instance family is equipped with high-capacity and high-throughput local SATA HDDs and can provide a maximum bandwidth of 17 Gbit/s between instances.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4, which is designed for big data scenarios.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Scenarios in which services such as Hadoop MapReduce, HDFS, Hive, and HBase are used
  - Machine learning scenarios such as Spark in-memory computing and MLlib
  - Scenarios in which customers in industries such as Internet and finance need to compute, store, and analyze big data
  - Search and log data processing scenarios in which solutions such as Elasticsearch are used

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d1.2xlarge	8	32.0	4 × 5500	3.0	300,000	1	4	10

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.d1.3xlarge	12	48.0	6 × 5500	4.0	400,000	1	6	10
ecs.d1.4xlarge	16	64.0	8 × 5500	6.0	600,000	2	8	20
ecs.d1.6xlarge	24	96.0	12 × 5500	8.0	800,000	2	8	20
ecs.d1-c8d3.8xlarge	32	128.0	12 × 5500	10.0	1,000,000	4	8	20
ecs.d1.8xlarge	32	128.0	16 × 5500	10.0	1,000,000	4	8	20
ecs.d1-c14d3.14xlarge	56	160.0	12 × 5500	17.0	1,800,000	6	8	20
ecs.d1.14xlarge	56	224.0	28 × 5500	17.0	1,800,000	6	8	20

Note

-
- For more information about these specifications, see Instance family.

4.5. Instance families with local SSDs

This topic describes the features of Elastic Compute Service (ECS) instance families with local SSDs and lists the instance types of each family.

- Recommended instance families
  - i3g, instance family with local SSDs
  - i3, instance family with local SSDs
  - i2, instance family with local SSDs
  - i2g, instance family with local SSDs
  - i2ne, instance family with local SSDs
  - i2gne, instance family with local SSDs
- Other available instance families (If these instance families are sold out, you can use the recommended ones.)

  i1, instance family with local SSDs

Overview

Instances with local SSDs provide high I/O performance. They are suitable for scenarios that place high demands on storage I/O performance and require a high availability architecture at the application layer. For example, they are suitable for NoSQL databases, massively parallel processing (MPP) data warehouses, and distributed file systems.

Instances with local SSDs are suitable for enterprises that provide online services such as online gaming, e-commerce, live video streaming, and media. These instances can satisfy the requirements that I/O-intensive applications have for low latency and high I/O performance of Elastic Block Storage (EBS) devices.

Instances with local SSDs have the following features:

- Deliver up to hundreds of thousands of low-latency random read/write IOPS for large databases.
- Offer a maximum sequential read/write throughput of several gibibytes per second in big data, parallel computing, and other large dataset scenarios.
- Use local Non-Volatile Memory Express (NVMe) SSDs to deliver hundreds of thousands of random read/write IOPS with single-digit microsecond latency.

When you use instances with local SSDs, take note of the following items:

- Instances with local SSDs do not support instance configuration changes or failovers.
- Local disks can be tied only to specific instance types. The number and capacity of local disks attached to an instance vary based on the instance type. You cannot separately purchase local disks, or detach local disks from the associated instances and then attach the disks to other instances.
- You cannot create snapshots for local disks. If you want to create an image from the system disk and data disks of an instance with local SSDs, we recommend that you create an image by combining the snapshots of both the system disk and data disks. In this case, the data disks must be cloud disks.
- You cannot create images that contain snapshots of system disks and data disks based on instance IDs.
- You can attach a standard SSD to an instance with local SSDs and extend the capacity of the standard SSD.
-
- Operations on an instance with local SSDs may affect the data stored on the local SSDs. For more information, see Impacts of instance operations on data stored on local disks.

i3g, instance family with local SSDs

Features:

- This instance family is attached with high-performance local NVMe SSDs that deliver high IOPS, high I/O throughput, and low latency.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4, which is designed for high-performance databases.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.2 GHz for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only enhanced SSDs (ESSDs).

- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Applicable scenarios:
  - Online transaction processing (OLTP) and high-performance relational databases
  - NoSQL databases such as Cassandra, MongoDB, and HBase
  - Search scenarios that use solutions such as Elasticsearch

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GB)	Baseline bandwidth (Gbit/s)	Burstable bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.i3g.2xlarge	8	32	1 × 480	3	10	1,750,000	250,000	8	4	15	52,500	2
ecs.i3g.4xlarge	16	64	1 × 960	5	10	3,500,000	300,000	8	8	15	84,000	3
ecs.i3g.8xlarge	32	128	2 × 960	12	None	7,000,000	600,000	8	8	30	157,500	5
ecs.i3g.13xlarge	52	192	3 × 960	16	None	12,000,000	900,000	16	8	30	252,000	8
ecs.i3g.26xlarge	104	384	6 × 960	32	None	24,000,000	1,800,000	32	15	30	500,000	16

Note

-
- This instance family supports only Linux images. When you create an instance of this instance family, you must select a Linux image.
- For more information about these specifications, see Instance family.
- For more information about the performance metrics of local SSDs, see Local disks.

i3, instance family with local SSDs

Features

- This instance family is attached with high-performance local NVMe SSDs that deliver high IOPS, high I/O throughput, and low latency, and allows damaged disks to be isolated online.
- Compute:
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.2 GHz for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Applicable scenarios:
  - OLTP and high-performance relational databases
  - NoSQL databases such as Cassandra and MongoDB
  - Search scenarios that use solutions such as Elasticsearch

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GB)	Baseline bandwidth (Gbit/s)	Burstable bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.i3.xlarge	4	32	1 × 960	1.5	10	1,000,000	250,000	4	4	15	40,000	1.5
ecs.i3.2xlarge	8	64	1 × 1920	2.5	10	1,600,000	250,000	8	4	15	50,000	2

Instance type	vCPUs	Memory (GiB)	Local storage (GB)	Baseline bandwidth (Gbit/s)	Burstable bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.i3.4xlarge	16	128	2 × 1920	5	10	3,000,000	300,000	8	8	30	80,000	3
ecs.i3.8xlarge	32	256	4 × 1920	10	None	6,000,000	600,000	16	8	30	150,000	5
ecs.i3.13xlarge	52	384	6 × 1920	16	None	9,000,000	900,000	32	7	30	240,000	8
ecs.i3.26xlarge	104	768	12 × 1920	32	None	24,000,000	1,800,000	32	15	30	480,000	16

Note

-
- This instance family supports only Linux images. When you create an instance of this instance family, you must select a Linux image.
- For more information about these specifications, see Instance family.
- For more information about the performance metrics of local SSDs, see Local disks.

i2, instance family with local SSDs

Features:

- This instance family is attached with high-performance local NVMe SSDs that deliver high IOPS, high I/O throughput, and low latency.
- Compute:
  - Offers a CPU-to-memory ratio of 1:8, which is designed for high-performance databases.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.

- Provides high network performance based on large computing capacity.

- Applicable scenarios:
  - OLTP and high-performance relational databases
  - NoSQL databases such as Cassandra, MongoDB, and HBase
  - Search scenarios that use solutions such as Elasticsearch

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI	Disk bandwidth (Gbit/s)
ecs.i2.xlarge	4	32	1 × 894	1	500,000	2	3	10	Up to 16
ecs.i2.2xlarge	8	64	1 × 1788	2	1,000,000	2	4	10	Up to 16
ecs.i2.4xlarge	16	128	2 × 1788	3	1,500,000	4	8	20	Up to 16
ecs.i2.8xlarge	32	256	4 × 1788	6	2,000,000	8	8	20	Up to 16
ecs.i2.16xlarge	64	512	8 × 1788	10	4,000,000	16	8	20	Up to 16
ecs.i2d.11xlarge	46	356	2 × 3570	10	2,000,000	16	8	20	Up to 16
ecs.i2d.21xlarge	84	712	4 × 3570	25	4,000,000	32	16	20	Up to 16
ecs.i2d.23xlarge	92	712	4 × 3570	25	4,000,000	32	15	20	Up to 16

> **Note**
>
> -
> - For more information about these specifications, see Instance family.
> - For more information about the performance metrics of local SSDs, see Local disks.

i2g, instance family with local SSDs

Features:

- This instance family is attached with high-performance local NVMe SSDs that deliver high IOPS, high

I/O throughput, and low latency.

- Compute:
  - Offers a CPU-to-memory ratio of 1:4, which is designed for high-performance databases.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large computing capacity.
- Applicable scenarios:
  - OLTP and high-performance relational databases
  - NoSQL databases such as Cassandra, MongoDB, and HBase
  - Search scenarios that use solutions such as Elasticsearch

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.i2g.2xlarge	8	32	1 × 894	2	1,000,000	2	4	10
ecs.i2g.4xlarge	16	64	1 × 1788	3	1,500,000	4	8	20
ecs.i2g.8xlarge	32	128	2 × 1788	6	2,000,000	8	8	20
ecs.i2g.16xlarge	64	256	4 × 1788	10	4,000,000	16	8	20

Note

-
- For more information about these specifications, see Instance family.
- For more information about the performance metrics of local SSDs, see Local disks.

i2ne, instance family with local SSDs

Features:

- This instance family is attached with high-performance local NVMe SSDs that deliver high IOPS, high I/O throughput, and low latency.
- Compute:

- Offers a CPU-to-memory ratio of 1:8, which is designed for high-performance databases.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
  - Provides a bandwidth of up to 20 Gbit/s.
- Applicable scenarios:
  - OLTP and high-performance relational databases
  - NoSQL databases such as Cassandra, MongoDB, and HBase
  - Search scenarios that use solutions such as Elasticsearch

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI	Disk bandwidth (Gbit/s)
ecs.i2ne.xlarge	4	32	1 × 894	1.5	500,000	2	3	10	Up to 16
ecs.i2ne.2xlarge	8	64	1 × 1788	2.5	1,000,000	2	4	10	Up to 16
ecs.i2ne.4xlarge	16	128	2 × 1788	5	1,500,000	4	8	20	Up to 16
ecs.i2ne.8xlarge	32	256	4 × 1788	10	2,000,000	8	8	20	Up to 16
ecs.i2ne.16xlarge	64	512	8 × 1788	20	4,000,000	16	8	20	Up to 16

Note

-
- For more information about these specifications, see Instance family.
- For more information about the performance metrics of local SSDs, see Local disks.

i2gne, instance family with local SSDs

Features:

- This instance family is attached with high-performance local NVMe SSDs that deliver high IOPS, high I/O throughput, and low latency.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4, which is designed for high-performance databases.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large computing capacity.
  - Provides a bandwidth of up to 20 Gbit/s.
- Applicable scenarios:
  - OLTP and high-performance relational databases
  - NoSQL databases such as Cassandra, MongoDB, and HBase
  - Search scenarios that use solutions such as Elasticsearch

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.i2gne.2xlarge	8	32	1 × 894	2.5	1,000,000	2	4	10
ecs.i2gne.4xlarge	16	64	1 × 1788	5	1,500,000	4	8	20
ecs.i2gne.8xlarge	32	128	2 × 1788	10	2,000,000	8	8	20
ecs.i2gne.16xlarge	64	256	4 × 1788	20	4,000,000	16	8	20

Note

-
- For more information about these specifications, see Instance family.
- For more information about the performance metrics of local SSDs, see Local disks.

i1, instance family with local SSDs

Features:

- This instance family is attached with high-performance local NVMe SSDs that deliver high IOPS, high I/O throughput, and low latency.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4, which is designed for high-performance databases.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large computing capacity.
- Applicable scenarios:
  - OLTP and high-performance relational databases
  - NoSQL databases such as Cassandra and MongoDB
  - Search scenarios that use solutions such as Elasticsearch

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.i1.xlarge	4	16	2 × 104	0.8	200,000	1	3	10
ecs.i1.2xlarge	8	32	2 × 208	1.5	400,000	1	4	10
ecs.i1.3xlarge	12	48	2 × 312	2	400,000	1	6	10
ecs.i1.4xlarge	16	64	2 × 416	3	500,000	2	8	20
ecs.i1-c5d1.4xlarge	16	64	2 × 1456	3	400,000	2	8	20

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.i1.6xlarge	24	96	2 × 624	4.5	600,000	2	8	20
ecs.i1.8xlarge	32	128	2 × 832	6	800,000	3	8	20
ecs.i1-c10d1.8xlarge	32	128	2 × 1456	6	800,000	3	8	20
ecs.i1.14xlarge	56	224	2 × 1456	10	1,200,000	4	8	20

Note

-
- For more information about these specifications, see Instance family.
- For more information about the performance metrics of local SSDs, see Local disks.

4.6. Instance families with high clock speeds

This topic describes the features of instance families with high clock speeds and lists the instance types of each instance family.

- Recommended instance families
  - hfc7, compute-optimized instance family with high clock speeds
  - hfc6, compute-optimized instance family with high clock speeds
  - hfg7, general-purpose instance family with high clock speeds
  - hfg6, general-purpose instance family with high clock speeds
  - hfr7, memory-optimized instance family with high clock speeds
  - hfr6, memory-optimized instance family with high clock speeds
- Other available instance families
  - hfc5, compute-optimized instance family with high clock speeds
  - hfg5, general-purpose instance family with high clock speeds

hfc7, compute-optimized instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of third-generation SHENLONG architecture to provide predictable and consistent ultra-high

performance and reduce virtualization overheads.

- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses Intel® Xeon® Cooper Lake processors that deliver an all-core turbo frequency of 3.8 GHz and have a minimum clock speed of 3.3 GHz for consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only enhanced SSDs (ESSDs) and provides ultra-high I/O performance.
  - Provides high storage I/O performance based on large compute capacity.

    Note

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance frontend server clusters
  - Frontend servers of massive multiplayer online (MMO) games
  - Data analysis, batch processing, and video encoding
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfc7.large	2	4	1.2/10	900,000	250,000	2	2	6	20,000	1
ecs.hfc7.xlarge	4	8	2/10	1,000,000	250,000	4	3	15	30,000	1.5

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfc7.2xlarge	8	16	3/10	1,600,000	250,000	8	4	15	45,000	2
ecs.hfc7.3xlarge	12	24	4.5/10	2,000,000	250,000	8	6	15	60,000	2.5
ecs.hfc7.4xlarge	16	32	6/10	2,500,000	300,000	8	8	30	75,000	3
ecs.hfc7.6xlarge	24	48	8/10	3,000,000	450,000	12	8	30	90,000	4
ecs.hfc7.8xlarge	32	64	10/none	4,000,000	600,000	16	8	30	105,000	5
ecs.hfc7.12xlarge	48	96	16/none	6,000,000	1,000,000	24	8	30	150,000	8
ecs.hfc7.24xlarge	96	192	32/none	12,000,000	1,800,000	32	15	30	300,000	16

Note

-
- For more information about these specifications, see Instance family.

hfc6, compute-optimized instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2.

- Uses 3.1 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.5 GHz for consistent computing performance.

  > **Note** The processors used by this instance family have a clock speed of 3.1 GHz. However, the Intel System Studio (ISS) feature may cause a lower clock speed to be displayed. Alibaba Cloud is working on this issue. This issue does not affect the actual clock speeds of your instances.
  >
  > You can separately run the following commands to use the turbostat tool to view the actual clock speeds:
  >
  > ```
  > yum install kernel-tools
  > ```
  >
  > ```
  > turbostat
  > ```

- Allows you to enable or disable Hyper-Threading.

  > **Note**

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Provides high storage I/O performance based on large compute capacity.

  > **Note**

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Web frontend servers
  - Frontend servers of MMO games
  - Data analysis, batch processing, and video encoding
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfc6.large	2	4	1/3	300,000	35,000	2	2	6	10,000	1
ecs.hfc6.xlarge	4	8	1.5/5	500,000	70,000	4	3	10	20,000	1.5
ecs.hfc6.2xlarge	8	16	2.5/8	800,000	150,000	8	4	10	25,000	2
ecs.hfc6.3xlarge	12	24	4/10	900,000	220,000	8	6	10	30,000	2.5
ecs.hfc6.4xlarge	16	32	5/10	1,000,000	300,000	8	8	20	40,000	3
ecs.hfc6.6xlarge	24	48	7.5/10	1,500,000	450,000	12	8	20	50,000	4
ecs.hfc6.8xlarge	32	64	10/none	2,000,000	600,000	16	8	20	60,000	5
ecs.hfc6.10xlarge	40	96	12.5/none	3,000,000	1,000,000	32	7	20	100,000	8
ecs.hfc6.16xlarge	64	128	20/none	4,000,000	1,200,000	32	8	20	120,000	10
ecs.hfc6.20xlarge	80	192	25/none	6,000,000	1,800,000	32	15	20	200,000	16

Note

-
- For more information about these specifications, see Instance family.

hfg7, general-purpose instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses Intel® Xeon® Cooper Lake processors that deliver an all-core turbo frequency of 3.8 GHz and have a minimum clock speed of 3.3 GHz for consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs and provides ultra-high I/O performance.
  - Provides high storage I/O performance based on large compute capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Enterprise-grade applications of various types and sizes
  - Game servers
  - Small and medium-sized database systems, caches, and search clusters
  - High-performance scientific computing
  - Video encoding applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfg7.large	2	8	1.2/10	900,000	250,000	2	2	6	20,000	1
ecs.hfg7.xlarge	4	16	2/10	1,000,000	250,000	4	3	15	30,000	1.5
ecs.hfg7.2xlarge	8	32	3/10	1,600,000	250,000	8	4	15	45,000	2
ecs.hfg7.3xlarge	12	48	4.5/10	2,000,000	250,000	8	6	15	60,000	2.5
ecs.hfg7.4xlarge	16	64	6/10	2,500,000	300,000	8	8	30	75,000	3
ecs.hfg7.6xlarge	24	96	8/10	3,000,000	450,000	12	8	30	90,000	4
ecs.hfg7.8xlarge	32	128	10/none	4,000,000	600,000	16	8	30	105,000	5
ecs.hfg7.12xlarge	48	192	16/none	6,000,000	1,000,000	24	8	30	150,000	8
ecs.hfg7.24xlarge	96	384	32/none	12,000,000	1,800,000	32	15	30	300,000	16

Note

-
- For more information about these specifications, see Instance family.

hfg6, general-purpose instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with

the use of SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.

- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 3.1 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.5 GHz for consistent computing performance.

    > **Note** The processors used by this instance family have a clock speed of 3.1 GHz. However, the Intel System Studio (ISS) feature may cause a lower clock speed to be displayed. Alibaba Cloud is working on this issue. This issue does not affect the actual clock speeds of your instances.
    >
    > You can separately run the following commands to use the turbostat tool to view the actual clock speeds:
    >
    > ```
    > yum install kernel-tools
    > ```
    >
    > ```
    > turbostat
    > ```

  - Allows you to enable or disable Hyper-Threading.

    > **Note**

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Provides high storage I/O performance based on large compute capacity.

    > **Note**

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Enterprise-grade applications of various types and sizes
  - Websites and application servers
  - Game servers
  - Small and medium-sized database systems, caches, and search clusters
  - Data analysis and computing
  - Computing clusters and memory-intensive data processing

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfg6.large	2	8	1/3	300,000	35,000	2	2	6	10,000	1
ecs.hfg6.xlarge	4	16	1.5/5	500,000	70,000	4	3	10	20,000	1.5
ecs.hfg6.2xlarge	8	32	2.5/8	800,000	150,000	8	4	10	25,000	2
ecs.hfg6.3xlarge	12	48	4/10	900,000	220,000	8	6	10	30,000	2.5
ecs.hfg6.4xlarge	16	64	5/10	1,000,000	300,000	8	8	20	40,000	3
ecs.hfg6.6xlarge	24	96	7.5/10	1,500,000	450,000	12	8	20	50,000	4
ecs.hfg6.8xlarge	32	128	10/none	2,000,000	600,000	16	8	20	60,000	5
ecs.hfg6.10xlarge	40	192	12.5/none	3,000,000	1,000,000	32	7	20	100,000	8
ecs.hfg6.16xlarge	64	256	20/none	4,000,000	1,200,000	32	8	20	120,000	10
ecs.hfg6.20xlarge	80	384	25/none	6,000,000	1,800,000	32	15	20	200,000	16

Note

-
- For more information about these specifications, see Instance family.

hfr7, memory-optimized instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses Intel® Xeon® Cooper Lake processors that deliver an all-core turbo frequency of 3.8 GHz and have a minimum clock speed of 3.3 GHz for consistent computing performance.
  - Allows you to enable or disable Hyper-Threading.

    Note
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs and provides ultra-high I/O performance.
  - Provides high storage I/O performance based on large compute capacity.

    Note
- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance databases and in-memory databases
  - Data analysis, data mining, and distributed memory caching
  - Hadoop clusters, Spark clusters, and other enterprise-level memory-intensive applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfr7.large	2	16	1.2/10	900,000	250,000	2	2	6	20,000	1

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfr7.xlarge	4	32	2/10	1,000,000	250,000	4	3	15	30,000	1.5
ecs.hfr7.2xlarge	8	64	3/10	1,600,000	250,000	8	4	15	45,000	2
ecs.hfr7.3xlarge	12	96	4.5/10	2,000,000	250,000	8	6	15	60,000	2.5
ecs.hfr7.4xlarge	16	128	6/10	2,500,000	300,000	8	8	30	75,000	3
ecs.hfr7.6xlarge	24	192	8/10	3,000,000	450,000	12	8	30	90,000	4
ecs.hfr7.8xlarge	32	256	10/none	4,000,000	600,000	16	8	30	105,000	5
ecs.hfr7.12xlarge	48	384	16/none	6,000,000	1,000,000	24	8	30	150,000	8
ecs.hfr7.24xlarge	96	768	32/none	12,000,000	1,800,000	32	15	30	300,000	16

Note

-
- For more information about these specifications, see Instance family.

hfr6, memory-optimized instance family with high clock speeds

Features

- This instance family offloads a large number of virtualization features to dedicated hardware with the use of SHENLONG architecture to provide predictable and consistent ultra-high performance and reduce virtualization overheads.
- Compute:

- Offers a CPU-to-memory ratio of 1:8.
- Uses 3.1 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver a turbo frequency of 3.5 GHz for consistent computing performance.

  > **Note** The processors used by this instance family have a clock speed of 3.1 GHz. However, the Intel System Studio (ISS) feature may cause a lower clock speed to be displayed. Alibaba Cloud is working on this issue. This issue does not affect the actual clock speeds of your instances.
  >
  > You can separately run the following commands to use the turbostat tool to view the actual clock speeds:
  >
  > ```
  > yum install kernel-tools
  > ```
  >
  > ```
  > turbostat
  > ```

- Allows you to enable or disable Hyper-Threading.

  > **Note**

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Provides high storage I/O performance based on large compute capacity.

    > **Note**

- Network:
  - Supports IPv6.
  - Provides ultra-high packet forwarding rates.
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance databases and in-memory databases
  - Data analysis, data mining, and distributed memory caching
  - Hadoop clusters, Spark clusters, and other enterprise-level memory-intensive applications

Instance types

Instance type	vCPUs	Memory (GiB)	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.hfr6.large	2	16	1/3	300,000	35,000	2	2	6	10,000	1
ecs.hfr6.xlarge	4	32	1.5/5	500,000	70,000	4	3	10	20,000	1.5
ecs.hfr6.2xlarge	8	64	2.5/8	800,000	150,000	8	4	10	25,000	2
ecs.hfr6.3xlarge	12	96	4/10	900,000	220,000	8	6	10	30,000	2.5
ecs.hfr6.4xlarge	16	128	5/10	1,000,000	300,000	8	8	20	40,000	3
ecs.hfr6.6xlarge	24	192	7.5/10	1,500,000	450,000	12	8	20	50,000	4
ecs.hfr6.8xlarge	32	256	10/none	2,000,000	600,000	16	8	20	60,000	5
ecs.hfr6.10xlarge	40	384	12.5/none	3,000,000	1,000,000	32	7	20	100,000	8
ecs.hfr6.16xlarge	64	512	20/none	4,000,000	1,200,000	32	8	20	120,000	10
ecs.hfr6.20xlarge	80	768	25/none	6,000,000	1,800,000	32	15	20	200,000	16

Note

-
- For more information about these specifications, see Instance family.

hfc5, compute-optimized instance family with high clock speeds

Features

- Compute:
  - Offers a CPU-to-memory ratio of 1:2.
  - Uses 3.1 GHz Intel® Xeon® Gold 6149 (Skylake) processors.
  - Offers consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - High-performance web frontend servers
  - High-performance scientific and engineering applications
  - MMO gaming and video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.hfc5.large	2	4	1	300,000	2	2	6
ecs.hfc5.xlarge	4	8	1.5	500,000	2	3	10
ecs.hfc5.2xlarge	8	16	2	1,000,000	2	4	10
ecs.hfc5.3xlarge	12	24	2.5	1,300,000	4	6	10
ecs.hfc5.4xlarge	16	32	3	1,600,000	4	8	20
ecs.hfc5.6xlarge	24	48	4.5	2,000,000	6	8	20
ecs.hfc5.8xlarge	32	64	6	2,500,000	8	8	20

Note

-
- For more information about these specifications, see Instance family.

hfg5, general-purpose instance family with high clock speeds

Features

- Compute:
  - Offers a CPU-to-memory ratio of 1:4 (excluding the instance type with 56 vCPUs).
  - Uses 3.1 GHz Intel® Xeon® Gold 6149 (Skylake) processors.
  - Offers consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large compute capacity.
- Applicable scenarios:
  - High-performance web frontend servers
  - High-performance scientific and engineering applications
  - MMO gaming and video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.hfg5.large	2	8	1	300,000	2	2	6
ecs.hfg5.xlarge	4	16	1.5	500,000	2	3	10
ecs.hfg5.2xlarge	8	32	2	1,000,000	2	4	10
ecs.hfg5.3xlarge	12	48	2.5	1,300,000	4	6	10
ecs.hfg5.4xlarge	16	64	3	1,600,000	4	8	20
ecs.hfg5.6xlarge	24	96	4.5	2,000,000	6	8	20

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.hfg5.8xlarge	32	128	6	2,500,000	8	8	20
ecs.hfg5.14xlarge	56	160	10	4,000,000	14	8	20

Note

-
- For more information about these specifications, see Instance family.

4.7. Security-enhanced instance family

4.7.1. Overview

Security-enhanced instance families provide the trusted computing capability based on Trusted Cryptography Module (TCM) or Trusted Platform Module (TPM) chips. Some security-enhanced instance families including g7t, c7t, and r7t also provide the Software Guard Extension (SGX) encrypted computing capability based on Intel® SGX for a trusted confidential environment that offers a higher degree of security.

Trusted computing

Security-enhanced instance families integrate TPM or TCM into the hardware platform, use TPM or TCM chips as Root of Trust (RoT), and use the Unified Extensible Firmware Interface (UEFI) firmware, vTPM or vTCM, and remote attestation service to implement instance startup measurement and integrity verification. These ensure that security-enhanced instances are secure and trusted.

For information about the features of the trusted computing capability and how to work with the trusted computing capability, see the following topics:

- Overview
- Use the trusted feature of security-enhanced instances
- Build a confidential computing environment by using Enclave

SGX encrypted computing

When you use RoT that is based on software and the software has security vulnerabilities, the data security cannot be guaranteed. RoT of SGX contains only hardware to improve the security level. In addition to instance startup measurement and integrity verification, some security-enhanced instance families including g7t, c7t, and r7t transfer the confidential computing capability of physical servers to instances based on Intel® SGX. vSGX instances are assigned encrypted memory.

> **Note** If you have stricter security and compliance requirements and do not want to share the physical resources of cloud hosts with other tenants, you can purchase dedicated hosts whose physical resources are reserved for the exclusive use of a single tenant to enhance security. For more information, see What is DDH?

SGX uses Memory Encryption Engines (MEEs) in CPUs to encrypt data in the encrypted memory. Encrypted data is decrypted into plaintext only after the data enters CPUs. CPUs protect your private data from being extracted by malicious code. Therefore, when you use a vSGX instance, data remains protected even if the operating system, virtualization stack, or BIOS becomes compromised. You need only to trust CPUs to keep your private data secure. For information about how to use SGX, see Build an SGX encrypted computing environment.

Support of secure computing capabilities of instance families

Instance family	Trusted computing	SGX encrypted computing
• g7t, security-enhanced general-purpose instance family • c7t, security-enhanced compute-optimized instance family • r7t, security-enhanced memory-optimized instance family	Supported	Supported
• g6t, security-enhanced general-purpose instance family • c6t, security-enhanced compute-optimized instance family	Supported	Not supported

4.7.2. Create security-enhanced instances

When you create a security-enhanced Elastic Compute Service (ECS) instance, you must select a specific operating system. When you use the Alibaba Cloud trusted system, you must also obtain the corresponding permissions so that the security-enhanced instance can report the trusted information to Alibaba Cloud Security Center when the instance starts. This topic describes how to create a security-enhanced instance.

Create a security-enhanced instance in the ECS console

The procedure for creating a security-enhanced instance in the ECS console is similar to that for creating a non-security-enhanced instance. However, you must pay attention to specific options when you create a security-enhanced instance. This procedure describes the specific configurations to make when you create a security-enhanced instance. For information about other general configurations, see Create an instance by using the wizard.

When you create a security-enhanced instance in the ECS console, you are prompted to perform the following operations:

- Activate Key Management Service (KMS). After KMS is activated, a service key is automatically created. You do not need to pay for this key.
- Create a RAM role and grant permissions to this role. Alibaba Cloud provides you with system policies for trusted services. Follow the steps in the wizard to complete the settings when you create an instance.

1.
2.
3. Click **Create Instance**.
4. Configure the settings in the Basic Configurations step.

   Take note of the following parameters:

   - **Instance Type**: Select a security-enhanced instance type. For more information about security-enhanced instance types, see Instance family.

     > **Note** The g7t, c7t, and r7t instance families support Software Guard Extensions (SGX) encrypted computing. When you create a g7t, c7t, or r7t instance in the ECS console, the Alibaba Cloud SGX runtime is automatically installed. For information about how to build an SGX encrypted computing environment on a g7t, c7t, or r7t instance, see Build an SGX encrypted computing environment.

   - **Image**: Select an image version based on the instance family. The following table describes the available image versions.

Instance family	Image version
g6t, c6t, and r6t	■ Alibaba Cloud Linux 2.1903 64-bit (Trusted) ■ CentOS 7.8 64-bit (Trusted)
g7t, c7t, and r7t	■ Alibaba Cloud Linux 2.1903 LTS 64-bit (UEFI) ■ CentOS 8.4 64-bit (UEFI) ■ CentOS 8.3 64-bit (UEFI) ■ Ubuntu 20.04 64-bit (UEFI) ■ Ubuntu 18.04 64-bit (UEFI)

     > **Note** If you select **Trusted System** when you create an instance, the Alibaba Cloud trusted system is used for the instance. The Alibaba Cloud trusted system verifies the instance when the instance starts. If you want to use a self-managed trusted service system, do not select **Trusted System**.

5. Click **Next: Networking**. If the **Enable Key Management Service (KMS)** dialog box appears, click **Enable**.

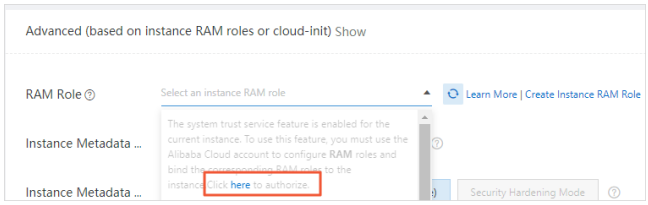
   KMS must be activated. Otherwise, the security-enhanced instance cannot be created. If you have activated KMS, the dialog box does not appear. Proceed with the Networking step.

6. Click **Next: System Configurations**.

   If **Trusted System** is selected, you must specify a RAM role for the instance. The RAM role must be granted permissions to access the trusted services. Alibaba Cloud provides you with the corresponding **AliyunECSInstanceForYundunSysTrustRole** service-linked role. We recommend that you configure and select this role by performing the following steps.

   > **Note** If you need more precise or customized configurations, create a role and grant it permissions based on your needs. When you create a RAM role, you must take some precautions. For more information, see Precautions on granting permissions to RAM roles.

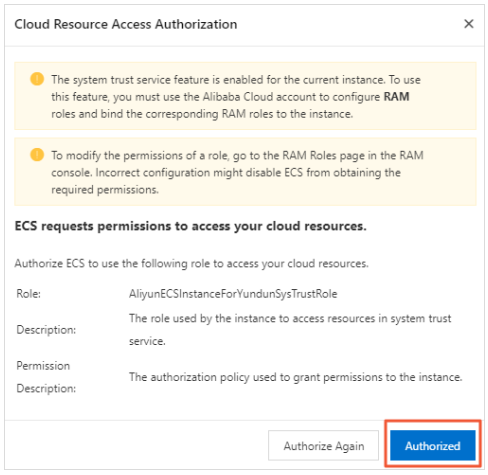
   i. Click **here** to authorize.

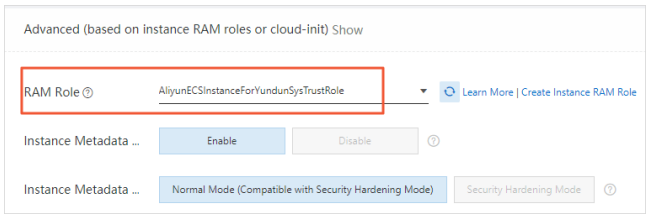
   ii. In the **Cloud Resource Access Authorization** dialog box, click **Confirm Authorization**.

   iii. In the dialog box that appears, click **Confirm Authorization Policy**.

   iv. Click **Authorized**.

   v. Select **AliyunECSInstanceForYundunSysTrustRole** as the RAM role.

   > **Note** You can also skip the authorization step and grant permissions after the instance is created. For more information, see Attach an instance RAM role to an ECS instance.

7. Follow the steps in the wizard to create the instance.

Create a security-enhanced instance by calling an API operation

When you call an API operation to create a security-enhanced instance, take note of the following items:

- KMS must be activated. Otherwise, the security-enhanced instance cannot be created. For more information, see Activate KMS.
- When you use the Alibaba Cloud trusted system, you must specify a RAM role for the security-enhanced instance to be created and this role must be granted permissions to access the trusted services. This way, the security-enhanced instance reports the trusted information to Alibaba Cloud Security Center when the instance starts. You can call an API operation to create a RAM role and grant permissions to this role. For more information, see Use an instance RAM role by calling API operations. When you create a RAM role, you must take some precautions. For more information, see Precautions on granting permissions to RAM roles.

> **Note** If you use a self-managed trusted service system, you do not need to specify the RAM role.

You can call the RunInstances or CreateInstance operation to create security-enhanced instances. The following table describes some parameters to take note of.

Parameter	Description	Example
InstanceType	The instance type of the security-enhanced instance. ECS provides the following security-enhanced instance families: • g7t • c7t • r7t • g6t • c6t	*ecs.c6t.large*
ImageId	The ID of the image that is used to create the security-enhanced instance. You can call the DescribeImages operation to query image IDs.	*aliyun_2_1903_x64_20G_secured_alibase_20210120.vhd*
SystemDisk.Category	The category of the system disk to attach to the security-enhanced instance. Only enhanced SSDs (ESSDs) can be used.	*cloud_essd*
VSwitchId	The ID of the vSwitch of the security-enhanced instance. This parameter is required because all security-enhanced instances reside in virtual private clouds (VPCs).	*vsw-bp134jzf285qg9u6w*****

Parameter	Description	Example
RamRoleName	The name of the RAM role. You can also call the AttachInstanceRamRole operation to attach a RAM role to the instance after the instance is created.	*AliyunECSInstanceForYundunSysTrustRole*
	The installation script used to install the Alibaba Cloud trusted	

Parameter	Description	Example
UserData	system, which must be encoded in Base64. For information about the script content in plaintext before the script is encoded in Base64, see Script for installing the Alibaba Cloud trusted system.	IyEvYmluL3NoCkNVUlBBVEg9Y HB3ZGAKU0NSSVBUX1BBVEg9Ii 9kb3dubG9hZC9saW51eC9zY3J pcHQvVHJ1c3RBZ2VudEluc3Rh bGwuc2giClJFR0lPTl9JRD1gY 3VybCAtcyAtLXJldHJ5IDEgLS 1tYXgtdGltZSAzIGh0dHA6Ly8 xMDAuMTAwLjEwMC4yMDAvbGF0 ZXN0L21ldGEtZGF0YS9yZWdpb 24taWRgClVQREFURV9TSVRFMT 1odHRwOi8vdHJ1c3RjbGllbnQ tJHtSRUdJT05fSUR9Lm9zcy0k e1JFR0lPTl9JRH0taW50ZXJuY WwuYWxpeXVuY3MuY29tClVQRE FURV9TSVRFMj1odHRwOi8vdHJ 1c3RjbGllbnQtJHtSRUdJT05f SUR9Lm9zcy0ke1JFR0lPTl9JR H0uYWxpeXVuY3MuY29tClVQRE FURV9TSVRFMz1odHRwOi8vdC1 0cnVzdGNsaWVudC0ke1JFR0lP Tl9JRH0ub3NzLXskUkVHSU9OX 0lEfS1pbnRlcm5hbC5hbGl5dW 5jcy5jb20KTVNHX0lORk89ImR vd25sb2FkaW5nIGluc3RhbGwg c2NyaXB0IGZyb20gc2l0ZSIKT VNHX0VSUj0iZG93bmxvYWQgZm lsZSBlcnJvci4iCk1TR19PSz0 idHJ1c3QgY2xpZW50IGluaXQg ZG9uZS4iCgppbnN0YWxsKCkKe wogIGVjaG8gIiR7TVNHX0lORk 99IiIgMS4uLiIKICBjdXJsIC1 mc1NMICIke1VQREFURV9TSVRF MX0iIiR7U0NSSVBUX1BBVEh9I nxzaAogIGlmIFsgJD8gPT0gMC BdOyB0aGVuCiAgICByZXR1cm4 gMQogIGZpCiAgZWNobyAiJHtN U0dfSU5GT30iIiAyLi4uIgogI GN1cmwgLWZzU0wgIiR7VVBEQV RFX1NJVEUyfSIiJHtTQ1JJUFR fUEFUSH0ifHNoCiAgaWYgWyAk PyA9PSAwIF07IHRoZW4KICAgI HJldHVybiAyCiAgZmkKICBlY2 hvICIke01TR19JTkZPfSIiIDM uLi4iCiAgY3VybCAtZnNTTCAi JHtVUERBVEVfU0lURTN9IiIke 1NDUklQVF9QQVRIfSJ8c2gKIC BpZiBbICQ/ID09IDAgXTsgdGh lbgogICAgcmV0dXJuIDMKICBm aQogIGVjaG8gIiIgMT4mMgogI GV4aXQgMQp9CgppbnN0YWxsCm VjaG8gIiR7TVNHX09LfSIKCmV 4aXQgMAo=

Parameter	Description	Example
SecurityOptions.TrustedSystemMode	The trusted system mode. When you call the RunInstances operation to create a security-enhanced instance, you must set the `SecurityOptions.TrustedSystemMode` parameter to vTPM if you set InstanceType to g7t, c7t, or r7t. **Note** You can call only the RunInstances operation to create an instance in trusted system mode. If you call the CreateInstance operation, you cannot set the trusted system mode parameter (`SecurityOptions.TrustedSystemMode`).	vTPM

Sample requests:

```
https://ecs.aliyuncs.com/?Action=RunInstances
&RegionId=cn-hangzhou
&InstanceType=ecs.c6t.large
&ImageId=aliyun_2_1903_x64_20G_secured_alibase_20210120.vhd
&SystemDisk.Category=cloud_essd
&VSwitchId=vsw-bp134jzf285qg9u6w****
&SecurityGroupId=sg-bp1c3o8hzd14dovh****
&RamRoleName=AliyunECSInstanceForYundunSysTrustRole
&UserData=IyEvYmluL3NoCkNVUlBBVEg9YHB3ZGAKU0NSSVBUX1BBVEg9Ii9kb3dubG9hZC9saW51eC9zY3JpcHQvV
HJ1c3RBZ2VudEluc3RhbGwuc2giClJFR0lPTl9JRD1gY3VybCAtcyAtLXJldHJ5IDEgLS1tYXgtdGltZSAzIGh0dHA6
Ly8xMDAuMTAwLjEwMC4yMDAvbGF0ZXN0L21ldGEtZGF0YS9yZWdpb24taWRgClVQREFURV9TSVRFMT1odHRwOi8vdHJ
1c3RjbGllbnQtJHtSRUdJT05fSUR9Lm9zcy0ke1JFR0lPTl9JRH0taW50ZXJuYWwuYWxpeXVuY3MuY29tClVQREFURV
9TSVRFMj1odHRwOi8vdHJ1c3RjbGllbnQtJHtSRUdJT05fSUR9Lm9zcy0ke1JFR0lPTl9JRH0uYWxpeXVuY3MuY29tC
lVQREFURV9TSVRFMz1odHRwOi8vdC10cnVzdGNsaWVudC0ke1JFR0lPTl9JRH0ub3NzLXskUkVHSU9OX0lEfS1pbnRl
cm5hbC5hbGl5dW5jcy5jb20KTVNHX0lORk89ImRvd25sb2FkaW5nIGluc3RhbGwgc2NyaXB0IGZyb20gc2l0ZSIKTVN
HX0VSUj0iZG93bmxvYWQgZmlsZSBlcnJvci4iCk1TR19PSz0idHJ1c3QgY2xpZW50IGluaXQgZG9uZS4iCgppbnN0YW
xsKCkKewogIGVjaG8gIiR7MVNHX0lORk99IiIgMS4uLiIKICBjdXJsIC1mc1NMICIke1VQREFURV9TSVRFMX0iIiR7U
0NSSVBUX1BBVEh9InxzaAogIGlmIFsgJD8gPT0gMCBdOyB0aGVuCiAgICByZXR1cm4gMQogIGZpCiAgZWNobyAiJHtN
U0dfSU5GT30iIiAyLi4uIgogIGN1cmwgLWZzU0wgIiR7VVBEQVRFX1NJVEUyfSIiJHtTQ1JJUFRfUEFUSH0ifHNoCiA
gaWYgWyAkPyA9PSAwIF07IHRoZW4KICAgIHJldHVybiAyCiAgZmkKICBlY2hvICIke01TR19JTkZPfSIiIDMuLi4iCi
AgY3VybCAtZnNTTCAiJHtVUERBVEVfU0lURTN9IiIke1NDUklQVF9QQVRIfSJ8c2gKICBpZiBbICQ/ID09IDAgXTsgd
GhlbgogICAgcmV0dXJuIDMKICBmaQogIGVjaG8gIiIgMT4mMgogIGV4aXQgMQp9CgppbnN0YWxsCmVjaG8gIiR7TVNH
X09LfSIKCmV4aXQgMAo=
&<Common request parameters>
```

Sample success responses:

- XML format

```
<RunInstancesResponse>
      <RequestId>04F0F334-1335-436C-A1D7-6C044FE73368</RequestId>
      <InstanceIdSets>
            <InstanceIdSet>i-bp16byi4f3fti5b3****</InstanceIdSet>
      </InstanceIdSets>
</RunInstancesResponse>
```

- JSON format

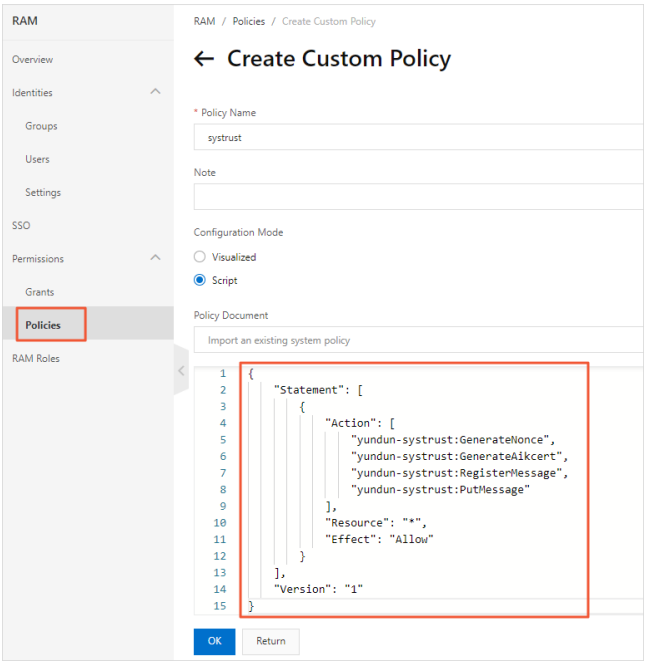
```
{
    "RequestId": "BB694A51-7860-4B5C-B906-9B4077798672",
    "InstanceIdSets": {
        "InstanceIdSet": [
            "i-bp16byi4f3fti5b3****"
        ]
    }
}
```

Precautions on granting permissions to RAM roles

We recommend that you create a custom policy that contains the minimum required permissions and attach the policy to the RAM role. You can set the permission type to **System Policy** (**AliyunSysTrustFullAccess**) corresponding to the trusted service. You can also set the permission type to **Custom Policy** for precise authorization. The following section shows the precise policy for accessing trusted services.

> **Note** You can select a system policy such as AdministratorAccess that grants greater permissions. However, permissions of RAM roles are related to information security risks. We strongly recommend that you grant permissions based on the principle of least privilege. For more information, see What is RAM?

```
{
    "Statement": [
        {
            "Action": [
                "yundun-systrust:GenerateNonce",
                "yundun-systrust:GenerateAikcert",
                "yundun-systrust:RegisterMessage",
                "yundun-systrust:PutMessage"
            ],
            "Resource": "*",
            "Effect": "Allow"
        }
    ],
    "Version": "1"
}
```



Script for installing the Alibaba Cloud trusted system

```
#!/bin/sh
CURPATH=`pwd`
SCRIPT_PATH="/download/linux/script/TrustAgentInstall.sh"
REGION_ID=`curl -s --retry 1 --max-time 3 http://100.100.100.200/latest/meta-data/region-id
`
UPDATE_SITE1=http://trustclient-${REGION_ID}.oss-${REGION_ID}-internal.aliyuncs.com
UPDATE_SITE2=http://trustclient-${REGION_ID}.oss-${REGION_ID}.aliyuncs.com
UPDATE_SITE3=http://t-trustclient-${REGION_ID}.oss-{$REGION_ID}-internal.aliyuncs.com
MSG_INFO="downloading install script from site"
MSG_ERR="download file error."
MSG_OK="trust client init done."
install()
{
echo "${MSG_INFO}"" 1..."
curl -fsSL "${UPDATE_SITE1}""${SCRIPT_PATH}"|sh
if [ $? == 0 ]; then
return 1
fi
echo "${MSG_INFO}"" 2..."
curl -fsSL "${UPDATE_SITE2}""${SCRIPT_PATH}"|sh
if [ $? == 0 ]; then
return 2
fi
echo "${MSG_INFO}"" 3..."
curl -fsSL "${UPDATE_SITE3}""${SCRIPT_PATH}"|sh
if [ $? == 0 ]; then
return 3
fi
echo "" 1>&2
exit 1
}
install
echo "${MSG_OK}"
exit 0
```

4.7.3. Trusted feature for security-enhanced instances

4.7.3.1. Overview

Security-enhanced instance families are provided by Alibaba Cloud to implement trusted boot based on Trusted Cryptography Module (TCM) or Trusted Platform Module (TPM) chips. During a trusted boot, each module in the boot chain from the underlying hardware to the guest OS is measured and verified. This topic describes how a security-enhanced instance works and basic concepts of the trusted computing technology.

How a security-enhanced instance works

Trusted computing is one of the main features used to achieve the high-level security of underlying computing environments for cloud tenants. TPM or TCM is integrated into the hardware platform to build a trusted chain that covers system startup and user-specified applications and implement a remote attestation mechanism. This guarantees a trusted environment for users in all aspects during the startup and running phases. The trust verification of systems and applications reduces vulnerability to attacks due to the use of unknown or tampered systems or software.

Security-enhanced instance families use the trusted computing technology to verify the integrity of each module. This ensures that instances are not compromised by startup-level or kernel-level malware or rootkits. Based on the TPM or TCM trusted hardware, the security-enhanced instance families can achieve measured boot and integrity verification by using the Unified Extensible Firmware Interface (UEFI) firmware, vTPM or vTCM, and remote attestation service. This ensures that security-enhanced instances are secure and trusted.

TPM or TCM trusted hardware

Trusted computing relies on TPM or TCM chips. TPM is standardized by ISO as ISO 11889, and TCM is standardized as GM/T 0012-2020 in China. TPM or TCM chips used as the root of trust has the following benefits:

- TPM or TCM uses its own internal firmware and logic circuits to process instructions. It does not rely on operating systems and is insulated from external software vulnerabilities.
- Attackers must have physical access to computers before they can attack TPM or TCM chips.
- Security-enhanced instances are equipped with TPM or TCM chips, startup firmware, and system software to build a chain of trust.

Firmware security

Alibaba Cloud supports secure firmware updates. Before firmware is updated, firmware signatures are verified to ensure that only authorized firmware can be updated. This can prevent malicious firmware from attacking the cloud infrastructure.

vTPM and vTCM

Alibaba Cloud also provides virtual roots of trust (vTPM and vTCM) for ECS instances to extend the trust system of servers to the ECS virtualization layer based on trusted hardware. A comprehensive security system is built on hardware and virtual roots of trust.

vTPM and vTCM are virtualized and trusted platform modules, which can be used to transmit trust from the trusted server hardware to the trusted instance. vTPM is fully compatible with the trusted computing specification TPM2.0. vTCM is fully compatible with TCM2.0. Security-enhanced instances enable vTPM or vTCM to build a virtual root of trust within these instances and implement a trusted boot chain and remote attestation mechanism similarly to the host layer. The baseline measurement data is generated when an instance is created. The measurement values collected on subsequent instance startups are compared with the baseline measurement data to determine whether the instance has changed. The comparison result indicates the trusted status of the instance and is displayed in the Security Center console.

UEFI firmware

Security-enhanced instances use trusted boot firmware that meets the UEFI for system boot. The UEFI firmware can measure the integrity of system firmware, system boot loader, and system kernel modules during the boot process of the operating system to build a chain of trust for system startup.

Measured boot

Components are measured stage by stage. The components started first measure the next stage components before starting them. If the measurement is successful, the chain of trust is extended to the next stage.

Each module in the boot chain from the underlying hardware to the guest OS is measured during the boot process of an instance. When these modules are loaded, trusted components calculate the hash value for each module and securely store the calculated values to the root of trust to form a chain of trust. Stage-by-stage measurement and verification of all modules in the boot chain ensures that the system has not changed since the last boot.

Integrity verification

Integrity verification helps you to learn the trusted status of instances and make decisions.

When an instance is started for the first time, the trusted components create the first set of hash values as baseline measurement and securely store the data. Then, these measurement and storage operations are performed each time the instance starts. Trusted components send the measurement values to the trusted service by using remote attestation. You can measure and verify the integrity of the instance by comparing the latest measurement data with the baseline measurement to determine whether the instance is running in the expected trusted state.

Integrity verification compares startup measurement information with the baseline measurement of an instance. If the information is matched, a pass result is returned, which indicates that the instance is trusted. If the information is not matched, a failure result is returned, which indicates that the instance is untrusted.

If an expected integrity verification failure occurs in specific scenarios such as when the system of the ECS instance was updated, you can update the instance baseline measurement by adding the trusted event to the whitelist. Subsequent integrity measurements are performed against the latest baseline measurement. For more information, see Handle trusted exceptions. If an unexpected integrity verification failure occurs, you must find the cause of the failure based on the trusted event details to prevent instances from running in an untrusted environment.

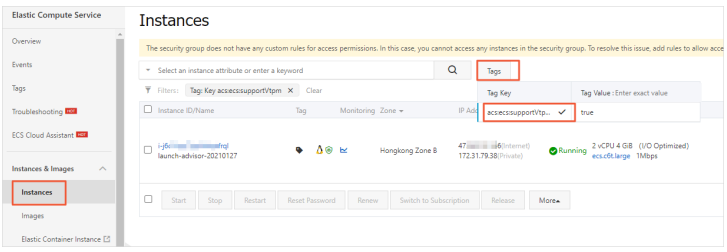
4.7.3.2. Use the trusted feature of security-enhanced instances

This topic describes how to use and maintain security-enhanced instances. These operations include quickly filtering instances, viewing the trusted state of instances, and handling state exceptions.

Filter security-enhanced instances

Security-enhanced instances are bound with the **acs:ecs:supportVtpm** tag. If you have a large number of security-enhanced instances within a region, you can filter the instances by tag.

On the **Instances** page, click **Tags** and select **acs:ecs:supportVtpm** from the Tag Key drop-down list.

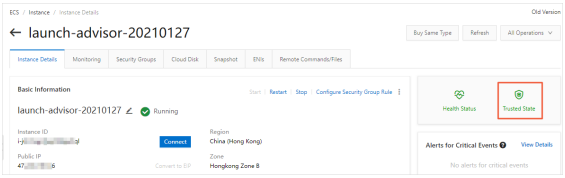


View the trusted state of an instance

The baseline measurement data is generated when an instance is created. The measurement values collected on subsequent instance startups are compared with the baseline measurement data to determine whether the instance has changed. The comparison result indicates the trusted status of the instance and is displayed in the Security Center console.

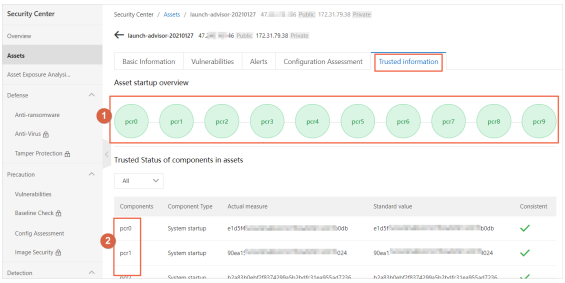
1. On the **Instances** or **Instance Details** page of the ECS console, find and click the icon above **Trusted State**.

   You are automatically redirected to the **Assets** page in the Security Center console.

   > **Note** If you move the pointer over the icon above **Trusted State** on the **Instance Details** page and **Unmeasured** appears, no valid measurement results have been reported for the security-enhanced instance for an extended period of time. In this case, no detailed trusted information is displayed in the Security Center console. For information about how to handle cases where no measurement is made, see Handle the unmeasured state.

2. Click the **Trusted Information** tab to view the trusted state of the instance.

   The circles in the ①**Asset startup overview** section are mapped with the component list in the ② **Trusted Status of components in assets** section. The color of a circle in the ①**Asset startup overview** section indicates whether the stage is normal:

   - If all of the circles are green, the instance startup process is normal. In this case, the **actual measurement value** (the actual status collected by the system trusted feature) is the same as the **standard value**.
   - If an error occurs at one stage during the instance startup process, the corresponding circle turns red and those that follow turn gray. You can view the specific information of this stage on the **Alerts** tab and try to fix it. For more information, see Handle trusted exceptions.

   Platform Configuration Registers (PCRs) are storage units of trusted security devices and are capable of reliably storing the status information collected during the instance startup process. Each PCR corresponds to a specific stage of the instance startup process and the PCR value represents the status of the measured object at each stage. If the actual measurement value stored in the PCR is the same as the expected standard value, this stage is considered to be as expected. The following objects are measured at each stage of the instance startup process:

   - pcr0: the SRTM, BIOS, embedded optional ROM, and PI driver.
   - pcr1: the host platform configurations.

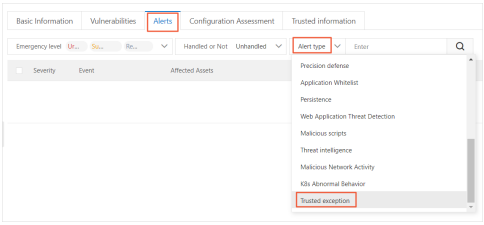
- pcr2: the UEFI driver and application code.
- pcr3: the UEFI driver, application configurations, and application data.
- pcr4: the UEFI startup management code (typically MBR).
- pcr5: the UEFI startup management code (typically MBR), startup-related data (data used by the UEFI startup management code), and GPT partition table.
- pcr6: the specific UEFI firmware defined by the platform manufacturer.
- pcr7: the secure startup policy.
- pcr8: the key commands to be run as provided in configuration files such as *grub.cfg* and command line information transmitted to the Linux kernel. Non-critical commands are not measured, such as the command used to define boot menu titles.
- pcr9: the GRUB module, Linux kernel, and initramfs.

> **Note** ISO provides detailed definitions. For more information, visit *ISO/IEC 11889:2015 Trusted Platform Module Library*.

Handle trusted exceptions

If an error occurs at one stage during the instance startup process, the corresponding circle on the **Trusted Information** tab turns red. You must go to the **Alerts** tab to view detailed alert information and fix the exceptions.

1. Click the **Alerts** tab and set **Alert type** to **Trusted exception**.



2. On the right side of alert information, click **Details** to view detailed error information.

> **Note** If the alert information has not been processed, alerts are periodically raised, but no more alerts are generated. Only the time of the latest alert is displayed in the **Latest Occurrence** column.

3. Contact the system administrator to check whether system upgrade and maintenance operations such as upgrading the operating system kernel, changing the operating system startup parameters, and modifying the initial file system (initramfs) have been performed recently. Then, take different measures to fix trusted exceptions based on specific situations.
   - **Scenario 1: If no system upgrade or maintenance operations have been performed recently, ignore the alert after you check and fix the exception.**

     In this scenario, an abnormal alert may occur because a security event has occurred to your instance. For example, the instance is damaged by malware such as rootkit or bootkit. We recommend that you contact the system administrator to perform a drill-down check on the system, fix the related exceptions, and then ignore the alert. Perform the following steps:

a. Enable and use the **Anti-Virus** and **Vulnerabilities** features in the Security Center console. Then, upgrade the latest virus library, check the status of malware in the system, and then fix the vulnerability.

b. On the **Alerts** tab, click **Handle**.

c. Select **Ignore** and click **Immediate processing**.

If an alert is generated on multiple instances, you can select **Handle the same alarms at the same time** to handle the same alert on each instance at a time.

> **Notice** Alerts handled in **ignore** mode are still displayed on the **Trusted Information** tab. The ignored alerts are continuously generated because Security Center periodically generates security alerts. These situations persist until you restart the system and pass the verification.

- **Scenario 2: If a system upgrade or maintenance operation is performed recently, add the exception to the whitelist after repair.**

  If a system upgrade or maintenance operation has been performed recently, the modified system status becomes the new standard status of your system. The status value of each stage during the instance startup process also becomes the new standard value of the corresponding PCR. In this case, you must select **Add whitelist**.

  After the collected actual measurement values are added to the whitelist, the values become the new benchmark measurement values.

Handle the unmeasured state

If you move the pointer over the icon next to **Trusted State** on the **Instance Details** page and **Unmeasured** appears, no valid measurement results are reported for the security-enhanced instance for an extended period of time. This is typically because the trusted client is unable to access the trusted service. In this case, you can perform the following steps to troubleshoot the problem:

1. Check the instance RAM role.

   If you have not specified a RAM role for the security-enhanced instance, specify one as required. If you have specified a RAM role for the security-enhanced instance, check whether the RAM role has the required permissions to access the trusted service. For more information, see Create security-enhanced instances.

2. Check the network connection.

   Run the following command in the security-enhanced instance to check the network connection:

   ```
   ping trusted-server-vpc.[region-id].aliyuncs.com
   ```

   Replace *[region-id]* with the ID of the region where the security-enhanced instance resides. If an output is returned, the network connection is normal.

3. Check the security group settings.

   Check the settings of the security group to which the security-enhanced instance belongs, and make sure that the access to trusted-server-vpc.[region-id].aliyuncs.com is not denied.

4.7.4. Build an SGX encrypted computing environment

This topic describes how to build a Software Guard Extensions (SGX) encrypted computing environment on a g7t, c7t, or r7t instance (vSGX instance) and run sample code to verify SGX features.

Prerequisites

An Elastic Compute Service (ECS) vSGX instance is created.

> Note

Context

Intel® SGX provides an encrypted computing environment at the physical level and ensures data security by providing hardware-based protections instead of firmware- or software-based protections. Intel® SGX uses instruction set extensions and an access control mechanism to isolate the runtime environment of SGX programs. This can protect the confidentiality and integrity of essential code and data against malware attacks. Compared with other security technologies, Intel® SGX uses the root of trust that contains only hardware. This can prevent defects caused by security vulnerabilities of software on which the root of trust is based, and improve system security.

The g7t, c7t, and r7t security-enhanced instance families provide encrypted memory based on Intel® SGX and support the SGX technology applicable to virtual machines. You can develop and run SGX programs in vSGX instances.

> Notice

Check whether SGX is enabled

Before you build an SGX encrypted computing environment, you can use CPUID to check whether SGX is enabled. This section describes how to check whether SGX is enabled. In this example, an Alibaba Cloud Linux 2 (UEFI) image or an Alibaba Cloud Linux 3 (UEFI) image is used.

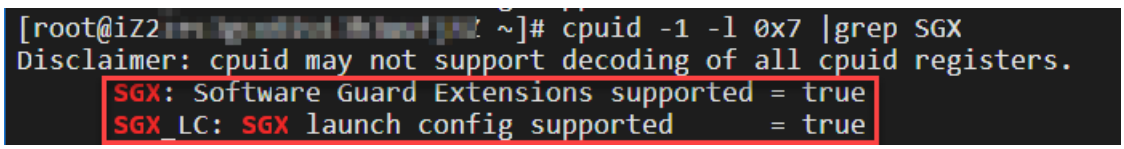
1. Install CPUID.

   ```
   yum install -y cpuid
   ```

2. Check whether SGX is enabled.

   ```
   cpuid -1 -l 0x7 |grep SGX
   ```

   The following figure shows that SGX is enabled.

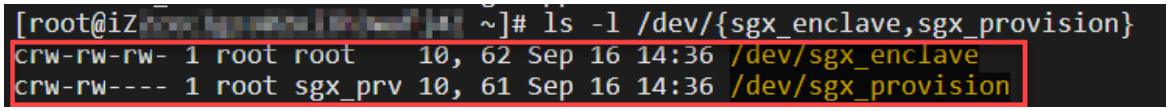


> **Note** After SGX is enabled, the SGX driver is required to run SGX programs. The dedicated images provided by Alibaba Cloud have a built-in SGX driver. If you do not use a dedicated image, install the SGX driver.

3. Check whether the SGX driver is installed.

```
ls -l /dev/{sgx_enclave,sgx_provision}
```

The following figure shows that the SGX driver is installed.



Build an SGX encrypted computing environment

Before you develop SGX programs, you must install the SGX runtime and SDK on a vSGX instance and configure the remote attestation service. We recommend that you use dedicated images provided by Alibaba Cloud for a better user experience. Dedicated images are equipped with the SGX driver and provide TEE SDK that is fully compatible with Intel® SGX SDK. This section describes how to build an SGX encrypted computing environment. In this example, an Alibaba Cloud Linux 2 (UEFI) image or an Alibaba Cloud Linux 3 (UEFI) image is used. If you use Ubuntu images, CentOS images, or other Linux images, install the SGX driver and Platform SoftWare (PSW). For more information, see Intel® SGX Software Installation Guide.

1. Install the Alibaba Cloud SGX runtime.

> **Note** When you create a vSGX instance in the ECS console, the Alibaba Cloud SGX runtime is automatically installed. You can skip this step and install Alibaba Cloud TEE SDK.

i. Import the YUM software repository for Alibaba Cloud encrypted computing.

- Public URLs of the repository are in the following format: `https://enclave-[Region-ID].oss-[Region-ID].aliyuncs.com/repo/alinux/enclave-expr.repo`.
- Internal URLs of the repository are in the following format: `https://enclave-[Region-ID].oss-[Region-ID]-internal.aliyuncs.com/repo/alinux/enclave-expr.repo`.

Replace *[Region-ID]* in the preceding URLs with the region ID of the vSGX instance. The following example shows the internal URL of a vSGX instance in the China (Hangzhou) region:

```
sudo yum install -y yum-utils && \
sudo yum-config-manager --add-repo \
https://enclave-cn-hangzhou.oss-cn-hangzhou-internal.aliyuncs.com/repo/alinux/enclave-expr.repo
```

ii. Install the Alibaba Cloud SGX runtime.

```
yum install -y \
libsgx-ae-le libsgx-ae-pce libsgx-ae-qe3 libsgx-ae-qve \
libsgx-aesm-ecdsa-plugin libsgx-aesm-launch-plugin libsgx-aesm-pce-plugin libsgx-ae
sm-quote-ex-plugin \
libsgx-dcap-default-qpl libsgx-dcap-ql libsgx-dcap-quote-verify \
libsgx-enclave-common libsgx-launch libsgx-pce-logic libsgx-qe3-logic libsgx-quote-
ex \
libsgx-ra-network libsgx-ra-uefi libsgx-uae-service libsgx-urts sgx-ra-service \
sgx-aesm-service
```

> **Note** SGX Architectural Enclave Service Manager (AESM) is used to manage services such as enclave start, key configuration, and remote attestation. The default installation path of SGX AESM is */opt/intel/sgx-aesm-service*.

2. Install Alibaba Cloud TEE SDK.

```
yum install -y sgxsdk
```

Alibaba Cloud TEE SDK is fully compatible with Intel® SGX SDK. After Alibaba Cloud TEE SDK is installed, you can refer to Intel® SGX Developer Reference to develop SGX programs.

> **Note** The default installation path of Intel® SGX SDK in Alibaba Cloud TEE SDK is */opt/alibaba/teesdk/intel/sgxsdk/*.

3. Configure the Alibaba Cloud SGX remote attestation service.

The Alibaba Cloud SGX remote attestation service is fully compatible with Intel® SGX Elliptic Curve Digital Signature Algorithm (ECDSA)-based remote attestation service and Intel® SGX SDK. Therefore, vSGX instances provided by Alibaba Cloud can gain trust from remote providers and producers by using remote attestation. For more information, see Intel® ECDSA remote attestation service. The Alibaba Cloud SGX remote attestation service provides the following information for SGX SDK:

- SGX certificates
- Revocation list: a list of revoked SGX certificates
- Trusted computing base information: information about the root of trust

> **Note**

The Alibaba Cloud SGX remote attestation service is deployed on a per-region basis. You can access this service deployed in the region where the vSGX instance is located for optimal stability. After Alibaba Cloud TEE SDK is installed, the default configuration file */etc/sgx_default_qcnl.conf* is automatically generated for the remote attestation service. You must manually adapt the file to the Alibaba Cloud SGX remote attestation service in the region where the vSGX instance is located.

- If the vSGX instance is assigned a public IP address, change the configurations in */etc/sgx_default_qcnl.conf* to the following content:

```
# PCCS server address
PCCS_URL=https://sgx-dcap-server.[Region-ID].aliyuncs.com/sgx/certification/v3/
# To accept insecure HTTPS cert, set this option to FALSE
USE_SECURE_CERT=TRUE
```

Note The Alibaba Cloud SGX remote attestation service is supported in regions and zones listed in the following table.

Supported region	Region ID
China (Qingdao)	cn-qingdao
China (Beijing)	cn-beijing
China (Zhangjiakou)	cn-zhangjiakou
China (Ulanqab)	cn-wulanchabu
China (Hangzhou)	cn-hangzhou
China (Shanghai)	cn-shanghai
China (Shenzhen)	cn-shenzhen
China (Heyuan)	cn-heyuan
China (Guangzhou)	cn-guangzhou
China (Chengdu)	cn-chengdu
China (Hong Kong)	cn-hongkong
Singapore (Singapore)	ap-southeast-1

Replace *[Region-ID]* with the region ID of the vSGX instance. Example for a vSGX instance in the China (Hangzhou) region:

```
# PCCS server address
PCCS_URL=https://sgx-dcap-server.cn-hangzhou.aliyuncs.com/sgx/certification/v3/
# To accept insecure HTTPS cert, set this option to FALSE
USE_SECURE_CERT=TRUE
```

- If the vSGX instance is in a virtual private cloud (VPC) and has only internal IP addresses, change the configurations in */etc/sgx_default_qcnl.conf* to the following content:

```
# PCCS server address
PCCS_URL=https://sgx-dcap-server-vpc.[Region-ID].aliyuncs.com/sgx/certification/v3/
# To accept insecure HTTPS cert, set this option to FALSE
USE_SECURE_CERT=TRUE
```

Replace *[Region-ID]* with the region ID of the vSGX instance. Example for a vSGX instance in the China (Hangzhou) region:

```
# PCCS server address
PCCS_URL=https://sgx-dcap-server-vpc.cn-hangzhou.aliyuncs.com/sgx/certification/v3/
# To accept insecure HTTPS cert, set this option to FALSE
USE_SECURE_CERT=TRUE
```

Example 1 of verifying SGX features: Start an enclave

Alibaba Cloud TEE SDK provides SGX sample code to verify SGX features. By default, the code is stored in the */opt/alibaba/teesdk/intel/sgxsdk/SampleCode* directory.

This section describes an example of how to start an enclave to verify whether the installed SGX SDK works normally. If the enclave is started, the SDK works normally. In this example, the sample code file named SampleEnclave is used.

1. Install a compiler.
   - If the Alibaba Cloud Linux 2 (UEFI) image is used, install devtoolset.
     a. Open the Alibaba Cloud experimental repository.

     ```
     rpmkeys --import http://mirrors.cloud.aliyuncs.com/epel/RPM-GPG-KEY-EPEL-7 && \
     yum install -y alinux-release-experimentals
     ```

     b. Install devtoolset.

     ```
     yum install -y devtoolset-9
     ```

     c. Set the environment variable related to devtoolset.

     ```
     source /opt/rh/devtoolset-9/enable
     ```

   - If the Alibaba Cloud Linux 3 (UEFI) image is used, install Development Tools.

   ```
   yum groupinstall -y "Development Tools"
   ```

2. Set the environment variable related to SGX SDK.

   ```
   source /opt/alibaba/teesdk/intel/sgxsdk/environment
   ```

3. Compile the sample code in SampleEnclave.
   i. Go to the SampleEnclave directory.

   ```
   cd /opt/alibaba/teesdk/intel/sgxsdk/SampleCode/SampleEnclave
   ```

   ii. Compile SampleEnclave.

   ```
   make
   ```

4. Run the compiled executable file.

   ```
   ./app
   ```

Example 2 of verifying SGX features: SGX remote attestation

Alibaba Cloud TEE SDK provides SGX sample code to verify SGX features. By default, the code is stored in the */opt/alibaba/teesdk/intel/sgxsdk/SampleCode* directory.

This section describes an example of the SGX remote attestation service. The expected result is that a quote is generated and verified (QuoteGenerationSample and QuoteVerificationSample). The example involves the challenged party (SGX programs that run in the vSGX instance) and the challenging party (the party that wants to verify whether the SGX programs are trusted). In this example, the sample code file named QuoteGenerationSample is used by the challenged party to generate a quote, and the sample code file named QuoteVerificationSample is used by the challenging party to verify the quote

1. Install a compiler.
   - If the Alibaba Cloud Linux 2 (UEFI) image is used, install devtoolset.
     a. Open the Alibaba Cloud experimental repository.

     ```
     rpmkeys --import http://mirrors.cloud.aliyuncs.com/epel/RPM-GPG-KEY-EPEL-7 && \
     yum install -y alinux-release-experimentals
     ```

     b. Install devtoolset.

     ```
     yum install -y devtoolset-9
     ```

     c. Set the environment variable related to devtoolset.

     ```
     source /opt/rh/devtoolset-9/enable
     ```

   - If the Alibaba Cloud Linux 3 (UEFI) image is used, install Development Tools.

   ```
   yum groupinstall -y "Development Tools"
   ```

2. Set the environment variable related to SGX SDK.

   ```
   source /opt/alibaba/teesdk/intel/sgxsdk/environment
   ```

3. Install the dependency package of SGX remote attestation.

   ```
   yum install -y libsgx-dcap-ql-devel libsgx-dcap-quote-verify-devel
   ```

4. Compile the sample code in QuoteGenerationSample used by the challenged party.
   i. Go to the QuoteGenerationSample directory.

   ```
   cd /opt/alibaba/teesdk/intel/sgxsdk/SampleCode/QuoteGenerationSample
   ```

   ii. Compile QuoteGenerationSample.

   ```
   make
   ```

5. Run the compiled executable file to generate Quote.

   ```
   ./app
   ```

6. Compile the sample code in QuoteVerificationSample used by the challenging party.
   i. Go to the QuoteVerificationSample directory.

   ```
   cd /opt/alibaba/teesdk/intel/sgxsdk/SampleCode/QuoteVerificationSample
   ```

   ii. Compile QuoteVerificationSample.

   ```
   make
   ```

7. Sign the QuoteVerificationSample enclave.

To release an official version of an enclave, you must provide the signature key to sign the enclave.

```
sgx_sign sign -key Enclave/Enclave_private_sample.pem -enclave enclave.so -out enclave.
signed.so -config Enclave/Enclave.config.xml
```

8. Run the compiled executable file to verify the quote.

```
./app
```

Known issues

The SGX driver that comes with Alibaba Cloud Linux 2 in the kernel of the 4.19.91-23.al7.x86_64 version experiences memory leaks in some specific cases. This issue is fixed in the latest version. We recommend that you upgrade the kernel to the latest version. If you want to continue using this kernel version, we recommend that you install patches to avoid this issue.

```
yum install -y alinux-release-experimentals && \
    yum install -y kernel-hotfix-5577959-23.al7.x86_64
```

4.7.5. Deploy the TensorFlow Serving online inference service on a security-enhanced instance

4.7.5.1. Overview

This topic describes how to deploy the TensorFlow Serving online inference service on an Intel® SGX-based security-enhanced instance and how to use TensorFlow Serving.

Context

TensorFlow Serving is part of the TensorFlow ecosystem, which is an open-source machine learning platform developed by Google. TensorFlow Serving can run trained models and provide APIs for other services to call. This way, the models can be used to perform inferences and make predictions.

Some security-enhanced Elastic Compute Service (ECS) instances provide encrypted computing capabilities based on Intel® Software Guard Extension (SGX) to create a hardware-level trusted confidential environment that offers a high degree of security. This ensures the confidentiality and integrity of essential code and data and protects them from malware attacks.

You can deploy TensorFlow Serving online inference scenarios in the trusted confidential environment of security-enhanced ECS instances. This ensures the security of data transmission and data usage, the security of data disks, and the integrity of artificial intelligence (AI) applications for online inference.

This practice provides reference implementations for developers to use Intel® SGX-based security-enhanced instances.

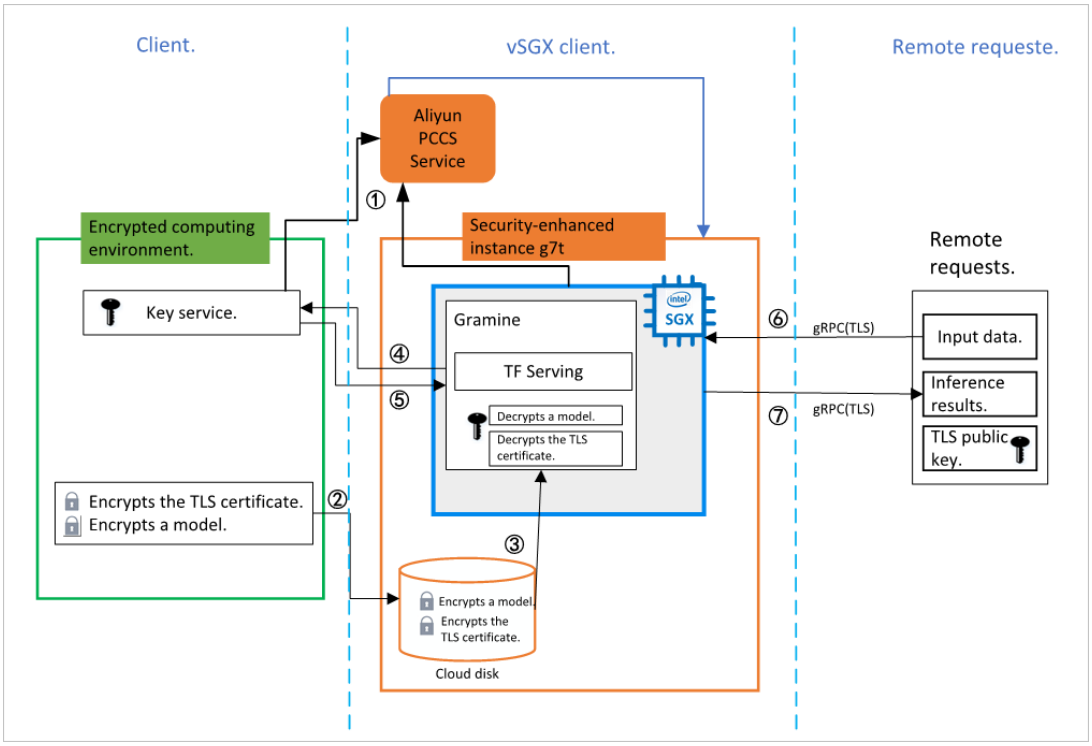
- You can have an overall understanding of end-to-end security solutions for full lifecycle data protection based on SGX.
- Developers who use TensorFlow Serving can refer to this practice to deploy and develop scripts.
- This practice provides a feasible reference framework and scripts for SGX developers who use

security-enhanced instances. You can follow the steps to get started with security-enhanced instances and build an SGX encrypted computing environment.

Architecture

The architecture in this practice is shown in the Architecture figure.

Architecture



This practice involves three roles: client, vSGX client, and remote requester.

- Client: A client encrypts a trained model and the Transport Layer Security (TLS) certificate that is used to establish secure connections, and uploads the encrypted files to the SGX encrypted computing environment. A key service is deployed on the same instance as the client to authenticate Alibaba Cloud vSGX instances and to ensure the integrity of the TensorFlow Serving inference service that runs on the cloud and the feasibility of cloud-based SGX environments. After a vSGX instance is authenticated, the client sends a key to the TensorFlow Serving inference service that runs on the instance.
- vSGX client: The vSGX client is deployed on Alibaba Cloud and provides an SGX encrypted computing environment in which the TensorFlow Serving inference service runs. When the inference service starts, it sends a remote attestation request to the client to verify the feasibility of the SGX environment and the integrity of the inference service. If the verification succeeds, the inference service receives a key from the client and decrypts the encrypted model and TLS certificate. Then, the inference service starts to run in the SGX environment and waits for remote access requests to perform inferences.
- Remote requester: Data is transferred over networks to the inference service that runs in the SGX encrypted computing environment. After inferences are complete, the inference service returns the inference results.

> **Note** In this practice, the client and the requester are deployed on the same ECS instance.

In this practice, the following components are used:

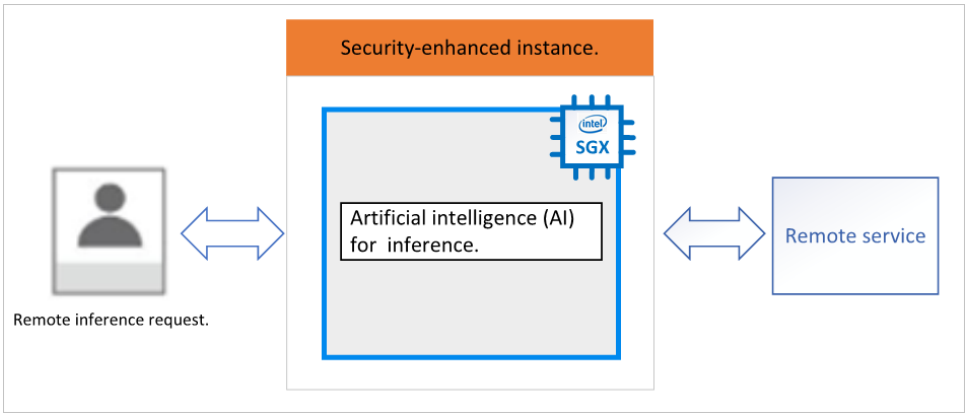
- Library operating system (LibOS): Gramine is a lightweight LibOS that combines Intel® SGX to support kernel customization. Gramine consumes a small amount of resources when it runs and is compatible with application binary interfaces (ABIs). Gramine greatly reduces the cost of porting native applications to the SGX environment and enables applications to run in the SGX environment with no or minimal modification. In this practice, Gramine is used to encapsulate and run the TensorFlow Serving inference service in the Alibaba Cloud vSGX instance. For more information, see Gramine.
- AI inference service: TensorFlow Serving is part of the TensorFlow ecosystem, which is an open-source machine learning platform developed by Google. TensorFlow Serving can run trained models and provide APIs for other services to call. This way, the models can be used to perform inferences and make predictions. For more information, see TensorFlow.
- Docker engine: To simplify the deployment of the inference service, the inference service is scheduled to run in a container by using docker commands.

Architecture of the TensorFlow Serving inference service:

- Alibaba Cloud Provisioning Certificate Caching Service (PCCS) is fully compatible with Intel SGX Elliptic Curve Digital Signature Algorithm (ECDSA) remote attestation and Intel SGX SDKs. As shown in Section ① of the Architecture figure, the Alibaba Cloud vSGX instance must request a provisioning certification key (PCK) certificate from Alibaba Cloud PCCS. Intel SGX has an enclave signing key that is unique to the processor or the platform. The public part of the key is the PCK public key. The client also requests quote verification data such as the trusted computing base (TCB) and certificate revocation list (CRL) from Alibaba Cloud PCCS to verify the SGX enclave.
- After you create an Alibaba Cloud vSGX instance, you can transfer the locally encrypted model file and TLS certificate over networks to a disk for backup, as shown in Section ② of the Architecture figure.
- When you start the TensorFlow Serving inference service by using Gramine, the encrypted model file is loaded, as shown in Section ③ of the Architecture figure.
- Gramine integrates the remote attestation feature. When Gramine recognizes that an encrypted file is loaded, it sends attestation requests to the configured remote IP address, as shown in Section ④ of the Architecture figure. In this practice, the client and the remote requester are deployed on the same ECS instance. The vSGX client is deployed on another instance.
- After the key service of the client determines that an SGX enclave quote is trusted, the TensorFlow Serving inference service sends the encryption key of the model file to Gramine, as shown in Section ⑤ of the Architecture figure. Gramine runs in the enclave, which can ensure that Gramine receives the key to decrypt the model in a secure manner.
- After the model is decrypted, TensorFlow Serving can run normally and wait for remote access requests. To establish a secure communication channel, the remote requester holds the public key of the TLS certificate. After a connection is established, the TLS certificate in TensorFlow Serving is verified, as shown in Section ⑥ of the Architecture figure.
- When TensorFlow Serving completes inferences on the data of the remote requester, it returns the inference results to the requester by using the secure communication channel, as shown in Section ⑦ of the Architecture figure.

Procedure

You can access the TensorFlow Serving inference service by using a domain name. The TensorFlow Serving inference service is deployed on an Alibaba Cloud security-enhanced instance. The model used by the inference service is encrypted and stored on a disk. When the SGX trusted confidential environment in the instance is initialized, a remote attestation request is sent to the remote service. After the verification succeeds, the remote attestation service sends the decryption key of the model to the TensorFlow Serving inference service that runs in the SGX trusted environment. After the model is decrypted, the TensorFlowServing inference service processes the request based on the decrypted model and returns the inference results.



The following section describes the procedure of deploying the TensorFlow Serving online inference service on a security-enhanced instance:

1. Step 1: Deploy a client

   The client encrypts a trained model and the TLS certificate that is used to establish secure connections, and uploads the encrypted files to the SGX encrypted computing environment. A key service is deployed on the same instance as the client to authenticate Alibaba Cloud vSGX instances.

2. Step 2: Deploy a vSGX client

   The vSGX client provides a vSGX encrypted computing environment in which the TensorFlow Serving inference service can run. When the inference service starts, it sends a remote attestation request to the client to verify the feasibility of the SGX environment and the integrity of the inference service. If the verification succeeds, the inference service receives a key from the client, decrypts the encrypted model and TLS certificate, and waits for remote access requests.

3. Step 3: Perform remote access

   The remote requester sends data to the inference service that runs in the SGX encrypted computing environment. After inferences are complete, the inference service returns the inference results.

4.7.5.2. Step 1: Deploy a client

This topic describes how to deploy a client. The operations to deploy a client include building a Software Guard Extensions (SGX) encrypted computing environment, creating an encryption model, and making a gRPC Transport Layer Security (TLS) certificate.

Context

The client encrypts a trained model and the TLS certificate that is used to establish secure connections, and uploads the encrypted files to the SGX encrypted computing environment. A key service is deployed on the same Elastic Compute Service (ECS) instance as the client to authenticate Alibaba Cloud virtual SGX (vSGX) ECS instances and to ensure the integrity of the TensorFlow Serving inference service that runs on the cloud and the feasibility of cloud-based SGX environments. After a vSGX instance is authenticated, the client sends a key to the TensorFlow Serving inference service that runs on the instance.

Procedure

1. Create a security-enhanced ECS instance and build an SGX encrypted computing environment on the instance.
   i. Create a security-enhanced ECS instance.

      For more information, see Create security-enhanced instances.

      Take note of the following parameters:

      - **Instance Type**: Select an instance type with 32 GiB or more of encrypted memory. In this example, the ecs.g7t.4xlarge instance type is used.
      - **Image**: Select the Alibaba Cloud Linux 2.1903 LTS 64-bit (UEFI) public image.
      - **Public IP Address**: Select Assign Public IPv4 Address.

   ii. Build an SGX encrypted computing environment.

      For more information, see Build an SGX encrypted computing environment.

   iii. Install Python 3 and configure environment variables.

      In this example, Python 3.6 is used. You can install another version of Python 3 based on your business requirements. For more information, visit the official Python website.

   iv. Install Docker.

      For more information, see Deploy and use Docker on Alibaba Cloud Linux 2 instances.

   v. Install required software packages such as opencv-python and mesa-libGL.

      ```
      pip install opencv-python
      yum install mesa-libGL
      pip install tensorflow-serving-api
      ```

2. Log on to the ECS instance.

   For more information, see Connection methodsGuidelines on instance connection.

3. Switch to the working directory (example: /home/tf), download the desired TensorFlow Serving script code, and then install required software packages such as argparse, aiohttp, and tensorflow.

   ```
   cd /home/tf
   git clone https://gitee.com/cloud_cc/confidential-computing.git
   pip3 install -r confidential-computing/Tensorflow_Serving/client/requirements.txt
   ```

4. Go to the TensorFlow_Serving/client directory and download a trained model.

   ```
   cd /home/tf/confidential-computing/Tensorflow_Serving/client
   ./download_model.sh
   ```

   The files of the trained model that you download are stored in the models/resnet50-v15-fp32 directory.

5. Convert the model.

   You must convert the format of the trained model files to make them compatible with TensorFlow Serving.

   ```
   pip3 install tensorflow==2.4.0  #Install TensorFlow, which is required by the script us
   ed to convert the model.
   python3 ./model_graph_to_saved_model.py --import_path `pwd -P`/models/resnet50-v15-fp32
   /resnet50-v15-fp32.pb --export_dir  `pwd -P`/models/resnet50-v15-fp32 --model_version 1
   --inputs input --outputs  predict
   ```

   The converted model files are stored as the models/resnet50-v15-fp32/1/saved_model.pb file.

6. Make a gRPC TLS certificate.

   In this practice, gRPC TLS is used to establish a connection between the client and TensorFlow Serving, and a TensorFlow Serving domain name is configured to create a unidirectional TLS key and to make a certificate for establishing a secure communications channel.

   The script creates a folder named ssl_configure. This folder contains the server.crt, server.key, and ssl.cfg files. server.crt is for use by the client, and ssl.cfg is for use by TensorFlow Serving

   ```
   service_domain_name=grpc.tf-serving.service.com
   ./generate_ssl_config.sh ${service_domain_name}
   ```

7. Create an encryption model.

   SGX SDK v1.9 and later versions come with the secure file I/O feature. This feature is provided in a component named Intel Protected File System Library and allows developers to securely perform I/O operations inside enclaves. For more information, see Understanding SGX Protected File System.

   SGX SDKs can ensure:

   - Confidentiality of user data. All user data is encrypted and then written to disks to prevent data leaks.
   - Integrity of user data. All user data is read from disks and then decrypted by using verified message authentication codes (MACs) to detect data tampering.
   - Matched file names. Before an existing file is opened, the metadata of the file is checked to ensure that the name of the file when created is the same as the name provided to the file open operation.

   LibOS Gramine is used in this practice and provides a reference tool based on the secure file I/O feature to encrypt and decrypt files. The sgx.protected_files.file_mode=file_name configuration option is defined in a template configuration file provided by LibOS Gramine to specify encrypted files for the reference tool to decrypt.

   TensorFlow Serving loads the model from the models/resnet50-v15-fp32/1/saved_model.pb path and uses the key stored in the files/wrap-key directory to encrypt the model file. You can also specify a password that is 128 characters in length. The paths of a file when the file is encrypted and when the file is used must be the same to meet the path matching rule. Run the following command in the gramine-sgx-pf-crypt tool to encrypt the model file:

   ```
   mkdir plaintext/
   mv models/resnet50-v15-fp32/1/saved_model.pb plaintext/
   LD_LIBRARY_PATH=./libs ./gramine-sgx-pf-crypt encrypt -w files/wrap-key -i  plaintext/s
   aved_model.pb -o  models/resnet50-v15-fp32/1/saved_model.pb
   ```

8. Start a key authentication service.

   In this practice, secret_prov_server_dcap provided by LibOS Gramine is used as a remote attestation service to verify SGX enclave quotes by calling the quote verification library of SGX Data Center Attestation Primitives (DCAP) at the underlying layer. secret_prov_server_dcap then provides quote verification data such as the trusted computing base (TCB) and certificate revocation list (CRL) to Alibaba Cloud Provisioning Certificate Caching Service (PCCS). After the remote attestation service determines that an SGX enclave quote is trusted, the TensorFlow Serving inference service sends a key from files/wrap-key in the current directory to the remote application. In this example, the remote application is LibOS Gramine in the vSGX environment. After LibOS Gramine receives keys from the inference service, LibOS Gramine decrypts the encrypted model file and TLS configuration file.

   i. Switch to the secrec_prov_server directory.

   ```
   cd /home/test/confidential-computing/Tensorflow_Serving/docker/secret_prov
   ```

   ii. Use the image of the key authentication service.

   You can use one of the following methods to use the image of the key authentication service:

   - Download the image of the key authentication service.

   ```
   docker pull registry.cn-beijing.aliyuncs.com/tee_sgx/secrec_prov_server:v1
   ```

   - Compile the image based on a script.

   ```
   ./build_secret_prov_image.sh
   ```

   iii. Obtain the image ID.

   ```
   docker images
   ```

   iv. Start the key authentication service.

   ```
   ./run_secret_prov.sh -i image_id
   ```

   After the service is started, it runs in the background to wait for remote attestation requests. After the service receives a remote attestation request from a remote end and authenticates the request to determine that the remote end is trusted, the service sends a key to the remote end.

   v. View the secret_prov_server logs.

   ```
   docker logs image_id
   ```

   A command output similar to the following one indicates that a remote attestation request has been received.

   ```
   Received the following measurements from the client:
     - MRENCLAVE:    cc24a9a84248a8c622e07bb36fa269c5cfaed84d908f80541a341c4b7
     - MRSIGNER:     2ebc89a3ef2ed34f9a498a9a495d49f49dd25c837ec0409adeac3c4d4
     - ISV_PROD_ID: 0
     - ISV_SVN:     0
   [ WARNING: In reality, you would want to compare against expected values! ]
   --- Sent secret1 = 'ffeeddccbbaa99887766554433' ---
   ```

What's next

After the client is deployed, it waits for the vSGX client to start the inference service and send a remote

attestation request. For information about how to deploy the vSGX client, see Step 2: Deploy a vSGX client.

4.7.5.3. Step 2: Deploy a vSGX client

After a client is deployed, you must deploy a virtual Software Guard Extensions (vSGX) client. A vSGX client is used to run the TensorFlow Serving inference service.

Context

The vSGX client provides a vSGX encrypted computing environment in which the TensorFlow Serving inference service can run. When the inference service starts, it sends a remote attestation request to a client to verify the feasibility of the current vSGX environment and the integrity of the inference service. If the verification succeeds, the inference service receives a key from the client and decrypts the encrypted model and Transport Layer Security (TLS) certificate. Then, the inference service starts to run in the vSGX environment and waits for remote access requests to perform inference.

Procedure

1. Create a security-enhanced Elastic Compute Service (ECS) instance and build an SGX encrypted computing environment.
   i. Create a security-enhanced ECS instance.

      For more information, see Create security-enhanced instances.

      Take note of the following parameters:

      - **Instance Type**: Select an instance type with 32 GiB or more of encrypted memory. In this example, the ecs.g7t.4xlarge instance type is used.
      - **Image**: Select the Alibaba Cloud Linux 2.1903 LTS 64-bit (UEFI) public image.
      - **Public IP Address**: Select Assign Public IPv4 Address.

   ii. Build an SGX encrypted computing environment.

      For more information, see Build an SGX encrypted computing environment.

   iii. Install Docker.

      For more information, see Deploy and use Docker on Alibaba Cloud Linux 2 instances.
2. Log on to the ECS instance.

   For more information, see Connection methodsGuidelines on instance connection.
3. Switch to the working directory (example: /home/tf) and download the desired TensorFlow Serving script code.

   ```
   cd /home/tf
   git clone https://gitee.com/cloud_cc/confidential-computing.git
   ```
4. Copy the ssl_configure and models folders from a client to the tf_serving folder on the vSGX client.

   ```
   scp -r user_name@192.168.XX.XX:{models} {tf_serving}
   scp -r user_name@192.168.XX.XX:{ssl_configure} {tf_serving}
   ```

user_name specifies the username of the client. 192.168.XX.XX indicates the IP address of the client. {models} specifies the path to the models folder on the client. *{ssl_configure}* specifies the path to the ssl_configure folder on the client. {tf_serving} specifies the path to the tf_serving folder on the vSGX client. Set the preceding parameters to the actual values. In this example, the following values are used:

```
scp -r root@192.168.XX.XX:/home/tf/confidential-computing/Tensorflow_Serving/client/mod
els  /home/tf/confidential-computing/Tensorflow_Serving/docker/tf_serving
scp -r root@192.168.XX.XX:/home/tf/confidential-computing/Tensorflow_Serving/client/ssl
_configure /home/tf/confidential-computing/Tensorflow_Serving/docker/tf_serving
```

5. Obtain a TensorFlow Serving image.

   You can use one of the following methods to obtain a TensorFlow Serving image:

   - Download a TensorFlow Serving container image.

     ```
     cd /home/tf/confidential-computing/Tensorflow_Serving/docker/tf_serving
     docker pull registry.cn-beijing.aliyuncs.com/tee_sgx/tf_serving:latest
     ```

   - Create a TensorFlow Serving image.

     You can run the build_gramine_tf_serving.sh script in the /Tensorflow_Serving/docker directory to create a TensorFlow Serving image.

     ```
     cd /home/test/confidential-computing/Tensorflow_Serving/docker
     ./build_gramine_tf_serving_image.sh image_tag    #You can specify the image_tag param
     eter.
     ```

     The Dockerfile named gramine_tf_serving.dockerfile is used to create the TensorFlow Serving image. This Dockerfile contains the following commands:

     - Commands used to install required dependency libraries in the container
     - Commands used to install TensorFlow Serving in the container
     - Commands used to install Gramine in the container

- Commands used to copy files from the host to the container, including the following files:
  - Makefile: used by Gramine to compile TensorFlow Serving.
  - tensorflow_model_server.manifest.template: used by Gramine to configure a TensorFlow Serving template file named tensorflow_model_server.manifest.attestation.template.

    Take note of the following items about tensorflow_model_server.manifest.attestation.template:

    - Gramine supports remote attestation based on SGX Remote Attestation (RA)-TLS. To use the remote attestation feature, you must enable this feature in the template file. The template used in this practice contains the following parameters:

      SECRET_PROVISION_SERVERS specifies the endpoint of the remote end that sends a remote attestation request. SECRET_PROVISION_SET_PF_KEY specifies the key that must be returned from the remote end to decrypt encrypted files.

      ```
      sgx.remote_attestation = 1
      loader.env.LD_PRELOAD = "libsecret_prov_attest.so"
      loader.env.SECRET_PROVISION_CONSTRUCTOR = "1"
      loader.env.SECRET_PROVISION_SET_PF_KEY = "1"
      loader.env.SECRET_PROVISION_CA_CHAIN_PATH ="certs/test-ca-sha256.crt"
      loader.env.SECRET_PROVISION_SERVERS ="attestation.service.com:4433"
      sgx.trusted_files.libsecretprovattest ="file:libsecret_prov_attest.so"
      sgx.trusted_files.cachain= "file:certs/test-ca-sha256.crt"
      ```

    - Gramine provides sgx.protected_file for you to specify the files that you want to encrypt for protection. Example:

      ```
      sgx.protected_files.model= "file:models/resnet50-v15-fp32/1/saved_model.pb"
      ```

      The encrypted model file and TLS configuration file that are generated on the user side are transmitted over the network and stored in the TensorFlow_Serving directory.
  - sgx_default_qcnl.conf: This file contains the endpoint of Alibaba Cloud Provisioning Certificate Caching Service (PCCS).
  - tf_serving_entrypoint.sh: This script is executed after the container starts.

6. Add the domain name of the client to the /etc/hosts file.

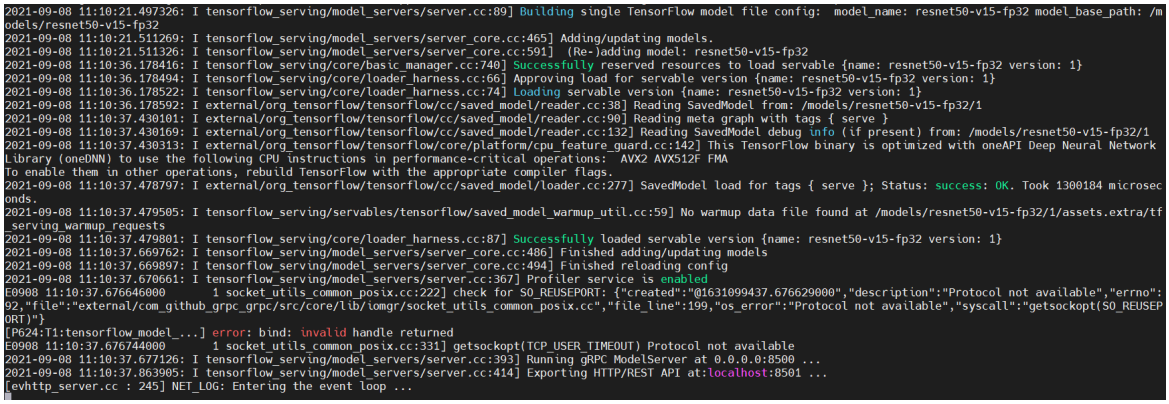
   ```
   echo "remote_ip attestation.service.com" >> /etc/hosts   #Set remote_ip to the IP addre
   ss of the client.
   ```

7. Run TensorFlow Serving.

   After TensorFlow Serving starts, it waits for remote access requests.

   ```
   cd /home/tf/confidential-computing/Tensorflow_Serving/docker/tf_serving
   cp ssl_configure/ssl.cfg .
   ./run_gramine_tf_serving.sh -i ${image_id} -p 8500-8501 -m resnet50-v15-fp32 -s ssl.cfg
   -a attestation.service.com:remote_ip
   ```

> **Note**
>
> - Set image_id to the ID of the TensorFlow Serving image.
> - -p 8500-8501 indicates the host ports that correspond to TensorFlow Serving.
> - Set remote_ip to the IP address of the client.



What's next

After TensorFlow Serving starts, it waits for remote access requests. For information about how to perform remote access, see Step 3: Perform remote access.

4.7.5.4. Step 3: Perform remote access

After the TensorFlow Serving inference service starts to run in the virtual Software Guard Extensions (vSGX) encrypted computing environment, you can configure your client to send data to the inference service. After inference is complete, the inference service returns the inference results.

Procedure

1. Log on to the Elastic Compute Service (ECS) instance on which a client is deployed.

   For more information, see Connection methodsGuidelines on instance connection.

   > **Note** In this example, the client is used as a remote end to initiate an access request.

2. Configure a TensorFlow Serving domain name.

   ```
   vSGX_ip_addr=192.168.XX.XX    #vSGX_ip_addr specifies the IP address of the vSGX instance. Set this parameter to the actual IP address of your vSGX instance.
   service_domain_name=grpc.tf-serving.service.com
   echo "${vSGX_ip_addr} ${service_domain_name}" >> /etc/hosts
   ```

3. Configure the client to send a remote access request.

   The remote access request carries data to the inference service that runs in the vSGX encrypted computing environment. After inference is complete, the inference service returns the inference results.

   ```
   cd /home/tf/confidential-computing/Tensorflow_Serving/client
   python3 ./resnet_client_grpc.py --url ${service_domain_name}:8500 --crt `pwd -P`/ssl_configure/server.crt --batch 1 --cnum 1 --loop 50
   ```

4.7.6. Deploy a PyTorch deep learning model on a security-enhanced instance

This topic describes how to deploy a PyTorch deep learning model on an Intel® SGX-based security-enhanced instance and how to use the PyTorch model.

Context

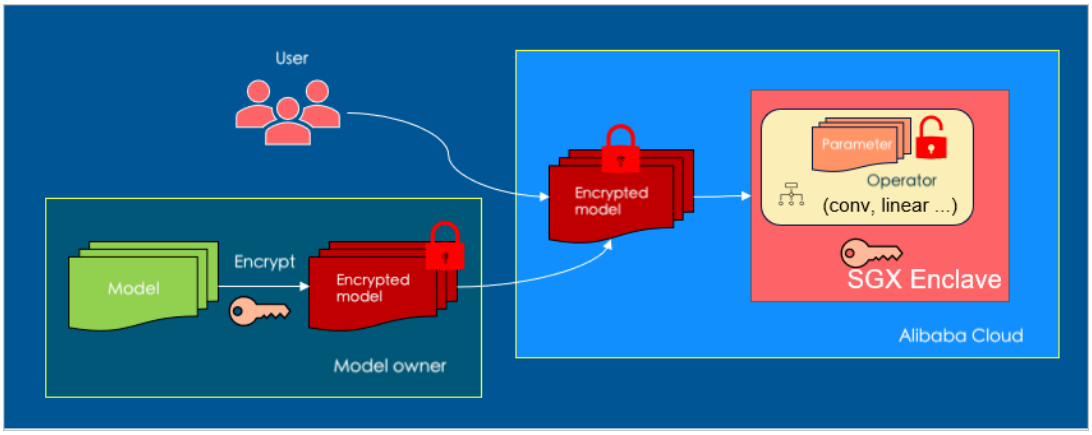
Artificial intelligence (AI) models are built upon large amounts of training data and computing power and are an extremely valuable form of intellectual property. PyTorch is widely recognized by AI developers for its flexible and dynamic programming environment, dynamic graph mechanism, and flexible networking architecture. Typically, PyTorch models are deployed on cloud servers provided by cloud service providers, such as Alibaba Cloud Elastic Compute Service (ECS) instances. All PyTorch model users and cloud service providers need to make sure that PyTorch models deployed on the public cloud are available but invisible and cannot be stolen by others.

Some security-enhanced ECS instances provide encrypted computing capabilities based on Intel® Software Guard Extension (SGX) to create a hardware-level trusted confidential environment that offers a high degree of security. This ensures the confidentiality and integrity of essential code and data and protects them from malware attacks.

You can deploy PyTorch deep learning models in the trusted confidential environment of security-enhanced ECS instances. This ensures the security of data transmission and data usage and the integrity of PyTorch deep learning applications.

Architecture

Architecture



The parameters of the SGX-based PyTorch end-to-end security model are shown in the Architecture figure. The model is stored in ciphertext at the deployment phase. Related operations are performed within the SGX enclave. The model parameters are decrypted only within the SGX enclave, and the keys are transmitted by using the secure remote attestation channel.

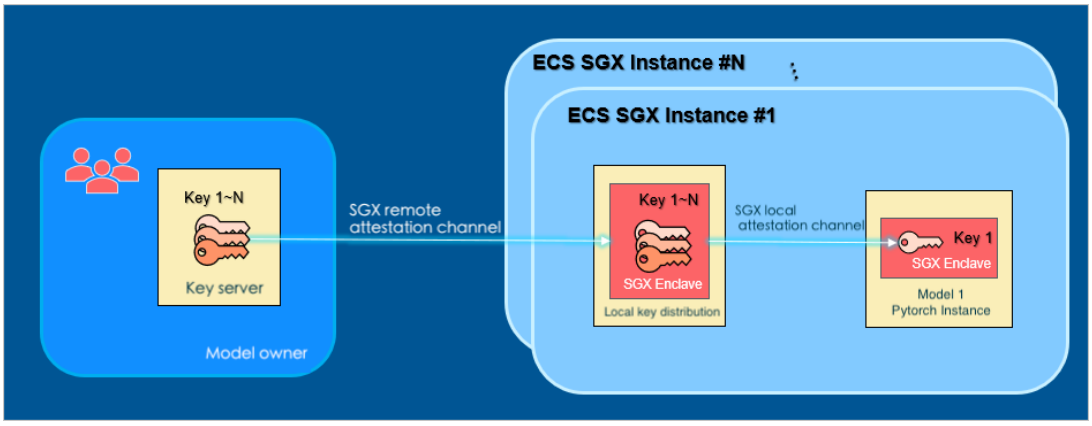
This practice involves three roles: dkeyserver, dkeycache, and PyTorch with SGX. Procedure shows these roles.

- dkeyserver: the key server, which is deployed on the on-premises computer of the PyTorch model user. The PyTorch model user first encrypts the PyTorch model parameters by using the tools provided by PyTorch with SGX and then builds the on-premises key server dkeyserver. The encrypted

model is then deployed on the Alibaba Cloud ECS SGX instance. The key server manages all model keys and model IDs and receives key requests from the key distribution service of the ECS SGX instance.

- dkeycache: the key distribution service, which is deployed on the ECS SGX instance. The key distribution service of the ECS SGX instance first requests all model keys from the key server. After the key server completes the SGX remote attestation, it sends the keys to the SGX enclave for the key distribution service of the ECS SGX instance by using the secure remote attestation channel. This operation is automatically completed after the key distribution server is started.
- PyTorch with SGX: the ECS SGX instance that runs PyTorch, which is deployed on the same server as dkeycache. When an Alibaba Cloud PyTorch instance uses models to make predictions or perform classification tasks for model inferences, it automatically sends a request for a model key to the key distribution service. The key is encrypted and sent to the SGX enclave of the PyTorch instance by using the SGX secure channel. The enclave started by PyTorch with SGX uses the key to decrypt the model parameters and perform model prediction operations. Model parameters are protected by SGX-based hardware throughout the process and are available but invisible, which ensures the security of data transmission and data usage.

Procedure



Procedure

In this practice, dkeyserver, dkeycache, and PyTorch with SGX are deployed on the same security-enhanced instance for easy verification.

1. Create a security-enhanced instance and install the packages required to run PyTorch.
   i. Create a security-enhanced instance.

      For more information about how to create a security-enhanced instance, see Create security-enhanced instances.

      Take note of the following parameters:

      - **Instance Type**: Select an instance type with at least 4 vCPUs and 16 GiB of memory. In this topic, ecs.g7t.4xlarge is used.
      - **Image**: Select Alibaba Cloud Linux 2.1903 LTS 64-bit (UEFI).
      - **Public IP Address**: Select Assign Public IPv4 Address.

   ii. Build an SGX encrypted computing environment.

      For more information about how to create a security-enhanced instance, see Build an SGX encrypted computing environment.

iii. Install the packages required to run PyTorch.

PyTorch has requirements on the version of software such as Python and GCC. Run the following commands to install the specified version of software:

```
yum update --skip-broken
sudo yum install -y teesdk git gcc-c++ scl-utils alinux-release-experimentals  pyth
on36-devel
sudo yum install -y devtoolset-7-gcc devtoolset-7-gdb devtoolset-7-binutils devtool
set-7-make devtoolset-7-gcc devtoolset-7-gcc-c++
scl -l devtoolset-7
ln -sf /opt/rh/devtoolset-7/root/bin/g++ /usr/bin/g++
ln -sf /opt/rh/devtoolset-7/root/bin/gcc /usr/bin/gcc
ln -sf /opt/rh/devtoolset-7/root/bin/c++ /usr/bin/c++
ln -sf /usr/bin/python3 /usr/bin/python
```

If the page shown in the following figure appears, the packages are installed.

iv. Install the PyTorch dependency library, encryption and decryption dependency library, and CMake.

```
pip3 install --upgrade pip
pip3 install astunparse numpy ninja pyyaml mkl mkl-include setuptools cmake cffi ty
ping_extensions future six requests dataclasses setuptools_rust pycryptodomex pycry
ptodome torchvision
ln -sf /usr/local/bin/cmake /usr/bin/cmake
ln -sf /usr/local/bin/cmake /bin/cmake
```

2. Log on to the ECS instance.

For more information, see Connection methodsGuidelines on instance connection.

3. Switch to the working directory such as /home/test and obtain the PyTorch sample code.

The sample code contains code of dkeyserver, dkeycache, and PyTorch with SGX.

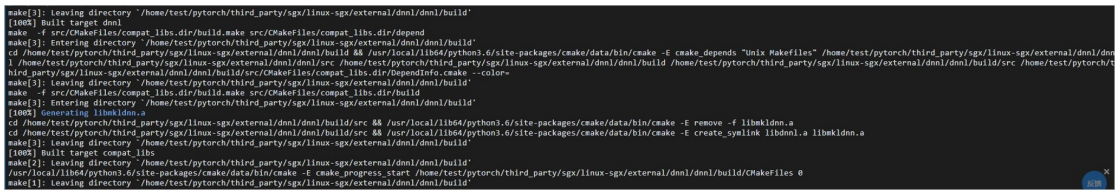
```
cd /home/PyTorch
git clone https://github.com/intel/sgx-pytorch -b sgx pytorch
cd /home/test/pytorch
git submodule sync && git submodule update --init --recursive
```

4. Compile PyTorch with SGX in the ECS SGX instance.

i. Compile oneAPI Deep Neural Network Library (oneDNN).

oneDNN is an open-source cross-platform performance library of basic building blocks for deep learning applications. This library is optimized for Intel Architecture Processors, Intel Processor Graphics, and Xe Architecture graphics. oneDNN is intended for developers of deep learning applications and models who are interested in improving application performance on Intel CPUs and GPUs.

```
source /opt/alibaba/teesdk/intel/sgxsdk/environment
cd /home/test/pytorch/third_party/sgx/linux-sgx
git am ../0001*
cd external/dnnl
make
sudo cp sgx_dnnl/lib/libsgx_dnnl.a /opt/alibaba/teesdk/intel/sgxsdk/lib64/libsgx_dnnl2.a
sudo cp sgx_dnnl/include/*  /opt/alibaba/teesdk/intel/sgxsdk/include/
```



ii. Compile the PyTorch enclave.

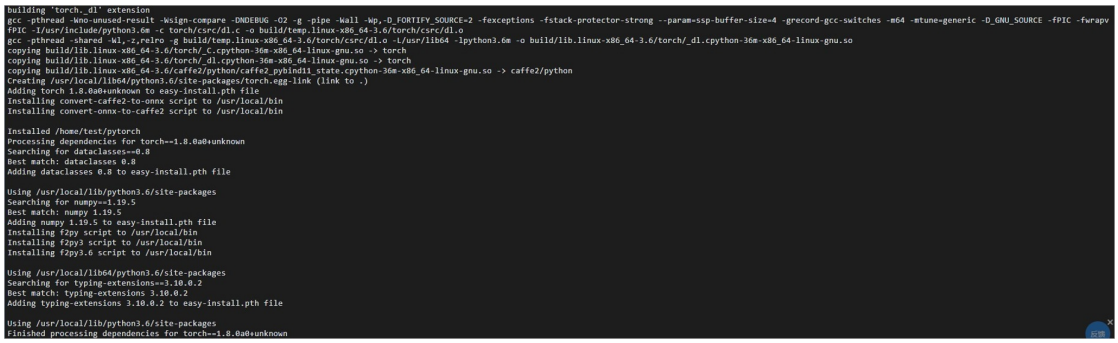
The enclave of PyTorch with SGX performs model parameter decryption and model prediction operations.

```
source /opt/alibaba/teesdk/intel/sgxsdk/environment
cd /home/test/pytorch/enclave_ops/ideep-enclave
make
```

```
LINK =>  libenclave.so
<EnclaveConfiguration>
    <ProdID>0</ProdID>
    <ISVSVN>0</ISVSVN>
    <StackMinSize>0x100000</StackMinSize>
    <StackMaxSize>0x100000</StackMaxSize>
    <HeapMinSize>0x100000</HeapMinSize>
    <HeapInitSize>0x4000000</HeapInitSize>
    <HeapMaxSize>0x9000000</HeapMaxSize>
    <TCSNum>8</TCSNum>
    <TCSMinPool>8</TCSMinPool>
    <TCSMaxNum>8</TCSMaxNum>
    <TCSPolicy>1</TCSPolicy>
    <!-- Recommend changing 'DisableDebug' to 1 to make the enclave undebuggable for enclave release -->
    <DisableDebug>0</DisableDebug>
    <MiscSelect>0</MiscSelect>
    <MiscMask>0xFFFFFFFF</MiscMask>
    <ReservedMemMinSize>0x1000000</ReservedMemMinSize>
    <ReservedMemInitSize>0x1000000</ReservedMemInitSize>
    <ReservedMemMaxSize> 0x1000000</ReservedMemMaxSize>
    <!-- On SGX1 platform, ReservedMemExecutable==1 means set reserved memory as read, write and execute (RWX) -->
    <ReservedMemExecutable>1</ReservedMemExecutable>
</EnclaveConfiguration>
tcs_num 8, tcs_max_num 8, tcs_min_pool 8
The required memory is 116727808B.
The required memory is 0x6f52000, 113992 KB.
Succeed.
SIGN =>  libenclave.signed.so
The project has been built in debug hardware mode.
make[1]: Leaving directory `/home/test/pytorch/enclave_ops/ideep-enclave'
```

iii. Compile PyTorch.

```
cd /home/test/pytorch
pip3 uninstall torch    #Uninstall the installed PyTorch. Then, install the self-co
mpiled PyTorch.
source /opt/alibaba/teesdk/intel/sgxsdk/environment
python setup.py develop --cmake-only
sudo python setup.py develop && python -c "import torch"
```



iv. Compile the secure PyTorch computing operator.

```
source /opt/alibaba/teesdk/intel/sgxsdk/environment
cd /home/test/pytorch/enclave_ops/secure_op && mkdir build && cd build
cmake -DCMAKE_PREFIX_PATH="$(python -c 'import torch.utils; print(torch.utils.cmake
_prefix_path)')" ..
make
```

5. Compile and generate the dkeyserver executable file on the key server and the dkeycache executable file on the ECS SGX instance.

```
cd /home/test/pytorch/enclave_ops/deployment
make
```

```
In file included from App/socket_server.cpp:50:0:
App/socket_server.h:48:0: note: this is the location of the previous definition
 #define SAFE_FREE(ptr) {if (NULL != (ptr)) {free(ptr); (ptr) = NULL;}}

CXX  <=  App/socket_server.cpp
CXX  <=  App/main.cpp
LINK =>  dkeyserver
GEN  =>  Enclave/enclave_t.h
CC   <=  Enclave/enclave_t.c
CXX  <=  Enclave/enclave.cpp
LINK =>  libenclave-dkeyserver.so
<EnclaveConfiguration>
    <ProdID>0</ProdID>
    <ISVSVN>0</ISVSVN>
    <StackMaxSize>0x40000</StackMaxSize>
    <HeapMaxSize>0xA00000</HeapMaxSize>
    <TCSNum>8</TCSNum>
    <TCSPolicy>1</TCSPolicy>
    <DisableDebug>0</DisableDebug>
    <MiscSelect>0</MiscSelect>
    <MiscMask>0xFFFFFFFF</MiscMask>
</EnclaveConfiguration>
tcs_num 8, tcs_max_num 8, tcs_min_pool 1
The required memory is 13131776B.
The required memory is 0xc86000, 12824 KB.
Succeed.
SIGN =>  libenclave-dkeyserver.signed.so
make[2]: Entering directory `/home/test/pytorch/enclave_ops/deployment/dkeyserver'
make[2]: Nothing to be done for `target'.
make[2]: Leaving directory `/home/test/pytorch/enclave_ops/deployment/dkeyserver'
make[1]: Leaving directory `/home/test/pytorch/enclave_ops/deployment/dkeyserver'
The project has been built in hardware debug mode.
```

6. Start the key service on the key server.

```
cd /home/test/pytorch/enclave_ops/deployment/bin/dkeyserver
sudo ./dkeyserver
```

The key server starts and waits for key requests from the dkeycache service deployed on the ECS SGX instance.

```
Waiting for incoming connections...
New Client(4) connected! IP=127.0.0.1
receive the msg type(1) from client.
Dispatching TYPE_RA_MSG1, body size: 68
send response success with msg type(2)
receive the msg type(3) from client.
Dispatching TYPE_RA_MSG3, body size: 5064
cert_key_type = 0x5
        Info: sgx_qv_get_quote_supplemental_data_size successfully returned.
        Info: App: sgx_qv_verify_quote successfully returned.
        Info: App: Verification quote_verification_result=0
send response success with msg type(4)
```

7. Compile dkeycache on the ECS SGX instance and start the key distribution service.

```
cd /home/test/pytorch/enclave_ops/deployment/bin/dkeycache
sudo ./dkeycache
```

After startup, dkeycache requests all model keys from dkeyserver. After dkeyserver completes the SGX remote attestation, dkeyserver sends the keys to the SGX enclave of dkeyserver by using the secure remote attestation channel.

```
Connect dkeyserver success!
Call sgx_ra_get_msg1_ex success, the MSG1 body generated.
Sending MSG1 to remote attestation service provider, and expecting MSG2 back...
MSG2 recieved success!
Call sgx_ra_proc_msg2_ex success.
Sending MSG3 to remote attestation service provider,expecting attestation result msg back...
Attestation result MSG recieved success!
Verify attestation result is succeed!
Enclave: Decrypt the serect success
Enclave: hyhyhy g_model_keys 0:0.
Enclave: hyhyhy g_model_keys 1:0.
Enclave: hyhyhy g_model_keys 2:0.
Enclave: hyhyhy g_model_keys 3:0.
Enclave: hyhyhy g_model_keys 4:0.
Enclave: hyhyhy g_model_keys 5:0.
Enclave: hyhyhy g_model_keys 6:0.
Enclave: hyhyhy g_model_keys 7:0.
Enclave: hyhyhy g_model_keys 8:0.
Enclave: hyhyhy g_model_keys 9:0.
Enclave: hyhyhy g_model_keys 10:0.
Enclave: hyhyhy g_model_keys 11:0.
Enclave: hyhyhy g_model_keys 12:0.
Enclave: hyhyhy g_model_keys 13:0.
Enclave: hyhyhy g_model_keys 14:0.
Enclave: hyhyhy g_model_keys 15:0.
Enclave: hyhyhy g_model_keys 16:0.
Enclave: hyhyhy g_model_keys 17:0.
Enclave: hyhyhy g_model_keys 18:0.
Enclave: hyhyhy g_model_keys 19:0.
Enclave: hyhyhy g_model_keys 20:1.
Enclave: hyhyhy g_model_keys 21:0.
Enclave: hyhyhy g_model_keys 22:0.
Enclave: hyhyhy g_model_keys 23:0.
Enclave: hyhyhy g_model_keys 24:255.
Enclave: hyhyhy g_model_keys 25:255.
Enclave: hyhyhy g_model_keys 26:255.
Enclave: hyhyhy g_model_keys 27:255.
Enclave: hyhyhy g_model_keys 28:255.
Enclave: hyhyhy g_model_keys 29:255.
Enclave: hyhyhy g_model_keys 30:255.
Enclave: hyhyhy g_model_keys 31:255.
Enclave: hyhyhy g_model_keys 32:255.
Enclave: hyhyhy g_model_keys 33:255.
Enclave: hyhyhy g_model_keys 34:255.
Enclave: hyhyhy g_model_keys 35:255.
Enclave: hyhyhy g_model_keys 36:255.
Enclave: hyhyhy g_model_keys 37:255.
Enclave: hyhyhy g_model_keys 38:255.
Enclave: hyhyhy g_model_keys 39:255.
Successfully received the DomainKey from deploy server.
Call enclave_ra_close success.
dkeycache service is ON...
```

8. Run ResNet-based test cases on the ECS SGX instance.

```
cd /home/test/pytorch/enclave_ops/test
sudo python whole_resnet.py
```

The ciphertext parameters of the PyTorch model are decrypted in the SGX enclave. The keys are obtained from dkeycache and then encrypted and transmitted to the enclave.

```
Samoyed 0.8732958436012268
Pomeranian 0.030270852148532867
white wolf 0.019671205431222916
keeshond 0.011073529720306396
Eskimo dog 0.009204281494021416
--- 0.25887560844421387 seconds ---
press any key to start mkldnn...
--- 0.08027315139770508 seconds ---
call in mkldnn inner product with biassgx_destroy_enclave ret is 0.
[root@iZuf6igaqeekgwethu3gieZ test]#
```

4.8. Compute optimized type family with GPU

4.8.1. GPU-accelerated compute-optimized and vGPU-accelerated instance families

This topic describes the features of GPU-accelerated compute-optimized and vGPU-accelerated instance families of Elastic Compute Service (ECS) and lists the instance types of each instance family.

- Recommended instance families
  - sgn7i-vws, vGPU-accelerated instance family with shared CPUs
  - vgn7i-vws, vGPU-accelerated instance family
  - gn7s, GPU-accelerated compute-optimized instance family
  - gn7i, GPU-accelerated compute-optimized instance family
  - gn7, GPU-accelerated compute-optimized instance family
  - vgn6i, vGPU-accelerated instance family
  - gn6i, GPU-accelerated compute-optimized instance family
  - gn6e, GPU-accelerated compute-optimized instance family
  - gn6v, GPU-accelerated compute-optimized instance family
- Other available instance families (If these instance families are sold out, you can use the recommended ones.)
  - vgn5i, vGPU-accelerated instance family
  - gn5, GPU-accelerated compute-optimized instance family
  - gn5i, GPU-accelerated compute-optimized instance family

sgn7i-vws, vGPU-accelerated instance family with shared CPUs

Features:

- This instance family uses third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance. This instance family utilizes fast path acceleration on chips to improve storage performance, network performance, and computing stability by an order of magnitude. This way, data storage and model loading can be performed more quickly.
- Instances of the sgn7i-vws instance family share CPU and network resources to maximize the utilization of underlying resources. Each instance has exclusive access to its memory and GPU memory

to ensure data isolation and high performance.

> **Note** If you want to use exclusive CPU resources, select the vgn7i-vws instance family.

- This instance family comes with a NVIDIA GRID vWS license and provides certified graphics acceleration capabilities for Computer Aided Design (CAD) software to meet the requirements of professional graphic design. Instances of this instance family can serve as lightweight GPU-accelerated compute-optimized instances to reduce the costs of small-scale AI inference tasks.
- Compute:
  - Uses NVIDIA A10 GPUs that have the following features:
    - Innovative NVIDIA Ampere architecture
    - Support for acceleration features (such as vGPU, RTX, and TensorRT) to provide all-purpose business support
  - Uses 2.9 GHz Intel® Xeon® Scalable (Ice Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports enhanced SSDs (ESSDs), standard SSDs, and ultra disks.

> **Note** For more information about the performance of cloud disks, see EBS performance.

- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Concurrent AI inference tasks that require high-performance CPUs, memory, and GPUs, such as image recognition, speech recognition, and behavior identification
  - Compute-intensive graphics processing tasks that require high-performance 3D graphics virtualization capabilities, such as remote graphic design and cloud gaming
  - 3D modeling in fields that require the use of Ice Lake processors, such as animation and film production, cloud gaming, and mechanical design

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Network interface controller (NIC) queues	Elastic network interfaces (ENIs)
ecs.sgn7i-vws-m2.xlarge	4	15.5	NVIDIA A10 × 1/12	24 GB × 1/12	1.5/5	500,000	4	2

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Baseline/burst bandwidth (Gbit/s)	Packet forwarding rate (pps)	Network interface controller (NIC) queues	Elastic network interfaces (ENIs)
ecs.sgn7i-vws-m4.2xlarge	8	31	NVIDIA A10 × 1/6	24 GB × 1/6	2.5/10	1,000,000	4	4
ecs.sgn7i-vws-m8.4xlarge	16	62	NVIDIA A10 × 1/3	24 GB × 1/3	5/20	2,000,000	8	4
ecs.sgn7i-vws-m2s.xlarge	4	8	NVIDIA A10 × 1/12	24 GB × 1/12	1.5/5	500,000	4	2
ecs.sgn7i-vws-m4s.2xlarge	8	16	NVIDIA A10 × 1/6	24 GB × 1/6	2.5/10	1,000,000	4	4
ecs.sgn7i-vws-m8s.4xlarge	16	32	NVIDIA A10 × 1/3	24 GB × 1/3	5/20	2,000,000	8	4

Note

-
- For more information about these specifications, see Instance family.

vgn7i-vws, vGPU-accelerated instance family

Features:

- This instance family uses third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance. This instance family utilizes fast path acceleration on chips to improve storage performance, network performance, and computing stability by an order of magnitude. This way, data storage and model loading can be performed more quickly.
- This instance family comes with a NVIDIA GRID vWS license and provides certified graphics acceleration capabilities for Computer Aided Design (CAD) software to meet the requirements of professional graphic design. Instances of this instance family can serve as lightweight GPU-accelerated compute-optimized instances to reduce the costs of small-scale AI inference tasks.
- Compute:

  - Uses NVIDIA A10 GPUs that have the following features:
    - Innovative NVIDIA Ampere architecture
    - Support for acceleration features (such as vGPU, RTX, and TensorRT) to provide all-purpose business support
  - Uses 2.9 GHz Intel® Xeon® Scalable (Ice Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.

    Note For more information about the performance of cloud disks, see EBS performance.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Concurrent AI inference tasks that require high-performance CPUs, memory, and GPUs, such as image recognition, speech recognition, and behavior identification
  - Compute-intensive graphics processing tasks that require high-performance 3D graphics virtualization capabilities, such as remote graphic design and cloud gaming
  - 3D modeling in fields that require the use of Ice Lake processors, such as animation and film production, cloud gaming, and mechanical design

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs
ecs.vgn7i-vws-m4.xlarge	4	30	NVIDIA A10 × 1/6	24 GB × 1/6	3	1,000,000	4	4
ecs.vgn7i-vws-m8.2xlarge	10	62	NVIDIA A10 × 1/3	24 GB × 1/3	5	2,000,000	8	6
ecs.vgn7i-vws-m12.3xlarge	14	93	NVIDIA A10 × 1/2	24 GB × 1/2	8	3,000,000	8	6

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs
ecs.vgn7i-vws-m24.7xlarge	30	186	NVIDIA A10 × 1	24 GB × 1	16	6,000,000	12	8

Note

-
- For more information about these specifications, see Instance family.

gn7s, GPU-accelerated compute-optimized instance family

Features

- This instance family uses new Intel Ice lake processors and NVIDIA A30 GPUs that are based on NVIDIA Ampere architecture. You can choose an appropriate mix of GPUs and CPU resources to meet various AI business requirements.
- This instance family uses third-generation SHENLONG architecture and doubles the average bandwidths of virtual private clouds (VPCs), networks, and disks compared with instance families of the previous generation.
- Compute:
  - Uses NVIDIA A30 GPUs that have the following features:
    - Innovative NVIDIA Ampere architecture
    - Support for the multi-instance GPU (MIG) feature and acceleration features (based on second-generation Tensor cores) to provide all-purpose business support
  - Uses 2.9 GHz Intel® Xeon® Scalable (Ice Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
  - Improves memory sizes significantly from instance families of the previous generation.
- Storage: Supports only ESSDs and ultra disks.

  Note For more information about the performance of cloud disks, see EBS performance.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios: concurrent AI inference tasks that require high-performance CPUs, memory, and GPUs, such as image recognition, speech recognition, and behavior identification.

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	IPv6 addresses per ENI	NIC queues	ENIs
ecs.gn7s-c8g1.2xlarge	8	60	NVIDIA A30 × 1	24 GB × 1	16	6,000,000	1	12	8
ecs.gn7s-c16g1.4xlarge	16	120	NVIDIA A30 × 1	24 GB × 1	16	6,000,000	1	12	8
ecs.gn7s-c32g1.8xlarge	32	250	NVIDIA A30 × 1	24 GB × 1	16	6,000,000	1	12	8
ecs.gn7s-c32g1.16xlarge	64	500	NVIDIA A30 × 2	24 GB × 2	32	12,000,000	1	16	15
ecs.gn7s-c32g1.32xlarge	128	1,000	NVIDIA A30 × 4	24 GB × 4	64	24,000,000	1	32	15
ecs.gn7s-c48g1.12xlarge	48	380	NVIDIA A30 × 1	24 GB × 1	16	6,000,000	1	12	8
ecs.gn7s-c56g1.14xlarge	56	440	NVIDIA A30 × 1	24 GB × 1	16	6,000,000	1	12	8

Note

-
- For more information about these specifications, see Instance family.

gn7i, GPU-accelerated compute-optimized instance family

Features:

- This instance family uses third-generation SHENLONG architecture to provide predictable and consistent ultra-high performance. This instance family utilizes fast path acceleration on chips to improve storage performance, network performance, and computing stability by an order of magnitude.
- Compute:
  - Uses NVIDIA A10 GPUs that have the following features:
    - Innovative NVIDIA Ampere architecture
    - Support for acceleration features such as RTX and TensorRT
  - Uses 2.9 GHz Intel® Xeon® Scalable (Ice Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
  - Provides memory of up to 752 GiB, which is much larger than the memory sizes of the gn6i instance family.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Concurrent AI inference tasks that require high-performance CPUs, memory, and GPUs, such as image recognition, speech recognition, and behavior identification
  - Compute-intensive graphics processing tasks that require high-performance 3D graphics virtualization capabilities, such as remote graphic design and cloud gaming

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.gn7i-c8g1.2xlarge	8	30	NVIDIA A10 × 1	24 GB × 1	16	1,600,000	8	4	5
ecs.gn7i-c16g1.4xlarge	16	60	NVIDIA A10 × 1	24 GB × 1	16	3,000,000	8	8	5
ecs.gn7i-c32g1.8xlarge	32	188	NVIDIA A10 × 1	24 GB × 1	16	6,000,000	12	8	5

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.gn7i-c32g1.16xlarge	64	376	NVIDIA A10 × 2	24 GB × 2	32	12,000,000	16	15	5
ecs.gn7i-c32g1.32xlarge	128	752	NVIDIA A10 × 4	24 GB × 4	64	24,000,000	32	15	10
ecs.gn7i-c48g1.12xlarge	48	310	NVIDIA A10 × 1	24 GB × 1	16	9,000,000	16	8	8
ecs.gn7i-c56g1.14xlarge	56	346	NVIDIA A10 × 1	24 GB × 1	16	12,000,000	16	12	8

Note

-
- For more information about these specifications, see Instance family.

gn7, GPU-accelerated compute-optimized instance family

Features:

- Compute:
  - Uses NVIDIA A100 GPUs. NVSwitches are used to establish connections between NVIDIA A100 GPUs. The GPUs have the following features:
    - Innovative NVIDIA Ampere architecture
    - 40 GB HBM2 memory per GPU
  - Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:

- Supports IPv6.
- Provides high network performance based on large computing capacity.

- Supported scenarios:
  - Deep learning applications such as training applications of AI algorithms used in image classification, autonomous vehicles, and speech recognition
  - Scientific computing applications that require robust GPU computing capabilities such as computational fluid dynamics, computational finance, molecular dynamics, and environmental analytics

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs
ecs.gn7-c12g1.3xlarge	12	95	NVIDIA A100 × 1	40 GB × 1	4	2,500,000	4	8
ecs.gn7-c13g1.13xlarge	52	380	NVIDIA A100 × 4	40 GB × 4	15	9,000,000	16	8
ecs.gn7-c13g1.26xlarge	104	760	NVIDIA A100 × 8	40 GB × 8	30	18,000,000	16	16

Note

-
- For more information about these specifications, see Instance family.

vgn6i, vGPU-accelerated instance family

Features:

- If you want your vgn6i instance to support graphics features such as Open Graphics Library (OpenGL), you must purchase a GRID license from NVIDIA. Then, after the instance is created, you must manually install a GRID driver and activate the license.
- Compute:
  - Uses NVIDIA T4 GPUs.
  - Uses vGPUs.
    - Supports the 1/4 and 1/2 computing capacity of NVIDIA Tesla T4 GPUs.
    - Supports 4 GB and 8 GB of GPU memory.
  - Offers a CPU-to-memory ratio of 1:5.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Real-time rendering for cloud gaming
  - Real-time rendering for augmented reality (AR) and virtual reality (VR) applications
  - AI (deep learning and machine learning) inference for elastic Internet service deployment
  - Educational environment of deep learning
  - Modeling experiment environment of deep learning

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.vgn6i-m4.xlarge	4	23	NVIDIA T4 × 1/4	16 GB × 1/4	3	500,000	2	4	10
ecs.vgn6i-m8.2xlarge	10	46	NVIDIA T4 × 1/2	16 GB × 1/2	4	800,000	4	5	20

Note

-
- For more information about these specifications, see Instance family.

gn6i, GPU-accelerated compute-optimized instance family

Features:

- Compute:
  - Uses NVIDIA T4 GPUs that have the following features:
    - Innovative NVIDIA Turing architecture
    - 16 GB memory (320 GB/s bandwidth) per GPU
    - 2,560 CUDA cores per GPU
    - Up to 320 Turing Tensor cores per GPU
    - Mixed-precision Tensor cores that support 65 FP16 TFLOPS, 130 INT8 TOPS, and 260 INT4 TOPS

  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports standard SSDs, ultra disks, and ESSDs that deliver millions of IOPS.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - AI (deep learning and machine learning) inference for computer vision, speech recognition, speech synthesis, natural language processing (NLP), machine translation, and recommendation systems
  - Real-time rendering for cloud gaming
  - Real-time rendering for AR and VR applications
  - Graphics workstations or overloaded graphics computing
  - GPU-accelerated databases
  - High-performance computing

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Baseline storage IOPS	NIC queues	ENIs	Private IP addresses per ENI
ecs.gn6i-c4g1.xlarge	4	15	NVIDIA T4 × 1	16 GB × 1	4	500,000	None	2	2	10
ecs.gn6i-c8g1.2xlarge	8	31	NVIDIA T4 × 1	16 GB × 1	5	800,000	None	2	2	10
ecs.gn6i-c16g1.4xlarge	16	62	NVIDIA T4 × 1	16 GB × 1	6	1,000,000	None	4	3	10
ecs.gn6i-c24g1.6xlarge	24	93	NVIDIA T4 × 1	16 GB × 1	7.5	1,200,000	None	6	4	10

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Baseline storage IOPS	NIC queues	ENIs	Private IP addresses per ENI
ecs.gn6i-c40g1.10xlarge	40	155	NVIDIA T4 × 1	16 GB × 1	10	1,600,000	None	16	10	10
ecs.gn6i-c24g1.12xlarge	48	186	NVIDIA T4 × 2	16 GB × 2	15	2,400,000	None	12	6	10
ecs.gn6i-c24g1.24xlarge	96	372	NVIDIA T4 × 4	16 GB × 4	30	4,800,000	250,000	24	8	10

Note

-
- For more information about these specifications, see Instance family.

gn6e, GPU-accelerated compute-optimized instance family

Features:

- Compute:
  - Uses NVIDIA V100 GPUs that each has 32 GB of GPU memory and support NVLink.
  - Uses NVIDIA V100 GPUs (SXM2-based) that have the following features:
    - Innovative NVIDIA Volta architecture
    - 32 GB HBM2 memory (900 GB/s bandwidth) per GPU
    - 5,120 CUDA cores per GPU
    - 640 Tensor cores per GPU
    - Support for up to six NVLink connections, which each have a bidirectional bandwidth of 50 GB/s for a total bandwidth of 300 GB/s
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.

- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Deep learning applications such as the training and inference applications of AI algorithms used in image classification, autonomous driving, and speech recognition
  - Scientific computing applications, such as computational fluid dynamics, computational finance, molecular dynamics, and environmental analytics

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.gn6e-c12g1.3xlarge	12	92	NVIDIA V100 × 1	32 GB × 1	5	800,000	8	6	10
ecs.gn6e-c12g1.12xlarge	48	368	NVIDIA V100 × 4	32 GB × 4	16	2,400,000	8	8	20
ecs.gn6e-c12g1.24xlarge	96	736	NVIDIA V100 × 8	32 GB × 8	32	4,800,000	16	8	20

Note

-
- For more information about these specifications, see Instance family.

gn6v, GPU-accelerated compute-optimized instance family

Features:

- Compute:
  - Uses NVIDIA V100 GPUs.

- Uses NVIDIA V100 GPUs (SXM2-based) that have the following features:
  - Innovative NVIDIA Volta architecture
  - 16 GB HBM2 memory (900 GB/s bandwidth) per GPU
  - 5,120 CUDA cores per GPU
  - 640 Tensor cores per GPU
  - Support for up to six NVLink connections, which each have a bidirectional bandwidth of 50 GB/s for a total bandwidth of 300 GB/s
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Deep learning applications such as the training and inference applications of AI algorithms used in image classification, autonomous driving, and speech recognition
  - Scientific computing applications, such as computational fluid dynamics, computational finance, molecular dynamics, and environmental analytics

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Baseline storage IOPS	NIC queues	ENIs	Private IP addresses per ENI
ecs.gn6v-c8g1.2xlarge	8	32	NVIDIA V100 × 1	16 GB × 1	2.5	800,000	None	4	4	10
ecs.gn6v-c8g1.8xlarge	32	128	NVIDIA V100 × 4	16 GB × 4	10	2,000,000	None	8	8	20
ecs.gn6v-c8g1.16xlarge	64	256	NVIDIA V100 × 8	16 GB × 8	20	2,500,000	None	16	8	20

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Baseline storage IOPS	NIC queues	ENIs	Private IP addresses per ENI
ecs.gn6v-c10g1.20xlarge	82	336	NVIDIA V100 × 8	16 GB × 8	32	4,500,000	250,000	16	8	20

Note

-
- For more information about these specifications, see Instance family.

vgn5i, vGPU-accelerated instance family

Features:

- If you want your vgn5i instance to support graphics features such as OpenGL, you must purchase a GRID license from NVIDIA. Then, after the instance is created, you must manually install a GRID driver and activate the license.
- Compute:
  - Uses NVIDIA P4 GPUs.
  - Uses vGPUs.
    - Supports the 1/8, 1/4, 1/2, and 1/1 computing capacity of NVIDIA Tesla P4 GPUs.
    - Supports 1 GB, 2 GB, 4 GB, and 8 GB of GPU memory.
  - Offers a CPU-to-memory ratio of 1:3.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Real-time rendering for cloud gaming
  - Real-time rendering for AR and VR applications
  - AI (deep learning and machine learning) inference for elastic Internet service deployment
  - Educational environment of deep learning
  - Modeling experiment environment of deep learning

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.vgn5i-m1.large	2	6	NVIDIA P4 × 1/8	8 GB × 1/8	1	300,000	2	2	6
ecs.vgn5i-m2.xlarge	4	12	NVIDIA P4 × 1/4	8 GB × 1/4	2	500,000	2	3	10
ecs.vgn5i-m4.2xlarge	8	24	NVIDIA P4 × 1/2	8 GB × 1/2	3	800,000	2	4	10
ecs.vgn5i-m8.4xlarge	16	48	NVIDIA P4 × 1	8 GB × 1	5	1,000,000	4	5	20

Note

-
- For more information about these specifications, see Instance family.

gn5, GPU-accelerated compute-optimized instance family

Features:

- Compute:
  - Uses NVIDIA P100 GPUs.
  - Offers multiple CPU-to-memory ratios.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Storage:
  - Supports high-performance local Non-Volatile Memory Express (NVMe) SSDs.
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Deep learning

- Scientific computing applications, such as computational fluid dynamics, computational finance, genomics, and environmental analytics
- Server-side GPU compute workloads such as high-performance computing, rendering, and multi-media encoding and decoding

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.gn5-c4g1.xlarge	4	30	440	NVIDIA P100 × 1	16 GB × 1	3	300,000	1	3	10
ecs.gn5-c8g1.2xlarge	8	60	440	NVIDIA P100 × 1	16 GB × 1	3	400,000	1	4	10
ecs.gn5-c4g1.2xlarge	8	60	880	NVIDIA P100 × 2	16 GB × 2	5	1,000,000	2	4	10
ecs.gn5-c8g1.4xlarge	16	120	880	NVIDIA P100 × 2	16 GB × 2	5	1,000,000	4	8	20
ecs.gn5-c28g1.7xlarge	28	112	440	NVIDIA P100 × 1	16 GB × 1	5	1,000,000	8	8	20
ecs.gn5-c8g1.8xlarge	32	240	1,760	NVIDIA P100 × 4	16 GB × 4	10	2,000,000	8	8	20
ecs.gn5-c28g1.14xlarge	56	224	880	NVIDIA P100 × 2	16 GB × 2	10	2,000,000	14	8	20

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.gn5-c8g1.14xlarge	54	480	3,520	NVIDIA P100 × 8	16 GB × 8	25	4,000,000	14	8	20

Note

-
- For more information about these specifications, see Instance family.

gn5i, GPU-accelerated compute-optimized instance family

Features:

- Compute:
  - Uses NVIDIA P4 GPUs.
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Deep learning inference
  - Server-side GPU compute workloads such as multi-media encoding and decoding

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.gn5i-c2g1.large	2	8	NVIDIA P4 × 1	8 GB × 1	1	100,000	2	2	6
ecs.gn5i-c4g1.xlarge	4	16	NVIDIA P4 × 1	8 GB × 1	1.5	200,000	2	3	10
ecs.gn5i-c8g1.2xlarge	8	32	NVIDIA P4 × 1	8 GB × 1	2	400,000	4	4	10
ecs.gn5i-c16g1.4xlarge	16	64	NVIDIA P4 × 1	8 GB × 1	3	800,000	4	8	20
ecs.gn5i-c16g1.8xlarge	32	128	NVIDIA P4 × 2	8 GB × 2	6	1,200,000	8	8	20
ecs.gn5i-c28g1.14xlarge	56	224	NVIDIA P4 × 2	8 GB × 2	10	2,000,000	14	8	20

Note

-
- For more information about these specifications, see Instance family.

4.8.2. Overview of heterogeneous computing services

Alibaba Cloud heterogeneous computing services is a complete service system with integrated software and hardware that helps you allocate and scale resources in a flexible and elastic manner, increase computing power, and control costs.

Heterogeneous computing

Heterogeneous computing is a systematic computing method that consists of computing units of different instruction set and architecture types. Alibaba Cloud heterogeneous computing service family includes Elastic GPU Service, FPGA as a Service (FaaS), and Elastic Accelerated Computing Instances (EAIS). In heterogeneous computing, dedicated hardware is used to serve their most suitable scenarios. This way, in specific scenarios, heterogeneous computing instances can achieve efficiency and cost-effectiveness higher than those of common Elastic Compute Service (ECS) instances by one or more orders of magnitude. Heterogeneous computing is a technology that offers a balance of performance, cost, and power consumption to optimize performance and costs.

The rapid development of AI technologies such as deep learning has given birth to increasingly complex and accurate AI computing models and a significant increase in demand for computing power and performance. Against this backdrop, more and more AI computing services accelerate their performance by using heterogeneous computing. Cloud-based AI accelerators developed by Alibaba Cloud for heterogeneous computing services use a unified framework to accelerate major AI computing frameworks such as TensorFlow, PyTorch, MxNet, and Caffe and optimize the performance of Ethernet and heterogeneous accelerators.

Heterogeneous computing service family

This section describes the Alibaba Cloud heterogeneous computing service family, which consists of the following services: Elastic GPU Service, FaaS, Apsara AI Accelerator (AIACC), FastGPU, cGPU, and EAIS.

- Elastic GPU Service

  GPU-accelerated instances are computing servers based on GPUs. GPUs have unique advantages over CPUs in mathematical and geometric computations such as floating-point and parallel computing and can provide 100 times the computing power of CPUs. GPU-accelerated instances combine the computing power of GPUs and CPUs and provide ready-to-use, scalable GPU compute resources for a variety of scenarios such as AI, high-performance computing, and professional graphics processing. For more information, see What is Elastic GPU Service?.

- AIACC

  AIACC is an AI acceleration engine developed by Alibaba Cloud based on Infrastructure as a Service (IaaS) resources. It optimizes the models built on mainstream AI computing frameworks to achieve significant gains in training and inference performance in deep learning scenarios. You can use AIACC in conjunction with the cluster deployment tool FastGPU to build AI computing tasks to increase R&D efficiency and GPU utilization, reduce computing times, and lower latency in AI inference. For more information, see What is AIACC?.

- FastGPU

  FastGPU is a set of fast deployment tools provided by Alibaba Cloud for AI computing. You can use the interfaces and automatic tools of FastGPU to build GPU clusters within minutes and set up an efficient training environment for deep learning. For more information, see What is FastGPU?.

- cGPU

  cGPU is a GPU-shared container technology developed by Alibaba Cloud. It provides kernel-based isolation of virtual GPU resources and allows you to deploy multiple containers on a single GPU. This way, you can better utilize your GPU hardware resources at lower costs while securely isolating your business. For more information, see What is the cGPU service?.

- FaaS

  FPGA-accelerated instances are instances that are equipped with FPGAs. FPGA hardware and acceleration applications can be reconfigured to obtain low-latency hardware and elastic resources. For more information, see What is FPGA as a Service?.

4.8.3. 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. • 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. **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. ◦ 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: ■ Create a GPU-accelerated instance that is not configured with a driver ■ 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: ◦ Create a GPU-accelerated instance that is not configured with a driver ◦ 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. ◦ 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. **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. ◦ 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: ■ Create a GPU-accelerated instance that is not configured with a driver ■ Install a GPU driver on a Linux GPU-accelerated compute-optimized instance

4.9. Compute optimized type family with FPGA

4.9.1. Overview

This topic describes the features of field programmable gate array (FPGA)-accelerated compute optimized instance families and lists the instance types of each family.

- Recommended instance families

  f3, FPGA-accelerated compute optimized instance family

- Other available instance families

  f1, FPGA-accelerated compute optimized instance family

f3, FPGA-accelerated compute optimized instance family

Features

- Uses Xilinx 16nm Virtex UltraScale+ VU9P FPGAs.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large computing capacity.

- Suits the following scenarios:
  - Deep learning and inference
  - Genomics research
  - Database acceleration
  - Image transcoding such as conversion of JPEG images to WebP images
  - Real-time video processing such as H.265 video compression

Instance types

Instance type	vCPUs	Memory (GiB)	FPGAs	Bandwidth (bidirectional), Gbit/s	Packet forwarding rate (bidirectional), Kpps	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.f3-c4f1.xlarge	4	16.0	1 × Xilinx VU9P	1.5	300	2	3	10
ecs.f3-c8f1.2xlarge	8	32.0	1 × Xilinx VU9P	2.5	500	4	4	10
ecs.f3-c16f1.4xlarge	16	64.0	1 × Xilinx VU9P	5.0	1,000	4	8	20
ecs.f3-c16f1.8xlarge	32	128.0	2 × Xilinx VU9P	10.0	2,000	8	8	20
ecs.f3-c16f1.16xlarge	64	256.0	4 × Xilinx VU9P	20.0	2,500	16	8	20
ecs.f3-c22f1.22xlarge	88	336.0	4 × Xilinx VU9P	30.0	4,500	16	8	20

> **Note**
>
> -
> - For more information about these specifications, see Instance family.

f1, FPGA-accelerated compute optimized instance family

Features

- Uses Intel® Arria® 10 GX 1150 FPGAs
- Compute:

- Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Offers a CPU-to-memory ratio of 1:7.5.

- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Supports IPv6.
  - Provides high network performance based on large computing capacity.
- Suits the following scenarios:
  - Deep learning and inference
  - Genomics research
  - Financial analysis
  - Image transcoding
  - Computational workloads such as real-time video processing and security management

Instance types

Instance type	vCPUs	Memory (GiB)	FPGAs	Bandwidth (bidirectional), Gbit/s	Packet forwarding rate (bidirectional), Kpps	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.f1-c8f1.2xlarge	8	60.0	Intel ARRIA 10 GX 1150	3.0	400	4	4	10
ecs.f1-c8f1.4xlarge	16	120.0	2 × Intel ARRIA 10 GX 1150	5.0	1,000	4	8	20
ecs.f1-c28f1.7xlarge	28	112.0	Intel ARRIA 10 GX 1150	5.0	2,000	8	8	20
ecs.f1-c28f1.14xlarge	56	224.0	2 × Intel ARRIA 10 GX 1150	10.0	2,000	14	8	20

Note

-
- For more information about these specifications, see Instance family.

4.9.2. Create an f1 instance

This topic describes how to create an f1 instance.

Prerequisites

An image pre-installed with the Intel development environment is obtained. The image is available only as a shared image. To obtain the image, submit a ticket.

Context

This topic describes the parameters for creating an f1 instance. For more information about other common parameters, see Create an instance by using the wizard.

Procedure

1.
2.
3. Click **Create Instance**.
4. In the Basic Configurations step, configure the parameters and click **Next: Networking**.

   When you configure the parameters, take note of the following items:

   - **Region**: Select a region from the following regions where f1 instance types are available. Note that the instance buy page shows the instance types available for your purchase in each zone and region.
     - China (Hangzhou)
     - China (Shenzhen)
     - China (Beijing)

     > **Note** If you can view subscription resources but not pay-as-you-go resources when you purchase an instance, see the "Why are some instance types not available on the instance buy page when I attempt to purchase a pay-as-you-go instance?" section in Instance FAQ.

   - **Instance Type**: Set Architecture to **Heterogeneous Computing** and Category to **Compute Optimized Type with FPGA**.
   - **Image**: Click **Shared Image** and select an image that is pre-installed with the Intel development environment.

     > **Note** Images pre-installed with the Intel development environment can be available only as shared images. These images are pre-installed with quartus17.0, vcs2017.3, and dcp sdk. You can view the files in the *opt* directory.

5. In the Networking step, configure the parameters and click **Next: System Configurations**.

   Only the network type of VPC is supported.
6. In the System Configurations (Optional) step, configure the parameters and click **Next: Grouping**.
7. In the Grouping (Optional) step, configure the parameters and click **Next: Preview**.
8. Confirm the order and click **Create Order**.

What's next

After the f1 instance is created, you can connect to the instance and run the following command to check whether the license is configured. For more information about remote connections, see Connection methodsGuidelines on instance connection.

```
echo $LM_LICENSE_FILE # Depends on whether the $LM_LICENSE_FILE variable is configured.
```

If the variable is configured, the actual value is displayed. Otherwise, no values are displayed.

Related information

References

- Use OpenCL on an f1 instance
- Use RTL Compiler on an f1 instance

4.9.3. Create an f3 instance

This topic describes how to create an f3 instance.

Context

This topic describes the parameters for creating an f3 instance. For more information about other common parameters, see Create an instance by using the wizard.

To facilitate testing, Alibaba Cloud provides an image pre-installed with the Xilinx development environment. The image is available only as a shared image. To obtain the image, submit a ticket.

Procedure

1.
2.
3. Click **Create Instance**.
4. In the Basic Configurations step, configure the parameters and click **Next: Networking**.

   When you configure the parameters, take note of the following items:

   - **Region**: Select a region from the following regions where f3 instance types are available. Note that the instance buy page shows the instance types available for your purchase in each zone and region.
     - China (Shanghai)
     - China (Beijing)
     - China (Zhangjiakou)

     > **Note** If you can view subscription resources but not pay-as-you-go resources when you purchase an instance, see the "Why are some instance types not available on the instance buy page when I attempt to purchase a pay-as-you-go instance?" section in Instance FAQ.

   - **Instance Type**: Set Architecture to **Heterogeneous Computing** and Category to **Compute Optimized Type with FPGA**.
   - **Image**: Click **Shared Image** and select an image that is pre-installed with the Xilinx development environment.

- **Storage**: Images that are pre-installed with the Xilinx development environment occupy some storage space. We recommend that you select an ultra disk of 200 GiB as the system disk.

5. In the Networking step, configure the parameters and click **Next: System Configurations**.

   Only the network type of VPC is supported.
6. In the System Configurations (Optional) step, configure the parameters and click **Next: Grouping**.
7. In the Grouping (Optional) step, configure the parameters and click **Next: Preview**.
8. Confirm the order and click **Create Order**.

Related information

References

- Use OpenCL on an f3 instance
- Use the RTL design on an f3 instance

4.10. ECS Bare Metal Instance types

4.10.1. Overview

This topic describes the features of Elastic Compute Service (ECS) Bare Metal Instance families and lists the instance types of each instance family.

- Recommended instance families
  - General-purpose instance families:
    - ebmg7, general-purpose ECS Bare Metal Instance family
    - ebmg7a, general-purpose ECS Bare Metal Instance family
    - ebmg6a, general-purpose ECS Bare Metal Instance family
    - ebmg6e, general-purpose ECS Bare Metal Instance family with enhanced performance
    - ebmg6, general-purpose ECS Bare Metal Instance family
  - Compute-optimized instance families:
    - ebmc7, compute-optimized ECS Bare Metal Instance family
    - ebmc7a, compute-optimized ECS Bare Metal Instance family
    - ebmc6me, compute-optimized ECS Bare Metal Instance family
    - ebmc6a, compute-optimized ECS Bare Metal Instance family
    - ebmc6e, compute-optimized ECS Bare Metal Instance family with enhanced performance
    - ebmc6, compute-optimized ECS Bare Metal Instance family

- Memory-optimized instance families:
  - ebmr7, memory-optimized ECS Bare Metal Instance family
  - ebmr7a, memory-optimized ECS Bare Metal Instance family
  - ebmr6a, memory-optimized ECS Bare Metal Instance family
  - ebmr6e, memory-optimized ECS Bare Metal Instance family with enhanced performance
  - ebmr6, memory-optimized ECS Bare Metal Instance family
  - ebmre6p, persistent memory-optimized ECS Bare Metal Instance family with enhanced performance
  - ebmre6-6t, memory-optimized ECS Bare Metal Instance family
- Instance families with high clock speeds:
  - ebmhfg7, general-purpose ECS Bare Metal Instance family with high clock speeds
  - ebmhfc7, compute-optimized ECS Bare Metal Instance family with high clock speeds
  - ebmhfr7, memory-optimized ECS Bare Metal Instance family with high clock speeds
  - ebmhfg6, general-purpose ECS Bare Metal Instance family with high clock speeds
  - ebmhfc6, compute-optimized ECS Bare Metal Instance family with high clock speeds
  - ebmhfr6, memory-optimized ECS Bare Metal Instance family with high clock speeds
- GPU-accelerated compute-optimized instance families:
  - ebmgn7e, GPU-accelerated compute-optimized ECS Bare Metal Instance family
  - ebmgn7i, GPU-accelerated compute-optimized ECS Bare Metal Instance family
  - ebmgn7, GPU-accelerated compute-optimized ECS Bare Metal Instance family
  - ebmgn6ia, GPU-accelerated compute-optimized ECS Bare Metal Instance family
  - ebmgn6e, GPU-accelerated compute-optimized ECS Bare Metal Instance family
  - ebmgn6v, GPU-accelerated compute-optimized ECS Bare Metal Instance family
  - ebmgn6i, GPU-accelerated compute-optimized ECS Bare Metal Instance family

- Other available instance families (If these instance families are sold out, you can use the recommended ones.)
  - ebmg5s, network-enhanced general-purpose ECS Bare Metal Instance family
  - ebmg5, general-purpose ECS Bare Metal Instance family
  - ebmc5s, network-enhanced compute-optimized ECS Bare Metal Instance family
  - ebmc4, compute-optimized ECS Bare Metal Instance family
  - ebmr5s, network-enhanced memory-optimized ECS Bare Metal Instance family

Introduction

ECS Bare Metal Instance is an innovative computing service developed by Alibaba Cloud based on state-of-the-art virtualization 2.0 technology. Virtualization 2.0 endows ECS bare metal instances with the elasticity of virtual machines (ECS instances), the performance and features of physical machines, and the full support for nested virtualization.

ECS bare metal instances combine the strengths of both physical machines and ECS instances to deliver powerful and robust computing capabilities. ECS Bare Metal Instance uses virtualization 2.0 to provide your business applications with direct access to the processors and memory resources of the underlying servers without virtualization overheads. ECS Bare Metal Instance retains the hardware feature sets (such as Intel VT-x) and resource isolation capabilities of physical machines, which is ideal for applications that need to run in non-virtualization environments.

ECS Bare Metal Instance integrates features from both physical and virtual machines based on the proprietary chips, hypervisor system software, and a redefined server hardware architecture. ECS Bare Metal Instance can seamlessly connect with other Alibaba Cloud services for storage, networking, and database tasks and is fully compatible with ECS images. These properties allow you to build resources to suit your business requirements.

When you use ECS Bare Metal Instance, take note of the following items:

- ECS bare metal instances do not support instance type changes.
- ECS bare metal instances support failover. When the physical machine that hosts an ECS bare metal instance fails, the instance is failed over. Data is retained in the disks of the instance.

Benefits

ECS Bare Metal Instance provides the following benefits based on technological innovations:

- Exclusive computing resources

  ECS Bare Metal Instance is a cloud-based elastic computing service that provides the same performance and resource isolation capabilities as physical machines. It can ensure the exclusivity of computing resources without virtualization overheads or performance loss. ECS Bare Metal Instance supports high clock speeds and configurations of 8, 32, 80, 96, and 104 vCPUs. An ECS bare metal instance that has eight vCPUs can have a maximum clock speed of 3.7 GHz to 4.1 GHz and provide better performance and faster response for gaming and finance scenarios than competing products.

- Chip-level security

  In addition to physical isolation, ECS Bare Metal Instance uses a chip-level trusted execution environment of Intel® Software Guard Extensions (SGX) to ensure that encrypted data can be computed only in a secure and trusted environment. This chip-level hardware security protection provides security for your data in the cloud and allows you to control all data encryption and key protection processes. For more information, see Install SGX.

- Compatibility with multiple private clouds

  ECS Bare Metal Instance can address high-performance computing needs and help you construct hybrid clouds. Thanks to the flexibility, elasticity, and other strengths inherited from the mix of physical and virtual machines, ECS Bare Metal Instance can implement re-virtualization and allow local private clouds to be seamlessly migrated to Alibaba Cloud without performance overheads arising from nested virtualization. This provides you a new method to move business to the cloud.

- Support for heterogeneous instruction set processors

  ECS Bare Metal Instance uses the virtualization 2.0 technology developed by Alibaba Cloud and supports instruction set processors such as Advanced RISC Machine (ARM) at no additional costs.

Comparison of ECS bare metal instances, physical machines, and virtual machines

An ECS bare metal instance delivers better performance than a physical machine that has the same configurations. During the yearly Double 11 sales event, ECS bare metal instances deliver robust computing capabilities with millions of vCPUs to handle spikes in traffic.

The following table compares the features of ECS bare metal instances, physical machines, and virtual machines. In this table, Y means supported, N means not supported, and N/A means not applicable.

Feature type	Feature	ECS bare metal instance	Physical machine	Virtual machine
Automated O&M	Delivery within minutes	Y	N	Y
Compute	Zero performance loss	Y	Y	N
	Zero feature loss	Y	Y	N
	Zero resource contention	Y	Y	N
Storage	Compatibility with ECS disks	Y	N	Y
	Startup from system disks	Y	N	Y
	Quick reset of system disks	Y	N	Y
	Use of ECS images	Y	N	Y
	Cold migration between physical and virtual machines	Y	N	Y
	No need to install the operating system	Y	N	Y
	No need for local redundant arrays of independent disks (RAIDs), and better protection of data in disks	Y	N	Y
Networking	Compatibility with virtual private clouds (VPCs)	Y	N	Y
	Compatibility with the classic network	Y	N	Y
	No communication bottlenecks between physical and virtual machine clusters in VPCs	Y	N	Y
Management	Compatibility with existing ECS management systems	Y	N	Y
	Consistent user experience on features such as Virtual Network Console (VNC) with that on virtual machines	Y	N	Y

Feature type	Feature	ECS bare metal instance	Physical machine	Virtual machine
	Out-of-band (OOB) network security	Y	N	N/A

ebmg7, general-purpose ECS Bare Metal Instance family

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.9 GHz Intel® Xeon® Platinum 8369B (Ice Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only enhanced SSDs (ESSDs) and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Enterprise-level applications of various types and sizes
  - Websites and application servers
  - Game servers
  - Small and medium-sized database systems, caches, and search clusters
  - Data analytics and computing
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmg7.32xlarge	128	512	64	24,000,000	2,400,000	32	15	600,000	32

Note

-
- For more information about these specifications, see Instance family.

ebmg7a, general-purpose ECS Bare Metal Instance family

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.55 GHz AMD EPYC™ MILAN processors that deliver a maximum single-core turbo frequency of 3.5 GHz to provide consistent computing performance.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Computing clusters and memory-intensive data processing
  - Video encoding, decoding, and rendering
  - Data analytics and computing

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmg7a.64xlarge	256	1024	64	24,000,000	4,000,000	32	31	15	600,000	32

Note

-
- For more information about these specifications, see Instance family.
- The boot mode of the images that are used by instances of this instance family must be Unified Extensible Firmware Interface (UEFI). If you want to use a custom image, make sure that the boot mode of the image is set to UEFI. For information about how to set the boot mode of a custom image, see Set the boot mode of custom images to the UEFI mode by calling API operations.
- Ubuntu 18 and Debian 9 operating system kernels do not support AMD EPYC ™ MILAN processors. Do not use Ubuntu 18 or Debian 9 images to create instances of this instance family. Otherwise, the instances cannot be started.

ebmg6a, general-purpose ECS Bare Metal Instance family

The instance family is in invitational preview. To use this instance family, submit a ticket.

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.6 GHz AMD EPYC™ ROME processors that deliver a turbo frequency of 3.3 GHz to provide consistent computing performance.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments

- Containers such as Docker, Clear Containers, and Pouch
- Video encoding, decoding, and rendering
- Computing clusters and memory-intensive data processing
- Data analytics and computing

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmg6a.64xlarge	256	1024	64	24,000,000	32	31	10	600,000	32

Note

-
- For more information about these specifications, see Instance family.
- The boot mode of the images that are used by instances of this instance family must be UEFI. If you want to use a custom image, make sure that the boot mode of the image is set to UEFI. For information about how to set the boot mode of a custom image, see Set the boot mode of custom images to the UEFI mode by calling API operations.

ebmg6e, general-purpose ECS Bare Metal Instance family with enhanced performance

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver an all-core turbo frequency of 3.2 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware

- Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
- Containers such as Docker, Clear Containers, and Pouch
- Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
- Enterprise-level applications of various types and sizes
- Websites and application servers
- Game servers
- Small and medium-sized database systems, caches, and search clusters
- Data analytics and computing
- Computing clusters and memory-intensive data processing
- High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmg6e.26xlarge	104	384	32	24,000,000	1,800,000	32	10	480,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmg6, general-purpose ECS Bare Metal Instance family

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver an all-core turbo frequency of 3.2 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 6,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware

- Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
- Containers such as Docker, Clear Containers, and Pouch
- Video encoding, decoding, and rendering
- Enterprise-level applications such as large and medium-sized databases
- Computing clusters and memory-intensive data processing
- Data analytics and computing

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmg6.26xlarge	104	384	32	6,000,000	1,800,000	32	10	200,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmc7, compute-optimized ECS Bare Metal Instance family

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:2.
- Uses 2.9 GHz Intel® Xeon® Platinum 8369B (Ice Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch

- Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
- Web frontend servers
- Frontend servers of massive multiplayer online (MMO) games
- Data analytics, batch processing, and video encoding
- High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmc7.32xlarge	128	256	64	24,000,000	2,400,000	32	15	600,000	32

> **Note**
>
> -
> - For more information about these specifications, see Instance family.

ebmc7a, compute-optimized ECS Bare Metal Instance family

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:2.
- Uses 2.55 GHz AMD EPYC™ MILAN processors that deliver a maximum single-core turbo frequency of 3.5 GHz to provide consistent computing performance.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Video encoding, decoding, and rendering

- Data analytics and computing

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmc7a.64xlarge	256	512	64	24,000,000	4,000,000	32	31	15	600,000	32

Note

-
- For more information about these specifications, see Instance family.
- The boot mode of the images that are used by instances of this instance family must be UEFI. If you want to use a custom image, make sure that the boot mode of the image is set to UEFI. For information about how to set the boot mode of a custom image, see Set the boot mode of custom images to the UEFI mode by calling API operations.
- Ubuntu 18 and Debian 9 operating system kernels do not support AMD EPYC™ MILAN processors. Do not use Ubuntu 18 or Debian 9 images to create instances of this instance family. Otherwise, the instances cannot be started.

ebmc6me, compute-optimized ECS Bare Metal Instance family

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:3.
- Uses 2.3 GHz Intel® Xeon® Gold 5218 (Cascade Lake) processors that deliver a turbo frequency of 3.9 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 6,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Video encoding, decoding, and rendering

- Frontend servers of MMO games
- High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmc6me.16xlarge	64	192	32	6,000,000	1,800,000	32	10	200,000	16

> **Note**
>
> -
> - For more information about these specifications, see Instance family.

ebmc6a, compute-optimized ECS Bare Metal Instance family

The instance family is in invitational preview. To use this instance family, submit a ticket.

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:2.
- Uses 2.6 GHz AMD EPYC™ ROME processors that deliver a turbo frequency of 3.3 GHz to provide consistent computing performance.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Video encoding, decoding, and rendering
  - Data analytics and computing

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmc6a.64xlarge	256	512	64	24,000,000	32	31	10	600,000	32

Note

-
- For more information about these specifications, see Instance family.
- The boot mode of the images that are used by instances of this instance family must be UEFI. If you want to use a custom image, make sure that the boot mode of the image is set to UEFI. For information about how to set the boot mode of a custom image, see Set the boot mode of custom images to the UEFI mode by calling API operations.

ebmc6e, compute-optimized ECS Bare Metal Instance family with enhanced performance

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:2.
- Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver an all-core turbo frequency of 3.2 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Web frontend servers
  - Frontend servers of MMO games

- Data analytics, batch processing, and video encoding
- High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmc6e.26xlarge	104	192	32	24,000,000	1,800,000	32	10	480,000	16

> **Note**
>
> -
> - For more information about these specifications, see Instance family.

ebmc6, compute-optimized ECS Bare Metal Instance family

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:2.
- Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver an all-core turbo frequency of 3.2 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 6,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Video encoding, decoding, and rendering
  - Frontend servers of MMO games
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmc6.26xlarge	104	192	32	6,000,000	1,800,000	32	10	200,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmr7, memory-optimized ECS Bare Metal Instance family

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:8.
- Uses 2.9 GHz Intel® Xeon® Platinum 8369B (Ice Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - High-performance databases and in-memory databases
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmr7.32xlarge	128	1024	64	24,000,000	2,400,000	32	15	600,000	32

Note

-
- For more information about these specifications, see Instance family.

ebmr7a, memory-optimized ECS Bare Metal Instance family

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:8.
- Uses 2.55 GHz AMD EPYC™ MILAN processors that deliver a maximum single-core turbo frequency of 3.5 GHz to provide consistent computing performance.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - In-memory databases
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmr7a.64xlarge	256	2048	64	24,000,000	4,000,000	32	31	15	600,000	32

Note

-
- For more information about these specifications, see Instance family.
- The boot mode of the images that are used by instances of this instance family must be UEFI. If you want to use a custom image, make sure that the boot mode of the image is set to UEFI. For information about how to set the boot mode of a custom image, see Set the boot mode of custom images to the UEFI mode by calling API operations.
- Ubuntu 18 and Debian 9 operating system kernels do not support AMD EPYC ™ MILAN processors. Do not use Ubuntu 18 or Debian 9 images to create instances of this instance family. Otherwise, the instances cannot be started.

ebmr6a, memory-optimized ECS Bare Metal Instance family

The instance family is in invitational preview. To use this instance family, submit a ticket.

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:8.
- Uses 2.6 GHz AMD EPYC™ ROME processors that deliver a turbo frequency of 3.3 GHz to provide consistent computing performance.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch

- In-memory databases
- Data analytics, data mining, and distributed memory caching
- Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmr6a.64xlarge	256	2048	64	24,000,000	32	31	10	600,000	32

Note

-
- For more information about these specifications, see Instance family.
- The boot mode of the images that are used by instances of this instance family must be UEFI. If you want to use a custom image, make sure that the boot mode of the image is set to UEFI. For information about how to set the boot mode of a custom image, see Set the boot mode of custom images to the UEFI mode by calling API operations.

ebmr6e, memory-optimized ECS Bare Metal Instance family with enhanced performance

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:8.
- Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver an all-core turbo frequency of 3.2 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments

- Containers such as Docker, Clear Containers, and Pouch
- Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
- High-performance databases and in-memory databases
- Data analytics, data mining, and distributed memory caching
- Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters
- High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmr6e.26xlarge	104	768	32	24,000,000	1,800,000	32	10	480,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmr6, memory-optimized ECS Bare Metal Instance family

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:8.
- Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver an all-core turbo frequency of 3.2 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 6,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - High-performance databases and in-memory databases
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmr6.26xlarge	104	768	32	6,000,000	1,800,000	32	10	200,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmre6p, persistent memory-optimized ECS Bare Metal Instance family with enhanced performance

To use ebmre6p,submit a ticket.

Features:

- Provides dedicated hardware resources and physical isolation.
- Uses the Intel® Optane™ persistent memory and is tuned for Redis applications in an end-to-end manner to provide cost-effectiveness.
- Supports a maximum total memory of 1,920 GiB (384 GiB DRAM + 1,536 GiB Intel® Optane™ persistent memory), offers a CPU-to-memory ratio of 1:20, and can meet the needs of memory-intensive applications.
- Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver an all-core turbo frequency of 3.2 GHz and provide consistent computing performance.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 6,000,000 pps.
- Suits the following scenarios:
  - In-memory databases such as Redis
  - High-performance databases such as SAP HANA
  - Other memory-intensive applications such as AI applications and smart search applications

Instance types

Instance type	vCPUs	Memory (GiB)	Persistent memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmre6p.26xlarge	104	384	1536	32	6,000,000	32	10	200,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmre6-6t, memory-optimized ECS Bare Metal Instance family

To use ebmre6-6t,submit a ticket.

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:30.
- Uses 2.5 GHz Intel® Xeon® Platinum 8269 (Cascade Lake) processors that deliver an all-core turbo frequency of 3.2 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 6,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - In-memory databases and high-performance databases such as SAP HANA
  - Memory-intensive applications
  - Big data processing engines such as Apache Spark and Presto

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmre6-6t.52xlarge	208	6144	32	6,000,000	1,800,000	32	10	200,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmhfg7, general-purpose ECS Bare Metal Instance family with high clock speeds

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:4.
- Uses third-generation Intel® Xeon® Scalable (Cooper Lake) processors that deliver a base frequency of not lower than 3.3 GHz and an all-core turbo frequency of 3.8 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Enterprise-level applications of various types and sizes
  - Game servers
  - Small and medium-sized database systems, caches, and search clusters
  - High-performance scientific computing
  - Video encoding applications

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmhfg7.48xlarge	192	768	64	24,000,000	32	31	10	600,000	32

Note

-
- For more information about these specifications, see Instance family.

ebmhfc7, compute-optimized ECS Bare Metal Instance family with high clock speeds

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:2.
- Uses third-generation Intel® Xeon® Scalable (Cooper Lake) processors that deliver a base frequency of not lower than 3.3 GHz and an all-core turbo frequency of 3.8 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance frontend server clusters
  - Frontend servers of MMO games
  - Data analytics, batch processing, and video encoding
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmhfc7.48xlarge	192	384	64	24,000,000	32	31	10	600,000	32

Note

-
- For more information about these specifications, see Instance family.

ebmhfr7, memory-optimized ECS Bare Metal Instance family with high clock speeds

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:8.
- Uses third-generation Intel® Xeon® Scalable (Cooper Lake) processors that deliver a base frequency of not lower than 3.3 GHz and an all-core turbo frequency of 3.8 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs and provides ultra-high I/O performance.
- Supports IPv6.
- Supports only VPCs.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - High-performance databases and in-memory databases
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmhfr7.48xlarge	192	1536	64	24,000,000	32	31	10	600,000	32

Note

-
- For more information about these specifications, see Instance family.

ebmhfg6, general-purpose ECS Bare Metal Instance family with high clock speeds

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:4.8.
- Uses 3.1 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 6,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Enterprise-level applications such as large and medium-sized databases
  - Video encoding, decoding, and rendering

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmhfg6.20xlarge	80	384	32	6,000,000	1,800,000	32	10	200,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmhfc6, compute-optimized ECS Bare Metal Instance family with high clock speeds

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:2.4.
- Uses 3.1 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 6,000,000 pps.
- Suits the following scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Video encoding, decoding, and rendering

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmhfc6.20xlarge	80	192	32	6,000,000	1,800,000	32	10	200,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmhfr6, memory-optimized ECS Bare Metal Instance family with high clock speeds

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:9.6.
- Uses 3.1 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 6,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - High-performance databases and in-memory databases
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmhfr6.20xlarge	80	768	32	6,000,000	1,800,000	32	10	200,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmgn7e, GPU-accelerated compute-optimized ECS Bare Metal Instance family

Features:

- Provides flexible and powerful software-defined compute based on the SHENLONG architecture.
- Uses NVIDIA A100 SXM4 80GB GPUs that support NVSwitch and deliver up to 312 TFLOPS of TensorFloat-32 (TF32) computing power.
- Uses 2.9 GHz Intel® Xeon® Scalable processors that deliver an all-core turbo frequency of 3.5 GHz and support PCIe 4.0 interfaces.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs. ESSDs at performance level (PL) 3 can deliver a maximum of 1,000,000 IOPS and 4,000 MB/s of throughput, which can meet the cache requirements of training and eliminate the need for local disks.

  Note For more information about the performance of ESSDs, see ESSDs.

- Supports IPv6.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Deep learning training and development
  - High-performance computing (HPC) and simulations

Instance types

Instance type	vCPUs	Memory (GiB)	GPU	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues (Primary ENI/Secondary ENI)	ENIs
ecs.ebmgn7e.32xlarge	128	1024	NVIDIA A100 * 8	80GB * 8	64	24,000,000	32/12	32

Note

-
- For more information about these specifications, see Instance family.

ebmgn7i, GPU-accelerated compute-optimized ECS Bare Metal Instance family

Features:

- Provides flexible and powerful software-defined compute based on the SHENLONG architecture.
- Uses NVIDIA A10 GPUs that have the following features:
  - Innovative Ampere architecture
  - Support for acceleration features such as vGPU, RTX technology, and TensorRT inference engine
- Uses 2.9 GHz Intel® Xeon® Scalable (Ice Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Provides ultra-high network performance with a packet forwarding rate of 24,000,000 pps.
- Suits the following scenarios:
  - Concurrent AI inference tasks that require high-performance CPUs, memory, and GPUs, such as image recognition, speech recognition, and behavior identification
  - Compute-intensive graphics processing tasks that require high-performance 3D graphics virtualization capabilities, such as remote graphic design and cloud gaming
  - Scenarios that require high network bandwidth and disk bandwidth, such as the creation of high-performance render farms
  - Small-scale deep learning and training applications that require high network bandwidth

Instance types

Instance type	vCPUs	Memory (GiB)	GPU	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs
ecs.ebmgn7i.32xlarge	128	768	NVIDIA A10 * 4	24GB * 4	64	24,000,000	32	32

Note

-
- For more information about these specifications, see Instance family.

ebmgn7, GPU-accelerated compute-optimized ECS Bare Metal Instance family

Features:

- Provides flexible and powerful software-defined compute based on the SHENLONG architecture.
- Uses NVIDIA A100 GPUs. NVSwitches are used to establish connections between NVIDIA A100 GPUs. The GPUs have the following features:
  - Innovative Ampere architecture
  - 40 GB HBM2 memory per GPU
- Uses 2.5 GHz Intel® Xeon® Platinum 8269CY (Cascade Lake) processors.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Provides high network performance based on large computing capacity.
- Suits the following scenarios:
  - Deep learning applications such as training applications of AI algorithms used in image classification, autonomous vehicles, and speech recognition
  - Scientific computing applications that require robust GPU computing capabilities such as computational fluid dynamics, computational finance, molecular dynamics, and environmental analysis

Instance types

Instance type	vCPUs	Memory (GiB)	GPU	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.ebmgn7.26xlarge	104	768	NVIDIA A100 * 8	30	18,000,000	16	15	10

Note

-
- For more information about these specifications, see Instance family.

ebmgn6ia, GPU-accelerated compute-optimized ECS Bare Metal Instance family

This instance family is in invitational preview. To use this instance family, .

Features:

- Uses the fast path acceleration feature of chips to provide predictable and consistent ultra-high computing, storage, and network performance based on the third-generation SHENLONG architecture.
- Uses NVIDIA T4 GPUs to offer GPU acceleration capabilities for graphics and AI applications and adopts container technology to start up to 60 virtual Android devices and provide hardware-accelerated video transcoding.
- Offers a CPU-to-memory ratio of 1:3.
- Uses 2.8 GHz Ampere® Altra® processors that deliver a turbo frequency of 3.0 GHz and provides high performance and high compatibility with applications for Android servers.
- Is an instance family in which all instances are I/O optimized.
- Supports only ESSDs.
- Supports IPv6.
- Suits the following scenarios: remote application services based on Android, such as always-on cloud-based services, cloud-based mobile games, cloud-based mobile phones, and Android service crawlers.

Instance types

Instance type	vCPUs	Memory (GiB)	GPU	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.ebmgn6ia.20xlarge	80	256	NVIDIA T4 * 2	16GB * 2	32	24,000,000	32	15	10

Note

-
- For more information about these specifications, see Instance family.
- Ampere® Altra® processors have specific requirements on the kernels of operating systems. Instances of this instance type can use Alibaba Cloud Linux 3 images or CentOS 8.4 or later images. We recommend that you use Alibaba Cloud Linux 3 images on the instances. If you want to use another operating system distribution, patch the kernel of an instance that runs an operating system of that distribution, create a custom image from the instance, and then use the custom image to create instances of this instance type. For information about kernel patches, visit Ampere Altra (TM) Linux Kernel Porting Guide.

ebmgn6e, GPU-accelerated compute-optimized ECS Bare Metal Instance family

Features:

- Provides flexible and powerful software-defined compute based on the SHENLONG architecture.
- Uses NVIDIA V100 (32 GB NVLink) GPUs.
- Uses NVIDIA V100 GPUs (SXM2-based) that have the following features:
  - Innovative Volta architecture
  - 32 GB HBM2 memory (900 GB/s bandwidth) per GPU
  - 5,120 CUDA cores per GPU
  - Up to 640 Tensor cores per GPU
  - Support for up to six NVLink connections, which each have a bidirectional bandwidth of 50 GB/s for a total bandwidth of 300 GB/s
- Offers a CPU-to-memory ratio of 1:8.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Provides high network performance based on large computing capacity.
- Suits the following scenarios:
  - Deep learning applications, such as training and inference applications of AI algorithms used in image classification, autonomous vehicles, and speech recognition
  - Scientific computing applications, such as computational fluid dynamics, computational finance, molecular dynamics, and environmental analysis

Instance types

Instance type	vCPUs	Memory (GiB)	GPU	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.ebmgn6e.24xlarge	96	768	NVIDIA V100 * 8	32GB * 8	32	4,800,000	16	15	20

Note

-
- For more information about these specifications, see Instance family.

ebmgn6v, GPU-accelerated compute-optimized ECS Bare Metal Instance family

Features:

- Provides flexible and powerful software-defined compute based on the SHENLONG architecture.
- Uses NVIDIA V100 GPUs.
- Uses NVIDIA V100 GPUs (SXM2-based) that have the following features:
  - Innovative Volta architecture
  - 16 GB HBM2 memory (900 GB/s bandwidth) per GPU
  - 5,120 CUDA cores per GPU
  - Up to 640 Tensor cores per GPU
  - Support for up to six NVLink connections, which each have a bidirectional bandwidth of 50 GB/s for a total bandwidth of 300 GB/s
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports IPv6.
- Provides high network performance based on large computing capacity.
- Suits the following scenarios:
  - Deep learning applications, such as training and inference applications of AI algorithms used in image classification, autonomous vehicles, and speech recognition
  - Scientific computing applications, such as computational fluid dynamics, computational finance, molecular dynamics, and environmental analysis

Instance types

Instance type	vCPUs	Memory (GiB)	GPU	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.ebmgn6v.24xlarge	96	384	NVIDIA V100 * 8	16GB * 8	30	4,500,000	8	32	10

Note

-
- For more information about these specifications, see Instance family.

ebmgn6i, GPU-accelerated compute-optimized ECS Bare Metal Instance family

Features:

- Provides flexible and powerful software-defined compute based on the SHENLONG architecture.
- Uses NVIDIA T4 GPUs that have the following features:
  - Innovative NVIDIA Turing architecture
  - 16 GB memory (320 GB/s bandwidth) per GPU
  - 2,560 CUDA cores per GPU
  - Up to 320 Turing Tensor cores per GPU
  - Mixed-precision Tensor cores that support 65 FP16 TFLOPS, 130 INT8 TOPS, and 260 INT4 TOPS
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Is an instance family in which all instances are I/O optimized.
- Supports standard SSDs, ultra disks, and ESSDs that deliver millions of IOPS.
- Supports IPv6.
- Provides high network performance based on large computing capacity.
- Suits the following scenarios:
  - AI (deep learning and machine learning) inference for computer vision, voice recognition, speech synthesis, natural language processing (NLP), machine translation, and reference systems
  - Real-time rendering for cloud games
  - Real-time rendering for AR and VR applications
  - Graphics workstations or overloaded graphics computing
  - GPU-accelerated databases
  - High-performance computing

Instance types

Instance type	vCPUs	Memory (GiB)	GPU	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.ebmgn6i.24xlarge	96	384	NVIDIA T4 * 4	16GB * 4	30	4,500,000	8	32	10

Note

-
- For more information about these specifications, see Instance family.

ebmg5s, network-enhanced general-purpose ECS Bare Metal Instance family

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors that deliver an all-core turbo frequency of 2.7 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 4,500,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Enterprise-level applications such as large and medium-sized databases
  - Video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmg5s.24xlarge	96	384	32	4,500,000	1,800,000	32	10	200,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmg5, general-purpose ECS Bare Metal Instance family

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors that deliver an all-core turbo frequency of 2.7 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only standard SSDs and ultra disks.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 4,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Enterprise-level applications such as large and medium-sized databases
  - Video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	ENIs	Private IP addresses per ENI
ecs.ebmg5.24xlarge	96	384	10	4,000,000	32	10

Note

-
- For more information about these specifications, see Instance family.

ebmc5s, network-enhanced compute-optimized ECS Bare Metal Instance family

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:2.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors that deliver an all-core turbo frequency of 2.7 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 4,500,000 pps.
- Suits the following scenarios:
  - Scenarios where large volumes of packets are received and transmitted, such as on-screen video comments and telecom data forwarding
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Video encoding, decoding, and rendering

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmc5s.24xlarge	96	192	32	4,500,000	1,800,000	32	10	200,000	16

Note

-
- For more information about these specifications, see Instance family.

ebmc4, compute-optimized ECS Bare Metal Instance family

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:2.
- Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors that deliver a turbo frequency of 3.0 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only standard SSDs and ultra disks.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 4,000,000 pps.
- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - Enterprise-level applications such as large and medium-sized databases
  - Video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	ENIs	Private IP addresses per ENI
ecs.ebmc4.8xlarge	32	64	10	4,000,000	12	10

Note

-
- For more information about these specifications, see Instance family.

ebmr5s, network-enhanced memory-optimized ECS Bare Metal Instance family

Features:

- Provides dedicated hardware resources and physical isolation.
- Offers a CPU-to-memory ratio of 1:8.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors that deliver an all-core turbo frequency of 2.7 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Supports only VPCs.
- Provides high network performance with a packet forwarding rate of 4,500,000 pps.

- Suits the following scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Scenarios that require compatibility with third-party hypervisors to implement hybrid-cloud and multi-cloud deployments
  - Containers such as Docker, Clear Containers, and Pouch
  - High-performance databases and in-memory databases
  - Data analytics, data mining, and distributed memory caching
  - Enterprise-level memory-intensive applications such as Hadoop clusters and Spark clusters

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	Connections	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.ebmr5s.24xlarge	96	768	32	4,500,000	1,800,000	32	10	200,000	16

Note

-
- For more information about these specifications, see Instance family.

Billing

ECS Bare Metal Instance supports pay-as-you-go and subscription billing methods. For more information, see Overview.

4.10.2. Create an ECS bare metal instance

The procedure to create an ECS bare metal instance is similar to the procedure to create a regular ECS instance.

Context

For information about how to create a regular ECS instance, see Create an instance by using the wizard.

Procedure

1.
2.
3. Click **Create Instance**.
4. In the Basic Configurations step, configure the parameters and click **Next: Networking**.

   Take note of the following parameters:

   - **Region**: ECS Bare Metal Instance families and types are available in specific regions and zones.

For the regions and zones where ECS Bare Metal Instance families and types are available, see the ECS Instance Types Available for Each Region page. Select a billing method and enter an instance type name to search for the instance type.

- **Instance Type**: Set Architecture to **ECS Bare Metal Instance**, set Category to **CPU Type** or **GPU Type**, and then select an instance type. For more information, see Instance family.
- **Image**: All available images are displayed in the Image section of the buy page.

5. In the Networking step, configure parameters for the network and security group, and click **Next: System Configurations**.

   Network Type: Only **VPC** is available.
6. In the System Configurations (Optional) step, configure the parameters and click **Next: Grouping**.
7. In the Grouping (Optional) step, configure the parameters and click **Next: Preview**.
8. Check your configurations, read and select *ECS Terms of Service*, and then click Create Order.

Related information

- RunInstances

4.11. Super Computing Cluster instance type family

4.11.1. Overview

This topic describes the features of Super Computing Cluster (SCC) instance families of Elastic Compute Service (ECS) and lists the instance types of each instance family.

- scchfc6, compute-optimized SCC instance family with high clock speeds
- scchfg6, general-purpose SCC instance family with high clock speeds
- scchfr6, memory-optimized SCC instance family with high clock speeds
- scch5, SCC instance family with high clock speeds
- sccg5, general-purpose SCC instance family
- sccgn7ex, GPU-accelerated compute-optimized SCC instance family
- sccgn6e, GPU-accelerated compute-optimized SCC instance family
- sccgn6, GPU-accelerated compute-optimized SCC instance family

Introduction

SCC is based on Elastic Compute Service (ECS) Bare Metal Instance and significantly improves the network performance and acceleration ratio of large-scale clusters by using high-speed Remote Direct Memory Access (RDMA) based interconnects. SCC has all the benefits of ECS Bare Metal Instance and can provide high-quality network performance with high bandwidth and low latency.

SCC is used in scenarios such as high-performance computing, artificial intelligence, machine learning, scientific computing, engineering computing, data analytics, and audio and video processing. In SCCs, nodes are connected over high-bandwidth and low-latency RDMA networks. This ensures the parallel efficiency of applications in areas such as high-performance computing, artificial intelligence, and machine learning. RDMA over Converged Ethernet (RoCE) networks rival InfiniBand networks in terms of connection speed and can support more Ethernet-based applications.

SCC can be used in conjunction with other Alibaba Cloud computing services such as ECS and Elastic GPU Service to provide ultimate high-performance parallel computing resources for Elastic High Performance Computing (E-HPC) and make cloud-based super computing possible.

Comparison of SCCs, physical machines, and virtual machines

The following table compares the features of SCCs, physical machines, and virtual machines. In this table, Y means supported, N means not supported, and N/A means not applicable.

Feature type	Feature	SCC	Physical machine	Virtual machine
Automated O&M	Delivery within minutes	Y	N	Y
Compute	Zero performance loss	Y	Y	N
	Zero feature loss	Y	Y	N
	Zero resource contention	Y	Y	N
Storage	Compatibility with ECS disks	Y	N	Y
	Startup from system disks	Y	N	Y
	Quick reset of system disks	Y	N	Y
	Use of ECS images	Y	N	Y
	Cold migration between physical and virtual machines	Y	N	Y
	No need to install the operating system	Y	N	Y
	No need for local RAIDs, and better protection of data in disks	Y	N	Y
	Compatibility with virtual private clouds (VPCs)	Y	N	Y
	Compatibility with the classic network	Y	N	Y

Network Feature type	Feature	SCC	Physical machine	Virtual machine
	No communication bottlenecks between physical and virtual machine clusters located in VPCs	Y	N	Y
Management	Compatibility with existing ECS management systems	Y	N	Y
	Consistent user experience on features such as Virtual Network Console (VNC) with that on virtual machines	Y	N	Y
	Out-of-band (OOB) network security	Y	N	N/A

scchfc6, compute-optimized SCC instance family with high clock speeds

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2.4.
  - Uses 3.1 GHz Intel® Xeon® Platinum 8269 (Cascade Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports enhanced SSDs (ESSDs), standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Supports both RoCE networks and VPCs. RoCE networks are dedicated to RDMA communication.
- Supported scenarios:
  - Large-scale machine learning training
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Physical cores	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	Elastic network interfaces (ENIs)	Private IP addresses per ENI
ecs.scchfc6.20xlarge	80	40	192.0	30	6,000,000	50	8	32	10

Note

- ecs.scchfc6.20xlarge provides 80 logical processors on 40 physical cores.
-
- For more information about these specifications, see Instance family.

scchfg6, general-purpose SCC instance family with high clock speeds

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.8.
  - Uses 3.1 GHz Intel® Xeon® Platinum 8269 (Cascade Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Supports both RoCE networks and VPCs. RoCE networks are dedicated to RDMA communication.
- Supported scenarios:
  - Large-scale machine learning training
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Physical cores	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.scchfg6.20xlarge	80	40	384.0	30	6,000,000	50	8	32	10

> **Note**
> - ecs.scchfg6.20xlarge provides 80 logical processors on 40 physical cores.
> -
> - For more information about these specifications, see Instance family.

scchfr6, memory-optimized SCC instance family with high clock speeds

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Offers a CPU-to-memory ratio of 1:9.6.
  - Uses 3.1 GHz Intel® Xeon® Platinum 8269 (Cascade Lake) processors that deliver an all-core turbo frequency of 3.5 GHz.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
- Network:
  - Supports IPv6.
  - Supports both RoCE networks and VPCs. RoCE networks are dedicated to RDMA communication.
- Supported scenarios:
  - Large-scale machine learning training
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Physical cores	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.scchfr6.20xlarge	80	40	768.0	30	6,000,000	50	8	32	10

Note

- ecs.scchfr6.20xlarge provides 80 logical processors on 40 physical cores.
-
- For more information about these specifications, see Instance family.

scch5, SCC instance family with high clock speeds

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Offers a CPU-to-memory ratio of 1:3.
  - Uses 3.1 GHz Intel® Xeon® Gold 6149 (Skylake) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports standard SSDs and ultra disks.
- Network:
  - Supports both RoCE networks and VPCs. RoCE networks are dedicated to RDMA communication.
- Supported scenarios:
  - Large-scale machine learning training
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Physical cores	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.scch5.16xlarge	64	32	192.0	10	4,500,000	50	8	32	10

Note
- ecs.scch5.16xlarge provides 64 logical processors on 32 physical cores.
-
- For more information about these specifications, see Instance family.

sccg5, general-purpose SCC instance family

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports standard SSDs and ultra disks.
- Network:
  - Supports both RoCE networks and VPCs. RoCE networks are dedicated to RDMA communication.
- Supported scenarios:
  - Large-scale machine learning training
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Physical cores	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.sccg5.24xlarge	96	48	384.0	10	4,500,000	50	8	32	10

Note
- ecs.sccg5.24xlarge provides 96 logical processors on 48 physical cores.
-
- For more information about these specifications, see Instance family.

sccgn7ex, GPU-accelerated compute-optimized SCC instance family

Features

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Uses eight NVIDIA A100 GPUs per instance that support NVSwitch and deliver up to 312 TFLOPS of TensorFloat-32 (TF32) computing power.
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses third-generation Intel® Xeon® Scalable (Ice Lake) processors that deliver a base frequency of 2.9 GHz and an all-core turbo frequency of 3.5 GHz and support PCIe 4.0 interfaces.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only ESSDs. ESSDs at performance level (PL) 3 can deliver a maximum of 1,000,000 IOPS and 4,000 MB/s of throughput, which can meet the cache requirements of training and eliminate the need for local disks.
- Network:
  - Supports IPv6.
  - Supports only VPCs.
  - Provides a bandwidth of 800 Gbit/s between sccgn7ex instances (100 Gbit/s per port on each of four dual-port RDMA network controller) to support GPUDirect. Each GPU is directly connected to a port with the bandwidth of 100 Gbit/s on RDMA network interface controllers.
- Supported scenarios: ultra-large-scale training for artificial intelligence.

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues (Primary ENI/Secondary ENI)	ENIs	Private IP addresses per ENI
ecs.sccgn7ex.32xlarge	128	1,024	NVIDIA A100 × 8	80GB × 8	64	24,000,000	800	32/12	32	15

Note

-
- For more information about these specifications, see Instance family.

sccgn6e, GPU-accelerated compute-optimized SCC instance family

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:

  - Uses NVIDIA V100 GPUs (SXM2-based) that have the following features:
    - Innovative Volta architecture
    - 32 GB HBM2 GPU memory
    - 5,120 CUDA cores
    - 640 Tensor cores
    - GPU memory bandwidth of up to 900 GB/s
    - Support for six NVLink links and a total bandwidth of 300 GB/s (25 GB/s per NVlink link per direction)
  - Offers a CPU-to-memory ratio of 1:8.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Supports high-performance Cloud Paralleled File System (CPFS).
- Network:
  - Supports IPv6.
  - Supports VPCs.
  - Supports RoCE v2 networks, which are dedicated to low-latency RDMA communication.
- Supported scenarios:
  - Ultra-large-scale training for machine learning on a distributed GPU cluster
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.scgn6e.24xlarge	96	768.0	NVIDIA V100 × 8	32GB × 8	32	4,800,000	50	8	32	10

Note

-
- For more information about these specifications, see Instance family.

sccgn6, GPU-accelerated compute-optimized SCC instance family

Features:

- Provides all features of ECS Bare Metal Instance. For more information, see Overview.
- Compute:
  - Uses NVIDIA V100 GPUs (SXM2-based) that have the following features:
    - Innovative Volta architecture
    - Up to 16 GB HBM2 GPU memory
    - 5,120 CUDA cores
    - 640 Tensor cores
    - GPU memory bandwidth of up to 900 GB/s
    - Support for six NVLink links and a total bandwidth of 300 GB/s (25 GB/s per NVlink link per direction)
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Supports high-performance CPFS.
- Network:
  - Supports IPv6.
  - Supports VPCs.
  - Supports RoCE v2 networks, which are dedicated to low-latency RDMA communication.
- Supported scenarios:
  - Ultra-large-scale training for machine learning on a distributed GPU cluster
  - Large-scale high-performance scientific computing and simulation calculation
  - Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	GPUs	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.sccgn6.24xlarge	96	384.0	NVIDIA V100 × 8	30	4,500,000	50	8	32	10

Note

-
- For more information about these specifications, see Instance family.

Billing methods

SCCs support pay-as-you-go and subscription billing methods. For more information, see Overview.

4.11.2. Create an SCC instance

This topic describes how to create a Super Computing Cluster (SCC) instance.

Context

This topic describes the configuration items you must pay attention to when you create an SCC instance. For more information about other general configuration items, see Create an instance by using the wizard.

To use remote direct memory access (RDMA), HPC schedulers, and cluster scaling services, log on to the E-HPC console, create an SCC cluster, and create an SCC instance. For more information, see Create an E-HPC cluster.

Procedure

1.
2.
3. Click **Create Instance**.
4. Complete basic configurations. Click **Next: Networking**.

   When you configure parameters, note that:

   - **Region**: Select the region and zone in which SCC instance families are available based on the Regions and zones table. Note that the purchase page displays the latest region and zone information, which may differ from information provided in this topic.

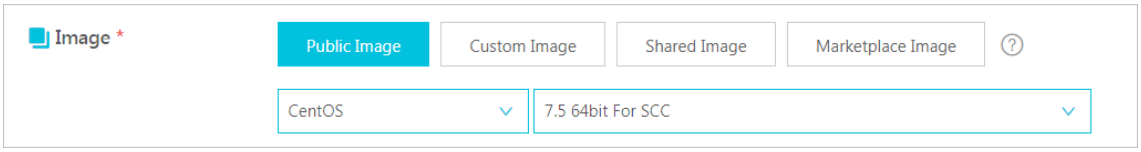
     > **Note** If your business needs to be deployed in a region or zone that is not included in the table, Submit a ticket for special deployment.

Instance family	Regions and zones
scch5	■ China (Hangzhou) Zone H ■ China (Shanghai) Zone D and Zone B ■ China (Qingdao) Zone C ■ China (Zhangjiakou-Beijing Winter Olympics) Zone A
sccg5	■ China (Hangzhou) Zone H ■ China (Shanghai) Zone B ■ China (Qingdao) Zone C
sccgn6	■ China (Shanghai) Zone G ■ China (Zhangjiakou-Beijing Winter Olympics) Zone A

> **Note** If you can view subscription resources but not pay-as-you-go resources when you purchase an instance, access the FAQ, as instructed in Instance FAQ.

- **Instance Type**: Select **Super Computing Cluster**. Currently, CPU-based instance families scch5 and sccg5 and GPU-based instance family sccgn6 are available. After selecting the SCC instance type, you must specify the SCC name and description as required.
- **Image**: Select **Public Image**. Currently, only the custom Linux CentOS 7.5 image for SCC is supported.

> **Note** The customized image supports the RDMA over Converged Ethernet (RoCE)-compliant driver and OpenFabrics Enterprise Distribution (OFED) stack. You can use SCC RDMA through the IB Verbs API and manage RDMA communication through the MPI.



- **Storage**: A maximum of 16 data disks can be attached. You can add a data disk when you create an SCC instance or attach an instance after the instance is created. For more information, see Create a disk and Attach a data disk.

5. Complete networking configurations. Click **Next: System Configurations (Optional)**.

   Only the network type of VPC is supported.
6. Complete system configurations. Click **Next: Grouping (Optional)**.
7. Complete grouping configurations. Click **Next: Preview**.
8. Confirm the order. Click **Create Order**.

4.11.3. sccgn instance family

To further optimize the network performance of GPU-accelerated servers that use the SHENLONG architecture, Alibaba Cloud provides GPU-accelerated compute-optimized Super Computing Cluster (SCC) instance families, which are named sccgn instance families. sccgn instances provide superior computing power and strong network communication capabilities. This topic describes how to use sccgn instances. In the example, the sccgn7ex.32xlarge instance type is used.

Prerequisites

Auto-install RDMA Software Stack is selected when you select an image to create a GPU-accelerated compute-optimized SCC instance. If GPUDirect RDMA is required for your business, **Auto-install GPU Driver** is also selected to install the required software stacks and toolkits.

GPUDirect RDMA is a technology introduced in Kepler-class GPUs and Compute Unified Device Architecture (CUDA) 5.0 to allow direct data exchange between GPUs and third-party devices that use standard PCI Express features. Examples of third-party devices: GPUs, network interfaces, video acquisition devices, and storage adapters. For more information, see NVIDIA documentation.

If the network interface controller (NIC) driver that you want to install is an OpenFabrics Enterprise Distribution (OFED) open source version (download URL), you can install the NIC driver and then install a GPU driver and CUDA.

> **Note** The nvidia-peermem kernel module is integrated into the GPU driver as of CUDA 11.4 and R470. For more information, visit nv_peer_memory.

Context

sccgn instances are equipped with GPUs and high-performance NVIDIA Mellanox ConnectX SmartNICs to deliver superior computing power and strong network communication capabilities. sccgn instances are suitable for scenarios that require high-intensity computing and communication, such as deep learning and high-performance computing.

Functional verification and bandwidth verification

- Functional verification

  This verification checks whether a RDMA software stack is properly installed and configured. Run the following command to perform the check. For information about the issues that you may encounter during the check, see FAQ.

  ```
  rdma_qos_check -V
  ```

  A command output similar to the following one indicates that the RDMA software stack is properly installed:

  ```
  ============================================================
  *    rdma_qos_check
  ------------------------------------------------------------
  * ITEM          DETAIL                                RESULT
  ============================================================
  * link_up       eth1: yes                                ok
  * mlnx_device   eth1: 1                                  ok
  * drv_ver       eth1: 5.2-2.2.3                          ok
  ...
  * pci           0000:c5:00.1                             ok
  * pci           0000:e1:00.0                             ok
  * pci           0000:e1:00.1                             ok
  ============================================================
  ```

- Bandwidth verification

  This verification checks whether RDMA network bandwidth meets the requirements of hardware.

- Run the following command on the server side:

```
ib_read_bw -a -q 20 --report_gbits -d mlx5_bond_0
```

A command output similar to the following one is returned:

```
---------------------------------------------------------------------------------------
                    RDMA_Read BW Test
 Dual-port       : OFF        Device         : mlx5_bond_0
 Number of qps   : 20         Transport type : IB
 Connection type : RC         Using SRQ      : OFF
 PCIe relax order: ON
 ibv_wr* API     : ON
 CQ Moderation   : 100
 Mtu             : 1024[B]
 Link type       : Ethernet
 GID index       : 3
 Outstand reads  : 16
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0000 QPN 0x11ca PSN 0x6302b0 OUT 0x10 RKey 0x17fddc VAddr 0x007f88e
1e5d000
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:200:00:46:14
 local address: LID 0000 QPN 0x11cb PSN 0x99aeda OUT 0x10 RKey 0x17fddc VAddr 0x007f88e
265d000
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:200:00:46:14
 local address: LID 0000 QPN 0x11cc PSN 0xf0d01c OUT 0x10 RKey 0x17fddc VAddr 0x007f88e
2e5d000
 ...
  remote address: LID 0000 QPN 0x11dd PSN 0x8efe92 OUT 0x10 RKey 0x17fddc VAddr 0x007f6
72004b000
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:200:00:45:14
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
 8388608    20000            165.65             165.63            0.002468
---------------------------------------------------------------------------------------
```

- Run the following command on the client side:

```
ib_read_bw -a -q 20 --report_gbits -d mlx5_bond_0 #server_ip
```

A command output similar to the following one is returned:

```
---------------------------------------------------------------------------------------
                    RDMA_Read BW Test
 Dual-port       : OFF        Device         : mlx5_bond_0
 Number of qps   : 20        Transport type : IB
 Connection type : RC        Using SRQ      : OFF
 PCIe relax order: ON
 ibv_wr* API     : ON
 TX depth        : 128
 CQ Moderation   : 100
 Mtu             : 1024[B]
 Link type       : Ethernet
 GID index       : 3
 Outstand reads  : 16
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0000 QPN 0x11ca PSN 0x787f05 OUT 0x10 RKey 0x17fddc VAddr 0x007f671
684b000
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:200:00:45:14
 local address: LID 0000 QPN 0x11cb PSN 0x467042 OUT 0x10 RKey 0x17fddc VAddr 0x007f671
704b000
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:200:00:45:14
 local address: LID 0000 QPN 0x11cc PSN 0xac262e OUT 0x10 RKey 0x17fddc VAddr 0x007f671
784b000
 ...
  remote address: LID 0000 QPN 0x11dd PSN 0xeb1c3f OUT 0x10 RKey 0x17fddc VAddr 0x007f8
8eb65d000
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:200:00:46:14
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
Conflicting CPU frequency values detected: 800.000000 != 3177.498000. CPU Frequency is
not max.
 2          20000           0.058511            0.058226            3.639132
Conflicting CPU frequency values detected: 799.996000 != 3384.422000. CPU Frequency is
not max.
...
Conflicting CPU frequency values detected: 800.000000 != 3166.731000. CPU Frequency is
not max.
 4194304    20000           165.55              165.55              0.004934
Conflicting CPU frequency values detected: 800.000000 != 2967.226000. CPU Frequency is
not max.
 8388608    20000           165.65              165.63              0.002468
---------------------------------------------------------------------------------------
```

When the preceding commands are being run, you can run the `rdma_monitor -s -t -G` command to monitor the bandwidth of each port on NICs in the ECS console.

A command output similar to the following one is returned:

```
------
2022-2-18 09:48:59 CST
tx_rate: 81.874 (40.923/40.951)
rx_rate: 0.092 (0.055/0.037)
tx_pause: 0 (0/0)
rx_pause: 0 (0/0)
tx_pause_duration: 0 (0/0)
rx_pause_duration: 0 (0/0)
np_cnp_sent: 0
rp_cnp_handled: 4632
num_of_qp: 22
np_ecn_marked: 0
rp_cnp_ignored: 0
out_of_buffer: 0
out_of_seq: 0
packet_seq_err: 0
tx_rate_prio0: 0.000 (0.000/0.000)
rx_rate_prio0: 0.000 (0.000/0.000)
tcp_segs_retrans: 0
tcp_retrans_rate: 0
cpu_usage: 0.35%
free_mem: 1049633300 kB
------
```

Use cases of nccl-tests

To test and verify the performance of an instance that uses RDMA networks in your application, the following section provides an example on how to use RDMA to accelerate your application. In the example, nccl-tests is used. For more information about nccl-tests, visit nccl-tests.

```
#!/bin/sh
# Use instances that run Alibaba Cloud Linux 2 operating systems.
# Install openmpi and a compiler.
wget https://ali-perseus-release.oss-cn-huhehaote.aliyuncs.com/openmpi-4.0.3-1.x86_64.rpm
rpm -ivh --force openmpi-4.0.3-1.x86_64.rpm --nodeps
yum install -y gcc-c++
# Modify ~/.bashrc.
export PATH=/usr/local/cuda-11.0/bin:$PATH
export LD_LIBRARY_PATH=/usr/local/cuda/lib64:/usr/local/lib:/usr/local/lib/openmpi:/usr/loc
al/cuda-11.0/lib64:$LD_LIBRARY_PATH
# Download and compile the test code.
git clone https://github.com/NVIDIA/nccl-tests
cd nccl-tests/
make MPI=1 CUDA_HOME=/usr/local/cuda
# Replace host1 and host2 with the IP addresses of the instances.
mpirun --allow-run-as-root -np 16 -npernode 8 -H {host1}:{host2}  \
  --bind-to none \
  -mca btl_tcp_if_include bond0 \
  -x PATH \
  -x CUDA_VISIBLE_DEVICES=0,1,2,3,4,5,6,7 \
  -x NCCL_SOCKET_IFNAME=bond0 \
  -x NCCL_IB_HCA=mlx5 \
  -x NCCL_IB_DISABLE=0 \
  -x NCCL_DEBUG=INFO \
  -x NCCL_NSOCKS_PERTHREAD=8 \
  -x NCCL_SOCKET_NTHREADS=8 \
  -x NCCL_IB_GID_INDEX=3 \
  -x NCCL_DEBUG_SUBSYS=NET,GRAPH \
  -x NCCL_IB_QPS_PER_CONNECTION=4 \
  ./build/all_reduce_perf -b 4M -e 4M -f 2 -g 1 -t 1 -n 20
```

A command output similar to the following one is returned:

```
# Instance output
# nThread 1 nGpus 1 minBytes 4194304 maxBytes 4194304 step: 2(factor) warmup iters: 5 iters
: 20 validation: 1
#
# Using devices
#   Rank  0 Pid  57655 on iZ2ze58t*****3vnehjdZ device  0 [0x54] NVIDIA A100-SXM-80GB
#   Rank  1 Pid  57656 on iZ2ze58t*****3vnehjdZ device  1 [0x5a] NVIDIA A100-SXM-80GB
#   Rank  2 Pid  57657 on iZ2ze58t*****3vnehjdZ device  2 [0x6b] NVIDIA A100-SXM-80GB
#   Rank  3 Pid  57658 on iZ2ze58t*****3vnehjdZ device  3 [0x70] NVIDIA A100-SXM-80GB
#   Rank  4 Pid  57659 on iZ2ze58t*****3vnehjdZ device  4 [0xbe] NVIDIA A100-SXM-80GB
#   Rank  5 Pid  57660 on iZ2ze58t*****3vnehjdZ device  5 [0xc3] NVIDIA A100-SXM-80GB
#   Rank  6 Pid  57661 on iZ2ze58t*****3vnehjdZ device  6 [0xda] NVIDIA A100-SXM-80GB
#   Rank  7 Pid  57662 on iZ2ze58t*****3vnehjdZ device  7 [0xe0] NVIDIA A100-SXM-80GB
#   Rank  8 Pid  58927 on iZ2ze58t*****3vnehjeZ device  0 [0x54] NVIDIA A100-SXM-80GB
#   Rank  9 Pid  58928 on iZ2ze58t*****3vnehjeZ device  1 [0x5a] NVIDIA A100-SXM-80GB
#   Rank 10 Pid  58929 on iZ2ze58t*****3vnehjeZ device  2 [0x6b] NVIDIA A100-SXM-80GB
#   Rank 11 Pid  58930 on iZ2ze58t*****3vnehjeZ device  3 [0x70] NVIDIA A100-SXM-80GB
#   Rank 12 Pid  58931 on iZ2ze58t*****3vnehjeZ device  4 [0xbe] NVIDIA A100-SXM-80GB
#   Rank 13 Pid  58932 on iZ2ze58t*****3vnehjeZ device  5 [0xc3] NVIDIA A100-SXM-80GB
#   Rank 14 Pid  58933 on iZ2ze58t*****3vnehjeZ device  6 [0xda] NVIDIA A100-SXM-80GB
#   Rank 15 Pid  58934 on iZ2ze58t*****3vnehjeZ device  7 [0xe0] NVIDIA A100-SXM-80GB
```

```
#  Rank 15 Pid  58934 on iZ2ze58t*****3vnehjeZ device  7 [0xe0] NVIDIA A100-SXM-80GB
iZ2ze6t9*****ssopZ:57655:57655 [0] NCCL INFO NCCL_SOCKET_IFNAME set to bond0
...
iZ2ze58t*****3vnehjeZ:58929:59248 [2] NCCL INFO NET/IB: Dev 1 Port 1 qpn 4573 mtu 3 GID 3 (
0/22D00C8FFFF0000)
iZ2ze58t*****3vnehjdZ:57657:58004 [2] NCCL INFO NET/IB: Dev 1 Port 1 qpn 4573 mtu 3 GID 3 (
0/22E00C8FFFF0000)
iZ2ze58t*****3vnehjeZ:58927:59225 [0] NCCL INFO Channel 04 : 0[54000] -> 8[54000] [receive]
via NET/IB/0/GDRDMA
iZ2ze58t*****3vnehjeZ:58927:59225 [0] NCCL INFO GPU Direct RDMA Enabled for GPU 54000 / HCA
0 (distance 4 <= 4), read 1
iZ2ze58t*****3vnehjeZ:58931:59227 [4] NCCL INFO NET/IB: Dev 2 Port 1 qpn 4573 mtu 3 GID 3 (
0/62D00C8FFFF0000)
iZ2ze58t*****3vnehjdZ:57659:58012 [4] NCCL INFO NET/IB: Dev 2 Port 1 qpn 4573 mtu 3 GID 3 (
0/62E00C8FFFF0000)
iZ2ze58t*****3vnehjeZ:58933:59183 [6] NCCL INFO NET/IB: Dev 3 Port 1 qpn 4573 mtu 3 GID 3 (
0/A2D00C8FFFF0000)
iZ2ze58t*****3vnehjeZ:58927:59225 [0] NCCL INFO Channel 00 : 8[54000] -> 0[54000] [send] vi
a NET/IB/0/GDRDMA
iZ2ze58t*****3vnehjeZ:58927:59225 [0] NCCL INFO GPU Direct RDMA Enabled for GPU 54000 / HCA
0 (distance 4 <= 4), read 1
iZ2ze58t*****3vnehjdZ:57661:58000 [6] NCCL INFO NET/IB: Dev 3 Port 1 qpn 4573 mtu 3 GID 3 (
0/A2E00C8FFFF0000)
iZ2ze58t*****3vnehjeZ:58927:59225 [0] NCCL INFO Channel 04 : 8[54000] -> 0[54000] [send] vi
a NET/IB/0/GDRDMA
iZ2ze58t*****3vnehjdZ:57655:57848 [0] NCCL INFO NET/IB: Dev 0 Port 1 qpn 4660 mtu 3 GID 3 (
0/E2E00C8FFFF0000)
iZ2ze58t*****3vnehjeZ:58927:59225 [0] NCCL INFO NET/IB: Dev 0 Port 1 qpn 4660 mtu 3 GID 3 (
0/E2D00C8FFFF0000)
iZ2ze58t*****3vnehjeZ:58927:59225 [0] NCCL INFO NET/IB: Dev 0 Port 1 qpn 4661 mtu 3 GID 3 (
0/E2D00C8FFFF0000)
iZ2ze58t*****3vnehjdZ:57655:57848 [0] NCCL INFO NET/IB: Dev 0 Port 1 qpn 4661 mtu 3 GID 3 (
0/E2E00C8FFFF0000)
#
#                                                              out-of-place
in-place
#       size         count      type   redop     time   algbw   busbw  error     time   alg
bw   busbw  error
#        (B)    (elements)                       (us)  (GB/s)  (GB/s)            (us)  (GB/
s)  (GB/s)
     4194304       1048576     float     sum    241.5   17.37   32.56  4e-07    235.2   17.
84   33.44  4e-07
# Out of bounds values : 0 OK
# Avg bus bandwidth    : 33.002
#
```

FAQ

- Problem 1

  When you run the `rdma_qos_check -V` verification command, the `drv_fw_ver eth1: 5.2-2.2.3/22.29.1016 fail` error message is returned.

  Solution:

The error message indicates that the Mellanox NIC firmware is not updated. You can perform the following operations:

- If the instance runs an Alibaba Cloud Linux 2 or CentOS 8.3 operating system, run the `/usr/share/nic-drivers-mellanox-rdma/sources/alifwmanager-22292302 --force --yes` command to update the NIC firmware of the instance.
- If the instance runs a Debian-base operating system, download the firmware update program (download URL) and then run the `./alifwmanager-22292302 --force --yes` command to update the NIC firmware of the instance.

- Problem 2

  When you run the `rdma_qos_check -V` verification command, the `* roce_ver : 0 fail` error message is returned.

  Solution:

  The error message indicates that kernel modules such as configfs and rdma_cm are missing. You can run the `modprobe mlx5_ib && modprobe configfs && modprobe rdma_cm` command to load the required kernel modules.

- Problem 3

  When you run the `systemctl start networking` command on an instance that runs a Debian operating system to start the network service, the system prompts that the bond interfaces cannot be found.

  Solution:

  The error may occur because the the mlx5_ib kernel module is not loaded. You can run the `modprobe mlx5_ib` command to load this kernel module.

- Problem 4

  When you run the `rdma_qos_check -V` verification command or the `ib_read_bw` bandwidth verification command, the `ERROR: RoCE tos isn't correct on mlx5_bond_3` error message is returned.

  Solution:

  You can run the `rdma_qos_init` command to initialize the network.

- Problem 5

  After you restart an instance that runs an Alibaba Cloud Linux 2 operating system, the `cm_tos mlx5_bond_1: 0 fail` error message is returned when you run the `rdma_qos_check -V` verification command.

  Solution:

  You can run the `rdma_qos_init` command to initialize the network.

- Problem 6

  After you restart an instance that runs a CentOS 8.3 operating system, the `trust_mod eth1: pcp fail` error message is returned when you run the `rdma_qos_check -V` verification command.

  Solution:

  You can run the `rdma_qos_init` command to initialize the network.

- Problem 7

  The IP address of the RDMA network interface bond* cannot be obtained.

  Solution:

  You can run the `ifdown bond*` and `ifup bond*` commands to obtain IP address of the bond interface.

  > **Note** Replace `*` with the serial number of the corresponding network interface.

4.12. Burstable instance types

4.12.1. Overview

Burstable instances are an economical instance type that is intended to meet burst performance requirements in entry-level computing scenarios. This topic describes the features, use scenarios, instance families and instance types, baseline performance, CPU credits, and performance modes of burstable instances.

What are burstable instances?

Burstable instances use CPU credits to maintain computing performance and are suitable for scenarios in which CPU utilization is typically low but experiences occasional bursts. Burstable instances can accumulate CPU credits and consume the credits to burst performance above their CPU baseline as needed by your workloads. This consumption pattern does not affect the environments or applications that are deployed on the instances. Burstable instances are more flexible and cost-effective than other types of instances in terms of CPU utilization.

The CPU credit mechanism allows you to save computing power during off-peak hours for use during peak hours and reduce costs. If you have unplanned performance requirements, you can enable unlimited mode for your burstable instances.

The following burstable instance families are available:

- t6, burstable instance family
- t5, burstable instance family

> **Note** Burstable instances are special shared instances. For more information about other shared instance families, see Shared instance families.

The following table describes the baseline performance, CPU credits, and performance modes of burstable instances.

Term	Description	References
baseline performance	The amount of vCPU capacity that is continuously provisioned to a burstable instance. Baseline performance varies based on instance types.	Baseline performance
initial CPU credit	A limited number of CPU credits that are allocated to a new burstable instance. 30 initial credits are allocated to each vCPU. These credits cannot be replenished after they are depleted.	CPU credits

Term	Description	References
CPU credit balance	The net credits that are accrued when the earned CPU credits exceed the consumed credits. You can consume these credits to run instances above their baseline performance levels.	CPU credits
max CPU credit balance	The maximum number of CPU credits that can be earned by a burstable instance within a 24-hour period. The CPU credit balance is valid for 24 hours. Each burstable instance earns CPU credits at a set rate that is determined by the instance type, and can accrue only a limited number of credits in its CPU credit balance.	CPU credits
performance mode	A burstable instance can run in standard or unlimited mode. • A burstable instance in standard mode runs below its baseline performance if its CPU credits are depleted. • A burstable instance in unlimited mode allows you to overdraw or pay for additional CPU credits to sustain the CPU utilization above its baseline performance level at anytime. You may be charged for using these CPU credits.	Performance modes
advance CPU credit	The CPU credits that a burstable instance will earn over the next 24-hour period. You may be charged for using these CPU credits. Advance CPU credits can be used only when the unlimited mode is enabled.	Performance modes
overdrawn CPU credit	The CPU credits that a burstable instance consumes after it depletes its advance credits to keep running above its baseline performance level. You are charged for using overdrawn CPU credits. Overdrawn CPU credits can be used only when the unlimited mode is enabled.	Performance modes

Use scenarios

When you purchase an enterprise-level instance, its vCPUs are exclusively reserved for your use. You are charged for the vCPUs regardless of whether you fully utilize their performance. If you require a high level of CPU performance for only a portion of a day, some vCPU resources are left idle for the remainder of the entire day but you are still charged for these unused resources. To prevent this situation, you can use burstable instances to better meet your business requirements.

Burstable instances apply to scenarios that require higher-than-normal performance for a specific period, such as stress testing service applications, lightweight applications, microservices, and web application servers. We recommend that you evaluate your business requirements to determine the performance levels required during off-peak and peak hours before you make a purchase. The baseline performance of the instances that you purchase must meet your business requirements during off-peak hours to achieve the required performance at significantly lower costs.

Note

- If the purchased burstable instances do not meet your requirements, you can change the configurations. For more information, see Configuration changes.
- The workloads of Windows applications and graphic UIs have high requirements on CPU utilization. If you create an instance of a t-series burstable instance type, the instance may run slowly or go down. We recommend that you change the instance type to better suit your business requirements. For example, you can select g-series general-purpose instance types, c-series compute-optimized instance types, or r-series memory-optimized instance types.

Baseline performance

Baseline performance is the amount of vCPU capacity that is continuously provisioned to a burstable instance. Baseline performance varies based on instance types. You can view the baseline performance of different instance types from the *Baseline CPU computing performance* column of the instance type tables.

CPU credits

CPU credits can be described as computing resources that are available for your use. These computing resources determine the computing performance that your burstable instances can deliver. The following section describes the terms and examples related to CPU credits:

- Initial CPU credit

  When you create a burstable instance, 30 CPU credits are allocated to each vCPU of the instance, which are *initial CPU credits*. These credits enable you to complete deployment tasks when you start the instance.

  For example, an ecs.t5-lc1m2.large instance has two vCPUs and earns 60 initial CPU credits when it is created. An ecs.t5-c1m1.xlarge instance has four vCPUs and earns 120 initial CPU credits when it is created.

- Rate of earning CPU credits

  When a burstable instance starts, it begins to consume CPU credits to maintain its computing performance. At the same time, the instance also earns CPU credits at a set rate that is determined by the instance type. The number of CPU credits that a vCPU can earn per hour varies based on instance types. The *CPU credits per hour* column in the instance type tables indicates the CPU credits that all the vCPUs of an instance can earn per hour.

  For example, 25% baseline performance of an ecs.t5-c1m1.large instance indicates that the CPU credits that a vCPU of the instance earns per hour can keep the vCPU running at 25% utilization for 1 hour or at 100% utilization for 15 minutes (60 × 25%). In response to its baseline performance, each vCPU earns 15 CPU credits per hour. An ecs.t5-c1m1.large instance has two vCPUs and earns 30 CPU credits per hour.

- CPU credit balance

If the earned CPU credits exceed the consumed credits, the net credits are accrued as *CPU credit balance*. The CPU credit balance is valid for 24 hours. Each burstable instance earns CPU credits at a set rate that is determined by the instance type, and can accrue only a limited number of credits in its CPU credit balance. The maximum CPU credit balance of a specific instance type is the maximum number of CPU credits that the instance can earn within a 24-hour period. For more information, see the *Max CPU credit balance* column in the instance type tables.

For example, an ecs.t5-c1m1.large instance can earn 30 CPU credits per hour. The maximum CPU credit balance that the instance can earn is 720 (30 × 24) credits.

- Rate of consuming CPU credits

  The rate at which a burstable instance consumes CPU credits is determined based on the number of vCPUs, CPU utilization, and operating hours of the instance. For example, one CPU credit is consumed in the following scenarios:

  - One vCPU runs at 100% utilization for 1 minute.
  - One vCPU runs at 50% utilization for 2 minutes.
  - Two vCPUs run at 25% utilization for 2 minutes.

  When a burstable instance starts, it begins to consume CPU credits to maintain its computing performance. Initial credits that cannot be replenished are consumed first. When the initial credits are depleted, the instance continues to consume the accrued CPU credits.

  - When the CPU utilization is below the baseline, the credits earned are more than the credits consumed and the CPU credit balance increases.
  - When the CPU utilization is equal to the baseline, the credits earned are equal to the credits consumed and the CPU credit balance remains unchanged.
  - When the CPU utilization is higher than the baseline, the credits earned are less than the credits consumed and the CPU credit balance decreases.

> **Note** The CPU utilization data is collected at the physical machine level and includes the CPU consumption when the privileged commands are run during the virtualization of the ECS instance. You can log on to the CloudMonitor console and click the instance ID on the **Host Monitoring** page, and then click the **Basic Monitoring** tab on the instance details page to view the CPU utilization of the instance. For more information, see Overview.

In different scenarios, stopping your instances may have different impacts on your CPU credits:

- If a pay-as-you-go instance is stopped in standard mode, the CPU credit balance of the instance is retained and the instance continues to earn CPU credits.
- If a pay-as-you-go instance is stopped in economical mode, the CPU credit balance of the instance becomes invalid and the instance cannot continue to earn credits. When the instance is started again, it receives initial credits and begins to earn credits again.
- If a pay-as-you-go instance is stopped due to an overdue payment, the CPU credit balance of the instance is retained but the instance cannot continue to earn credits until you complete the payment.
- If a subscription instance expires and is stopped, the CPU credit balance of the instance is retained but the instance cannot continue to earn credits. When the instance is reactivated, it begins to earn credits again.

Performance modes

A burstable instance can run in standard or unlimited mode.

- Standard mode

  The performance of a burstable instance in standard mode is limited by the availability of CPU credits. After the instance depletes all of its initial credits and CPU credit balance, the instance cannot run above its baseline performance level. When the CPU credit balance is low, the instance gradually reduces performance to its baseline level within 15 minutes. This way, the instance does not experience a sharp performance drop-off when its CPU credit balance is depleted.

  The standard mode is applicable to scenarios such as lightweight web servers, development and testing environments, and databases that have low and medium performance. In these scenarios, workloads are stable, instances do not run above the baseline performance level for an extended period of time, and bursts in performance are only occasionally required.

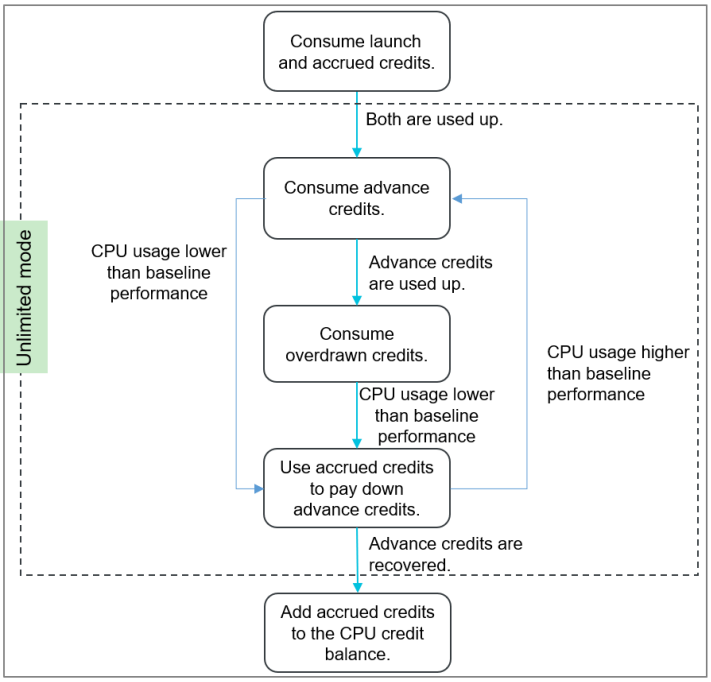
- Unlimited mode

  The performance of a burstable instance in unlimited mode is not limited by the availability of CPU credits. You can overdraw or pay for additional CPU credits to obtain performance boosts at any time. If your instances continue to run above the baseline performance level after the initial CPU credits and accrued credits are depleted, the instance begins to consume advance CPU credits and overdrawn CPU credits.

  - Advance CPU credits: Advance CPU credits are the credits that a burstable instance will earn over the next 24-hour period. You may be charged for using these CPU credits.
  - Overdrawn CPU credits: Overdrawn CPU credits are the CPU credits that a burstable instance consumes after it depletes its advance credits to keep running above its baseline performance level. You are charged for using overdrawn credits.

  > **Note** For more information about the fees and billing rules of using advance and overdrawn CPU credits, see Additional fees.

  The following figure shows how the CPU credits change when an instance is running in unlimited mode.



> **Note** If the instance has consumed advance CPU credits and is stopped in economical mode, is released, has its configurations changed, or switches to standard mode before the advance CPU credits are replenished, you are charged a lump sum amount for the consumed advance CPU credits.

You can enable unlimited mode for your burstable instances if you want to consume advance or overdrawn CPU credits in addition to your credit balance to meet burst performance requirements. Examples:

- Some events such as new feature releases, e-commerce promotions, and website promotions cause a substantial increase in your workloads. High CPU performance is required during this period of time. In this case, you can enable unlimited mode for your burstable instances. You can disable this mode to reduce costs when the workload peak ends.
- Some web applications may require CPU bursts for a specific period during a day, but the daily average CPU utilization is below the baseline. In this case, you can enable unlimited mode for your instances during peak hours to ensure positive user experience. If the CPU credits that you have earned during off-peak hours can offset the advance CPU credits that you consumed during peak hours, you can continue to provide positive user experience at no additional costs.

By default, the standard mode is enabled when you create a burstable instance. For more information about how to enable the unlimited mode, see Enable the unlimited mode.

For more information about how CPU credits change when an instance is running in different performance modes, see CPU credit change examples.

Configuration changes

When you monitor a burstable instance, you may find that its CPU utilization remains above or below the baseline for an extended period of time. This indicates that the instance type is not suitable for your business. We recommend that you re-evaluate the instance type to decide whether to select another burstable or enterprise-level instance type. The operation of changing instance types varies based on the billing methods. For more information, see Overview of instance configuration changes.

t6, burstable instance family

Features:

- Provides baseline CPU performance and is burstable but limited by accrued CPU credits.
- Is more cost-effective when compared with the t5 burstable instance family.
- Compute:
  - Uses 2.5 GHz Intel® Xeon® Cascade Lake processors that deliver a turbo frequency of 3.2 GHz.
  - Uses DDR4 memory.
- Storage:
  - Supports enhanced SSDs (ESSDs), standard SSDs, and ultra disks.

    > **Note** ESSDs at performance level (PL) 2 and 3 cannot provide maximum performance due to the specification limits of burstable instances. We recommend that you use enterprise-level instances or ESSDs that are at lower performance levels.
- Network:

- Supports IPv6.
- Supports only virtual private clouds (VPCs).
- Delivers a bandwidth up to 4 Gbit/s.

- Supported scenarios:
  - Web application servers
  - Lightweight applications and microservices
  - Development and testing environments

Instance types

Instance type	vCPU	Memory (GiB)	Baseline CPU performance	CPU credits per hour	Max CPU credit balance	Base bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.t6-c4m1.large	2	0.5	5%	6	144	0.08	40,000	1	2	2
ecs.t6-c2m1.large	2	1.0	10%	12	288	0.08	60,000	1	2	2
ecs.t6-c1m1.large	2	2.0	20%	24	576	0.08	100,000	1	2	2
ecs.t6-c1m2.large	2	4.0	20%	24	576	0.08	100,000	1	2	2
ecs.t6-c1m4.large	2	8.0	30%	36	864	0.08	100,000	1	2	2
ecs.t6-c1m4.xlarge	4	16.0	40%	96	2304	0.16	200,000	1	2	6
ecs.t6-c1m4.2xlarge	8	32.0	40%	192	4608	0.32	400,000	1	2	6

Note

- Secondary elastic network interfaces (ENIs) cannot be bound to instances of this instance family while the instances are being created and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from instances of the following instance types, the instances must be in the Stopped state: ecs.t6-c1m1.large, ecs.t6-c1m2.large, ecs.t6-c1m4.large, ecs.t6-c2m1.large, and ecs.t6-c4m1.large.
-
- For more information about these specifications, see Instance family.

t5, burstable instance family

Features:

- Provides baseline CPU performance and is burstable but limited by accrued CPU credits.
- Offers a balance between compute, memory, and network resources.
- Compute:
  - Offers multiple CPU-to-memory ratios.
  - Uses 2.5 GHz Intel® Xeon® processors.
  - Uses DDR4 memory.
- Network:
  - Supports IPv6.
  - Supports only VPCs.
- Supported scenarios:
  - Web application servers
  - Lightweight applications and microservices
  - Development and testing environments

Instance types

Instance type	vCPU	Memory (GiB)	Baseline CPU performance	CPU credits per hour	Max CPU credit balance	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.t5-lc2m1.nano	1	0.5	20%	12	288	0.1	40,000	1	2	2
ecs.t5-lc1m1.small	1	1.0	20%	12	288	0.2	60,000	1	2	2

Instance type	vCPU	Memory (GiB)	Baseline CPU performance	CPU credits per hour	Max CPU credit balance	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.t5-lc1m2.small	1	2.0	20%	12	288	0.2	60,000	1	2	2
ecs.t5-lc1m2.large	2	4.0	20%	24	576	0.4	100,000	1	2	2
ecs.t5-lc1m4.large	2	8.0	20%	24	576	0.4	100,000	1	2	2
ecs.t5-c1m1.large	2	2.0	25%	30	720	0.5	100,000	1	2	2
ecs.t5-c1m2.large	2	4.0	25%	30	720	0.5	100,000	1	2	2
ecs.t5-c1m4.large	2	8.0	25%	30	720	0.5	100,000	1	2	2
ecs.t5-c1m1.xlarge	4	4.0	25%	60	1440	0.8	200,000	1	2	6
ecs.t5-c1m2.xlarge	4	8.0	25%	60	1440	0.8	200,000	1	2	6
ecs.t5-c1m4.xlarge	4	16.0	25%	60	1440	0.8	200,000	1	2	6

Instance type	vCPU	Memory (GiB)	Baseline CPU performance	CPU credits per hour	Max CPU credit balance	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.t5-c1m1.2xlarge	8	8.0	25%	120	2880	1.2	400,000	1	2	6
ecs.t5-c1m2.2xlarge	8	16.0	25%	120	2880	1.2	400,000	1	2	6
ecs.t5-c1m4.2xlarge	8	32.0	25%	120	2880	1.2	400,000	1	2	6
ecs.t5-c1m1.4xlarge	16	16.0	25%	240	5760	1.2	600,000	1	2	6
ecs.t5-c1m2.4xlarge	16	32.0	25%	240	5760	1.2	600,000	1	2	6

Note

- Secondary ENIs cannot be bound to instances of this instance family while the instances are being created and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from instances of the following instance types, the instances must be in the Stopped state: ecs.t5-lc2m1.nano, ecs.t5-c1m1.large, ecs.t5-c1m2.large, ecs.t5-c1m4.large, ecs.t5-lc1m1.small, ecs.t5-lc1m2.large, ecs.t5-lc1m2.small, and ecs.t5-lc1m4.large.
-
- For more information about these specifications, see Instance family.

4.12.2. Benefits

Burstable instances are the only instances that use CPU credits for maximum cost efficiency. This topic describes the benefits of burstable instances.

Lower cost

The cost of a burstable instance is 10% to 48% lower than that of a shared instance with the same configurations.

> **Note** The actual performance of a burstable instance depends on the accumulation and usage of CPU credits. Before you purchase a burstable instance, make sure that you fully understand the relevant terms such as baseline performance and CPU credits of burstable instances. For more information, see Overview.

Finer-granularity specifications

ecs.t5-lc2m1.nano is an instance type of burstable instances, which has the baseline level specifications of one vCPU and 0.5 GiB memory. The instance specifications of burstable instances allow you to flexibly combine instance types to meet finer-granularity requirements. For more information, see t5, burstable instance family.

Predictable CPU performance

Burstable instances provide a baseline level of CPU performance (*baseline performance*). You can predict the burstable performance of a burstable instance based on its baseline performance and accrued CPU credits, and select appropriate compute capabilities.

Higher network performance

Burstable instances use the most advanced network-based technologies to reduce network latency to one-third of that of the last-generation instances.

4.12.3. CPU credit change examples

After a burstable instance is created, its CPU credits change based on the relationship between its CPU utilization and the baseline performance. This topic describes how the CPU credits change for a burstable instance in different performance modes.

Background information

The examples given in this topic are used to describe concepts such as baseline performance and CPU credits. Actual business scenarios may be more complex and variable than those provided in these examples. For example, the CPU utilization is unlikely to remain at a constant value for very long. We recommend that you select appropriate instances based on your understanding of burstable instance-related concepts and manage the performance modes or configurations of the instances to suit your needs. For more information, see Switch the performance mode of a burstable instance or Configuration changes.

Before you read the examples, take note of the following items:

- After a burstable instance is created, each of its vCPUs can earn 30 initial CPU credits.
- The consumption rate of CPU credits of a burstable instance is subject to the number of vCPUs, CPU utilization, and instance running hours. One CPU credit is equal to one vCPU at 100% utilization for one minute. When the actual performance is other values, the running time is converted proportionally.
- When a burstable instance runs at the baseline performance level, the CPU credits it earns are equal to the CPU credits it consumes.

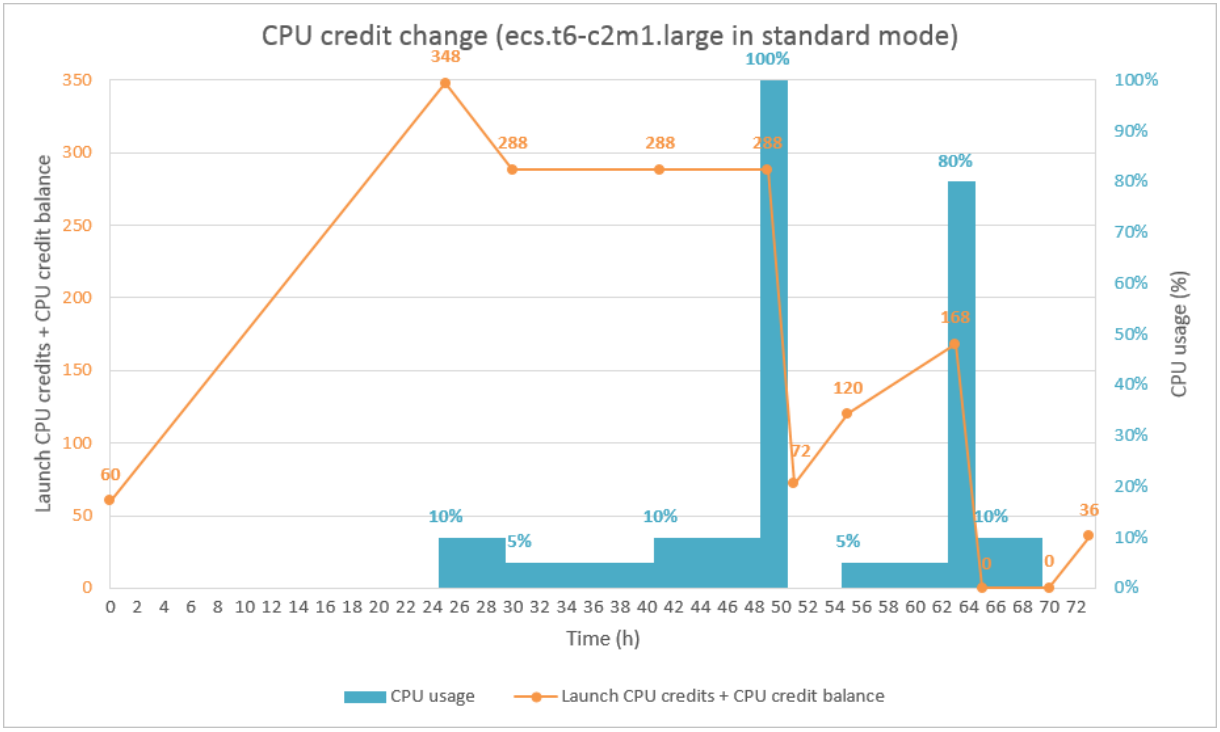
For more information, see Baseline performance and CPU credits.

Standard mode

A burstable instance in standard mode cannot burst above the performance baseline after its initial CPU credits and CPU credit balance are depleted.

The following figure shows how the CPU credits of an ecs.t6-c2m1.large instance that runs in standard mode change. The instance has 2 vCPUs and 1 GiB memory. Take note of the following items:

- The instance has two vCPUs and can earn 60 initial CPU credits.
- The performance baseline of the instance is 10%.
- The instance earns 12 CPU credits per hour and has a CPU credit balance limit of 288. For more information about the instance type, see the "t6, burstable instance family" section of the Overview topic.
- The instance has two vCPUs and consumes 12 CPU credits per hour when it runs at the baseline performance level.



The following section describes how CPU credits change in different phases shown in the preceding figure.

- 0 h ~ 24 h

  Phase A: 60 initial CPU credits are earned on instance startup. The CPU utilization is 0%. The CPU credit balance keeps increasing until it reaches the limit in a 24-hour period after startup.

  At the end of these hours, the instance has a CPU credit balance of 348, which is calculated based on the following formula: 60 initial credits + 288 maximum credit balance.

- 25 h ~ 48 h

  i. Phase B: The CPU utilization is 10%, which is equal to the performance baseline. The initial CPU credits are consumed first. When the instance is running, 12 CPU credits are consumed per hour. After the 60 initial CPU credits are consumed, the instance cannot obtain more initial CPU credits.

At the end of these hours, the instance has a CPU credit balance of 288, which is calculated based on the following formula: 348 available credits at the end of Phase A - 60 consumed initial credits.

ii. Phase C: The CPU utilization is 5%, which is below the performance baseline. However, the CPU credit balance has reached the upper limit and remains constant.

At the end of these hours, the credit balance is 288 and is equal to the maximum CPU credit balance.

iii. Phase D: The CPU utilization is 10%, which is equal to the performance baseline. The instance earns and consumes equal CPU credits, and the CPU credit balance remains constant.

At the end of these hours, the credit balance is 288 and is equal to the maximum CPU credit balance.

- 49 h ~ 72 h

i. Phase E: The CPU utilization is 100%. The instance runs for 2 hours and consumes 120 CPU credits per hour. In this case, the baseline performance cannot meet requirements, and the instance begins to consume the CPU credit balance.

At the end of these hours, the instance has a CPU credit balance of 72, which is calculated based on the following formula: 288 maximum CPU credit balance - 2 × 120 credits consumed per hour + 2 × 12 credits earned per hour.

ii. Phase F: The CPU utilization is 0%. The instance is idle for 4 hours and earns 12 CPU credits per hour. All earned credits are accrued in the credit balance.

At the end of these hours, the instance has a CPU credit balance of 120, which is calculated based on the following formula: 72 credit balance at the end of Phase E + 4 × 12 credits earned per hour.

iii. Phase G: The CPU utilization is 5%. The instance runs for 8 hours and consumes 6 CPU credits per hour. The unconsumed credits are accrued in the credit balance.

At the end of these hours, the instance has a CPU credit balance of 168, which is calculated based on the following formula: 120 credit balance at the end of Phase F - 8 × 6 credits consumed per hour + 8 × 12 credits earned per hour.

iv. Phase H: The CPU utilization is 80%, and the baseline performance cannot meet requirements. The instance runs for 2 hours and consumes 96 CPU credits per hour. The CPU credit balance is depleted. When the instance runs in standard mode and has no available CPU credits, it cannot burst beyond the performance baseline.

> **Note** When the CPU credit balance is low, the instance gradually reduces performance to its baseline level within 15 minutes. This way, the instance does not experience a sharp performance drop-off when its accrued CPU credit balance is depleted.

At the end of these hours, the instance has a CPU credit balance of zero, which is calculated based on the following formula: 168 credit balance at the end of Phase G - 2 × 96 credits consumed per hour + 2 × 12 credits earned per hour.

v. Phase I: The CPU utilization is 10%, which is equal to the performance baseline. The instance earns and consumes equal CPU credits, and the CPU credit balance remains constant.

At the end of this these hours, the instance has a CPU credit balance of zero, which is calculated based on the following formula: 0 credit balance at the end of Phase H - 5 × 12 credits consumed per hour + 5 × 12 credits earned per hour.

vi. Phase J: The CPU utilization is 0%. The instance is idle for 3 hours and earns 12 CPU credits per hour. All earned credits are accrued in the credit balance.

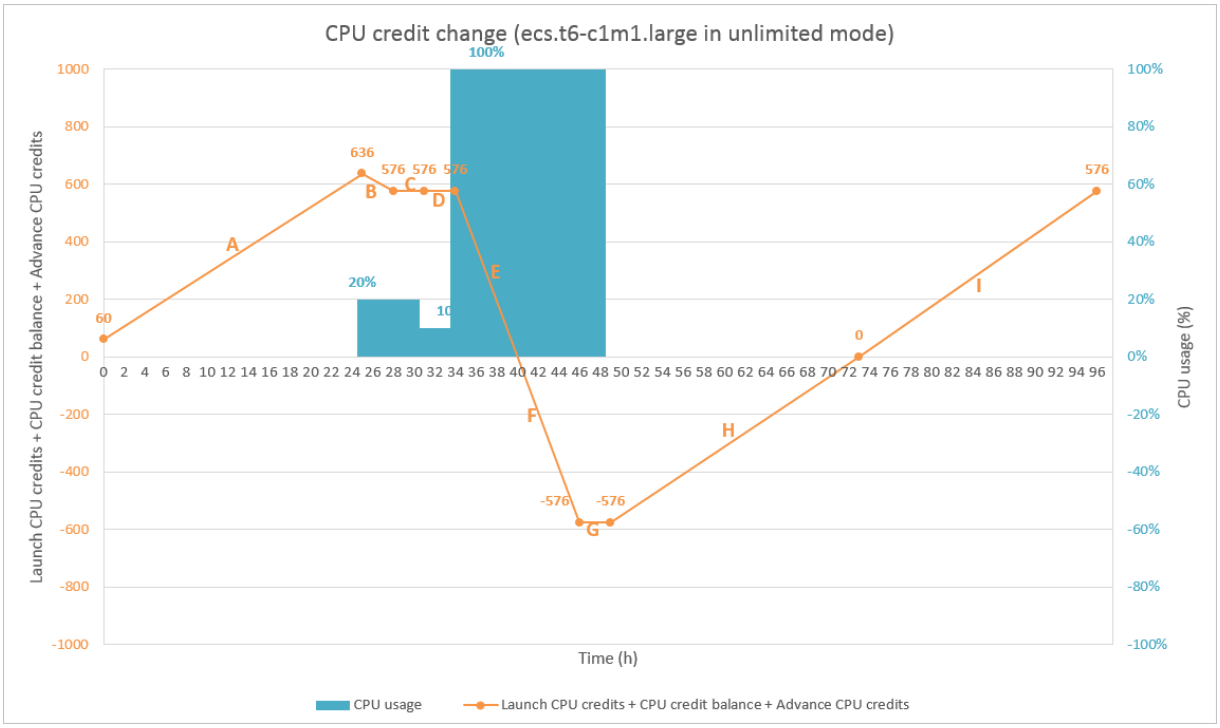
At the end of these hours, the instance has a CPU credit balance of 36, which is calculated based on the following formula: 0 credit balance at the end of Phase I + 3 × 12 credits earned per hour.

Unlimited mode

The performance of a burstable instance in unlimited mode is not limited by the availability of CPU credits. You can overdraw or pay for additional CPU credits to obtain performance boosts at any time.

The following figure shows how the CPU credits of an ecs.t6-c1m1.large instance that runs in unlimited mode change. The instance has 2 vCPUs and 2 GiB memory. Take note of the following items:

- The instance has two vCPUs and can earn 60 initial CPU credits.
- The instance has a performance baseline of 20%.
- The instance earns 24 CPU credits per hour and has a CPU credit balance limit of 576. For more information about the instance type, see the "t6, burstable instance family" section of the Overview topic.
- The instance has two vCPUs and consumes 24 CPU credits per hour when it runs at the baseline performance level.



The following section describes how CPU credits change in different phases shown in the preceding figure.

- 0 h ~ 24 h

  Phase A: 60 initial CPU credits are earned on instance startup. The CPU utilization is 0%. The CPU credit balance keeps increasing until it reaches the limit in a 24-hour period after startup.

  At the end of these hours, the instance has a CPU credit balance of 636, which is calculated based on the following formula: 60 initial credits + 576 maximum CPU credit balance.

- 25 h ~ 48 h

i. Phase B: The CPU utilization is 20%, which is equal to the performance baseline. The initial CPU credits are consumed first. When the instance is running, 24 CPU credits are consumed per hour. After the 60 initial CPU credits are consumed, the instance cannot obtain more initial CPU credits.

   At the end of these hours, the instance has a CPU credit balance of 576, which is calculated based on the following formula: 636 credits at the end of Phase A - 60 consumed initial credits.

ii. Phase C: The CPU utilization is 20%, which is equal to the performance baseline. The instance earns and consumes equal CPU credits, and the CPU credit balance remains constant.

   At the end of these hours, the CPU credit balance of the instance is 576 and is equal to the maximum CPU credit balance.

iii. Phase D: The CPU utilization is 10%, which is below the performance baseline. However, the CPU credit balance has reached the upper limit and remains constant.

   At the end of these hours, the CPU credit balance of the instance is 576 and is equal to the maximum CPU credit balance.

iv. Phase E: The CPU utilization is 100%. When the instance is running, it consumes 120 CPU credits per hour. In this case, the baseline performance cannot meet requirements, and the instance begins to consume the CPU credit balance.

   At the end of these hours, the CPU credit balance is depleted.

v. Phase F: The CPU utilization is 100%. When the instance is running, it consumes 120 CPU credits per hour. In this case, the baseline performance cannot meet requirements, and the instance begins to consume the advance CPU credits. For more information, see the "performance modes" section of the Overview topic.

   At the end of these hours, the advance CPU credits are depleted, and a total of 576 CPU credits are overdrawn.

vi. Phase G: The CPU utilization is 100%. When the instance is running, it consumes 120 CPU credits per hour. In this case, the baseline performance cannot meet requirements, and the instance begins to consume the overdrawn CPU credits. For more information, see the "performance modes" section in Overview.

   At the end of these hours, the CPU credit balance remain constant, and 576 CPU credits are overdrawn.

- 49 h ~ 72 h

  Phase H: The CPU utilization is 0%. The earned CPU credits preferentially pay down the consumed advance CPU credits and can pay off the advanced CPU credits in 72 hours.

  At the end of these hours, no overdrawn CPU credits are available, but the CPU credit balance is still 0.

- 73 h ~ 96 h

  Phase H: The CPU utilization is 0%. The instance is idle for 24 hours and earns 24 CPU credits per hour. All earned credits are accrued in the credit balance until the credit balance reaches the limit in a 96-hour period.

  At the end of these hours, the CPU credit balance is 576 and is equal to the maximum CPU credit balance.

4.12.4. Billing

Burstable instances consume CPU credits to maintain running performance. The CPU credit balance accrued during off-peak hours can be used to improve running performance during peak hours and reduce costs. If your burstable instances consume advance or overdrawn CPU credits, you are charged additional fees. The fees for a burstable instance include the instance purchase fee and additional fees.

Instance purchase fee

Burstable instances support the pay-as-you-go, subscription, and preemptible instance billing methods. For more information, see Overview. For information about the pricing of burstable instances, see the Pricing tab of the Elastic Compute Service (ECS) product page.

If the billing method of your burstable instances is pay-as-you-go, you can purchase reserved instances to offset the bills of these pay-as-you-go instances. For more information, see Overview. However, if the instance types of your reserved instances belong to the t5 instance family, the following limits apply:

- Only zonal reserved instances can be purchased.
- Reserved instances cannot be split or merged.

Additional fees

The billing of a burstable instance is related to its performance mode. The following section describes the differences between the billing of burstable instances in different performance modes.

> **Note** When you purchase or use a burstable instance, you can choose whether to enable the unlimited mode for the burstable instance. For information about the consumption rules of CPU credits in unlimited mode, see the "Unlimited mode" section in Overview.

- Standard mode: You are charged only the burstable instance purchase fee. You are charged no additional fees when you use a burstable instance in standard mode.
- Unlimited mode: You are charged both the burstable instance purchase fee and additional fees that may be incurred in the following cases:
  - The burstable instance has used up advance CPU credits and started to consume overdrawn CPU credits. You are billed and charged for the consumed overdrawn CPU credits by the hour.
  - If a burstable instance has consumed advance CPU credits and is stopped in economical mode, is released, has its configurations changed, or switches to standard mode before the advance CPU credits are replenished, you are charged a lump sum amount for the consumed advance CPU credits.

The following table describes how the additional fees are charged when burstable instances that run in unlimited mode consume CPU credits in different regions.

Region	Windows instance (USD/credit)	Linux instance (USD/credit)
Region inside the Chinese mainland	0.0008	0.0008
Region outside the Chinese mainland	0.0016	0.0008

Example scenarios for charging additional fees:

- Scenario 1: If a Linux instance in a region inside the Chinese mainland consumes 100 overdrawn CPU credits after it depletes its advance CPU credits, the additional fee for the instance is calculated based on the following formula: 100 (Number of consumed overdrawn CPU credits) × 0.0008 (Unit price per CPU credit for a Linux instance in a region inside the Chinese mainland) = USD 0.08.
- Scenario 2: If a Windows instance in a region outside the Chinese mainland first consumes 100 advance CPU credits in unlimited mode and then consumes another 100 advance CPU credits in standard mode, the additional fee is calculated based on the following formula: 100 (Number of consumed advance CPU credits in unlimited mode) × 0.0016 (Unit price per CPU credit for a Windows instance in a region outside the Chinese mainland) = USD 0.16.

4.12.5. Create a burstable instance

Burstable instances can use CPU credits to burst performance when required by your workloads. You can enable the unlimited mode for burstable instances. The minimum memory of burstable instances can be 0.5 GiB. This topic describes the configuration items of which you must take note when you create a burstable instance.

Context

For information about other general configuration items, see Create an instance by using the wizard.

Procedure

1.
2.
3. Click **Create Instance**.
4. Configure the parameters in the Basic Configurations step. Click **Next: Networking**.

   When you configure parameters, note the following parameters:

   - **Region**: Instance types that are available to your account in each region are displayed on the instance creation page.

     > Note

   - **Instance Type**: Select **x86-Architecture > Entry-Level (Shared)**. You can select **Enable Unlimited Mode for t5 Instances** when you create a burstable instance or after the instance is created. For more information, see Enable the unlimited mode.

   - **Image**: Burstable instance types have a minimum memory of 0.5 GiB. They support only Linux and Windows Server Version 1809, and do not support operating systems that require more than 0.5 GiB memory, such as Windows Server 2016. For more information, see Select an image.
5. Configure the network and security group parameters. Click **Next: System Configurations**.

   Only the network type of VPC is supported.
6. Configure the system parameters. Click **Next: Grouping**.
7. Configure the grouping parameters. Click **Next: Preview**.
8. Read and select *ECS Service Terms*, and click Create Order.

Related information

- RunInstances

4.12.6. Switch the performance mode of a burstable instance

Burstable instances support the standard and unlimited modes to suit the requirements of different business scenarios. This topic describes how to query and switch the performance mode of a burstable instance.

Context

The performance of a burstable instance in standard mode is limited by the availability of its CPU credits. After its initial credits and CPU credit balance are exhausted, the instance is unable to burst beyond its performance baseline. The performance of a burstable instance in unlimited mode is not limited by the availability of CPU credits. You can overdraw or pay for additional CPU credits to obtain performance boosts at any time.

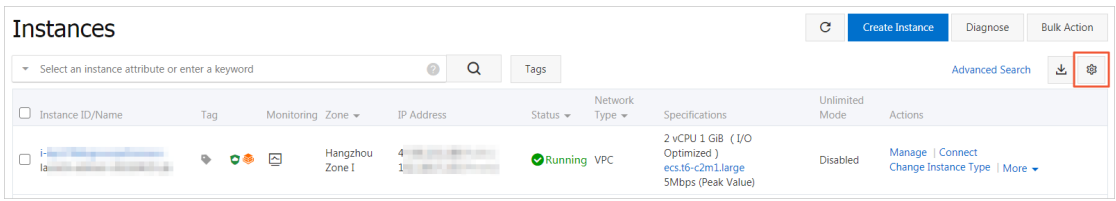
> **Note** In this case, you may be charged for the consumption of these CPU credits. For more information, see Additional fees.

In the following scenarios, the system selects or switches the performance mode for a burstable instance:

- By default, the standard mode is enabled for new burstable instances.
- If a burstable instance is in the **Stopped** state and in economical mode, the instance runs in standard mode after it is started again.
- If a burstable instance is in the **Stopped** state and is not in economical mode, the performance mode used before the instance is stopped continues to take effect after the instance is started again.
- If you have overdue payments within your account, the unlimited mode is automatically disabled for burstable instances and is not re-enabled until you settle the payments.

Query the performance mode of a burstable instance

1.
2.
3. (Optional)If the **Unlimited Mode** column is not displayed on the **Instances** page, configure the column to be displayed.
   i. Click the ⚙ icon in the upper-right corner.



   ii. In the Column Filters dialog box, select **Unlimited Mode** and click **OK**.

4. In the **Unlimited Mode** column, view the mode of the burstable instance.
   - **Disabled**: indicates that the instance runs in standard mode.
   - **Enabled**: indicates that the instance runs in unlimited mode.

Enable the unlimited mode

If a burstable instance is running in standard mode, you can enable the unlimited mode for the instance.

> **Note** Make sure that the burstable instance is in the **Running** state. Otherwise, you cannot switch the performance mode for the instance.

1.
2.
3. Find the burstable instance and use one of the following methods to enable the unlimited mode for the instance:
   - To enable the unlimited mode for a single burstable instance at a time, choose **More > Instance Settings > Enable Unlimited Mode** in the **Actions** column corresponding to the instance.
   - To enable the unlimited mode for one or more burstable instances at a time, select the instances and choose **More > Instance Settings > Enable Unlimited Mode** in the lower-left corner of the Instances page.
4. In the **Enable Unlimited Mode** message, click **OK**.

Disable the unlimited mode

If a burstable instance is running in unlimited mode, you can disable the unlimited mode for the instance.

> **Note** Make sure that the burstable instance is in the **Running** state. Otherwise, you cannot switch the performance mode for the instance.

1.
2.
3. Find the burstable instance and use one of the following methods to disable the unlimited mode for the instance:
   - To disable the unlimited mode for a single burstable instance at a time, choose **More > Instance Settings > Disable Unlimited Mode** in the **Actions** column corresponding to the instance.
   - To disable the unlimited mode for one or more burstable instances at a time, select the instances and choose **More > Instance Settings > Disable Unlimited Mode** in the lower-left corner of the Instances page.
4. In the **Disable Unlimited Mode** message, click **OK**.

Related information

- ModifyInstanceAttribute

4.12.7. Monitor burstable instances

This topic describes how to query the CPU utilization and credits of a burstable instance in the Elastic Compute Service (ECS) console and how to configure alert rules on CPU credits in the CloudMonitor console.

Prerequisites

Before you can configure contacts to receive notifications, you must create a contact group. For more information, see Create an alert contact or alert contact group.

Context

Changes to the number of CPU credits of a burstable instance directly affect the CPU utilization and load performance of the instance. You can configure alert rules that include the following monitoring metrics for one or more burstable instances in the CloudMonitor console: BurstCredit, TotalCredit, NotpaidSurplusCredit, and AdvanceCredit. The following table describes the monitoring metrics for CPU credits of burstable instances.

Monitoring metric	Description
BurstCredit	The changes in CPU credit consumption. Consumption trends are consistent with CPU utilization. For more information, see CPU credits.
TotalCredit	The changes in CPU credit balance. CPU credit balance can be consumed to maintain CPU utilization. For more information, see CPU credits.
NotpaidSurplusCredit	The changes in the number of overdrawn CPU credits. Overdrawn CPU credits can be used only after the unlimited mode is enabled. For more information, see Performance modes.
AdvanceCredit	The changes in the number of advance CPU credits. Advance CPU credits can be used only after the unlimited mode is enabled. For more information, see Performance modes.

View CPU credit usage information

This section describes how to view the real-time CPU credit trends of a burstable instance in the ECS console.

1.
2.
3.
4. Find the burstable instance and click its ID to go to the **Instance Details** page.
5. Click the **Monitoring** tab and view the CPU credit usage and CPU utilization of the instance.

> Note
>
> - CPU utilizations displayed in the ECS console are the values measured by instance operating systems and are not used to calculate CPU credit usage.
> - CPU utilizations that are collected from physical machines incorporate the simulation overheads of privileged instructions and are used to calculate CPU credit usage. You can log on to the CloudMonitor console, click **Host Monitoring** in the left-side navigation pane, and then click the ID of an ECS instance on the Host Monitoring page. Then, you can click the **Basic Monitoring** tab and view the CPU utilization of the instance. For more information, see Overview.

Create alert rules on CPU credits

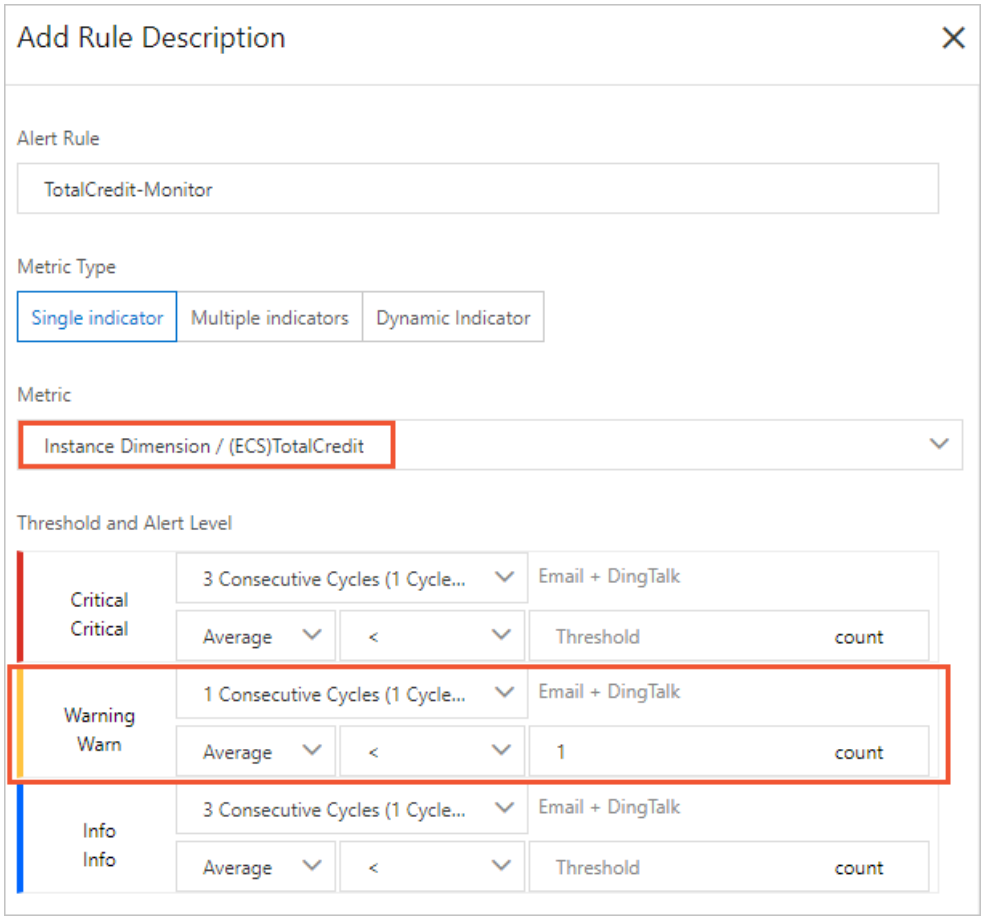
This section describes how to create alert rules that include the **TotalCredit** and **NotpaidSurplusCredit** monitoring metrics in the CloudMonitor console. Take note of the following items:

- In standard mode, if a burstable instance does not have available CPU credits, its CPU utilization cannot burst above the baseline level. You can monitor the **TotalCredit** metric to receive notifications when instance performance is limited and determine whether to enable the unlimited mode.
- In unlimited mode, after a burstable instance consumes all of its advance CPU credits, the instance consumes overdrawn CPU credits to continue to run above its performance baseline. The consumed overdrawn CPU credits are billed and charged on an hourly basis. You can monitor the **NotpaidSurplusCredit** metric to receive notifications when overdrawn CPU credits are billed and determine whether to disable the unlimited mode.

1.
2.
3. In the left-side navigation pane, choose **Alerts > Alert Rules**.
4. On the **Alert Rules** page, click **Create Alert Rule**.
5. In the **Create Alert Rule** panel, configure parameters.
   i. Configure instance-related parameters:
      - **Product**: Select **ECS**. from the drop-down list.
      - **Resource Range**: Select **Instances**.
      - **Associated Resources**: Select one or more burstable instances from the drop-down list.
   ii. Click **Add Rules**. In the **Add Rule Description** panel, configure parameters to create an alert rule. If you want to create multiple alert rules, repeat this step.
      - **Alert Rule**: Enter a name for the alert rule.
      - **Metric Type**: Select **Single indicator**.
      - **Metric** and **Threshold and Alert Level**: Configure alert rules and judgment standards.

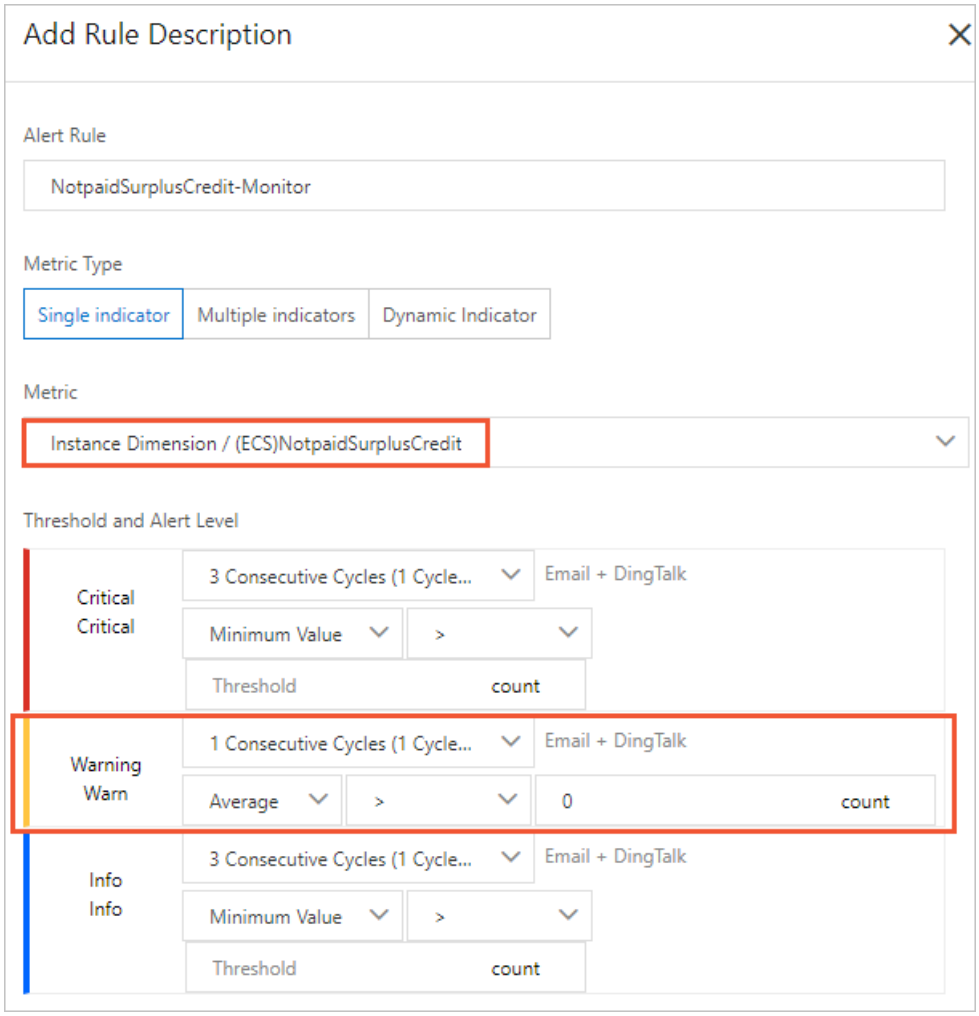
- Monitoring of **TotalCredit**: Choose **Instance Dimension > (ECS)TotaCredit** to monitor accrued CPU credits. In this example, 1Minute cycle, Continue for 1 periods, Average, <, and 1 are used as the values of the alert triggering condition fields. If the average value of **TotalCredit** remains less than 1 for at least 1 minute, an alert is triggered.

> **Note** In standard mode, if the number of accrued CPU credits of a burstable is less than 1, the CPU utilization of the instance cannot exceed the performance baseline. In unlimited mode, a burstable instance can consume advance or overdrawn CPU credits to run at a CPU utilization higher than the performance baseline. You can also configure the average value for multiple consecutive periods as the alert triggering condition based on your actual requirements on CPU performance.



- Monitoring of **NotpaidSurplusCredit**: Choose **Instance Dimension > (ECS)NotpaidSurplusCredit** to monitor overdrawn CPU credits. In this example, 1Minute cycle, Continue for 1 periods, Average, >, and 0 are used as the values of the alert triggering condition fields. If the average value of **NotpaidSurplusCredit** remains greater than 0 for at least 1 minute, an alert is triggered.

  > **Note** If the number of overdrawn CPU credits is greater than 0, overdrawn CPU credits are being used and billed. You can also configure the average value for multiple consecutive periods as the alert triggering condition based on your actual requirements on billing of overdrawn CPU credits.



- Monitoring of **BurstCredit**: Choose **Instance Dimension > (ECS)BurstCredit** to monitor consumed CPU credits.
- Monitoring of **AdvanceCredit**: Choose **Instance Dimension > (ECS)AdvanceCredit** to monitor advance CPU credits.

iii. **Mute for**: Select the interval at which notifications are pushed.

iv. **Effective Time**: Select the time range during which you can receive notifications.

v. **Alert Contact Group**: Select a contact group to receive notifications.

vi. Configure parameters in the **Advanced Settings** section.

- Set **WebHook And Alert Trigger**.
  - **Alert Callback**: The callback URL that can be accessed over the Internet. CloudMonitor sends a POST request to push an alert notification to the specified callback URL.
  - **Auto Scaling**: If you turn on **Auto Scaling** and an alert is triggered, the specified scaling rule is enabled.
  - **Log Service**: If you turn on **Log Service** and an alert is triggered, the alert information is written to the specified Logstore in Log Service.
  - **Message Service - topic**: If you turn on **Message Service - topic** and an alert is triggered, the alert information is written to the specified Message Service (MNS) topic.
- **No data alarm processing method**: The method that is used to handle alerts when no monitoring data is found. Default value: **Do not do anything**.

vii. (Optional)Set **Tags**.

Alert labels are added to the alert content.

6. Click **OK**.

Related information

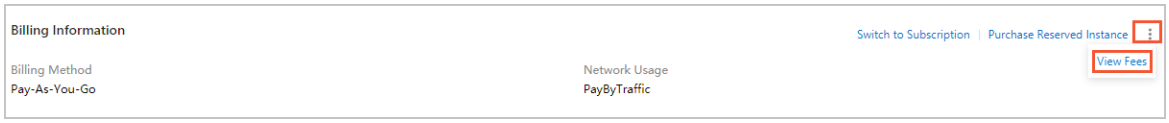
- Create an alert rule
- Manage the monitoring charts of a custom dashboard
- Switch the performance mode of a burstable instance

4.12.8. View bills of a burstable instance

After you purchase a burstable instance, additional costs are charged to your account when you enable the unlimited mode and use overdrawn CPU credits. This topic describes how to query whether your burstable instance incurs additional fees.

Procedure

1.
2.
3. Find the burstable instance that you want to query and click the instance ID.
4. In the **Billing Information** section, choose > **View Fees**.



5. On the **Details** tab, view the bill of the instance.

4.13. Shared instance families

This topic describes the features of shared instance families and lists the instance types of each instance family.

Previous-generation shared instance families xn4, n4, mn4, and e4

Shared instances use a CPU-unbound scheduling scheme. Each vCPU is randomly allocated to an idle CPU hyperthread. vCPUs of different instances compete for CPU resources, which causes computing performance to fluctuate when traffic loads are heavy. Shared instances can guarantee availability but cannot guarantee the performance that may be required in the service level agreement (SLA). Different from enterprise-level instances that have exclusive resources, shared instances share resources. Therefore, shared instances cannot ensure consistent computing performance but offer lower costs.

> **Note** Burstable instances are also shared instances. For more information, see Overview.

Previous-generation shared instance families xn4, n4, mn4, and e4

Features

- Offers multiple CPU-to-memory ratios.
- Uses 2.5 GHz Intel® Xeon® processors.
- Uses DDR4 memory.

Instance family	Description	vCPU-to-memory ratio	Scenario
xn4	Shared compact instance family	1:1	• Frontend web applications • Lightweight applications and microservices • Development and testing environments
n4	Shared compute instance family	1:2	• Websites and web applications • Development environments, servers, code repositories, microservices, and testing and staging environments • Lightweight enterprise applications
mn4	Shared general-purpose instance family	1:4	• Websites and web applications • Lightweight databases and caches • Integrated applications and lightweight enterprise services

Instance family	Description	vCPU-to-memory ratio	Scenario
e4	Shared memory instance family	1:8	• Applications that require a large memory • Lightweight databases and caches

xn4

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.xn4.small	1	1.0	0.5	50,000	1	2	2

Note

- Secondary elastic network interfaces (ENIs) cannot be bound to instances of this instance family while the instances are being created, and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from an ecs.xn4.small instance, the instance must be in the Stopped state.
-
- For more information about these specifications, see Instance family.

n4

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.n4.small	1	2.0	0.5	50,000	1	2	2
ecs.n4.large	2	4.0	0.5	100,000	1	2	2
ecs.n4.xlarge	4	8.0	0.8	150,000	1	2	6
ecs.n4.2xlarge	8	16.0	1.2	300,000	1	2	6
ecs.n4.4xlarge	16	32.0	2.5	400,000	1	2	6

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.n4.8xlarge	32	64.0	5.0	500,000	1	2	6

Note

- Secondary ENIs cannot be bound to instances of this instance family while the instances are being created, and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from an ecs.n4.small or ecs.n4.large instance, the instance must be in the Stopped state.
-
- For more information about these specifications, see Instance family.

mn4

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.mn4.small	1	4.0	0.5	50,000	1	2	2
ecs.mn4.large	2	8.0	0.5	100,000	1	2	2
ecs.mn4.xlarge	4	16.0	0.8	150,000	1	2	6
ecs.mn4.2xlarge	8	32.0	1.2	300,000	1	2	6
ecs.mn4.4xlarge	16	64.0	2.5	400,000	1	2	6
ecs.mn4.8xlarge	32	128.0	5	500,000	2	8	6

Note

- Secondary ENIs cannot be bound to instances of this instance family while the instances are being created, and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from an ecs.mn4.small or ecs.mn4.large instance, the instance must be in the Stopped state.
-
- For more information about these specifications, see Instance family.

e4

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.e4.small	1	8.0	0.5	50,000	1	2	2
ecs.e4.large	2	16.0	0.5	100,000	1	2	2
ecs.e4.xlarge	4	32.0	0.8	150,000	1	2	6
ecs.e4.2xlarge	8	64.0	1.2	300,000	1	3	6
ecs.e4.4xlarge	16	128.0	2.5	400,000	1	8	6

Note

- Secondary ENIs cannot be bound to instances of this instance family while the instances are being created, and can be bound after the instances are created. When you bind secondary ENIs to or unbind them from an ecs.e4.small or ecs.e4.large instance, the instance must be in the Stopped state.
-
- For more information about these specifications, see Instance family.

4.14. Retired instance types

This topic describes all retired instance types on the China site (aliyun.com). However, the sn1, sn2, n1, n2, and e3 instance types are still available for purchase on the International site (alibabacloud.com).

- g5se, storage-enhanced instance family
- sn2, general-purpose instance family
- sn1, compute-optimized instance family
- c4, ce4, and cm4, compute-optimized instance families with high clock speeds
- gn4, GPU-accelerated compute-optimized instance family
- ga1, GPU-accelerated compute-optimized instance family
- ebmhfg5, ECS Bare Metal Instance family with high clock speeds
- sccgn6ne, GPU-accelerated compute-optimized SCC instance family
- n1, n2, and e3, shared instance families
- Generation I instance families

Instance type change

If you are using a retired instance type, we recommend that you change it to another instance type that is available for purchase. For more information about the changes between instance types, see Instance families that support instance type changes.

g5se, storage-enhanced instance family

Features:

- g5se instances can be created only on dedicated hosts.

  > **Note** For more information about instances of other instance types that can be created on dedicated hosts, see Dedicated host types.

- A single g5se instance attached with enhanced SSDs (ESSDs) can deliver a random IOPS of up to 1,000,000 and a sequential read and write performance of up to 32 Gbit/s.
- Compute:
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Provides high storage I/O performance based on large computing capacity.

  > **Note**
- Network:
  - Supports IPv6.
- Supported scenarios:
  - I/O-intensive scenarios such as large and medium-sized online transactional processing (OLTP) core databases
  - Large and medium-sized NoSQL databases
  - Search and real-time log analytics
  - Traditional large enterprise-level commercial software such as SAP

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.g5se.large	2	8.0	1.0	300,000	2	2	6	30,000	1.5

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI	Disk IOPS	Disk bandwidth (Gbit/s)
ecs.g5se.xlarge	4	16.0	1.5	500,000	2	3	6	60,000	2
ecs.g5se.2xlarge	8	32.0	2.0	800,000	2	4	8	85,000	3
ecs.g5se.4xlarge	16	64.0	4.0	1,000,000	4	8	10	150,000	5
ecs.g5se.8xlarge	32	128.0	7.0	2,000,000	8	8	10	300,000	10
ecs.g5se.16xlarge	64	256.0	14.0	3,000,000	16	7	10	750,000	25
ecs.g5se.18xlarge	70	336.0	16.0	4,000,000	16	15	10	1,000,000	32

sn2, general-purpose instance family

Features:

- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) or E5-2680 v3 (Haswell) processors for consistent computing performance.

  > **Note** Instances of this instance family may be deployed on different server platforms. If your business requires all instances to be deployed on the same server platform, we recommend that you use the g6, g6e, or g7 instance family instead.

- Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Enterprise-level applications of various types and sizes
  - Small and medium-sized database systems, caches, and search clusters
  - Data analytics and computing

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs
ecs.sn2.medium	2	8.0	0.5	100,000	1	2
ecs.sn2.large	4	16.0	0.8	200,000	1	3
ecs.sn2.xlarge	8	32.0	1.5	400,000	1	4
ecs.sn2.3xlarge	16	64.0	3.0	500,000	2	8
ecs.sn2.7xlarge	32	128.0	6.0	800,000	3	8
ecs.sn2.13xlarge	56	224.0	10.0	1,200,000	4	8

sn1, compute-optimized instance family

Features:

- Offers a CPU-to-memory ratio of 1:2.
- Uses 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) or E5-2680 v3 (Haswell) processors for consistent computing performance.

  > **Note** Instances of this instance family may be deployed on different server platforms. If your business requires all instances to be deployed on the same server platform, we recommend that you use the c6, c6e, or c7 instance family instead.

- Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Web frontend servers
  - Frontend servers of massively multiplayer online (MMO) games
  - Data analytics, batch processing, and video encoding
  - High-performance scientific and engineering applications

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs
ecs.sn1.medium	2	4.0	0.5	100,000	1	2
ecs.sn1.large	4	8.0	0.8	200,000	1	3

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs
ecs.sn1.xlarge	8	16.0	1.5	400,000	1	4
ecs.sn1.3xlarge	16	32.0	3.0	500,000	2	8
ecs.sn1.7xlarge	32	64.0	6.0	800,000	3	8

c4, ce4, and cm4, compute-optimized instance families with high clock speeds

Features:

- Uses 3.2 GHz Intel Xeon E5-2667 v4 (Broadwell) processors.
- Provides consistent computing performance.
- Is an instance family in which all instances are I/O optimized.
- Supports only standard SSDs and ultra disks.
- Provides high network performance based on large computing capacity.
- Supported scenarios:
  - High-performance web frontend servers
  - High-performance scientific and engineering applications
  - MMO gaming and video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs
ecs.c4.xlarge	4	8.0	1.5	200,000	1	3
ecs.c4.2xlarge	8	16.0	3.0	400,000	1	4
ecs.c4.3xlarge	12	24.0	4.5	600,000	2	6
ecs.c4.4xlarge	16	32.0	6.0	800,000	2	8

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs
ecs.ce4.xlarge	4	32.0	1.5	200,000	1	3
ecs.ce4.2xlarge	8	64.0	3.0	400,000	1	3

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs
ecs.cm4.xlarge	4	16.0	1.5	200,000	1	3
ecs.cm4.2xlarge	8	32.0	3.0	400,000	1	4
ecs.cm4.3xlarge	12	48.0	4.5	600,000	2	6
ecs.cm4.4xlarge	16	64.0	6.0	800,000	2	8
ecs.cm4.6xlarge	24	96.0	10.0	1,200,000	4	8

gn4, GPU-accelerated compute-optimized instance family

Features:

- Uses NVIDIA M40 GPUs.
- Compute:
  - Offers multiple CPU-to-memory ratios.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:
  - Provides high network performance based on large computing capacity.
- Supported scenarios:
  - Deep learning
  - Scientific computing applications, such as computational fluid dynamics, computational finance, genomics, and environmental analysis

- Server-side GPU compute workloads such as high-performance computing, rendering, and multi-media encoding and decoding

Instance types

Instance type	vCPUs	Memory (GiB)	GPU	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.gn4-c4g1.xlarge	4	30.0	NVIDIA M40 * 1	12GB * 1	3.0	300,000	1	3	10
ecs.gn4-c8g1.2xlarge	8	30.0	NVIDIA M40 * 1	12GB * 1	3.0	400,000	1	4	10
ecs.gn4.8xlarge	32	48.0	NVIDIA M40 * 1	12GB * 1	6.0	800,000	3	8	20
ecs.gn4-c4g1.2xlarge	8	60.0	NVIDIA M40 * 2	12GB * 2	5.0	500,000	1	4	10
ecs.gn4-c8g1.4xlarge	16	60.0	NVIDIA M40 * 2	12GB * 2	5.0	500,000	1	8	20
ecs.gn4.14xlarge	56	96.0	NVIDIA M40 * 2	12GB * 2	10.0	1,200,000	4	8	20

ga1, GPU-accelerated compute-optimized instance family

Features:

- Uses AMD S7150 GPUs.
- Supports high-performance local Non-Volatile Memory Express (NVMe) SSDs.
- Compute:
  - Offers a CPU-to-memory ratio of 1:2.5.
  - Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports only standard SSDs and ultra disks.
- Network:

- Provides high network performance based on large computing capacity.

- Supported scenarios:
  - Rendering and multi-media encoding and decoding
  - Machine learning, high-performance computing, and high-performance databases
  - Server-side workloads that require powerful parallel floating-point computing capacity

Instance types

Instance type	vCPUs	Memory (GiB)	Local storage (GiB)	GPU	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	NIC queues	ENIs	Private IP addresses per ENI
ecs.ga1.xlarge	4	10.0	1 * 87	AMD S7150 * 1/4	8GB * 1/4	1.0	200,000	1	3	10
ecs.ga1.2xlarge	8	20.0	1 * 175	AMD S7150 * 1/2	8GB * 1/2	1.5	300,000	1	4	10
ecs.ga1.4xlarge	16	40.0	1 * 350	AMD S7150 * 1	8GB * 1	3.0	500,000	2	8	20
ecs.ga1.8xlarge	32	80.0	1 * 700	AMD S7150 * 2	8GB * 2	6.0	800,000	3	8	20
ecs.ga1.14xlarge	56	160.0	1 * 1400	AMD S7150 * 4	8GB * 4	10.0	1,200,000	4	8	20

ebmhfg5, ECS Bare Metal Instance family with high clock speeds

Features:

- Provides dedicated hardware resources and physical isolation.
- Supports encrypted computing based on Intel® SGX.
- Has failover disabled by default.

  You can call the ModifyInstanceMaintenanceAttributes operation to modify the maintenance action. Set ActionOnMaintenance to *AutoRedeploy* to enable failover.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 3.7 GHz Intel® Xeon® E3-1240v6 (Skylake) processors that deliver a turbo frequency of 4.1 GHz.
- Is an instance family in which all instances are I/O optimized.
- Supports only standard SSDs and ultra disks.
- Supports only virtual private clouds (VPCs).

- Provides high network performance with a packet forwarding rate of 2,000,000 pps.
- Supported scenarios:
  - Workloads that require direct access to physical resources or that require a license to be bound to the hardware
  - Gaming and finance applications that require high performance
  - High-performance web servers
  - Enterprise-level applications such as high-performance databases

Instance types

Instance type	vCPUs	Memory (GiB)	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	ENIs	Private IP addresses per ENI
ecs.ebmhfg5.2xlarge	8	32	6	2,000,000	6	8

sccgn6ne, GPU-accelerated compute-optimized SCC instance family

Features:

- Provides all features of ECS Bare Metal Instance.
- Compute:
  - Uses NVIDIA V100 GPUs (SXM2-based) that have the following features:
    - Innovative Volta architecture
    - 32 GB HBM2 GPU memory
    - CUDA Cores 5120
    - Tensor Cores 640
    - GPU memory bandwidth of up to 900 GB/s
    - Support for six NVLink links and a total bandwidth of 300 GB/s (25 GB/s per NVlink link per direction)
  - Offers a CPU-to-memory ratio of 1:4.
  - Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors for consistent computing performance.
- Storage:
  - Is an instance family in which all instances are I/O optimized.
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Supports high-performance Cloud Paralleled File System (CPFS).
- Network:
  - Supports IPv6.
  - Supports VPCs.
  - Supports RoCE v2 networks, which are dedicated to low-latency RDMA communication.
- Supported scenarios:
  - Ultra-large-scale training for machine learning on a distributed GPU cluster

- Large-scale high-performance scientific computing and simulation calculation
- Large-scale data analytics, batch processing, and video encoding

Instance types

Instance type	vCPUs	Memory (GiB)	GPU	GPU memory	Bandwidth (Gbit/s)	Packet forwarding rate (pps)	RoCE bandwidth (Gbit/s)	NIC queues	ENIs	Private IP addresses per ENI
ecs.scgn6ne.24xlarge	96	768.0	NVIDIA V100 * 8	32GB * 8	32.0	4,800,000	100	16	8	20

n1, n2, and e3, shared instance families

Features

- Uses 2.5 GHz Intel Xeon E5-2680 v3 (Haswell) processors.
- Is an instance family in which all instances are I/O optimized.
- Supports standard SSDs and ultra disks.
- Provides high network performance based on large computing capacity.

Instance family	Description	CPU-to-memory ratio	Supported scenario
n1	Shared compute instance family	1:2	• Small and medium-sized web servers • Batch processing • Distributed analysis • Advertisement services
n2	Shared general-purpose instance family	1:4	• Medium-sized web servers • Batch processing • Distributed analysis • Advertisement services • Hadoop clusters

Instance family	Description	CPU-to-memory ratio	Supported scenario
e3	Shared memory instance family	1:8	• Cache/Redis • Search applications • In-memory databases • Databases with high I/O requirements, such as Oracle and MongoDB • Hadoop clusters • Large-volume data processing

Instance types

Instance type	vCPU	Memory (GiB)	ENIs
ecs.n1.tiny	1	1.0	1
ecs.n1.small	1	2.0	1
ecs.n1.medium	2	4.0	1
ecs.n1.large	4	8.0	2
ecs.n1.xlarge	8	16.0	2
ecs.n1.3xlarge	16	32.0	2
ecs.n1.7xlarge	32	64.0	2

Instance types

Instance type	vCPU	Memory (GiB)	ENIs
ecs.n2.small	1	4.0	1
ecs.n2.medium	2	8.0	1
ecs.n2.large	4	16.0	2
ecs.n2.xlarge	8	32.0	2
ecs.n2.3xlarge	16	64.0	2
ecs.n2.7xlarge	32	128.0	2

Instance types

Instance type	vCPU	Memory (GiB)	ENIs
ecs.e3.small	1	8.0	1
ecs.e3.medium	2	16.0	1
ecs.e3.large	4	32.0	2
ecs.e3.xlarge	8	64.0	2
ecs.e3.3xlarge	16	128.0	2

Generation I instance families

Generation I instance families include t1, s1, s2, s3, m1, m2, c1, and c2. All these instance families are legacy shared instance families. They are categorized based on the number of cores such as 1, 2, 4, 8, or 16 cores.

Features:

- Uses Intel Xeon E5-2420 processors with clock speeds of no less than 1.9 GHz.
- Uses the latest DDR3 memory.
- Is an instance family in which instances can be I/O optimized or non-I/O optimized.

I/O optimized instance types support standard SSDs and ultra disks. The following table describes the instance types and their specifications.

Category	Instance type	vCPU	Memory (GiB)
Standard	ecs.s2.large	2	4
	ecs.s2.xlarge	2	8
	ecs.s2.2xlarge	2	16
	ecs.s3.medium	4	4
	ecs.s3.large	4	8
High Memory	ecs.m1.medium	4	16
	ecs.m2.medium	4	32
	ecs.m1.xlarge	8	32
High CPU	ecs.c1.small	8	8
	ecs.c1.large	8	16
	ecs.c2.medium	16	16
	ecs.c2.large	16	32
	ecs.c2.xlarge	16	64

Non-I/O optimized instance types support only basic disks. The following table describes the instance types and their specifications.

Category	Instance type	vCPU	Memory (GiB)
Tiny	ecs.t1.small	1	1
Standard	ecs.s1.small	1	2
	ecs.s1.medium	1	4
	ecs.s1.large	1	8
	ecs.s2.small	2	2
	ecs.s2.large	2	4
	ecs.s2.xlarge	2	8
	ecs.s2.2xlarge	2	16
	ecs.s3.medium	4	4
	ecs.s3.large	4	8
High Memory	ecs.m1.medium	4	16
	ecs.m2.medium	4	32
	ecs.m1.xlarge	8	32
High CPU	ecs.c1.small	8	8
	ecs.c1.large	8	16
	ecs.c2.medium	16	16
	ecs.c2.large	16	32
	ecs.c2.xlarge	16	64

Instance type specifications

Specification	Description
Local storage	Local storage (also called local disks or cache disks) refers to the disks attached to the physical servers on which ECS instances are hosted. Local storage provides temporary block storage for instances. Local storage capacity is measured in GiB. When the computing resources (vCPUs and memory) of an instance are released or when an instance is failed over, data stored on its local disks may be lost. For more information, see Local disks.

Specification	Description
Bandwidth	The maximum sum of inbound and outbound bandwidth values. For information about how to test the network bandwidth of an instance, see Test the network bandwidth. **Note** Instance type specifications are all verified and obtained within a test environment. In actual scenarios, the performance of an instance may vary based on other factors such as instance load and networking model. We recommend that you perform business stress tests on instances to choose appropriate instance types.
Packet forwarding rate	The maximum sum of inbound and outbound packet forwarding rates. For information about how to test the packet forwarding rate of an instance, see Best practices for testing network performance. **Note** Instance type specifications are all verified and obtained within a test environment. In actual scenarios, the performance of an instance may vary based on other factors such as instance load, image version, and networking model. We recommend that you perform business stress tests on instances to choose appropriate instance types.
Connections	A connection (also called a session) is the process of connecting a client and a server and transferring data between them. A connection is uniquely defined by the network communication quintuple that consists of a source IP address, a destination IP address, a source port, a destination port, and a protocol. Connections of an ECS instance include TCP, UDP, and Internet Control Message Protocol (ICMP) connections.
NIC queues	The maximum number of network interface controller (NIC) queues supported by the primary NIC of an instance. For the instance types other than ECS Bare Metal Instance types, the maximum number of NIC queues supported by a secondary NIC is the same as that supported by the primary NIC.
ENI	The number of elastic network interfaces (ENIs) per instance that include one primary ENI.

4.15. Instance families that do not support advanced VPC features

This topic describes instance families that do not support advanced VPC features.

Advanced VPC features include network ACLs, route tables, and flow logs. For more information, see Overview of network ACLs, Route table overview, and Overview of the flow log feature.

Limits of advanced VPC features on ECS instances:

- Instance creation: If advanced features are enabled on a VPC, you cannot create instances of instance families listed in the following table within the VPC.
- Instance upgrade or downgrade: If the instance type of an ECS instance supports advanced features and those advanced features are enabled on the corresponding VPC, the instance cannot be changed to an instance that belongs to instance families listed in the following table. The methods to change instance configurations include upgrading, downgrading, renewal and downgrading, and changing instance types.

When you enable advanced features on a VPC, the system automatically detects the instances within the VPC that must be upgraded or downgraded. Then, you can upgrade or downgrade configurations of these instances. For more information, see Change the instance type of a pay-as-you-go instance or Upgrade the instance types of subscription instances.

The following table describes instance families that do not support advanced VPC features.

Instance family type	Instance family
General purpose instance families	sn2 (retired)
Compute optimized instance families	sn1 (retired)
Memory optimized instance families	se1
Big data instance families	d1
Instance families with local SSDs	i1
Instance families with high clock speed	• c4 (retired) • ce4 (retired) • cm4 (retired)
GPU-accelerated compute optimized instance families	• gn4 • gn5
GPU-accelerated instance families with graphics acceleration capabilities	ga1 (retired)

Instance family type	Instance family
Shared instance families	• n1 (retired) • n2 (retired) • e3 (retired) • xn4 • n4 • mn4 • e4
Generation I instance families	• t1 (retired) • s1 (retired) • s2 (retired) • s3 (retired) • m1 (retired) • m2 (retired) • c1 (retired) • c2 (retired)

5.Instance purchasing options

5.1. Subscription

Subscription is a billing method that allows you to pay in advance for the use of resources. This billing method allows you to reserve resources at discounted rates and reduce costs. This topic describes the billing rules for subscription Elastic Compute Service (ECS) resources.

Overview

Before you can use subscription resources, you must create a subscription instance. When you create a subscription instance on the Custom Launch or Quick Launch tab in the ECS console, supported subscription durations are displayed on the instance buy page.

When you create a subscription instance, subscription resources are billed separately to generate a total price. Subscription resources can be used only after the total price has been paid. For more information about how prices are calculated, see Billing.

After a subscription instance is created, you can change its configurations or extend the subscription cloud disks that are attached to the instance. For more information, see Overview of instance configuration changes and Overview.

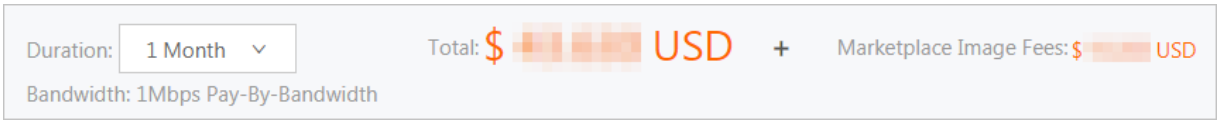
After a subscription instance expires, you can renew the instance to continue to use it. For more information, see Renewal overview.

Applicable resources

The combination of computing resources (vCPUs and memory), Elastic Block Storage (EBS) devices, an image, and a network type uniquely determines the service form of an ECS instance. The following table describes the ECS resources that support the subscription billing method.

Resource	Description
Computing resource (vCPUs and memory)	When you create an instance, you must specify whether to use the subscription billing method.
Image	The image that you select when you create a subscription instance also uses the subscription billing method.
Cloud disk	Cloud disks created along with a subscription instance also use the subscription billing method. After a subscription instance is created, you can create subscription disks for the instance or attach pay-as-you-go disks that were separately created to the instance. For more information, see Create a subscription disk and Attach a data disk.
Public bandwidth (pay-by-bandwidth)	If you select pay-by-bandwidth as the billing method for network usage when you create a subscription instance, the amount of public bandwidth allocated to the instance is also billed on a subscription basis. For more information, see Public bandwidth.

You can view the total price of the preceding resources in the lower part of the instance buy page in the ECS console, as shown in the following figure.



- **Total** indicates the total price of the following resources:
  - Computing resources (vCPUs and memory)
  - Cloud disks
  - Public bandwidth (pay-by-bandwidth)

    > **Note** If you select pay-by-traffic as the billing method for network usage, the total price does not include the price of public bandwidth. For more information, see Public bandwidth.

- **Marketplace Image Fees**: If you selected a paid Alibaba Cloud Marketplace image, Marketplace Image Fees are displayed.

You can use the ECS TCO Calculator to analyze your cloud migration costs.

Billing

The billing cycle of a subscription instance is the time commitment you made when you purchased the instance (based on UTC+8:00). The billing cycle begins at the time when the subscription instance is purchased or renewed (accurate to the second) and ends at 00:00:00 on the day after the expiration date.

> **Note** The billing cycles of subscription resources are in the unit of calendar years or months.

For example, assume that you created a one-month subscription instance at 13:00:00 on August 9, 2019 and that the instance had computing resources (vCPUs and memory), an image, and a cloud disk (system disk). Assume that you manually renewed the instance for another one month. The following billing cycles apply:

- The first billing cycle is from 13:00:00 on August 9, 2019 to 00:00:00 on September 10, 2019.
- The second billing cycle is from 00:00:00 on September 10, 2019 to 00:00:00 on October 10, 2019.

Resources are billed separately. You must pay for the resources before you can use them. You can calculate the total price for each billing cycle based on your selected configurations. The following table describes the formulas used to calculate the fee of each resource.

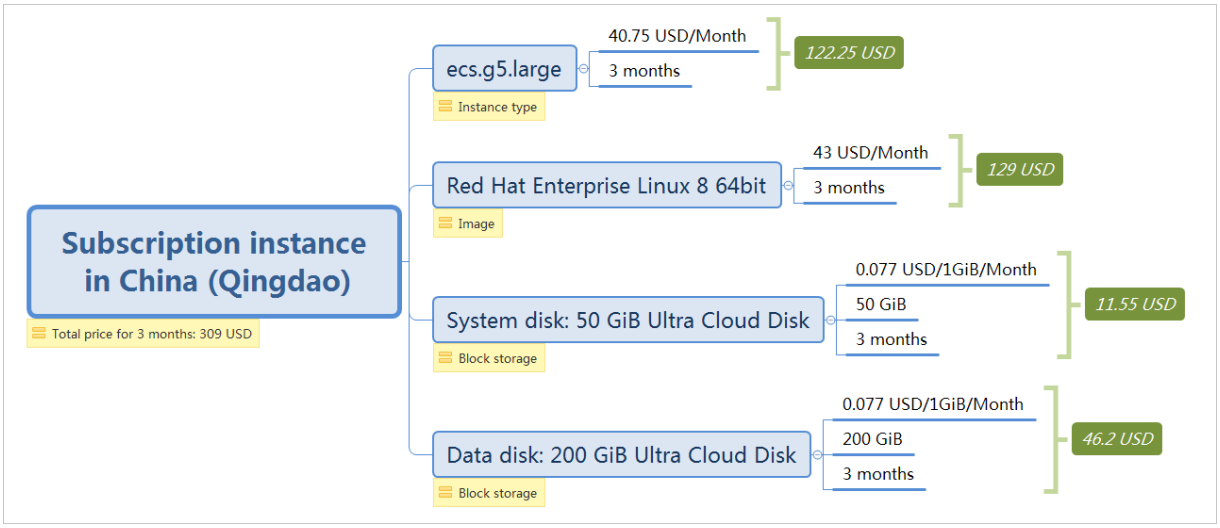
Resource	Formula	Unit price

Resource	Formula	Unit price
Computing resource (vCPUs and memory)	Unit price of an instance type × Subscription duration	For more information, see the **Instance** section on the Pricing tab of the Elastic Compute Service product page. **Note** Local disks are tied to specific instance types. The prices of local disks are included in the prices of corresponding instance types.
Image	Unit price of an image × Subscription duration	You can view the price on the instance buy page in the ECS console or in Alibaba Cloud Marketplace.
Cloud disk (system disk)	Unit price of a disk × Disk capacity × Subscription duration	For more information, see **System Cloud Disk Fee** in the **Storage** section on the Pricing tab of the Elastic Compute Service product page. **Note** The prices for pay-as-you-go disks displayed on the page are prices in USD per 100 GiB-hour, divided by 100 to obtain the unit prices.
Cloud disk (data disk)	Unit price of a disk × Disk capacity × Subscription duration	For more information, see **Data Cloud Disk Fee** in the **Storage** section on the Pricing tab of the Elastic Compute Service product page. **Note** The prices for pay-as-you-go disks displayed on the page are prices in USD per 100 GiB-hour, divided by 100 to obtain the unit prices.

Resource	Formula	Unit price
Public bandwidth (pay-by-bandwidth)	Unit price of bandwidth × Bandwidth value × Subscription duration For more information, see Public bandwidth.	A tiered billing model is used for bandwidth. You can select a bandwidth value on the instance buy page to view the fee schedule.

For example, assume that you created a three-month subscription instance in the China (Qingdao) region. The following figure shows the process of calculating the instance price.

> **Note** The prices in the following figure are for reference only. For more information about actual prices, visit the URLs in the preceding table.



Changes in resource states after a subscription instance expires

> **Notice** After a subscription instance expires, it may be stopped. The system sends you notifications for renewing the instance. Renew your instance at your earliest convenience to ensure service availability. If you have other questions, submit a ticket.

If the auto-renewal feature is not enabled for a subscription instance, the instance stops providing services at some point from 00:00:00 on the expiration date to 00:00:00 the next day.

> **Note** You cannot enable the auto-renewal feature for an expired subscription instance.

The following table describes the resource states for an expired subscription instance.

Resource	Within 15 days after the instance expires	More than 15 days after the instance expires

Resource	Within 15 days after the instance expires	More than 15 days after the instance expires
Computing resource (vCPUs and memory)	The computing resources (vCPUs and memory) are retained, but the instance stops providing services. **Note** After an instance is stopped, you cannot connect to the instance or access websites deployed on the instance, and service errors may occur.	The computing resources (vCPUs and memory) are released.
Image	The image is unavailable.	The image is unavailable.
EBS device	• Cloud disks and their data are retained, but the cloud disks cannot be used. • Local disks and their data are retained, but the local disks cannot be used.	• Subscription disks are released and their data cannot be recovered. **Note** If you have manually attached pay-as-you-go cloud disks to the instance and have not set the release mode of the disks to Release with Instance, these disks stop working. • Local disks are released and their data cannot be recovered.

Public IP address	• If the instance is located in the classic network, its public IP address is retained.
	• If the instance is located in a virtual private cloud (VPC), the following rules apply:
	  ◦ The public IP address is retained.
	  ◦ The elastic IP address (EIP) associated with the instance remains unchanged.
		• If the instance is located in the classic network, its public IP address is released.
		• If the instance is located in a VPC, the following rules apply:
		  ◦ The public IP address is released.
		  ◦ The EIP is disassociated from the instance.

If the auto-renewal feature is enabled for a subscription instance but the instance fails to renew, the instance stops providing services at some point from 00:00:00 on the 15th day after it expires to 00:00:00 on the 16th day after it expires.

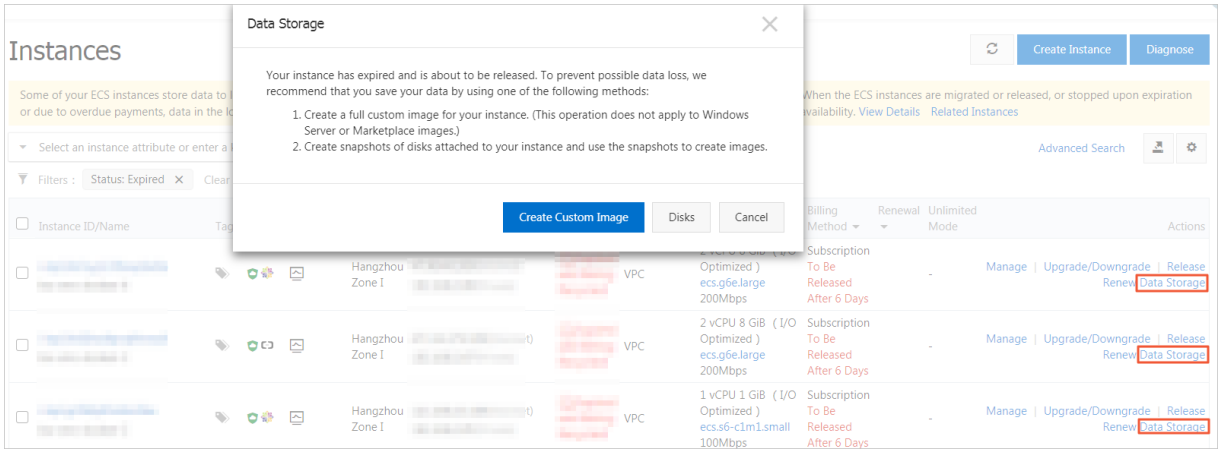
The following table describes the resource states for an expired subscription instance.

Resource	Within 15 days after the instance expires	16 to 30 days after the instance expires	More than 30 days after the instance expires
Computing resource (vCPUs and memory)	The computing resources (vCPUs and memory) are retained, and the instance works normally. **Note** When an instance works normally, you can start or stop the instance and connect to the instance by using management terminals or other connection methods.	The computing resources (vCPUs and memory) are retained, but the instance stops providing services. **Note** After an instance is stopped, you cannot connect to the instance or access websites deployed on the instance, and service errors may occur.	The computing resources (vCPUs and memory) are released.
Image	The image is available.	The image is unavailable.	The image is unavailable.
EBS device	• Cloud disks and their data are retained. The cloud disks can work normally. • Local disks and their data are retained. The local disks can work normally.	• Cloud disks and their data are retained, but the cloud disks cannot be used. • Local disks and their data are retained, but the local disks cannot be used.	• Subscription disks are released and their data cannot be recovered. **Note** If you have manually attached pay-as-you-go cloud disks to the instance and have not set the release mode of the disks to Release with Instance, these disks stop working. • Local disks are released and their data cannot be recovered.

Resource	Within 15 days after the instance expires	16 to 30 days after the instance expires	More than 30 days after the instance expires

Public IP address	• If the instance is located in the classic network, its public IP address is retained.
	• If the instance is located in a VPC, the following rules apply:
	  ○ The public IP address is retained.
	  ○ The EIP associated with the instance remains unchanged.
		• If the instance is located in the classic network, its public IP address is retained.
		• If the instance is located in a VPC, the following rules apply:
		  ○ The public IP address is retained.
		  ○ The EIP associated with the instance remains unchanged.
			• If the instance is located in the classic network, its public IP address is released.
			• If the instance is located in a VPC, the following rules apply:
			  ○ The public IP address is released.
			  ○ The EIP is disassociated from the instance.

After an instance expires, **Data Storage** is displayed in the **Actions** column corresponding to the instance on the **Instances** page. Before the instance is released, you can create a custom image from the instance or create snapshots to back up disk data.



Overdue payments

If you have overdue payments within your account, you can use your subscription resources normally but cannot perform operations that incur charges on them, such as purchasing instances, upgrading instance configurations, or renewing resources. For more information, see Overdue payments.

5.2. Pay-as-you-go

Pay-as-you-go is a billing method that allows you to use resources first and pay for them afterward. Pay-as-you-go Elastic Compute Service (ECS) resources can be purchased and released on demand and provide cost savings of 30% to 80% compared with traditional hosts. This topic describes the billing and settlement rules for pay-as-you-go ECS resources.

Overview

You are billed for pay-as-you-go resources on an hourly basis. If you have a quota agreement with Alibaba Cloud, fees are deducted only when the cumulative consumption of your account exceeds the quota. You must complete the payment at least once a month.

You can change the configurations of your pay-as-you-go resources. For more information, see Change the instance type of a pay-as-you-go instance and Modify the bandwidth configurations of pay-as-you-go instances.

You can change the billing methods of your pay-as-you-go resources. For more information, see Change the billing method of an ECS instance from pay-as-you-go to subscription.

You can use one of the following methods to view the consumption details of your pay-as-you-go ECS resources:

- For information about how fees are calculated, see the Billing section of this topic.
- For information about how the resource state affects the billing duration, see Billing duration section of this topic.

  > **Note** If you stop an instance but do not release its resources, you continue to be charged for these resources.

- For information about settlement, see the Settlement cycle section of this topic.

Applicable resources

The pay-as-you-go billing method is applicable to the following ECS resources:

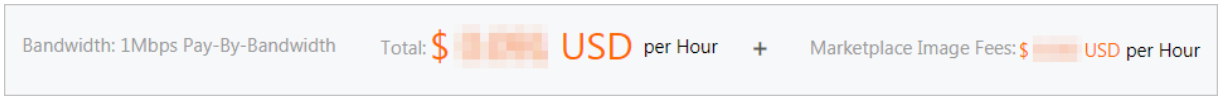
- Computing resources (vCPUs and memory)
- Image
- Cloud disks
- Public bandwidth (pay-by-bandwidth)
- Snapshots

The combination of computing resources (vCPUs and memory), Elastic Block Storage (EBS) devices, an image, and a network type uniquely determines the service form of an ECS instance. When you create a pay-as-you-go instance, the image and disks created along with the instance use the pay-as-you-go billing method. However, you can also use a billing method for network usage.

> **Note** After you create a pay-as-you-go instance, you can attach pay-as-you-go disks that are separately created to the instance. For more information, see Attach a data disk.

After you create a snapshot, you are immediately charged for the snapshot.

You can view the total price of the preceding resources in the lower part of the instance buy page in the ECS console.



- **Total** indicates the total price of the following resources:
  - Computing resources (vCPUs and memory)
  - Cloud disks

- Public bandwidth (pay-by-bandwidth)

  > **Note** If you select pay-by-traffic as the billing method for network usage, the total price does not include the price of public bandwidth. For more information, see Public bandwidth.

- **Marketplace Image Fees**: If you selected a paid Alibaba Cloud Marketplace image, Marketplace Image Fees are displayed.

Billing

You are charged for pay-as-you-go resources based on their billing cycles. The billing cycle of a pay-as-you-go resource immediately begins after the resource is created. You can calculate the total fee that you must pay for a period of time based on the configurations of your choice. The following table describes the billing cycle of each ECS resource and the formula used to calculate the fee of each resource.

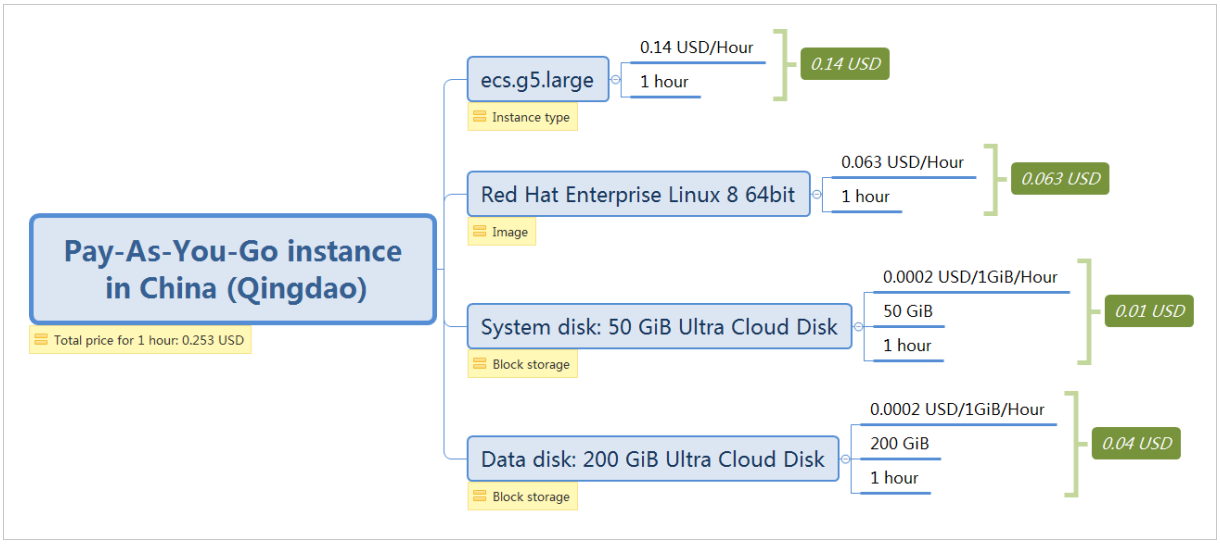
Resource	Billing cycle	Formula	Unit price
Computing resources (vCPUs and memory)	Varies based on the number of the vCPUs of the instance type. • 1 vCPU: 10 minutes, with a minimum of 10 minutes • 2 vCPUs: 5 minutes, with a minimum of 5 minutes • 4 vCPUs: 2 minutes, with a minimum of 2 minutes • More than 4 vCPUs: 1 second	Unit price of an instance type × Billing duration	For more information, see the Instance tab on the ECS Pricing page. **Note** Local disks are tied to specific instance types. The prices of local disks are included in the prices of corresponding instance types.
Image	1 second	Unit price of an image × Billing duration	You can view the price on the instance buy page in the ECS console or in Alibaba Cloud Marketplace.
Cloud disk (system disk)	Varies based on the number of the vCPUs of the instance type. • 1 vCPU: 10 minutes, with a minimum of 10 minutes • 2 vCPUs: 5 minutes, with a minimum of 5 minutes • 4 vCPUs: 2 minutes, with a minimum of 2 minutes • More than 4 vCPUs: 1 second	Unit price of a cloud disk × Disk capacity × Billing duration	For more information, see the Storage tab on the ECS Pricing page. **Note** The prices for pay-as-you-go disks are displayed on the page in the unit of USD per 100 GiB-hour. You can divide the unit price by 100 to obtain the unit price per GiB.

Resource	Billing cycle	Formula	Unit price
Cloud disk (data disk)	1 second	Unit price of a cloud disk × Disk capacity × Billing duration	For more information, see the Storage tab on the ECS Pricing page. **Note** The prices for pay-as-you-go disks are displayed on the page in the unit of USD per 100 GiB-hour. You can divide the unit price by 100 to obtain the unit price per GiB.
Public bandwidth (pay-by-bandwidth)	1 second	Unit price of bandwidth × Bandwidth value × Billing duration For more information, see Public bandwidth.	A tiered billing model is used for bandwidth. You can select a bandwidth value on the instance buy page to view the fee schedule.
Snapshot	1 hour, with a minimum of 1 hour	Unit price of a snapshot × Snapshot capacity × Billing duration For more information, see Snapshots.	For more information, see the Snapshot tab on the ECS Pricing page.

Note If the billing cycle is 1 second, you are charged for resources in 1-second increments. If an hourly price is displayed, you can divide the price by 3600 to obtain the price per second.

For example, you created a pay-as-you-go instance in the China (Qingdao) region and used this instance from 11:00:00 to 12:00:00 on August 8, 2019. The following figure shows the process of calculating the instance price.

Note The prices in the following figure are for reference only. For more information about actual prices, visit the URLs in the preceding table.



Billing duration

If a pay-as-you-go instance is automatically stopped due to overdue payments, the billing of some resources suspends. The billing resumes after you complete the overdue payment and reactivate the instance.

The following table describes the billing duration of each resource type on the premise that you have no overdue payments within your account.

Resource	Billing duration

Computing resources (vCPUs and memory)	The billing duration of computing resources (vCPUs and memory) is affected by the network type of the instance.
	• For an instance of the classic network type, billing begins when the instance is created and ends when the instance is released.
	• For an instance of the virtual private cloud (VPC) type, the billing duration varies based on whether the economical mode is enabled.
	  ◦ If the economical mode is disabled, billing begins when the instance is created and ends when the instance is released.
	  ◦ If the economical mode is enabled, billing begins when the instance is created and suspends when the instance is stopped in the ECS console. Billing resumes when the instance is started again in the ECS console and ends when the instance is released. For more information, see Economical mode.
	Notice If you stop the instance by shutting down its operating system, the economical mode cannot be not triggered.
	You can purchase reserved instances to minimize costs. For more information, see Overview.

Resource	Billing duration
Image	Billing begins when the instance is created and ends when the instance is released.
Cloud disk (system disk)	Billing begins when the instance is created and ends when the system disk is released along with the instance.
Cloud disk (data disk)	Billing begins when the data disk is created and ends when the data disk is released.
Public bandwidth (pay-by-bandwidth)	Billing begins when public bandwidth (pay-by-bandwidth) is enabled and ends when public bandwidth is disabled or when the instance is released. For information about how to disable public bandwidth, see Modify the bandwidth configurations of pay-as-you-go instances.
Snapshots	Billing begins when the snapshot is created and ends when the snapshot is deleted.

Note If a pay-as-you-go instance incurs charges of less than USD 0.01 during its entire lifecycle, you are charged USD 0.01.

To prevent unexpected charges incurred when a pay-as-you-go instance is not released in a timely manner, we recommend that you enable the automatic release feature. If automatic release is enabled, billing ends when the instance is released. The automatic release time is accurate to the second.

Settlement cycle

You are billed for pay-as-you-go ECS resources on an hourly basis. The fees for pay-as-you-go resources are paid together with those for other pay-as-you-go resources within your account. If you have a quota agreement with Alibaba Cloud, fees are deducted only when the cumulative consumption of your account exceeds the quota. If the cumulative monthly consumption of your account is less than the quota, fees are deducted on the first day of the next month.

- If your default payment method is bank card, the quota is USD 1,000.
- If your default payment method is PayPal or Paytm (India), the quota varies based on your ECS usage.

The system attempts to deduct fees three times: on the due date (T), day T+7, and day T+14. If the fees of a pay-as-you-go instance cannot be deducted on the due date (T), payment becomes overdue for the instance and the system attempts to deduct the fees again on day T+7 and day T+14. If all three attempts to deduct the fees fail, the instance is stopped on day T+15 and its billing also stops. For more information, see the "Pay-as-you-go resources" section in Pay-as-you-go resources

Overdue payments

When you have overdue payments within your account, pay-as-you-go resources cannot be used. Your pay-as-you-go instances may be stopped and the resources may be released. To prevent the consequences of overdue payments, such as instance stop or release from affecting your business, add funds to your account to complete the overdue payments at your earliest convenience. For information about changes in resource states when payments become overdue, see Overdue payments.

5.3. Preemptible instances

5.3.1. Overview

Preemptible instances are a type of on-demand instances that are offered at a discounted price compared to pay-as-you-go instances. Preemptible instances are designed to minimize Elastic Compute Service (ECS) instance costs in specific scenarios.

Introduction

The following table describes the features of preemptible instances.

Feature	Description
Bidding mode	The market price of a preemptible instance fluctuates based on changes in supply and demand for the instance type. When you create a preemptible instance, you must set a maximum hourly price to bid for a specified instance type. If your bid price is higher than the market price at purchase and the stock of the instance type is sufficient, a preemptible instance can be created. After a preemptible instance is created, it can be used in the same manner as a pay-as-you-go instance. You can also use it with other cloud resources such as disks or elastic IP addresses (EIPs).
Protection period	By default, each preemptible instance is created with a protection period of 1 hour. You can also configure a preemptible instance to not have a protection period when you create the instance. • During the protection period of a preemptible instance, the preemptible instance is not released by the system regardless of the fluctuations in market price. You can continue to perform operations on the preemptible instance. • Preemptible instances without a protection period are more cost-effective and have a discount of about 10% off compared with preemptible instances with a protection period.

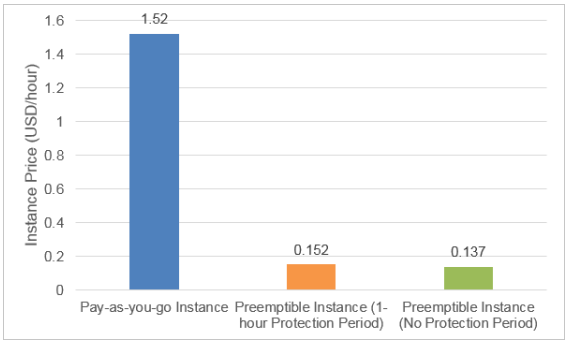
Feature	Description
Recycling	After the protection period ends, the system checks the market price and resource stock of the instance type every 5 minutes. If the market price is higher than your bid price or if the stock of the instance type is insufficient, the preemptible instance is released. **Notice** After an instance is released, its data cannot be recovered. We recommend that you create a snapshot for the instance to back up its data before the instance is released. For more information, see Snapshot overview. Regardless of whether you attempt to create preemptible instances with or without a protection period, you can check the release rate of each instance type on the instance buy page in the ECS console. The release rate of each instance type is determined by the bidding policy and the supply-demand relationships of resources. A lower release rate indicates that preemptible instances are less likely to be recycled.

In this example, the following instances are of the ecs.c5.8xlarge instance type and reside in Silicon Valley Zone B:

- The price of a pay-as-you-go instance is USD 1.52/hour.
- A preemptible instance with a protection period of 1 hour has a discount of 90% off compared with a pay-as-you-go instance of the same type. The price of the preemptible instance is USD 0.152/hour.

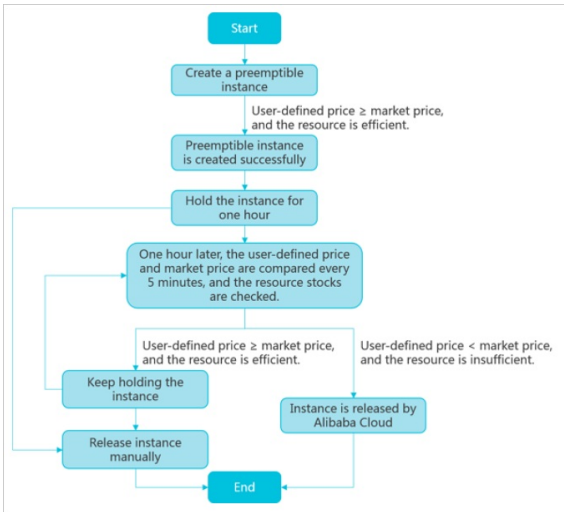
If the protection period feature is disabled when the preemptible instance is created, the price of the instance is discounted an additional 10%. As a result, the price for the instance is 0.152 × 0.9 = USD 0.137/hour.

Note The original price of pay-as-you-go instances and the discount for preemptible instances are subject to changes. Prices in this example are for reference only.



Lifecycle

The following figure shows the lifecycle of a preemptible instance that has a protection period of 1 hour.



After a preemptible instance is created, you can release the instance at any time. When the market price exceeds your bid price or when the stock of the instance type is insufficient, your instance enters the To Be Released state. After 5 minutes, the instance is automatically released. You can check whether the instance enters the To Be Released state based on the instance metadata or the `OperationLocks` value returned by the DescribeInstances operation.

You can check whether your preemptible instance enters the To Be Released state and store a small amount of data while you wait for the release of the instance. However, to make sure that the application runs normally after your preemptible instance is released, we recommend that you optimize the application design. You can manually release your preemptible instance to check whether the application runs normally after the preemptible instance is released.

Typically, the system first releases the instance that has the lowest bid price. If multiple preemptible instances have the same bid price, the system randomly determines the order in which the instances are released.

Limits

- Whether you can create a preemptible instance is determined by your ECS instance resource usage. If the Preemptible Instance option is not available for the Billing Method parameter when you create an instance, you cannot create a preemptible instance.
- Preemptible instances cannot be converted to subscription instances.
- The instance types of preemptible instances cannot be changed.

Scenarios

Preemptible instances are ideal for stateless scenarios such as scalable web services, image rendering, big data analytics, and large-scale parallel computing. Preemptible instances can be applied to applications that require a high level of distribution, scalability, and fault tolerance capabilities. Preemptible instances help reduce costs and increase the throughput of these applications.

You can use preemptible instances in the following scenarios:

- Real-time analytics
- Big data
- Geological surveys

- Image and media coding
- Scientific computing
- Scalable websites and web crawlers
- Tests

Preemptible instances are not suitable for stateful applications such as databases. When a preemptible instance is released due to a failed bid or other reasons, application status data cannot be stored.

Pricing and billing

- Prices

  The price of a preemptible instance covers only the price of the instance type (including vCPUs and memory) and does not include the prices of resources such as the system disk, data disks, and network bandwidth.

  - System disks and data disks are billed on a pay-as-you-go basis. For more information, see Pay-as-you-go.
  - Network bandwidth is billed based on the billing rule for network usage of pay-as-you-go instances. For more information, see Public bandwidth.

- Market prices

  The market price of a preemptible instance fluctuates based on changes in supply and demand for the instance type. If your bid price is higher than the market price at purchase and the stock of the instance type is sufficient, a preemptible instance can be created.

  If the protection period feature is enabled for a preemptible instance, the instance is billed based on the market price at the time of purchase. After 1 hour, the instance is billed based on the real-time market price.

  We recommend that you evaluate the fluctuations in market price to minimize computing costs and increase throughput when you purchase preemptible instances.

- Billing methods

  Preemptible instances are billed by second. The market price of a preemptible instance is an hourly price. You can divide the hourly price by 3,600 to obtain the price per second.

  The fees for a preemptible instance from creation to release are accurate to two decimal places. Accrued fees of less than USD 0.01 are not charged.

- Billing duration

  A preemptible instance is billed based on its usage duration, which lasts from the time when the instance is created to that when it is released. If you stop an instance only by calling the StopInstance operation or by using the ECS console, the instance continues to be billed. When a preemptible instance is no longer needed, we recommend that you create snapshots to back up your data and environment and then release the instance. You can purchase new preemptible instances at any time.

References

- Create a preemptible instance
- Query the interruption events of preemptible instances
- View bills of a preemptible instance
- Stop a preemptible instance

- Release an instance
- Instance limits
- Best practices for preemptible instances

FAQ

- To what resources are the prices of preemptible instances applicable?

  The prices of preemptible instances are applicable only to instance types. Other resources such as disks and network bandwidth are billed at the same prices as those of pay-as-you-go instances.

- Am I notified when my preemptible instances are about to be released?

  Yes, you are notified when your preemptible instances are about to be released. When your preemptible instances are about to be released due to a spot price change or insufficient resources, the instances first enter the To Be Released state and then are automatically released after 5 minutes.

- Can the instance type of a preemptible instance be changed?

  No, the instance types of preemptible instances cannot be changed.

- Which is more cost-effective: a preemptible instance with a protection period or a preemptible instance without a protection period?

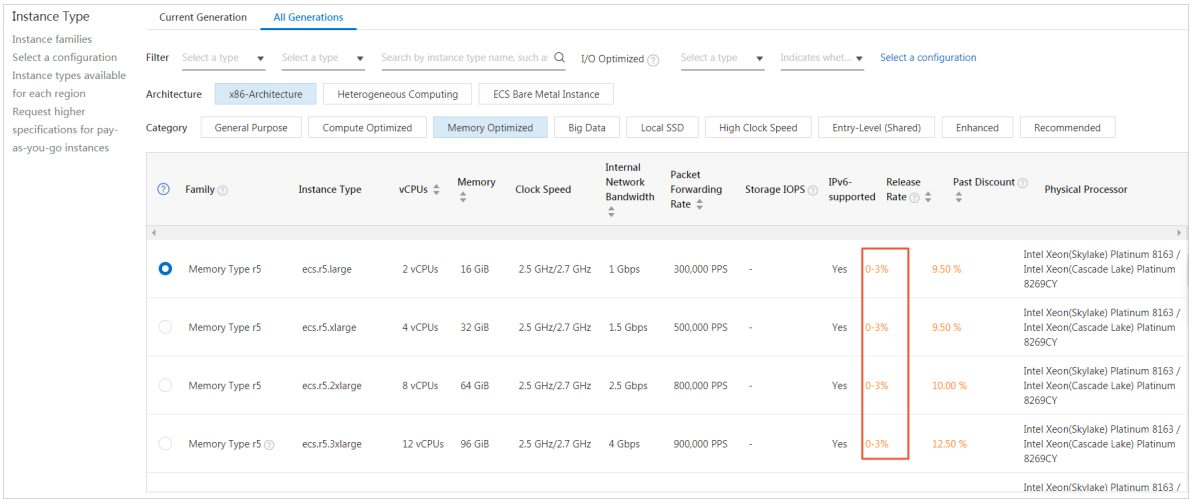
  Preemptible instances without a protection period are more cost-effective and have a discount of about 10% off compared with preemptible instances with a protection period.

- Can I convert between preemptible instances with and without a protection period?

  No, you cannot convert between preemptible instances with and without a protection period. By default, each preemptible instance is created with a protection period of 1 hour. The protection period for a preemptible instance can be configured only when the instance is created.

- Is the release rate of preemptible instances without a protection period higher than that of preemptible instances with a protection period?

  You can check the release rate of each instance type on the instance buy page in the ECS console regardless of whether you attempt to create preemptible instances with or without a protection period. For each instance type, the release rate is determined by the bidding policy and the supply-demand relationships of resources.



For more information about preemptible instances, see Instance FAQ.

5.3.2. Create a preemptible instance

This topic describes how to create a preemptible instance in the ECS console.

Context

To create and use a preemptible instance, take note of the following items:

- Set an appropriate bidding price and take into account the estimated market price fluctuations. An appropriate bidding price can result in a higher probability of creating a preemptible instance and ensure that the instance is not released due to small changes in price. The bidding price must also meet your business expectations.

  > **Note** If you do not know what price to bid for your preemptible instances, we recommend that you use the market price at the time of purchase as the bidding price.

- Use an image that contains the configurations of all the required software to ensure that the instance can be started any time after it is created. You can also use user data of the instance to run commands when you start the instance. For more information, see Overview of ECS instance user data.
- To prevent data loss caused by instance release, store important data in a storage medium that is not affected when preemptible instances are released, such as separately created cloud disks, OSS buckets, or ApsaraDB for RDS instances.
- Break your jobs down into smaller tasks by using grids, Hadoop, or queue-based architecture, or use checkpoints to save calculation results.
- Monitor the status of a preemptible instance by checking the instance release notifications from ECS. ECS updates the instance metadata five minutes before ECS releases a preemptible instance. You can obtain the status of a preemptible instance every minute by checking instance metadata. For more information, see Overview of ECS instance metadata.
- Run your applications on a pay-as-you-go instance and release the instance to verify whether your applications can automatically adjust themselves when the instance is released.

You can use developer tools such as Alibaba Cloud CLI, OpenAPI Explorer, and Alibaba Cloud SDK to call the RunInstances operation and create a preemptible instance.

> **Note** You can set the SpotStrategy parameter to *SpotAsPriceGo* to use the market price at the time of purchase. Alternatively, you can set the SpotStrategy parameter to *SpotWithPriceLimit* to use the acceptable maximum price.

This topic describes configurations to create a preemptible instance. For more information about other configurations for creating an instance, see Create an instance by using the wizard.

Procedure

1.
2.
3. On the **Instances** page, click **Create Instance**.
4. Set **Billing Method** to **Preemptible Instance**.
5. In the **Maximum Price for Instance Type** section, specify your bidding price.

The preemptible instance you request is created at the market price only if your bidding price is higher than or equal to the market price and resources are sufficient. You can bid for a preemptible instance only once. The following bidding modes are supported:

- **Use Automatic Bid**: The market price at the time of purchase is used as the bidding price.
- **Set Maximum Price**: You must set the highest price you are willing to pay for the instance type.

> **Note** In the displayed price range, the maximum price is equal to the price for the pay-as-you-go instance of the same instance type. Your bidding price must be based on the displayed price range, your business requirements, and the estimated price fluctuations. If you set your bidding price while taking the estimated price fluctuations into consideration, you can retain a preemptible instance for a long period of time. Otherwise, the instance may be released at any time after the protection period ends.

6. Select or enter the quantity of instances that you want to purchase.
7. Complete other settings.
8. Confirm the order information and click **Create Instance**.

Result

After a preemptible instance is created, you can view its information in the instance list. A preemptible instance is marked as **Pay-As-You-Go Preemptible Instance** in the Billing Method column. Click the instance ID to go to the Instance Details page. In the **Payment Information** section, you can view the bidding policy configured when you create the instance.

Related information

- RunInstances

5.3.3. Query the interruption events of preemptible instances

If a preemptible instance is forcibly recycled due to a market price change or insufficient resources, an interruption event is triggered. Before the preemptible instance is recycled, it enters the locked state, and you are prompted that the instance will be recycled. This topic describes how to query the interruption events of preemptible instances. You can automate the instance interruption and recycle processes based on the recycle status of instances.

Query the interruption events of preemptible instances by using CloudMonitor SDK

This section describes how to use CloudMonitor SDK for Java to query the interruption events of preemptible instances.

1. Access CloudMonitor SDK for Java.

   For more information, see CloudMonitor SDK for Java.
2. Query the interruption events of preemptible instances by using the SDK.

   The following section shows the sample code:

```
import com.aliyuncs.DefaultAcsClient;
import com.aliyuncs.IAcsClient;
import com.aliyuncs.exceptions.ClientException;
import com.aliyuncs.exceptions.ServerException;
import com.aliyuncs.profile.DefaultProfile;
import com.google.gson.Gson;
import java.util.*;
import com.aliyuncs.cms.model.v20190101.*;
public class DescribeSystemEventAttribute {
    public static void main(String[] args) {
        // Initialize a DefaultAcsClient instance.
        DefaultProfile profile = DefaultProfile.getProfile("cn-hangzhou", "<AccessKeyID
>", "<AccessKeySecret>");
        IAcsClient client = new DefaultAcsClient(profile);
        // Query the interruption events of preemptible instances.
        DescribeSystemEventAttributeRequest request = new DescribeSystemEventAttributeR
equest();
        request.setRegionId("cn-hangzhou");
        request.setProduct("ECS");
        request.setEventType("StatusNotification");
        request.setName("Instance:PreemptibleInstanceInterruption");
        try {
            // Obtain the responses.
            DescribeSystemEventAttributeResponse response = client.getAcsResponse(reque
st);
            System.out.println(new Gson().toJson(response));
        } catch (ServerException e) {
            e.printStackTrace();
        } catch (ClientException e) {
            System.out.println("ErrCode:" + e.getErrCode());
            System.out.println("ErrMsg:" + e.getErrMsg());
            System.out.println("RequestId:" + e.getRequestId());
        }
    }
}
```

3. Check the interruption events of preemptible instances based on the responses.

   The following section shows the event notification in the JSON format:

```
{
  "ver": "1.0",
  "id": "2256A988-0B26-4E2B-820A-8A********E5",
  "product": "ECS",
  "resourceId": "acs:ecs:cn-hangzhou:169070********30:instance/i-bp1ecr********5go2go",
  "level": "INFO",
  "name": "Instance:PreemptibleInstanceInterruption",
  "userId": "169070********30",
  "eventTime": "20190409T121826.922+0800",
  "regionId": "cn-hangzhou",
  "content": {
    "instanceId": "i-bp1ecr********5go2go",
    "action": "delete"
  }
}
```

The following table describes parameters in the content field.

Subfield	Description	Example
instanceId	The ID of the preemptible instance.	*i-bp1ecr********5go2go*
action	The action on the preemptible instance. If the parameter is set to *delete*, the interrupted preemptible instance is forcibly recycled.	*delete*

Query the interruption events of preemptible instances by using instance metadata

1.
2. Run the following command to query instance metadata:

```
curl 'http://100.100.100.200/latest/meta-data/instance/spot/termination-time'
```

   - If the HTTP status code of the response is 404, the instance can still be used.
   - If information in the yyyy-MM-ddTHH:mm:ssZ (UTC+0) format is returned, the preemptible instance is recycled at that point in time. Example: `2015-01-05T18:02:00Z`.

Query the interruption events of preemptible instances by calling an API operation

This section describes how to call the DescribeInstances operation to determine whether the instance has entered the *To Be Released* state based on the returned OperationLocks parameter.

The following section shows the sample code of DescribeInstancesSample.java:

```
import com.alibaba.fastjson.JSONArray;
import com.aliyuncs.AcsRequest;
import com.aliyuncs.AcsResponse;
import com.aliyuncs.DefaultAcsClient;
import com.aliyuncs.IAcsClient;
import com.aliyuncs.ecs.model.v20140526.DescribeInstancesRequest;
import com.aliyuncs.ecs.model.v20140526.DescribeInstancesResponse;
```

```
import com.aliyuncs.profile.DefaultProfile;
import com.aliyuncs.profile.IClientProfile;
import java.util.*;
public class DescribeInstancesSample {
    public static void main(String[] args) throws InterruptedException {
        // Initialize a DefaultAcsClient instance.
        OpenApiCaller caller = new OpenApiCaller();
        // Specify the IDs of one or more ECS instances to be queried.
        JSONArray allInstances = new JSONArray();
        allInstances.addAll(Arrays.asList("i-bp1i9c3qiv1qs6nc****"));
        while (!allInstances.isEmpty()) {
            DescribeInstancesRequest request = new DescribeInstancesRequest();
            // Specify the region where the instances are located.
            request.setRegionId("cn-hangzhou");
            // Specify the instance IDs for higher query efficiency.
            request.setInstanceIds(allInstances.toJSONString());
            // Obtain the responses.
            DescribeInstancesResponse response = caller.doAction(request);
            // Obtain instance-related results.
            List<DescribeInstancesResponse.Instance> instanceList = response.getInstances()
;
            if (instanceList != null && !instanceList.isEmpty()) {
                for (DescribeInstancesResponse.Instance instance : instanceList) {
                    // Show the IDs and zone information of the queried instances.
                    System.out.println("result:instance:" + instance.getInstanceId() + ",az
:" + instance.getZoneId());
                    if (instance.getOperationLocks() != null) {
                        for (DescribeInstancesResponse.Instance.LockReason lockReason : ins
tance.getOperationLocks()) {
                            // If an instance is locked, specify the instance ID and the co
rresponding lock reason.
                            System.out.println("instance:" + instance.getInstanceId() + "--
>lockReason:" + lockReason.getLockReason() + ",vmStatus:" + instance.getStatus());
                            if ("Recycling".equals(lockReason.getLockReason())) {
                                // Specify the IDs of the instances to be recycled.
                                System.out.println("spot instance will be recycled immediat
ely, instance id:" + instance.getInstanceId());
                                allInstances.remove(instance.getInstanceId());
                            }
                        }
                    }
                }
                // If a preemptible instance is not locked, it is queried every 2 minutes.
                System.out.print("try describeInstances again later ...");
                Thread.sleep(2 * 60 * 1000);
            } else {
                break;
            }
        }
    }
}
// Initialize a DefaultAcsClient instance.
class OpenApiCaller{
    IClientProfile profile;
```

```
    IAcsClient client;
    public OpenApiCaller() {
        profile = DefaultProfile.getProfile("cn-hangzhou", "<AccessKeyID>", "<AccessKeySecr
et>");
        client = new DefaultAcsClient(profile);
    }
    public  <T extends AcsResponse> T doAction(AcsRequest<T> var1) {
        try {
            return  client.getAcsResponse(var1);
        } catch (Throwable e) {
            e.printStackTrace();
            return null;
        }
    }
}
```

The following result is returned if the recycle is triggered:

```
result:instance:i-bp1i9c3qiv1qs6nc****,az:cn-hangzhou-i
instance:i-bp1i9c3qiv1qs6nc****-->lockReason:Recycling,vmStatus:Stopped
spot instance will be recycled immediately, instance id:i-bp1i9c3qiv1qs6nc****
```

5.3.4. View bills of a preemptible instance

This topic describes how to view the bills of a preemptible instance.

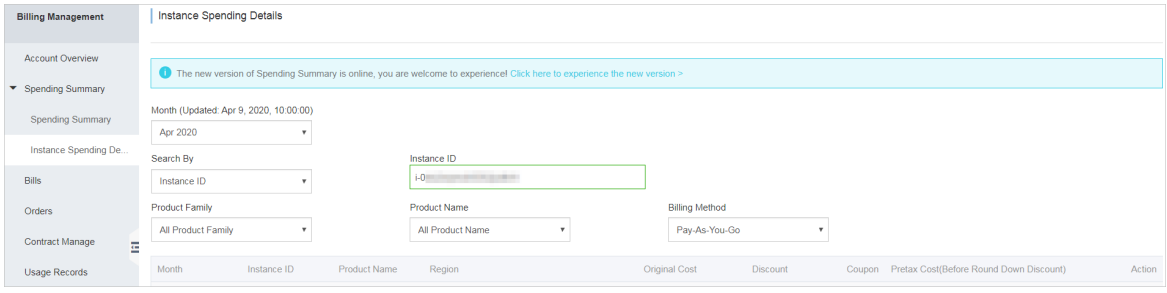
Context

The price of each pay-as-you-go instance is fixed at the time of purchase, whereas the price of each preemptible instance fluctuates with the supply and demand changes.

Procedure

1.
2. Choose **Billing > User Center**.
3. In the left-side navigation pane, choose **Spending Summary > Instance Spending Detail**.
4. Enter the ID of the preemptible instance whose bills you want to view to search for the instance. Find the instance in the searching results and click **Detail** in the **Action** column.

   You can also search for instance bills by product family, product name, or billing method.



5.3.5. Stop a preemptible instance

This topic describes how to stop a preemptible instance and whether the stopped instance can be started under different scenarios.

Context

The following table describes whether a stopped preemptible instance can be started based on its network type, bidding mode, and stop mode.

Network type	Bidding mode	Stop mode	Startup of a stopped preemptible instance
Classic network	SpotWithPriceLimit	Standard mode	The stopped instance can be started within the protection period. After the protection period ends: • If your bid price for the instance type is higher than or equal to the market price and if the resource stock of the instance type is sufficient, the stopped instance can be started. • If the market price exceeds your bid price or if the resource stock of the instance type is insufficient, the stopped instance cannot be started.
	SpotAsPriceGo	Standard mode	The stopped instance can be started within the protection period. After the protection period ends: • If the resource stock of the instance type is sufficient, the stopped instance can be started. • If the resource stock of the instance type is insufficient, the stopped instance cannot be started.
VPC	SpotWithPriceLimit	Standard mode	The stopped instance can be started within the protection period. After the protection period ends: • If your bid price for the instance type is higher than or equal to the market price and if the resource stock of the instance type is sufficient, the stopped instance can be started. • If the market price exceeds your bid price or if the resource stock of the instance type is insufficient, the stopped instance cannot be started.
		Economical mode	If the resource stock of the instance type is sufficient, the stopped instance can be started within the protection period. After the protection period ends: • If your bid price for the instance type is higher than or equal to the market price and if the resource stock of the instance type is sufficient, the stopped instance can be started. • If the market price exceeds your bid price or if the resource stock of the instance type is insufficient, the stopped instance cannot be started.

Network type	Bidding mode	Stop mode	Startup of a stopped preemptible instance
	SpotAsPriceGo	Standard mode	The stopped instance can be started within the protection period. After the protection period ends: • If the resource stock of the instance type is sufficient, the stopped instance can be started. • If the resource stock of the instance type is insufficient, the stopped instance cannot be started.
		Economical mode	If the resource stock of the instance type is sufficient, the stopped instance can be started within the protection period. After the protection period ends: • If the resource stock of the instance type is sufficient, the stopped instance can be started. • If the resource stock of the instance type is insufficient, the stopped instance cannot be started.

Note The economical mode can be enabled only for instances that meet specific conditions. For information about the conditions, see Economical mode.

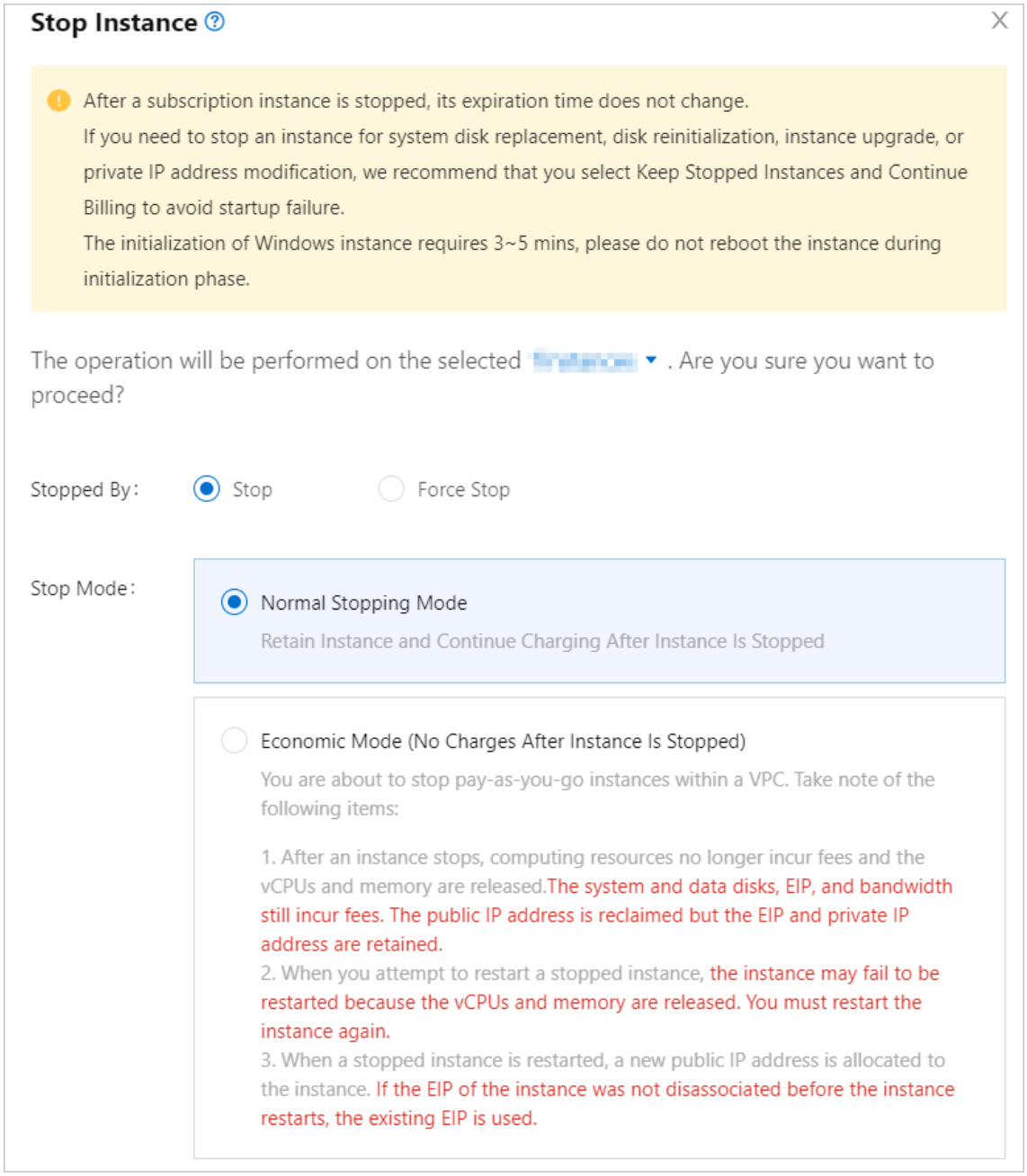
Procedure

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**.



Related information

- StopInstance

5.4. Reserved Instances

5.4.1. Overview

A reserved instance is a discount coupon that can be automatically applied to one or more pay-as-you-go instances, excluding preemptible instances. A reserved instance can also be used to reserve instance resources. A combination of reserved instances and pay-as-you-go instances provides similar cost-effectiveness to subscription instances but with a higher degree of flexibility.

Comparison between reserved instances, pay-as-you-go instances, and subscription instances

The following table describes the differences between reserved instances, pay-as-you-go instances, and subscription instances.

Item	Reserved instance	Pay-as-you-go instance	Subscription instance
Form	A discount coupon. Reserved instances are classified into regional and zonal reserved instances.	An instance that uses the pay-as-you-go billing method. A pay-as-you-go instance is equivalent to a virtual machine. For more information, see Pay-as-you-go.	An instance that uses the subscription billing method. A subscription instance is equivalent to a virtual machine. For more information, see Subscription.
Purpose	Reserved instances cannot be independently used. They must be matched to pay-as-you-go instances to offset bills.	Pay-as-you-go instances can be independently used. They can be used as simple web servers, or used in combination with other Alibaba Cloud services to deliver powerful solutions.	Subscription instances can be independently used. They can be used as simple web servers, or used in combination with other Alibaba Cloud services to deliver powerful solutions.

Scenarios

The following table describes some scenarios where a combination of reserved instances and pay-as-you-go instances is the optimal solution.

Scenario	Combination of reserved instances and pay-as-you-go instances	Pay-as-you-go instance	Subscription instance

<table>
<tr><th>Scenario</th><th>Combination of reserved instances and pay-as-you-go instances</th><th>Pay-as-you-go instance</th><th>Subscription instance</th></tr>
<tr><td>You may need to change the region for your business. You must release Elastic Compute Service (ECS) instances in the original zone and create ECS instances in the zones of the destination region.</td><td rowspan="4">A combination of reserved instances and pay-as-you-go instances has the following benefits:
After you purchase reserved instances, you make commitments to using pay-as-you-go instances for a period of time. Reserved instances provide significant discounts compared with pay-as-you-go pricing.
Purchased reserved instances deliver computing power instead of specific instances. Reserved instances can be matched to eligible pay-as-you-go instances and are more flexible than subscription instances.
You can split or merge reserved instances to offset the bills of pay-as-you-go instances of different instance types.
You can modify the zone of a reserved instance anytime. A regional reserved</td><td rowspan="4">The unit prices of pay-as-you-go instances are higher than those of subscription instances. Pay-as-you-go instances may fail to be created if resources are insufficient. However, pay-as-you-go instances are easier to manage. For example, after you configure automated O&M, instances are automatically created and released. No manual refunds are involved after the instances are released. Pay-as-you-go instances can be used together with Auto Scaling.</td><td>Bills are associated with subscription instances. You may need to pay service fees when you apply for refunds.</td></tr>
<tr><td>During automated O&M, the number of subscription instances needs to be automatically adjusted.</td><td>Refunds must be manually implemented.</td></tr>
<tr><td>You use Auto Scaling to manage ECS instances and have a large number of pay-as-you-go instances. You want to lower your costs.</td><td>You must manually change pay-as-you instances to subscription instances. This process is inefficient and prone to errors.</td></tr>
<tr><td>You want to simplify the lifecycle management operations of subscription instances, such as renewing, releasing, and synchronizing the expiration dates of subscription instances.</td><td>You must perform a large number of operations.</td></tr>
</table>

Scenario	Combination of reserved instances and pay-as-you-go instances	Pay-as-you-go instance	Subscription instance
You want to pay for resources by installments to mitigate financial pressure.	• You can pay by hour by selecting the Partial Upfront or No Upfront payment option to avoid financial pressure caused by one-time payment.		The unit prices are lower. However, one-time payment is required.

Attributes

The following figure shows the key attributes of a reserved instance.

The following table describes the key attributes of the preceding reserved instance.

Section	Attribute	Description

Section	Attribute	Description
①	Instance types	The following instance families support reserved instances: • General-purpose instance families: g7, g6e, g6, g5, g5ne, and sn2ne • Compute-optimized instance families: c7, c6e, c6, c5, ic5, and sn1ne • Memory-optimized instance families: r7, r6e, r6, r5, re6, re4, and se1ne • Big data instance family: d2s • Instance families with local SSDs: i3, i3g, i2, i2g, and i2gne • Instance families with high clock speeds: hfg7, hfc7, hfr7, hfg6, hfc6, hfr6, hfg5, and hfc5 • GPU-accelerated compute-optimized instance families: gn7, gn6i, gn6e, gn6v, gn5, and gn5i • ECS Bare Metal Instance families: ebmgn7, ebmgn6i, ebmgn6e, ebmg6, ebmc6, ebmr6, ebmhfg6, ebmhfc6, and ebmhfr6 • Burstable instance families: t6 and t5
②	Region and zone	For a regional reserved instance, you need only to specify a region. Regional reserved instances provide zone flexibility and instance size flexibility. For a zonal reserved instance, you must specify a region and a zone. Zonal reserved instances provide resource reservation. The matching conditions are determined by the type of a reserved instance. For more information, see Match between reserved instances and pay-as-you-go instances. After you purchase a reserved instance, you can modify its region and zone as needed. For more information, see Modify a reserved instance.
③	Operating system	Reserved instances are classified into Linux and Windows reserved instances. Windows reserved instances can be used to offset the image bills of Windows pay-as-you-go instances.
④	Normalization factor	A normalization factor indicates the performance level of an instance type and also the computing power. Normalization factors are determined by the number of vCPUs. For information about detailed specifications, see View normalization factors.
⑤	Instance quantity	The instance quantity is used for the following purposes: • Calculates the computing power of reserved instances. • Specifies the number of reserved resources for zonal reserved instances. For example, in the preceding figure, the instance quantity is two and the instance type in use is ecs.r6e.xlarge. This indicates that two pay-as-you-go instances of the ecs.r6e.xlarge instance type are reserved.

Section	Attribute	Description
⑥	Computing power	Reserved instances deliver computing power in advance. Pay-as-you-go instances consume the computing power. Computing power of a reserved instance = Normalization factor of an instance type × Instance quantity. The computing power is used for the following purposes: • Evaluates whether the computing power is the same before and after you split or merge reserved instances. • Evaluates the usage of a reserved instance when the size of a regional reserved instance is different from that of the matched pay-as-you-go instance.
⑦	Term	You must specify the term when you purchase a reserved instance. After you split, merge, and modify reserved instances, the terms of the original and new reserved instances are also changed. For more information, see Split a reserved instance, Merge reserved instances, and Modify a reserved instance. **Note** You cannot shorten the term of a reserved instance to deliver more computing power. A reserved instance takes effect and expires on the hour. For example, assume that you purchase a reserved instance with a term of one year at 13:45:00 of May 29, 2020. The reserved instance takes effect at 13:00:00 of May 29, 2020 and expires at 24:00:00 of May 30, 2021. If you have pay-as-you-go instances that match the reserved instance, the reserved instance is applied to offset the bills of the matching pay-as-you-go instances from 13:00:00 of May 29, 2020 by hour until it expires. Expired reserved instances cannot continue to offset the bills of pay-as-you-go instances. However, the pay-as-you-go instances are not released. This ensures continuity of your service. **Note** Make sure that your account balance is sufficient to ensure service availability of your pay-as-you-go instances.

Limits

Limited object	Item	Description

Limited object	Item	Description
Reserved instance	Maximum number of reserved instances	The maximum number of reserved instances is subject to the type of reserved instances. • Maximum number of regional reserved instances: Each account can have up to 20 regional reserved instances across all regions. • Maximum number of zonal reserved instances: Each account can have up to 20 zonal reserved instances in each zone. For example, you can purchase 10 regional reserved instances in the China (Hangzhou) region and 10 regional reserved instances in the China (Qingdao) region, but the total number of regional reserved instances cannot exceed 20. You can purchase 20 zonal reserved instances in Hangzhou Zone B and 20 zonal reserved instances in Hangzhou Zone H. If you need more reserved instances, submit a ticket.
ECS instance	Billing method	• Reserved instances can be matched only to pay-as-you-go instances. Reserved instances cannot be matched to preemptible instances. • Reserved instances can be used to offset the bills for computing resources of pay-as-you-go instances. They cannot be used to offset the bills for network and the storage resources of pay-as-you-go instances.
	Instance family	The gn6i and t5 instance families support only zonal reserved instances, not regional reserved instances. gn6i and t5 reserved instances cannot be split or merged.

Billing

Reserved instances support the All Upfront, Partial Upfront, and No Upfront payment options. Billing standards vary with payment options. For more information, see Reserved instances.

> **Note** Your ECS usage determines whether you can use the No Upfront payment option. If you want to use the No Upfront payment option, submit a ticket.

References

- Match between reserved instances and pay-as-you-go instances
- Purchase reserved instances
- Split a reserved instance
- Merge reserved instances
- Modify a reserved instance
- Instance FAQ

5.4.2. Match between reserved instances and pay-as-you-go instances

After you purchase reserved instances, you make commitments to using pay-as-you-go instances for a period of time. Reserved instances can be used to offset bills of pay-as-you-go instances only when they match the pay-as-you-go instances. If you do not have any pay-as-you-go instances in your account, the reserved instances are idle and continue to be charged. This topic describes the matching rules of reserved instances and provides some examples.

Matching rules

The match between reserved instances and pay-as-you-go instances cannot be manually managed. After you purchase a reserved instance, the reserved instance automatically matches one or more pay-as-you-go instances that have some attributes within the term. After the reserved instance is matched, it checks for eligible pay-as-you-go bills on an hourly basis and deducts fees based on its computing power. You can view pay-as-you-go instances that match the reserved instance. For more information, see View matched pay-as-you-go instances.

The matching conditions of a regional reserved instance are different from those of a zonal reserved instance. The following table describes the attributes that determine the match between reserved instances and pay-as-you-go instances.

Attribute	Regional reserved instance	Zonal reserved instance
Region and zone	A regional reserved instance can match one or more pay-as-you-go instances across zones within a specific region.	A zonal reserved instance can match one or more pay-as-you-go instances only in the same zone.
Instance type	The following section describes the instance size flexibility and resource reservation status: • Within the same instance family, regional reserved instances of small instance types can match pay-as-you-go instances of large instance types or regional reserved instances of large instance types can match pay-as-you-go instances of small instance types. • Resource reservation is not supported.	The following section describes the instance size flexibility and resource reservation status: • A zonal reserved instance can match only pay-as-you-go instances of the same instance type. • Resource reservation is supported. A specific number of pay-as-you-go instances of the specified instance types are reserved within the term. This ensures that you can create pay-as-you-go instances anytime in the specified zone.
Operating system	A regional reserved instance can match only pay-as-you-go instances that have the same operating system.	A zonal reserved instance can match only one or more pay-as-you-go instances that have the same operating system.

Examples of regional reserved instances

Regional reserved instances and pay-as-you-go instances must meet the following requirements to match:

- They must reside within the same region.

- They must use instances types that belong to the same instance family. Within the same instance family, regional reserved instances of small instance types can match pay-as-you-go instances of large instance types or regional reserved instances of large instance types can match pay-as-you-go instances of small instance types.

  > **Note** You can evaluate the usage of reserved instances based on the computing power when the instance sizes of the reserved instance and the pay-as-you-go instance are different.

- They must run the same operating system.

The following table describes examples of regional reserved instances.

Scenario	Pay-as-you-go instance	Regional reserved instance	How bills are offset
Regional reserved instances of small instance types match pay-as-you-go instances of large instance types	You have a pay-as-you-go instance that has the following configurations: • Qingdao Zone B • ecs.g5.4xlarge • Linux	You have two active regional reserved instances that have the following attributes: • All zones in China (Qingdao) • ecs.g5.xlarge • Linux • Instance quantity: 2	The following section describes the normalization factors: • The normalization factor of ecs.g5.4xlarge is 16. • The normalization factor of ecs.g5.xlarge is 4. Comparison between the delivered and consumed computing powers: • Pay-as-you-go instance: The pay-as-you-go instance consumes 16 units of computing power per hour (1 instance × 16). • Reserved instance: The reserved instance delivers 8 units of computing power per hour (2 instances × 4). Therefore, one reserved instance offsets 50% of the pay-as-you-go instance bills per hour, and two reserved instances offset 100% of the pay-as-you-go instance bills per hour.

Scenario	Pay-as-you-go instance	Regional reserved instance	How bills are offset

Regional reserved instance of large instance types match pay-as-you-go instances of small instance types	You have six pay-as-you-go instances.
	• The following section describes the configurations of one of the pay-as-you-go instances:
	  ○ Qingdao Zone B
	  ○ ecs.g5.xlarge
	  ○ Linux
	• The following section describes the configurations of the other five pay-as-you-go instances:
	  ○ Qingdao Zone C
	  ○ ecs.g5.xlarge
	  ○ Linux
		You have two active regional reserved instances.
		• The following section describes the attributes of one of the reserved instances:
		  ○ All zones in China (Qingdao)
		  ○ ecs.g5.4xlarge
		  ○ Linux
		  ○ Instance quantity: 1
		• The following section describes the attributes of the other reserved instance:
		  ○ All zones in China (Qingdao)
		  ○ ecs.g5.2xlarge
		  ○ Linux
		  ○ Instance quantity: 1
			The following section describes the normalization factors:
			• The normalization factor of ecs.g5.xlarge is 4.
			• The normalization factor of ecs.g5.2xlarge is 8.
			Comparison between the delivered and consumed computing powers:
			• Pay-as-you-go instance: The six pay-as-you-go instances consume 24 units of computing power per hour (6 instances × 4).
			• Reserved instance: One reserved instance delivers 16 units of computing power (1 instance × 16) per hour and the other reserved instance delivers 8 units of computing power (1 instance × 8) per hour.
			Therefore, the two reserved instances offset 100% of bills of the six pay-as-you-go instances per hour.
			Note The pay-as-you-go instances belong to two zones. However, the regional reserved instances support the zone flexibility and can offset bills of pay-as-you-go instances across zones.

Scenario	Pay-as-you-go instance	Regional reserved instance	How bills are offset

Failed to match	You have two pay-as-you-go instances.
	• The following section describes the configurations of one of the pay-as-you-go instances:
	  ○ Qingdao Zone B
	  ○ ecs.g5.xlarge
	  ○ Windows
	• The following section describes the configurations of the other pay-as-you-go instance:
	  ○ Hangzhou Zone B
	  ○ ecs.c5.xlarge
	  ○ Linux
		You have two active regional reserved instances.
		• The following section describes the attributes of one of the reserved instances:
		  ○ All zones in China (Qingdao)
		  ○ ecs.g5.4xlarge
		  ○ Linux
		  ○ Instance quantity: 1
		• The following section describes the attributes of the other reserved instance:
		  ○ All zones in China (Qingdao)
		  ○ ecs.g5.xlarge
		  ○ Linux
		  ○ Instance quantity: 1
			The reserved instances failed to match the pay-as-you-go instances due to the following causes:
			• The operating system of one of the pay-as-you-go instances is Windows.
			• One of the pay-as-you-go instances resides in China (Hangzhou) and uses an instance type that belongs to the c5 instance family.
			Therefore, the reserved instances remain idle and continue to be charged. The bills of pay-as-you-go instances are paid by account balance.

Examples of zonal reserved instances

Zonal reserved instances and pay-as-you-go instances must meet the following requirements to match:

- They must reside within the same region and zone.
- They must use instance types that belong to the same instance family and have the same size.
- They must run the same operating system.

The following table describes examples of zonal reserved instances.

Scenario	Pay-as-you-go instance	Zonal reserved instance	How bills are offset
Matched	You have five pay-as-you-go instances that have the following configurations: • Qingdao Zone B • ecs.g5.xlarge • Windows	You have an active zonal reserved instance that has the following attributes: • Qingdao Zone B • ecs.g5.xlarge • Windows • Instance quantity: 5	The reserved instance matches the pay-as-you-go instances. The reserved instance offsets 100% of bills of the five pay-as-you-go instances per hour.

Scenario	Pay-as-you-go instance	Zonal reserved instance	How bills are offset
Resource reservation	You have no pay-as-you-go instances.	You have an active zonal reserved instance that has the following attributes: • Qingdao Zone B • ecs.g5.2xlarge • Linux • Instance quantity: 10	The reserved instance is idle and continues to be charged. However, 10 pay-as-you-go instances of the ecs.g5.2xlarge instance type are reserved for you within the term of the reserved instance. This ensures that you can create pay-as-you-go instances anytime in Qingdao Zone B.
Failed to match	You have two pay-as-you-go instances. • The following section describes the configurations of one of the pay-as-you-go instances: ◦ Qingdao Zone B ◦ ecs.g5.xlarge ◦ Windows • The following section describes the configurations of the other pay-as-you-go instance: ◦ Qingdao Zone C ◦ ecs.g5.4xlarge ◦ Linux	You have two active zonal reserved instances. • The following section describes the attributes of one of the reserved instances: ◦ Qingdao Zone B ◦ ecs.g5.xlarge ◦ Linux ◦ Instance quantity: 1 • The following section describes the attributes of the other reserved instance: ◦ Qingdao Zone B ◦ ecs.g5.xlarge ◦ Linux ◦ Instance quantity: 1	The reserved instances failed to match the pay-as-you-go instances due to the following causes: • The operating system of one of the pay-as-you-go instances is Windows. • One of the pay-as-you-go instances resides in Qingdao Zone C and uses the ecs.g5.4xlarge instance type. Therefore, the reserved instances remain idle and continue to be charged. The bills of pay-as-you-go instances are paid by account balance.

Match between a single reserved instance and multiple pay-as-you-go instances

A single regional or zonal reserved instance can match multiple pay-as-you-go instances. However, you cannot shorten the term of a reserved instance to deliver more computing power.

The following table describes an example that involves six pay-as-you-go instances and an active reserved instance.

Six pay-as-you-go instances	An active reserved instance
The following section describes the configurations of each pay-as-you-go instance: • Qingdao Zone B • ecs.g5.6xlarge • Linux • Normalization factor of the instance type: 24	The following section describes the attributes of the reserved instance: • Qingdao Zone B • ecs.g5.6xlarge • Linux • Normalization factor of the instance type: 24 • Instance quantity: 1 • Term: one year

The six pay-as-you-go instances all match the reserved instance. The following table describes examples of how the reserved instances offset bills of the pay-as-you-go instances based on how long each pay-as-you-go instance exists.

Six pay-as-you-go instances	A reserved instance	How bills are offset
The following section describes the consumed computing power: • Duration of each pay-as-you-go instance: one hour (each consumes 24 units of computing power) • Computing power consumed by the six pay-as-you-go instances per hour: 144 (6 instances × 24)	The following section describes the delivered computing power: Computing power delivered by the reserved instance per hour: 24 (1 instance × 24)	The computing power delivered by the reserved instance is equal to that consumed by a pay-as-you-go instance. The reserved instance offsets the bills of a pay-as-you-go instance randomly. You cannot shorten the term of the reserved instance to two months to offset bills of the six pay-as-you-go instances at the same time.
The following section describes the consumed computing power: • Duration of each pay-as-you-go instance: 10 minutes (each consumes 24/6 units of computing power) • Computing power consumed by the six pay-as-you-go instances per hour: 24 (6 instances × 24/6)	The following section describes the delivered computing power: Computing power delivered by the reserved instance per hour: 24 (1 instance × 24)	The computing power delivered by the reserved instance is equal to that consumed by the six pay-as-you-go instances. The reserved instance offsets bills of the six pay-as-you-go instances.
The following section describes the consumed computing power: • Duration of each pay-as-you-go instance: 15 minutes (each consumes 24/4 units of computing power) • Computing power consumed by the six pay-as-you-go instances per hour: 36 (6 instances × 24/4)	The following section describes the delivered computing power: Computing power delivered by the reserved instance per hour: 24 (1 instance × 24)	The computing power delivered by the reserved instance is less than that consumed by the six pay-as-you-go instances. The reserved instance offsets bills of the six pay-as-you-go instances for an hour. The deducted amount of each pay-as-you-go instance is random.

5.4.3. Purchase reserved instances

This topic describes how to purchase reserved instances in the ECS console.

Prerequisites

- Before you purchase reserved instances, make sure that the pay-as-you-go instances that you want to match meet the requirements for applying reserved instances. For more information, see Overview.
- You cannot manually manage how reserved instances and pay-as-you-go instances are matched. Make sure that you understand the matching rules for reserved instances. For more information, see Match between reserved instances and pay-as-you-go instances.

Procedure

1.
2.
3.
4. Click **Purchase Reserved Instance**.
5. Configure region-related parameters.
   i. Select a region.
   ii. Set Resource Reservation.

   > **Note** Only zonal reserved instances support resource reservation. Regional reserved instances apply to pay-as-you-go instances of different sizes in different zones within the same region.

   iii. Select a zone.
6. Configure instance-related parameters.
   i. Select an instance type.

   > **Note** You must select an instance type when you purchase a regional reserved instance. The regional reserved instance can match any pay-as-you-go instances of the specified instance family within the specified region regardless of size.

   ii. Set Operating System Platform.

   You can select **Linux** or **Windows**.

   > **Note** The reserved instance matches only pay-as-you-go instances that use the selected type of operating system. The operating system of an reserved instance cannot be changed after you purchase the reserved instance.

   To apply a reserved instance to pay-as-you-go instances created from Bring Your Own License (BYOL) images, submit a ticket.

   iii. Set Payment Option.

   All Upfront, **Partial Upfront**, and **No Upfront** are available. For more information, see Reserved instance billing.

7. Configure purchase-related parameters.
   i. (Optional)Set Name.
   ii. Set Term.

   You can select **1 Month**, **1 Year**, **3 Years**, or **5 Years**.

   iii. Set Quantity.

   The reserved instance can match the specified number of pay-as-you-go instances of the specified instance type. For example, if the instance type is ecs.g5.large and Quantity is set to 3, the reserved instance can match three pay-as-you-go instances of the ecs.g5.large instance type.
8. Configure tags.

   You can add tags for multiple types of Alibaba Cloud resources. You can use tags to perform cost sharing, monitoring by group, and automated O&M by group. For more information, see Overview.
9. Read and select *ECS Terms of Service* and then click **Purchase**.
10. In the message that appears, confirm that the parameters are correct and click **Create Order**.
11. Read the payment information and then click **Subscribe**.

Result

After you purchase a reserved instance, you can immediately use it to offset bills when the reserved instance matches one or more pay-as-you-go instances. You can also manage the reserved instance in response to pay-as-you-go instance changes.

Related information

- PurchaseReservedInstancesOffering
- Split a reserved instance
- Merge reserved instances
- Modify a reserved instance

5.4.4. Split a reserved instance

You can split a single reserved instance into multiple reserved instances to match smaller pay-as-you-go instances. For ease of description, the reserved instance to be split is referred to as the original reserved instance. The resulting reserved instances are referred to as the destination reserved instances.

Prerequisites

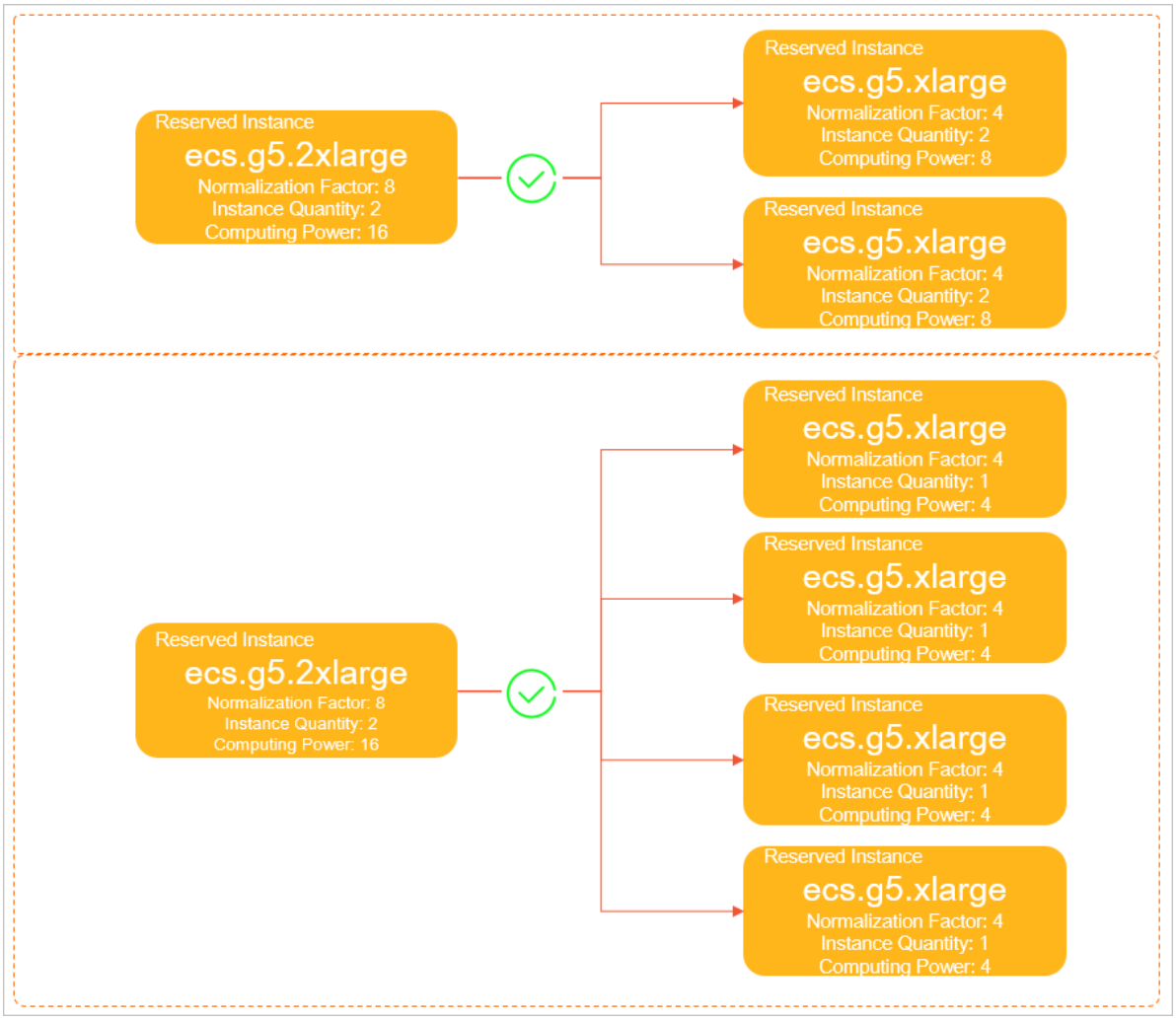
- The original reserved instance is in the **Active** state.
- No ongoing requests for splitting, merging, or modifying reserved instances exist.

Context

When you split a reserved instance, take note of the following items:

- Reserved instances that belong to the gn6i and t5 instance families cannot be split.
- You can change the instance type of a reserved instance. However, you cannot change the instance family of a reserved instance.
- You cannot change the zone or region of a reserved instance.

- The total computing power of destination reserved instances must be equal to that of the original reserved instance. For more information about computing power, see Match between reserved instances and pay-as-you-go instances. The following figure shows an example of splitting a reserved instance.



Procedure

1.
2.
3. On the **Reserved Instances** page, find the original reserved instance and click **Split** in the **Actions** column.
4. In the **Split Reserved Instance** pane, configure the names, instance types, and quantities of the destination reserved instances.
5. Click **OK**.

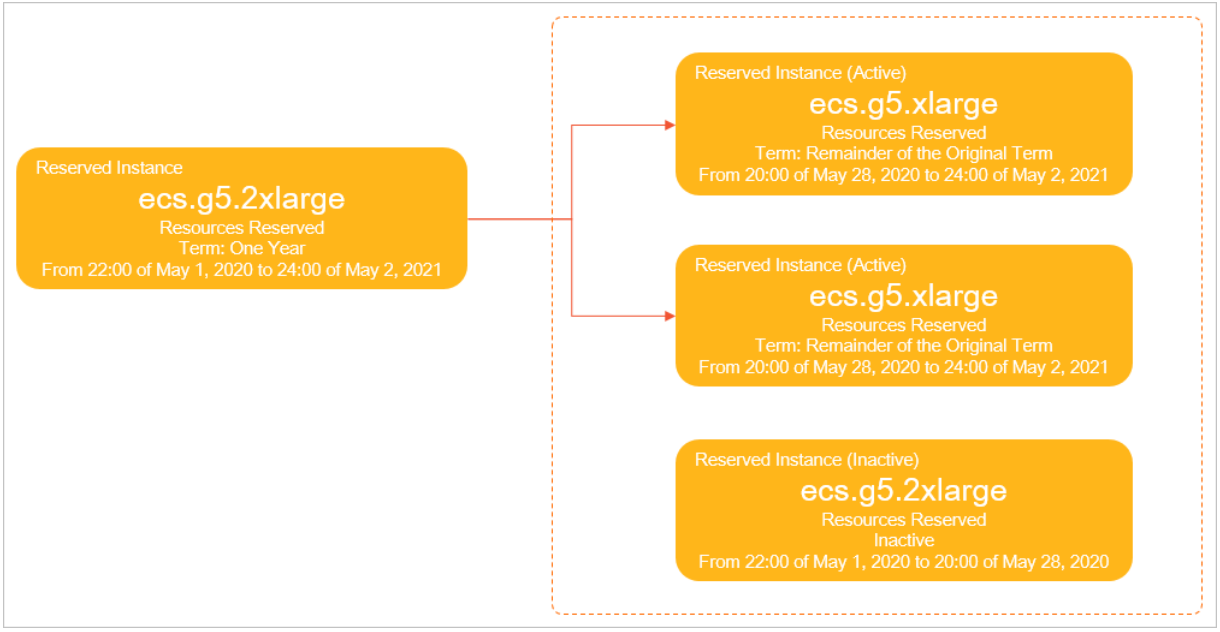
Result

After you submit the request for instance splitting, the original reserved instance enters the **Updating** state, and the destination reserved instances in the Creating state are displayed. You cannot cancel the ongoing request for splitting a reserved instance. If you want to roll back the change made by the splitting operation, you can merge the destination reserved instances to obtain the original reserved instance.

After the request for splitting a reserved instance is processed, you can obtain one of the following results:

- If the reserved instance is split:
  - The original reserved instance enters the **Inactive** state and expires on the hour when it is split. The price becomes USD 0.
  - The destination reserved instances enter the **Active** state and take effect on the hour when the original reserved instance is split. If the destination reserved instances are zonal reserved instances, the type of reserved resources is updated automatically.
  - If the destination reserved instances match pay-as-you-go instances, the billing discounts provided by these reserved instances are applied to the matched pay-as-you-go instances starting from the hour when the destination reserved instances take effect.
- If the original reserved instance fails to be split, this reserved instance remains active.

Assume that you split an ecs.g5.2xlarge zonal reserved instance into two ecs.g5.xlarge zonal reserved instances at 20:30 of May 28, 2020. The following figure shows the time when the original reserved instance expires and the time when the destination reserved instances take effect.



Related information

- ModifyReservedInstances
- DescribeReservedInstances

5.4.5. Merge reserved instances

This topic describes how to merge reserved instances. If traffic to your instances increases, you can merge multiple reserved instances into a single reserved instance to match larger pay-as-you-go instances. For ease of description, the reserved instances to be merged are referred to as the original reserved instances. The resulting merged reserved instance is referred to as the destination reserved instance.

Prerequisites

- The original reserved instances are purchased by using the same currency.
- The original reserved instances are in the **Active** state.
- The original reserved instances belong to the same region if they are regional reserved instances. The original reserved instances belong to the same zone if they are zonal reserved instances.
- The original reserved instances have the same operating system and end time.

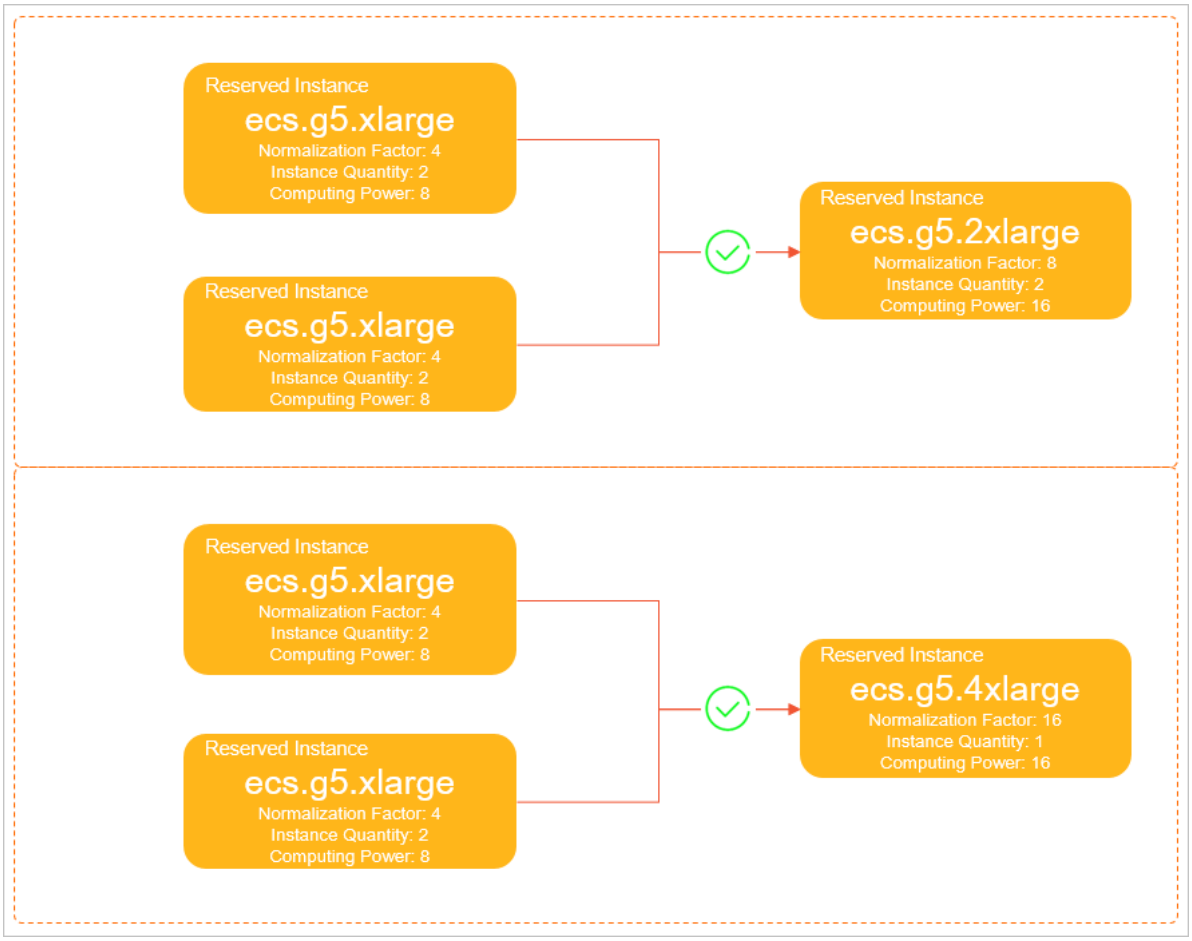
  > **Note** The end time of a reserved instance is determined when you purchase the reserved instance, and is not necessarily the time when the reserved instance becomes inactive. For example, after you merge the reserved instances, the original reserved instances become inactive. However, the end time of the original reserved instances remains unchanged.

- No ongoing requests for splitting, merging, or modifying reserved instances exist.

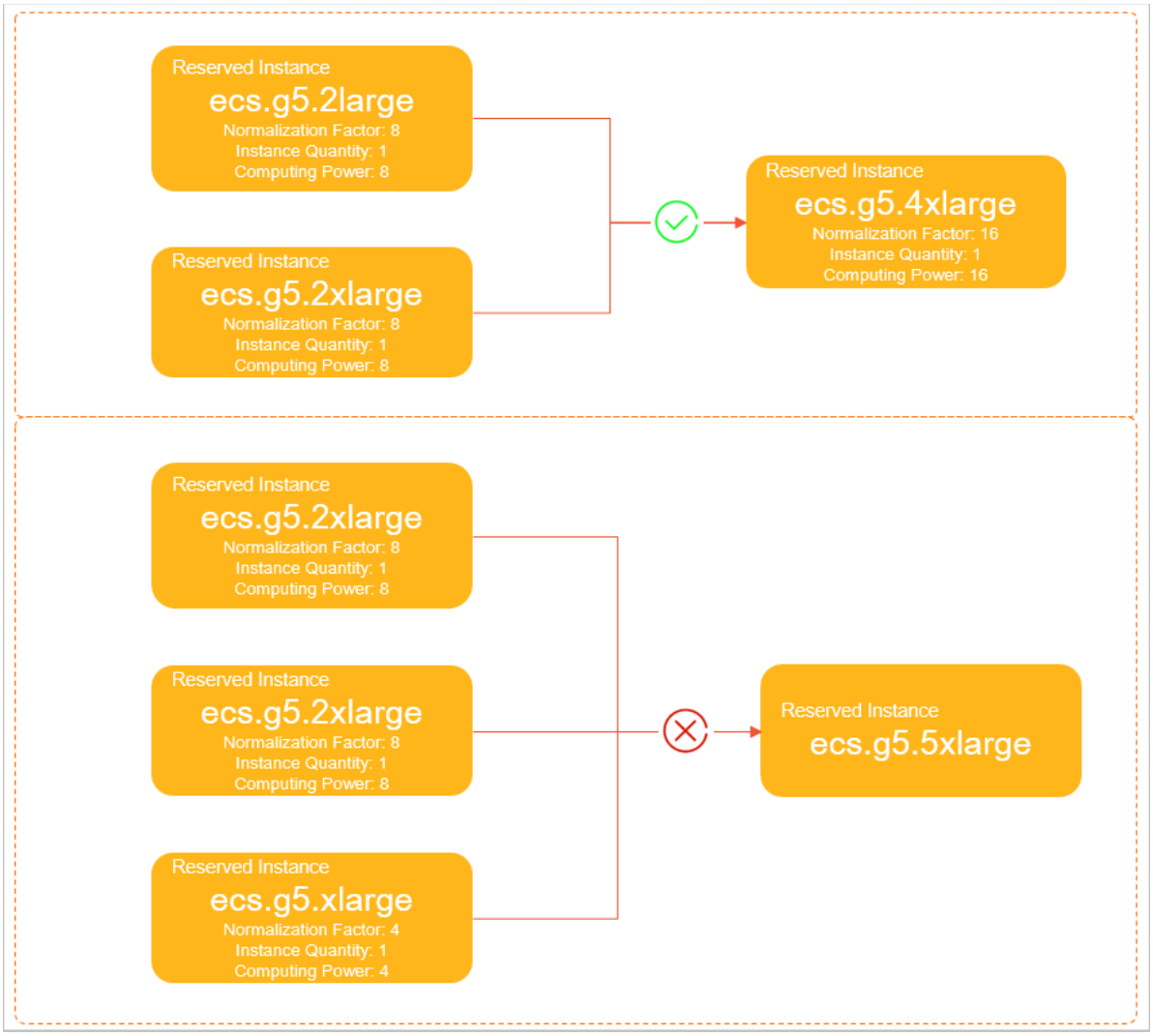
Context

When you merge reserved instances, take note of the following items:

- Reserved instances that belong to the gn6i and t5 instance families cannot be merged.
- You can change the instance type of a reserved instance. However, you cannot change the instance family of a reserved instance.
- You cannot change the zone or region of a reserved instance.
- The number of pay-as-you-go instances to which the destination reserved instance can be applied cannot exceed 100.
- The total computing power of the destination reserved instance must be equal to that of the original reserved instances. For more information about computing power, see Match between reserved instances and pay-as-you-go instances. The following figure shows an example of merging reserved instances.



- The instance type of the destination reserved instance must be valid. For example, you cannot obtain an ecs.g5.5xlarge reserved instance because the ecs.g5.5xlarge instance type is not included in the g5 instance family, as shown in the following figure.



Procedure

1.
2.
3. On the **Reserved Instances** page, find an original reserved instance and click **Merge** in the **Actions** column.
4. In the **Merge Reserved Instances** pane, select the reserved instances that you want to merge with the current reserved instance. Configure the name, instance type, and quantity of the destination reserved instance.
5. Click **OK**.

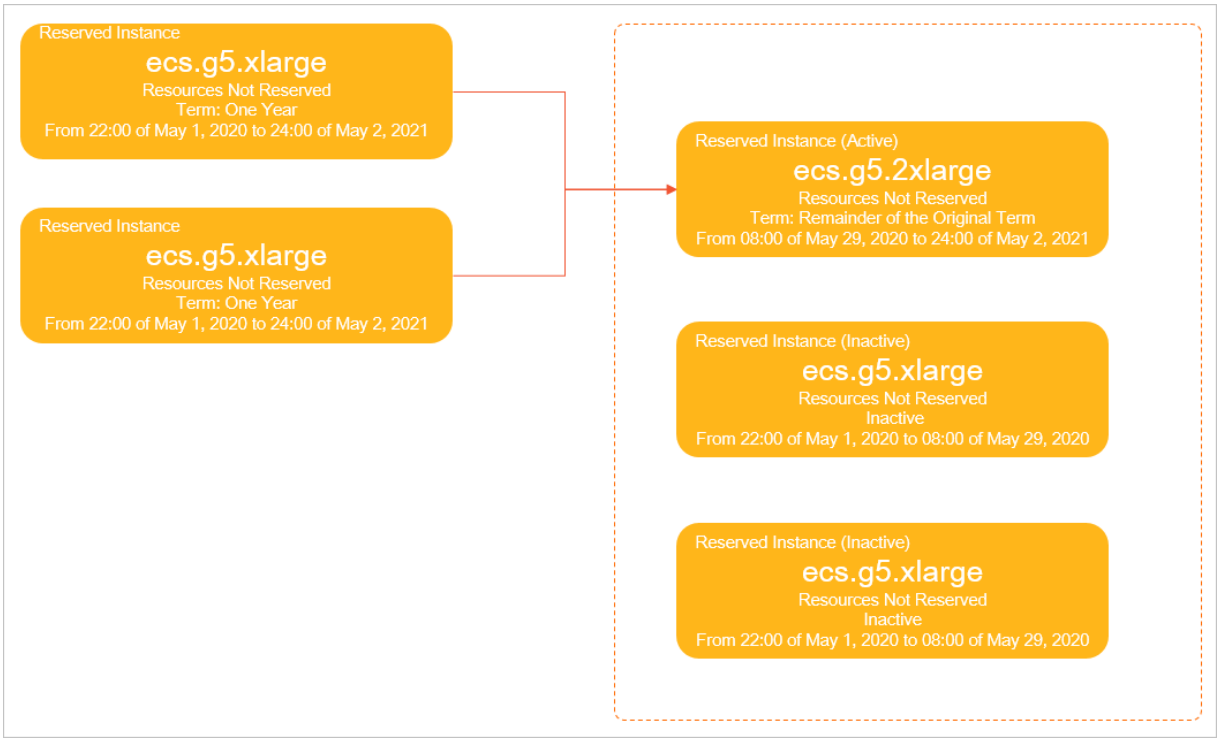
Result

After you submit the request for instance merging, the original reserved instances enter the **Updating** state, and the destination reserved instance in the Creating state is displayed. You cannot cancel the ongoing request for merging reserved instances. If you want to roll back the change made by the merging operation, you can split the merged reserved instance to obtain the original reserved instances.

After the request for merging reserved instances is processed, you can obtain one of the following results:

- If the reserved instances are merged:
  - The original reserved instances enter the **Inactive** state and expires on the hour when they are merged. The prices become USD 0.
  - The destination reserved instance enters the **Active** state and takes effect on the hour when the original reserved instances are merged. If the destination reserved instance is a zonal reserved instance, the type of reserved resources is updated automatically.
  - If the destination reserved instance matches pay-as-you-go instances, the billing discount provided by the reserved instance is applied to the matched pay-as-you-go instances starting from the hour when the destination reserved instance takes effect.
- If the original reserved instances fail to be merged, they remain active.

Assume that you merged two ecs.g5.xlarge regional reserved instances into an ecs.g5.2xlarge regional reserved instance at 08:30 of May 29, 2020. The following figure shows the time when the original reserved instances expire and the time when the destination reserved instance takes effect.



Related information

- ModifyReservedInstances
- DescribeReservedInstances

5.4.6. Modify a reserved instance

You can modify the name of a reserved instance. You can modify the scope (zone or region) of a reserved instance to meet your service requirements. For ease of description, the reserved instance to be modified is referred to as the original reserved instance. The resulting modified reserved instance is referred to as the target reserved instance.

Prerequisites

- The original reserved instance is in the **Active** state.

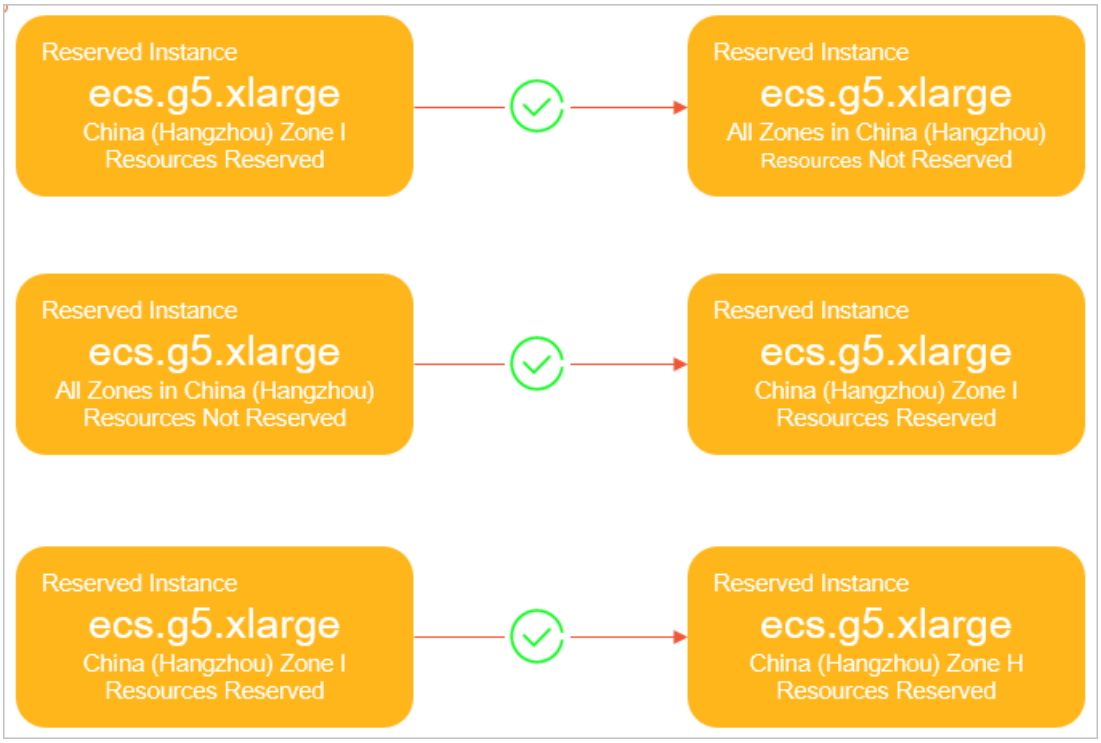
- No ongoing requests for splitting, merging, or modifying reserved instances exist.

Context

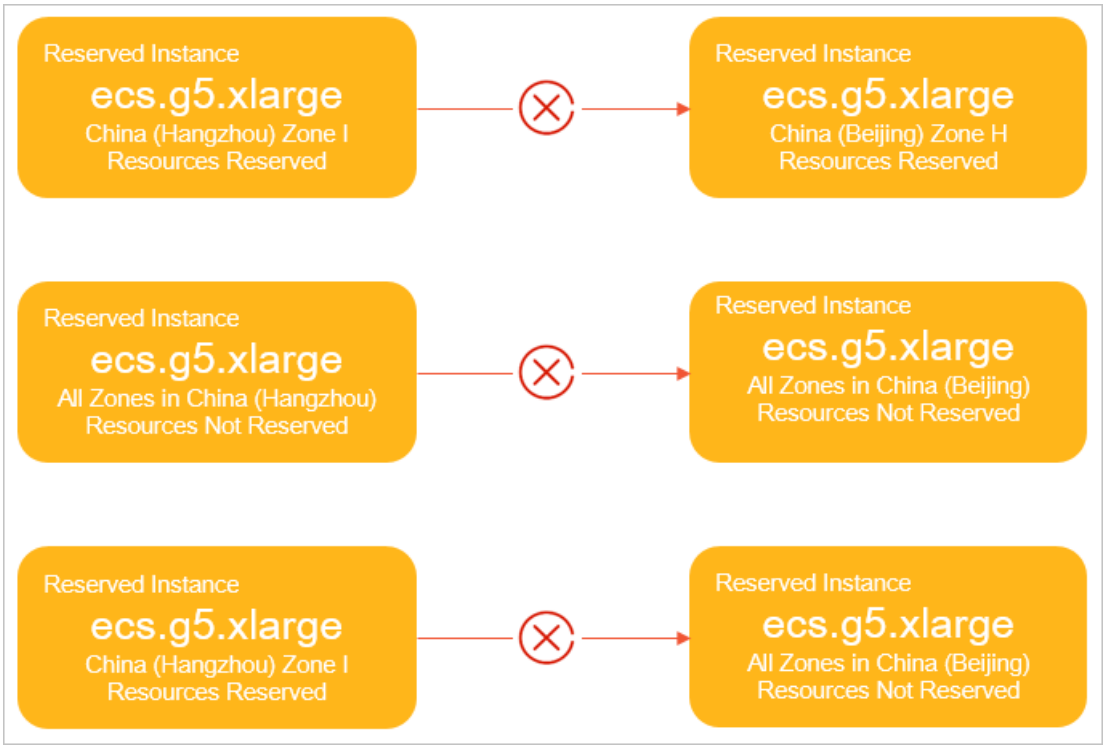
When you modify the zone or region of a reserved instance, you can make the following changes:

- From a zonal reserved instance to a regional one within the same region
- From a regional reserved instance to a zonal one within the same region
- From one zone to another within the same region

The following figure shows the modified reserved instances.



You cannot modify the scope of a reserved instance across regions. The following figure shows reserved instances that fail to be modified.



Procedure

1.
2.
3. On the **Reserved Instances** page, find the original reserved instance and click **Modify** in the **Actions** column.
4. In the **Modify Reserved Instance** pane that appears, reset the name, region, and zone.

   > **Note** If instance types used by reserved instances are insufficient in the target zone, you cannot select this zone.

5. Click **OK**.

Result

After you submit the modification request, the original reserved instance enters the **Updating** state, and the target reserved instance in the Creating state is displayed. You cannot cancel the ongoing request for modifying a reserved instance. If you want to roll back the change made by the modification operation, you can modify the reserved instance again.

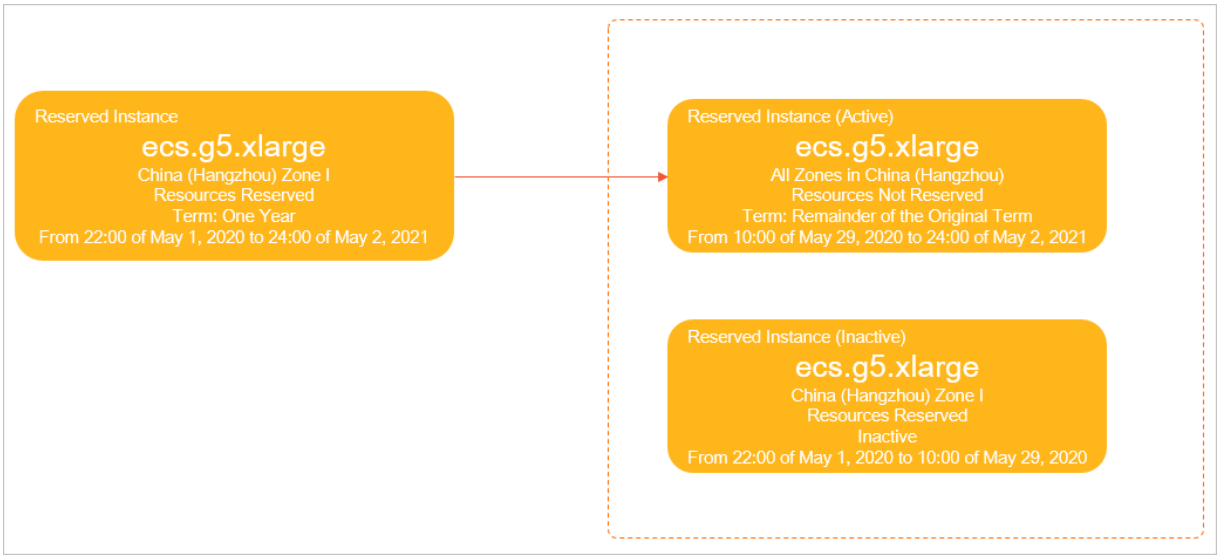
After the request for modifying a reserved instance is processed, you can obtain one of the following results:

- If the reserved instance is modified:
  - The original reserved instance enters the **Inactive** state and expires at the top of the hour when it is modified. The price becomes USD 0.

- The target reserved instance enters the **Active** state and takes effect at the top of the hour when the original reserved instance is modified. If the target reserved instance is a zonal reserved instance, the zone where resources are reserved will also be updated automatically.
- If the target reserved instance matches one or more pay-as-you-go instances, the billing discount provided by this reserved instance is applied to the matched pay-as-you-go instances starting from the hour when the target reserved instance takes effect.

- If the original reserved instance fails to be modified, it remains active.

Assume that you modified a zonal reserved instance with reserved resources to a regional reserved instance without reserved resources at 10:50 of May 29, 2020. The following figure shows the time when the original reserved instance expires and the time when the target reserved instance takes effect.



Related information

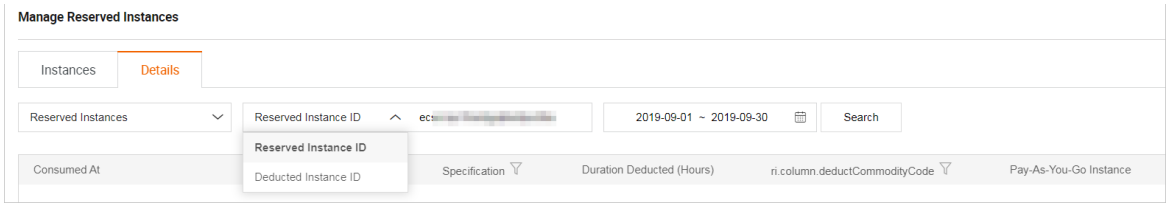
- ModifyReservedInstances
- DescribeReservedInstances

5.4.7. View the usage details of a reserved instance

After you purchase a reserved instance (RI), you can view its matching pay-as-you-go instances and usage details in the ECS console. The bills of pay-as-you-go instances that are offset by RIs are refreshed every hour. This topic describes how to view the usage details of an RI.

Procedure

1.
2.
3. Find the target RI and click **View Bill** in the **Actions** column.
4. On the page that appears, set the search conditions and click **Search**.



5.4.8. View normalization factors

Reserved instances deliver computing power in advance. Pay-as-you-go instances consume the computing power. A normalization factor indicates the performance level of an instance type and also the computing power. This topic describes how to view and download the normalization factor table.

Context

Computing power of a reserved instance = Normalization factor of an instance type × Instance quantity.

The computing power of a reserved instance can be used for the following purposes:

- Evaluates whether the computing power is the same before and after you split or merge reserved instances.
- Evaluates the usage of a reserved instance when the size of a regional reserved instance is different from that of the matched pay-as-you-go instance.

Procedure

1.
2.
3. In the upper-right corner of the **Reserved Instances** page, click **View Normalization Factor Table**.
4. View the normalization factor of each instance type.

   You can view normalization factors by instance family. You can also click **Download** to save the normalization factor table to your computer for later use.

5.4.9. View matched pay-as-you-go instances

Reserved instances allow you to view the purchased pay-as-you-go instances that the reserved instances match. If pay-as-you-go instances are displayed in the matched instance list of a reserved instance, these pay-as-you-go instances match the reserved instance. However, the reserved instance is not necessarily applied to all these pay-as-you-go instances. You can check your bills for usage details of the reserved instance.

Procedure

1.
2.
3. In the upper-left corner of the **Reserved Instances** page, click **View Reserved Instances**.
4. Select the region and zone where the target reserved instance resides.
5. Find the reserved instance and click **View the matching instances** that corresponds to the reserved instance.

> **Note** The pay-as-you-go instances created by E-MapReduce (EMR) or Alibaba Cloud Container Service for Kubernetes (ACK) are not displayed in the matching list of reserved instances. However, these pay-as-you-go instances can also benefit from billing discounts of reserved instances.

Result

You are directed to the Instances page where the matched pay-as-you-go instances are displayed.

5.4.10. Edit the tags of a reserved instance

Tags can be used to identify resources with the same characteristics (such as reserved instances that belong to the same organization or that serve the same purpose) for easy search and management. This topic describes how to edit the tags of an existing reserved instance.

Context

For information about how to use tags, the resources that support tags, and the limits on tags, see Overview and the "Tag limits" section of the Limits topic.

Procedure

1.
2.
3.
4. Find the reserved instance whose tags you want to edit, move the pointer over the icon in the **Tag** column, and then click **Edit Tags**.
5. On the **Tag Editor** tab, find the reserved instance and click **Edit Tags**.
6. On the **Manage Tags** page, click **Add Tag**.
7. Select an existing tag or create a tag. Then, click **Submit**.
8. In the **Confirmation** message, click **OK**.

What's next

After tags are added to your reserved instances, you can filter the reserved instances by tag to perform different operations. For example, you can renew reserved instances that have a set of tags and change the regions or zones of reserved instances that have a different set of tags.

5.4.11. Automatically purchase identical replacements when existing reserved instances expire

To continuously benefit from the billing discounts provided by reserved instances, we recommend that you configure the automatic purchase of reserved instances before your reserved instances expire. This topic describes how to configure the automatic purchase of reserved instances in Operation Orchestration Service (OOS) to ensure that identical replacements are in place when existing reserved instances expire.

Context

OOS is an automated O&M service provided by Alibaba Cloud to help you manage and execute O&M tasks. You can create templates to define execution tasks, the order in which to execute the tasks, input parameters, and output parameters, and use the templates to automate O&M tasks. For more information, see Introduction to OOS.

You can use OOS to automatically purchase identical replacements when existing reserved instances expire. When a replacement reserved instance is automatically purchased and matched to pay-as-you-go instances, it can be applied to offset the bills of the pay-as-you-go instances. Automatic purchase of reserved instances helps you automate the management of reserved instances and respond to pay-as-you-go instance changes. To configure the automatic purchase of reserved instances, perform the following steps:

- Step 1: Create a template in the OOS console
- Step 2: Create a scheduled O&M task to execute the template on a regular basis

Step 1: Create a template in the OOS console

1. Log on to the OOS console.
2. In the left-side navigation pane, click **My Templates**.
3. On the **My Templates** page, click **Create Template**.
4. On the **Create Template** page, configure parameters.
   i. In the **Basic Information** section, enter a name for the template.

      You can click **Show More** to add tags, select a resource group, or enter a description for the template version based on your business requirements.

   ii. Click the **YAML** tab and copy a template that is used to batch clone reserved instances to the code editor.

      Sample YAML-formatted template used to batch clone reserved instances:

```
FormatVersion: OOS-2019-06-01
Description:
  en: Bulky clone reserved instances
  name-en: ACS-ECS-BulkyCloneReservedInstances
  categories:
    - instance_manage
Parameters:
  regionId:
    Type: String
    Description:
      en: The id of region.
    AssociationProperty: RegionId
    Default: '{{ ACS::RegionId }}'
  tags:
    Description:
      en: 'The resource tag(example:{"k1":"v1","k2":"v2"}).'
```

```
    Type: Json
    AssociationProperty: Tags
  period:
    Description:
      en: The term of the reserved instance.
    Type: Number
    MinValue: 1
    MaxValue: 1
    Default: 1
  periodUnit:
    Description:
      en: The term unit of the reserved instance.
    Type: String
    AllowedValues:
      - Month
      - Year
    Default: Year
  OOSAssumeRole:
    Description:
      en: The RAM role to be assumed by OOS.
    Type: String
    Default: OOSServiceRole
RamRole: '{{ OOSAssumeRole }}'
Tasks:
  - Name: describeReservedInstances
    Action: 'ACS::ExecuteAPI'
    Description:
      en: Query purchased reserved instances.
    Properties:
      Service: ECS
      API: DescribeReservedInstances
      Parameters:
        RegionId: '{{ regionId }}'
        Tags: '{{ tags }}'
    Outputs:
      reservedInstanceIds:
        Type: List
        ValueSelector: '.ReservedInstances.ReservedInstance[] | select(.ExpiredTime
[0:11] == "{{ACS::CurrentDate}}") | .ReservedInstanceId'
  - Name: bulkyCloneReservedInstance
    Action: 'ACS::ECS::CloneReservedInstance'
    Description:
      en: Clone reserved instances.
    Properties:
      regionId: '{{ regionId }}'
      reservedInstanceId: '{{ ACS::TaskLoopItem }}'
      period: '{{ period }}'
      periodUnit: '{{ periodUnit }}'
    Loop:
      RateControl:
        Mode: Concurrency
        MaxErrors: 0
        Concurrency: 10
      Items: '{{ describeReservedInstances.reservedInstanceIds }}'
```

```
      Outputs:
        reservedInstanceIds:
          AggregateType: 'Fn::ListJoin'
          AggregateField: reservedInstanceId
    Outputs:
      reservedInstanceId:
        Type: String
        ValueSelector: reservedInstanceId
Outputs:
  reservedInstanceIds:
    Type: List
    Value: '{{ bulkyCloneReservedInstance.reservedInstanceIds }}'
```

iii. Click **Create Template**.
On the **My Templates** page, you can view the information of the created template. This information includes the name, version description, and format of the template.

For more information about how to create templates, see Create a template.

Step 2: Create a scheduled O&M task to execute the template on a regular basis

1. In the left-side navigation pane, click **Scheduled O&M**.
2. On the **Scheduled O&M** page, click **Create**.
3. On the **Create Scheduled O&M** page, configure parameters.
   i. In the Set Scheduled Task section, configure the parameters described in the following table.

Parameter	Description
Scheduled Task Type	Select **Executed Periodically**.
Execution Frequency	Specify a cron expression to execute the template at 00:00 every day. Set the fields in the cron expression to the following values: ■ Minute (0 to 59): 0 ■ Hour (0 to 23): 0 ■ Day (1 to 31): * ■ Month (1 to 12): * ■ Day (1 to 7, with 1 standing for Sunday and 7 for Saturday): * For more information about cron expressions, see Configure cron expressions.
Time Zone for Periodic Execution	Select a time zone based on your region. Example: **(GMT+08:00) Asia/Shanghai**.
End Time for Period Execution	Select a point in time at which to terminate the scheduled O&M task that executes the template. Example: `Jun 23, 2022, 18:00:00`. You can also preview the scheduled execution times in the **Upcoming Execution Time Preview** section to determine whether these times meet your business requirements.

ii. Click **Select Template**. In the Select Template section, select **My Templates** from the drop-down list and then select the template created in Step 1: Create a template in the OOS console from the template list.

iii. Click **Configure Parameters** . In the Configure Parameters section, configure the parameters described in the following table.

Parameter	Description
regionId	Select the region of the reserved instance. Example: China (Hangzhou).
tags	Enter the tag keys and values of the template. Example: ■ Tag key: k1 ■ Tag value: v1 Select **Attach Resource Tag to Execution** so that you can search for the execution by tag on the Executions page to view execution results.
period	Specify the term of the reserved instance. Default value: 1.
periodUnit	Select a term unit for the reserved instance. Default value: Year.
Permissions	Select a Resource Access Management (RAM) role. The selected role is assumed by OOS to grant it the permissions required to execute the template. **Note** Move your pointer over **View Authorization Policies** to view the permissions required to execute the template. Make sure that your selected RAM role has been granted the relevant permissions. For more information, see Modify the document and description of a custom policy and Grant permissions to a RAM role.

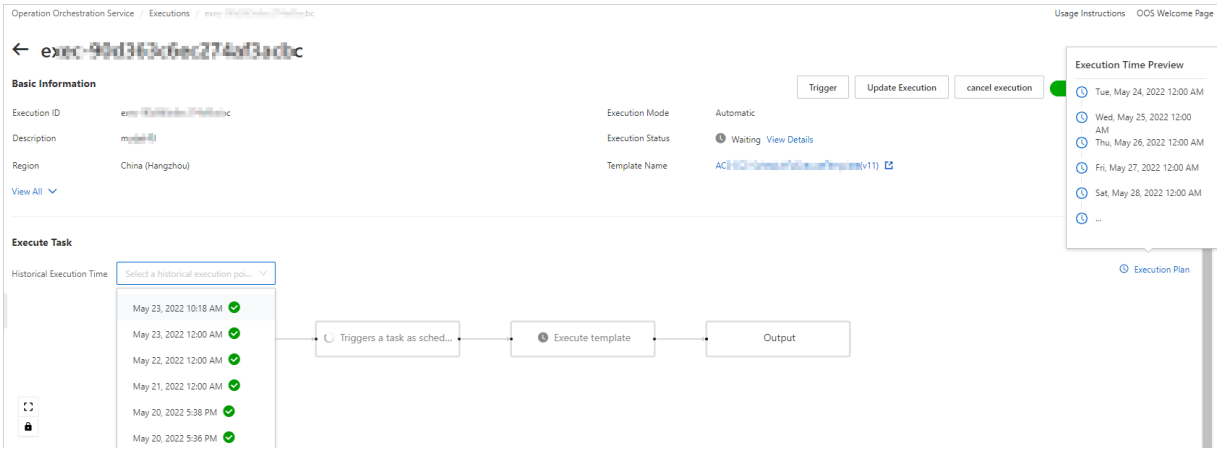
iv. *(Optional)* Click **Advanced** and configure parameters in the Advanced section.

Accept the default values for the parameters or configure the parameters based on your business requirements. For example, you can set **Tags for Execution** or **Tags for Resource Group**. You can also select Save Template to save the schedule O&M task as a template. Then, you can find the saved template on the **My Templates** page and execute this template again.

4. Click **Execute Now**.

Result

After you create the scheduled O&M task, you can click **Executions** in the left-side navigation pane and view the execution created for the task on the Executions page. If the scheduled O&M task is created, an execution is created for the task and in the Waiting state. You can also view the details of the scheduled O&M task. The task details include basic information, execution status, historical execution times, execution plan (upcoming execution times), and logs.



5.5. Savings plans

5.5.1. Overview

A savings plan is a discount plan that can be applied to offset the bills of pay-as-you-go instances, excluding preemptible instances. A combination of savings plans and pay-as-you-go instances is more flexible in use than subscription instances or a combination of reserved instances and pay-as-you-go instances.

What is a savings plan?

A savings plan is a discount plan that allows you to receive pay-as-you-go billing discounts in exchange for a commitment to use a consistent amount (measured in USD/hour) of resources over a one-year or three-year period. After you purchase a savings plan, the hourly bills of your pay-as-you-go instances are covered up to the amount of the plan.

When you use a savings plan, your pay-as-you-go instances of each instance type have a regular pay-as-you-go unit price and a savings plan unit price. For more information, see the Discount Details page. You are charged for resource usage within your commitment based on the savings plan unit price. You are charged for resource usage beyond your commitment at the regular pay-as-you-go unit price.

For example, Alex has several ecs.g6.xlarge instances in the China (Shanghai) region and can use a three-year general-purpose savings plan to obtain the following discounts.

> **Note** The prices used in this example are for demonstration only. For more information about the actual prices and discounts, see the Pricing tab on the Elastic Compute Service page and the Discount Details page.

For example, the regular pay-as-you-go unit price of ecs.g6.xlarge instances is `USD 0.155/instance/hour`, and a three-year savings plan provides savings of `54.5%` off the pay-as-you-go price for the ecs.g6 instance family in the China (Shanghai) region. The savings plan unit price of the ecs.g6.xlarge instances is calculated based on the following formula: `USD 0.155/instance/hour × 0.455 = USD 0.0705/instance/hour`.

If Alex makes a commitment of `USD 0.31/hour`, the savings plan can be applied to offset the hourly bills of 4.397 pay-as-you-go ecs.g6.xlarge instances, which is calculated based on the following formula: `0.31/0.0705 = 4.397`.

The following table compares the regular pay-as-you-go prices and the savings plan prices.

Billing method	The first hour (based on the assumption that six instances are running)	The second hour (based on the assumption that five instances are running)	The third hour (based on the assumption that four instances are running)
Total regular pay-as-you-go price without applying the savings plan	6 × 0.155 = 0.93 USD	5 × 0.155 = 0.775 USD	4 × 0.155 = 0.62 USD
Total price after the savings plan is applied	You are charged for the instances that exceed the maximum number of instances (4.397 instances) to which the savings plan can be applied on a pay-as-you-go basis. 0.31 + 0.155 × (6 - 0.31/0.0705) = 0.558 USD	You are charged for the instances that exceed the maximum number of instances (4.397 instances) to which the savings plan can be applied on a pay-as-you-go basis. 0.31 + 0.155 × (5 - 0.31/0.0705) = 0.403 USD	The total price is calculated based on the commitment because the number of running instances is less than the maximum number of instances (4.397 instances) to which the savings plan can be applied. 0.31 USD

Select an hourly commitment

For each savings plan, you must select an **hourly commitment**. You are charged for resource usage within your commitment based on the savings plan unit price. You are charged for resource usage beyond your commitment at the regular pay-as-you-go unit price.

You can purchase a desired savings plan on the Savings Plan page or purchase a recommended savings plan on the Recommended page.

Savings plan types

Savings plans are available in two types: general-purpose and Elastic Compute Service (ECS) compute. The following table compares the two types of savings plans. ECS compute savings plans are less flexible but offer higher discounts than general-purpose savings plans.

Comparison item	General-purpose	ECS compute
Use across services	Can be used across services and supports the following services and resources: • ECS: instance computing resources (vCPUs and memory), system disks, and public bandwidth • Elastic Container Instance: instance computing resources (vCPUs and memory)	Can be used only for the following ECS resources: instance computing resources (vCPUs and memory), system disks, and public bandwidth.
Region limits	Has no limits on regions.	Can be applied only within a single region.

Comparison item	General-purpose	ECS compute
Instance limits	Has no limits on instance families, instance sizes, or operating systems.	Can be applied only to specific instance families. You can choose to place an order by instance family set. In this case, this savings plan type can be applied to all instance families in the set. You can also place an order by instance family. In this case, this savings plan type can be applied only to the specified instance family. Has no limits on instance sizes or operating systems.

Scenarios

Both savings plans and reserved instances can be applied to offset the bills of pay-as-you-go instances. The following table describes scenarios where a combination of savings plans and pay-as-you-go instances serves as an optimal billing solution.

Scenario	Combination of savings plans and pay-as-you-go instances	Combination of reserved instances and pay-as-you-go instances
You want to change the instance family of your instances for business adjustment purposes, or you want to upgrade your instances to a next-generation instance family.	Has no limits on instance sizes or operating systems. General-purpose savings plans support more services and have no limits on regions or instance families.	Does not support changes to the instance family.
You want to deploy your instances in multiple regions.		A reserved instance can be used only within a single region.
You want to make it easy to select resources in the budgeting phase.	Requires an estimate of the approximate usage range but not details about the instance family or the operating system.	Requires details about the region, instance family, and operating system.

For more information about instance billing, see Instance types.

Billing

Savings plans support three payment options: All Upfront, Partial Upfront, and No Upfront. You can receive different discounts based on the selected payment option of your savings plan. Your savings plans are automatically applied to offset the bills of your pay-as-you-go instances based on specific rules within the terms of the plans. For more information, see Savings plans.

Lifecycle management

The following rules apply to the lifecycle management of savings plans:

- Management process from taking effect to expiration

  When a savings plan is purchased, the savings plan immediately takes effect at the hour of purchase. Savings plans take effect and expire on the hour.

  For example, you purchased a one-year savings plan at 13:45:00 of May 29, 2020. The savings plan takes effect at 13:00:00 of May 29, 2020 and expires at 24:00:00 of May 30, 2021. If you have eligible pay-as-you-go instances, the savings plan is applied to offset the bills of your pay-as-you-go instances starting from 13:00:00 of May 29, 2020 by hour until it expires.

  > **Notice** Expired savings plans cannot continue to offset the bills of pay-as-you-go instances. However, the pay-as-you-go instances are not released, which ensures that your business can continue. Make sure that you have sufficient balance within your account to ensure that the services on your pay-as-you-go instances are available.

- Overdue payment, suspension, payment, and resumption

  If you select the Partial Upfront or No Upfront payment option and you are unable to pay your hourly commitment because you have overdue payments within your account, the overdue payment management procedure is triggered. 72 hours after a payment within your account becomes overdue, your savings plans are suspended. The discounts of your savings plans are no longer applied starting from the next hour and cannot be resumed until the overdue bills are paid.

  > **Notice**
  > - If your savings plans are suspended, you are still charged by hour based on the commitment.
  > - If your savings plans are suspended multiple times or for an extended period of time, you may not be able to use the No Upfront payment option for other Alibaba Cloud services. Make sure that you have sufficient balance within your account.

- Unsubscription

  You cannot unsubscribe from your savings plans on your own. If you no longer need your purchased savings plans, submit a ticket .

Limits

The following table describes the limits of savings plans.

Object	Item	Description
Savings plan	Maximum allowable number	You can purchase up to 40 savings plans for each account.
	Order of application	When multiple discount plans take effect, they are applied in the following order: 1. Reserved instances and resource plans 2. Savings plans 3. Coupons 4. Vouchers

Object	Item	Description

ECS instance	Billing method	• Savings plans are applicable only to pay-as-you-go instances, excluding preemptible instances.
		• Savings plans are applicable to the following resources:
		  ◦ ECS instance computing resources (vCPUs and memory)
		  ◦ System disks
		  ◦ Public bandwidth

	Instance family	ECS compute savings plans can be applied only to pay-as-you-go instances of a specific instance family within a specific region.

Use across accounts

If you want to manage the financial relationships of multiple Alibaba Cloud accounts in a centralized manner, you can use the corporate finance feature in the User Center. The corporate finance feature allows you to establish trusteeship between multiple Alibaba Cloud accounts. You can use the main account to pay the bills of linked accounts. After the trusteeship is established between multiple accounts, you can share savings plans within the main account to offset the bills of pay-as-you-go instances within the linked accounts.

> **Note** In corporate finance, the main account and linked accounts are independent Alibaba Cloud accounts that are used to grant financial management permissions. The relationships between these accounts are different from those between regular Alibaba Cloud accounts and Resource Access Management (RAM) users.

The following limits apply to the use of savings plans across accounts:

- Savings plans within linked accounts cannot be shared.
- If you want to use savings plans within a linked account, the original savings plans within the linked account take priority over shared ones.

References

- Savings plans
- Purchase and apply savings plans
- Billing FAQ

5.5.2. Purchase and apply savings plans

This topic describes how to calculate and select an hourly commitment to purchase appropriate savings plans, and how to view the results of applying savings plans.

Purchase savings plans

You can purchase a savings plan recommended by the system or purchase a savings plan based on your own consumption trends.

- Purchase a system-recommended savings plan

If you want to optimize the costs of your pay-as-you-go instances, you can go to the Recommended page for system recommendations. You need to configure only the savings plan type, duration, and payment option, and the system recommends a savings plan with an appropriate hourly commitment based on your configurations. The recommendation feature is available for Elastic Compute Service (ECS) instances and elastic container instances.

- Select and purchase a savings plan on your own

  If you have not created pay-as-you-go instances, you can calculate an hourly commitment before you purchase a savings plan on the Savings Plans page.

  For more information about how to calculate and select an appropriate hourly commitment, see the following Select an hourly commitment for ECS instances and Select an hourly commitment for elastic container instances sections in this topic. The prices provided in this topic are for reference only. The actual prices displayed on the buy page prevail.

The following table describes the attributes of savings plans.

Attribute	Description
Savings plan type	Savings plans are classified into the following two types: • General-purpose savings plan: General-purpose savings plans can be automatically applied to eligible pay-as-you-go instances regardless of region, instance family, instance size, or operating system. General-purpose savings plans are more flexible than ECS compute savings plans and can offset up to 72% of the pay-as-you-go bill amount. • ECS compute savings plan: ECS compute savings plans can be applied to only pay-as-you-go instances of a specified instance family within a specified region regardless of instance size or operating system. ECS compute savings plans are more cost-effective than general-purpose savings plans and can offset up to 76% of the pay-as-you-go bill amount.
Region	When you select ECS compute savings plans, you must specify a region.
Purchase method	When you select ECS compute savings plans, you must specify a purchase method.
Instance family	When you select ECS compute savings plans, you must specify an instance family.
Hourly commitment	The hourly usage amount to which you commit. Minimum value: USD0.01 /hour. Resource usage within your commitment is calculated and offset based on the savings plan unit price. Resource usage beyond your commitment is billed at the regular pay-as-you-go unit price. **Notice** When you select the hourly commitment, use the savings plan discount price for calculation. For information about specific discounts, see the Discount Details page.
Payment option	Three payment options are provided: All Upfront, Partial Upfront, and No Upfront. The All Upfront payment option provides the largest discount.
Duration	You can choose a duration of one or three years. A duration of three years provides the largest discount.

Apply savings plans

After you purchase a savings plan, the savings plan is automatically applied to your pay-as-you-go instances. For information about the application rules, see the "Billing methods" section in Savings plans.

You can go to the Savings Plan page to view the information about how your savings plans are applied, including the saving amount, plan details, usage, and coverage.

Select an hourly commitment for ECS instances

For example, assume that John plans to purchase a three-year general-purpose savings plan with the All Upfront payment option and apply the plan to the following ECS instances.

> **Note** The calculation procedure described in this example only shows how to calculate the required hourly commitment. It does not mean that subsequent resources must belong to the specified instance family within the specified region.

Instance	Region	Instance type	System disk	Network bandwidth	Eligible quantity
Instance A	China (Shanghai)	ecs.g6.xlarge	40 GiB ESSD PL0	3 Mbit/s	15
Instance B	China (Beijing)	ecs.c5.large	40 GiB ESSD PL0	3 Mbit/s	5

Perform the following operations to calculate an appropriate hourly commitment:

1. Go to the Pricing tab of the Elastic Compute Service page and the Discount Details page to obtain the regular pay-as-you-go and savings plan prices of Instance A.

   The following table describes the prices of a single instance.

Billable item	Pay-as-you-go price (USD/hour)	Savings plan discount	Savings plan price (USD/hour)
ECS instance type (computing resources)	0.155	54.5% off	0.0705
System disk	0.0064	58.8% off	0.0026
Network bandwidth	0.054	57.5% off	0.0229

   After the savings plan is applied, the total cost of Instance A is calculated based on the following formula: (0.0705 + 0.0026 + 0.0229) × 15 = USD 1.44/hour.

2. Query and calculate the prices of Instance B in the same way.

   The following table describes the prices of a single instance.

Billable item	Pay-as-you-go price (USD/hour)	Savings plan discount	Savings plan price (USD/hour)
ECS instance type (computing resources)	0.1	72.7% off	0.0273

Billable item	Pay-as-you-go price (USD/hour)	Savings plan discount	Savings plan price (USD/hour)
System disk	0.0064	58.8% off	0.0026
Network bandwidth	0.054	57.5% off	0.0229

After the savings plan is applied, the total cost of Instance B is calculated based on the following formula: (0.0273 + 0.0026 + 0.0229) × 5 = USD 0.264/hour.

3. Calculate an appropriate hourly commitment, which is the sum of total costs of Instance A and Instance B.

   In this example, the recommended hourly commitment is calculated based on the following formula: 1.44 + 0.264 ≈ USD 1.70/hour.

Select an hourly commitment for elastic container instances

For example, assume that William wants to optimize the costs for his elastic container instances. The hourly bill of the elastic container instances is USD 8/hour. William wants to purchase a three-year general-purpose savings plan with the All Upfront payment option.

1. Go to the Discount Details page to obtain the discounts for elastic container instances.

   In this example, the savings plan discount is 54.5%.

2. Calculate an appropriate hourly commitment.

   In this example, after the savings plan is applied, the hourly cost for per elastic container instance is calculated based on the following formula:8 × 0.455 = USD 3.64/hour. Therefore, when William purchases a three-year general-purpose savings plan with the All Upfront payment option, an hourly commitment of USD3.64 /hour is recommended to offset the existing hourly bills of the elastic container instances.

References

- Savings plans
- Billing FAQ

5.6. Resource assurances

5.6.1. Overview

Resource Assurance is a service that guarantees the provision of Elastic Compute Service (ECS) resources as your needs change. It allows you to quantify the amount of available resources, reserve resources, plan private pools, and gain a better experience when you query, reserve, purchase, and use resources.

Introduction

Resource Assurance provides the Quota Management, Resource Reservation, Private Pool and Privilege services.The following table describes these services and their features.

Service	Feature	Description
Quota Management	Elastic Quota	Shows quotas on resources such as instance types, images, disks, and security groups, and guarantees the provision of instance resources.
Resource Reservation	• Elasticity Assurance • Capacity Reservation	Allows you to reserve resources for different scenarios. After you purchase an elasticity assurance or capacity reservation, Alibaba Cloud reserves resources that match the attributes of the elasticity assurance or capacity reservation as a private pool.
Private Pool	Private Pool	Provides guaranteed access to resources for you to create instances.
Privilege	Privilege	Allows you to view the privileges that your account has, including feature privileges and network privileges.

Quota Management

The following table describes the different types of elastic quotas.

Elastic quota type	Description
Instance quota	Instance quotas are allocated based on zones, instance types, billing methods, and network types. Instance quotas are classified into the following types based on how well the provision of resources is guaranteed: • Base quota: specifies the minimum amount of guaranteed instance resources. You can create instances within a base quota with a high success rate. Base quotas are adjusted and allocated before the tenth day of each month based on the usage of your ECS resources. You cannot apply to increase the base quotas. • Reserved quota: specifies the amount of instance resources reserved by resource reservations. When you create instances within a reserved quota, you have guaranteed access to the resources that you request. You can create resource reservations to increase reserved quotas. • Total quota: specifies the maximum amount of instance resources for which access can be guaranteed. A total quota includes a base quota, a reserved quota, and other quotas. The system periodically adjusts your instance quotas based on your instance usage to ensure that growing demands can be met. If a total quota is insufficient for your demands, you can apply to increase it. For more information, see View and increase instance quotas.
Resource quota	Resource quotas are quotas on other ECS resources such as images, disks, and security groups. You can apply to increase these quotas. For more information, see View and increase resource quotas.

Resource Reservation

You can purchase resource reservations to gain guaranteed access to resources. Resource reservations are classified into the following types based on use scenarios:

- Elasticity assurances: are suitable for irregular peak and off-peak demands for resources. Elasticity

assurances offer guaranteed resources only for pay-as-you go instances. Elasticity assurances can take effect immediately after creation or at the specified time. For more information, see Overview of Elasticity Assurance

- Capacity reservations: are suitable for stable and large demands for resources. Capacity reservations have different types. Capacity reservations offer guaranteed resources for subscription and pay-as-you go instances. Capacity reservations can take effect immediately after creation or at the specified time. For more information, see Overview of Immediate Capacity Reservation.

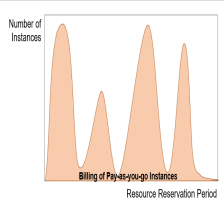
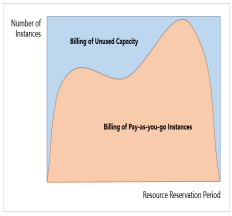
> **Note** Resource reservations do not offer guaranteed resources for preemptible instances.

The following figure shows the classification of resource reservations.

res-reserve

When you use resource reservations, you can make different resource reservation plans based on your requirements on use scenarios, resource types, and effective time. The following table compares these plans.

Item	Elasticity assurance	Immediate capacity reservation	Capacity reservation with Savings Plan	Capacity reservation for subscription resources

Item	Elasticity assurance	Immediate capacity reservation	Capacity reservation with Savings Plan	Capacity reservation for subscription resources
Scenario	Resource usage shows discrete spikes over time, but the overall resources usage is not high, as shown in the following figure. Instances need to be released and then created from time to time.  Elasticity assurances are suitable for the following scenarios: • A financial Software as a Service (SaaS) service provider needs a large number of resources to perform account checks at the beginning of each month. • A rendering enterprise needs to process a number of rendering tasks at the beginning of each week. • An Internet media enterprise needs to report hot news that appears from time to time.	The overall resource usage is stable and high, as shown in the following figure. Instances need to be released and then created from time to time.  Immediate capacity reservations are suitable for the following scenarios: • System upgrades and cluster deployment. • Guaranteed provision of resources for existing regional reserved instances or savings plans.	The overall resource usage is stable and high, as shown in the following figure. Instances need to be released and then created from time to time. sp-acr Capacity reservations with Savings Plan are suitable for scenarios in which data is migrated to the cloud. You must reserve resources in advance and upgrade systems from time to time after migration. Instances need to be released and then created.	The overall resource usage is stable and high, as shown in the following figure. Instances do not need to be released and then created. ocr Capacity reservations for subscription resources are suitable for scenarios in which resources need to be filed.
Resource	Pay-as-you-go instances	Pay-as-you-go instances	Pay-as-you-go instances	Subscription instances

Item	Elasticity assurance	Immediate capacity reservation	Capacity reservation with Savings Plan	Capacity reservation for subscription resources
Flexibility in time	• Effective time: Elasticity assurances can take effect immediately after creation or at the specified time. • Use time: You do not need to plan when to use resources over an extended period of time but create instances as needed. • Release time: Elasticity assurances can be released only upon expiration.	• Effective time: Immediate capacity reservations take effect only immediately after creation. • Use time: To make resources cost-effective, you must plan when to use resources and continuously use them over an extended period of time. • Release time: Immediate capacity reservations can be manually released or released upon expiration.	• Effective time: Capacity reservations with Savings Plan take effect only at the specified time. • Use time: To make resources cost-effective, you must plan when to use resources and continuously use them over an extended period of time. • Release time: Capacity reservations with Savings Plan can be manually released (they must be in the Active state) or released upon expiration.	• Effective time: Capacity reservations for subscription resources take effect only at the specified time. • Use time: You must plan when to use resources and create subscription instances within the validity period of the capacity reservation. • Release time: Capacity reservations for subscription resources can be released only upon expiration.
Billing	• Assurance fee. • Price of created pay-as-you-go instances.	• Price of unused reserved capacity (which is billed at the pay-as-you-go instance rate). • Price of created pay-as-you-go instances.	• Price of saving plans. • Price of unused reserved capacity (which is billed at the pay-as-you-go instance rate and offset by the associated savings plans). • Price of created pay-as-you-go instances (which is offset by the associated savings plans).	• Price of unused reserved capacity (which is billed at the pay-as-you-go instance rate). • Price of created subscription instances.

Item	Elasticity assurance	Immediate capacity reservation	Capacity reservation with Savings Plan	Capacity reservation for subscription resources
Cost optimization	If you have purchased regional reserved instances or savings plans, you can apply them to created pay-as-you-go instances.	If you have purchased regional reserved instances or savings plans, you can apply them to the unused reserved capacity and created pay-as-you-go instances.	When you create a capacity reservation, you purchase savings plans which can be used to offset the bills of resources reserved by the capacity reservation. **Note** If you manually release a capacity reservation with Savings Plan, the savings plans can continue to offset the bills of resources, but cannot be used to create a new capacity reservation with Savings Plan.	N/A. **Note** Regional reserved instances or savings plans cannot offset the bills of the capacity reservation for subscription resources and the unused capacity.

Note Regional reserved savings plans can provide reserved resources but cannot be used to offset the bills of resources reserved by resource reservations.

Private Pool

Automatically scheduled resources constitute a public pool. All users have access to the resources within the public pool. Instances may fail to be created when resources within the public pool are insufficient. When you purchase a resource reservation, Alibaba Cloud reserves resources that have matching attributes as a private pool for your use to create instances.

Note The capacity in private pools is displayed as reserved quotas to provide an overview of guaranteed resources. For more information about reserved quotas, see View and increase instance quotas.

You can purchase resource reservations based on your planned demands to create private pools to accommodate reserved instances. Then, you can use the reserved resources in the private pools to create instances. You can perform the following private pool-related operations:

- When you purchase an elasticity assurance or capacity reservation, configure private pool attributes including the following ones:

- Private pool type: open or targeted private pool. Open private pools are suited to common business that requires guaranteed provision of resources. You can use an open private pool of your choice by specifying its ID or by matching the tags of its associated elasticity assurance or capacity reservation, or use an open private pool selected by the system. Targeted private pools are suited to key business that requires dedicated reserved resources. To use a targeted private pool, you must specify its ID.
- Tag matching: You can enable tag matching to use open private pools based on tags.

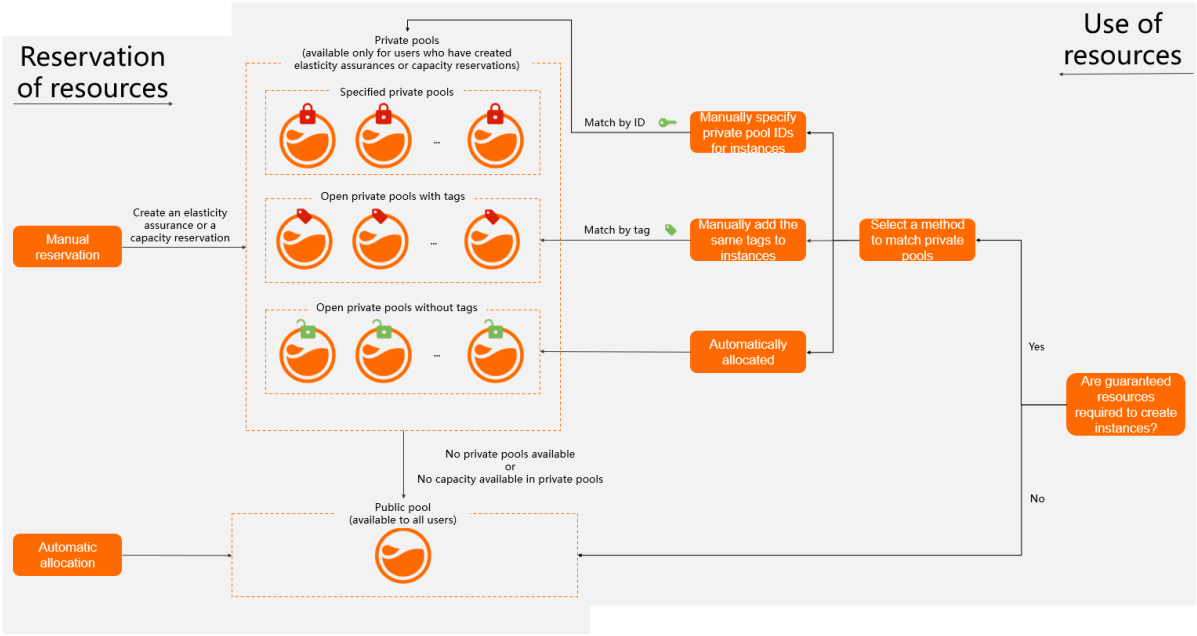
  For example, assume that you have developed a script to create instances with specific tags for your key business. If your key business requires guaranteed provision of resources, purchase an elasticity assurance or capacity reservation. When you purchase the elasticity assurance or capacity reservation, set the private pool type to Open and add the instance tags specified in the script. After the elasticity assurance or capacity reservation is purchased, tag matching is automatically enabled for the associated open private pool. When you run the script to create instances, the associated open private pool is automatically matched and used. This eliminates the need to modify the instance creation script and reduces unnecessary communication and O&M costs.

For more information, see Purchase an elasticity assurance.

- When you create instances, you can use one of the following methods to configure which private pool to use:
  - Manually specify the ID of an open or targeted private pool.
  - Manually add tags to use the open private pool associated with an elasticity assurance or capacity reservation that has the tags.
  - Set the private pool type to Open to use an open private pool selected by the system.

  For more information, see Use a private pool to create instances.

The following figure shows the workflow for creating and using private pools.



Alibaba Cloud utilizes the capacity of private pools as much as it possibly can.

- When you purchase elasticity assurances or capacity reservations, Alibaba Cloud takes the instances that you have into account to evaluate whether the current resource inventory can meet your demands.

For example, assume that you have 100 pay-as-you-go instances and want to release them to create new ones. When you attempt to purchase a capacity reservation to reserve resources, your purchase request is denied due to insufficient resources. You can set the private pool type of your existing instances to Open so that the instances are deemed as matchable resources. Then, you can purchase a capacity reservation.

- After a private pool is created, it is automatically used by existing instances that match its attributes.

  > **Note** Only private pools that apply to pay-as-you-go instances can match existing instances. Private pools that apply to subscription instances can be used only to create instances.

- If instances no longer match a private pool after instance attributes such as the instance type and operating system are changed, the system checks the instances and matches them to other private pools.

The following section describes some best practices for using private pools:

- Resources allocated based on business types

  A system administrator creates multiple targeted private pools and informs O&M personnel and developers of different pool IDs. This ensures that different private pools are used to create instances for the O&M and development purposes.

- Resources allocated regardless of business types

  A system administrator creates open private pools. O&M personnel and developers use the reserved resources in the open private pools to create instances. When the reserved resources in the private pools have been used up, resources in the public pool are used.

- Resources allocated by tag

  A system administrator creates open private pools and adds tags based on business. When you create instances for a specific business, add tags to the instances. The instances automatically match the open private pools that have the same tags.

- Resources allocated exclusively to a specific business

  A system administrator creates targeted private pools and allows instances to be created only by using these pools for a specific business. When the reserved resources in the private pools have been used up, instances cannot be created.

Privilege

You can view the privileges that your account has, including feature privileges and network privileges.

- Feature privileges: include the privileges on the instance configuration downgrade, image copy, image import, and image export features. If the privileges on a feature are not displayed, you are not granted the privileges.
- Network privileges: allow you to check whether the classic network is available in a specified region.

For more information about how to view privileges, see View privileges.

5.6.2. Resource reservation

5.6.2.1. Overview of Elasticity Assurance

Elasticity Assurance offers guaranteed resources to flexibly meet your daily requirements. If you have consistent resource requirements, we recommend that you use Capacity Reservation.

Introduction

Elasticity Assurance allows you to pay a small assurance fee for guaranteed access to resources for a duration of one month up to five years. When you purchase an elasticity assurance, you must specify attributes such as zone and instance type. The system generates a private pool for the created elasticity assurance to reserve resources that match the specified attributes. For example, you can purchase an elasticity assurance to reserve resources of the ecs.c6.large instance type in Hangzhou Zone I. You can have guaranteed access to the reserved capacity in the private pool to create pay-as-you-go instances.

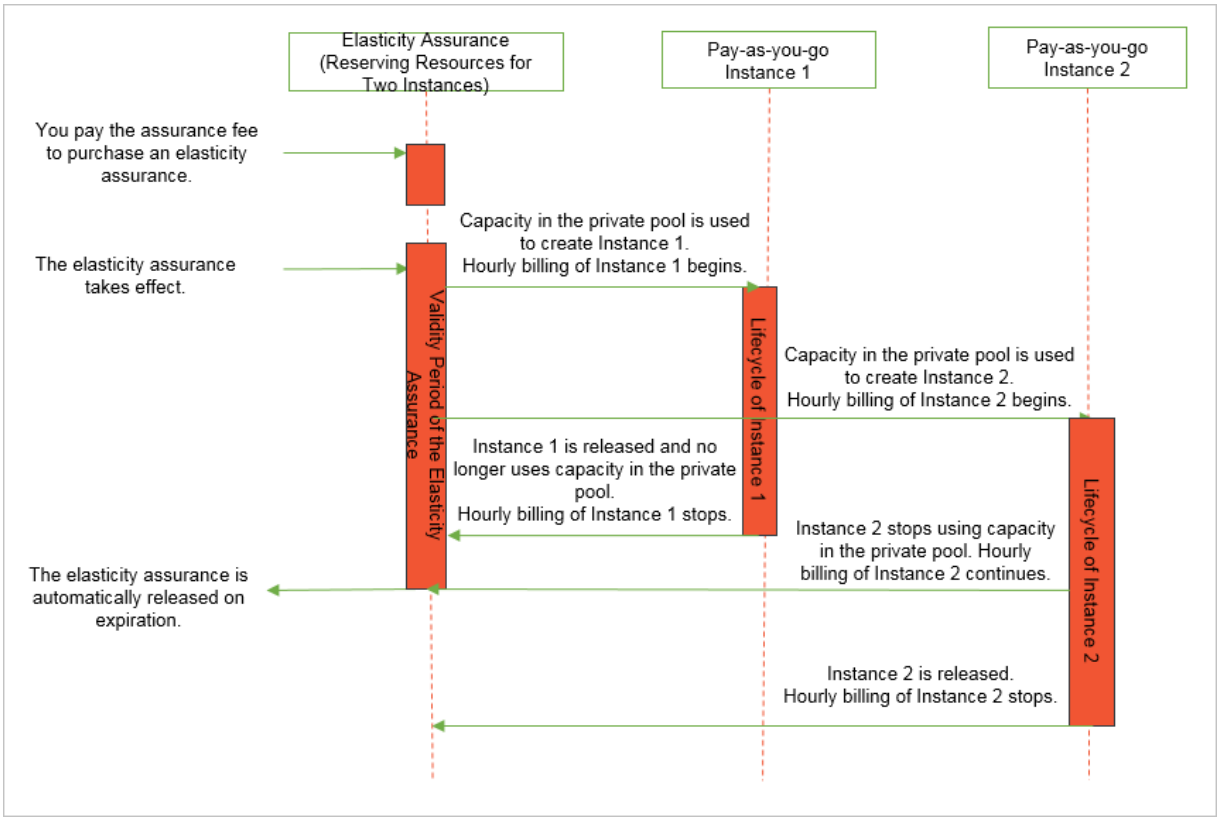
> **Note** Elasticity Assurance offers guaranteed resources only to create pay-as-you go instances but not preemptible instances.

An elasticity assurance transitions through the following phases during its lifecycle:

1. The elasticity assurance is created after you pay for its upfront costs (assurance fee).
2. At any time during the assurance period (the validity period of the elasticity assurance), you can use the reserved capacity in the private pool associated with the elasticity assurance to create pay-as-you-go instances.
3. The elasticity assurance is automatically released when it expires.

   > **Note** Created pay-as-you-go instances are not affected when the associated elasticity assurance is released, and continue to run normally. The instances are billed at the pay-as-you-go rate after they are created.

The following figure shows how an elasticity assurance that reserves resources for two instances is used.



Benefits

- Low-cost guaranteed provision of resources: You can purchase elasticity assurances at low costs to reserve resources for specified durations. During these durations, you can use the reserved resources to create pay-as-you-go instances that have matching attributes.
- Flexibility in resource use time: Elasticity assurances provide guaranteed access to resources and allow you to create and release pay-as-you-go instances within the reserved capacity at any time over an extended period of time.
- Use in conjunction with discount plans: Pay-as-you-go instances created from reserved resources in private pools can match savings plans or regional reserved instances to benefit from the billing discounts.

Billing

Purchased elasticity assurances cannot be manually released. When you use an elasticity assurance, you must pay the following fees:

- Assurance fee generated when you create the elasticity assurance
- Hourly fees of the pay-as-you-go instances that were created from the reserved resources of the elasticity assurance

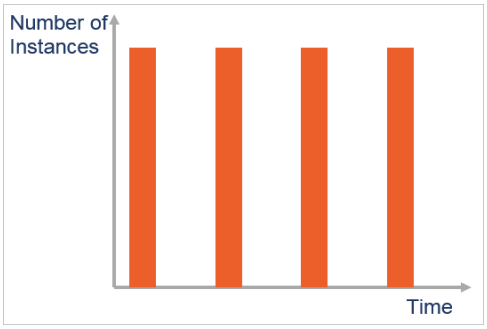
> **Note** If you have purchased applicable savings plans or regional reserved instances, you can apply them to the pay-as-you-go instances.

Limits

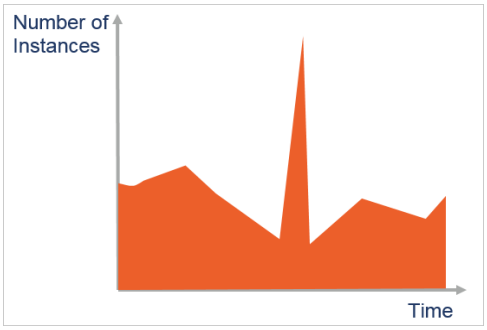
- Elasticity Assurance is available only for specific instance types in specific regions. For more information, see the buy page.
- You cannot cancel elasticity assurances or release them before they expire.
- Reserved resources in a private pool can be used to create only pay-as-you-go instances of the same instance type in the same zone as the associated elasticity assurance.
- Zonal reserved instances cannot be applied to the pay-as-you-go instances that were created from reserved resources in private pools.

Scenarios

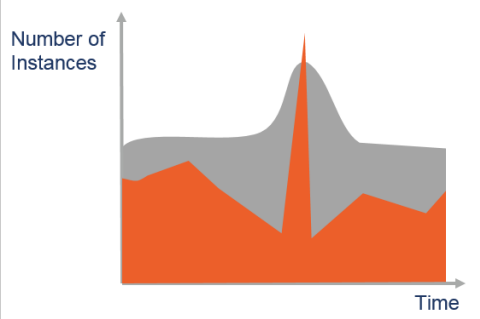
- Periodic short-term resource requirements: Elasticity Assurance guarantees the provision of resources when you want to scale your computing resources during a fixed period of time every day, week, or month. Your business is affected if no sufficient resources are available during that period of time. However, you have only small resource requirements during the other periods of time, and consume a small amount of resources in total. In this context, discount plans such as reserved instances are underused. For example, a financial SaaS service provider requires a large amount of resources to perform an account check at the beginning of each month, or a rendering enterprise needs to process a number of rendering tasks at the beginning of each week.



- Occasional large resource requirements: Elasticity Assurance enables you to reserve resources for urgent use so that you can have fast access to resources and ensure business continuity in case of unexpected events. For example, an Internet media company that needs to occasionally report breaking news or enterprises that need to reserve resources for disaster recovery are scenarios for which Elastic Assurance is suited.



- Resource guarantee during special periods: In high-traffic periods such as Double 11 and Spring Festival when resources are strained, Elasticity Assurance helps ensure that key business runs smoothly and avoid risks caused by resource contention. For example, resources required for key business such as live video streaming, ticket-grabbing, and giveaways need to be guaranteed.



Best practices

If a company is attempting to run computing tasks while resources are insufficient, their business is significantly affected and they may encounter the following issues:

- In use of preemptible instances, clusters may fail to be scaled out due to insufficient resources in a zone.
- They use pay-as-you-go instances in place of preemptible instances for guaranteed provision of resources, which results in cost increases.
- In high-traffic periods such as Double 11, it may not be possible to create pay-as-you-go instances due to unavailability of resources. This failure may affect their key business.

Solutions:

1. Make a resource requirements plan and select zones where resources are sufficient.
2. Purchase zonal reserved instances and elasticity assurances to reduce pay-as-you-go instance costs while guaranteeing the provision of resources.
   - Zonal reserved instances can cover resources in consistent use to reduce pay-as-you-go instance costs.
   - Elasticity assurances can be used to meet daily requirements for elastic resources and reserve resources at lower costs than reserved instances do. Elasticity assurances provide guaranteed resources for your exclusive use to create pay-as-you-go instances even when resources are strained.
   - You can choose to create pay-as-you-go instances only for key business by using reserved resources in private pools.

5.6.2.2. Overview of Immediate Capacity Reservation

You can purchase immediate capacity reservations to reserve and lock down capacity for pay-as-you-go instances. Immediate capacity reservations take effect as soon as they are created, and are applicable to scenarios where resource requirements are large.

Introduction

You can reserve capacity at any time by purchasing an immediate capacity reservation. After you purchase an immediate capacity reservation, the specified resources are reserved and locked down for your exclusive use. As soon as an capacity reservation takes effect, the reservation begins to be billed at the pay-as-you-go instance rate . Billing continues regardless of whether the capacity reservation is actually used to create instances. Billing stops when the capacity reservation is released.

When you purchase an immediate capacity reservation, you must specify attributes such as zone, instance type, and operating system. The system generates a private pool in which to reserve resources that match the specified attributes. You have guaranteed access to the reserved capacity in the private pool to create pay-as-you-go instances.

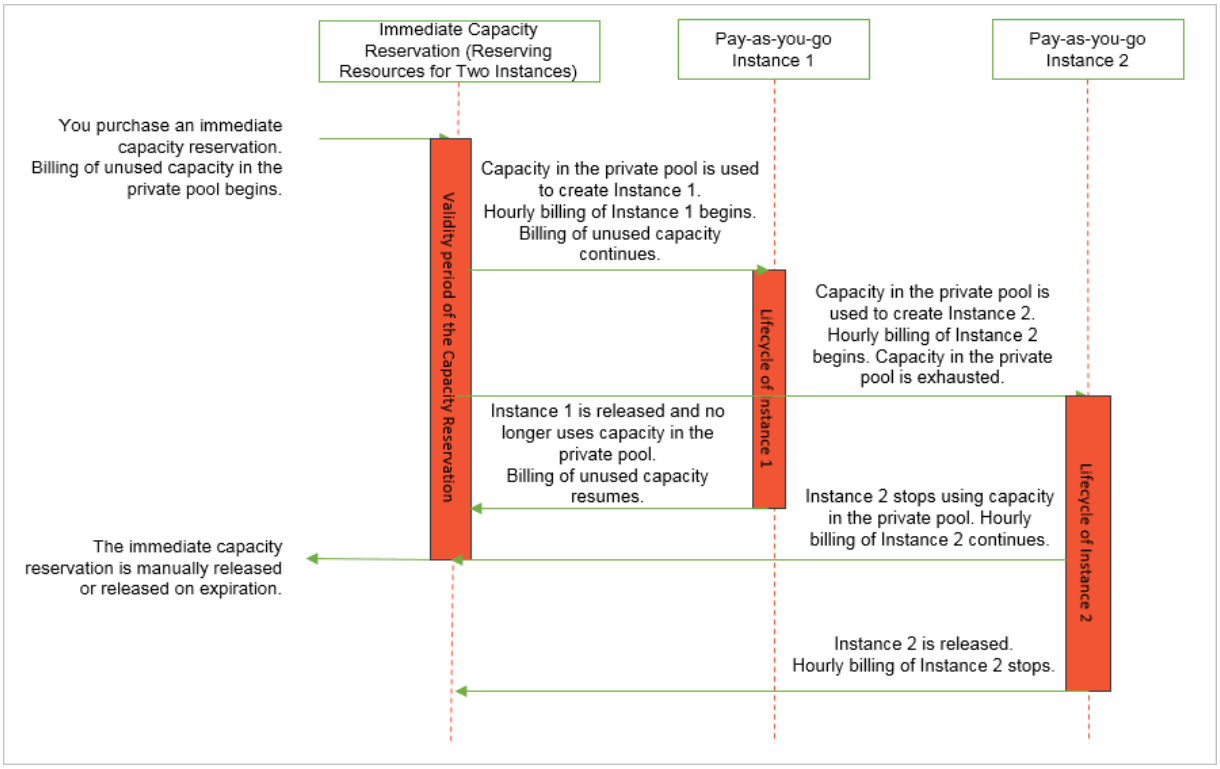
> **Note** Immediate Capacity Reservation offers guaranteed resources only to create pay-as-you go instances. Preemptible instances are not supported.

An immediate capacity reservation transitions through the following phases throughout its lifecycle:

1. As soon as the capacity reservation is purchased, it begins to be billed at the pay-as-you-go instance rate.
2. At any time during the validity period of the capacity reservation, you can use the reserved capacity in the private pool associated with the capacity reservation to create pay-as-you-go instances.
3. You can manually release the capacity reservation or wait for it to expire and be automatically released.

> **Note** Created pay-as-you-go instances are not affected when the associated immediate capacity reservation is released and continue to run as expected. The instances are billed at the pay-as-you-go rates after they are created.

The following figure shows how an immediate capacity reservation that reserves resources for two instances is used.



Billing

An immediate capacity reservation begins to be billed at the pay-as-you-go instance rate as soon as the capacity reservation is purchased. Billing continues regardless of whether the capacity reservation is actually used to create instances. Billing stops when the capacity reservation is automatically released upon expiration or manually released.

> **Note** Before the capacity reserved by an immediate capacity reservation is used to create pay-as-you-go instances, you are charged only for the instance type. After the capacity reserved by the capacity reservation is used to create pay-as-you-go instances, you are charged based on the instance configurations including the instance type, disks, and public bandwidth.

Savings plans and regional reserved instances can be applied to offset the hourly bills of the pay-as-you-go instances associated with immediate capacity reservations, but zonal reserved instances cannot.

Limits

- Immediate Capacity Reservation is available only for specific instance types in specific regions. For more information, see the buy page.
- The reserved capacity in a private pool can be used to create only pay-as-you-go instances that match the instance type, zone, and operating system attributes of the associated capacity reservation.
- Zonal reserved instances cannot be applied to the pay-as-you-go instances that were created from reserved resources in private pools.

Scenarios

- Large requirements for elastic resources: When an immediate capacity reservation is purchased, billing starts based on the pay-as-you-go rate of the instance type. To use resources in the most cost-effective manner, you must make continuous and full use of the reserved resources during the validity period of the capacity reservation.
- Requirements for reserved resources in use of savings plans or regional reserved instances: Savings plans and regional reserved instances can significantly reduce costs but cannot provide resource reservations. To reserve resources when you have savings plans or regional reserved instances, you can purchase immediate capacity reservations.

Usage examples

- Manual release

  Requirements:

  - Apply one regional reserved instance to offset the hourly bills of pay-as-you-go instances. The reserved instance is an ecs.g6.large Linux reserved instance in the China (Hangzhou) region and can be applied to 10 ecs.g6.large instances.
  - Immediately reserve resources in Hangzhou Zone H and Hangzhou Zone I.
  - Switch to other zones to create pay-as-you-go instances based on business needs.

  Solution:

  - Create an immediate capacity reservation that can be manually released in both Hangzhou Zone H and Hangzhou Zone I. Make sure that the created immediate capacity reservations match the instance type, instance quantity, and operating system attributes of the regional reserved instance.

- To reserve resources in other zones, manually release the immediate capacity reservations in Hangzhou Zone H and Hangzhou Zone I and create immediate capacity reservations within the zones where you want to reserve resources.

- Scheduled release

  Requirements:

  - Apply a general-purpose savings plan that you purchased to offset the hourly bills of pay-as-you-go instances.
  - Reserve resources of different instance types for specific periods of time. Reserve resources of the ecs.c6e.large instance type for the first half of a month and resources of the ecs.c6.large instance type for the second half of the month.

  Solution:

  - At the beginning of the month, purchase immediate capacity reservations for the ecs.c6e.large instance type and schedule the reservations to be released at the specified time.
  - In the middle of the month, purchase immediate capacity reservations for the ecs.c6.large instance type and schedule the reservations to be released at the specified time.

5.6.2.3. Purchase an elasticity assurance

When you purchase an elasticity assurance, you must specify attributes such as instance type and zone. Alibaba Cloud reserves resources that have matching attributes in a private pool for your use to create pay-as-you-go instances.

Context

When you purchase an elasticity assurance, you must pay a small assurance fee to gain guaranteed access to resources. After you use the reserved capacity in the associated private pool to create pay-as-you-go instances, you are charged for the instances on an hourly basis at the pay-as-you-go rate. The actual price of an elasticity assurance is displayed on the elasticity assurance buy page.

Procedure

1.
2.
3.
4. In the upper part of the Resource Assurance page, click the **Elasticity Assurance** tab.
5. Click **Purchase Elasticity Assurance**.
6. In the **Query Solutions** step, configure parameters and click **Next: Confirm Information**.

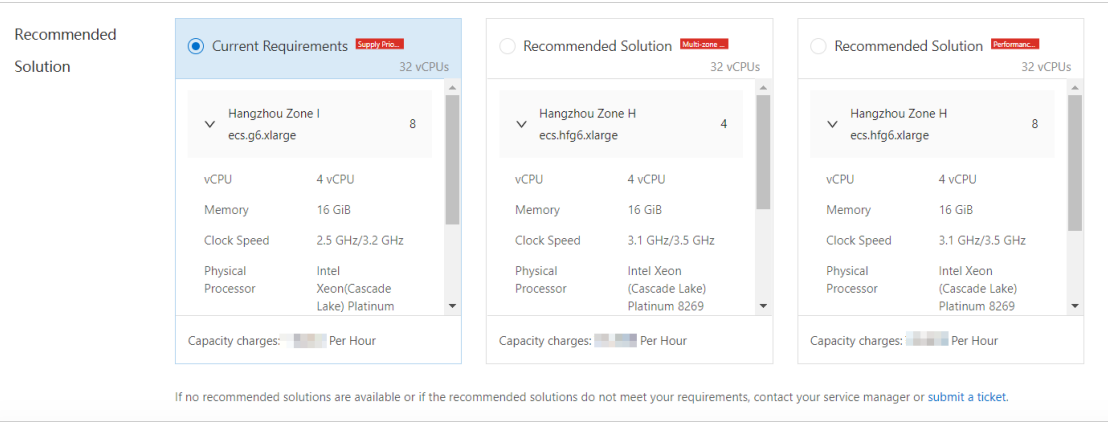
i. Configure the attributes of the elasticity assurance.

An elasticity assurance can be used to create only pay-as-you-go instances that match the attributes of the elasticity assurance, such as instance type and zone. The following table describes parameters used to specify the attributes of an elasticity assurance.

Parameter	Description
Region and Zone	The region and zone in which to reserve resources. In the Recommended Solution section, other zones may be recommended based on resource availability. The zone in the solution that you select is used.
Instance Type	The instance type for which to reserve resources. In the Recommended Solution section, other alternative instance types may be recommended based on resource availability. The instance type in the solution that you select is used.
Reserved Quantity	The number of instances for which to reserve resources.
Billing Method	The billing method of instances. After an elasticity assurance is purchased, you can use the reserved capacity in the associated private pool to create only **pay-as-you-go** instances.
Effective At	The time at which the elasticity assurance takes effect. The end time of the validity period cannot be specified but is automatically changed with the specified **Duration** value.
Duration	The validity period of the elasticity assurance. You can select a duration of one month up to five years. The durations available for selection are displayed on the elasticity assurance buy page.

ii. View and select a solution in the Recommended Solution section.

The recommended solutions may deliver optimal performance, be backed by the most sufficient supply of resources, or provide multi-zone disaster recovery. You can select the solution that best suits your needs.



Note If no recommended solutions are available or if the recommended solutions cannot meet your requirements, you can **submit a ticket**.

iii. Configure the private pool.

After an elasticity assurance is purchased, Alibaba Cloud reserves resources that match the attributes such as instance type and zone of the elasticity assurance in a private pool.

Parameter	Description
Private Pool Type	The private pool type. Valid values: **Open** and **Targeted**. ■ You can use one of the following methods to specify which open private pool to use: ■ Specify the ID of an open private pool to use when you create instances. ■ If tag matching is enabled for an open private pool, you can add the tags of the associated elasticity assurance when you create instances to use the open private pool. ■ If tag matching is not enabled for one or more open private pools and you select Open as the private pool type without specifying a private pool ID, the system selects for you an open private pool whose associated elasticity assurance does not have tags. ■ To use a targeted private pool to create instances, you must manually specify the ID of the private pool. We recommend that you prepare a number of open and targeted private pools based on your business types. For example, you can prepare targeted private pools dedicated to key business. For more information, see Private Pool.
Private Pool Name	The name of the private pool. The name must be 2 to 128 characters in length and can contain letters, digits, colons (:), underscores (_), periods (.), and hyphens (-). It must start with a letter and cannot start with http:// or https://.
Description	The description of the private pool. Enter an informative description for easy management. The description must be 2 to 256 characters in length and cannot start with http:// or https://.

iv. Add tags.

You can perform operations such as batch executing O&M tasks and financial sharing based on tags. For more about the use scenarios of tags and how to work with tags, see Overview.

In addition to common use scenarios, you can use open private pools in a fine-grained manner based on tags. After you add tags to an elasticity assurance, tag matching is automatically enabled for the associated private pool. If you select Open as the private pool type and add the same tags when you create instances, the associated private pool is automatically matched and used.

Notice After an elasticity assurance takes effect, its associated private pool is always matched based on the tags added when the elasticity assurance was purchased. We recommend that you do not edit the tags of purchased elasticity assurances.

7. In the **Confirm Information** step, confirm you configurations, click Purchase, and then complete the payment.
   i. View the parameters in the Overview of Elastic Security section.
   ii. Select one or more elastic assurances in the Elastic guarantee splitting section.

   Each elasticity assurance is applicable only to a specific instance type within a specific zone. If you select a solution that includes multiple instance types or zones, the solution is automatically split into multiple elasticity assurances. Select the elasticity assurances that suit your needs.
   iii. After you confirm that all configurations are correct, read and select the notes, and then click **Purchase**.
   Complete the payment as prompted.

What's next

On the **Elasticity Assurance** tab, view the elasticity assurance that you purchased. When the elasticity assurance is in the **Active** state, you can use its associated private pool to create pay-as-you-go instances. For more information, see Use a private pool to create instances.

5.6.2.4. Purchase an immediate capacity reservation

When you purchase an immediate capacity reservation, you must specify attributes such as instance type, zone, and operating system. Alibaba Cloud reserves resources that have matching attributes in a private pool for your use to create pay-as-you-go instances.

Context

An immediate capacity reservation begins to be billed at the pay-as-you-go instance rate of the specified instance type as soon as the capacity reservation is purchased, regardless of whether it is used to create pay-as-you-go instances. Billing continues until the capacity reservation is manually released or automatically released on expiration.

> **Note** Before the capacity reserved by an immediate capacity reservation is used to create pay-as-you-go instances, you are charged only for the instance type. After the capacity reserved by the capacity reservation is used to create pay-as-you-go instances, you are charged based on instance configurations including the instance type, disks, and public bandwidth. For more information about billing rules, see Pay-as-you-go.

Procedure

1.
2.
3.
4. In the upper part of the Resource Assurance page, click the **Capacity Reservation** tab.
5. Click **Purchase Capacity Reservation**.
6. In the **Query Solutions** step, configure parameters and click **Next: Confirm Information**.

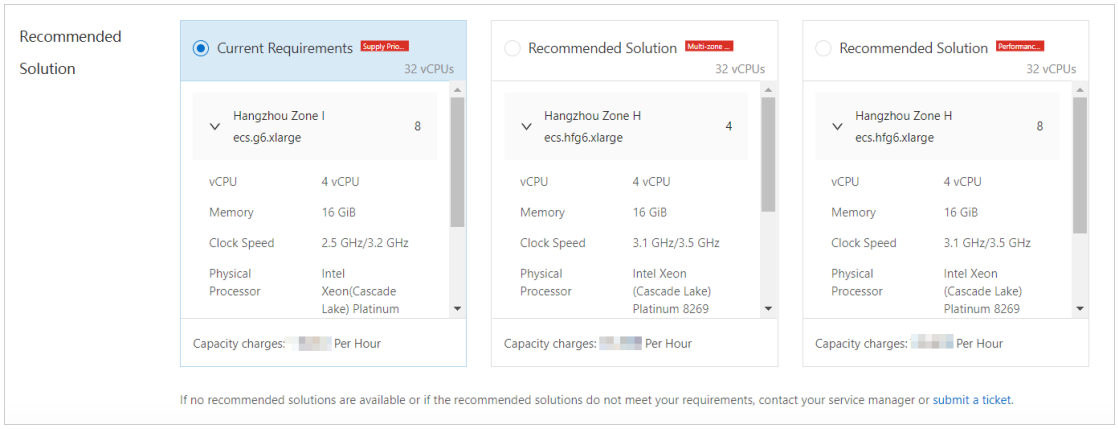
i. Configure the attributes of the immediate capacity reservation.

A capacity reservation can be used to create only pay-as-you-go instances that match the attributes of the capacity reservation, such as instance type, zone, and operating system. The following table describes parameters used to specify the attributes of a capacity reservation.

Parameter	Description
Application Mode	Only a value of **Effective Immediately** is available.
Expiration Method	After an immediate capacity reservation is released, instances that are using its associated private pool are not affected and the unused reserved capacity is no longer billed. Valid values: ■ **Manual Release**: After you purchase the immediate capacity reservation, it exists until you manually release it. ■ **Release Upon Expiration**: The immediate capacity reservation is automatically released when it expires. If you select **Release Upon Expiration**, you must set **Expired At** to specify an expiration time for the capacity reservation. The expiration time must be more than 1 hour later than the current time.
Billing Method	After an immediate capacity reservation is purchased, you can use the reserved capacity in the associated private pool to create only **pay-as-you-go** instances.
Region and Zone	Specify the region and zone in which to reserve resources. In the Recommended Solution section, other zones may be recommended based on resource availability. The zone in the solution that you select is used.
Instance Type	Specify the instance type for which to reserve resources. In the Recommended Solution section, other alternative instance types may be recommended based on resource availability. The instance type in the solution that you select is used.
Operating System	Specify the operating system type for instances. Valid values: **linux** and **windows**. If you want to use an immediate capacity reservation in conjunction with a regional reserved instance to reduce costs, make sure that their operating system types are the same.
Reserved Quantity	Specify the number of instances for which to reserve resources.

ii. View and select a solution in the Recommended Solution section.

The recommended solutions may deliver optimal performance, be backed by the most sufficient supply of resources, or provide multi-zone disaster recovery. You can select the solution that best suits your needs.



Note If no recommended solutions are available or if the recommended solutions cannot meet your requirements, you can **submit a ticket**.

iii. Configure the private pool.

After an immediate capacity reservation is purchased, Alibaba Cloud reserves resources that match the attributes (such as instance type, zone, and operating system) of the capacity reservation in a private pool.

Parameter	Description
Private Pool Type	The private pool type. Valid values: **Open** and **Targeted**. ■ You can use one of the following methods to specify which open private pool to use: ■ Specify the ID of an open private pool to use when you create instances. ■ If tag matching is enabled for an open private pool, you can add the tags of the associated capacity reservation when you create instances to use the open private pool. ■ If tag matching is not enabled for one or more open private pools and you select Open as the private pool type without specifying a private pool ID, the system selects for you an open private pool whose associated capacity reservation does not have tags. ■ To use a targeted private pool to create instances, you must manually specify the ID of the private pool. We recommend that you prepare a number of open and targeted private pools based on your business types. For example, you can prepare targeted private pools dedicated to key business. For more information, see Private Pool.
Private Pool Name	The name of the private pool. The name must be 2 to 128 characters in length and can contain letters, digits, colons (:), underscores (_), periods (.), and hyphens (-). It must start with a letter and cannot start with http:// or https://.
Description	The description of the private pool. Enter an informative description for easy management. The description must be 2 to 256 characters in length and cannot start with http:// or https://.

iv. Add tags.

You can perform operations such as batch executing O&M tasks and financial sharing based on tags. For more about the use scenarios of tags and how to work with tags, see Overview.

In addition to common use scenarios, you can use open private pools in a fine-grained manner based on tags. After you add tags to an immediate capacity reservation, tag matching is automatically enabled for the associated private pool. If you select Open as the private pool type and add the same tags when you create instances, the associated private pool is automatically matched and used.

> **Notice** After an immediate capacity reservation takes effect, its associated private pool is always matched based on the tags added when the capacity reservation was purchased. We recommend that you do not edit the tags of purchased immediate capacity reservations.

7. In the **Confirm Information** step, confirm you configurations, click Purchase, and then complete the payment.
   i. View the parameters in the Capacity reservation overview.
   ii. Select one or more capacity reservations in the Capacity reservation split section.

   Each capacity reservation is applicable only to a specific instance type within a specific zone. If you select a solution that includes multiple instance types or zones, the solution is automatically split into multiple capacity reservations. Select the capacity reservations that suit your needs.

   iii. After you confirm that all configurations are correct, read and select the notes, and then click **Purchase**.
   Complete the payment as prompted.

What's next

On the **Capacity Reservation** tab, view the capacity reservation that you purchased. When the capacity reservation is in the **Active** state, you can use its associated private pool to create pay-as-you-go instances. For more information, see Use a private pool to create instances.

5.6.2.5. View and modify an elasticity assurance

This topic describes how to view information of an elasticity assurance, such as its status, instance type, zone, resource usage of the associated private pool, and associated instances. This topic also demonstrates how to modify the information of an elasticity assurance.

Procedure

1.
2.
3.
4. In the upper part of the Resource Assurance page, click the **Elasticity Assurance** tab.
5. View and modify information of an elasticity assurance.
   - On the Elasticity Assurance tab, view information of an elasticity assurance, such as its status, instance type, zone, and resource usage of the associated private pool.
   - Click the ID of an elasticity assurance to view more information such as associated instances on

the details page of the elasticity assurance.

Instances that were created by using the capacity in the private pool associated with the elasticity assurance are displayed in the **Associated Instances** section on the details page of the elasticity assurance. You can find all instances that are associated with the elasticity assurance in the section.

- Click the ID of an elasticity assurance to go to the details page of the elasticity assurance. Click the icon on the right of a parameter to modify the parameter value, such as Description or Private Pool Name.

5.6.2.6. View and modify a capacity reservation

This topic describes how to view information of a capacity reservation, such as its status, instance type, zone, resource usage of the associated private pool, associated instances, and operating system. This topic also demonstrates how to modify the information of a capacity reservation.

Procedure

1.
2.
3.
4. In the upper part of the Resource Assurance page, click the **Capacity Reservation** tab.
5. View and modify information of a capacity reservation.
   - On the Capacity Reservation tab, view information of a capacity reservation, such as its status, zone, instance type, resource usage of the associated private pool, and operating system.
   - Click the ID of a capacity reservation to view more information such as associated instances on the details page of the capacity reservation.

     Instances that were created by using the capacity in the private pool associated with the capacity reservation are displayed in the **Associated Instances** section on the details page of the capacity reservation. You can find all instances that are associated with the capacity reservation in the section.
   - Click the ID of a capacity reservation to go to the details page of the capacity reservation. Click the icon on the right of a parameter to modify the parameter value, such as Expiration Method, Instances, Description, or Private Pool Name.

5.6.2.7. Release an immediate capacity reservation

This topic describes how to manually release an immediate capacity reservation whose Expiration Method is Manual Release.

Context

Capacity reservations are used only to reserve resources. When a capacity reservation is released, the instances created from the reserved resources are not affected and continue to run properly.

Procedure

1.

2.
3.
4. In the upper part of the Resource Assurance page, click the **Capacity Reservation** tab.
5. Find the capacity reservation that you want to release and click **Release** in the **Actions** column.
6. Click **OK**.

5.6.3. Private pools

5.6.3.1. View a private pool

After you purchase elasticity assurances or immediate capacity reservations, Alibaba Cloud reserves resources that have matching attributes in the form of private pools for your use. This topic describes how to view your private pool and its associated instances.

Procedure

1.
2.
3.
4. In the upper part of the page, click the **Private Pools** tab.
5. View information of a private pool.
   - On the Private Pools tab, view information of a private pool, such as its acquisition method, type, status, and capacity usage.
   - Click the ID of a private pool to go to the details page of the private pool. In the **Associated Instances** section, view the instances that are using the private pool.

5.6.3.2. Use a private pool to create instances

After you purchase elasticity assurances or immediate capacity reservations, Alibaba Cloud reserves resources that have matching attributes in the form of private pools. You can use these private pools to create pay-as-you-go Elastic Compute Service (ECS) instances. This topic describes how to use a private pool to create instances.

Prerequisites

- The private pool that you want to use is in the **Active** state.
- The private pool has available capacity. For more information, see View a private pool.

Context

This topic describes only private pool-related parameters involved in a procedure to create ECS instances. For information about other parameters involved in the procedure, see Create an instance by using the wizard.

Procedure

1. Go to the Custom Launch tab of the instance buy page in the ECS console.
2. Complete the settings in the Basic Configurations step and click **Next**.

Take note of the following items:

- Set Billing Method to **Pay-As-You-Go**.
- To use the private pool associated with an elasticity assurance, make sure that the selected zone and instance type are the same as those of the elasticity assurance. Otherwise, instances cannot be created.
- To use the private pool associated with an immediate capacity reservation, make sure that the selected zone, instance type, and operating system are the same as those of the capacity reservation. Otherwise, instances cannot be created.

3. Complete the settings in the Networking step and click **Next**.
4. Complete the settings in the System Configurations (Optional) step and click **Next**.
5. Complete the settings in the Grouping (Optional) step and click **Next**.

   Configure whether to use a private pool and which private pool to use.

   - Manually specify the ID of a private pool to use. When you create instances, set **Private Pool** to **Targeted** and specify the ID of a targeted or open private pool. If the specified private pool has no available capacity, instances cannot be created.

     > **Note** The ID of a private pool is the same as the ID of the elasticity assurance or immediate capacity reservation that is associated with the private pool.

   - Manually add tags to match an open private pool. If you added tags at the time of purchase for an elasticity assurance or immediate capacity reservation, tag matching is automatically enabled for the associated private pool. When you create instances, set **Private Pool** to **Open** and add the tags of your elasticity assurance or immediate capacity reservation. The open private pool that is associated with your elasticity assurance or immediate capacity reservation is automatically used. If the associated private pool has no available capacity, the system attempts to use the public pool.

     > **Note** If you specify the ID of an open private pool and add the tags of the elasticity assurance or capacity reservation associated with another open private pool to instances, the open private pool whose ID is specified is used. For example, if you specify the ID of Open Private Pool A and add the tags of the elasticity assurance or capacity reservation associated with Open Private Pool B, Open Private Pool A is used.

   - Use an open private pool that is selected by the system. When you create instances, set **Private Pool** to **Open**. The system selects an open private pool whose associated assurance or capacity reservation does not have tags. If no open private pools are available, the system attempts to use the public pool.
   - Specify not to use private pools. When you create instances, set **Private Pool** to **None**. The public pool instead of private pools is used.

6. Check your configurations and read the terms of service. Confirm to create the instances and complete the payment.

What's next

After you use a private pool to create instances, you can check whether and which instances are associated with the private pool on the **Private Pools** tab. For more information, see View a private pool.

5.6.3.3. Configure a private pool for existing instances

You can configure existing instances to use a private pool to increase the capacity usage of the private pool. You can also configure existing instances not to use a private pool so that the private pool can be used to create instances.

Context

Some examples on how to work with private pools:

- Assume that you have used a private pool to create instances and want to create more instances when resources are strained. You can configure the created instances not to use the private pool and then use the private pool to create new instances.
- Assume that you do not want to use the private pool associated with a purchased capacity reservation to create instances but you want to use the private pool for existing instances. You can configure the existing instances to use the private pool. This helps improve resource utilization and reduce cost wastes.
- Assume that you want to release existing instances to create new ones and want to purchase a capacity reservation to reserve resources. Your request to purchase a capacity reservation is rejected due to insufficient resources. You can set the private pool type to Open for your existing instances so that the instances are deemed as matchable resources. Then, you can purchase a capacity reservation.

Procedure

1.
2.
3.
4. On the **Instances** page, use one of the following methods to configure a private pool for one or more instances:
   - To configure a private pool for a single pay-as-you-go instance at a time, find the instance and choose **More > Instance Settings > Configure Private Pool** in the **Actions** column.
   - To configure a private pool for multiple pay-as-you-go instances at a time, select the instances and choose **More > Instance Settings > Configure Private Pool** in the lower part of the page.
5. In the **Configure Private Pool** dialog box, configure whether to use a private pool and which private pool to use.
   - Manually specify the ID of a private pool to use. Set **Private Pool Type** to **Target** and specify the ID of an open or targeted private pool.

     > **Note** The ID of a private pool is the same as the ID of the elasticity assurance or immediate capacity reservation that is associated with the private pool.

   - Manually add tags to match an open private pool. Set **Private Pool Type** to **Open** and add the tags of an elasticity assurance or capacity reservation to the instances. The instances match and use the open private pool that is associated with the elasticity assurance or capacity reservation.

> **Note** If you specify the ID of an open private pool and add the tags of the elasticity assurance or capacity reservation associated with another open private pool to instances, the open private pool whose ID is specified is used. For example, if you specify the ID of Open Private Pool A and add the tags of the elasticity assurance or capacity reservation associated with Open Private Pool B, Open Private Pool A is used.

For information about how to edit tags of existing instances, see Edit the tags of an instance. After an instance has matched an open private pool based on tags, if you want to edit the instance tags to match the instance to a different private pool, make sure that the following conditions are met:

- The economical mode is enabled for the instance.
- The instance is manually stopped and then started.

> **Note** The instance cannot be matched to another open private pool based on new tags when it is restarted.

- Use an open private pool that is selected by the system. Set **Private Pool Type** to **Open**. The system selects an open private pool whose associated elasticity assurance or capacity reservation does not have tags.
- Specify not to use private pools. Set **Private Pool Type** to **None**. The public pool instead of private pools is used.

6. Click **OK**.

What's next

After you configure a private pool for existing instances, you can check whether and which instances are associated with the private pool on the **Private Pools** tab. For more information, see View a private pool.

5.6.4. Privileges & Quotas

5.6.4.1. Overview

You can view the privileges and quotas available for your Alibaba Cloud account in the Elastic Compute Service (ECS) console.

The following table describes the privileges and quotas in ECS.

Type	Description	Operations in the ECS console
Instance quota	Instance quotas are allocated based on zones, instance types, billing methods, and network types. You can request an instance quota increase based on your business needs.	• View and increase instance quotas. • View quota increase requests.

Type	Description	Operations in the ECS console
Resource quota	Resource quotas are quotas for ECS resources such as images, disks, and security groups. You can request a resource quota increase based on your business needs.	• View and increase resource quotas. • View quota increase requests.
Privilege	• Feature privileges specify whether you can downgrade the configurations of instances in real time, copy images, or import and export images. • Network privileges specify whether the classic network is available in a specified region.	View privileges.

5.6.4.2. View and increase instance quotas

Instance quotas are allocated based on regions, zones, instance types, billing methods, and network types. This topic describes the different types of instance quotas and how to view and increase instance quotas.

Context

View instance quotas

1.
2.
3.
4. On the **Privileges and Quotas** page, click the **Instance Quota** tab to view instance quotas.

Increase instance quotas

When you request an instance quota increase, the recommended quota is displayed. The request is submitted for fast and automatic approval first. If the request is rejected, it can be submitted again for manual approval.

- Typically, a requested quota that is less than the recommended quota can pass automatic approval.
- A requested quota that is greater than the recommended quota may be rejected by automatic approval. If the request is rejected, you can submit it again for manual approval.

1.
2.
3.
4. On the **Privileges and Quotas** page, click the **Instance Quota** tab.
5. On the **Instance Quota** tab, click **Increase Quota** in the **Actions** column corresponding to the instance quota that you want to increase.
6. On the Instance Quota tab, click **Increase Quota** in the **Actions** column corresponding to the instance quota that you want to increase.

7. In the Calculate Quotas and Request dialog box, configure parameters described in the following table. Then, click **Submit**.

Parameter or option	Description
Target Quota	Enter your expected quota based on the number of instances or vCPUs.
Reason	Enter the reason of your request.
Notify Adjustment Results	If you select Yes, the system notifies you of the adjustment resultsby email.

8. In the message that appears, confirm the result of the quota increase operation.
9. On the Instance Quota tab, click **Quota Increase Requests** in the **Actions** column corresponding to the quota.
   - **Approved**: Your request is approved and cannot be revoked.
   - **Deny**: Your request is rejected. You can click **Submit Again** in the **Actions** column and select Manual Approval.

   > **Note** You can view all quota increase requests on the **My Quota Requests** tab. For more information, see View quota increase requests.

5.6.4.3. View and increase resource quotas

This topic describes how to view and increase the quotas for Elastic Compute Service (ECS) resources such as images, disks, and security groups.

View resource quotas

1.
2.
3.
4. On the **Privileges and Quotas** page, click the **Resource Quota** tab to view quotas for different resources.

Increase resource quotas

When you request a quota increase for a resource, the recommended quota is displayed. The request is submitted for fast and automatic approval first. If the request is rejected, it can be submitted again for manual approval.

- Typically, a requested quota that is less than the recommended quota can pass automatic approval.
- A requested quota that is greater than the recommended quota may be rejected by automatic approval. If the request is rejected, you can submit it again for manual approval.

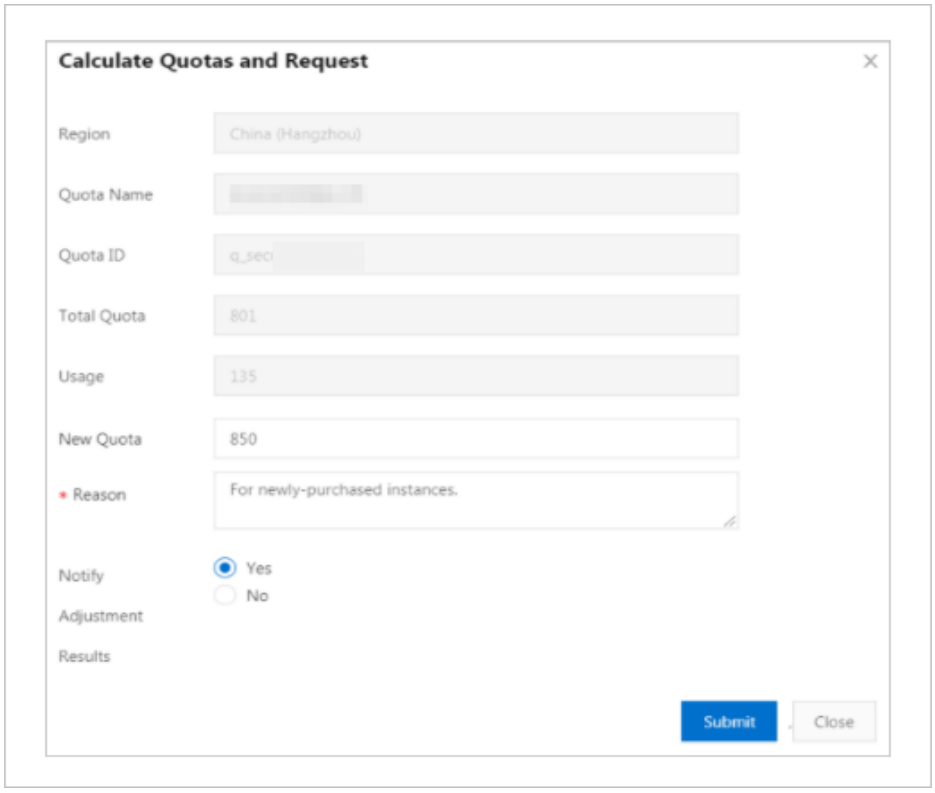
1.
2.
3.
4. On the **Privileges and Quotas** page, click **Resource Quota**.

5. On the Resource Quota tab, click **Increase Quota** in the **Actions** column corresponding to the quota that you want to increase.

   You can select multiple quota names and click **Calculate Quotas and Request** in the lower part of the page to batch increase resource quotas.

   > **Note** If you cannot click Increase Quota, this may be because the quota cannot be increased or because an increase request is being processed. You cannot request a quota increase for a resource when a previous quota increase request for the resource is being processed.

6. In the Calculate Quotas and Request dialog box, configure parameters described in the following table. Then, click **Submit**.



Parameter or option	Description
New Quota	Enter your expected quota. The unit is based on the resource type.
Reason	Enter the reason of your request.
Notify Adjustment Results	If you select Yes, the system notifies you of the adjustment resultsby email.

7. Click **Quota Increase Requests** in the **Actions** column to view the state of your request.
   - **Approved**: Your request is approved and cannot be revoked.
   - **Deny**: Your request is rejected. You can click **Submit Again** in the **Actions** column and select Manual Approval.
   - **Pending**: You request is being processed.

> **Note** You can view all quota increase requests on the **My Quota Requests** tab. For more information, see View quota increase requests.

Related information

-
-
- DescribeAccountAttributes (This operation queries only resource quotas and the zone-specific vCPU-based total quota for all instance types.)
-
-

5.6.4.4. View privileges

This topic describes how to view privileges that your Alibaba Cloud account has.

Procedure

1.
2.
3.
4. On the **Privileges and Quotas** page, click **Privileges** to view the privileges that your account has.

   > **Note** Your privileges change in response to the usage of your Elastic Compute Service (ECS) resources.

   The following table describes the sections on the Privileges tab.

Section	Description
Feature Privileges	Privileges contain **Downgrade**, **Copy Image**, and **Export Custom Image**. Only feature privileges that your account has are displayed in the section.
Network Privileges	The network privileges indicates whether the classic network is available in the selected region.

5.6.4.5. View quota increase requests

This topic describes how to view quota increase requests for Elastic Compute Service (ECS) instances and resources. If a quota increase request is rejected, you can submit the request again.

Procedure

1.
2.
3.
4. On the **Privileges and Quotas** page, click **My Quota Requests** to view all quota increase requests for instances and resources.

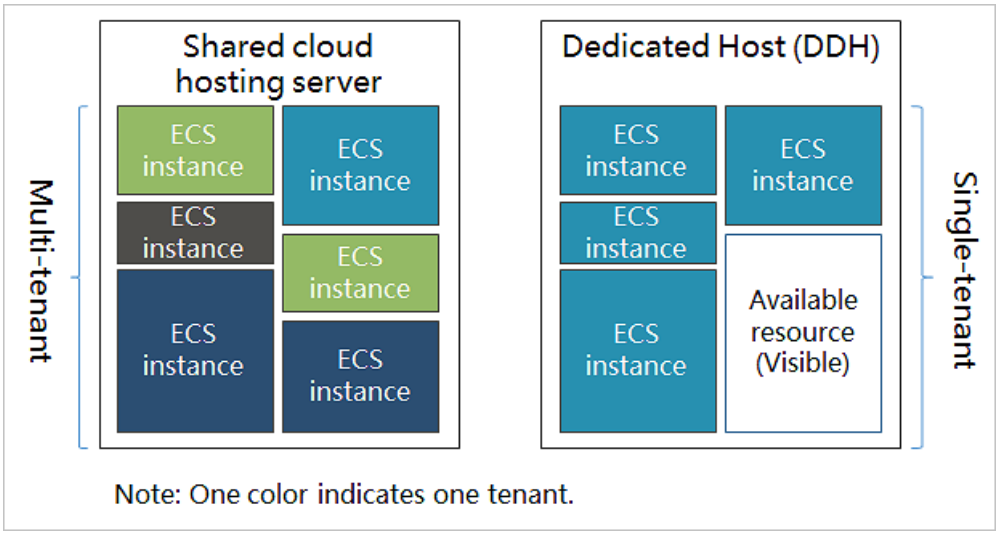
- To view quota increase requests for instances, click **Instance Quota**.
- To view quota increase requests for resources, click **Resource Quota**.

5. (Optional)If **Deny** is displayed in the Status column corresponding to a request, you can click **Submit Again** in the **Actions** column and select Manual Approval.

5.7. Dedicated hosts

A dedicated host is a cloud host whose physical resources are exclusively reserved for a single tenant.

A shared host is a cloud host whose physical resources are shared by multiple tenants. When you deploy Elastic Compute Service (ECS) instances on shared hosts, you cannot choose which shared hosts to use but can only use the shared hosts assigned by the system. Unlike shared hosts, dedicated hosts provide physical resources for your exclusive use and allow you to plan these physical resources at your discretion. For example, you can deploy ECS instances on specified dedicated hosts or view the physical attributes of dedicates hosts, including the number of sockets (CPUs) and the number of physical cores. The following figure shows the differences between dedicated hosts and shared hosts.



ECS instances deployed on dedicated hosts can meet the following typical requirements:

- Security compliance: ECS instances are physically isolated and meet the regulatory requirements of sensitive business.
- Support for bring your own license (BYOL): If you have purchased licenses for sockets and physical cores, you can import BYOL images to ECS and use the images to create ECS instances without the need to purchase the licenses again. This way, you can reduce cloud migration costs.
- Lower deployment costs: The CPU overprovisioned dedicated host type is provided to reduce unit deployment costs by increasing the number of available vCPUs under the same conditions of physical resources.
- Self-planning of physical resources: You can deploy ECS instances on specified dedicated hosts or migrate ECS instances between shared and dedicated hosts or between dedicated hosts. You can also associate ECS instances with dedicated hosts to ensure that the instances always reside on the same dedicated hosts. For example, when you reactivate a pay-as-you-go instance that was stopped in economical mode and has its computing resources released, the instance remains on the associated dedicated host.

After you purchase a dedicated host, you are not charged for the computing resources (vCPUs and memory) or local disks when you create ECS instances on the dedicated host. However, if the instances use other billable resources such as paid images, cloud disks, and public bandwidth, you are charged for these resources.

For more information, see What is DDH?

5.8. Switch billing method

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

After you create a pay-as-you-go Elastic Compute Service (ECS) instance, you can change its billing method to subscription to reserve resources at a discounted rate. This topic describes how to change the billing method of an instance from pay-as-you-go to subscription in the ECS console.

Prerequisites

The ECS instance for which you want to change the billing method meets the following requirements:

- The instance type of the instance is not retired. For more information, see Retired instance types.
- The instance is not a preemptible instance.
- You do not have unpaid orders for the instance.

  If you have unpaid orders for the instance, you must pay for the orders or cancel the orders before you can change the billing method of the instance.
- The automatic release time is not set for the instance.

  If the automatic release time is set for the instance, you must cancel the automatic release of the instance before you change its billing method. For more information, see Disable automatic release.
- The instance is in the **Running** or **Stopped** state.

  Note: An order to change the billing method of an ECS instance must be placed when the ECS instance is in the Running or Stopped state. If the instance state changes before the payment completes, the order fails and the billing method does not change. You can go to the Billing Management console and pay for the order when the instance is in the Running or Stopped state again.

Procedure

1.
2.
3.
4. Use one of the following methods to change the billing method of an instance from pay-as-you-go to subscription:
   - Change the billing method of a single instance: Find the instance and choose **More > Configuration Change > Switch to Subscription** in the **Actions** column.
   - Change the billing method of multiple instances: Select the instances and click **Switch to Subscription** in the lower part of the page.

> **Note** You can change the billing method of up to 20 pay-as-you-go instances at a time.

5. In the Switch to Subscription dialog box, complete the settings for the billing method change.
   i. Select the duration of the subscription.

      Instances whose billing methods are changed at the same time must have the same subscription duration.

   ii. Select Switch to Subscription in the Data Disk section if you want to change the billing method of the data disks attached to the instances.

   iii. Read and select *ECS Service Terms*.
6. Click **Create Order** and then complete the payment.

5.8.2. Change the billing method of an instance from subscription to pay-as-you-go

This topic describes how to change the billing method of an Elastic Compute Service (ECS) instance from subscription to pay-as-you-go. After you create a subscription instance, you can change its billing method to pay-as-you-go to recover some costs and use the instance in a more flexible manner. After the billing method of the instance is changed from subscription to pay-as-you-go, make sure that your account balance is sufficient. Otherwise, overdue payments occur and services provided by the instance are affected.

Prerequisites

The instance whose billing method you want to change is in the **Running** or **Stopped** state.

Context

Alibaba Cloud determines whether the billing method of your instance can be changed based on your calculated ECS instance usage. You can perform the operations described in this topic to check whether an entry point for changing the billing method of your instance exists in the ECS console. If the entry point does not exist, the billing method of your instance cannot be changed.

The following section describes the results after you change the billing method of an ECS instance from subscription to pay-as-you-go:

- The billing methods of the ECS instance, system disk, and subscription data disks attached to the instance are changed to pay-as-you-go. The billing method for network usage remains unchanged.
- The subscription duration that was previously offered for reasons such as the Internet Content Provider (ICP) filing, system failures, or migration from data centers is automatically invalidated.

> **Note** If the economical mode is enabled and the subscription instance is in the Stopped state before the billing method of the instance is changed, the economical mode is not automatically triggered after the billing method of the instance is changed to pay-as-you-go. You must manually start and then stop the pay-as-you-go instance to trigger the economical mode.

The following refund rules apply after you change the billing method of the instance from subscription to pay-as-you-go:

- A refund generated from the billing method change counts against the monthly refund quota. If the refund quota within your account has been used up, you cannot apply for new refunds until the refund quota is set on the first day of the next month. For information about the refund quota, see 使用限制.

  When you change the billing method of the instance, the refund amount for the instance is calculated based on the number of vCPUs and the remaining hours in the current billing cycle. Example: 1 refund unit = `1 vCPU × 1 hour`.

  For example, you have purchased a six-month subscription instance that has four vCPUs. Four months later, you want to change the billing method of the instance to pay-as-you-go. In this case, the refund amount for the instance is calculated based on the following formula: `4 (vCPUs) × 60 (remaining days) × 24 (hours/day) = 5760 (refund)`.

- If you have renewal or upgrade orders that have not taken effect for the instance, full refunds are made for the orders. If you have orders that have already taken effect for the instance, only partial refunds are made for the orders.

Procedure

1.
2.
3.
4. Use one of the following methods to change the billing method of an instance from subscription to pay-as-you-go:
   - To change the billing method of a single instance at a time, find the instance and choose **More > Configuration Change > Switch to Pay-As-You-Go** in the **Actions** column.
   - To change the billing methods of multiple instances at a time, find the instances and choose **More > Configuration Change > Switch to Pay-As-You-Go** in the lower part of the page.
5. Read the notes. Read and select *ECS Service Terms* and then click **Switch**.

Result

After the billing method is changed, you can go to the ECS console to view the billing method of the instance:

- On the **Instances** page, the billing method of the instance has been changed to **Pay-As-You-Go** in the **Billing Method** column.
- Click the instance ID to go to the **Instance Details** page. Click the **Cloud Disk** tab. In the **Billing Method(All)** column, the billing methods of the system disk and data disks (if any) are changed to **Pay-As-You-Go**.

What's next

You can set the automatic release time for the instance to be automatically released when the instance is no longer needed to minimize costs. For more information, see Release an instance.

Related information

- ModifyInstanceChargeType

6.Create an instance

6.1. Creation method overview

This topic describes several methods of creating an ECS instance, from basic creation operations to advanced customization operations.

You can create an ECS instance and choose configurations by following the instructions prompted by the wizard on the buy page. For more information, see Create an instance by using the provided wizard.

> **Note** If you require custom configurations such as a specific operating system or application, you can create a custom image and then select that image during the creation of an instance to improve configuration efficiency. For more information, see Create an instance by using a custom image.

If you need a new instance to have the same configuration as the current instance, you can create an instance of the same configuration. For more information, see Purchase an ECS instance of the same configuration.

You can also create a launch template in advance, and then use it to create an instance in one click. For more information, see Create an instance by using a launch template. For more information about launch templates, see 实例启动模板概述.

6.2. Create an instance by using the wizard

The Elastic Compute Service (ECS) console provides a wizard for creating instances. This wizard lists all configuration information used to create an instance and guides you through creating an instance.

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.

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.

   Different billing and charging rules apply to the instance based on the selected billing method. The state changes of instance resources also vary based on the billing method.

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 on demand. **Note** We recommend that you use the pay-as-you-go billing method together with savings plans and reserved instances to reduce costs.	○ Pay-as-you-go ○ Savings plans ○ Reserved instances
Preemptible Instance	A billing method in which you use resources first and pay for them afterward. You can place a bid for available instance resources to create preemptible instances at a discount compared with pay-as-you-go instance pricing. Preemptible instances may be automatically released due to fluctuations in 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 complete the relevant configurations.

   i. Select an instance type.

   Available instance types vary based on the selected region. You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.

   You may have specific configuration requirements for the instance. For example, you may want the instance to have multiple elastic network interfaces (ENIs) bound, use enhanced SSDs (ESSDs), or use local disks. In this case, make sure that the selected instance type meets your requirements. For more information about the features, supported scenarios, and specifications of instance types, see Instance family.

   If you purchase instances for specific scenarios, you can click the **Scenario-based Selection** tab to view the recommended instance types. For example, you can set Business Scenario to Web Development and Test, Big Data Cluster, or AI Machine Learning.

ii. 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.

iii. 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 required to run instances. Alibaba Cloud provides a variety of image types for easy access to images. The following table describes the image types.

Image type	Description	References
Public image	Public images are base images provided by Alibaba Cloud that are fully licensed. These images 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 initial system environments, application environments, and software configurations. This eliminates the need for repeated manual configurations.	Overview
Shared image	Shared images are custom images shared by other Alibaba Cloud accounts for you to create instances.	Share or unshare a custom image
Alibaba Cloud Marketplace image	An extensive range of images are provided in Alibaba Cloud Marketplace. Alibaba Cloud Marketplace images are thoroughly reviewed by Alibaba Cloud and can be used to create instances for website building and application development purposes without additional configurations.	Alibaba Cloud Marketplace images
Community image	Community images are publicly available. Custom images can be published as community images for other users to obtain and use.	Overview

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 Create a General-purpose NAS file system in the NAS console. 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

You can make network and security group configurations to allow the instance to communicate with the Internet and other Alibaba Cloud resources and safeguard the instance on the network. After you complete the settings in the Networking step, click **Next**.

1. Select a network type.

   We recommend that you select VPC. Virtual private clouds (VPCs) are physically isolated from each other to ensure security and support features such as elastic IP addresses (EIPs) and ENIs.

Network type	Description	References
VPC	A VPC is an isolated network dedicated for your use. You have full control over your VPC. For example, you can specify a private CIDR block and configure route tables and gateways for the VPC. If you have not created a VPC in the region selected in the Basic Configurations step, you can skip this step. The system creates a default VPC and vSwitch in that region. Select an existing VPC and an existing vSwitch. Alternatively, 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 query the most recent VPC and vSwitch lists.	○ What is a VPC? ○ Create a VPC and a vSwitch ○ Create a vSwitch
Classic Network	Instances of the classic network type are deployed in the public infrastructure of Alibaba Cloud, and are planned and managed by Alibaba Cloud. **Note** If you purchase an ECS instance for the first time after 12:00 on June 16, 2016 (UTC+8), you can no longer select the classic network.	Network types

2. (Optional)Assign a public IP address to the instance.

   To enable the instance to access the Internet, you must assign a public IP address to the instance. You can select Assign Public IPv4 Address when you create an instance to have a public IP address automatically assigned to the instance. Alternatively, you can configure an EIP or a NAT gateway after an instance is created to provide Internet access for the instance. You can purchase EIPs and NAT gateways on your own. For more information, see What is an EIP? and What is NAT Gateway?.

   i. Select **Assign Public IPv4 Address**.

ii. Select a billing method for network usage.

For more information about the billing methods for network usage, see Public bandwidth.

Billing method for network usage	Description
Pay-By-Bandwidth	You are charged based on the specified bandwidth. This billing method for network usage is applicable to scenarios that require stable network bandwidth.
Pay-By-Traffic	You are charged for the amount of data actually transferred. You can configure maximum bandwidths for inbound and outbound traffic to prevent unmanageable fees incurred by bursts of traffic. This billing method for network usage is suitable for scenarios that require highly variable bandwidth, such as the scenarios where traffic is low in most cases but spikes occasionally occur.

iii. Set Bandwidth or Peak Bandwidth.

3. 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**.

4. 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. For more information, see Create an SSH key pair. **Note** Key pairs can be used to log on only to Linux instances.
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 ○ Batch create ECS instances ○ 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.

Result

After the instance is created, go to the Instances page to check the state of the instance. When the state of the instance changes to **Running**, the instance can be accessed.

What's next

- You can connect to an instance to format its data disks. You can connect to instances by using Workbench, Virtual Network Computing (VNC), or third-party client tools. For more information about how to connect to instances by using VNC, see Connect to a Linux instance by using a password and Connect to a Windows instance by using a password.
- If you add data disks when you create instances, you must partition and format the data disks before you can use them. For more information, see Partition and format a data disk on a Linux instance and Partition and format a data disk on a Windows instance.
- You can build an FTP site on an ECS instance to upload local files to the instance. For more information about how to deploy the FTP service, see Manually build an FTP site on a CentOS 8 instance and Manually build an FTP site on a Windows instance.

Related information

- RunInstances
- Instance FAQ

6.3. Create an ECS instance by using a custom image

This topic describes how to create an ECS instance by using a custom image. You can use a custom image to create an ECS instance that has the same operating system, applications, and data as those of the custom image to make the process more efficient.

Prerequisites

A custom image is created in the account and region where you want to create an instance.

Context

If you do not have a custom image in the account and region where you want to create an instance, you can use one of the following solutions.

Scenario	Solution
You have an image on the local device.	Import the local image to Alibaba Cloud as a custom image. For more information, see Image import procedure.
You do not have custom images but have an instance as a template.	For more information, see Create a custom image from an instance.
You do not have custom images but have a snapshot as a template.	For more information, see Create a custom image from a snapshot.
You have a custom image in another region.	Copy the custom image to the region where you want to create an instance. For more information, see Copy a custom image.
You have a custom image in another account.	Share the custom image with the account under which you want to create an instance. For more information, see Share or unshare a custom image.

Procedure

1.
2.
3.
4. Based on the image source, use one of the following methods to go to the Images page:
   - Custom image created or exported: Go to the **Custom Images** tab.
   - Custom image obtained by copying: Go to the **Custom Images** tab.
   - Custom image obtained by sharing: Go to the **Shared Images** tab.
5. Find the image that you want to use. Click **Create Instance** in the **Actions** column.

6. Configure the parameters and create the instance.

   Information of the region and image sections is automatically filled. Configure other parameters based on your needs. For more information, see Create an instance by using the wizard.

   > **Note** If the selected custom image contains one or more data disk snapshots, an equal number of data disks are automatically created from these snapshots. Each disk has the same size as the snapshot from which the disk is created. You can extend a data disk but cannot shrink it.

What's next

If you add data disks when you create the instance, you must format the partitions before you can use the data disks. For more information, see Partition and format a data disk on a Windows instance or Partition and format a data disk on a Linux instance.

Related information

- RunInstances

6.4. Purchase an ECS instance of the same configuration

You can purchase an ECS instance that has the same configuration as that of your existing instance. This allows you to improve the efficiency of creating ECS instances.

Procedure

1.
2.
3.
4. Find the target instance and choose **More > Buy Same Type** in the **Actions** column.
5. Confirm the configuration that is automatically selected.

   You can modify the configuration that is automatically selected. After you modified the configuration, you can click **Compare the original configuration** in the lower part of the page to view the same and different configurations.
6. Specify time options based on the billing method.
   - Subscription instance: Specify the duration and whether to enable auto-renewal.
   - Pay-as-you-go instance: Specify whether to enable automatic release. If you enabled automatic release, you must specify when to release the instance.
7. Read and select *ECS Terms of Service*.
8. Confirm instance creation based on the billing method.
   - Subscription instance: Click **Create Order**.
   - Pay-as-you-go instance: Click **Create Instance**.

Related information

- RunInstances

6.5. Create an instance by using a launch template

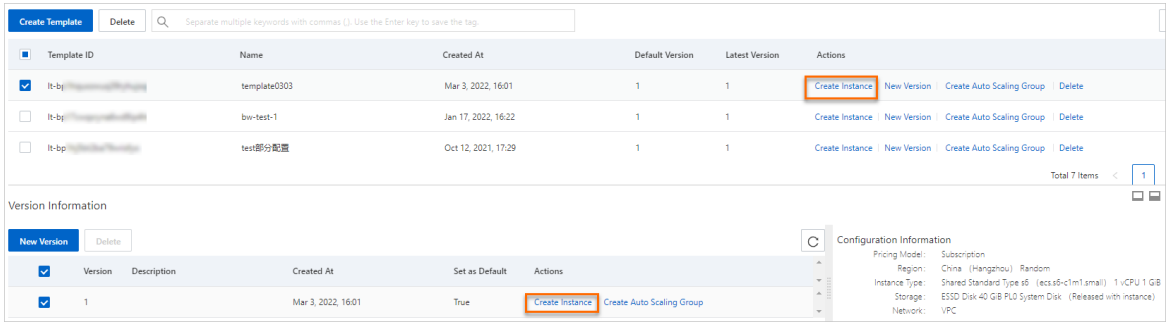
You can use an existing launch template to quickly create ECS instances.

Prerequisites

A launch template or a new version of an existing launch template is created. For more information, see Create a launch template and Create a launch template version.

Procedure

1.
2.
3. Find the launch template or version that you want to use. Click **Create Instance** in the **Actions** column.



4. On the **Custom Launch** tab that appears, select the template and version. Check all the configurations after they are loaded.

> **Note** If you want to modify the parameters, or if the selected template does not contain the required parameters, you can modify the configurations by clicking the Edit icon.

5. Create an instance.
   - To create a subscription instance, set Duration and select ECS Terms of Service and Product Terms of Service. Click **Create Order**.
   - To create a pay-as-you-go instance, select ECS Terms of Service and Product Terms of Service. Click **Create Instance**.

   After the instance is created, you can view its details in the ECS console.

Related information

- RunInstances

6.6. Instructions for purchase

This topic describes the information that you must understand before you purchase Elastic Compute Service (ECS) instances.

Resource upgrades

For information about how to upgrade the configurations of ECS instances, see Overview of instance configuration changes. Before you upgrade the configurations of ECS instances, take note of the following items:

- Except ECS instances that use local storage, ECS instances can have their CPU and memory resources scaled and their bandwidths upgraded while the instances are running. You can also downgrade the configurations of ECS instances.
- Typically, a maximum of 16 data disks can be attached to each ECS instance. You cannot reduce the size of a data disk after the data disk is extended.
- The bandwidth of each ECS instance is measured in Mbit/s and can range from 0 Mbit/s to 200 Mbit/s. You can modify the bandwidth or change the billing method for network usage.

References

The following topics describe the basic terms of ECS and how to use ECS:

- For information about the terms and services related to ECS, see What is ECS?
- For information about how to choose ECS instance families, see Instance family.
- For information about how to choose images, see Select an image.
- For information about the performance of Elastic Block Storage devices, see EBS performance.
- For information about the usage considerations of ECS instances, see Usage notes.
- For information about the features of pay-as-you-go resources, see Pay-as-you-go.
- For information about the features of subscription resources, see Subscription.
- For information about the limits that apply to ECS, see 使用限制.
- If you want to select instance types based on their benchmark data for compute performance, contact the Service Manager or Submit a ticket to contact Alibaba Cloud after-sales support.

7.Connect to instances

7.1. Connection methods

You can use a variety of methods to connect to an Elastic Compute Service (ECS) instance, such as Workbench, Virtual Computing Console (VNC) and third-party client tools. You can choose a method to connect to your instance based on the operating system of your instance, 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 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. • VNC For more information, see Connect to a Linux instance by using a password. • Client tools such as PuTTY ◦ 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. ◦ 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 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. • VNC For more information, see Connect to a Linux instance by using a password. • SSH commands ◦ 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). ◦ 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 For more information, see Connect to a Linux instance from a mobile device.
	Windows	• Workbench 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. • VNC For more information, see Connect to a Windows instance by using a password. • Client tools such as Remote Desktop Connection (formerly called MSTSC) 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 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. • VNC For more information, see Connect to a Windows instance by using a password. • Client tools such as rdesktop For more information, see Connect from a local client that runs a Linux operating system.
	macOS	• Workbench 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. • VNC For more information, see Connect to a Windows instance by using a password. • Client tools such as Microsoft Remote Desktop Connection for Mac 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 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. **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.

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

7.2. Connect to an instance by using Workbench

7.2.1. 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 in virtual private clouds (VPCs), you can use the public or private IP addresses of the instances. ◦ To connect to instances in the classic network, you can use the public or internal IP addresses of the instances.
Username, **Password**, and **PrivateKey**	Enter a username such as root and select an authentication method. The following authentication methods are supported: ◦ **Password-based**: Enter the password of your specified username. ◦ **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.

Parameter	Description
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.
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 want to log on to the Linux instance as the root user, make sure that `PermitRootLogin yes` and `PasswordAuthentication yes` are 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 Mismatch Between OS and Protocol 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 their public or private IP addresses. ■ To connect to instances 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 want to log on to the Linux instance as the root user, make sure that `PermitRootLogin yes` and `PasswordAuthentication yes` are 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. For more information, see Connect to a Linux instance by using password authentication.
2. Open the SSH configuration file.

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

3. Configure the following parameters:
   - Change `PermitRootLogin no` to `PermitRootLogin yes`.
   - Change `PasswordAuthentication no` to `PasswordAuthentication 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)**. ◦ 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. ◦ 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. 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. ◦ If you want to connect to the instance by using its private IP address, specify 100.104.0.0/16. **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)**. ○ 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. ○ 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. 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. **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)**. ◦ 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. ◦ 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 161.117.90.22. **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.

7.2.2. Use Workbench to manage files in a Linux instance

Workbench provides a GUI for you to manage files in Elastic Compute Service (ECS) Linux instances in a visual manner. This topic describes how to use Workbench to view, add, delete, and modify files in a Linux instance.

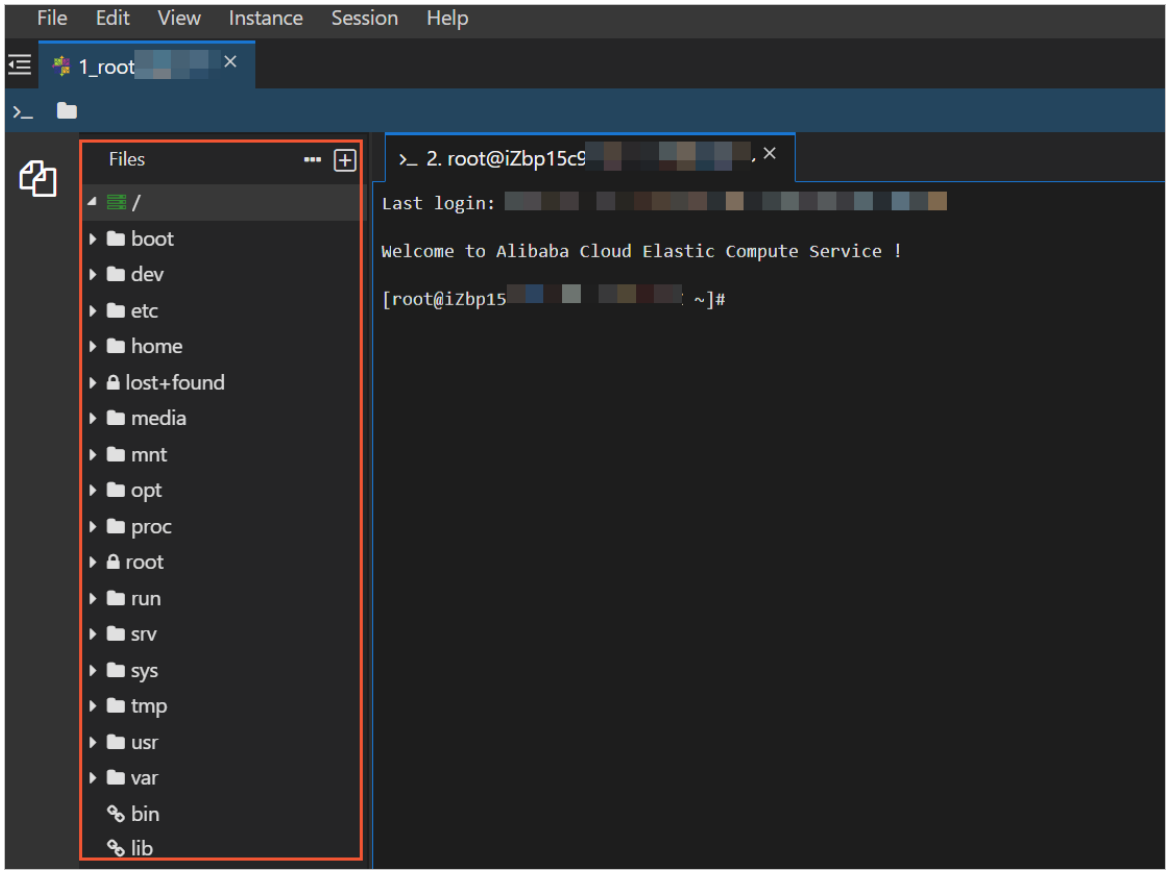
Prerequisites

A Linux instance is connected by using Workbench.

View files

1. In the left part of the top navigation bar, choose **File > New Navigator**.
2. View files.

   The following figure shows a file navigator.

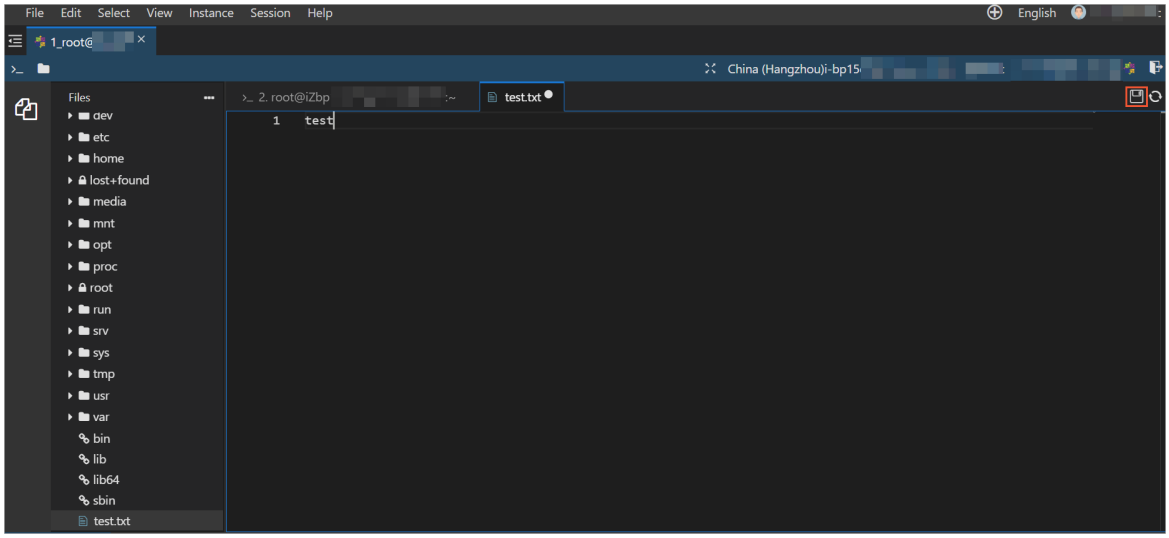


Create a file

1. In a file navigator, right-click the location where to create a file and select **New File**.
2. In the **New File** dialog box, enter a name for the file and click **OK**.

Edit a file

1. In a file navigator, right-click a file and select **Open**.
2. Enter content in the editor and click the Save icon in the upper-right corner.



3. In the **Confirm Saving** message, click **OK**.

Delete a file

1. In a file navigator, right-click a file and select **Delete**.
2. In the **Confirm Deletion** message, click **OK**.

7.2.3. 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. ○ 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. ■ 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: ■ **Password-based**: Enter the password of your specified username. ■ **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)**. ◦ 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. ◦ 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. ◦ If you want to connect to the instance by using its private IP address, specify 100.104.0.0/16. **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)**. ◦ 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. ◦ 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. **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)**. ◦ 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. ◦ 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. **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.

7.3. Connect to an instance by using session management

7.3.1. How session management works

The session management feature is provided by Cloud Assistant. Compared with SSH and Virtual Network Computing (VNC), session management makes your connections to Elastic Compute Service (ECS) instances more convenient and secure.

Establish connections

The following clients and server are used in session management:

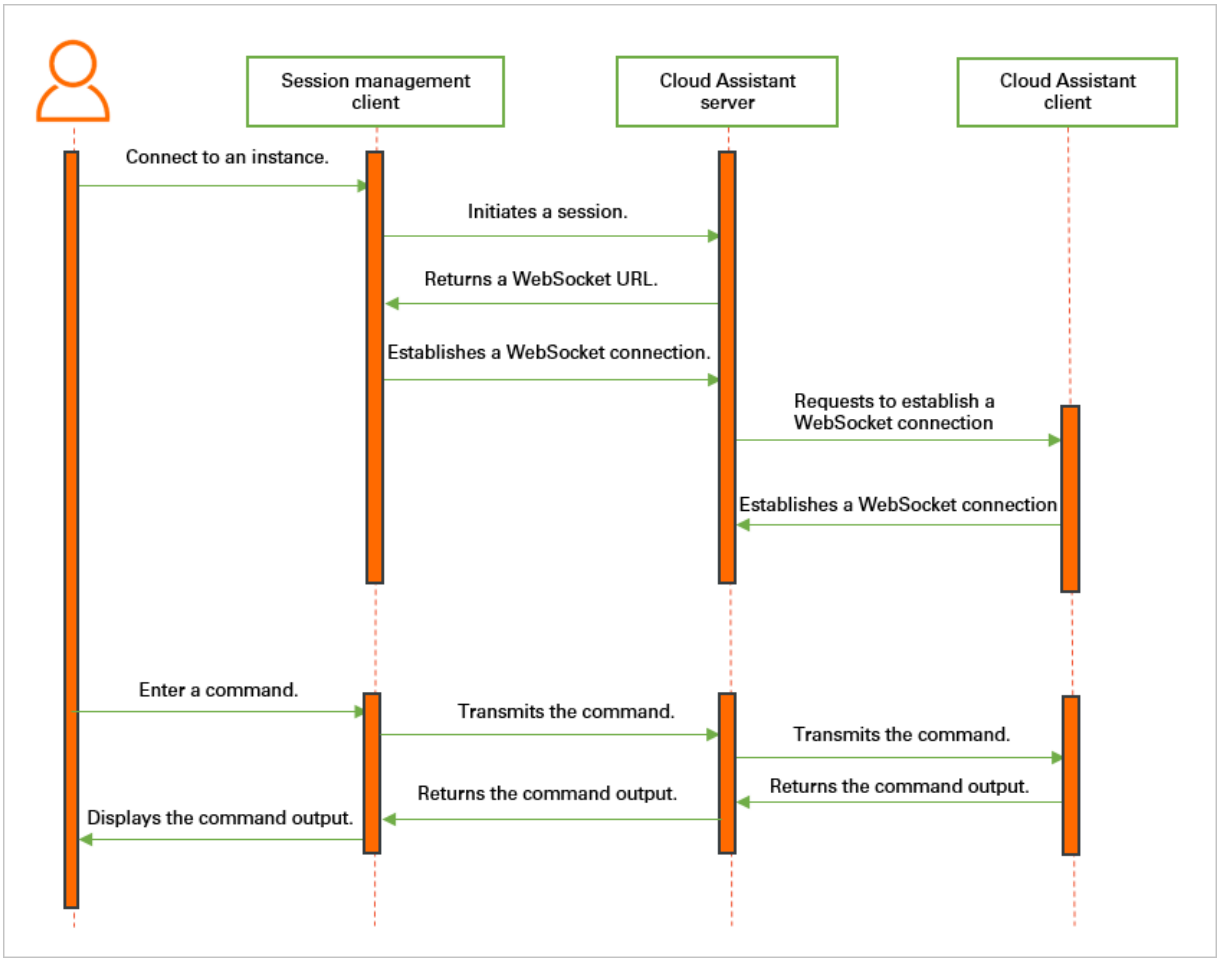
- Session management client: initiates sessions, receives commands sent by users, and displays command outputs.
- Cloud Assistant server: controls permissions and manages session status.
- Cloud Assistant client installed on an instance: runs commands sent by users.

The following section describes the procedure to establish a connection by using session management:

1. The session management client initiates a session.
2. The Cloud Assistant server authenticates the session request. After the request is authenticated, the server generates a WebSocket URL for connection and a token that remains valid for 10 minutes

and returns the URL and the token to the session management client.

3. The session management client establishes a WebSocket connection to the Cloud Assistant server by using the URL and the token.
4. The Cloud Assistant server requests to establish a WebSocket connection to the Cloud Assistant client that is installed on an instance.
5. A WebSocket connection is established between the Cloud Assistant server and the Cloud Assistant client.
6. After the WebSocket connection is established, you can enter a command in the session management client. The command is streamed to the instance on which the Cloud Assistant client is installed and is run on the instance. Then, the command output is displayed in the session management client.



Security

The Web Socket Secure (WSS) protocol is used to establish persistent WebSocket connections between the session management client and the Cloud Assistant server as well as between the Cloud Assistant server and the Cloud Assistant client. The WSS protocol encrypts the persistent WebSocket connections by using the Secure Socket Layer (SSL) protocol.

When you use session management to connect to instances, you do not need to manage the instance passwords. Unlike SSH and VNC that use the username and password authentication, session management uses the Resource Access Management (RAM)-based authorization. You can use your Alibaba Cloud account to enable or disable the session management feature for all instances within the account. After the session management feature is enabled, both Alibaba Cloud accounts and RAM users can connect to instances by using this feature.

If you want to use the session management feature as a RAM user, the RAM user must be attached policies to call the StartTerminalSession operation. RAM policies allow you to control permissions by using a variety of dimensions such as tags, regions, ECS instances, and connection IP addresses. Thanks to RAM policies, you can connect to instances and manage the instances in a secure manner without using passwords. For more information, see Connect to an instance by using session management.

After WebSocket connections are established between the Cloud Assistant clients installed on instances and the Cloud Assistant servers, you can use session management instead of SSH and VNC to connect to instances. In this case, ports that allow inbound traffic on instances can be disabled to improve the security of the instances.

7.3.2. Connect to an instance by using session management

When you connect to an Elastic Compute Service (ECS) instance by using session management, passwords and public IP addresses are not required, and SSH ports and Remote Desktop Protocol (RDP) ports do not need to be enabled. Compared with the SSH or RDP connection method, session management allows you to connect to instances in a more convenient manner. This topic describes how to connect to an instance by using session management.

Prerequisites

The instance meets the following requirements:

- The instance is in the **Running** state.
- The Cloud Assistant client is installed on the instance. The version of the client supports session management. The version of the client installed on a Linux instance must be 2.2.3.196 or later, and that on a Windows instance must be 2.1.3.196 or later. For more information about how to install the Cloud Assistant client, see Install the Cloud Assistant client.

> **Note** Session management is enabled. The session management feature is in public preview.

Context

Session management offers security and convenience. For information about how session management works, see How session management works.

Procedure

1. Log on to the ECS console.
   - Both Alibaba Cloud accounts and RAM users can use the session management feature. However, you can use only Alibaba Cloud accounts to enable and disable this feature. If the session management feature is not enabled, use an Alibaba Cloud account or contact the owner of an Alibaba Cloud account to enable this feature.

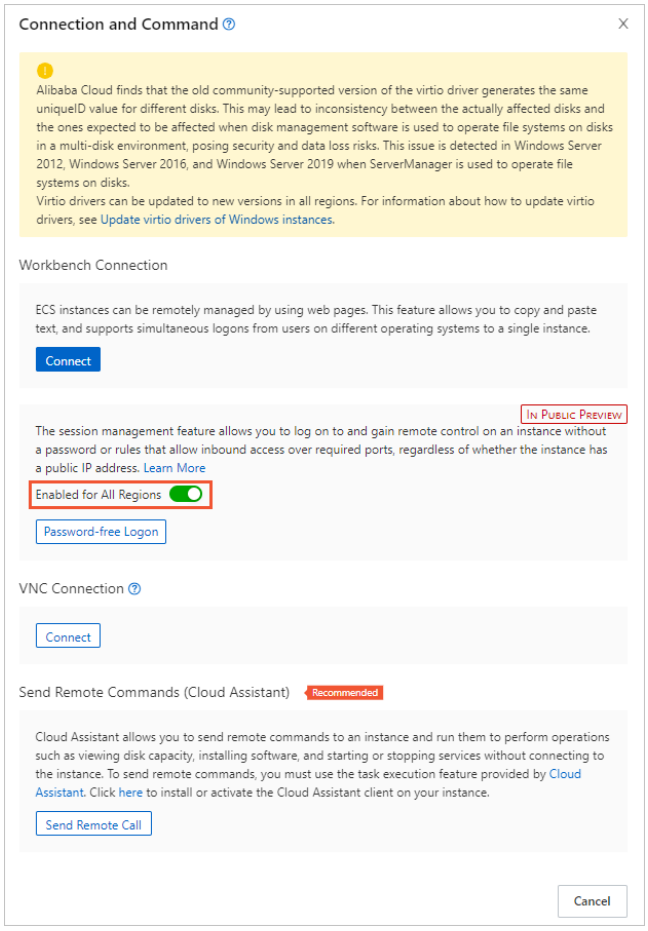
- If you want to use the session management feature as a RAM user, make sure that the RAM user is attached policies to call the StartTerminalSession operation. For information about sample policies, see the Sample policies section of this topic. Proceed with caution when you attach policies to RAM users. Otherwise, unauthorized operations may be performed due to improper management of or unintended authorizations to RAM users.

2.

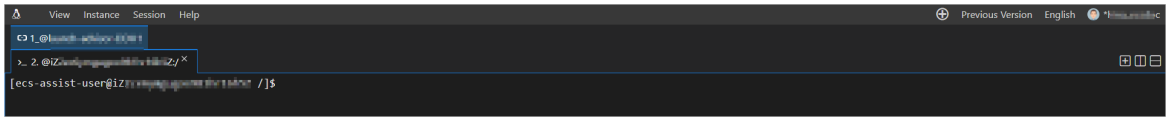
3.

4. On the **Instances** page, find the instance to which you want to connect and click **Connect** in the **Actions** column.
5. In the **Workbench Connection** section of the Connection and Command dialog box, make sure that **Enabled for All Regions** is displayed above the **Password-free Logon** button. If **Disabled** is displayed, turn on the switch of the session management feature.

> **Note** You can enable the session management feature to connect to instances without using passwords, which is more convenient. However, if you use RAM users to connect to instances without using passwords, proceed with caution when you attach policies to the RAM users. Otherwise, unauthorized operations may be performed due to improper management of or unintended authorizations to RAM users.



6. Click **Password-free Logon**.
The instance is connected by using the user named ecs-assist-user by default, as shown in the following figure.



Sample policies

After RAM users are attached policies to call the StartTerminalSession operation, the RAM users can be used to use session management to connect to instances. For information about how to create policies and attach policies to RAM users, see Create a custom policy and Grant permissions to a RAM user. The following sections provide examples of policies:

- The policy that allows a RAM user to connect to all instances

```
{
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "ecs:StartTerminalSession"
            ],
            "Resource": [
                "acs:ecs:*:*:instance/*"
            ]
        }
    ],
    "Version": "1"
}
```

- The policy that allows a RAM user to connect a specified instance

```
{
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "ecs:StartTerminalSession"
            ],
            "Resource": [
                "acs:ecs:*:*:instance/i-****",
                "acs:ecs:*:*:instance/i-****"
            ]
        }
    ],
    "Version": "1"
}
```

> **Note** Replace *i-**** with the ID of the instance to which you want to connect.

- The policy that allows a RAM user to connect to an instance that has a specified tag added

```
{
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "ecs:StartTerminalSession"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {
                    "ecs:tag/key-****": "value-****"
                }
            }
        }
    ],
    "Version": "1"
}
```

Note Replace *key-***** with the key of the specified tag and replace *value-***** with the value of the specified tag.

- The policy that allows a RAM user to connect to an instance from a specified IP address

```
{
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "ecs:StartTerminalSession"
            ],
            "Resource": "*",
            "Condition": {
                "IpAddress": {
                    "acs:SourceIp": [
                        "192.168.XX.XX",
                        "192.168.XX.XX/24"
                    ]
                }
            }
        }
    ],
    "Version": "1"
}
```

Note Replace *192.168.XX.XX* with the specified IP address or replace *192.168.XX.XX/24* with the specified CIDR block.

7.3.3. Connect to an instance by using ali-instance-cli

Session management is a feature provided by Cloud Assistant that allows you to connect to Elastic Compute Service (ECS) instances in a secure and convenient manner. ali-instance-cli is a CLI used for session management. This topic describes how to use ali-instance-cli to connect to an ECS instance.

Prerequisites

- The Cloud Assistant client is installed on the ECS instance to which you want to connect. For a Windows instance, the installed client version must be 2.1.3.256 or later. For a Linux instance, the installed client version must be 2.2.3.256 or later. For more information, see Install the Cloud Assistant client.
- For information about how to enable the session management feature, see Connect to an instance by using session management.

Context

When you use ali-instance-cli to connect to an instance, you need to only provide the ID and password of the instance. You do not need to expose the public IP address and port number of the instance. This connection method is more convenient and secure than using SSH or Virtual Network Console (VNC). For more information about session management, see How session management works.

Session management clients support Linux, macOS, and Windows operating systems and are used differently on these operating systems. For more information, see the following sections in this topic:

- Linux and macOS operating systems
- Windows operating systems

Linux and macOS operating systems

1. Log on to a session management client.
2. Install ali-instance-cli on the session management client.

   Run commands to install ali-instance-cli based on the following operating system types:

   - Linux

     ```
     curl -O https://aliyun-client-assist.oss-accelerate.aliyuncs.com/session-manager/linux/ali-instance-cli
     chmod a+x ali-instance-cli
     ```

   - macOS

     ```
     curl -O https://aliyun-client-assist.oss-accelerate.aliyuncs.com/session-manager/mac/ali-instance-cli
     chmod a+x ali-instance-cli
     ```

3. Configure an AccessKey pair, a Security Token Service (STS) token, or CredentialsURI.

   For information about how to obtain an AccessKey pair or STS token, see Obtain an AccessKey pair or What is STS?.

   i. Switch to the test directory.

   ```
   cd /home/test
   ```

ii. Configure an authentication method.

The following authentication methods are supported:

- AccessKey pair-based authentication

  Run the following command and enter an AccessKey ID, AccessKey secret, and region ID as prompted:

  ```
  ./ali-instance-cli configure --mode AK
  ```

- STS token-based authentication

  > **Note** Replace *region*, *ak*, *sk*, and *token* with the actual region ID, AccessKey ID, AccessKey secret, and STS token.

  ```
  ./ali-instance-cli configure set --mode StsToken --region "region" --access-key-id "ak"  --access-key-secret "sk"   --sts-token "token"
  ```

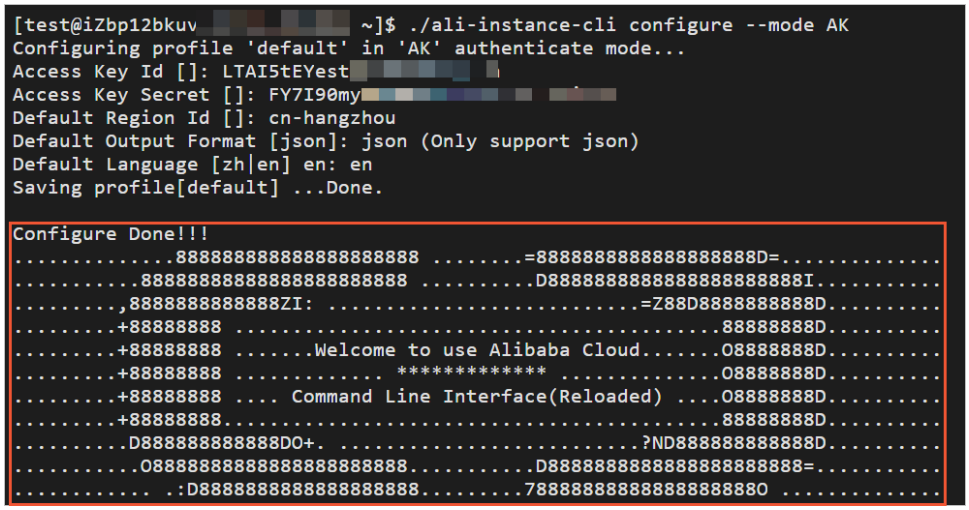
- CredentialsURI-based authentication

  Run the following command and specify CredentialsURI and RegionID as prompted.

  > **Note** Set the CredentialsURI value to the IP address of the authentication server that you configure.

  ```
  ./ali-instance-cli configure --mode=CredentialsURI
  ```

A command output similar to the following one indicates that the AccessKey pair-based authentication method is configured.



4. Run the following command to connect to an instance:

   ```
   ./ali-instance-cli session --instance instance-id
   ```

   > **Note** Replace *instance id* with the ID of the instance to which you want to connect.

A command output similar to the following one indicates that you are connected to the instance by using session management.

```
Linux instance to which you want to connect
```

Windows operating systems

Before you use a session management client that runs a Windows operating system to connect to an ECS instance, make sure that OpenSSH is installed on the client.For more information, see Use Cloud Assistant to install OpenSSH on an ECS Windows instance.

1. Log on to a session management client.

   For more information, see Connection methodsGuidelines on instance connection.

2. Download ali-instance-cli to the session management client.

   Download and save ali-instance-cli.exe for Windows to a directory on the session management client. In this example, the `C:\Users\test` directory is used.

3. Create a file named `config` and add configurations to the file.

   i. In the *C:\Users\<Username>* directory, create a folder named `.ssh`.

   > **Note** Replace *C:\Users\<Username>* with the actual directory. In this example, `C:\Users\test` is used.

   ii. In the `.ssh` folder, create a file named `config`.

   iii. Add the following content to the `config` file.

   Replace *ali-instance-cli.exe* with the absolute path of the ali-instance-cli.exe file. In this example, `C:\Users\test\ali-instance-cli.exe` is used.

   ```
   host i-*
       ProxyCommand C:\Windows\System32\WindowsPowerShell\v1.0\powershell.exe "ali-instance-cli.exe ssh -i '%h' --port  '%p'"
   ```

4. Configure an AccessKey pair or an STS token.

   For information about how to obtain an AccessKey pair or STS token, see Obtain an AccessKey pair or What is STS?.

   i. Choose **Start** > **Run**, enter **cmd**, and then press the `Enter` key to open the Command Prompt window.

   ii. Switch to the test directory.

   ```
   cd C:\Users\test
   ```

iii. Configure an authentication method.

The following authentication methods are supported:

- AccessKey pair-based authentication

  Run the following command and enter an AccessKey ID, AccessKey secret, and region ID as prompted:

  ```
  ali-instance-cli.exe configure --mode AK
  ```

- STS token-based authentication

  > **Note** Replace *region*, *ak*, *sk*, and *token* with the actual region ID, AccessKey ID, AccessKey secret, and STS token.

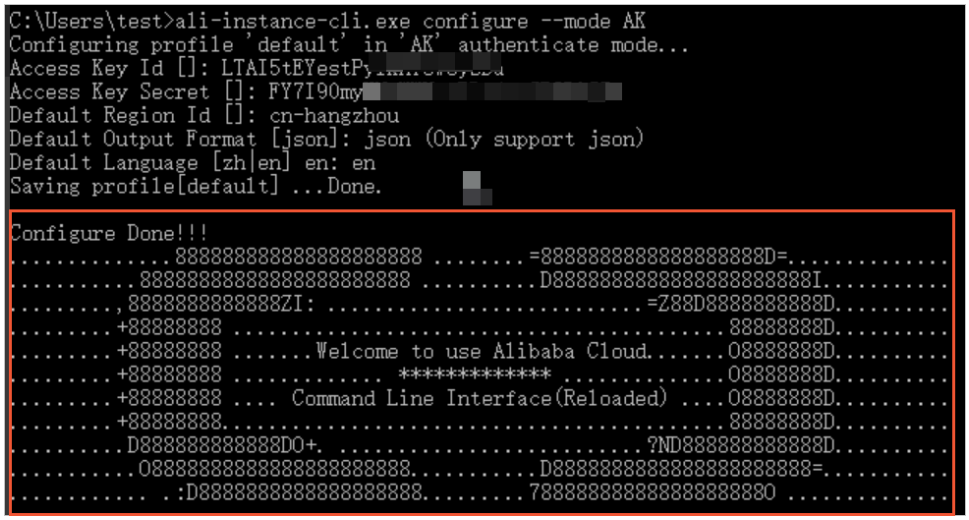
  ```
  ali-instance-cli.exe configure set --mode StsToken --region "region" --access-key
  -id "ak"  --access-key-secret "sk"   --sts-token "token"
  ```

- CredentialsURI-based authentication

  Run the following command and specify CredentialsURI and RegionID as prompted:

  ```
  ./ali-instance-cli configure --mode=CredentialsURI
  ```

A command output similar to the following one indicates that the AccessKey pair-based authentication method is configured.



5. Run the following command to connect to an instance:

   ```
   .\ali-instance-cli.exe session --instance instance-id
   ```

   > **Note** Replace *instance id* with the ID of the instance to which you want to connect.

   A command output similar to the following one indicates that you are connected to the instance by using session management.

   Windows instance to which you want to connect

FAQ

If an error occurs when you use a session management client, you can view logs to identify the error cause.

- View the log generated at the current time for the session management client. Example: `/home/test/log/aliyun_ecs_session_log.2022XXXX`.
- View logs of the Cloud Assistant client in one of the following directories based on the operating system type.
  - Linux

    ```
    /usr/local/share/aliyun-assist/<Version number of Cloud Assistant>/log/
    ```

  - Windows

    ```
    C:\ProgramData\aliyun\assist\<Version number of Cloud Assistant>\log
    ```

If the session management feature is not enabled when you use the session management client to connect to an instance, the `ssh_exchange_identification: Connection closed by remote host` error is reported. Additionally, the `session manager is disabled, please enable first` entry appears in the session management client log. You can enable the session management feature in the ECS console. For more information, see Connect to an instance by using session management.

7.3.4. Connect to an instance over SSH by using ali-instance-cli

Session management is a feature provided by Cloud Assistant that allows you to connect to Elastic Compute Service (ECS) instances in a secure and convenient manner. ali-instance-cli is a CLI used for session management. This topic describes how to use ali-instance-cli to connect to an ECS instance over SSH.

Prerequisites

- The Cloud Assistant client is installed on the ECS instance to which you want to connect. For a Windows instance, the installed client version must be 2.1.3.256 or later. For a Linux instance, the installed client version must be 2.2.3.256 or later. For more information, see Install the Cloud Assistant client.
- For information about how to enable the session management feature, see Connect to an instance by using session management.

Context

When you use ali-instance-cli to connect to an ECS instance over SSH, you need to only provide the ID and password of the instance. You do not need to expose the public IP address and port number of the instance. This connection method is more convenient and secure than using SSH or Virtual Network Console (VNC). For more information about session management, see How session management works.

Session management clients support Linux, macOS, and Windows operating systems and are used differently on these operating systems. For more information, see the following sections in this topic:

- Linux and macOS operating systems
- Windows operating systems

Linux and macOS operating systems

1. Log on to a session management client.
2. Install ali-instance-cli on the session management client.

   Run commands to install ali-instance-cli based on the following operating system types:

   - Linux

     ```
     curl -O https://aliyun-client-assist.oss-accelerate.aliyuncs.com/session-manager/linux/ali-instance-cli
     chmod a+x ali-instance-cli
     ```

   - macOS

     ```
     curl -O https://aliyun-client-assist.oss-accelerate.aliyuncs.com/session-manager/mac/ali-instance-cli
     chmod a+x ali-instance-cli
     ```

3. Create a file named `config` and add configurations to the file.

   i. Create the `.ssh` directory in the current working directory. In this example, */home/test* is used as the working directory.

   ```
   mkdir .ssh
   ```

   ii. Switch to the `.ssh` directory.

   ```
   cd .ssh
   ```

   iii. Create and open the `config` file.

   ```
   vim config
   ```

   iv. Press the `I` key to enter the edit mode.

   v. Add the following content to the `config` file.

   > **Note** Replace *ali-instance-cli* in the following command with the absolute path of the ali-instance-cli file. In this example, `/home/test/ali-instance-cli` is used.

   ```
   host i-*
       ProxyCommand sh -c "ali-instance-cli ssh -i '%h' --port  '%p'"
   ```

   vi. Press the `Esc` key to exit the edit mode.

   vii. Enter `:wq` and press the `Enter` key to save and close the file.

   viii. Grant the execute permissions on the `config` file.

   ```
   chmod 755 config
   ```

4. Configure an AccessKey pair, a Security Token Service (STS) token, or CredentialsURI.

   For information about how to obtain an AccessKey pair or STS token, see Obtain an AccessKey pair or What is STS?.

   i. Switch to the test directory.

   ```
   cd /home/test
   ```

ii. Configure an authentication method.

The following authentication methods are supported:

- AccessKey pair-based authentication

  Run the following command and enter an AccessKey ID, AccessKey secret, and region ID as prompted:

  ```
  ./ali-instance-cli configure --mode AK
  ```

- STS token-based authentication

  > **Note** Replace *region*, *ak*, *sk*, and *token* with the actual region ID, AccessKey ID, AccessKey secret, and STS token.

  ```
  ./ali-instance-cli configure set --mode StsToken --region "region" --access-key-id "ak"  --access-key-secret "sk"   --sts-token "token"
  ```

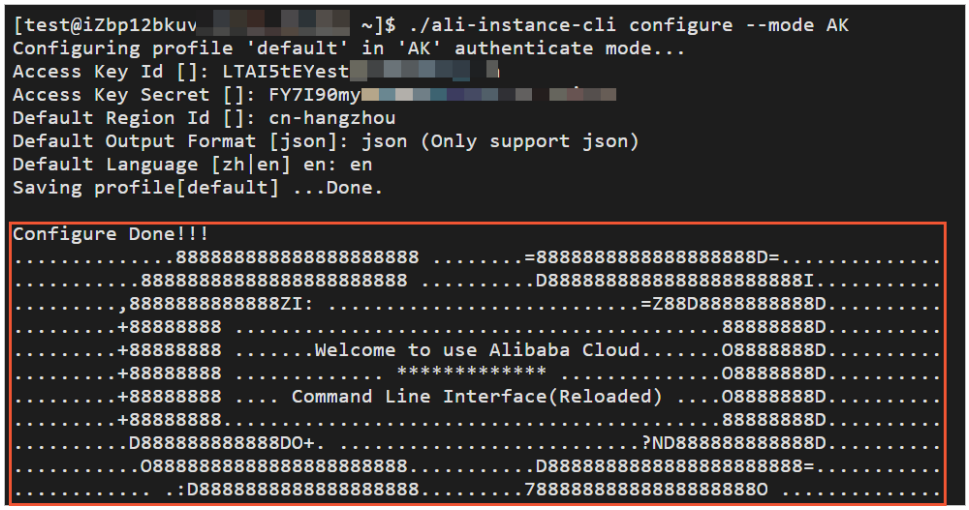
- CredentialsURI-based authentication

  Run the following command and specify CredentialsURI and RegionID as prompted.

  > **Note** Set the CredentialsURI value to the IP address of the authentication server that you configure.

  ```
  ./ali-instance-cli configure --mode=CredentialsURI
  ```

A command output similar to the following one indicates that the AccessKey pair-based authentication method is configured.



5. Run an SSH command to connect to an ECS instance.

You can use a username and password pair or a key pair to connect to the instance.

> **Note** Replace *user* and *aliyun instance id* with the actual username and ID of the instance.

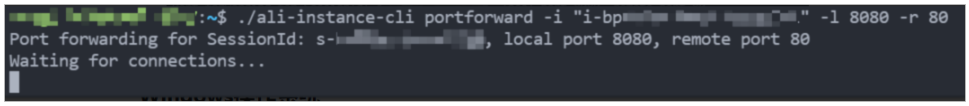
- Run the following command to connect to the instance with a username and password pair:

```
ssh user@aliyun instance id
```

- Run the following command to connect to the instance with a key pair:

```
ssh -i key.pem user@aliyun instance id
```

A command output similar to the following one indicates that you are connected to the instance over SSH by using session management.



Windows operating systems

Before you use a session management client that runs a Windows operating system to connect to an ECS instance, make sure that OpenSSH is installed on the client.For more information, see Use Cloud Assistant to install OpenSSH on an ECS Windows instance.

1. Log on to a session management client.

   For more information, see Connection methodsGuidelines on instance connection.

2. Download ali-instance-cli to the session management client.

   Download and save ali-instance-cli.exe for Windows to a directory on the session management client. In this example, the `C:\Users\test` directory is used.

3. Create a file named `config` and add configurations to the file.

   i. In the *C:\Users\<Username>* directory, create a folder named `.ssh` .

   > **Note** Replace *C:\Users\<Username>* with the actual directory. In this example, `C:\Users\test` is used.

   ii. In the `.ssh` folder, create a file named `config` .

   iii. Add the following content to the `config` file.

   Replace *ali-instance-cli.exe* with the absolute path of the ali-instance-cli.exe file. In this example, `C:\Users\test\ali-instance-cli.exe` is used.

```
host i-*
    ProxyCommand C:\Windows\System32\WindowsPowerShell\v1.0\powershell.exe "ali-ins
tance-cli.exe ssh -i '%h' --port  '%p'"
```

4. Configure an AccessKey pair or an STS token.

   For information about how to obtain an AccessKey pair or STS token, see Obtain an AccessKey pair or What is STS?.

   i. Choose **Start > Run**, enter **cmd**, and then press the `Enter` key to open the Command Prompt window.

   ii. Switch to the test directory.

```
cd C:\Users\test
```

iii. Configure an authentication method.

The following authentication methods are supported:

- AccessKey pair-based authentication

  Run the following command and enter an AccessKey ID, AccessKey secret, and region ID as prompted:

  ```
  ali-instance-cli.exe configure --mode AK
  ```

- STS token-based authentication

  > **Note** Replace *region*, *ak*, *sk*, and *token* with the actual region ID, AccessKey ID, AccessKey secret, and STS token.

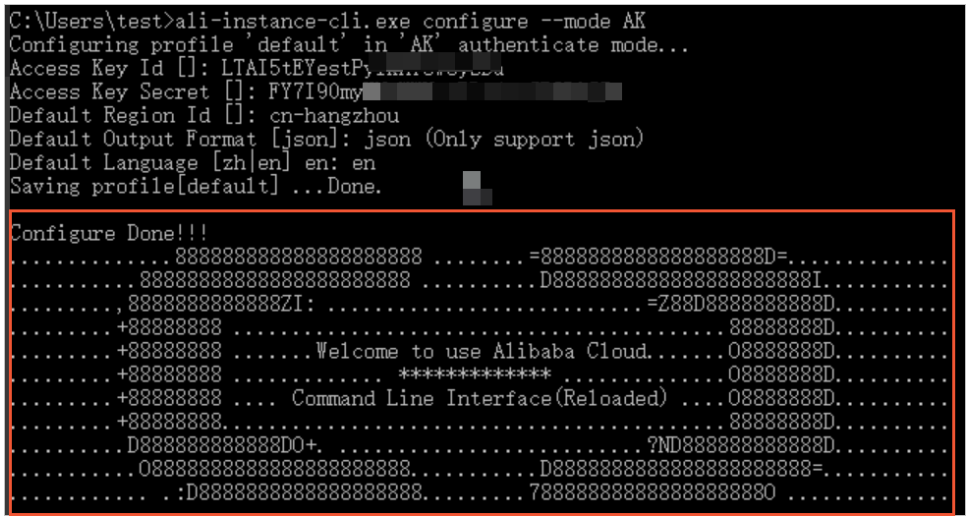
  ```
  ali-instance-cli.exe configure set --mode StsToken --region "region" --access-key
  -id "ak"  --access-key-secret "sk"   --sts-token "token"
  ```

- CredentialsURI-based authentication

  Run the following command and specify CredentialsURI and RegionID as prompted:

  ```
  ./ali-instance-cli configure --mode=CredentialsURI
  ```

A command output similar to the following one indicates that the AccessKey pair-based authentication method is configured.



5. Run an SSH command to connect to an ECS instance.

   You can use a username and password pair or a key pair to connect to the instance.

   > **Note** Replace *user* and *aliyun instance id* with the actual username and ID of the instance.

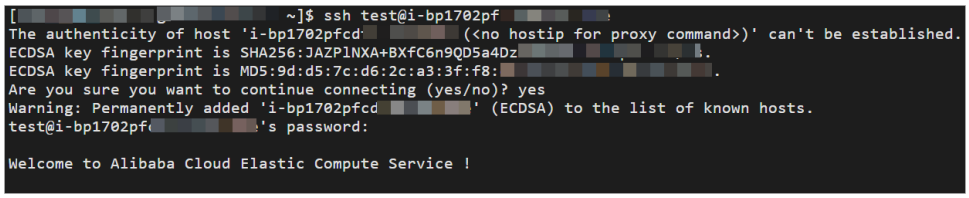
   - Run the following command to connect to the instance with a username and password pair:

     ```
     ssh user@aliyun instance id
     ```

   - Run the following command to connect to the instance with a key pair:

```
ssh -i key.pem user@aliyun instance id
```

A command output similar to the following one indicates that you are connected to the instance over SSH by using session management.



FAQ

If an error occurs when you use a session management client, you can view logs to identify the error cause.

- View the log generated at the current time for the session management client. Example: `/home/test/log/aliyun_ecs_session_log.2022XXXX`.
- View logs of the Cloud Assistant client in one of the following directories based on the operating system type.
  - Linux

    ```
    /usr/local/share/aliyun-assist/<Version number of Cloud Assistant>/log/
    ```

  - Windows

    ```
    C:\ProgramData\aliyun\assist\<Version number of Cloud Assistant>\log
    ```

If the session management feature is not enabled when you use the session management client to connect to an instance, the `ssh_exchange_identification: Connection closed by remote host` error is reported. Additionally, the `session manager is disabled, please enable first` entry appears in the session management client log. You can enable the session management feature in the ECS console. For more information, see Connect to an instance by using session management.

7.3.5. Perform port forwarding by using ali-instance-cli

Session management is a feature provided by Cloud Assistant that allows you to connect to Elastic Compute Service (ECS) instances in a secure and convenient manner. ali-instance-cli is a CLI tool provided by the session management feature. This topic describes how to use ali-instance-cli to forward network traffic from a local port of your computer to an ECS instance.

Prerequisites

- The Cloud Assistant client is installed on the ECS instance to which you want to connect. For a Windows instance, the installed client version must be 2.1.3.256 or later. For a Linux instance, the installed client version must be 2.2.3.256 or later. For more information, see Install the Cloud Assistant client.
- For information about how to enable the session management feature, see Connect to an instance by using session management.

> **Note** The ali-instance-cli port forwarding feature is in invitational preview. To use this feature, submit a ticket.

Context

When you use ali-instance-cli to set up port forwarding on an instance, you do not need to provide the public IP address of the instance but need only to provide the ID and a port number of the instance. Then, you can use a session management client to forward network traffic from a local port of your computer to the instance. This way, you can access the services on the ECS instance in a secure and convenient manner. For more information about session management, see How session management works.

Session management clients support Linux, macOS, and Windows operating systems and are used differently on these operating systems. For more information, see the following sections in this topic:

- Linux and macOS operating systems
- Windows operating systems

Linux and macOS operating systems

1. Log on to a session management client.
2. Install ali-instance-cli on the session management client.

   Run commands to install ali-instance-cli based on the following operating system types:

   - Linux

     ```
     curl -O https://aliyun-client-assist.oss-accelerate.aliyuncs.com/session-manager/linux/ali-instance-cli
     chmod a+x ali-instance-cli
     ```

   - macOS

     ```
     curl -O https://aliyun-client-assist.oss-accelerate.aliyuncs.com/session-manager/mac/ali-instance-cli
     chmod a+x ali-instance-cli
     ```

3. Configure an AccessKey pair, a Security Token Service (STS) token, or CredentialsURI.

   For information about how to obtain an AccessKey pair or STS token, see Obtain an AccessKey pair or What is STS?.

   i. Switch to the test directory.

      ```
      cd /home/test
      ```

ii. Configure an authentication method.

The following authentication methods are supported:

- AccessKey pair-based authentication

  Run the following command and enter an AccessKey ID, AccessKey secret, and region ID as prompted:

  ```
  ./ali-instance-cli configure --mode AK
  ```

- STS token-based authentication

  > **Note** Replace *region*, *ak*, *sk*, and *token* with the actual region ID, AccessKey ID, AccessKey secret, and STS token.

  ```
  ./ali-instance-cli configure set --mode StsToken --region "region" --access-key-id "ak"  --access-key-secret "sk"   --sts-token "token"
  ```

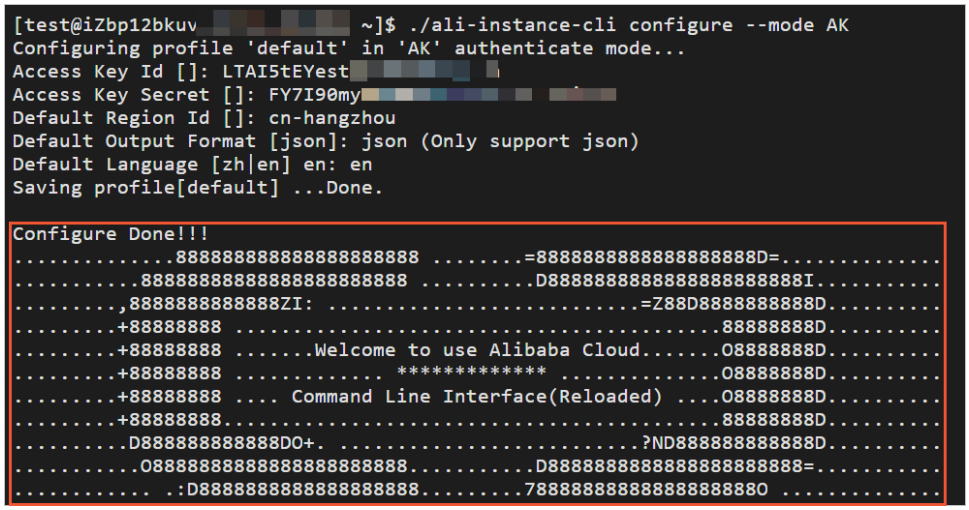
- CredentialsURI-based authentication

  Run the following command and specify CredentialsURI and RegionID as prompted.

  > **Note** Set the CredentialsURI value to the IP address of the authentication server that you configure.

  ```
  ./ali-instance-cli configure --mode=CredentialsURI
  ```

A command output similar to the following one indicates that the AccessKey pair-based authentication method is configured.

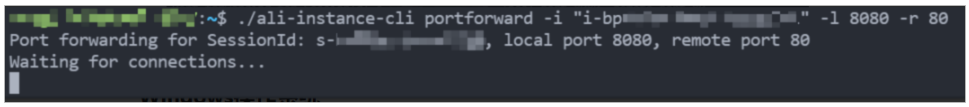


4. Run the following command to forward network traffic from a local port of your computer to an ECS instance:

   ```
   ./ali-instance-cli portforward -i "instance-id" -l 8080 -r 80
   ```

> **Note** This example demonstrates how to use session management for port forwarding. In this example, local port 8080 and port 80 of an instance are used. You can specify another port based on your needs and replace *instance id* with the actual instance ID.

A command output similar to the following one indicates that a forwarding path is established from the specified local port to the specified instance port by using session management.



Windows operating systems

Before you use a session management client that runs a Windows operating system to connect to an ECS instance, make sure that OpenSSH is installed on the client.For more information, see Use Cloud Assistant to install OpenSSH on an ECS Windows instance.

1. Log on to a session management client.

   For more information, see Connection methodsGuidelines on instance connection.

2. Download ali-instance-cli to the session management client.

   Download and save ali-instance-cli.exe for Windows to a directory on the session management client. In this example, the `C:\Users\test` directory is used.

3. Create a file named `config` and add configurations to the file.

   i. In the *C:\Users\<Username>* directory, create a folder named `.ssh`.

   > **Note** Replace *C:\Users\<Username>* with the actual directory. In this example, `C:\Users\test` is used.

   ii. In the `.ssh` folder, create a file named `config`.

   iii. Add the following content to the `config` file.

   Replace *ali-instance-cli.exe* with the absolute path of the ali-instance-cli.exe file. In this example, `C:\Users\test\ali-instance-cli.exe` is used.

   ```
   host i-*
       ProxyCommand C:\Windows\System32\WindowsPowerShell\v1.0\powershell.exe "ali-instance-cli.exe ssh -i '%h' --port  '%p'"
   ```

4. Configure an AccessKey pair or an STS token.

   For information about how to obtain an AccessKey pair or STS token, see Obtain an AccessKey pair or What is STS?.

   i. Choose **Start > Run**, enter **cmd**, and then press the `Enter` key to open the Command Prompt window.

   ii. Switch to the test directory.

   ```
   cd C:\Users\test
   ```

iii. Configure an authentication method.

The following authentication methods are supported:

- AccessKey pair-based authentication

  Run the following command and enter an AccessKey ID, AccessKey secret, and region ID as prompted:

  ```
  ali-instance-cli.exe configure --mode AK
  ```

- STS token-based authentication

  > **Note** Replace *region*, *ak*, *sk*, and *token* with the actual region ID, AccessKey ID, AccessKey secret, and STS token.

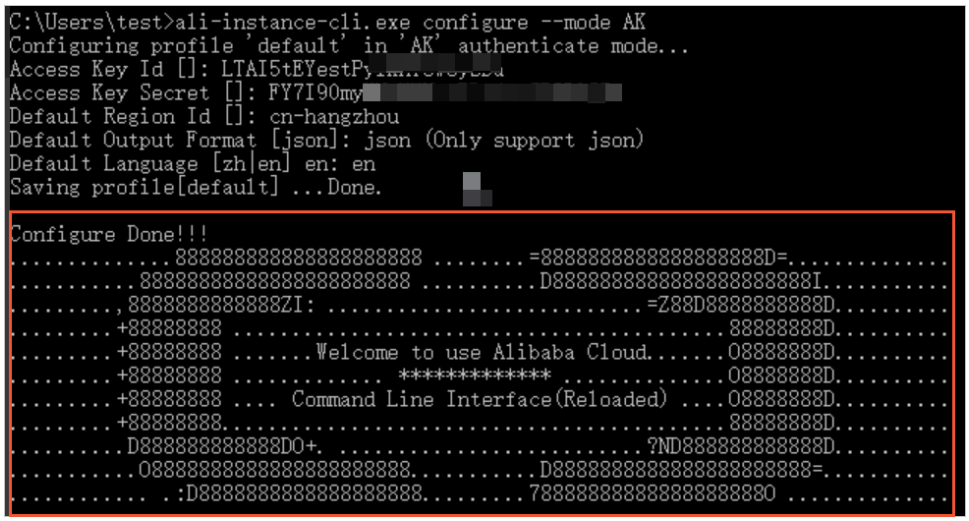
  ```
  ali-instance-cli.exe configure set --mode StsToken --region "region" --access-key
  -id "ak"  --access-key-secret "sk"   --sts-token "token"
  ```

- CredentialsURI-based authentication

  Run the following command and specify CredentialsURI and RegionID as prompted:

  ```
  ./ali-instance-cli configure --mode=CredentialsURI
  ```

A command output similar to the following one indicates that the AccessKey pair-based authentication method is configured.



5. Run the following command to forward network traffic from a local port of your computer to an ECS instance:

   ```
   ali-instance-cli.exe portforward -i "instance-id" -l 8080 -r 80
   ```

   > **Note** This example demonstrates how to use session management for port forwarding. In this example, local port 8080 and port 80 of an instance are used. You can specify another port based on your needs and replace *instance id* with the actual instance ID.

A command output similar to the following one indicates that a forwarding path is established from the specified local port to the specified instance port by using session management.

Use case: Access the MySQL service on an ECS instance

Assume that the MySQL service is running on port 3306 on your instance. Your computer on which the session management client is installed runs a Linux operating system. You can use ali-instance-cli to access services on an instance from local port 33306 of your computer.

1. Run the following command to forward traffic from local port 33306 to port 3306 of the instance:

   ```
   ./ali-instance-cli portforward -i "instance-id" -l 33306 -r 3306
   ```

   A command output similar to the following one indicates that a forwarding path is established from the specified local port to the specified instance port by using session management.

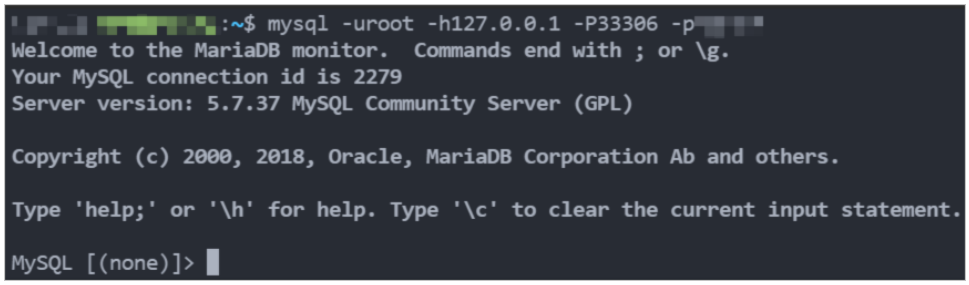
   

2. Run the following command to access the MySQL service on the instance from your computer:

   ```
   mysql -uroot -h127.0.0.1 -uroot -ppassword --port=33306
   ```

   > **Note** Replace *password* with the MySQL password.

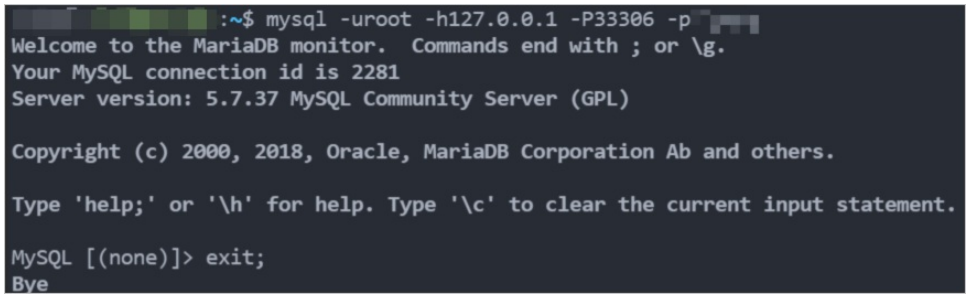
   A command output similar to the following one indicates that the MySQL service on the instance is accessed.

   The ali-instance-cli log shows that a new connection is accepted, which is the connection initiated by the MySQL client.

3. Run the `exit` command to close the connection to MySQL.

The ali-instance-cli log shows that the connection is closed.

FAQ

If an error occurs when you use a session management client, you can view logs to identify the error cause.

- View the log generated at the current time for the session management client. Example: `/home/test/log/aliyun_ecs_session_log.2022XXXX`.
- View logs of the Cloud Assistant client in one of the following directories based on the operating system type.
  - Linux

    ```
    /usr/local/share/aliyun-assist/<Version number of Cloud Assistant>/log/
    ```

  - Windows

    ```
    C:\ProgramData\aliyun\assist\<Version number of Cloud Assistant>\log
    ```

If the session management feature is not enabled when you use the session management client to connect to an instance, the `ssh_exchange_identification: Connection closed by remote host` error is reported. Additionally, the `session manager is disabled, please enable first` entry appears in the session management client log. You can enable the session management feature in the ECS console. For more information, see Connect to an instance by using session management.

7.3.6. Connect to a Linux instance by using the config_ecs_instance_connect plug-in

This topic describes how to use the config_ecs_instance_connect plug-in to connect to an Elastic Compute Service (ECS) instance that runs a Linux operating system.

Prerequisites

- The session management feature is enabled for the ECS instance to which you want to connect in your Alibaba Cloud account. For more information, see Connect to an instance by using session management.
- The Cloud Assistant client is installed on the ECS instance, and the version of the client supports session management. If an instance runs a Linux operating system, the installed client version must be 2.2.3.196 or later. If an instance runs a Windows operating system, the installed client version must be 2.1.3.196 or later. For more information, see Install the Cloud Assistant client.
- The ECS instance runs a Linux operating system.

Context

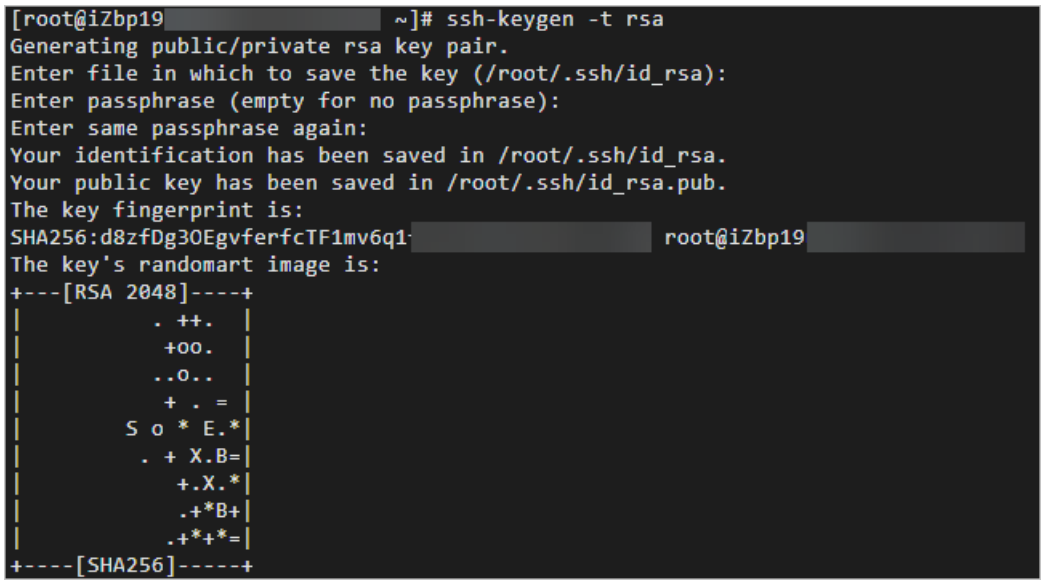
You can use the config_ecs_instance_connect plug-in and a common Cloud Assistant command to send an SSH public key to a specified instance for a specified user to use. The SSH public key is stored on the instance for 60 seconds. During these 60 seconds, you can use the SSH public key to log on to the instance as the specified user without a password.

Procedure

1. Run the following command on a session management client to generate a Rivest-Shamir-Adleman (RSA) public key and key file:

   ```
   ssh-keygen -t rsa
   ```

   Press the Enter key as prompted. A command output similar to the following one indicates that the public key and key file are generated.

   > **Note** The default path of the generated public key is ~/.ssh/id_rsa.pub.

2. Use Cloud Assistant to run the following command to install and enable the config_ecs_instance_connect plug-in on the instance to which you want to connect.

   For more information, see Run a command.

   ```
   acs-plugin-manager -e -P config_ecs_instance_connect --params --install
   ```

3. Send the SSH public key to the instance.

   You can use one of the following methods to send the SSH public key to the instance:

   - Call API operations.
     a. Call the DescribeCommands operation to query the common Cloud Assistant command named `ACS-ECS-SendSshPublicKey-linux.sh`.

```
import com.aliyuncs.DefaultAcsClient;
import com.aliyuncs.IAcsClient;
import com.aliyuncs.exceptions.ClientException;
import com.aliyuncs.exceptions.ServerException;
import com.aliyuncs.profile.DefaultProfile;
import com.google.gson.Gson;
import java.util.*;
import com.aliyuncs.ecs.model.v20140526.*;
public class DescribeCommands {
    public static void main(String[] args) {
        DefaultProfile profile = DefaultProfile.getProfile("cn-beijing", "<access
KeyId>", "<accessSecret>");
        IAcsClient client = new DefaultAcsClient(profile);
        DescribeCommandsRequest request = new DescribeCommandsRequest();
        request.setRegionId("cn-beijing");
        request.setProvider("AlibabaCloud");
        request.setName("ACS-ECS-SendSshPublicKey-linux.sh");
        try {
            DescribeCommandsResponse response = client.getAcsResponse(request);
            System.out.println(new Gson().toJson(response));
        } catch (ServerException e) {
            e.printStackTrace();
        } catch (ClientException e) {
            System.out.println("ErrCode:" + e.getErrCode());
            System.out.println("ErrMsg:" + e.getErrMsg());
            System.out.println("RequestId:" + e.getRequestId());
        }
    }
}
```

In the response, find the array in which the `Latest` value is `true` and obtain the `CommandId` value.

```
{
    "TotalCount": 1,
    "PageSize": 10,
    "RequestId": "8D7DC6FF-6849-5927-XXXX-FBE1027FEFDE",
    "PageNumber": 1,
    "Commands": {
        "Command": [
            {
                "Description": "Sends SSH public keys.",
                "Category": "Alibaba Cloud-ECS-Application Installation",
                "ParameterNames": {
                    "ParameterName": [
                        "username",
                        "sshpublickey"
                    ]
                },
                "Timeout": 60,
                "Provider": "AlibabaCloud.ECS.Applications",
                "Name": "ACS-ECS-SendSshPublicKey-linux.sh",
                "WorkingDir": "",
                "CommandContent": "c2VuZF9zc2hfcHVibGljX2tleSAtLXVzZXJuYW1lIHt7dX
Nlcm5hbWV9fSAtLXNzaC1wdWJsaWMta2V5IHt7c3NoUHVibGljS2V5****",
                "Type": "RunShellScript",
                "Version": 1,
                "InvokeTimes": 40,
                "CreationTime": "2022-04-13T07:28Z",
                "Latest": true,
                "EnableParameter": true,
                "CommandId": "c-xxxxxx"
            }
        ]
    }
}
```


b. Call the InvokeCommand operation to send the SSH public key to the instance to which you want to connect.

   Take note of the following parameters. For more information about the parameters of the InvokeCommand operation, see InvokeCommand.

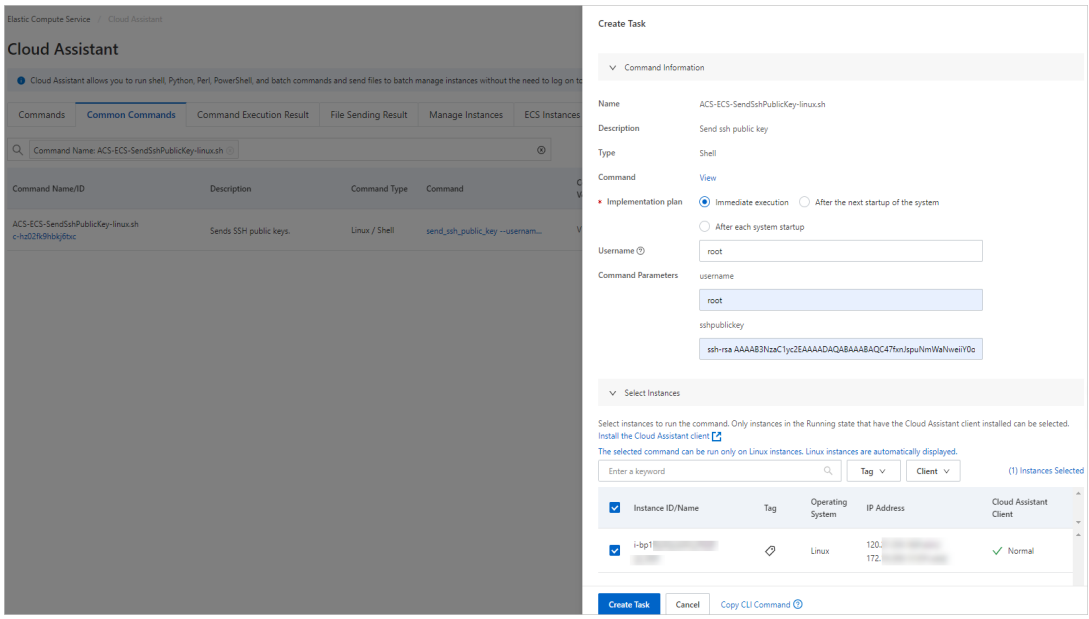
   - CommandId: the ID of the command. Set the value to the CommandId value obtained in the previous step.
   - username: the username used to connect to the instance. Default value: root.
   - sshpublickey: required. The content of the SSH public key. Set the value to the content of the id_rsa.pub file generated in Step 1.

```
import com.aliyuncs.DefaultAcsClient;
import com.aliyuncs.IAcsClient;
import com.aliyuncs.exceptions.ClientException;
import com.aliyuncs.exceptions.ServerException;
import com.aliyuncs.profile.DefaultProfile;
import com.google.gson.Gson;
import java.util.*;
import com.aliyuncs.ecs.model.v20140526.*;
public class InvokeCommand {
    public static void main(String[] args) {
        DefaultProfile profile = DefaultProfile.getProfile("cn-beijing", "<access
KeyId>", "<accessSecret>");
        IAcsClient client = new DefaultAcsClient(profile);
        InvokeCommandRequest request = new InvokeCommandRequest();
        request.setRegionId("cn-beijing");
        request.setCommandId("c-xxxxxx");
        List<String> instanceIdList = new ArrayList<String>();
        instanceIdList.add("i-xxxxxx");
        request.setInstanceIds(instanceIdList);
        request.setParameters("{\"username\":\"root\", \"sshpublickey\":\"ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAABgQDftEm8H5A19FXv5SCVzHqmS9vg+8B4wsp9M/U/vKwPMlM0fJr8Z
52ErGnEnpFA24hLSf/Ffpht19tp+QtsYhVcg xxx\"}");
        try {
            InvokeCommandResponse response = client.getAcsResponse(request);
            System.out.println(new Gson().toJson(response));
        } catch (ServerException e) {
            e.printStackTrace();
        } catch (ClientException e) {
            System.out.println("ErrCode:" + e.getErrCode());
            System.out.println("ErrMsg:" + e.getErrMsg());
            System.out.println("RequestId:" + e.getRequestId());
        }
    }
}
```

- Use Cloud Assistant.
  a. Log on to the ECS console.
  b. In the left-side navigation pane, choose **Maintenance & Monitoring > ECS Cloud Assistant**.
  c. Click the **Common Commands** tab.
  d. Find ACS-ECS-SendSshPublicKey-linux.sh in the Command Name/ID column and click **Create Task** in the Actions column.

e. In the **Create Task** panel, configure parameters.

- Take note of the following parameters:
  - **username**: the username used to connect to the instance. Default value: root.
  - **sshpublickey**: required. The content of the SSH public key. Set the value to the content of the id_rsa.pub file generated in Step 1.
- Select Instances: Select the instance to which you want to connect.



f. Click **Create Task**.

4. Connect to the instance without a password.

   You can use the public IP address or ID of the instance to connect to the instance without a password.

   - Connect to the instance by using its public IP address without a password.

     ```
     ssh -i ~/.ssh/id_rsa username@instance_ip
     ```

     Note In the preceding command, replace ~/.ssh/id_rsa with the actual path of the SSH public key, *username* with the username used to connect to the instance, and *instance_ip* with the public IP address of the instance.

   - Connect to the instance by using its ID without a password.

     ```
     ssh -i ~/.ssh/id_rsa username@instance_id
     ```

     Note

     - Make sure that ali-instance-cli is installed on the session management client. For more information, see Linux and macOS operating systems or Windows operating systems.
     - In the preceding command, replace ~/.ssh/id_rsa with the actual path of the SSH public key, *username* with the username used to connect to the instance, and *instance_id* with the ID of the instance.

FAQ

When I attempt to connect to the instance in password-free mode, why am I still prompted for a password?

A public key remains valid for only 60 seconds after it is registered with an instance. Check whether your public key has expired.

An error is reported when I install the config_ecs_instance_connect plug-in by using Cloud Assistant or use the common Cloud Assistant command. How do I identify the issue?

You can view logs to identify the issue.

- View logs of the Cloud Assistant client in the following path:

```
/usr/local/share/aliyun-assist/Version number of Cloud Assistant/log/aliyun_assist_main.log.*
```

- View acs_plugin_manager logs in the following path:

```
/usr/local/share/aliyun-assist/Version number of Cloud Assistant/log/acs_plugin_manager.log
```

7.4. Connect to an instance by using Alibaba Cloud Client

7.4.1. Overview of Alibaba Cloud Client

Alibaba Cloud Client is a client provided by Alibaba Cloud that can be used to query, view, and connect to Elastic Compute Service (ECS) instances, elastic container instances, simple application servers, and instances that are managed by Alibaba Cloud. Alibaba Cloud Client allows you to use cloud services in a more convenient and efficient manner. Alibaba Cloud Client can only be installed on on-premises machines that run a macOS operating system with M1 chips or a 64-bit macOS operating system.

Features

The following table describes the features that Alibaba Cloud Client provides to query, view, and connect to specific cloud services.

Cloud service	Feature description

Cloud service	Feature description

ECS	• You can view ECS instances by region, and query ECS instances in a specified region or in all regions.
	• The client supports the following methods to connect to instances:
	  ◦ Connect to instances over SSH.
	  ▪ For an instance that has a public IP address, connect to the instance by using the public IP address.
	  ▪ For an instance that has no public IP address, connect to the instance over SSH by using session management that Cloud Assistant provides.
	  ◦ Connect to instances by using session management.
	  ◦ Connect to instances by using port forwarding.
	• You can use the client to manage passwords and certificates.
	You can choose to store instance logon information to Alibaba Cloud Key Management Service (KMS) or on-premises files. For information about KMS, see What is Key Management Service?

Elastic Container Instance	• You can view elastic container instances by region. • You can view events of container groups. • You can connect to a container in a container group.
Simple Application Server	• You can view simple application servers by region. • You can connect to simple application servers over SSH.

Alibaba Cloud managed instance	• You can view instances that are managed by Alibaba Cloud by region.
	• The client supports the following methods to connect to instances:
	  ◦ Connect to instances over SSH.
	  ◦ Connect to instances by using session management.

Download methods

You can click one of the following links to download Alibaba Cloud Client based on the operating system of your machine and your business requirements:

- Download link for 64-bit macOS operating systems
- Download link for macOS operating systems with M1 chips

7.4.2. Add one or more accounts to Alibaba Cloud Client

The first time you use Alibaba Cloud Client, you must add your account information to it. You can add multiple Alibaba Cloud accounts and then switch between these accounts when you use Alibaba Cloud Client.

Prerequisites

Alibaba Cloud Client is downloaded and installed. For information about how to download Alibaba Cloud Client for different operating system versions, see Download methods.

Context

If you have configured account information in Alibaba Cloud CLI, Alibaba Cloud Client imports the account information from Alibaba Cloud CLI.

Procedure

1. In the upper-left corner of the Alibaba Cloud Client homepage, click **Config Profile**.
2. On the Profiles page, set **Name**.

   Specify an appropriate name to make the account easy to identify.
3. Set **Mode** to specify account management mode.

   The following table describes the valid values of this parameter.

Value	Description	Related parameters
AccessKey	AccessKey pairs are the credentials used by Alibaba Cloud accounts or Resource Access Management (RAM) users to access Alibaba Cloud APIs and have full permissions on resources within the accounts. An AccessKey pair consists of an AccessKey ID and an AccessKey secret. For information about how to obtain an AccessKey pair, see Obtain an AccessKey pair.	You must configure the **AccessKeyId**, **AccessSecret**, and **Default Region** parameters. ○ **AccessKeyId**: Enter the AccessKey ID of your account. AccessKey IDs are used to verify user identities. ○ **AccessSecret**: Enter the AccessKey secret of your account. AccessKey secrets are used to verify user keys. ○ **Default Region**: Select a region to display by default when your account uses Alibaba Cloud Client. When your account accesses the list of Elastic Compute Service (ECS) instances in Alibaba Cloud Client, the information of instances that reside within the selected region is displayed by default.

Value	Description	Related parameters
StsToken	Alibaba Cloud Security Token Service (STS) allows you to manage temporary credentials for your Alibaba Cloud resources. RAM provides two types of identity: RAM users and RAM roles. A RAM role does not have permanent identity credentials and can only be assumed to access Alibaba Cloud resources by using an STS token. When the STS token is issued, you can specify a validity period and access permissions for the token. For more information, see What is STS?.	You must configure the **AccessKeyId**, **AccessSecret**, **STS Token**, and **Default Region** parameters. ○ **AccessKeyId**: Enter the AccessKey ID of your account. AccessKey IDs are used to verify user identities. ○ **AccessSecret**: Enter the AccessKey secret of your account. AccessKey secrets are used to verify user keys. ○ **STS Token**: Enter an STS token. STS tokens are temporary identity credentials and have user-defined validity periods and access permissions. ○ **Default Region**: Select a region to display by default when your account uses Alibaba Cloud Client. When your account accesses the list of ECS instances in Alibaba Cloud Client, the information of instances that reside within the selected region is displayed by default.
		You must configure the **AccessKeyId**, **AccessSecret**, **RamRoleARN**, **RoleSessionName**, and **Default Region** parameters. ○ **AccessKeyId**: Enter the AccessKey ID of your account. AccessKey IDs are used to verify user identities. ○ **AccessSecret**: Enter the AccessKey secret of your account. AccessKey secrets are used to verify user keys. ○ **RamRoleARN**: Enter the Alibaba Cloud Resource Name (ARN) of a RAM role. ■ The trusted entity of the RAM role is an Alibaba Cloud account. For more information, see Create a RAM role for a trusted Alibaba Cloud account or CreateRole. ■ Format: *acs:ram::<account_id>:role/<role_name>*. ■ You can view the ARNs of RAM roles by using the RAM console or by calling API operations. For more information, see FAQ about RAM roles and STS tokens, GetRole, or ListRoles.

Value	Description	Related parameters
RamRoleArn	Use a RAM user to assume a RAM role to automatically apply for and maintain an STS token. For more information, see AssumeRole.	◦ (Optional)**RoleSessionName**: Enter a name for the role session. ■ You can specify the value of this parameter based on your business requirements. In most cases, you can set this parameter to the identity of the user who calls API operations. For example, set this parameter to a username. In ActionTrail logs, you can use the RoleSessionName value to distinguish different users who assume the same RAM role to perform operations. This way, you can perform user-specific auditing. ■ The value must be 2 to 64 characters in length and can contain letters, digits, periods (.), at signs (@), hyphens (-), and underscores (_). ◦ Default Region: Select a region to display by default when your account uses Alibaba Cloud Client. When your account accesses the list of ECS instances in Alibaba Cloud Client, the information of instances that reside within the selected region is displayed by default.

Value	Description	Related parameters
CredentialsURI	Obtain identity credentials from the local or remote uniform resource identifier (URI). You can use Alibaba Cloud CLI to verify the CredentialsURI parameter. You can use the CredentialsURI parameter to create an account in Alibaba Cloud Client only if a status code of 200 and a JSON string in the following format are returned. For more information, see aliyun/aliyun-cli: Alibaba Cloud CLI. `{ "Code": "Success", //The status code. "AccessKeyId": "<ak id>", //The AccessKey ID. "AccessKeySecret": "<ak secret>", //The AccessKey secret. "SecurityToken": "<security token>", //The STS token. "Expiration" "2006-01-02T15:04:05Z" //The expiration time of the STS token in UTC. }`	You must configure the **CredentialsURI** and **Default Region** parameters. ○ **CredentialsURI**: Enter the local or remote URI from which to obtain identity credentials. ○ **Default Region**: Select a region to display by default when your account uses Alibaba Cloud Client. When your account accesses the list of ECS instances in Alibaba Cloud Client, the information of instances that reside within the selected region is displayed by default.
ExternalCommand	Run an external command to obtain identity credentials. Alibaba Cloud Client runs the command and returns the command output as identity credentials for you to create an account. You can use Alibaba Cloud CLI to verify the external command. You can use the command to create an account in Alibaba Cloud Client only if a JSON string that contains identity credentials is returned. For more information, see aliyun/aliyun-cli: Alibaba Cloud CLI.	You must configure the **External Command** and **Default Region** parameters. ○ **External Command**: Enter the external command that you want to run to obtain identity credentials. ○ **Default Region**: Select a region to display by default when your account uses Alibaba Cloud Client. When your account accesses the list of ECS instances in Alibaba Cloud Client, the information of instances that reside within the selected region is displayed by default.

4. Click **Test** to verify whether the account information that you specified is valid.
5. If all of the account information is valid, click **Save**.

7.4.3. Manage ECS instances by using Alibaba Cloud Client

You can use Alibaba Cloud Client to view and connect to Elastic Compute Service (ECS) instances, elastic container instances, simple application servers, and instances managed by Alibaba Cloud. This topic describes how to use Alibaba Cloud Client to view, connect to, and manage ECS instances.

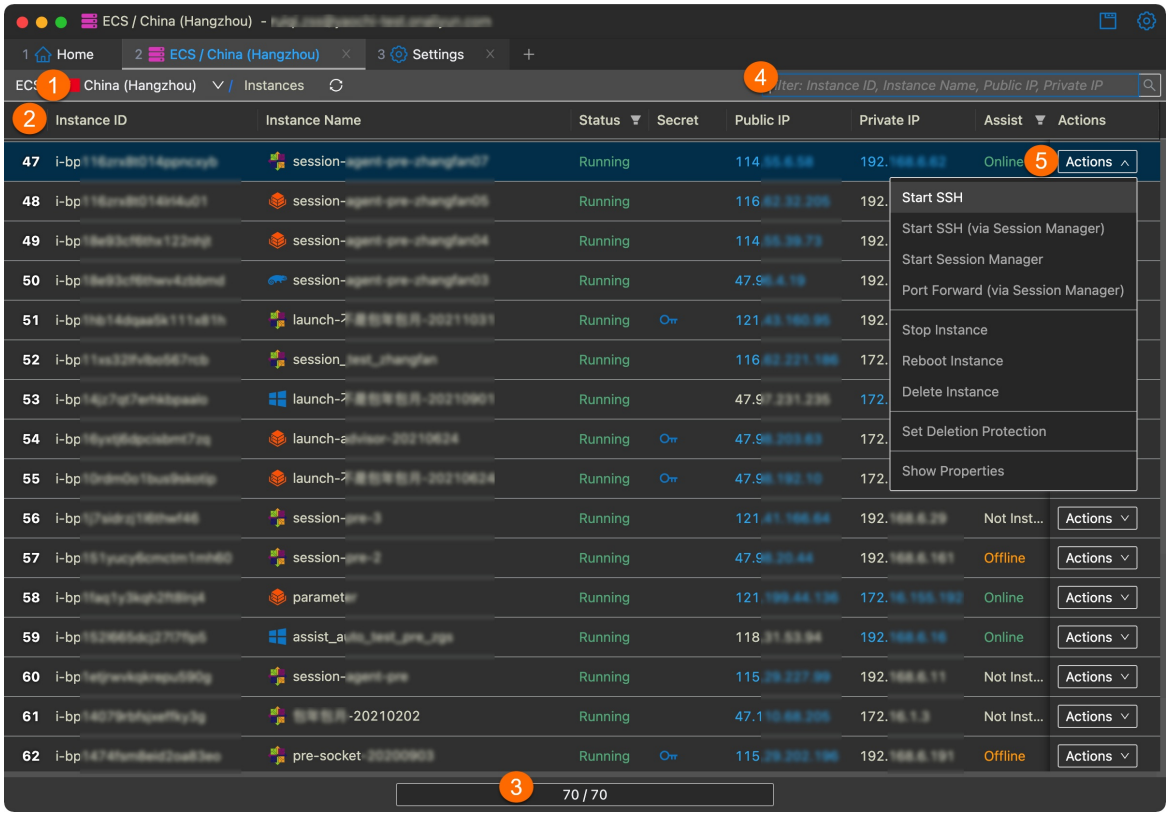
Prerequisites

- One or more accounts are added to Alibaba Cloud Client. For more information, see Add one or more accounts to Alibaba Cloud Client.
- Session management can be used to connect to ECS instances in a variety of manners, such as over SSH and by using port forwarding. Both Alibaba Cloud accounts and Resource Access Management (RAM) users can use session management, but only Alibaba Cloud accounts have permissions to enable session management.
  - If the session management feature is not enabled, use an Alibaba Cloud account or contact the owner of an Alibaba Cloud account to enable this feature. For more information, see Connect to an instance by using session management.
  - If you want to use the session management feature as a RAM user, make sure that you are authorized to call the StartTerminalSession operation. For more information about sample policies attached to RAM users, see the "Sample policies" section in Sample policies. Make sure that your RAM users are granted secure and controllable permissions. Proceed with caution when you attach policies to RAM users to prevent unauthorized operations caused by improper management of or unintended authorizations to RAM users.

> **Note** If one or more of the preceding prerequisites are not met, **Start Session Manager** becomes unavailable in the **Actions** column on the Instances page of Alibaba Cloud Client.

View ECS instances

1. On the homepage of Alibaba Cloud Client, click **ECS**.
2. On the Instances page, you can view information of ECS instances, as shown in the following figure.



- ①: This section shows the location information of ECS instances. You can click the icon on the right side of a region to switch to another region.
- ②: This section shows the list of ECS instances.
- ③: The pagination toolbar. You can navigate through pages in the pagination toolbar.
- ④: This section shows the instance search box.
  a. Enter an ECS instance by instance ID, public IP address, private IP address, or instance name.
  b. Press the **Enter** key to search for instances in the current region. You can enter an instance name for fuzzy search.
- ⑤: This section shows the Actions column.
  - Actions to connect to instances
  - Actions to start, stop, and release instances
  - Actions to configure release protection settings (only for pay-as-you-go instances)
  - Actions to view more instance details

Connect to an ECS instance

Connect to an ECS instance over SSH

1. On the homepage of Alibaba Cloud Client, click **ECS**.
2. In the top navigation bar, select the region of the instance to which you want to connect.
3. Find the instance to which you want to connect and choose **Actions > Start SSH** in the **Actions** column.
4. Specify **Username**. Default value: root. You can use the default value. Specify **Port Num**. Default

value: 22. You can use the default value. Set **Certifier**.

Valid values of Certifier:

- Password: Enter the password of the ECS instance.

  > **Note** In Alibaba Cloud Client, you can click the **Settings** icon in the upper-right corner. Then, on the Settings page, click **SSH** in the left-navigation pane. Set the Save Password to parameter to **Local File** to save the instance password to the /.aliyun/secrets.json file on your computer, or to **KMS** to save the instance password to Key Management Service (KMS).

- Identity: Select an SSH private key file from the *~/.ssh/* directory or select a private key file on your computer.
- Temp KeyPair: A temporary key pair is generated and sent to the instance. The key pair is valid within 1 minute. This method eliminates the need to manage passwords and key files.

5. Click **Connect**.

Connect to an instance by using session management

You can use the session management feature on Alibaba Cloud Client to connect to instances without the need to specify passwords or public IP addresses, or enable SSH or Remote Desktop Protocol (RDP) ports. Session management features high security and ease to use. For information about how session management works, see How session management works.

1. On the homepage of Alibaba Cloud Client, click **ECS**.
2. In the top navigation bar, select the region of the instance to which you want to connect.
3. Find the instance to which you want to connect and choose **Actions** > **Start Session Manager** in the **Actions** column.

   By default, the ecs-assist-user username is used to connect to the instance.

Connect to an instance over SSH by using session management

You can connect to instances over SSH by using session management on Alibaba Cloud Client. When you use session management to connect to an instance, the logon username is ecs-assist-user by default. When you use session management to connect to an instance over SSH, you can specify the username to log on to the instance, which is more flexible in some scenarios.

1. On the homepage of Alibaba Cloud Client, click **ECS**.
2. In the top navigation bar, select the region of the instance to which you want to connect.
3. Find the instance to which you want to connect and choose **Actions** > **Start SSH (via Session Manager)** in the **Actions** column.
4. Specify **Username**. Default value: root. You can use the default value. Specify **Port Num**. Default value: 22. You can use the default value. Set **Certifier**.

   Valid values of Certifier:

   - Password: Enter the password of the ECS instance.

     > **Note** In Alibaba Cloud Client, you can click the **Settings** icon in the upper-right corner. Then, on the Settings page, click **SSH** in the left-navigation pane. Set the Save Password to parameter to **Local File** to save the instance password to the /.aliyun/secrets.json file on your computer, or to **KMS** to save the instance password to KMS.

- Identity: Select an SSH private key file from the *~/.ssh/* directory or select a private key file on your computer.
- Temp KeyPair: A temporary key pair is generated and sent to the instance. The key pair is valid within 1 minute. This method eliminates the need to manage passwords and key files.

We recommend that you select Temp KeyPair to connect to instances without the need to specify passwords or public IP addresses. You can use RAM to control the connection permissions.

Connect to an instance by using the port forwarding feature provided by session management

You can use the port forwarding feature provided by session management to forward network traffic from a port on your computer to an instance without the need to specify the public IP address of the instance. This way, you can access the services that are running on the instance in a secure and convenient manner.

> **Note** Assume that the MySQL service is running on port 3306 on your instance. Your computer on which Alibaba Cloud Client is installed runs a Linux operating system. You can use the port forwarding feature to access the MySQL service on the instance from port 13306 on your computer.

The fort forwarding feature has the following benefits:

- The operations on Alibaba Cloud Client can be audited and are secure and controllable.
- RAM can be used to control connection permissions.
- The port forwarding feature can be used to access port 22 on instances to allow specified users to connect to the instances over SSH.
- The port forwarding feature can be used to access HTTP ports and web applications on instances without the need to specify the public IP addresses of the instances.

1. On the homepage of Alibaba Cloud Client, click **ECS**.
2. In the top navigation bar, select the region of the instance to which you want to connect.
3. Find the instance to which you want to connect and choose **Actions** > **Port Forward (via Session Manager)** in the **Actions** column.
4. Specify Remote Port. The port is used to access services on the instance. Specify Local Port. The port is listened on your computer. Turn on or off Print Request and Print Response. Enable or disable Open http://localhost:8080/ after started.
5. Click **Start**.

Manage ECS instances

Start an instance

1. On the homepage of Alibaba Cloud Client, click **ECS**.
2. In the top navigation bar, select the region of the instance to which you want to connect.
3. Find the instance that you want to start and choose **Actions** > **Start Instance** in the **Actions** column.
4. In the message that appears, check instance information and click **Start Instance**.

Stop an instance

1. On the homepage of Alibaba Cloud Client, click **ECS**.

2. In the top navigation bar, select the region of the instance to which you want to connect.
3. Find the instance that you want to stop and choose **Actions** > **Stop Instance** in the **Actions** column.
4. In the message that appears, check instance information and click **Stop Instance**.

Restart an instance

1. On the homepage of Alibaba Cloud Client, click **ECS**.
2. In the top navigation bar, select the region of the instance to which you want to connect.
3. Find the instance that you want to restart and choose **Actions** > **Reboot Instance** in the **Actions** column.
4. In the message that appears, check instance information and click **Reboot Instance**.

Release an instance

1. On the homepage of Alibaba Cloud Client, click **ECS**.
2. In the top navigation bar, select the region of the instance to which you want to connect.
3. Find the instance that you want to release and choose **Actions** > **Delete Instance** in the **Actions** column.
4. In the message that appears, check instance information and click **Delete Instance**.

Enable release protection for an instance

You can enable the release protection feature for your pay-as-you-go instances that run critical workloads. This feature prevents your pay-as-you-go instances instance from being manually released due to accidental operations. For more information about instance release protection, see Enable or disable release protection for ECS instances.

> **Note** This feature is available for only pay-as-you-go instances.

1. On the homepage of Alibaba Cloud Client, click **ECS**.
2. In the top navigation bar, select the region of the instance to which you want to connect.
3. Find the instance for which you want to enable release protection and choose **Actions** > **Set Deletion Protection** in the **Actions** column.
4. In the message that appears, check instance information and click **Set Deletion Protection**.

View instance attributes

1. On the homepage of Alibaba Cloud Client, click **ECS**.
2. In the top navigation bar, select the region of the instance to which you want to connect.
3. Find the instance whose attributes you want to view and choose **Actions** > **Show Properties** in the **Actions** column.
   You can view the attributes of the instance, including the instance name, hostname, instance type, operating system, and Cloud Assistant state.

7.4.4. Use Alibaba Cloud Client to manage elastic container instances

Alibaba Cloud Client allows you to query, view, and connect to Elastic Compute Service (ECS) instances, elastic container instances, simple application servers, and Alibaba Cloud managed instances. This topic describes how to use Alibaba Cloud Client to manage elastic container instances. You can view container groups, connect to containers, and manage container groups.

Prerequisites

An Alibaba Cloud account is added to Alibaba Cloud Client. For more information, see Add one or more accounts to Alibaba Cloud Client.

Context

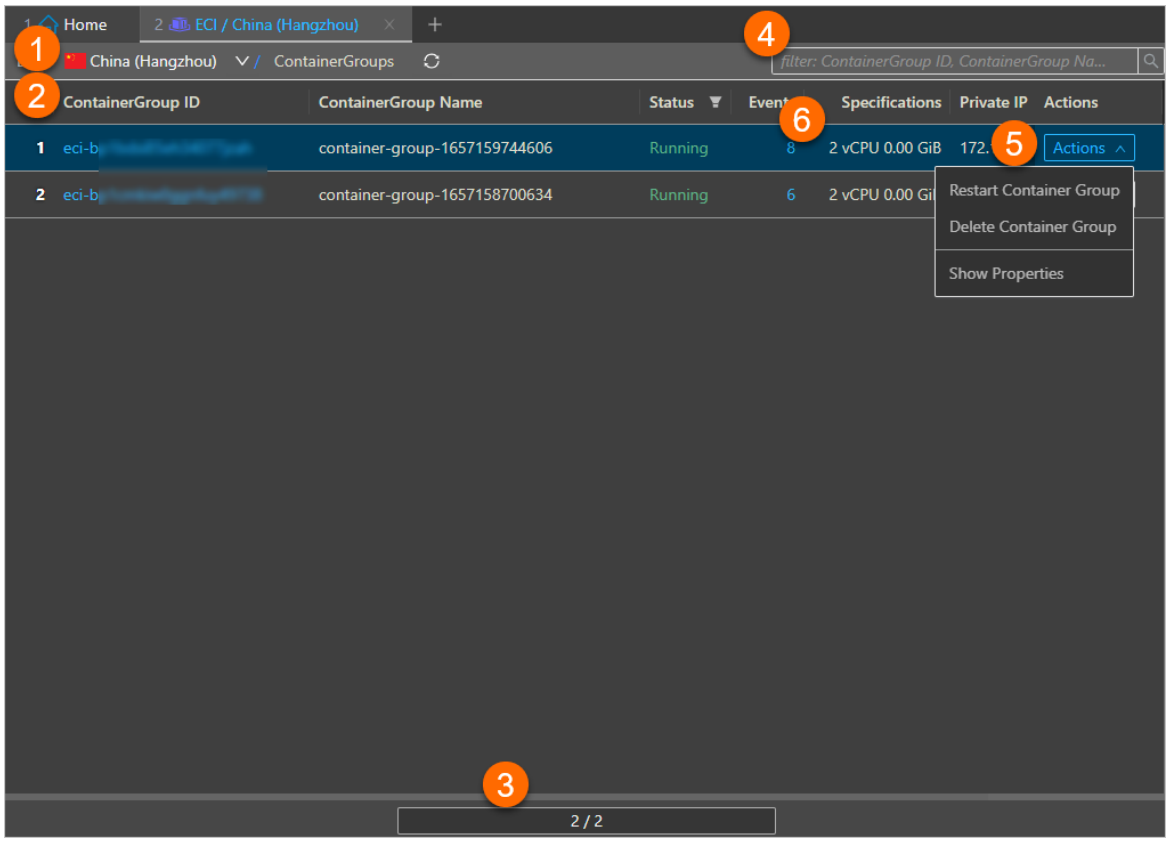
The terms of container groups and containers are described in the documents of Elastic Container Instance:

- Container group: a set of containers that can be scheduled to the same host. The lifecycle of a container group is determined based on all containers in the group. These containers share the network and storage resources of the container group. A container group is an elastic container instance and is similar to a pod in Kubernetes.
- Container: a lightweight, executable, and standalone software package. A container is the running entity of an image.

For more information, see What is Elastic Container Instance?

View container groups

1. On the homepage of Alibaba Cloud Client, click **ECI**.
2. On the instance list page, you can view the following information of elastic container instances:



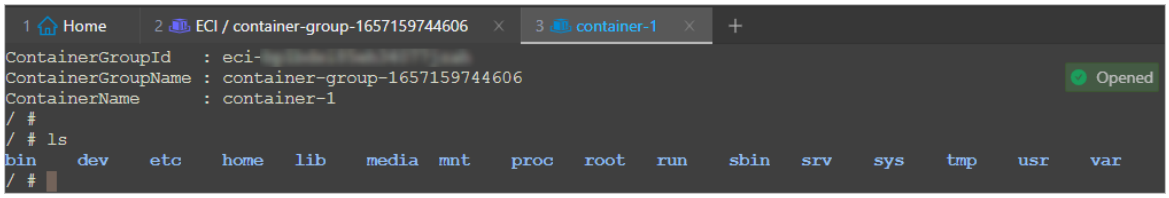
- ①: The region name of the current elastic container instance. You can click the icon to the

right of the region name to switch to another region.

- ②: List of elastic container instances.
- ③: The pagination. You can click the pagination to paginate the instances.
- ④:The search box.
  a. Enter a container group ID or container group name in the search box.
  b. Press the **Enter** key to search the instance in the current region.
- ⑤: The menu of the Actions column. You can perform the following operations:
  - Restart or delete a container group.
  - View the details of a container group.
  - Connect to an instance. You need to click the container group ID and then perform operations on the Containers page.
- ⑥: The events of the instances.

Connect to a container

1. On the homepage of Alibaba Cloud Client, click **ECI**.
2. In the upper-left corner of the page, select the region where the desired instance resides from the drop-down list.
3. Find the container group and click the container group ID.
4. Find the container that you want to connect to and select **Terminal** in the **Actions** column.
5. Select the Shell command type and click **Connect**.
   If the following information is displayed, the connection is successful.



View container properties

1. On the homepage of Alibaba Cloud Client, click **ECI**.
2. In the upper-left corner of the page, select the region where the desired instance resides from the drop-down list.
3. Find the container group and click the container group ID.
4. Find the container of which properties you want to view. Select the ··· icon in the **Actions** column, and then click **View Properties**.
   You can view the basic information of the container, such as the name, image, and command type.

Manage a container group

Restart a container group

1. On the homepage of Alibaba Cloud Client, click **ECI**.
2. In the upper-left corner of the page, select the region where the desired instance resides from the drop-down list.

3. Find the container group that you want to restart and then select **Restart** in the **Actions** column.
4. In the dialog box that appears, check the information of the instance and then click **Restart**.

Delete a container group

1. On the homepage of Alibaba Cloud Client, click **ECI**.
2. In the upper-left corner of the page, select the region where the desired instance resides from the drop-down list.
3. Find the container group that you want to delete and then select **Delete** in the **Actions** column.
4. In the dialog box that appears, check the information of the instance and then click **Delete**.

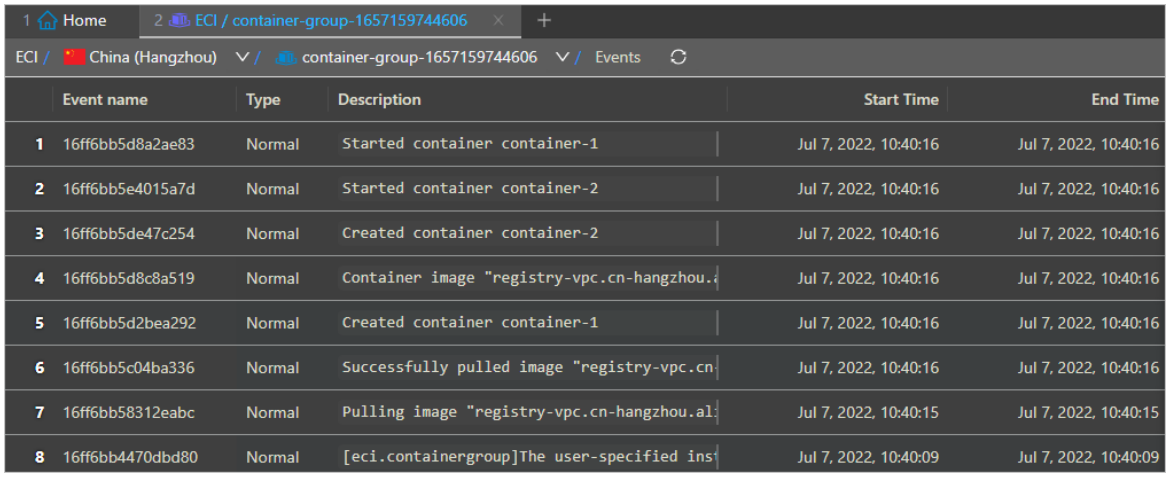
View container group properties

1. On the homepage of Alibaba Cloud Client, click **ECI**.
2. In the upper-left corner of the page, select the region where the desired instance resides from the drop-down list.
3. Find the container group of which properties you want to view. Select **View Properties** in the **Actions** column.
   You can view the basic information of the container group, such as the name and specification of the container group, the security group that the container group belongs, and the vSwitch to which to connect the container group.

View the events of an elastic container instance

1. On the homepage of Alibaba Cloud Client, click **ECI**.
2. In the upper-left corner of the page, select the region where the desired instance resides from the drop-down list.
3. In the Container Group ID column, find the container group of which events you want to view. Click the quantity of the events in the **Events** column.
4. On the Events page, view the details of the events.

   The following figure shows the most recent eight entries of events of the current elastic container instance. For more information about common events and their solutions, see Custom events of Elastic Container Instance.



	Event name	Type	Description	Start Time	End Time
1	16ff6bb5d8a2ae83	Normal	Started container container-1	Jul 7, 2022, 10:40:16	Jul 7, 2022, 10:40:16
2	16ff6bb5e4015a7d	Normal	Started container container-2	Jul 7, 2022, 10:40:16	Jul 7, 2022, 10:40:16
3	16ff6bb5de47c254	Normal	Created container container-2	Jul 7, 2022, 10:40:16	Jul 7, 2022, 10:40:16
4	16ff6bb5d8c8a519	Normal	Container image "registry-vpc.cn-hangzhou.a	Jul 7, 2022, 10:40:16	Jul 7, 2022, 10:40:16
5	16ff6bb5d2bea292	Normal	Created container container-1	Jul 7, 2022, 10:40:16	Jul 7, 2022, 10:40:16
6	16ff6bb5c04ba336	Normal	Successfully pulled image "registry-vpc.cn	Jul 7, 2022, 10:40:16	Jul 7, 2022, 10:40:16
7	16ff6bb58312eabc	Normal	Pulling image "registry-vpc.cn-hangzhou.al	Jul 7, 2022, 10:40:15	Jul 7, 2022, 10:40:15
8	16ff6bb4470dbd80	Normal	[eci.containergroup]The user-specified ins	Jul 7, 2022, 10:40:09	Jul 7, 2022, 10:40:09

7.5. Connect to an instance by using VNC

7.5.1. 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.

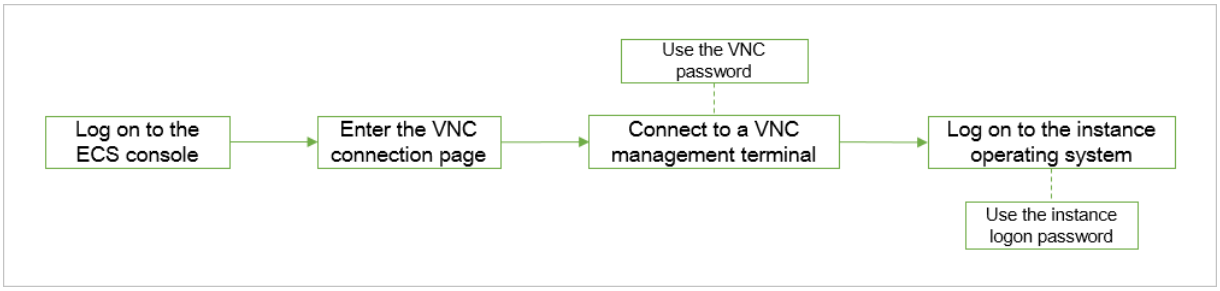
By default, a VNC connection session lasts for about 300 seconds. If you do not perform operations within these 300 seconds, the connection to the instance is automatically closed. You must connect to the instance again.

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**.

7.5.2. 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.

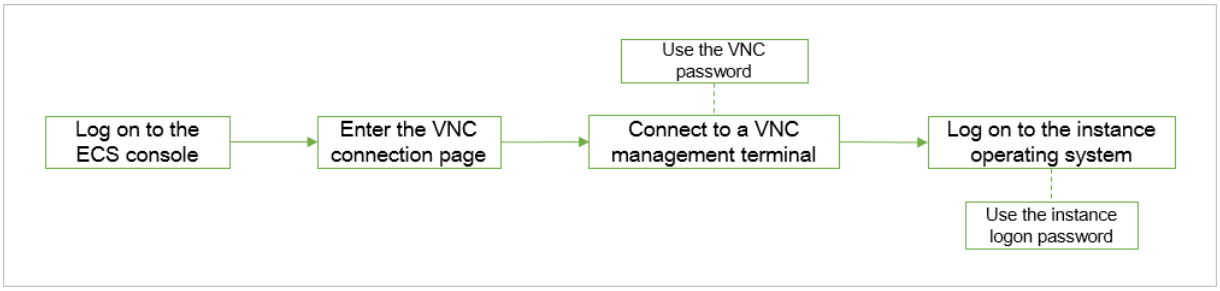
By default, a VNC connection session lasts for about 300 seconds. If you do not perform operations within these 300 seconds, the connection to the instance is automatically closed. You must connect to the instance again.

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.



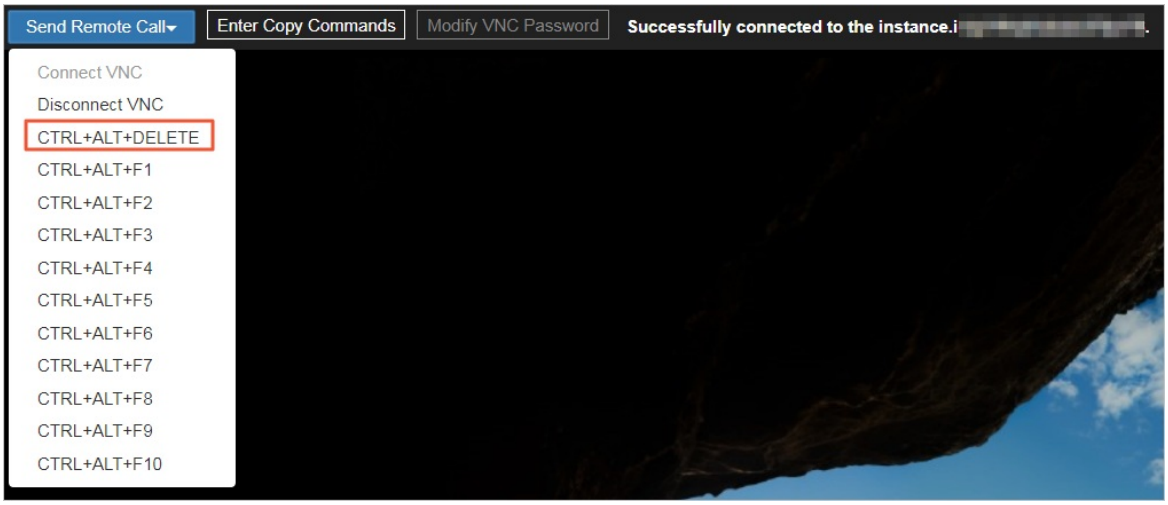
1.
2.
3.
4. On the **Instances** page, find the instance to be connected and click **Connect** in the **Actions** column.
5. Connect to a VNC management terminal.

   > **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?.

7.6. Connect to an instance by using third-party client tools

7.6.1. Connect to a Linux instance by using an SSH key pair

Secure Shell (SSH) key pairs are a secure and convenient method to authenticate logons. This topic describes how to use an SSH key pair to connect to a Linux instance from a Windows device or a device that supports SSH commands, such as a Linux client or MobaXterm for Windows.

Prerequisites

- An SSH key pair is created and the *.pem* private key file is downloaded. For more information, see Create an SSH key pair.
- The Linux instance to which you want to connect is in the Running state.
- An SSH key pair is bound to the instance.
- A public IP address or an elastic IP address (EIP) is associated with the instance.
- A security group rule is added to the security group of the instance to allow traffic on the corresponding port, such as the default port 22 for SSH. For more information, see Add a security group rule.

Network type	NIC type	Direction	Action	Protocol type	Port range	Priority	Authorization type	Authorization object
Virtual Private Cloud (VPC)	N/A	Inbound	Allow	SSH(22)	22/22	1	IPv4 CIDR blocks	0.0.0.0/0
Classic network	Public							

Context

You can use one of the following methods to connect to the Linux instance by using an SSH key pair based on the operating system of your device:

- Use an SSH key pair to connect to a Linux instance from a Windows device
- Use an SSH key pair to connect to a Linux instance from a device that supports SSH commands (configure information by using commands)
- Use an SSH key pair to connect to a Linux instance from a device that supports SSH commands (configure information by using the config file)

Use an SSH key pair to connect to a Linux instance from a Windows device

The following section describes how to convert the format of a private key file from *.pem* to *.ppk* and how to use an SSH key pair to connect to a Linux instance. PuTTYgen is used in this example.

1. Download and install PuTTYgen and PuTTY.

   Download PuTTYgen and PuTTY from the following links:

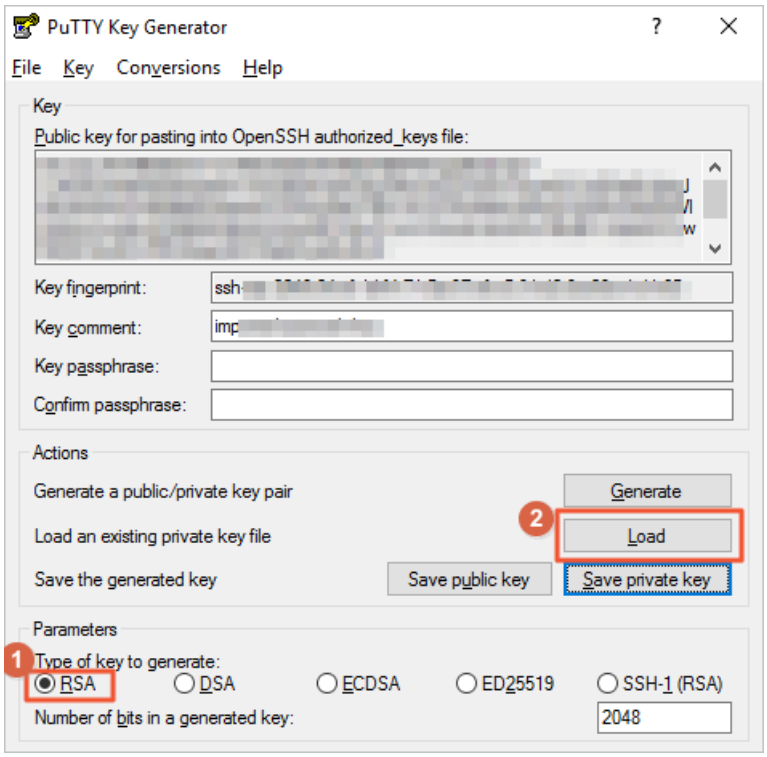
   - PuTTYgen
   - PuTTY

2. Convert the format of a private key file from *.pem* to *.ppk*.

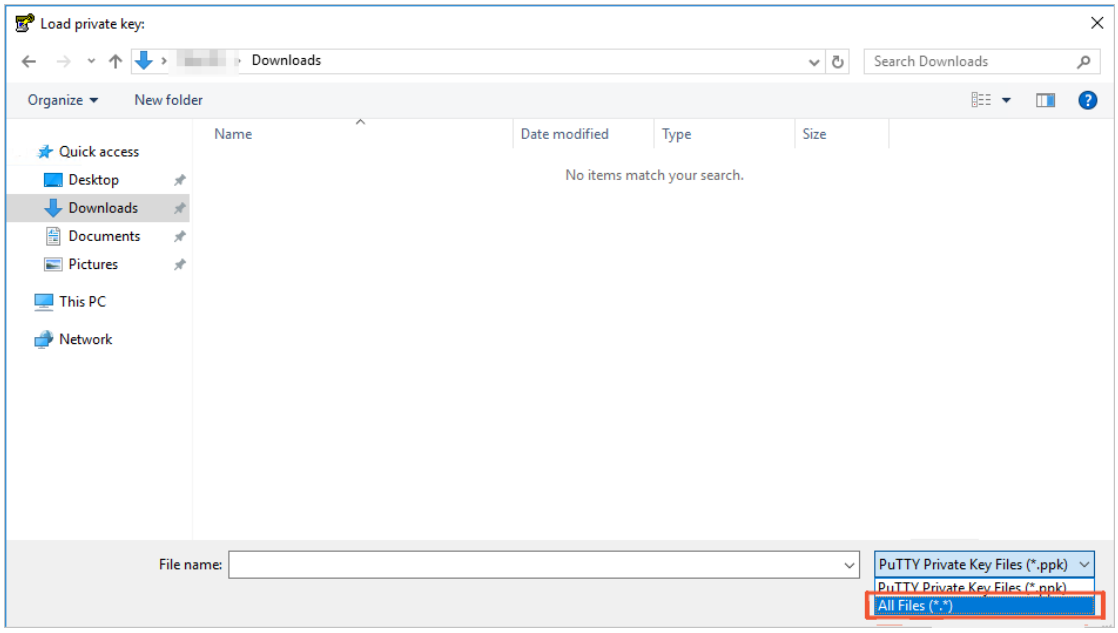
   i. Start PuTTYgen.

   In this example, PuTTYgen 0.71 is used.

   ii. Set **Type of key to generate** to **RSA** and click **Load**.



iii. Select **All Files**.



iv. Select the *.pem* private key file that you want to convert.

v. In the dialog box that appears, click **OK**.

vi. Click **Save private key**.

vii. In the dialog box that appears, click **Yes**.

viii. Specify a name for the *.ppk* private key file and click **Save**.

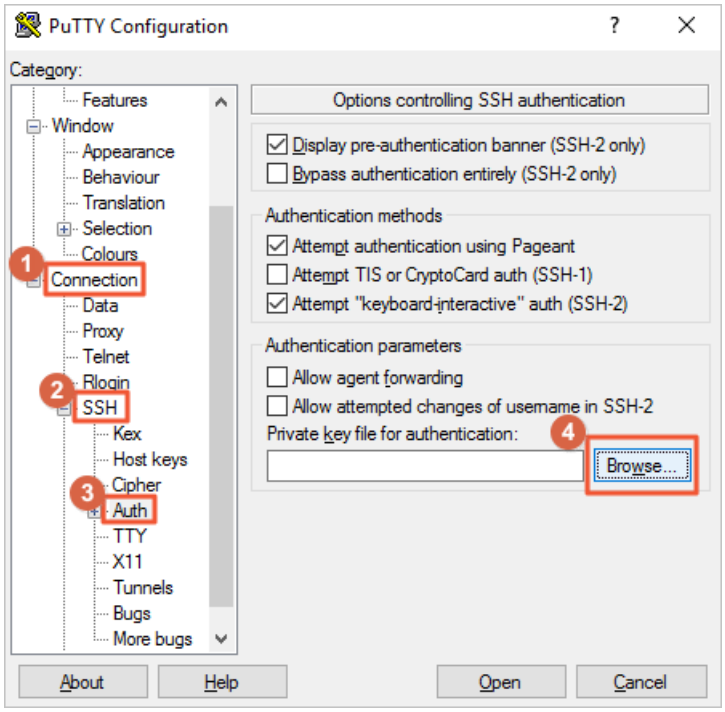
3. Start PuTTY.

4. Configure the private key file used for authentication.

   i. Choose **Connection > SSH > Auth**.

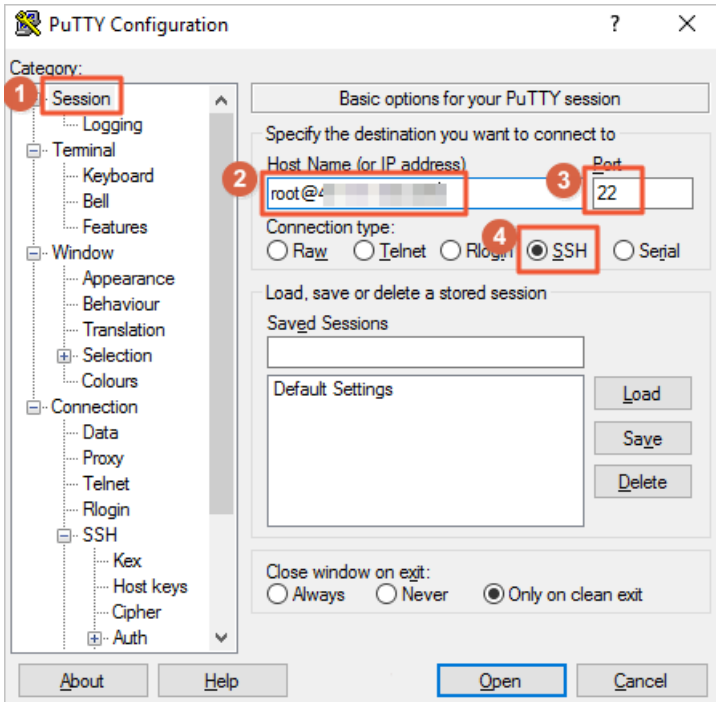
   ii. Click **Browse...**

   iii. Select the resulting *.ppk* private key file.



5. Configure the required parameters to connect to the Linux instance.
   i. Click **Session**.
   ii. In the **Host Name (or IP address)** field, enter the logon account and public IP address of the instance.

   The format is **root@<IP address>**. Example: root@10.10.xx.xxx.
   iii. In the **Port** field, enter 22.
   iv. Set **Connection type** to **SSH**.



6. Click **Open**.

If the following message appears, you are logged on to the instance.

Use an SSH key pair to connect to a Linux instance from a device that supports SSH commands (configure information by using commands)

The following section describes how to use commands to configure required information on a device that supports SSH commands (such as a Linux client or MobaXterm for Windows) and then how to use SSH commands to connect to the Linux instance from the device.

1. Find the path where the *.pem* private key file is located. Example: *~/.ssh/ecs.pem*.

   The path and file name used are for reference only. You can modify the information in subsequent commands based on actual conditions.

2. Run the following command to modify the attribute of the private key file:

   ```
   chmod 400 [Path of the .pem private key file on your PC]
   ```

   Example:

   ```
   chmod 400 ~/.ssh/ecs.pem
   ```

3. Run the following command to connect to the instance:

   ```
   ssh -i [Path of the .pem private key file on your PC] root@[Public IP address]
   ```

   Example:

   ```
   ssh -i ~/.ssh/ecs.pem root@10.10.xx.xxx
   ```

Use an SSH key pair to connect to a Linux instance from a device that supports SSH commands (configure information by using the config file)

The following section describes how to use commands to configure required information on a device that supports SSH commands (such as a Linux client or MobaXterm for Windows) and then how to use SSH commands to connect to the Linux instance from the device.

1. Find the path where the *.pem* private key file is located. Example: *~/.ssh/ecs.pem*.

   The path and file name used are for reference only. You can modify the information in subsequent commands based on actual conditions.

2. Run the following command to modify the attribute of the private key file:

   ```
   chmod 400 [Path of the .pem private key file on your PC]
   ```

Example:

```
chmod 400 ~/.ssh/ecs.pem
```

3. Run the following commands to go to the *.ssh* directory in the home directory and create a *config* file:

```
cd ~/.ssh
vim config
```

4. In the `config` file, press the /key to enter the edit mode and add the following configuration items:

```
# Enter the alias of the ECS instance to connect to the instance by using an SSH key pa
ir.
Host ecs
# Enter the public IP address of the instance.
HostName 121.196.**.**
# Enter the port number. The default port number is 22.
Port 22
# Enter the logon account.
User root
# Enter the address of the .pem private key file on your PC.
IdentityFile ~/.ssh/ecs.pem
```

If you have multiple ECS instances, you can use the `config` file to configure password-free logon in a centralized manner. The following example demonstrates how to configure password-free logon for two ECS instances:

```
# Enter the alias of one ECS instance to connect to the instance by using an SSH key pa
ir.
Host ecs1
# Enter the public IP address of the instance.
HostName 121.196.**.**
# Enter the port number. The default port number is 22.
Port 22
# Enter the logon account.
User root
# Enter the address of the .pem private key file on your PC.
IdentityFile ~/.ssh/ecs.pem
# Enter the alias of the other ECS instance to connect to the instance by using an SSH
key pair.
Host ecs2
# Enter the public IP address of the instance.
HostName 121.196.**.**
# Enter the port number. The default port number is 22.
Port 22
# Enter the logon account.
User root
# Enter the address of the .pem private key file on your PC.
IdentityFile ~/.ssh/ecs.pem
```

After the configuration items are added, press the *Esc* key and enter `:wq` to save the *config* file.

5. Run the following command to restart the SSH service:

```
service sshd restart
```

6. Run the following command to connect to the instance:

```
ssh [Alias of the instance]
```

Example:

```
ssh ecs
```

7.6.2. Connect to a Linux instance by using a password

This topic describes how to use a username and password to connect to an Elastic Compute Service (ECS) Linux instance from a Windows, Linux, Mac OS X, Android, or iOS device.

Prerequisites

- A Linux instance is created.
- A logon password is set for the instance.
- A public IP address or an elastic IP address (EIP) is associated with the instance.
- The Linux instance to which you want to connect is in the Running state.
- Rules are added to the security group to which the instance belongs to allow access to the corresponding ports. For more information, see Add a security group rule.

Network type	Network interface controller (NIC) type	Rule direction	Action	Protocol type	Port range	Priority	Authorization type	Authorization object
Virtual Private Cloud (VPC)	N/A	Inbound	Allow	SSH (22)	22/22	1	IPv4 CIDR blocks	0.0.0.0/0
Classic network	Public							

Context

You can use one of the following methods based on the operating system of your device to connect to a Linux instance with a username and password:

- Use a username and password to connect to a Linux instance from a Windows device
- Use a username and password to connect to a Linux instance from a Linux or Mac OS X device

Use a username and password to connect to a Linux instance from a Windows device

The following section describes how to use a username and password to connect to a Linux instance from a Windows device. In this example, PuTTY is used.

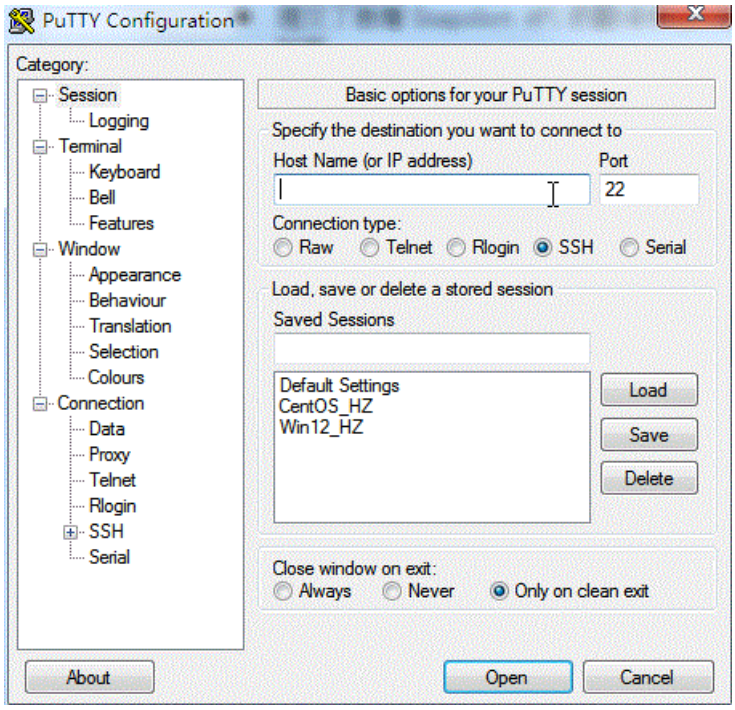
1. Download and install PuTTY.

   Download link: PuTTY.

2. Start PuTTY.

3. Configure required parameters to connect to the Linux instance.

   - **Host Name (or IP address)**: Specify the static public IP address of the instance or the EIP associated with the instance.
   - **Port**: Enter **22**.
   - **Connection type**: Select **SSH**.
   - **Saved Sessions**: optional. Enter a name that helps you identify the session and click **Save** to save the session. This way, you do not need to enter session information such as the public IP address when you connect to the instance again.

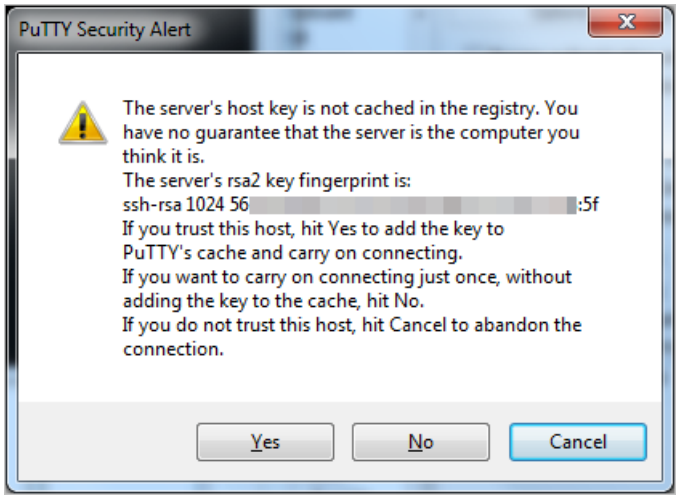


> **Note** To prevent the PuTTY connection from closing due to timeout, we recommend that you click **Connection** in the Category section and set **Seconds between keepalives (0 to turn off)** to 60, which indicates that the server sends a message to the client every 60 seconds (equivalent to 1 minute) to maintain the connection. The default value of Seconds between keepalives (0 to turn off) is 0.

4. Click **Open**.

   The first time you connect to the instance, the **PuTTY Security Alert** message appears. This message indicates that PuTTY cannot verify the authenticity of the remote server (instance) and can provide only the public key fingerprint of the server instead. Click **Yes** to indicate that you trust this server. PuTTY then adds the public key fingerprint to the registry of your device.

> **Note** If the **PuTTY Security Alert** message appears the next time you connect to the instance, the instance may have suffered from man-in-the-middle attacks. For more information, see PuTTY User Manual.



5. Enter the username and press Enter. The default username is root.
6. Enter the logon password of the instance and press the Enter key.

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

   If the `Welcome to Alibaba Cloud Elastic Compute Service !` message appears, you are connected to the instance.

Use a username and password to connect to a Linux instance from a Linux or Mac OS X device

1. Run the following SSH command:

   ```
   ssh root@<Public IP address or EIP of the instance>
   ```

   Example:

   ```
   ssh root@47.99.XX.XX
   ```

2. Enter the logon password of the instance.
   If the `Welcome to Alibaba Cloud Elastic Compute Service !` message appears, you are connected to the instance.

7.6.3. Connect to a Windows instance by using a username and password

This topic describes how to connect to a Windows instance from a local client.

Prerequisites

Before you connect to a Windows instance, make sure that the following requirements are met:

- The instance is in the **Running** state. If not, you must start the instance. For more information, see

Start an instance.

- A logon password is set for the instance. If you have not set a password or if you have forgotten the password, you must reset the password for the instance. For more information, see Reset the logon password of an instance.
- The instance can access the Internet:
  - In a virtual private cloud (VPC), you can assign a public IP address to an instance when you create the instance and purchase bandwidth for the instance. You can also associate an elastic IP address (EIP) with an instance after you create the instance. For more information, see Create an IPv4 VPC.
  - In the classic network, a public IP address is assigned to the instance by using one of the following methods:
    - If you select Assign Public IPv4 Address when you create a subscription or pay-as-you-go instance, a public IP address is assigned to the instance.
    - If you do not select Assign Public IPv4 Address when you create a subscription instance, you can upgrade the bandwidth to obtain a public IP address for the instance. For more information, see Overview of instance upgrade and downgrade.
- The following security group rules are added to the security group to which the instance belongs. For more information, see Add a security group rule.

Network Type	ENI type	Rule direction	Authorization policy	Protocol type	Port	Authorization type	Authorization object	Priority
VPC	N/A	Inbound	Allow	RDP(3389)	3389/3389	CIDR block	0.0.0.0/0	1
Classic network	Internet							

Procedures

You can connect to a Windows instance by using different remote connection software based on the operating system of your local client:

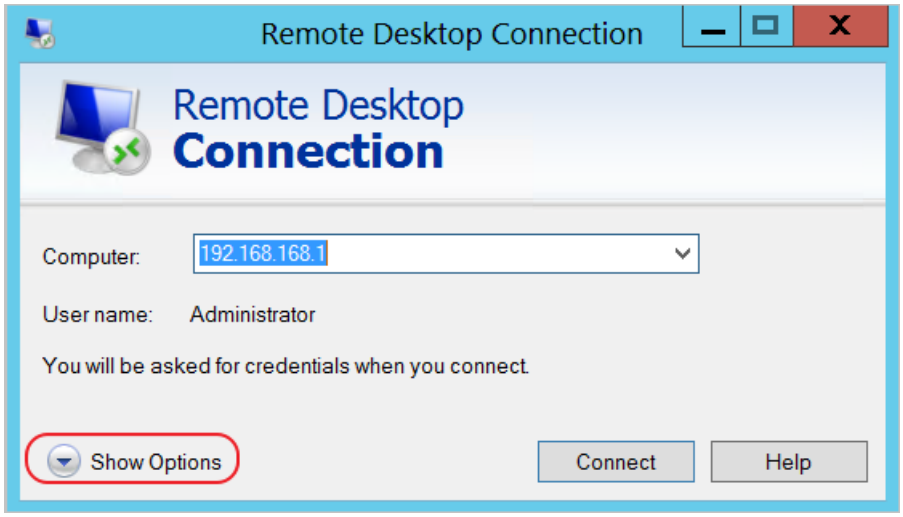
- Connect from a local client that runs a Windows operating system
- Connect from a local client that runs a Linux operating system
- Connect from a local client that runs a macOS operating system

Connect from a local client that runs a Windows operating system

If the local client runs a Windows operating system, you can use the Microsoft Terminal Services Client (MSTSC) that comes with the Windows operating system to connect to a Windows instance from the local client.

1. Use one of the following methods to start **Remote Desktop Connection** (MSTSC):
   - Choose **Start > Windows Accessories > Remote Desktop Connection**.
   - Click the **Start** icon, enter **mstsc** in the search box, and then press the Enter key.
   - Press **Win+R** to open the **Run** dialog box, enter **mstsc**, and then press the Enter key.
2. In the **Remote Desktop Connection** dialog box, perform the following operations in sequence:

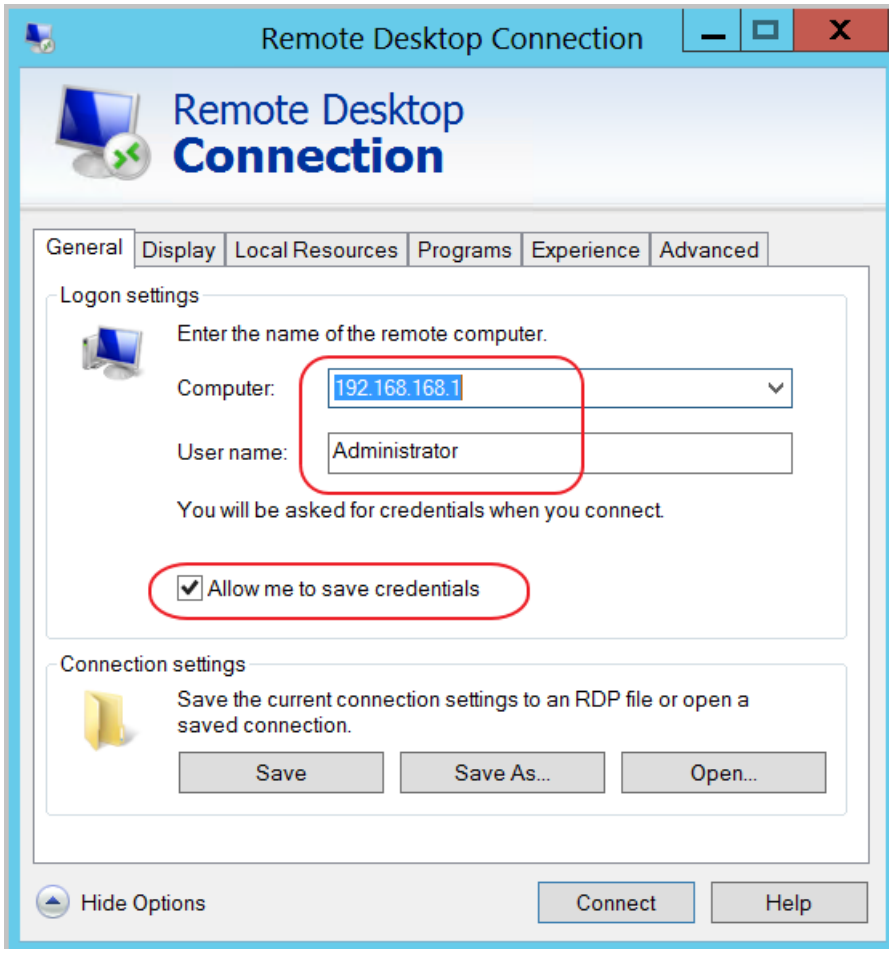
i. Click **Show Options**.



ii. Set Computer to the public IP address or EIP of the instance.

iii. Set Username. The default username is Administrator.

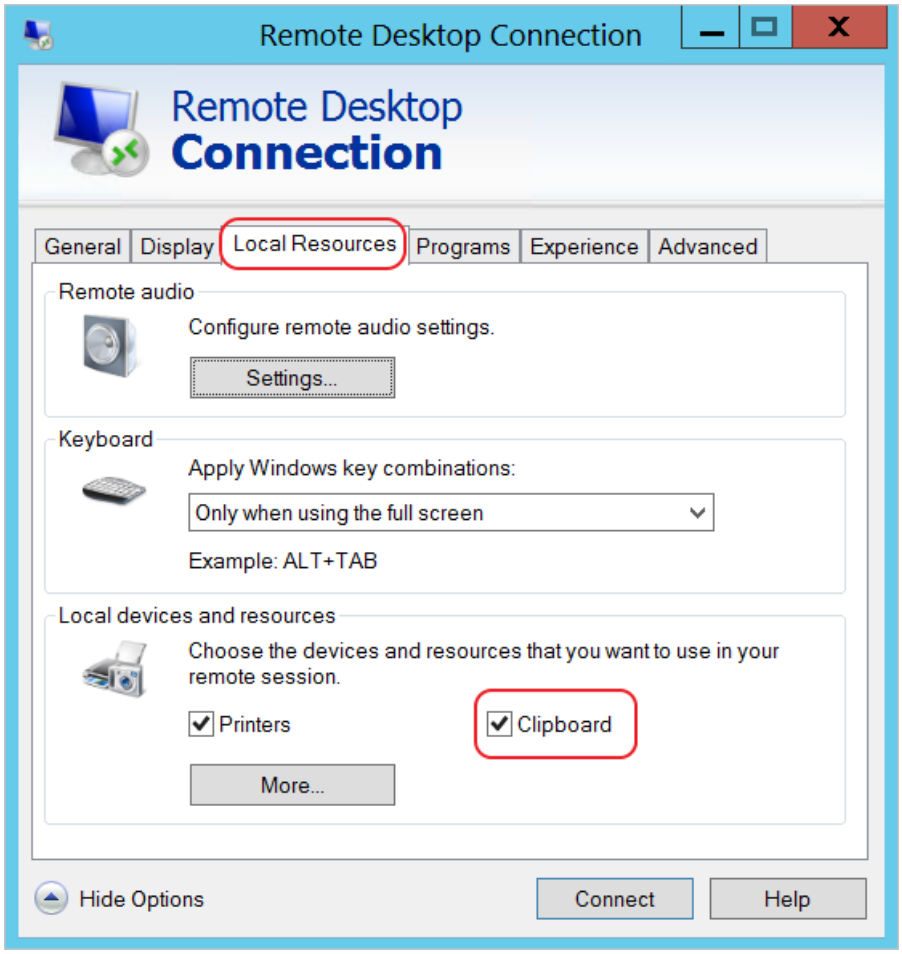
If you do not want to manually enter your username and password again the next time you connect to the instance, you can select **Allow me to save credentials**.

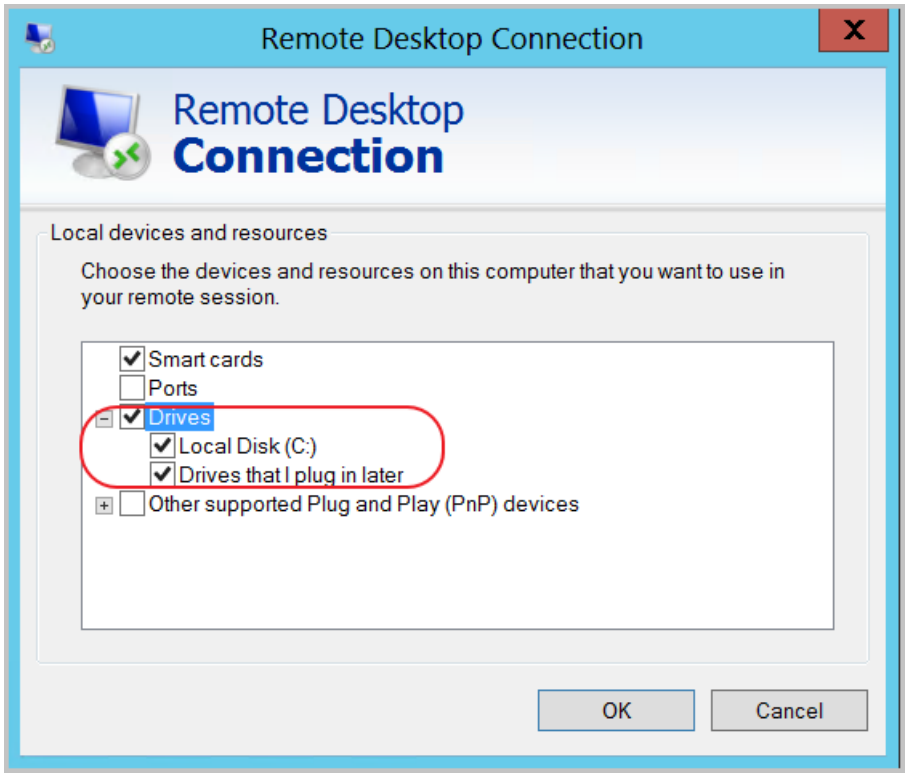


iv. (Optional)If you want to copy files from your local client to the instance, click the **Local Resources** tab to view the options for sharing local computer resources.

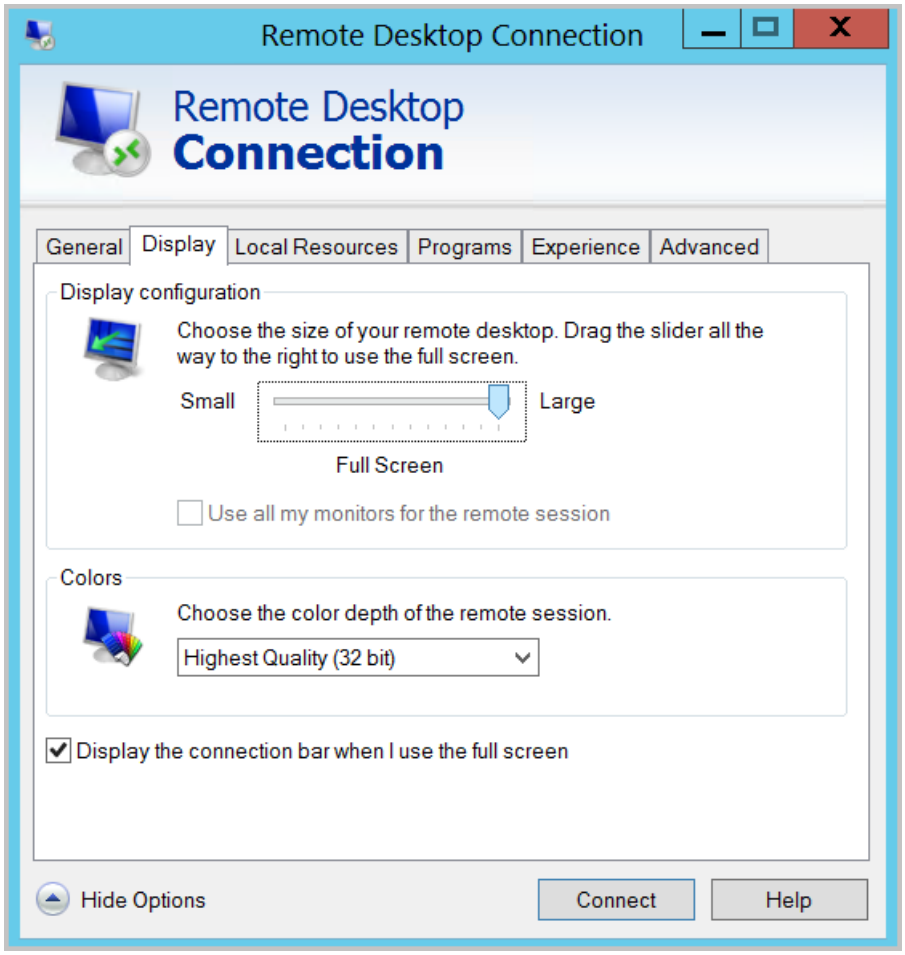
- If you want to copy only text, select **Clipboard**.

- If you want to copy files, click **More...**, select Drives, and then select the letters of the drives from which you want to copy files.





v. (Optional)If you have specific requirements on the size of the remote desktop window, click the **Display** tab to resize the remote desktop window. We recommend that you use Full Screen.



vi. Click **Connect**.

Connect from a local client that runs a Linux operating system

If the local client runs a Linux operating system, you can use a remote connection tool to connect to a Windows instance from the local client. In this example, rdesktop is used.

1. Download and start rdesktop.
2. Run the following command to connect to the Windows instance.

   This sample command is for your reference. Replace the parameters in the command based on your needs.

   ```
   rdesktop -u administrator -p password -f -g 1024*720 192.168.1.1 -r clipboard:PRIMARYCL
   IPBOARD -r disk:sunray=/home/yz16184
   ```

   The following table describes the parameters.

Parameter	Description
-u	The username to use to log on to the Windows instance. The default username is Administrator.

Parameter	Description
-p	The password to use to log on to the Windows instance.
-f	The full-screen mode. You can press **Ctrl+Alt+Enter** to switch the mode.
-g	The screen resolution. An asterisk (*) is used between the pixel width and height. This parameter can be left empty. If this parameter is not specified, the full-screen mode is used.
192.168.1.1	The IP address of the instance. Replace it with the public IP address or EIP of your Windows instance.
-d	The domain name. For example, if the domain name is INC, set the parameter to `-d inc`.
-r	Multimedia redirection. Examples: ○ Turn on the sound: `-r sound`. ○ Use a local sound card: `-r sound : local`. ○ Enable USB flash drive: `-r disk:usb=/mnt/usbdevice`.
-r clipboard:PRIMARYCLIPBOARD	Allows text including Chinese to be copied between the local client that runs a Linux operating system and the Windows instance.
-r disk:sunray=/home/yz16184	Maps a directory in the Linux operating system on the local client to a disk in the Windows instance. This way, Samba and FTP are not required for transferring files.

Connect from a local client that runs a macOS operating system

For information about how to connect to a Windows instance from a local client that runs a macOS operating system, visit Get started with the macOS client.

7.6.4. Connect to a Linux instance from a mobile device

This topic describes how to use a username and password to connect to a Linux instance from an iOS or Android mobile device.

Prerequisites

- A Linux instance is created.
- A logon password is set for the instance.
- A public IP address or an elastic IP address (EIP) is associated with the instance.
- The Linux instance to which you want to connect is in the Running state.
- Rules are added to the security group to which the instance belongs to allow access to the corresponding ports. For more information, see Add a security group rule.

Network type	Network interface controller (NIC) type	Rule direction	Action	Protocol type	Port range	Priority	Authorization type	Authorization object
Virtual Private Cloud (VPC)	N/A	Inbound	Allow	SSH (22)	22/22	1	IPv4 CIDR blocks	0.0.0.0/0
Classic network	Public							

Context

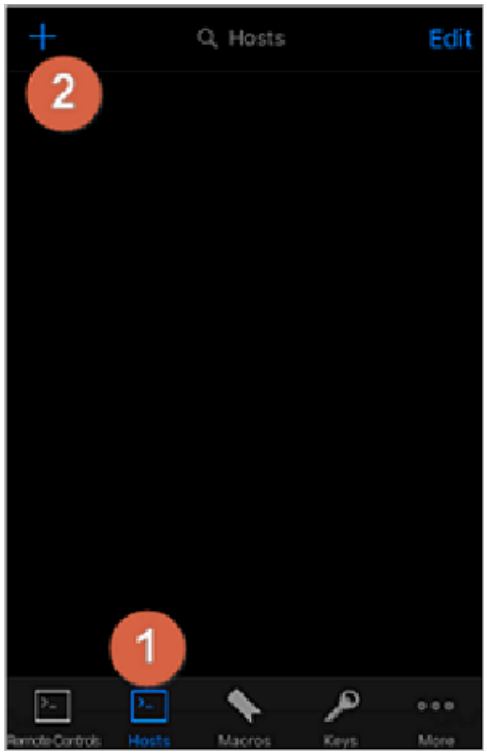
You can use one of the following methods to connect to a Linux instance based on the operating system of your mobile device:

- Use SSH Control Lite to connect to a Linux instance from an iOS device
- Use JuiceSSH to connect to a Linux instance from an Android device

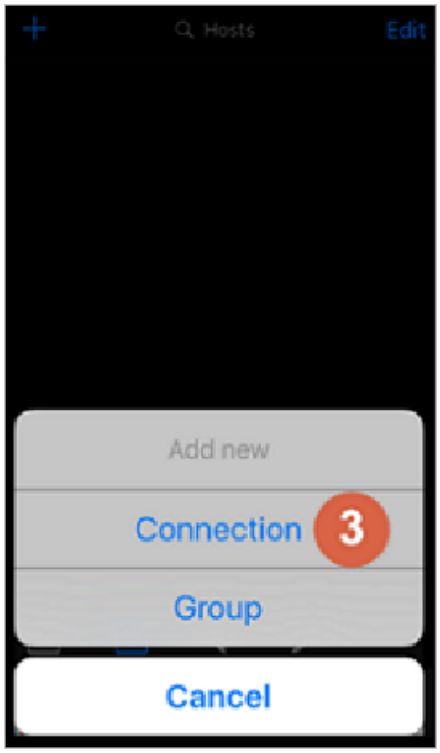
Use SSH Control Lite to connect to a Linux instance from an iOS device

In this example, a username and password is used for authentication.

1. Download SSH Control Lite.
2. Start SSH Control Lite.
3. In the lower part of the page, tap **Hosts**.
4. In the upper-left corner of the page, tap the **+** icon.



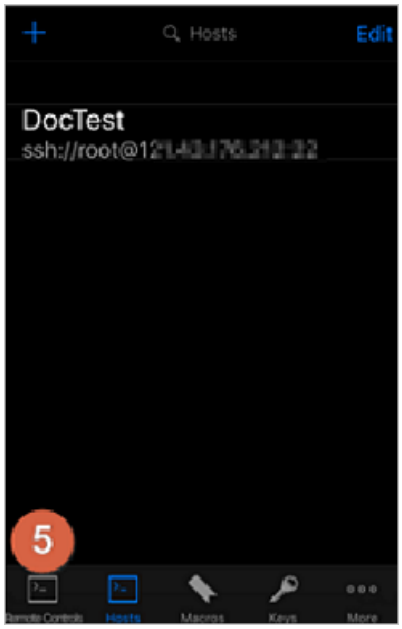
5. Tap **Connection**.



6. Configure the connection parameters and tap **Save**.
   - **Name**: Specify the host name. In this example, **DocTest** is used.
   - **Protocol**: Use the default value **SSH**.
   - **Host**: Specify the public IP address or EIP of the Linux instance to which you want to connect.
   - **Port**: Enter the port number **22**.

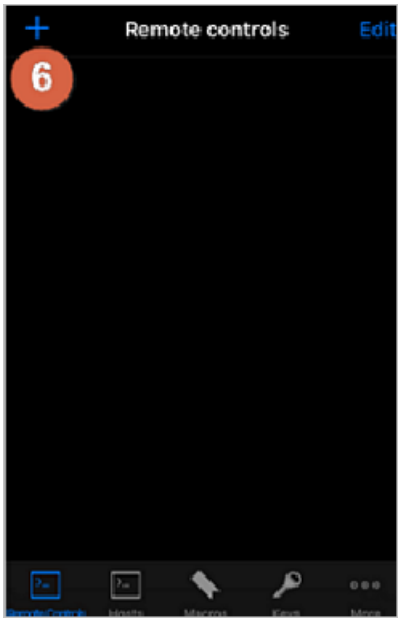
- **Username**: Enter the username **root**.
- **Password**: Enter the password to log on to the instance.

7. In the lower part of the page, tap **Remote Controls**.

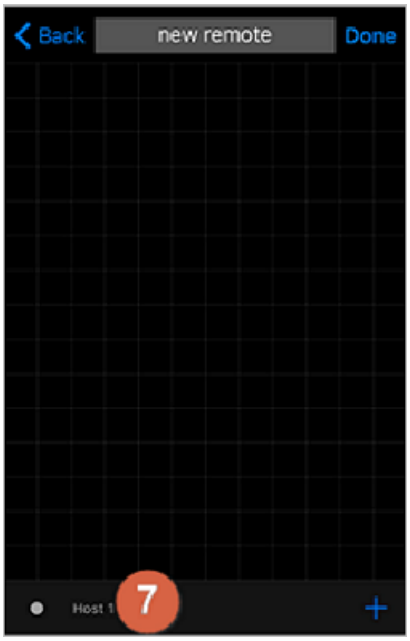


8. In the upper-left corner of the page, tap the **+** icon.

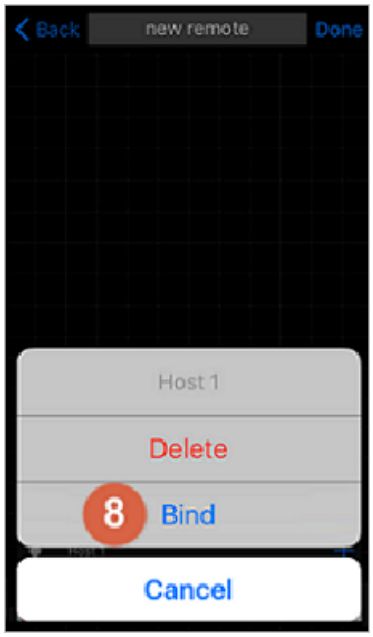
   Create a remote connection session. In this example, the session name is specified as **New remote**.



9. Tap **Host 1**.

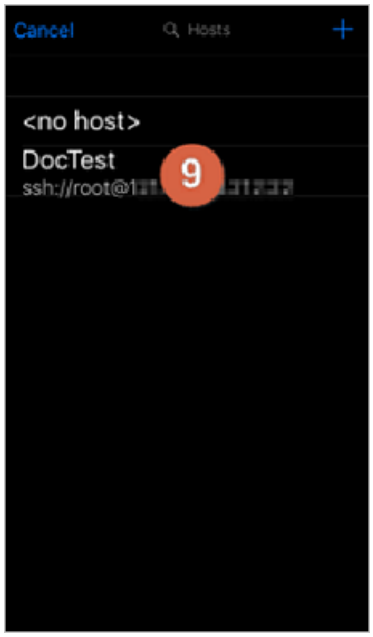


10. Tap **Bind**.

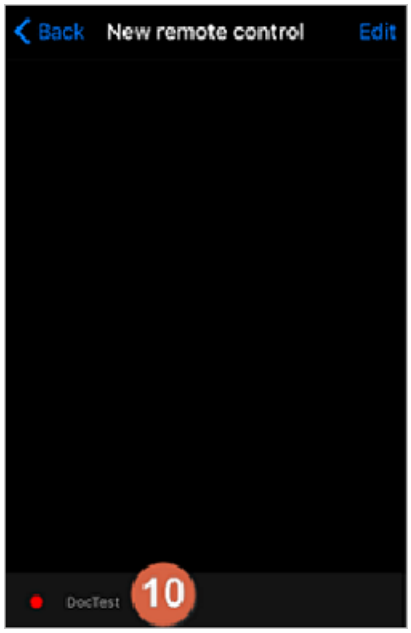


11. Select the newly added Linux instance.

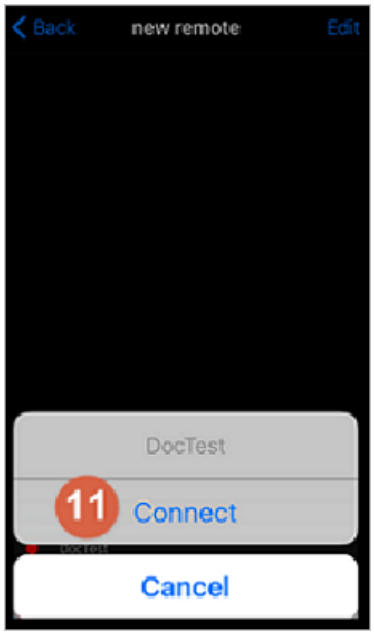
    In this example, **DocTest** is used.



12. In the upper-right corner of the page, tap **Done**. When **Edit** is displayed in the upper-right corner of the page, tap **DocTest**.

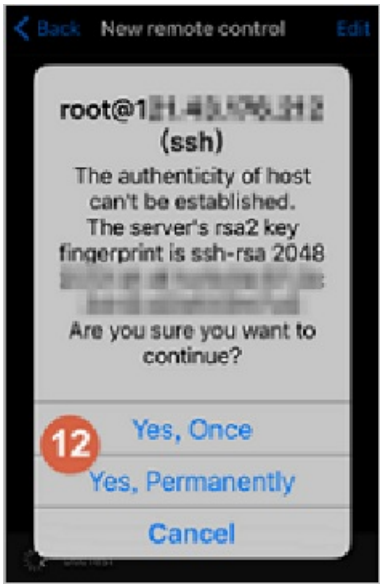


13. Tap **Connect**.

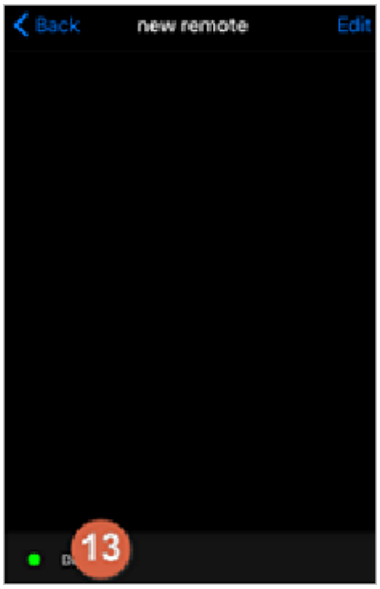


14. Select **Yes, Once** or **Yes, Permanently**.

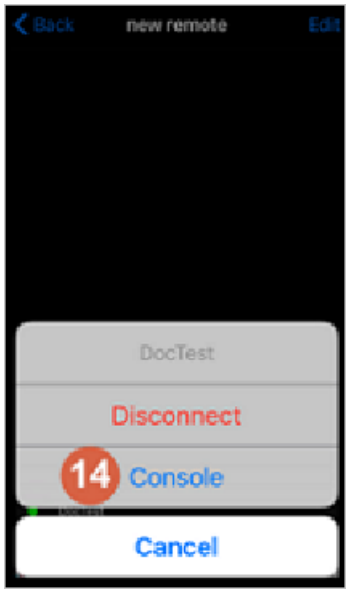
    If the connection is successful, the indicator icon next to **DocTest** becomes green.



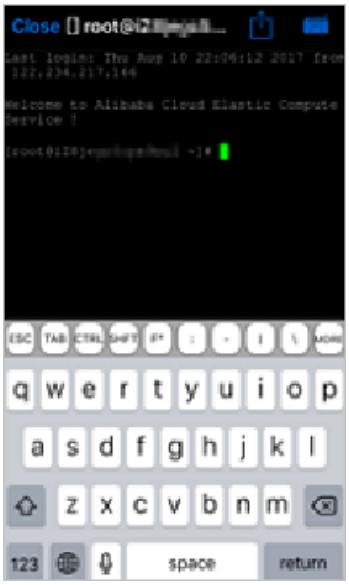
15. Tap **DocTest**.



16. Tap **Console** to go to the Linux instance management page.



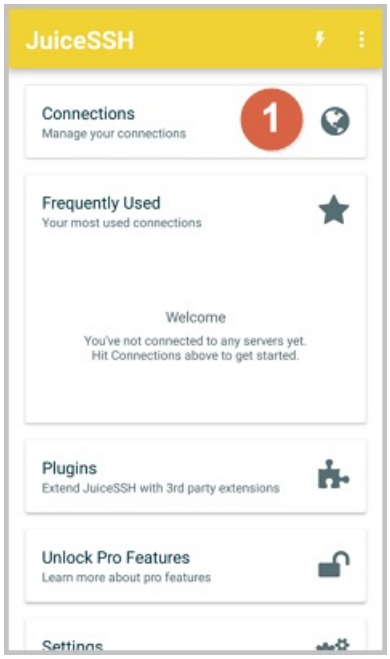
You have connected to the Linux instance.



Use JuiceSSH to connect to a Linux instance from an Android device

In this example, a username and password is used for authentication.

1. Install JuiceSSH.
2. Start JuiceSSH.
3. Tap **Connections**.



4. Tap the **+** icon.

5. Configure the connection parameters and tap the

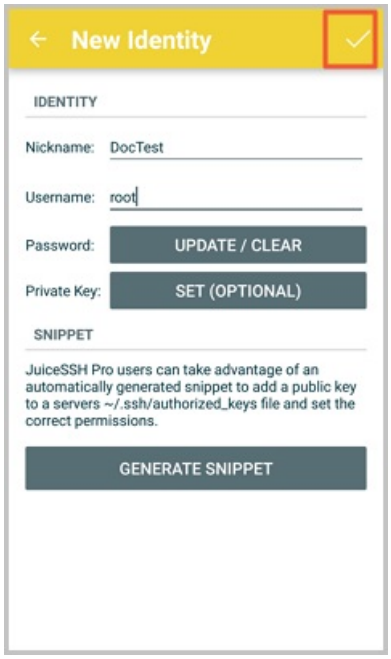
   icon.

   - **Nickname**: Specify the session name. In this example, **DocTest** is used.
   - **Type**: Use the default value **SSH**.
   - **Address**: Specify the public IP address or EIP of the Linux instance to which you want to connect.
   - Set **Identity**.
     a. Tap **Identity** and select **New** from the drop-down list.

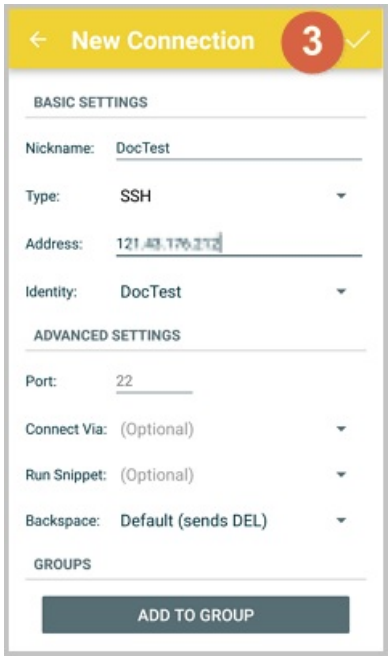
b. Specify the following parameters and tap the

icon.

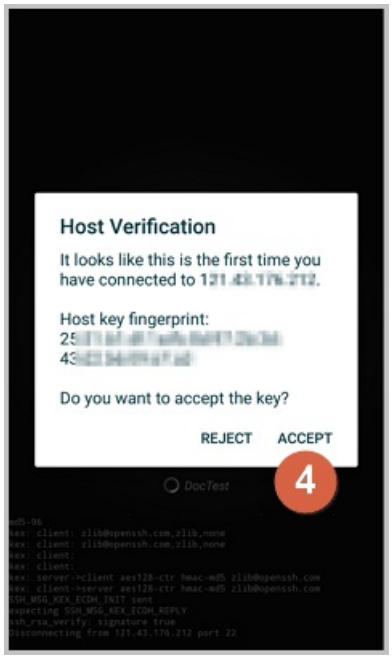
- **NickName**: optional. You can specify an identity name based on your management needs. In this example, **DocTest** is used.
- **Username**: Enter the username **root**.
- **Password**: Tap **SET (OPTIONAL)** and enter the password to log on to the instance.



- **Port**: Enter the port number **22**.

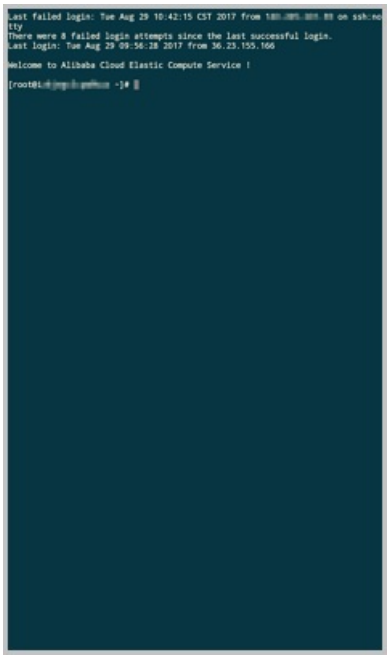


6. Read the prompt and tap **ACCEPT**.



7. (Optional)When you connect to the instance for the first time, a message appears to remind you to set information such as font. Read the information and tap **OK - I'VE GOT IT!**.

You have connected to the Linux instance.



Related information

- Connection methodsGuidelines on instance connection

7.6.5. Connect to a Windows instance from a mobile device

This topic describes how to connect to a Windows Elastic Compute Service (ECS) instance from an Android device. Microsoft Remote Desktop is used in this topic.

Prerequisites

Before you connect to a Windows instance, make sure that the following requirements are met:

- The instance is in the **Running** state.
- The instance has a public IP address and is accessible from the Internet.
- A logon password is set for the instance. If you forget the password, reset the password. For more information, see Reset the logon password of an instance.
- Microsoft Remote Desktop is installed on the instance. You can download Microsoft Remote Desktop from the official Microsoft website.
- The rule described in the following table is added to the security group of the instance based on the network type of the instance. For information about how to add a rule to a security group, see Add a security group rule.

Network type	NIC type	Direction	Action	Protocol type	Port range	Authorization type	Authorization object	Priority
Virtual Private Cloud (VPC)	N/A	Inbound	Allow	RDP (3389)	3389/3389	IPv4 CIDR block	0.0.0.0/0	1
Classic network	Public							

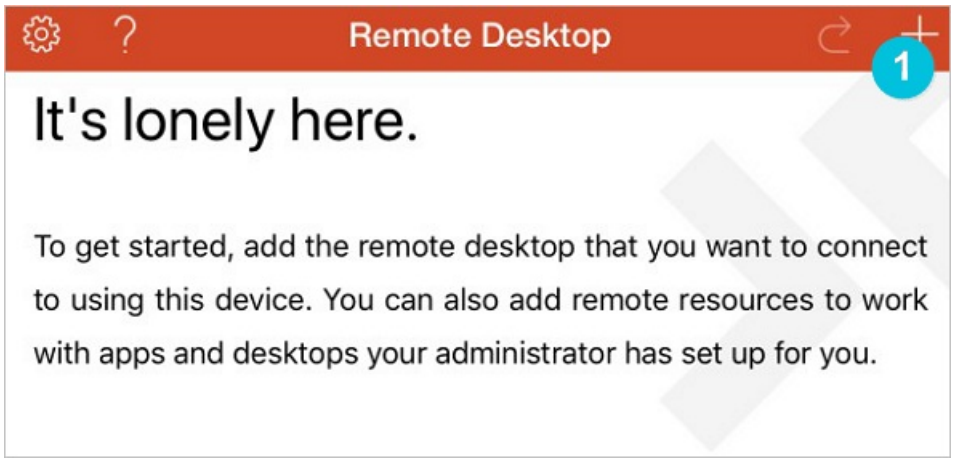
- If you log on to the Windows instance as a non-administrator user, the user must belong to the Remote Desktop Users group.

Procedure

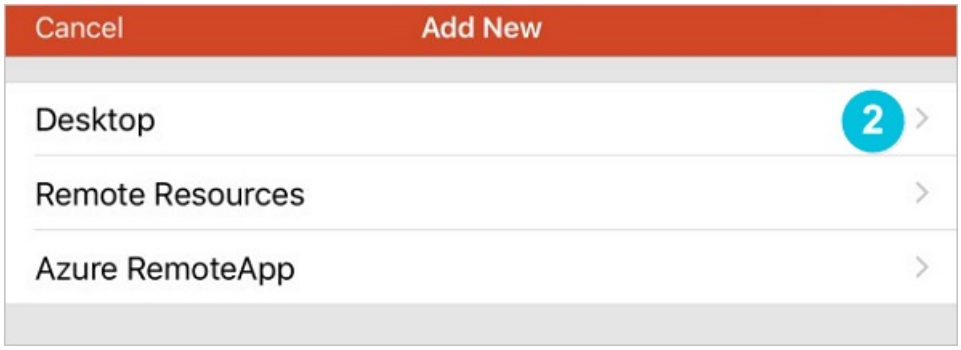
1. Open the Microsoft Remote Desktop app.

   In this example, RD Client 8.1.56.294 is used.

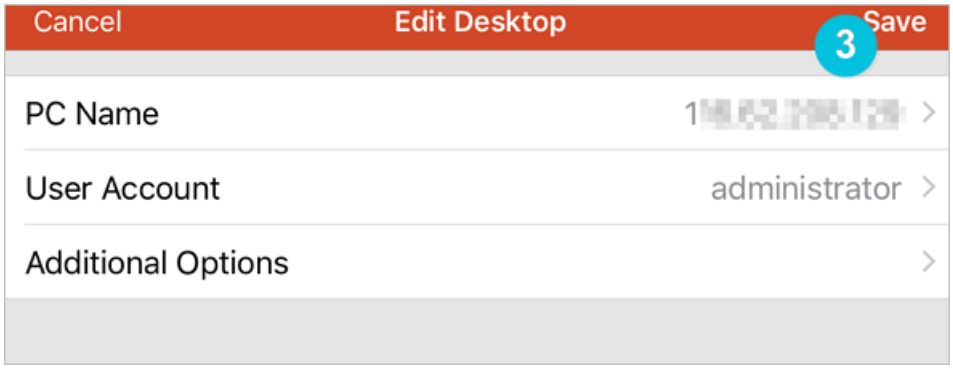
2. In the upper-right corner of the **Remote Desktop** page, tap the **+** icon.



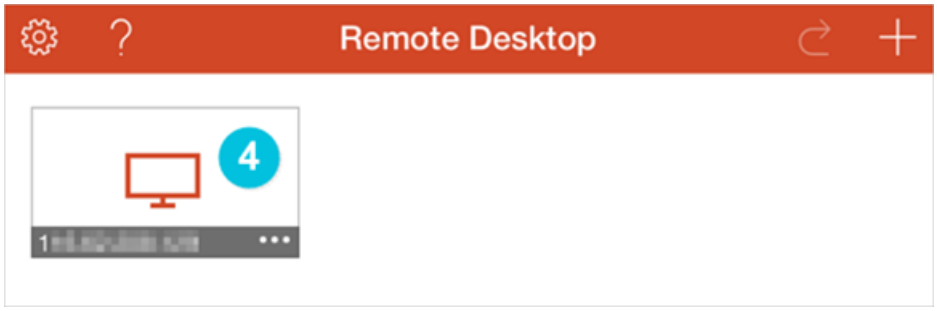
3. Tap **Desktop**.
4. On the **Add desktop** page, enter the hostname or public IP address of the Windows instance to which you want to connect in the **PC name** field and tap **SAVE**.



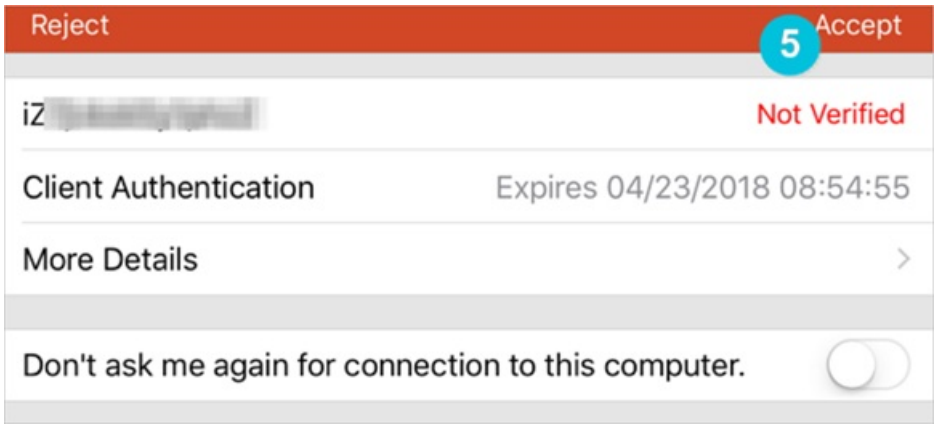
5. On the **Remote Desktop** page, tap the Windows instance.



6. If the **Certificate can't be verified. Do you want to connect anyway?** message appears, confirm that certificate information and connection information are correct and tap **CONNECT**.

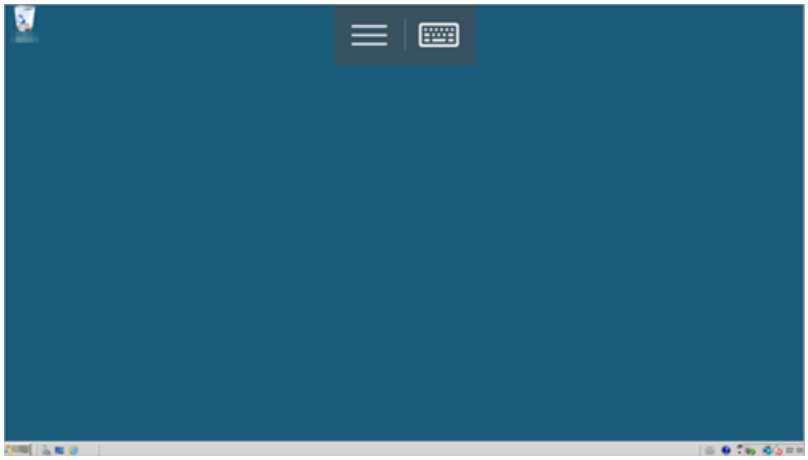


7. On the **Logon** page, enter your username such as **administrator** and your password and tap **CONNECT** to connect to the Windows instance.



Result

You are connected to the Windows instance.



Related information

- Connection methodsGuidelines on instance connection

8.Manage instance status

8.1. Start an instance

If an Elastic Compute Service (ECS) instance is in a state in which it cannot provide services, such as the Stopped state, you must start the instance first before you can use it. This topic describes how to start an instance in the ECS console.

Prerequisites

The instance that you want to start meets one of the following requirements:

- The instance is in the **Stopped** state.
- The instance is a pay-as-you-go instance that entered the **Expired** state due to an overdue payment. The overdue payment is settled but the instance cannot be automatically reactivated.

> Note

Context

If you have settled an overdue payment that caused a pay-as-you-go instance to stop but the instance cannot be automatically reactivated, the instance will still be released. You must manually reactivate the instance in a timely manner to prevent the instance from being released and affecting your business. For more information, see Pay-as-you-go.

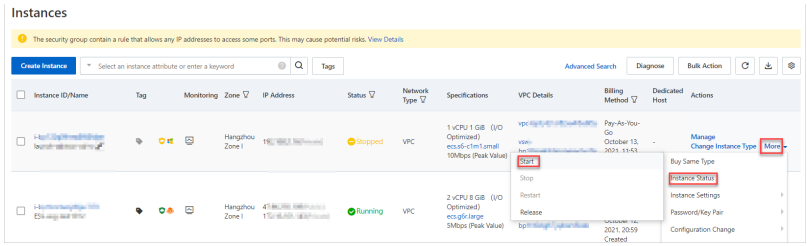
After the overdue payment for a pay-as-you-go instance is settled, the system reactivates the instance. If the instance cannot be automatically reactivated, wait 10 minutes and check whether the instance is reactivated and enters the **Running** state. If the instance still cannot be automatically reactivated, manually reactivate it in a timely manner.

> **Note** After you reactivate a pay-as-you-go instance that was stopped due to an overdue payment, the instance begins to run again and resumes billing on a pay-as-you-go basis. Release instances that are no longer needed to avoid unnecessary costs.

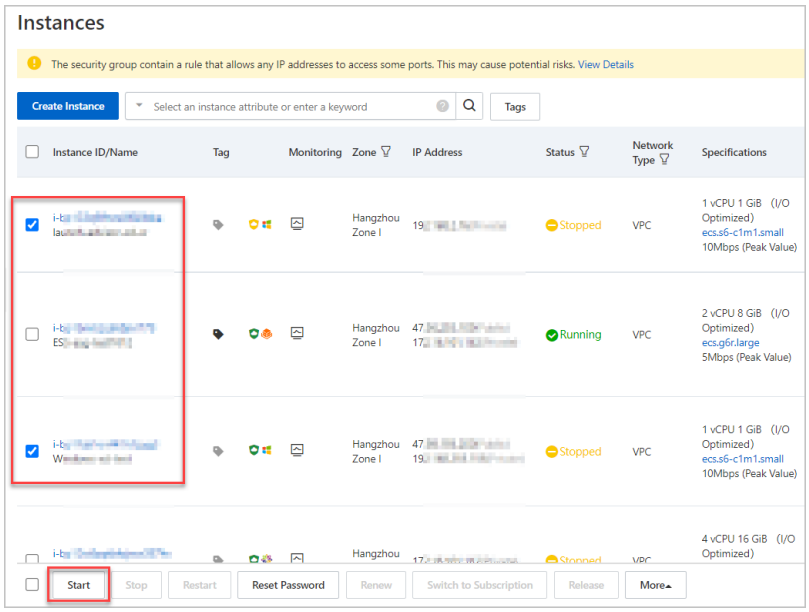
The instance may fail to be manually reactivated if resources for the instance type are insufficient. Change the instance type. For more information, see Change the instance type of a pay-as-you-go instance. If the problem persists, submit a ticket.

Procedure

1.
2.
3.
4. Start instances.
   - To start a single instance, find the instance and choose **More > Instance Status > Start** in the **Actions** column.



- To start multiple instances at a time, select the instances that you want to start and click **Start** in the lower part of the Instances page.



5. In the **Start Instance** dialog box, confirm the instance information and click **OK**.

Result

After the instance is started, it enters the **Running** state.

Related information

- StartInstance

8.2. Stop an instance

This topic describes how to stop an Elastic Compute Service (ECS) instance and how to enable economical mode for instances that are located in 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.

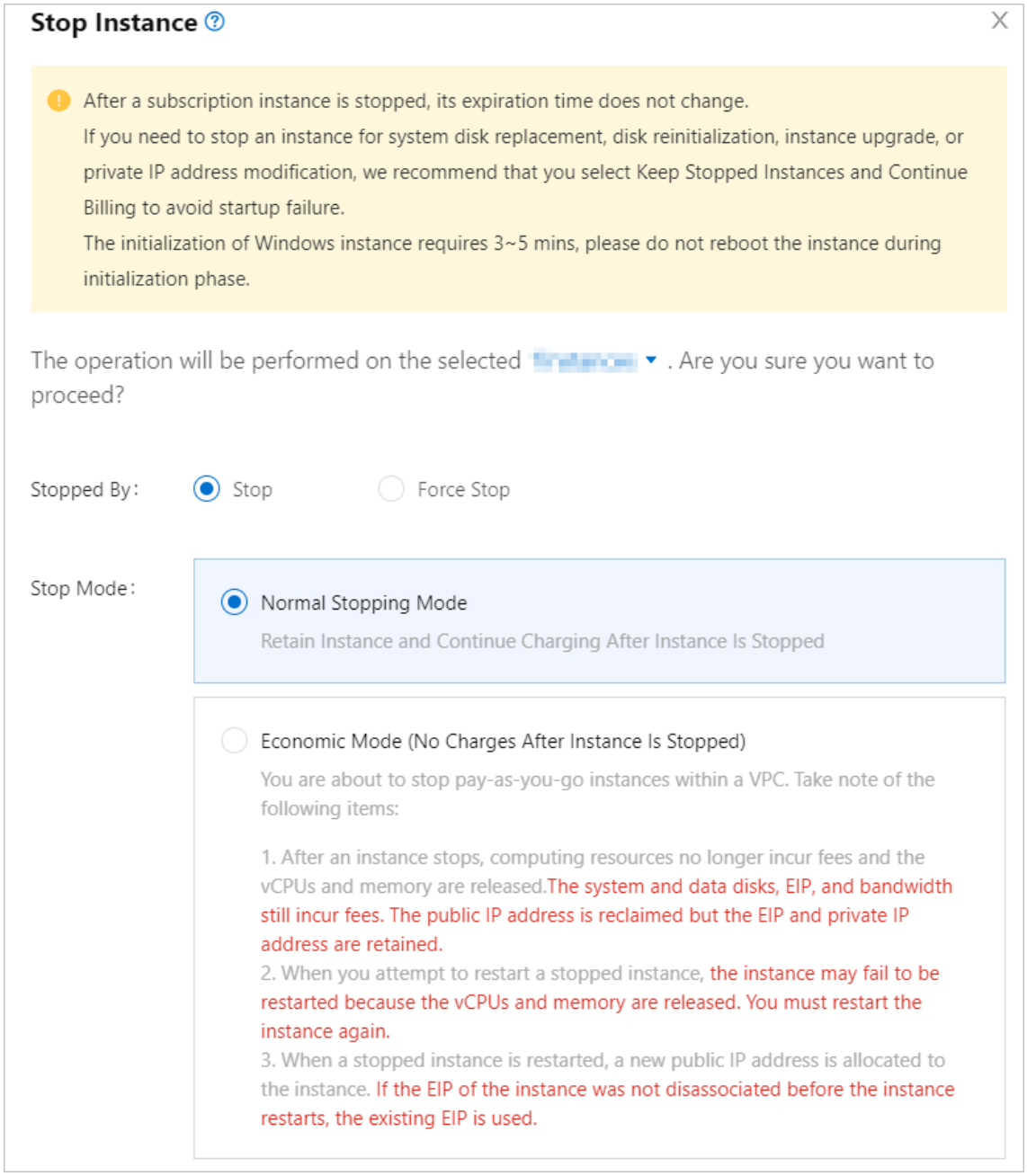
b. Click **OK**.

- 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

8.3. Hibernate an instance

If you do not require the use of an ECS instance for a period of time, but you still want to retain the instance without performing operations such as configuration upgrade or downgrade on the instance, we recommend that you hibernate the instance. A hibernated instance is different from a stopped instance. A hibernated instance automatically restores the applications on the instance to the status before hibernation when the instance is waked. This allows the instance to resume providing services in a short time.

Context

When you hibernate an instance, the operating system of the instance saves data from the memory to the system disk of the instance. The saved data includes the applications that run in the operating system and the usage status of the applications. When you wake the instance, the operating system reads the data saved in the system disk, automatically restores the applications to the status before hibernation, and resumes the running state of the instance. In comparison, when you stop and restart an instance, the operating system restarts the backend services and applications.

> **Note** If the instance fails to be hibernated, the instance is automatically shut down. Data in the memory is not saved to the system disk. When the instance is started again, the operating system of the instance restarts the backend services and applications. The operating system cannot restore the applications to the status before hibernation.

Hibernation has different impacts on the billing of instances that use different billing methods:

- Subscription instance: The expiration time and billing of the hibernated instance are not affected.
- Pay-as-you-go instance: Whether the billing of the hibernated instance is affected is based on whether you select the **No Fees for Hibernated Instances** option when you hibernate the instance. The following table describes the billing details of resources.

Billing of resources on a hibernated instance

Resource	No Fees for Hibernated Instances	Retain Instance and Continue Charging After Instance Is Hibernated
Computing resource (vCPUs and memory)	Release and stop billing	Retain and continue billing
Disk (system disk and data disk)	Retain and continue billing	Retain and continue billing
Internal IP address	Retain and stop billing	Retain and stop billing
Public IP address	Release and stop billing. After the instance is started, a new public IP address is obtained.	Retain and stop billing
EIP	Retain and continue billing	Retain and continue billing
Bandwidth	Continue billing	Continue billing

Limits

- The instance hibernation feature is now available only in the US (Silicon Valley) and Germany (Frankfurt) regions, and will be gradually supported in other regions.

- Before you can hibernate an instance, the instance must meet the following requirements:
  - The instance hibernation feature is enabled when the instance is created.

    > **Note** The instance hibernation feature cannot be disabled after it is enabled. If you do not enable the instance hibernation feature when you create an instance, you cannot hibernate the instance.

  - The hibernation agent is installed on the instance.
- You can enable the instance hibernation feature only when you create an ECS instance by using an encrypted custom image. The following image versions are supported:
  - Windows Server 2016 or later
  - Ubuntu 18 or later
  - CentOS 7 or later
- If the instance hibernation feature is enabled for an ECS instance when the instance is created, you cannot perform the following operations on the instance:
  - Create custom images.
  - Create snapshots.
  - Change the instance type.
  - Change the operating system or system disk.
  - Change the bandwidth of subscription instances.
- If the instance hibernation feature is enabled for a preemptible instance, you can select only the No Fees for Hibernated Instances option when you hibernate the instance.
- You cannot hibernate ECS instances in scaling groups.

Step 1: Enable the instance hibernation feature

You must enable the instance hibernation feature when you create an ECS instance. Otherwise, you cannot hibernate the instance. When you create the instance, you must use an encrypted image.

1. Obtain an encrypted custom image.

   You can use one of the following methods to obtain an encrypted custom image:

   - Prepare an encrypted custom image that meets the hibernation requirements.
   - Copy an image and encrypt it at the same time. For more information, see Copy a custom image.

   > **Note** For more information about the limits on images, see Limits.

2. Create an ECS instance and enable the instance hibernation feature when you create the instance.

   For more information, see Create an instance by using the wizard. Take note of the parameters described in the following table.

Parameter	Description	Example

Parameter	Description	Example
Instance	○ Instance Type: Select an I/O optimized instance type, except ECS Bare Metal Instance. ○ Memory: ■ Windows: Set the memory size to a value less than 16 GiB. ■ Linux: Set the memory size to a value less than 150 GiB.	ecs.g6e.large
Image	○ Select the encrypted custom image created in the previous step or an existing encrypted custom image that meets the hibernation requirements. ○ Select **Instance Hibernation** to enable the instance hibernation feature.	○ encrypted.windows2016 ○ Select **Instance Hibernation**.
Disk	○ System Disk: required. The system disk must meet the following requirements: ■ Category: ultra disk, standard SSD, or enhanced SSD (ESSD). ■ Capacity: The system disk capacity must be sufficient. We recommend that you set the system disk capacity to at least twice the memory size. This is because when the instance hibernation feature is enabled, the system disk reserves some space to store memory data. Therefore, the system disk capacity must be sufficient to ensure normal running of the operating system and applications when the system disk stores the memory data. ■ Encryption: By default, the system disk is encrypted if an encrypted image is used. ○ Data Disk: optional. To create data disks for an instance when you create the instance, you must select the disk categories and specify the sizes and quantity of the disks. You must also determine whether to encrypt the disks.	○ System Disk: Select Enhanced SSD (ESSD), set Disk Capacity to 60 GiB, select Disk Encryption, and then select **Default Service CMK** from the drop-down list. ○ Data Disk: Select Enhanced SSD (ESSD), set Disk Capacity to 40 GiB, and do not select Disk Encryption.

Parameter	Description	Example
Network	Select a virtual private cloud (VPC). **Note** ECS instances in the classic network do not support the instance hibernation feature.	[Default]vpc-bp1opxu1zkhn00g****

Step 2: Install the hibernation agent

After you enable the instance hibernation feature for the instance, you must install the hibernation agent on the instance before you can hibernate the instance.

1. Create and run one of the following commands to install the hibernation agent on the instance. For more information, see Use the immediate execution feature.
   - Windows instance:

     ```
     acs-plugin-manager.exe --exec --plugin ecs-hibernate-win --params "install"
     ```

   - Linux instance:

     ```
     acs-plugin-manager --exec --plugin ecs-hibernate-linux --params "install"
     ```

2. Restart the instance to make the hibernation agent take effect. For more information, see Restart an instance.

Step 3: Hibernate the instance

After the instance hibernation feature is enabled for the instance and the hibernation agent is installed on the instance, you can hibernate the instance in the Running state. You are unable to connect to the instance when the instance is hibernated.

1.
2.
3.
4. Find the instance that you want to hibernate and choose **More > Instance Status > Stop** in the **Actions** column.
5. In the **Stop Instance** dialog box, configure the parameters.
   i. Set **Stopped By** to **Hibernate**.

ii. Set **Stop Mode** to **Retain Instance and Continue Charging After Instance Is Hibernated** or **No Fees for Hibernated Instances** based on the billing method of the instance.

- You are charged for subscription instances even after the instances are hibernated. Select **Retain Instance and Continue Charging After Instance Is Hibernated**.
- For pay-as-you-go instances, you can select **Retain Instance and Continue Charging After Instance Is Hibernated** or **No Fees for Hibernated Instances**.

> **Note** For preemptible instances, select **No Fees for Hibernated Instances**.

For more information about the difference between **Retain Instance and Continue Charging After Instance Is Hibernated** and **No Fees for Hibernated Instances**, see Billing of resources on a hibernated instance in this topic.

iii. Click **OK**.

> **Note** The instance is stopped and enters the **Stopped** state. To start the instance, see Start an instance.

References

You can use Operation Orchestration Service (OOS) to hibernate and wake ECS instances at the scheduled time. This way, the hibernation and wake time of a large number of instances can be managed in an automated manner and the costs can be reduced by using the No Fees for Hibernated Instances feature. For more information, see Start and shut down ECS instances at the scheduled time.

8.4. Restart an instance

This topic describes how to restart an instance through the ECS console. You can also choose to call the RebootInstance API action.

Limits

- Only instances in the **Running** state can be restarted.
- When you restart an instance, the instance is stopped. As a result, services provided by the instance are disrupted.

Procedure

1. In the left-side navigation pane, click **Instances**.
2. Select the target region.
3. Find the target instance, and then choose **More > Instance Status > Restart** in the **Actions** column.

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

Related API: RebootInstance

8.5. Release an instance

Only pay-as-you-go Elastic Compute Service (ECS) instances (including preemptible instances) and expired subscription instances can be released. This topic describes how to release pay-as-you-go instances manually or automatically.

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.

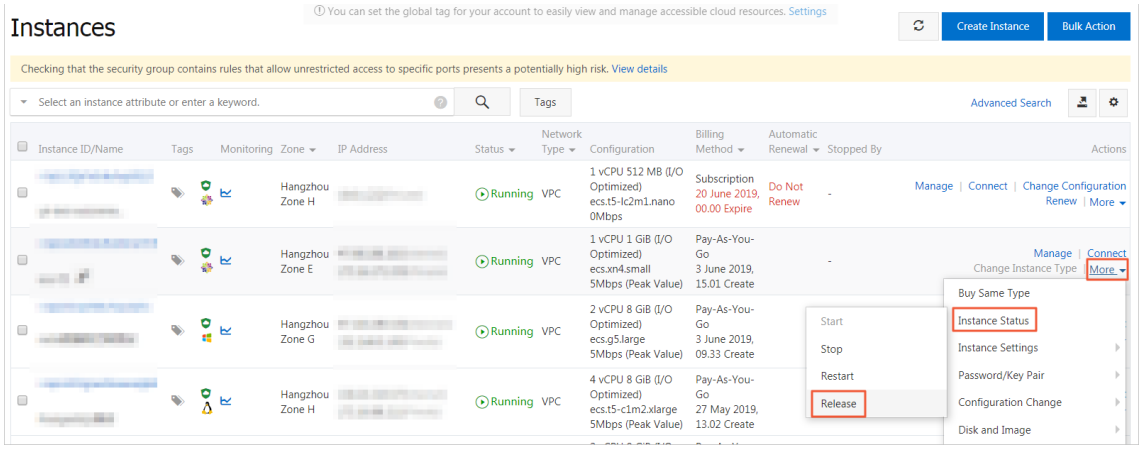
Context

- 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.

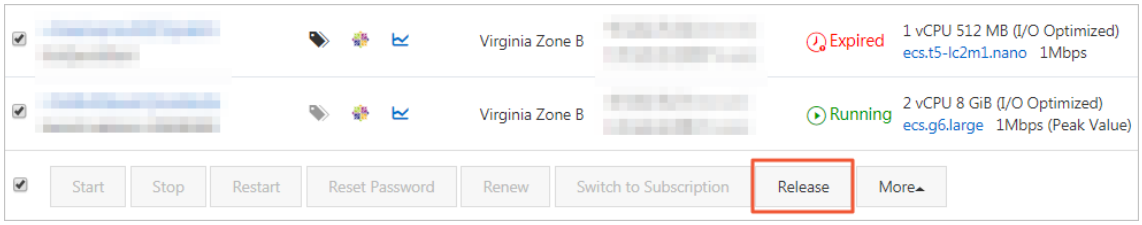
Manually release instances

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.



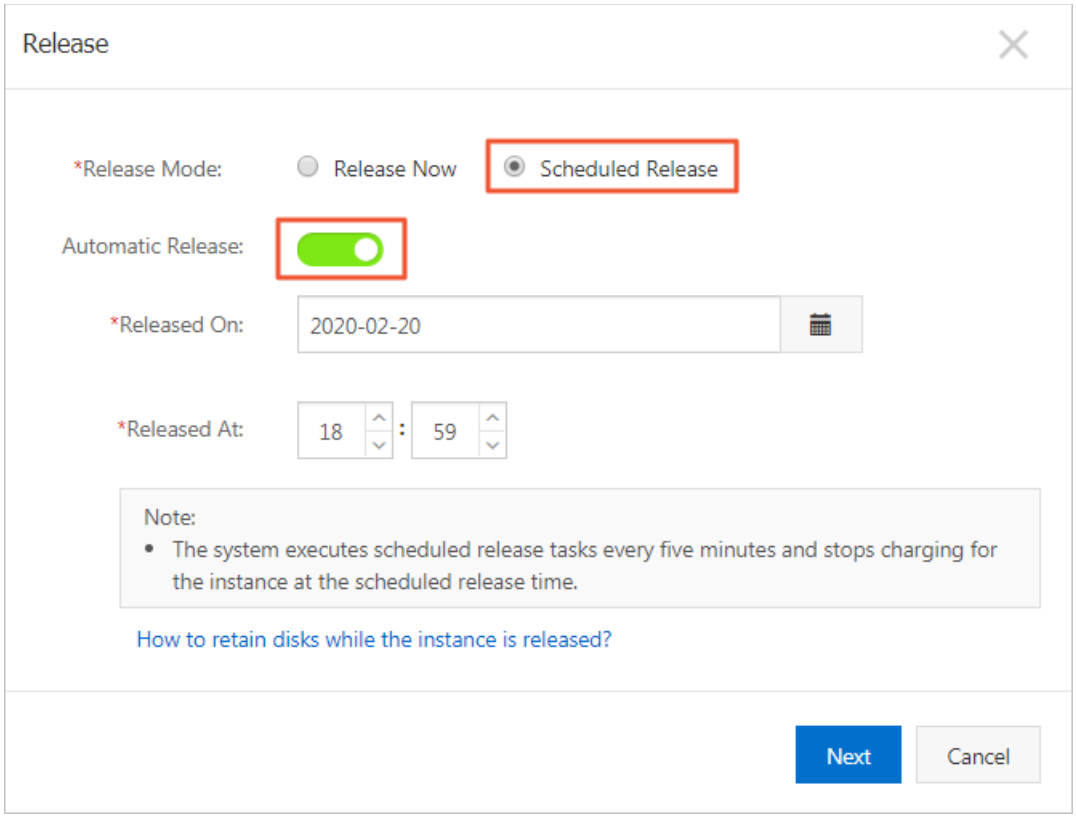
5. In the Release dialog box, select **Release Now**.
6. Click **Next**. Then, click **OK**.

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

9.Manage instance attributes

9.1. View instance information

This topic describes how to obtain an overview of instances and details of a single instance in your account.

View information of instances on the Overview page

When you log on to the ECS console, the **Overview** page appears.

On the **Overview** page, you can view the following information of ECS instances in your account:

- **Pending Events**

  Lists all pending events and instances that are associated with the events.

- **My Resources**

  Lists ECS instances and other resources in each region.

View information of instances on the Instances page

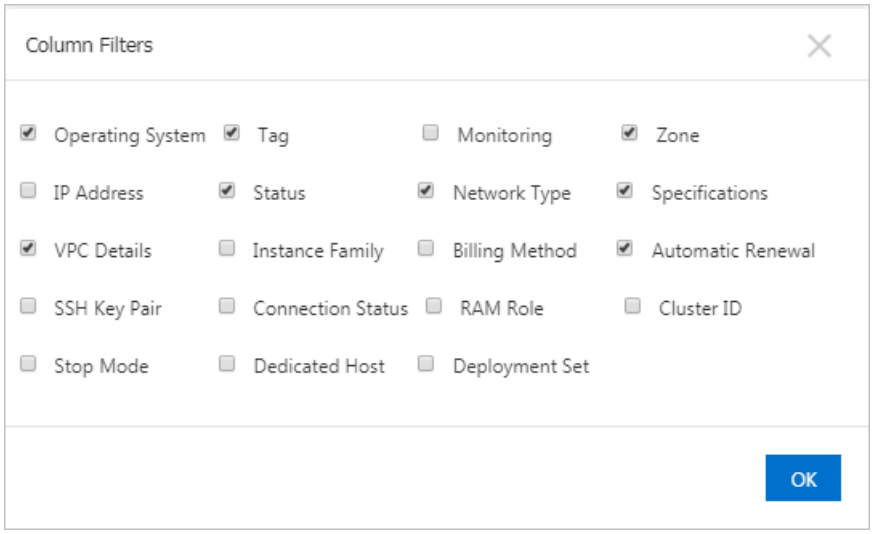
To go to the **Instances** page, perform the following steps:

1.
2.
3.
4. On the **Instances** page, view the information of all ECS instances in a specific region, such as **Instance ID/Name**, **Zone**, **IP Address**, **Status**, **Network Type**, **Billing Method**, and **Actions**.

   To configure **Column Filters**, perform the following steps:

   i. In the upper-right corner of the **Instances** page, click the icon.

   ii. In the **Column Filters** dialog box, select the instance information that you want to view and click **OK**.



View information of a single instance on the Instance Details page

To go to the **Instances Details** page and view the information of a single instance, perform the following steps:

1.
2.
3.
4. Find the ECS instance that you want to view and click the instance ID. Alternatively, click **Manage** in the corresponding **Actions** column.

   The following table describes the information that is displayed on the **Instance Details** page.

Information	Description
Basic Information	The information related to instance identifier such as the instance ID, public IP address, security group, region, zone, and hostname.
Configuration Information	The information related to instance configurations such as the CPU and memory, operating system, instance type, instance family, cloud disk, snapshot, image ID, current bandwidth value, and VPC (only for VPC-type instances).
Network Information	The information related to instance network such as the network type, elastic network interfaces (ENIs), VPC, vSwitch, and primary private IP address.
Billing Information	The information related to instance billing such as the billing method, auto-renewal, and billing method for network usage.
Other Information	The information related to instance O&M such as the maintenance property, instance I/O optimization type, cluster ID, and release protection configuration.

You can switch from the **Instance Details** tab to the **Cloud Disks**, **Snapshot**, or **Security Groups** tab to view other types of instance resources.

Related information

- DescribeInstances

9.2. Modify the properties of an instance

After an Elastic Compute Service (ECS) instance is created, you can modify its name, hostname, and description. If the instance is a pay-as-you-go instance, you can enable or disable release protection for the instance.

Procedure

1.
2.

3.
4. Find the instance whose properties you want to modify and choose **More > Instance Settings > Modify Instance Properties** in the **Actions** column.
5. Modify the properties of the instance.
6. Click **OK**.
7. If the hostname of the instance is modified, restart the instance for the new hostname to take effect.

> Note

Related information

- ModifyInstanceAttribute

9.3. Customize CPU options

9.3.1. Specify and view CPU options

The CPU options of an Elastic Compute Service (ECS) instance include the number of physical CPU cores and the number of threads per core. For some ECS instance types, you can specify these options when you call the RunInstances operation to create an instance.

CPU and vCPU

CPUs are central processing units. A single CPU can contain several physical cores. The Hyper-Threading (HT) technology can be used to create two virtual processing cores for each physical core that is present in a CPU. Virtual CPUs (vCPUs) are virtual processing cores of ECS instances.

Alibaba Cloud ECS supports multi-threading based on HT of the x86 architecture. HT enables two threads to concurrently run on a single physical core. Each thread can be considered as a vCPU.

The following table describes the CPU options for ECS instances.

CPU option	API parameter	Description	Scenario	Supported instance type
Number of physical CPU cores	CpuOptions.Core	Specifies the number of physical CPU cores to use.	You can use a smaller number of physical CPU cores to improve the CPU-to-memory ratio of an instance. This can reduce the number of billable items and software licensing costs.	For more information, see Limits.

CPU option	API parameter	Description	Scenario	Supported instance type
Number of threads per core	CpuOptions.ThreadsPerCore	Specifies whether to enable HT on the CPU. Number of vCPUs = Number of physical CPU cores × Number of threads per core.	In most cases, the default configuration of an ECS instance type provides sufficient performance. You can disable HT in the following scenarios: • High-performance computing (HPC) scenarios. In these scenarios, you can disable HT to improve the performance of instances. • Memory-intensive business scenarios. You can disable HT to reduce the number of vCPUs and increase the CPU-to-memory ratio. This can also reduce the number of billable items and software licensing costs.	For more information, see Limits.

Billing

You can specify CPU options at no additional costs.

Limits

- The following instance families support custom CPU options.

  > **Note** Click the following links to check the default and valid values for the number of physical CPU cores (CpuOptions.Core) and the number of threads per core (CpuOptions.ThreadsPerCore). Instance types that are not listed do not support custom CPU options.

  - General-purpose instance families:g7a, g7, g7t, g7ne, g6t, g6a, g6e, and g6
  - Compute-optimized instance families:c7a, c7, c7t, c6t, c6a, c6e, and c6
  - Memory-optimized instance families: r7a, r7, r7t, re6p, r6a, r6e, and r6
  - Instance families with high clock speeds: hfg7, hfc7, hfr7, hfg6, hfc6, and hfr6
  - Instance families with local SSDs: i3g and i3
- CPU options can be specified only when you create an ECS instance. You cannot modify CPU options after the instance is created.
- If you upgrade or downgrade the configurations of an instance, the custom CPU options are changed to the default CPU options of the new instance type.
- The instance type of an instance determines the number of physical cores available for the instance. You can specify the number of physical CPU cores to be enabled within the specified value range.

Enable or disable HT

You can call the RunInstances operation to specify the CPU options of an ECS instance. If you want to use an Alibaba Cloud ECS SDK, upgrade the SDK to the latest version.

- By default, HT is enabled on ECS instances. You can enable HT by using Alibaba Cloud CLI. The

following code shows a sample request:

```
aliyun ecs RunInstances --RegionId cn-hangzhou --CpuOptions.Core 2 --CpuOptions.ThreadsPerCore 2 --ImageId ubuntu_18_04_64_20G_alibase_20190624.vhd --InstanceType ecs.g6.xlarge --SecurityGroupId sg-bp67acfmxazb4ph*** --VSwitchId vsw-bp1s5fnvk4gn2tws03*** --Amount 1 --SystemDisk.AutoSnapshotPolicyId sp-bp67acfmxazb4ph***
```

- To disable HT, set the CpuOptions.ThreadsPerCore parameter to *1* by using Alibaba Cloud CLI. The following code shows a sample request:

```
aliyun ecs RunInstances --RegionId cn-hangzhou --CpuOptions.Core 2 --CpuOptions.ThreadsPerCore 1 --ImageId ubuntu_18_04_64_20G_alibase_20190624.vhd --InstanceType ecs.g6.xlarge --SecurityGroupId sg-bp67acfmxazb4ph*** --VSwitchId vsw-bp1s5fnvk4gn2tws03*** --Amount 1 --SystemDisk.AutoSnapshotPolicyId sp-bp67acfmxazb4ph***
```

For example, the ecs.g6.xlarge instance type provides 2 physical CPU cores by default.

- If you enable HT for an instance of this instance type and set the number of threads per core to 2, the instance has 4 vCPUs. The number of vCPUs is calculated by using the formula described in the preceding table: 4 (Number of vCPUs) = 2 (Number of physical CPU cores) × 2 (Number of threads per core). By default, HT is enabled for this instance type.
- If you disable HT for an instance of this instance type, only one thread can run on each physical CPU core. This way, the instance has 2 vCPUs, which is equal to the number of its physical CPU cores.

View CPU options

You can call the DescribeInstances operation to view the CPU options of an ECS instance. If you want to use an Alibaba Cloud ECS SDK, upgrade the SDK to the latest version.

For example, you can run the following sample request in Alibaba Cloud CLI to view the CPU options of an instance:

```
aliyun ecs DescribeInstances --InstanceIds '["i-bp19rxmzeocge2z57***"]' --output cols=CpuOptions rows=Instances.Instance[]
```

Sample response:

```
CpuOptions
----------
map[CoreCount:1 ThreadsPerCore:2]
```

In the response, `CoreCount:1` indicates that the number of physical CPU cores is 1, and `ThreadsPerCore:2` indicates that the number of threads per core is 2.

9.3.2. CPU options of general-purpose instance families

This topic lists the default and valid values for the number of physical CPU cores and the number of threads per core of general-purpose instance families. You can use these values when you specify CPU options.

Default and valid values for the number of physical CPU cores and the number of threads per core of the g7a instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.g7a.large	2	1	2	1, 2
ecs.g7a.xlarge	4	2	2	1, 2
ecs.g7a.2xlarge	8	2, 4	2	1, 2
ecs.g7a.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.g7a.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.g7a.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.g7a.32xlarge	128	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the g7 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.g7.large	2	1	2	1, 2
ecs.g7.xlarge	4	2	2	1, 2
ecs.g7.2xlarge	8	2, 4	2	1, 2
ecs.g7.3xlarge	12	2, 4, 6	2	1, 2
ecs.g7.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.g7.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.g7.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.g7.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.g7.32xlarge	128	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the g7t instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.g7t.large	2	1	2	1, 2
ecs.g7t.xlarge	4	2	2	1, 2
ecs.g7t.2xlarge	8	2, 4	2	1, 2
ecs.g7t.3xlarge	12	2, 4, 6	2	1, 2
ecs.g7t.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.g7t.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.g7t.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.g7t.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.g7t.32xlarge	128	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the g7ne instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.g7ne.large	2	1	2	1, 2
ecs.g7ne.xlarge	4	2	2	1, 2
ecs.g7ne.2xlarge	8	2, 4	2	1, 2
ecs.g7ne.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.g7ne.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.g7ne.12xlarge	48	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24	2	1, 2
ecs.g7ne.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.g7ne.24xlarge	96	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the g6t instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.g6t.large	2	1	2	1, 2
ecs.g6t.xlarge	4	2	2	1, 2
ecs.g6t.2xlarge	8	2, 4	2	1, 2
ecs.g6t.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.g6t.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.g6t.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.g6t.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the g6a instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.g6a.large	2	1	2	1, 2
ecs.g6a.xlarge	4	2	2	1, 2
ecs.g6a.2xlarge	8	2, 4	2	1, 2
ecs.g6a.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.g6a.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.g6a.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.g6a.32xlarge	128	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the g6e instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.g6e.large	2	1	2	1, 2
ecs.g6e.xlarge	4	2	2	1, 2
ecs.g6e.2xlarge	8	2, 4	2	1, 2

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.g6e.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.g6e.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.g6e.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2
ecs.g6e.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the g6 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.g6.large	2	1	2	1, 2
ecs.g6.xlarge	4	2	2	1, 2
ecs.g6.2xlarge	8	2, 4	2	1, 2
ecs.g6.3xlarge	12	2, 4, 6	2	1, 2
ecs.g6.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.g6.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.g6.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.g6.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2
ecs.g6.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2

9.3.3. CPU options of compute optimized instance families

This topic lists the default and valid values for the number of physical CPU cores and the number of threads per core of compute optimized instance families. You can query these values when you customize CPU options.

Default and valid values for the number of physical CPU cores and the number of threads per core of the c7a instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.c7a.large	2	1	2	1, 2
ecs.c7a.xlarge	4	2	2	1, 2
ecs.c7a.2xlarge	8	2, 4	2	1, 2
ecs.c7a.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.c7a.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.c7a.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.c7a.32xlarge	128	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the c7 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.c7.large	2	1	2	1, 2
ecs.c7.xlarge	4	2	2	1, 2
ecs.c7.2xlarge	8	2, 4	2	1, 2
ecs.c7.3xlarge	12	2, 4, 6	2	1, 2

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.c7.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.c7.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.c7.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.c7.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.c7.32xlarge	128	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the c7t instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.c7t.large	2	1	2	1, 2
ecs.c7t.xlarge	4	2	2	1, 2
ecs.c7t.2xlarge	8	2, 4	2	1, 2
ecs.c7t.3xlarge	12	2, 4, 6	2	1, 2
ecs.c7t.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.c7t.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.c7t.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.c7t.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.c7t.32xlarge	128	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the c6t instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.c6t.large	2	1	2	1, 2
ecs.c6t.xlarge	4	2	2	1, 2
ecs.c6t.2xlarge	8	2, 4	2	1, 2
ecs.c6t.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.c6t.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.c6t.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2
ecs.c6t.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the c6a instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.c6a.large	2	1	2	1, 2
ecs.c6a.xlarge	4	2	2	1, 2
ecs.c6a.2xlarge	8	2, 4	2	1, 2

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.c6a.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.c6a.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.c6a.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.c6a.32xlarge	128	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the c6e instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.c6e.large	2	1	2	1, 2
ecs.c6e.xlarge	4	2	2	1, 2
ecs.c6e.2xlarge	8	2, 4	2	1, 2
ecs.c6e.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.c6e.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.c6e.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2
ecs.c6e.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the c6 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.c6.large	2	1	2	1, 2
ecs.c6.xlarge	4	2	2	1, 2
ecs.c6.2xlarge	8	2, 4	2	1, 2
ecs.c6.3xlarge	12	2, 4, 6	2	1, 2
ecs.c6.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.c6.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.c6.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.c6.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2
ecs.c6.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2

9.3.4. CPU options of memory optimized instance families

This topic lists the default and valid values for the number of physical CPU cores and the number of threads per core of memory optimized instance families. You can query these values when you customize CPU options.

Default and valid values for the number of physical CPU cores and the number of threads per core of the r7a instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.r7a.large	2	1	2	1, 2
ecs.r7a.xlarge	4	2	2	1, 2
ecs.r7a.2xlarge	8	2, 4	2	1, 2
ecs.r7a.4xlarge	16	2, 4, 6, 8	2	1, 2

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.r7a.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.r7a.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.r7a.32xlarge	128	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the r7 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.r7.large	2	1	2	1, 2
ecs.r7.xlarge	4	2	2	1, 2
ecs.r7.2xlarge	8	2, 4	2	1, 2
ecs.r7.3xlarge	12	2, 4, 6	2	1, 2
ecs.r7.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.r7.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.r7.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.r7.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.r7.32xlarge	128	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the r7t instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.r7t.large	2	1	2	1, 2
ecs.r7t.xlarge	4	2	2	1, 2
ecs.r7t.2xlarge	8	2, 4	2	1, 2
ecs.r7t.3xlarge	12	2, 4, 6	2	1, 2
ecs.r7t.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.r7t.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.r7t.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.r7t.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.r7t.32xlarge	128	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the re6p instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.re6p.large	2	1	2	1, 2
ecs.re6p.xlarge	4	2	2	1, 2
ecs.re6p.2xlarge	8	2, 4	2	1, 2
ecs.re6p.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.re6p.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.re6p.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2
ecs.re6p-redis.large	2	1	2	1, 2
ecs.re6p-redis.xlarge	4	2	2	1, 2
ecs.re6p-redis.2xlarge	8	2, 4	2	1, 2
ecs.re6p-redis.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.re6p-redis.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2
ecs.re6p-redis.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the r6a instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.r6a.large	2	1	2	1, 2
ecs.r6a.xlarge	4	2	2	1, 2
ecs.r6a.2xlarge	8	2, 4	2	1, 2
ecs.r6a.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.r6a.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.r6a.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the r6e instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.r6e.large	2	1	2	1, 2
ecs.r6e.xlarge	4	2	2	1, 2
ecs.r6e.2xlarge	8	2, 4	2	1, 2
ecs.r6e.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.r6e.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.r6e.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2
ecs.r6e.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the r6 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.r6.large	2	1	2	1, 2
ecs.r6.xlarge	4	2	2	1, 2
ecs.r6.2xlarge	8	2, 4	2	1, 2
ecs.r6.3xlarge	12	2, 4, 6	2	1, 2
ecs.r6.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.r6.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.r6.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.r6.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.r6.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2

9.3.5. CPU options of instance families with high clock speeds

This topic lists the default and valid values for the number of physical CPU cores and the number of threads per core of instance families with high clock speeds. You can query these values when you customize CPU options.

Default and valid values for the number of physical CPU cores and the number of threads per core of the hfg7 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.hfg7.large	2	1	2	1, 2
ecs.hfg7.xlarge	4	2	2	1, 2
ecs.hfg7.2xlarge	8	2, 4	2	1, 2
ecs.hfg7.3xlarge	12	2, 4, 6	2	1, 2
ecs.hfg7.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.hfg7.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.hfg7.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.hfg7.12xlarge	48	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24	2	1, 2
ecs.hfg7.24xlarge	96	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the hfc7 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.hfc7.large	2	1	2	1, 2
ecs.hfc7.xlarge	4	2	2	1, 2
ecs.hfc7.2xlarge	8	2, 4	2	1, 2
ecs.hfc7.3xlarge	12	2, 4, 6	2	1, 2
ecs.hfc7.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.hfc7.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.hfc7.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.hfc7.12xlarge	48	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24	2	1, 2
ecs.hfc7.24xlarge	96	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the hfr7 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.hfr7.large	2	1	2	1, 2
ecs.hfr7.xlarge	4	2	2	1, 2
ecs.hfr7.2xlarge	8	2, 4	2	1, 2
ecs.hfr7.3xlarge	12	2, 4, 6	2	1, 2
ecs.hfr7.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.hfr7.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.hfr7.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.hfr7.12xlarge	48	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24	2	1, 2
ecs.hfr7.24xlarge	96	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the hfg6 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.hfg6.large	2	1	2	1, 2
ecs.hfg6.xlarge	4	2	2	1, 2
ecs.hfg6.2xlarge	8	2, 4	2	1, 2
ecs.hfg6.3xlarge	12	2, 4, 6	2	1, 2
ecs.hfg6.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.hfg6.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.hfg6.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.hfg6.10xlarge	40	2, 4, 6, 8, 10, 12, 14, 16, 18, 20	2	1, 2
ecs.hfg6.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.hfg6.20xlarge	80	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the hfc6 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.hfc6.large	2	1	2	1, 2
ecs.hfc6.xlarge	4	2	2	1, 2
ecs.hfc6.2xlarge	8	2, 4	2	1, 2
ecs.hfc6.3xlarge	12	2, 4, 6	2	1, 2
ecs.hfc6.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.hfc6.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.hfc6.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.hfc6.10xlarge	40	2, 4, 6, 8, 10, 12, 14, 16, 18, 20	2	1, 2
ecs.hfc6.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.hfc6.20xlarge	80	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the hfr6 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.hfr6.large	2	1	2	1, 2
ecs.hfr6.xlarge	4	2	2	1, 2
ecs.hfr6.2xlarge	8	2, 4	2	1, 2
ecs.hfr6.3xlarge	12	2, 4, 6	2	1, 2
ecs.hfr6.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.hfr6.6xlarge	24	2, 4, 6, 8, 10, 12	2	1, 2
ecs.hfr6.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.hfr6.10xlarge	40	2, 4, 6, 8, 10, 12, 14, 16, 18, 20	2	1, 2
ecs.hfr6.16xlarge	64	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32	2	1, 2
ecs.hfr6.20xlarge	80	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40	2	1, 2

9.3.6. CPU options of instance families with local SSDs

This topic lists the default and valid values for the number of physical CPU cores and the number of threads per core of instance families with local SSDs. You can query these values when you customize CPU options.

Default and valid values for the number of physical CPU cores and the number of threads per core of the i3g instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.i3g.2xlarge	8	2, 4	2	1, 2
ecs.i3g.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.i3g.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.i3g.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2
ecs.i3g.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2

Default and valid values for the number of physical CPU cores and the number of threads per core of the i3 instance family

Instance type	Default value for the number of vCPUs	Valid values for the number of physical CPU cores	Default value for the number of threads per core	Valid values for the number of threads per core
ecs.i3.xlarge	4	2	2	1, 2
ecs.i3.2xlarge	8	2, 4	2	1, 2
ecs.i3.4xlarge	16	2, 4, 6, 8	2	1, 2
ecs.i3.8xlarge	32	2, 4, 6, 8, 10, 12, 14, 16	2	1, 2
ecs.i3.13xlarge	52	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26	2	1, 2
ecs.i3.26xlarge	104	2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52	2	1, 2

9.4. Reset the logon password of an instance

Change the password

If you did not set a logon password when you created an Elastic Compute Service (ECS) instance or if you forget the logon password of an instance, you must use the password reset feature to set a logon password for the instance. This topic describes how to reset the logon password of an ECS instance in the ECS console.

Prerequisites

The ECS instance whose logon password you want to reset is in a stable state, such as the **Stopped** or **Running** state. For more information about stable instance states, see Instance lifecycle.

Context

This topic focuses on how to reset the logon password of an ECS instance in the ECS console. You can also use Cloud Assistant to reset the logon password of an ECS instance, or connect to an ECS instance and reset its logon password. The following table compares the different methods to reset the logon password of an ECS instance.

Method	Description	References

Method	Description	References
Reset the logon password of an ECS instance in the ECS console	• The instance must be in the **Running** or **Stopped** state. • You do not need to log on to the instance. • After you reset the password, you must restart the instance for the new password to take effect. This may affect the state of the business that is running on the instance.	This topic
Use Cloud Assistant to reset the logon password of an ECS instance	• The instance must be in the **Running** state. • You do not need to log on to the instance. • After you reset the password, you do not need to restart the instance for the new password to take effect.	Change the logon password of an instance
Connect to the instance and reset its logon password	• The instance must be in the **Running** state. • You must connect to the instance. The procedure to connect to an instance is complex. • After you reset the password, you do not need to restart the instance for the new password to take effect.	Change the logon password of an instance by connecting to the instance

Considerations

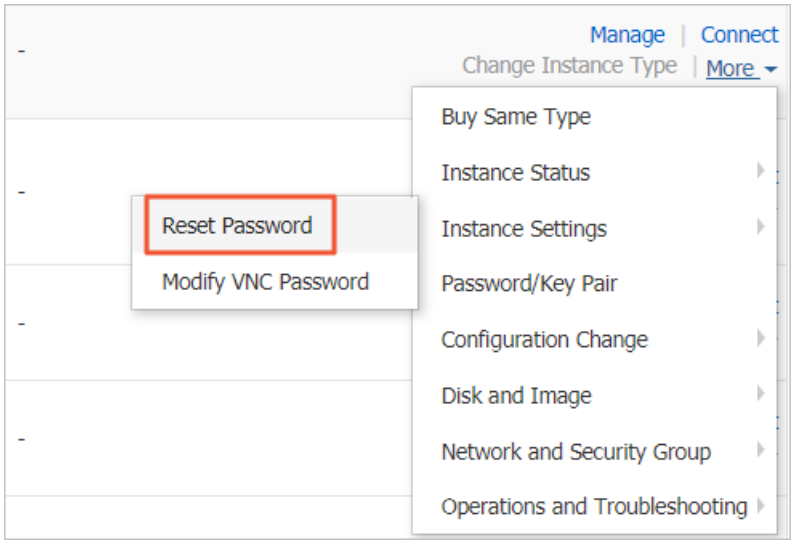
When you reset the logon password of an ECS instance in the ECS console, take note of the following items:

- Select a proper time to reset the logon password. If the instance is in the **Running** state, you must restart the instance after the logon password is reset for the new password to take effect. The instance restart may affect your business. We recommend that you reset the password during off-peak hours to minimize the impact on business.
- The key pair authentication method overrides the username and password authentication method. Linux instances can be logged on to by using key pairs or by using usernames and passwords. If you bind a key pair to a Linux instance that already has a logon password, the password becomes invalid and you can only use the key pair to log on to the instance. If you want to use a password to log on to the Linux instance, you must reset the logon password.

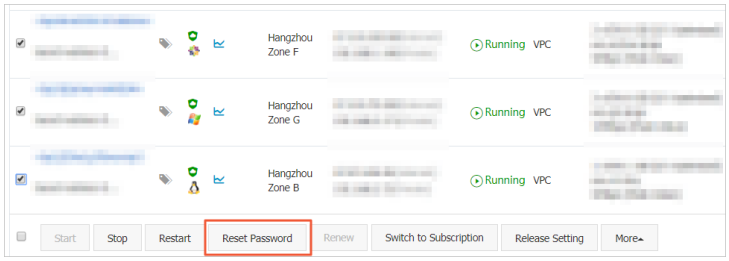
Procedure

1.

2.
3.
4. Perform one of the following operations based on the number of instances whose logon passwords you want to reset at a time:
   - To reset the logon password of a single instance at a time, find the instance and choose **More > Password/Key Pair > Reset Password** in the **Actions** column.



   - To reset the password of multiple instances at a time, select the instances and click **Reset Password** below the instance list.



5. In the **Reset Password** dialog box, enter and confirm a new valid password and then click **OK**.

   The entered password must be 8 to 30 characters in length and contain at least three of the following character types:

   - Uppercase letters
   - Lowercase letters
   - Digits
   - Special characters. The supported special characters include `( ) ` ~ ! @ # $ % ^ & * - _ + = | { } [ ] : ; ' < > , . ? /`

   > **Note** For Windows instances, logon passwords cannot start with a forward slash (/).

   At this point, the logon password is reset. Then, you must start or restart the instance for the new password to take effect.
6. Perform one of the following operations based on the instance state for the new password to take effect:

- If the instance is in the **Running** state, click **Restart Now**.
- If the instance is in the **Stopped** state, click Cancel. When you manually start the instance, the new password takes effect immediately.

Related information

- ModifyInstanceAttribute

9.5. Change the logon password of an instance by connecting to the instance

You can change the logon password of an instance without going to the console when you perform operations on the instance. The change takes effect immediately and you do not need to restart the instance. This topic describes how to change the logon passwords of a Linux instance and a Windows instance by connecting to the instances. Windows and CentOS are used in the examples.

Change the logon password of a Linux instance

An instance that runs CentOS 7.6 is used in this example. Perform the following operations to change the logon password of the Linux instance:

1. Log on to the instance. For more information, see Overview.
2. Run the **passwd <username>** command. Example: **passwd root**.
3. Enter a new password.
4. Enter the new password again to confirm the password.

Change the logon password of a Windows instance

An instance that runs Windows Server 2012 is used in this example. Perform the following operations to change the logon password of the Windows instance:

1. Log on to the instance. For more information, see Overview.
2. Choose **Start > Run**, enter **compmgmt.msc**, and then press the Enter key.
3. In the **Computer Management** window, choose **System Tools > Local Users and Groups > Users**.
4. Right-click the username for which you want to change the password. Example: **Administrator**.
5. Select **Set Password**.
6. In the **Set Password for Administrator** dialog box, click **Proceed**. In the dialog box that appears, enter a new password in the **New password** and **Confirm password** fields, and then click **OK**.

Related information

- Restart an instance
- RebootInstance

9.6. Enable or disable release protection for ECS instances

You can enable release protection for pay-as-you-go instances to prevent potential irreversible consequences arising from accidental manual instance release. This topic describes how to enable and disable release protection for ECS instances, how to check whether release protection is enabled, and how release protection is implemented.

Prerequisites

The instance is a pay-as-you-go instance.

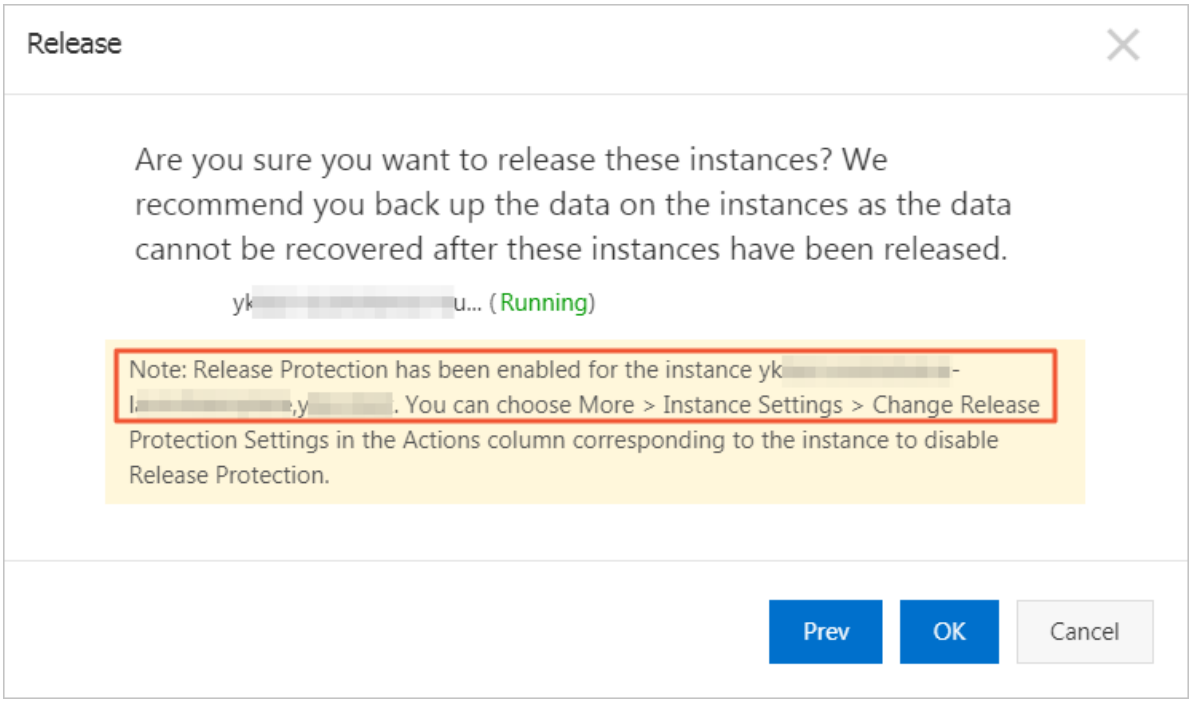
Context

The release protection feature cannot prevent the automatic release of instances in normal scenarios such as the following ones:

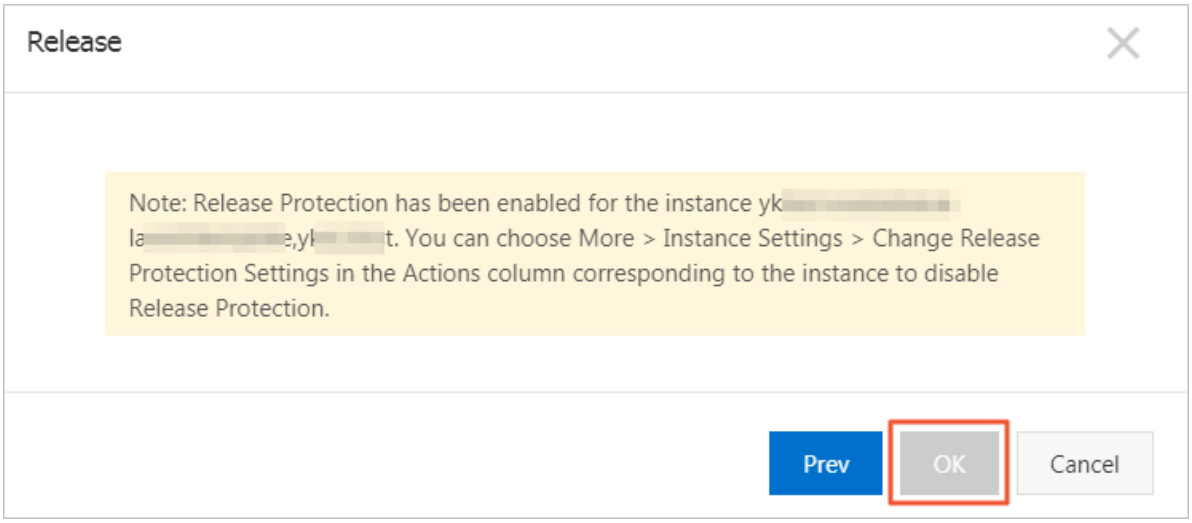
- A payment in your account is overdue for more than 15 days.
- The automatic release time that you set for the instance has been reached.
- The instance does not comply with the applicable security compliance policies.
- The instance was automatically created by Auto Scaling and is removed by subsequent scale-in events.

The following examples show how release protection is implemented:

- When you attempt to manually release instances in the ECS console, instances with release protection enabled are automatically skipped.



- When you attempt to manually release instances in the ECS console, the selected instances cannot be released if they all have release protection enabled.

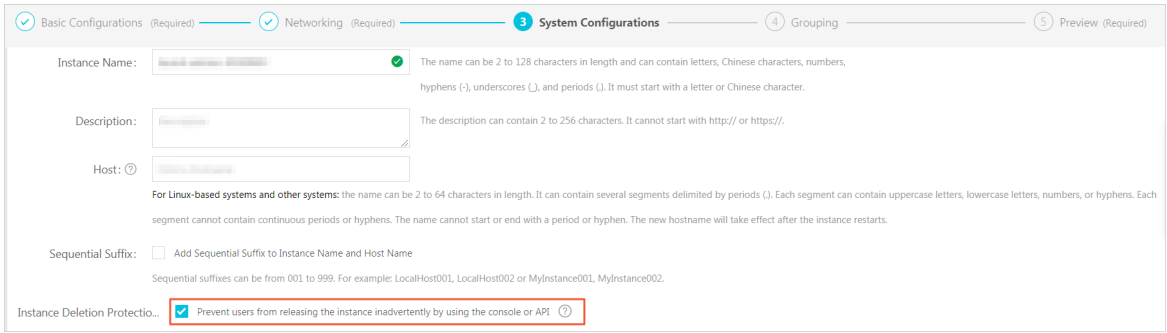


- The `InvalidOperation.DeletionProtection` error code is returned if you attempt to call the DeleteInstance operation to release an instance with release protection enabled.

Enable release protection when you create an instance

This section describes how to configure release protection settings when you create an instance. For more information about how to create an instance, see Create an instance by using the provided wizard.

1.
2.
3.
4. On the **Instances** page, click **Create Instance**.
5. In the **Basic Configurations** step, set **Billing Method** to **Pay-As-You-Go** and complete the remaining configurations. Click **Next: Networking**.
6. In the **Networking** step, complete all configurations. Click **Next: System Configurations**.
7. In the **System Configurations** step, select **Prevent users from releasing the instance inadvertently by using the console or API** and complete the remaining configurations. Click **Next: Grouping**.



8. Complete the remaining configurations until the instance is created.

When you call the RunInstances or CreateInstance operation to create an instance, you can enable or disable release protection for the instance by setting the DeletionProtection parameter.

Change the release protection settings

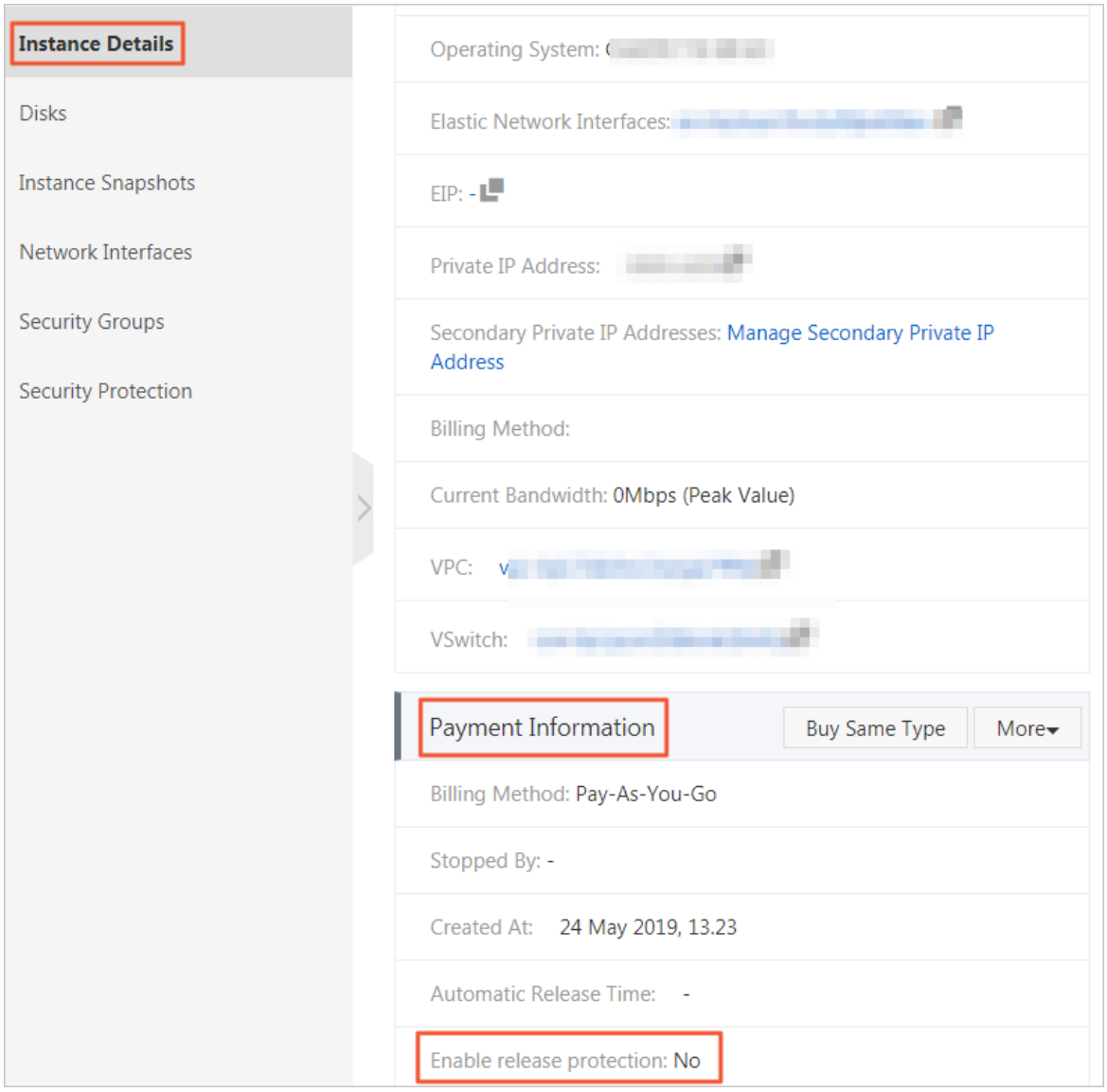
You can also enable or disable release protection for an instance by modifying the attributes of the instance.

1.
2.
3.
4. On the **Instances** page, use one of the following methods to change the release protection settings of instances:
   - Change the release protection settings of a single instance: Find the instance for which you want to change the release protection setting, and choose **More > Instance Settings > Change Release Protection Setting** in the **Actions** column.
   - Change the release protection settings of one or more instances: Select multiple instances and choose **More > Instance Settings > Change Release Protection Setting** in the lower part of the Instances page.
5. In the Change Release Protection Setting dialog box, turn on or off **Release Protection**.
6. Click **OK**.

When you call the ModifyInstanceAttribute operation to modify the attributes of an instance, you can enable or disable release protection for the instance by setting the DeletionProtection parameter.

Check whether release protection is enabled

1.
2.
3.
4. On the **Instances** page, use one of the following methods to view details of an instance:
   - In the **Instance ID/Name** column, click the ID of the instance.
   - Find the instance and click **Manage** in the **Actions** column.
5. On the **Instance Details** tab, check whether release protection is enabled in the **Release Protection** item of the **Other Information** section.



Related information

References

- DeleteInstance
- RunInstances
- CreateInstance
- ModifyInstanceAttribute

9.7. Edit the tags of an instance

Tags can be used to identify resources with the same characteristics (such as instances that belong to the same organization or that serve the same purpose) for easy search and management. This topic describes how to edit the tags of an existing instance.

Context

For information about how to use tags, the resources that support tags, and the limits on tags, see

Overview and the "Tag limits" section of the Limits topic.

Procedure

1.
2.
3.
4. Find the instance whose tags you want to edit, move the pointer over the  icon in the **Tag** column, and then click **Edit Tags**.
5. In the **Edit Tags** dialog box, click Available Tags to select existing tags or click Create to create tags. Then, click **OK**.

What's next

After tags are added to your instances, you can filter the instances by tag to perform different O&M operations. For example, you can grant RAM users access to instances that have a set of tags and configure instances that have a different set of tags to automatically start or stop at specified points in time.

10.Manage instance configurations

10.1. Send remote commands

You can send remote commands to perform O&M operations on one or more instances without logging on to the instances.

Prerequisites

- The instances that are used to receive commands are in the **Running** state.
- The Cloud Assistant client is started on the instances. For more information, see Start or stop the Cloud Assistant client.

Context

You must use Cloud Assistant to send and run remote commands, and the remote commands consume the quota of Cloud Assistant commands. For more information about Cloud Assistant and its limits, see Overview.

Procedure

1.
2.
3.
4. Select one of the following methods to send a remote command:
   - To send a remote command to a single instance, find the instance on the Instances page and click its ID to go to the **instance details** page. Click the **Remote Commands** tab and then click **Send Command**.
   - To send a remote command to multiple instances, select the instances on the Instances page. In the lower part of the page, choose **More > Send Command**.
5. In the dialog box that appears, perform the following operations.

   Note If some instances run Linux and others run Windows, you must configure the instances based on their operating systems when you perform batch operations on multiple instances.

   i. Select a command type.
      - Linux instances: **Shell** is selected by default.
      - Windows instances: Select **Bat** or **PowerShell**.
   ii. Specify whether to retain commands.

   Note You can view the retained commands on the Cloud Assistant page and run these commands repeatedly. For more information about how to use Cloud Assistant to run remote commands on ECS instances, see Run a command.

iii. In the **Command Content** code editor, enter a command.

> **Note**
> - The command must be able to return the results of a single execution. Interactions with returned information are not allowed.
> - For more information about shell commands, see View instance configurations.

iv. Click **Run**.

- You can click **Stop** to cancel the execution of a command.
- After the execution is complete, you can view the command output in the **Command Output** section.

  To view the command output and execution results of each instance, click the instance ID.

> **Note** After an execution is complete, you can enter another command in the **Command Content** code editor to run the command.

10.2. Manage instance metadata

10.2.1. Overview of ECS instance metadata

Elastic Compute Service (ECS) instance metadata contains the information of ECS instances in Alibaba Cloud. You can view the metadata of running instances and configure or manage the instances based on their metadata.

Instance metadata types

The following table describes the types of instance metadata. For more information about the instance metadata items and their definitions, see Instance metadata items.

Type	Description	References
Basic metadata	The basic metadata of an instance includes the following information: • Basic information such as the instance ID, IP address, media access control (MAC) address of the network interface controller (NIC), and operating system. • System events, including underlying operations and maintenance (O&M) events or unexpected maintenance events. You can use system events to have a timely understanding of the running status of the instance.	• View instance metadata • Overview

Type	Description	References
Dynamic metadata	The dynamic metadata of an instance includes only the identifier of the instance. Each instance identifier consists of an instance identity document and an instance identity signature. Instance identifiers are generated in real time and are typically used to identify instances. This can provide an important trust foundation for application permission control and software activation. You can view the identifier of an instance based on the metadata of the instance.	• View instance metadata • Use instance identities

You can also use the user data of instances to manage startups of the instances in a flexible manner. For more information, see Overview of ECS instance user data.

Limits

The instance metadata feature is supported only for instances that reside in virtual private clouds (VPCs).

10.2.2. Instance metadata items

This topic describes the basic and dynamic metadata items that you can obtain from an Elastic Compute Service (ECS) instance.

Basic metadata items

Item	Description	Example
/meta-data/dns-conf/nameservers	The Domain Name System (DNS) configurations of the instance.	100.100.XX.XX
/meta-data/hostname	The hostname of the instance.	iZbp13znx0m0me8cquu****
/meta-data/instance/instance-type	The instance type of the instance.	ecs.g6e.large
/meta-data/image-id	The ID of the image used to create the instance.	aliyun_3_x64_20G_alibase_20210425.vhd
/meta-data/image/market-place/product-code	The product code of the Alibaba Cloud Marketplace image.	cmjj01****
/meta-data/image/market-place/charge-type	The billing method of the Alibaba Cloud Marketplace image.	PrePaid

Item	Description	Example
/meta-data/instance-id	The ID of the instance.	i-bp13znx0m0me8cquu****
/meta-data/mac	The media access control (MAC) address of the instance. If the instance has multiple network interface controllers (NICs), only the MAC address of eth0 is displayed.	00:16:3e:0f:XX:XX
/meta-data/network-type	The network type of the instance. Only instances that reside in virtual private clouds (VPCs) are supported.	vpc
/meta-data/network/interfaces/macs/	The MAC addresses of the NICs.	00:16:3e:0f:XX:XX
/meta-data/network/interfaces/macs/[mac]/network-interface-id	The identifier of the NIC. Replace [mac] with the MAC address of the instance.	eni-bp1b2c0jvnj0g17b****
/meta-data/network/interfaces/macs/[mac]/netmask	The subnet mask of the NIC.	255.255.XX.XX
/meta-data/network/interfaces/macs/[mac]/vswitch-cidr-block	The IPv4 CIDR block of the vSwitch to which the NIC is connected.	192.168.XX.XX/24
/meta-data/network/interfaces/macs/[mac]/vpc-cidr-block	The IPv4 CIDR block of the VPC to which the NIC belongs.	192.168.XX.XX/16
/meta-data/network/interfaces/macs/[mac]/private-ipv4s	The private IPv4 addresses assigned to the NIC.	["192.168.XX.XX"]
/meta-data/network/interfaces/macs/[mac]/vswitch-id	The ID of the vSwitch that resides within the same VPC as the security group of the NIC.	vsw-bp1ygryo03m39xhsy****
/meta-data/network/interfaces/macs/[mac]/vpc-id	The ID of the VPC to which the security group of the NIC belongs.	vpc-bp1e0g399hkd7c8q3****

Item	Description	Example
/meta-data/network/interfaces/macs/[mac]/primary-ip-address	The primary private IP address of the NIC.	192.168.XX.XX
/meta-data/network/interfaces/macs/[mac]/gateway	The IPv4 gateway address of the VPC to which the NIC belongs.	192.168.XX.XX
/meta-data/instance/max-netbw-egress	The maximum outbound internal bandwidth of the instance. Unit: Kbit/s.	1228800
/meta-data/instance/max-netbw-ingress	The maximum inbound internal bandwidth of the instance. Unit: Kbit/s.	1228800
/meta-data/private-ipv4	The private IPv4 address of the primary NIC.	192.168.XX.XX
/meta-data/eipv4	This metadata item is used to obtain the following information: • The public IPv4 address of the instance • The elastic IPv4 address associated with the primary NIC	120.55.XX.XX
/meta-data/ntp-conf/ntp-servers	The domain name of the Network Time Protocol (NTP) server.	ntp1.aliyun.com
/meta-data/owner-account-id	The ID of the Alibaba Cloud account to which the instance belongs.	1609****
/meta-data/public-keys/	The public keys of the instance.	skp-bp1brtqj5sw1vq****/
/meta-data/region-id	The region ID of the instance.	cn-hangzhou
/meta-data/zone-id	The zone ID of the instance.	cn-hangzhou-i
/meta-data/serial-number	The serial number of the instance.	4acd2b47-b328-4762-852f-998****
/meta-data/source-address	The address of the YUM or APT image repository. The package management software of a Linux instance can obtain updates from the image repository.	http://mirrors.cloud.aliyuncs.com
/meta-data/kms-server	The Key Management Service (KMS) server that activates the Windows instance.	kms.cloud.aliyuncs.com
/meta-data/wsus-server/wu-server	The server that updates the Windows instance.	http://update.cloud.aliyuncs.com

Item	Description	Example
/meta-data/wsus-server/wu-status-server	The server that monitors the update status of the Windows instance.	http://update.cloud.aliyuncs.com
/meta-data/vpc-id	The ID of the VPC to which the instance belongs.	vpc-bp1e0g399hkd7c8q****
/meta-data/vpc-cidr-block	The CIDR block of the VPC to which the instance belongs.	192.168.XX.XX/16
/meta-data/vswitch-cidr-block	The CIDR block of the vSwitch to which the instance is connected.	192.168.XX.XX/24
/meta-data/vswitch-id	The ID of the vSwitch to which the instance is connected.	vsw-bp1ygryo03m39xhsy****
• /meta-data/ram/security-credentials/[role-name] • /meta-data/ram/security-credentials/	The temporary Security Token Service (STS) credentials generated for the Resource Access Management (RAM) role of the instance. You can obtain the STS credentials only after the instance assumes a RAM role. Replace [role-name] with the name of the RAM role. If the [role-name] parameter is not specified, the name of the RAM role is returned. **Note** A new STS credential is available 30 minutes prior to the expiration time of the previous one. During these 30 minutes, both STS credentials are valid.	AliyunECSImageExportDefaultRole
/meta-data/instance/spot/termination-time	The stop and release time specified in the operating system of a preemptible instance. The time is in the yyyy-MM-ddThh:mm:ssZ format. The time is displayed in UTC.	2020-04-07T17:03:00Z
/meta-data/instance/virtualization-solution	The ECS virtualization solution. Virt 1.0 and Virt 2.0 are supported.	ECS Virt
/meta-data/instance/virtualization-solution-version	The version of the ECS virtualization solution.	2.0
/maintenance/active-system-events	The active system events of the instance.	None

Dynamic metadata items

Item	Description	Example
/dynamic/instance-identity/document	The instance identity document that provides the identity information of the instance. This document contains instance information such as the ID and IP addresses of the instance.	{"zone-id":"cn-hangzhou-i","serial-number":"4acd2b47-b328-4762-852f-99****","instance-id":"i-bp13znx0m0me8cq****","region-id":"cn-hangzhou","private-ipv4":"192.168.XX.XX","owner-account-id":"1609****","mac":"00:16:3e:0f:XX:XX","image-id":"aliyun_3_x64_20G_alibase_20210425.vhd","instance-type":"ecs.g6e.large"}
/dynamic/instance-identity/pkcs7	The instance identity signature that is used to verify the authenticity and content of the instance identity document.	MIIDJwYJKoZIhvcNAQcCoIIDGDCCAxQCAQExCzAJBgUrDgMCGgUAMIIBYQYJKoZIhvcNAQcBoIIBUgSCAU57InpvbmUtaWQiOiJjbi1oYW5nemhvdS1oIiwic2VyaWFsLW****

10.2.3. View instance metadata

This topic describes the differences between instance metadata viewed in normal mode and in security hardening mode. This topic also demonstrates how to view the metadata of an Elastic Compute Service (ECS) instance.

Prerequisites

- The instance resides in a virtual private cloud (VPC).
- You are connected to the instance. For more information, see Guidelines on instance connection.

Context

You can view instance metadata by using an endpoint in the `http://100.100.100.200/latest/[metadata]` format. Replace *[metadata]* with the instance metadata item. For more information, see Instance metadata items.

You can access the endpoint in normal or security hardening mode. The following table compares the two modes.

Comparison item	Normal mode	Security hardening mode
Interaction mode	Interacts with requests and responses.	Interacts in sessions.
Security verification	Verifies source IP addresses within the same VPC.	Verifies tokens for authentication.
Access method	Uses cURL commands to access the endpoint.	Uses cURL commands to access the endpoint. Requests must include token headers.

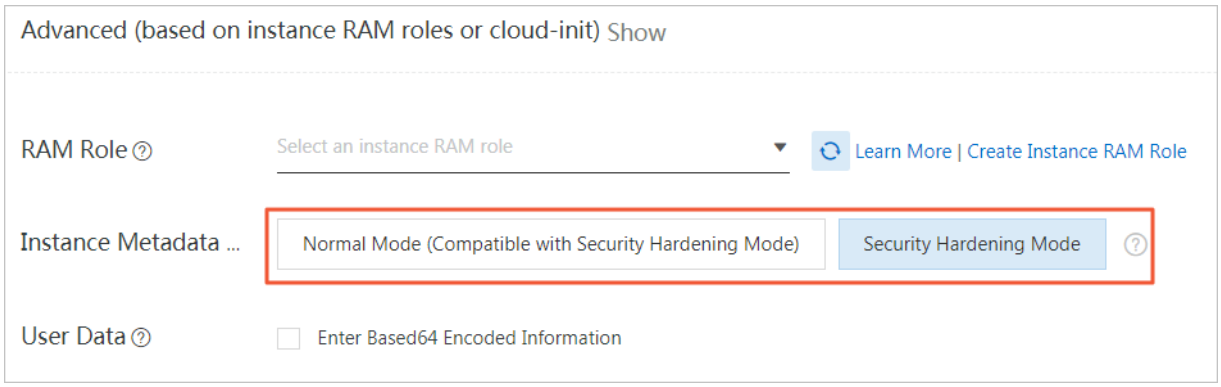
In normal mode, a new connection is established with each request to view instance metadata, and the connection is immediately released after the request is complete. This mode uses a simple verification method. If the instance metadata server is attacked and sensitive data such as Resource Access Management (RAM) roles leaks, your data and assets are at risk.

A server-side request forgery (SSRF) is an attack in which an attacker exploits vulnerabilities in a server to send forged resource requests to the server and access resources located within the same internal network. When a request for instance metadata is received, the instance metadata server shares the requested metadata in URLs. These URLs are vulnerable to tampering and may be used to attack internal systems that are inaccessible to external networks. In security hardening mode, instance metadata is restricted and can be viewed only by using token-based authentication. The security hardening mode provides better protection against SSRF attacks than the normal mode. In scenarios such as self-managed network firewall applications, self-managed reverse proxy applications, and self-managed web applications that provide transcoding and download services, we recommend that you view instance metadata in security hardening mode to prevent SSRF attacks and improve the security of applications.

You must specify an instance metadata access mode when you create an instance in the ECS console.

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

> Note



View instance metadata in normal mode

The following section provides examples of the shell commands that you can run to view the metadata of Linux instances:

- View the root directory of instance metadata:

```
curl http://100.100.100.200/latest/meta-data
```

- View the instance ID:

```
curl http://100.100.100.200/latest/meta-data/instance-id
```

- View the active system events:

```
curl http://100.100.100.200/latest/maintenance/active-system-events
```

- View the instance identity document:

```
curl http://100.100.100.200/latest/dynamic/instance-identity/document
```

- View the user data of the instance:

```
curl http://100.100.100.200/latest/user-data
```

The following section provides examples of the PowerShell commands that you can run to view the metadata of Windows instances:

- View the root directory of instance metadata:

```
Invoke-RestMethod http://100.100.100.200/latest/meta-data
```

- View the instance ID:

```
Invoke-RestMethod http://100.100.100.200/latest/meta-data/instance-id
```

- View the active system events:

```
Invoke-RestMethod http://100.100.100.200/latest/maintenance/active-system-events
```

- View the instance identity document:

```
Invoke-RestMethod http://100.100.100.200/latest/dynamic/instance-identity/document
```

- View the user data of the instance:

```
Invoke-RestMethod http://100.100.100.200/latest/user-data
```

View instance metadata in security hardening mode

In security hardening mode, you can establish a session between the ECS instance and the instance metadata server. When you attempt to view instance metadata, the instance metadata server authenticates your identity based on a token. When the token expires, the instance metadata server closes the session and deletes the token. The following limits apply to tokens:

- Each token can be used only for a single ECS instance. If you attempt to use the token of one instance to access a different instance, you are denied access.
- Each token must have a validity period that ranges from 1 to 21,600 seconds (6 hours). Tokens can be repeatedly used until they expire. This helps implement a balance between security and user experience.
- Proxy access is not supported. If a request used to create a token contains the `X-Forwarded-For` header, the instance metadata server refuses to issue the token.
- An unlimited number of tokens can be issued to each instance.

Perform the following steps to view instance metadata in security hardening mode:

1. Use the PUT method to initiate a request to create a token. You must specify the token validity period in the header in the following format: `X-aliyun-ecs-metadata-token-ttl-seconds:<Token validity period>`.
2. The instance metadata server issues the token.
3. Enter the endpoint of the instance metadata server and the token header. Enter the token header in the following format: `X-aliyun-ecs-metadata-token: $TOKEN`.
4. After successful authentication, the instance metadata server returns the requested instance metadata.

The following section provides examples of the commands that you can run to view instance metadata:

- Linux instance:

```
TOKEN=`curl -X PUT "http://100.100.100.200/latest/api/token" -H "X-aliyun-ecs-metadata-token-ttl-seconds: 21600"` \
&& curl -H "X-aliyun-ecs-metadata-token: $TOKEN"  http://100.100.100.200/latest/meta-data/instance-id
```

- Windows instance:

```
$token = Invoke-RestMethod -Headers @{"X-aliyun-ecs-metadata-token-ttl-seconds" = "21600"} -Method PUT -Uri http://100.100.100.200/latest/api/token
Invoke-RestMethod -Headers @{"X-aliyun-ecs-metadata-token" = $token} -Method GET -Uri http://100.100.100.200/latest/meta-data/instance-id
```

The preceding sample commands involve the following steps:

- Use the PUT method to create a token that has a validity period of 21,600 seconds (6 hours).
- Use the TOKEN variable to store the token.
- View the instance ID in the instance metadata and include the $TOKEN variable in the request.

Tokens can be repeatedly used until they expire. The following section provides examples of the commands that you can run to use an existing token:

- Linux instance:

```
curl -H "X-aliyun-ecs-metadata-token: $TOKEN"  http://100.100.100.200/latest/meta-data/instance-id
```

- Windows instance:

```
Invoke-RestMethod -Headers @{"X-aliyun-ecs-metadata-token" = $token} -Method GET -Uri http://100.100.100.200/latest/meta-data/instance-id
```

Error examples:

- The validity period is 21,700 seconds, which exceeds the maximum allowed length.

```
curl -X PUT "http://100.100.100.200/latest/api/token" -H "X-aliyun-ecs-metadata-token-ttl-seconds: 21700"
```

- The request used to create a token contains the X-Forwarded-For header.

```
curl -X PUT "http://100.100.100.200/latest/api/token" -H "X-Forwarded-For: www.ba****.com"
```

- The specified token to use to view the instance metadata is invalid.

```
curl -H "X-aliyun-ecs-metadata-token: aaa" -v http://100.100.100.200/latest/meta-data/
```

10.2.4. Use instance identities

This topic describes instance identities and how to use instance identities. This topic also provides examples on how to use instance identities with custom parameters specified and unspecified.

Context

Instance identities are part of the metadata of instances and can be used to identify and differentiate instances. They provide an important trust foundation to control application permissions and activate software. Instance identities are generated in real time and dynamically change with instances.

Each instance identity consists of an instance identity document (document) and an instance identity signature (signature).

The instance identity document is used to describe information of an instance and contains instance properties described in the following table.

Property	Description	Changeable
account-id	The ID of the Alibaba Cloud account to which the instance belongs.	No
instance-id	The ID of the instance.	No
mac	The media access control (MAC) address of the primary elastic network interface (ENI) of the instance.	No
region-id	The ID of the region where the instance resides.	No
serial-number	The serial number of the instance.	No
zone-id	The ID of the zone where the instance resides.	No
instance-type	The instance type.	Yes. This property changes when the instance type of the instance is changed. For more information, see Overview of instance upgrade and downgrade.
image-id	The ID of the image used by the instance.	Yes. This property changes when the system disk of the instance is replaced. For more information, see Replace the operating system of an instance by using a public image.

Property	Description	Changeable
private-ip	The private IP address of the instance.	Yes. This property changes when the private IP address of the instance in a virtual private cloud (VPC) is changed. For more information, see Modify a private IP address.

The instance identity signature is encrypted by using the PKCS #7 standard and is secure and reliable.

You can specify the audience parameter in the instance identity signature. The value of the audience parameter can be a random string, a timestamp, regularly changing data, or data generated by a specific algorithm. After the audience parameter is specified, it is difficult for other users to guess the value of the audience parameter even if they have obtained information about the identity document and the identity signature. This effectively prevents fraudulent use of signatures.

If you specify the audience parameter, you must simultaneously set the instance identity document and signature. For example, if you specify the audience parameter when you obtain the identity signature, you must add the audience value to the end of the dynamically obtained instance identity document in the following format before you verify the signature by using OpenSSL: `"audience":"Value of audience"`. Separate multiple values with commas (,).

In the following scenarios, you can use instance identities (`instance-identity`) for authentication, authorization, or identifying runtime environment.

- Software is traditionally activated by using a single serial number for a single device. This practice is not suitable for using software on the cloud because software is used at varying points in time and in different scenarios. You can use instance identities for user authorization when you publish application software in Alibaba Cloud Marketplace. For more information, see Example 1: Use instance identities without specifying the audience parameter.
- When you write sensitive data to an instance, you can use instance identities to ensure that you are writing the sensitive data to the exact instance that you want to use.
- Scenarios where you want to confirm the source of the instance.

Use instance identities

OpenSSL is required if you want to use instance identities. If you do not have OpenSSL configured in your instance, you must go to the OpenSSL official website to download and install OpenSSL. The following example demonstrates how to use instance identities on a Linux instance that runs CentOS 7.4.

1. Connect to the Linux instance.
2. Run the following command to obtain the instance identity document:

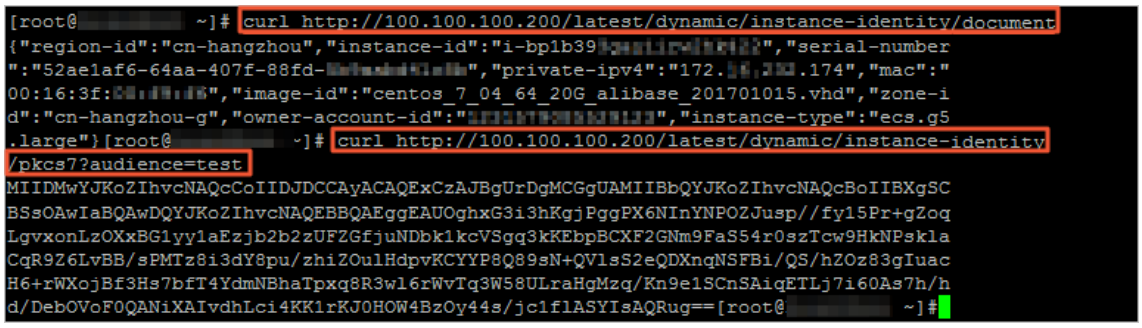
   ```
   curl http://100.100.100.200/latest/dynamic/instance-identity/document
   ```

3. Use one of the following methods to obtain the instance identity signature:
   - Do not specify the audience parameter:

     ```
     curl http://100.100.100.200/latest/dynamic/instance-identity/pkcs7
     ```

- Specify the audience parameter:

```
curl http://100.100.100.200/latest/dynamic/instance-identity/pkcs7?audience=XXXX
```



4. Verify the instance identity by using OpenSSL.

```
openssl smime -verify -in $signature -inform PEM -content $DOCUMENT -certfile AliyunPubkey -noverify > /dev/null
```

The following section describes the parameters in the preceding command:

- $signature specifies the identity signature that you obtained.
- $DOCUMENT specifies the identity document that you obtained.

  (Optional) If you specified the audience parameter in Step 3, you must add the audience value to the end of the dynamically obtained instance identity document in the following format: `"audience":"Value of audience"`. Separate multiple values with commas (,).
- AliyunPubkey specifies the Alibaba Cloud public certificate.

The following code demonstrates the Alibaba Cloud public certificate:

```
-----BEGIN CERTIFICATE-----
MIIDdzCCAl+gAwIBAgIEZmbRhzANBgkqhkiG9w0BAQsFADBsMRAwDgYDVQQGEwdV
bmtub3duMRAwDgYDVQQIEwdVbmtub3duMRAwDgYDVQQHEwdVbmtub3duMRAwDgYD
VQQKEwdVbmtub3duMRAwDgYDVQQLEwdVbmtub3duMRAwDgYDVQQDEwdVbmtub3du
MB4XDTE4MDIyMzAxMjkzOFoXDTM4MDIxODAxMjkzOFowbDEQMA4GA1UEBhMHVW5r
bm93bjEQMA4GA1UECBMHVW5rbm93bjEQMA4GA1UEBxMHVW5rbm93bjEQMA4GA1UE
ChMHVW5rbm93bjEQMA4GA1UECxMHVW5rbm93bjEQMA4GA1UEAxMHVW5rbm93bjCC
ASIwDQYJKoZIhvcNAQEBBQADggEPADCCAQoCggEBAIJwy5sbZDiNyX4mvdP32pqM
YMK4k7+5lRnVR2Fky/5uwyGSPbddNXaXzwEm+u4wIsJiaAN3OZgJpYIoCGik+9lG
5gVAIr0+/3rZ61IbeVE+vDenDd8g/m/YIdYBfC2IbzgS9EVGAf/gJdtDODXrDfQj
Fk2rQsvpftVOUs3Vpl9O+jeCQLoRbZYm0c5v7jP/L2lK0MjhiywPF2kpDeisMtnD
/ArkSPIlg1qVYm3F19v3pa6ZioM2hnwXg5DibYlgVvsIBGhvYqdQ1KosNVcVGGQa
HCUuVGdS7vHJYp3byH0vQYYygzxUJT2TqvK7pD57eYMN5drc7e19oyRQvbPQ3kkC
AwEAAaMhMB8wHQYDVR0OBBYEFAwwrnHlRgFvPGo+UD5zS1xAkC91MA0GCSqGSIb3
DQEBCwUAA4IBAQBBLhDRgezd/OOppuYEVNB9+XiJ9dNmcuHUhjNTnjiKQWVk/YDA
v+T2V3t9yl8L8o61tRIVKQ++lDhjlVmur/mbBN25/UNRpJllfpUH6oOaqvQAze4a
nRgyTnBwVBZkdJ0d1sivL9NZ4pKelJF3Ylw6rp0YMqV+cwkt/vRtzRJ31ZEeBhs7
vKh7F6BiGCHL5ZAwEUYe8O3akQwjgrMUcfuiFs4/sAeDMnmgN6Uq8DFEBXDpAxVN
sV/6Hockdfinx85RV2AUwJGfClcVcu4hMhOvKROpcH27xu9bBIeMuY0vvzP2VyOm
DoJeqU7qZjyCaUBkPimsz/1eRod6d4P5qxTj
-----END CERTIFICATE-----
```

Example 1: Use instance identities without specifying the audience parameter

The following example demonstrates how to use instance identities as an application software seller if an image is published to Alibaba Cloud Marketplace.

1. Connect to the instance.
2. Verify whether the image used by the instance is from Alibaba Cloud Marketplace.

   You can check the product-code and charge-type items in the instance metadata. The product-code item indicates the product code of the Alibaba Cloud Marketplace image, and the charge-type item indicates the billing method of the Alibaba Cloud Marketplace image. For more information, see Overview of ECS instance metadata.

   ```
   curl http://100.100.100.200/latest/meta-data/image/market-place/product-code
   curl http://100.100.100.200/latest/meta-data/image/market-place/charge-type
   ```

3. Create a temporary file named *cert.cer* in the current directory and save the Alibaba Cloud public certificate to the file.
4. Identify the instance.

   Example script:

   ```
   #!/usr/bin/bash
   function verify_signature_without_audience(){
   curl 100.100.100.200/latest/dynamic/instance-identity/document > document
   echo "-----BEGIN CERTIFICATE-----" > signature
   curl 100.100.100.200/latest/dynamic/instance-identity/pkcs7 >> signature
   echo "" >> signature
   echo "-----END CERTIFICATE-----" >> signature
   openssl smime -verify -in signature -inform PEM -content document -certfile cert.cer -n
   overify > /dev/null
   }
   verify_signature_without_audience
   ```

5. If `Verification successful` is returned, you have permissions to use the application software.

Example 2: Use instance identities while specifying the audience parameter

The following example demonstrates how to use instance identities as an application software seller if an image is published to Alibaba Cloud Marketplace. You can specify the audience parameter to identify the instance and implement policy control at the application server to allow authenticated users to use the software. This can protect the authorization code (license) against unauthorized use. The value of the audience parameter can be a random string, a timestamp, regularly changing data, or data generated by a specific algorithm.

1. Connect to the instance.
2. Verify whether the image used by the instance is from Alibaba Cloud Marketplace.

   You can check the product-code and charge-type items in the instance metadata. The product-code item indicates the product code of the Alibaba Cloud Marketplace image, and the charge-type item indicates the billing method of the Alibaba Cloud Marketplace image.

```
curl http://100.100.100.200/latest/meta-data/image/market-place/product-code
curl http://100.100.100.200/latest/meta-data/image/market-place/charge-type
```

3. Create a temporary file named *cert.cer* in the current directory and save the Alibaba Cloud public certificate to the file.
4. Identify the instance.

   Example script:

```
#!/usr/bin/bash
function verify_signature_with_specified_audience(){
audience='your audience' # Specify the audience parameter.
document=$(curl 100.100.100.200/latest/dynamic/instance-identity/document)
audience_json=',"audience":'"'${audience}'"}'
echo -n ${document%?}${audience_json} > document
echo "-----BEGIN CERTIFICATE-----" > signature
curl 100.100.100.200/latest/dynamic/instance-identity/pkcs7?audience=${audience} >> sig
nature
echo "" >> signature
echo "-----END CERTIFICATE-----" >> signature
openssl smime -verify -in signature -inform PEM -content document -certfile cert.cer -n
overify > /dev/null
}
verify_signature_with_specified_audience
```

5. If `Verification successful` is returned, you have permissions to use the application software.

10.3. Manage instance user data

10.3.1. Overview of ECS instance user data

The user data of Elastic Compute Service (ECS) instances can be used to manage startups of the instances or pass data into the instances.

Introduction

Both Linux and Windows instances support the user data feature. User data can be used in the following ways:

- User data can be run as scripts on instance startup to automate instance configurations such as automatically obtaining software resource packages, enabling services, printing logs, installing dependencies, and initializing web environments.
- User data can be used as common data and passed into instances for use.

You can prepare user data by using different types of scripts. After you prepare the user data, you can pass it into an instance by entering the script content when you create the instance. For more information, see Manage user data of Linux instances and Manage user data of Windows instances.

You can view the user data that has been passed into an instance by using the metadata of the instance. For more information, see View instance metadata.

Limits

- The user data feature is supported only for instances that reside in virtual private clouds (VPCs).

- The instances must be created from the following public images or custom images derived from public images:
  - Alibaba Cloud Linux, CentOS, Ubuntu, SUSE Linux Enterprise, OpenSUSE, and Debian
  - Windows Server 2008 R2 and later
- The user data feature is supported for all available instance types. For retired instance types, the user data feature is supported only for I/O-optimized instances. For more information, see Retired instance types.
- The user data that you want to run must be encoded in Base64. The size of the user data cannot exceed 16 KB before it is encoded.

> **Note** You can enter the user data that has not been encoded in Base64 in the console. The console automatically encodes the user data in Base64. If you do not want to enter the user data in the console, you must encode it in Base64 on your own.

10.3.2. Manage the user data of Linux instances

This topic describes how to prepare user data for Linux instances and how to pass in user data and verify its running results.

Prerequisites

If you want to modify the user data of an instance, the instance must be in the **Stopped** state.

Context

The user data feature for Linux instances uses the open source cloud-init architecture. After you pass user data into an instance in the Elastic Compute Service (ECS) console or by calling an API operation, you can access the user data by using instance metadata. Metadata is also used by cloud-init to automatically configure Linux instances. When an instance starts, the system uses the administrator or root permissions to run user data.

The following limits apply to user data:

- The user data feature is supported only for instances that reside in virtual private clouds (VPCs).
- The instances must be created from the following public images or custom images derived from public images:
  - Alibaba Cloud Linux, CentOS, Ubuntu, SUSE Linux Enterprise, OpenSUSE, and Debian
  - Windows Server 2008 R2 and later
- The user data feature is supported for all available instance types. For retired instance types, the user data feature is supported only for I/O-optimized instances. For more information, see Retired instance types.
- The user data that you want to run must be encoded in Base64. The size of the user data cannot exceed 16 KB before it is encoded.

> **Note** You can enter the user data that has not been encoded in Base64 in the console. The console automatically encodes the user data in Base64. If you do not want to enter the user data in the console, you must encode it in Base64 on your own.

Procedure

1. Prepare user data.

   You can run various scripts to prepare the user data of Linux instances. For more information about the characteristics of different scripts with examples, see the following sections:

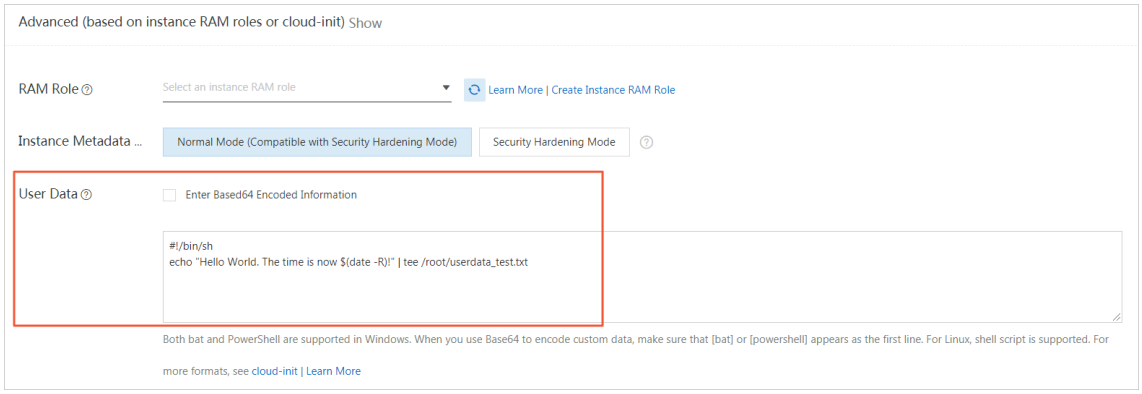
   - User-data scripts
   - Cloud-config data
   - Include files
   - Gzip compressed content
   - Upstart Job

   > **Note** If you want to use include files or Gzip compressed content in your user data scripts, you must upload script files to available storage services, obtain the links of the script files, and set a validity period for the links. We recommend that you use Alibaba Cloud Object Storage Service (OSS). For more information, see Upload objects and Configure lifecycle rules. You can also learn more about the ways to prepare user data from the cloud-init documentation. For more information, see User-Data Formats.

2. Pass the user data into the instance.

   - You can pass in the user data when you create an instance. In the **System Configurations (Optional)** step, click **Advanced** to show the parameters and enter the user data in the **User Data** field. If the user data is Base64 encoded, select **Enter Base64 Encoded Information**.

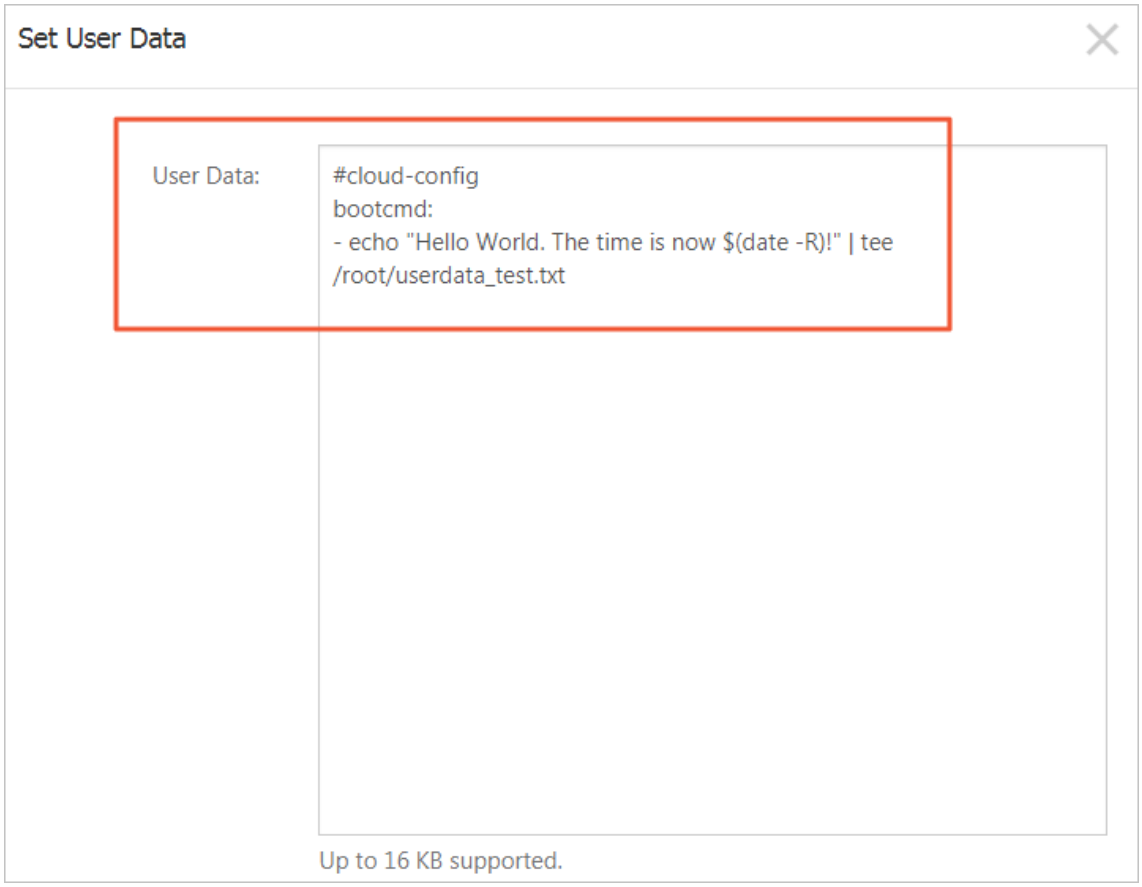
     The following figure shows an example of how to write the system time to a specified file when the instance starts for the first time.

   - You can modify the user data of an existing instance. On the **Instances** page, find the instance for which you want to modify the user data and choose **More > Instance Settings > Set User Data**. In the Set User Data dialog box, enter the new user data in the **User Data** field.

     > **Note** If you want to start a pay-as-you-go instance immediately after you modify the user data of the instance, we recommend that you set the stop mode of the instance to **Standard Mode**.

The following figure provides an example of how to write the system time to a specified file on each instance startup.



After the user data is modified for a Linux instance, the script type and module type determine whether new user data is run. Examples:

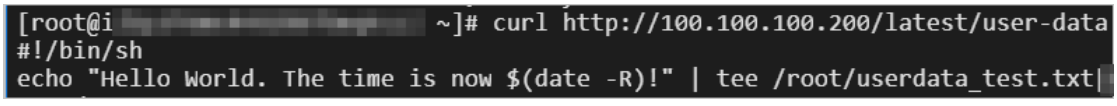
- User-data scripts are not run.
- Cloud-config data is not run if modules such as Byobu and Set Passwords are configured.
- Cloud-config data is run if modules such as Bootcmd, Update Etc Hosts, and Yum Add Repo are configured.

For information about the characteristics of the modules, see the module frequency line of each module in Modules

3. View the content passed into the instance and the result of running the script.
   i. Connect to the instance. For more information, see Connection methods.
   ii. View the content by using instance metadata.

```
curl http://100.100.100.200/latest/user-data
```

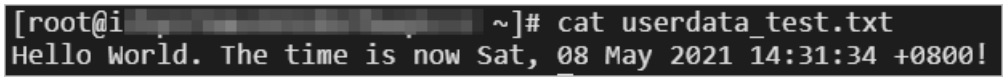
In this example, the user data is passed into the instance in Step 2. If the user data is passed into the instance, the user data is included in the command output, as shown in the following figure.



iii. View the running results.

```
cat userdata_test.txt
```

The result of running a script is related to its content. The following figure provides an example of the result of writing the system time to a specified file.



User-data scripts

User-data scripts are directly executed as shell scripts after they are passed into Linux instances. User-data scripts have the following characteristics:

- The first line starts with `#!` .
- User-data scripts are run once only when instances start for the first time.

Example:

```
#!/bin/sh
echo "Hello World. The time is now $(date -R)!" | tee /root/userdata_test.txt
```

The example user-data script can be run to write the system time to the *userdata_test.txt* file when the instance starts for the first time.

Cloud-config data

Cloud-config data is a convenient way to configure services such as YUM repository update, SSH key import, and instance dependency installation. Cloud-config data has the following characteristics:

- The first line starts with `#cloud-config` , and the header cannot have spaces.
- The script must follow the YAML syntax.
- The frequency at which the user data is run varies based on the modules that are configured. For example, if you configure the Apt Configure module, the user data is run only once for each instance. If you configure the Bootcmd module, the user data is run each time the instance starts.

Example:

```
#cloud-config
apt:
 primary:
  - arches: [default]
    uri: http://us.archive.ubuntu.com/ubuntu/
    bootcmd:
     - echo "Hello World. The time is now $(date -R)!" | tee /root/userdata_test.txt
```

The cloud-config data in this example can be run to modify the default software source and write the latest system time to the *userdata_test.txt* file each time the instance starts.

Include files

An include file contains one or more script links, one per line. When the instance starts, cloud-init reads each script link and its content. If an error occurs while a script is being read, the remaining scripts are not read. Include files have the following characteristics:

- The first line starts with `#include`, and the header cannot have spaces.
- The size of each script cannot exceed 16 KB before it is encoded in Base64.
- The frequency at which the user data is run varies based on the types of scripts and modules.

Example:

```
#include
http://ecs-image-test.oss-cn-hangzhou.aliyuncs.com/userdata/myscript.sh
```

The include file in this example contains a script link. The running frequency is determined by the script type. For example, if the script is a user-data script, the script is run once only when the instance starts for the first time.

Gzip compressed content

If the size of your user-data script, cloud-config data, or include file may exceed 16 KB, you can use gzip compressed content in the `.gz` format to compress the script into a link. Then, you can pass in the link as an include file. cloud-init automatically decompresses the gzip compressed content. The result of running the decompressed content shows no difference from that of running a script that is directly passed in. Gzip compressed content has the following characteristics:

- The first line starts with `#include`, and the header cannot have spaces.
- The size of the gzip compressed content cannot exceed 16 KB before it is encoded in Base64.
- The frequency at which the user data is run varies based on the types of scripts and modules.

Example:

```
#include
http://ecs-image-test.oss-cn-hangzhou.aliyuncs.com/userdata/myscript.gz
```

The include file in this example contains a link to gzip compressed content. cloud-init reads the gzip compressed content and automatically decompresses and runs it. The running frequency is determined by the script type. For example, if the gzip compressed content is obtained by compressing a user-data script, the gzip compressed content is run once only when the instance starts for the first time.

Upstart Job

The content of upstart job scripts is placed into a file in the */etc/init* directory. Upstart job scripts have the following characteristics:

- The first line starts with `#upstart-job`, and the header cannot have spaces.
- Upstart job scripts are run each time the instance starts.

> **Note** To use upstart job scripts, you must install the upstart service for the instance. The upstart service is supported for instances that run the CentOS 6, Ubuntu 10, Ubuntu 12, Ubuntu 14, Debian 6, or Debian 7 operating system.

Example:

```
#upstart-job
description "upstart test"
start on runlevel [2345] #Starts at run levels 2, 3, 4, and 5.
stop on runlevel [!2345] #Stops at a run level that is not 2, 3, 4, or 5.
exec echo "Hello World. The time is now $(date -R)!" | tee /root/output.txt
```

Example 1: Use user-data scripts to customize YUM repositories and NTP and DNS services

When an instance starts, the system configures the default YUM repository and Network Time Protocol (NTP) and Domain Name System (DNS) services. You can use the user data of the instance to change the default YUM repository and NTP and DNS services that are configured. Take note of the following items:

- If you customize a YUM repository, Alibaba Cloud stops providing YUM repository support.
- If you customize the NTP service, Alibaba Cloud stops providing time synchronization services.

The following code provides an example of a user-data script that can be run on an instance that runs the CentOS 7.2 operating system:

```
#!/bin/sh
# Modify DNS
echo "nameserver 8.8.8.8" | tee /etc/resolv.conf
# Modify yum repo and update
rm -rf /etc/yum.repos.d/*
touch myrepo.repo
echo "[base]" | tee /etc/yum.repos.d/myrepo.repo
echo "name=myrepo" | tee -a /etc/yum.repos.d/myrepo.repo
echo "baseurl=http://mirror.centos.org/centos" | tee -a /etc/yum.repos.d/myrepo.repo
echo "gpgcheck=0" | tee -a /etc/yum.repos.d/myrepo.repo
echo "enabled=1" | tee -a /etc/yum.repos.d/myrepo.repo
yum update -y
# Modify NTP Server
echo "server ntp1.aliyun.com" | tee /etc/ntp.conf
systemctl restart ntpd.service
```

> **Note**
>
> - In the preceding example, the URL is for reference only. You can replace it to suit your needs.
> - You can also use cloud-config data to change the YUM repository. However, cloud-config data is not as flexible as user-data scripts and is not applicable to scenarios where Alibaba Cloud pre-configures some YUM repositories. We recommend that you use user-data scripts.

Pass in the user data when you create the instance. After the instance starts, log on to the instance to view the running result. Check whether the configurations of the YUM repository and NTP and DNS services are as expected. The following is an example script:

```
[root@iZbp1csxtw7jo9zp12s**** ~]# cat /etc/yum.repos.d/myrepo.repo
[base]
name=myrepo
baseurl=http://mirror.centos.org/centos
gpgcheck=0
enabled=1
[root@iZbp1csxtw7jo9zp12s**** ~]# cat /etc/resolv.conf
nameserver 8.8.8.8
[root@iZbp1csxtw7jo9zp12s**** ~]# ping www.baidu.com
PING www.a.shifen.com (14.215.XX.XX) 56(84) bytes of data.
64 bytes from 14.215.XX.XX (14.215.XX.XX): icmp_seq=1 ttl=52 time=26.3 ms
64 bytes from 14.215.XX.XX (14.215.XX.XX): icmp_seq=2 ttl=52 time=26.3 ms
64 bytes from 14.215.XX.XX (14.215.XX.XX): icmp_seq=3 ttl=52 time=26.2 ms
^Z
[2]+  Stopped                 ping www.baidu.com
[root@iZbp1csxtw7jo9zp12s**** ~]# cat /etc/ntp.conf
server ntp1.aliyun.com
[root@iZbp1csxtw7jo9zp12s**** ~]# systemctl status ntpd.service
● ntpd.service-Network Time Service
   Loaded: loaded (/usr/lib/systemd/system/ntpd.service; enabled; vendor preset: disabled)
   Active: active (running) since Mon 2021-09-06 14:53:19 CST; 13min ago
 Main PID: 5795 (ntpd)
   CGroup: /system.slice/ntpd.service
           └─5795 /usr/sbin/ntpd -u ntp:ntp -g
Sep 06 14:53:19 iZbp1cjdaurreftzgpgvqoZ ntpd[5795]: Listen and drop on 1 v6wildcard :: UDP
123
Sep 06 14:53:19 iZbp1cjdaurreftzgpgvqoZ ntpd[5795]: Listen normally on 2 lo 127.0.XX.XX UDP
123
Sep 06 14:53:19 iZbp1cjdaurreftzgpgvqoZ ntpd[5795]: Listen normally on 3 eth0 192.168.XX.XX
UDP 123
Sep 06 14:53:19 iZbp1cjdaurreftzgpgvqoZ ntpd[5795]: Listening on routing socket on fd #20 f
or interface updates
Sep 06 14:53:19 iZbp1cjdaurreftzgpgvqoZ ntpd[5795]: 0.0.XX.XX c016 06 restart
Sep 06 14:53:19 iZbp1cjdaurreftzgpgvqoZ ntpd[5795]: 0.0.XX.XX c012 02 freq_set kernel 0.000
PPM
Sep 06 14:53:19 iZbp1cjdaurreftzgpgvqoZ ntpd[5795]: 0.0.XX.XX c011 01 freq_not_set
Sep 06 14:56:34 iZbp1cjdaurreftzgpgvqoZ ntpd[5795]: 0.0.XX.XX c61c 0c clock_step +0.464773
s
Sep 06 14:56:35 iZbp1cjdaurreftzgpgvqoZ ntpd[5795]: 0.0.XX.XX c614 04 freq_mode
Sep 06 14:56:36 iZbp1cjdaurreftzgpgvqoZ ntpd[5795]: 0.0.XX.XX c618 08 no_sys_peer
```

Example 2: Use user-data scripts to customize the administrator account

By default, Linux instances use the root user as the administrator. You can use the user data of an instance to configure another user as the administrator.

The following code provides an example of a user-data script that can be run on an instance that runs the CentOS 7.2 operating system:

```
#!/bin/sh
useradd test
echo "test   ALL=(ALL)        NOPASSWD:ALL" | tee -a /etc/sudoers
mkdir /home/test/.ssh
touch /home/test/.ssh/authorized_keys
echo "ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABAQCRnnUveAis****" | tee -a /home/test/.ssh/author
ized_keys
```

Note Replace *ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABAQCRnnUveAis****** in the preceding example with your public key.

The sample user-data script can be run to obtain the following results:

- A user named test is created and used as the administrator account.
- The user can use only SSH key pairs to log on to the instance and cannot use passwords for logon.
- If the user wants to perform operations that require administrator permissions, the user needs only to run the **sudo** command, without the need to enter the password.

Pass in the user data when you create the instance. After the instance starts, log on to the instance by using the test user and the SSH key pair. An error is reported if you attempt to use the password for logon. After you connect to the instance, you can run the **sudo** command to perform operations that require administrator permissions, as shown below.

```
[test@iZbp1csxtw7jo9zp12s**** ~]$ cd /root
-bash: cd: /root: Permission denied
[test@iZbp1csxtw7jo9zp12s**** ~]$ sudo cd /root
[test@iZbp1csxtw7jo9zp12s**** ~]$
```

10.3.3. Manage the user data of Windows instances

This topic describes how to prepare user-data scripts for Windows instances and how to pass in user data and verify the result of running the user data.

Prerequisites

If you want to modify the user data of an instance, the instance is in the **Stopped** state.

Context

The user data feature enables Windows instances to run initialization scripts. When an instance starts, the system uses the administrator permissions to run the user data of the instance.

The following limits apply to user data:

- The user data feature is supported only for instances that reside in virtual private clouds (VPCs).
- The instances must be created from the following public images or custom images derived from public images:
  - Alibaba Cloud Linux, CentOS, Ubuntu, SUSE Linux Enterprise, OpenSUSE, and Debian
  - Windows Server 2008 R2 and later

- The user data feature is supported for all available instance types. For retired instance types, the user data feature is supported only for I/O-optimized instances. For more information, see Retired instance types.
- The user data that you want to run must be encoded in Base64. The size of the user data cannot exceed 16 KB before it is encoded.

> **Note** You can enter the user data that has not been encoded in Base64 in the console. The console automatically encodes the user data in Base64. If you do not want to enter the user data in the console, you must encode it in Base64 on your own.

Procedure

1. Prepare user data.

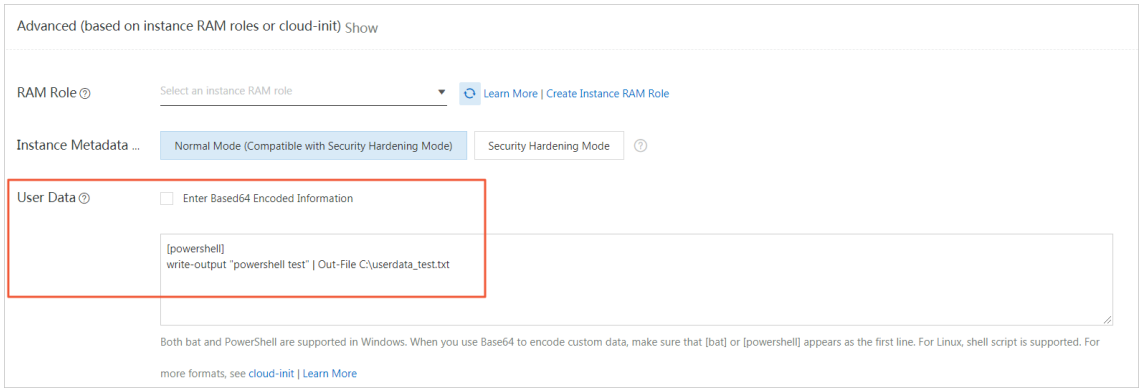
   You can run batch and PowerShell scripts to prepare user data of Windows instances. For more information about the characteristics of different scripts and their examples, see the following sections:

   - Bat
   - PowerShell

2. Pass the user data into the instance.

   - Pass in the user data when you create the instance: In the **System Configurations (Optional)** step, click **Advanced** to show the parameters and enter the user data in the **User Data** section. If the user data is encoded in Base64, select **Enter Base64 Encoded Information**.

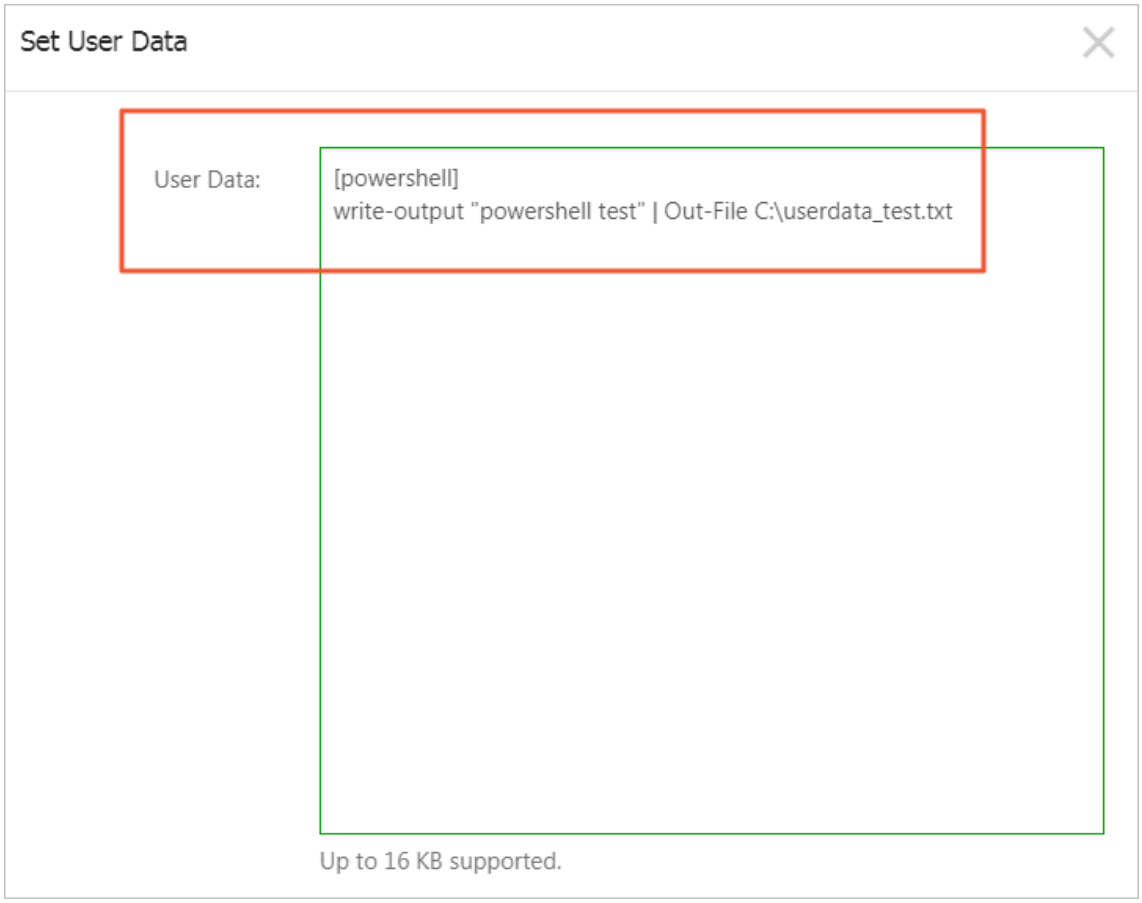
     The following figure shows an example of content written to a specified file.

   - Modify the user data of an existing instance: On the **Instances** page, find the instance for which you want to modify the user data and choose **More > Instance Settings > Set User Data**. In the Set User Data dialog box, enter new user data in the **User Data** section.

     > **Note** If you want to start a pay-as-you-go instance immediately after you modify the user data of the instance, we recommend that you set the stop mode of the instance to **Keep Instances and Continue Billing**.

The following figure shows an example of the content written to a specified file.



Note After the user data is modified for a Windows instance, the new user data is not run when the instance is started.

3. View the content passed into the instance and the script results.
   i. Connect to the instance. For more information, see Overview.
   ii. View the content by using the metadata of the instance.

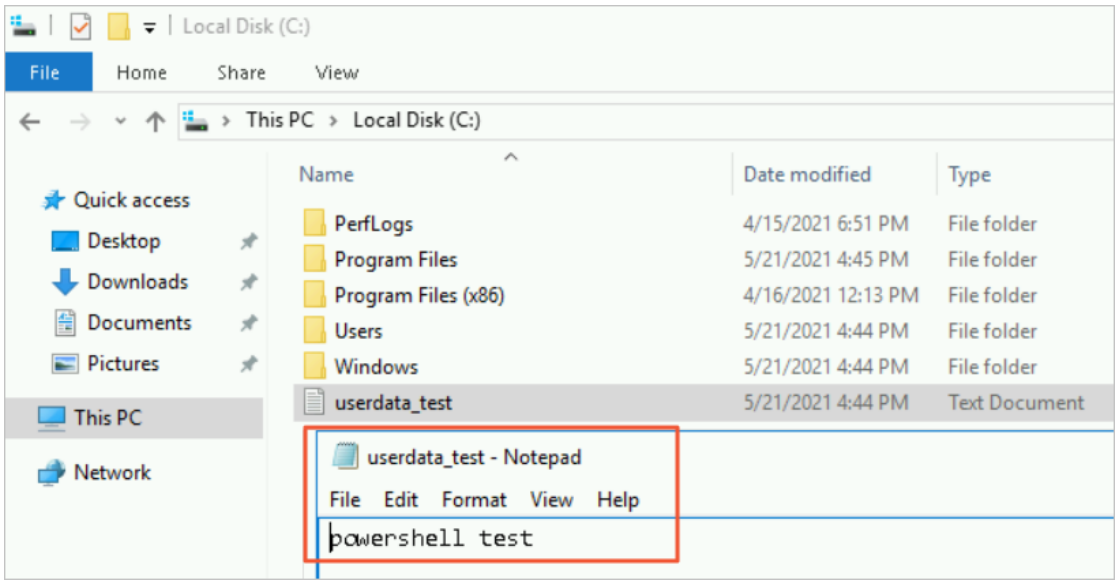
```
Invoke-RestMethod http://100.100.100.200/latest/user-data
```

In this example, the user data that is passed in in Step 2 is used as an example. If the user data is included in the output, the user data is passed in, as shown in the following figure.

```
PS C:\Users\Administrator> Invoke-RestMethod http://100.100.100.200/latest/user-data
[powershell]
write-output "powershell test" | Out-File C:\userdata_test.txt
```

iii. View the running results.

The result of running a script is related to its content. The following figure shows an example of the result of writing content to the specified file.



Bat

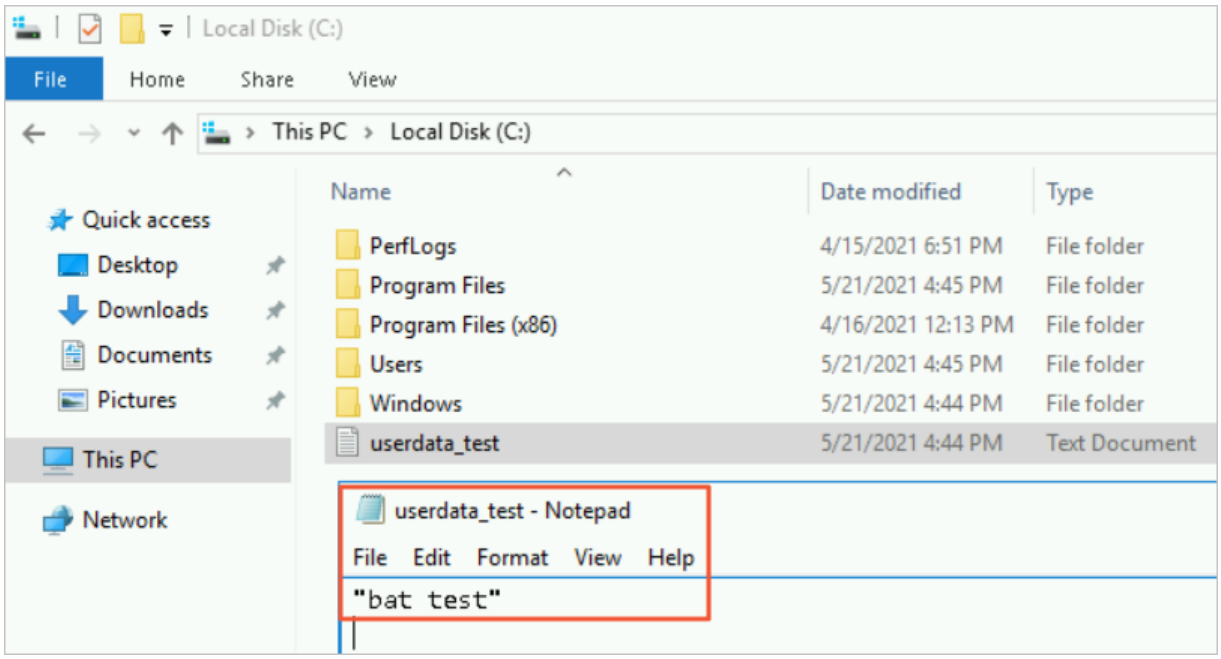
Batch scripts have the following characteristics:

- The first line starts with `[bat]`, and the header cannot have spaces.
- Only half-width letters can be entered, and no additional characters are allowed.

Example:

```
[bat]
echo "bat test" > C:\userdata_test.txt
```

The example batch script can be run to write `"bat test"` to the *userdata_test.txt* file when the instance starts for the first time, as shown in the following figure.



PowerShell

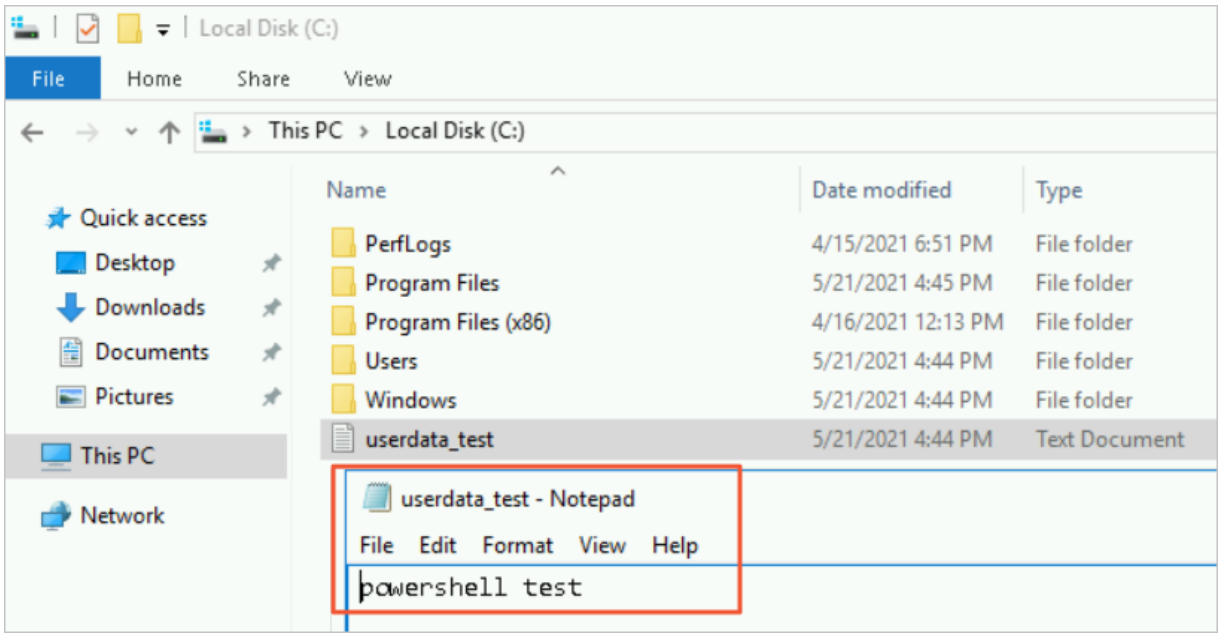
PowerShell scripts have the following characteristics:

- The first line starts with `[powershell]`, and the header cannot have spaces.
- Only half-width letters can be entered, and no additional characters are allowed.

Example:

```
[powershell]
write-output "powershell test" | Out-File C:\userdata_test.txt
```

The example PowerShell script can be run to write `powershell test` to the *userdata_test.txt* file when the instance starts for the first time, as shown in the following figure.



10.4. Build a confidential computing environment by using Enclave

This topic describes how to create a trusted isolation space by using the Enclave feature to protect your applications and data.

Context

Typically, data is classified into three states: data at rest, data in transit, and data in use. Data at rest and data in transit can be protected by using encryption, but you may encounter difficulties in ensuring security for data in use. Typically, confidential computing is used to protect data in use.

The Enclave feature provides a trusted isolation space inside Elastic Compute Service (ECS) instances to encapsulate the secure operations of legitimate software within an enclave. This ensures the confidentiality and integrity of your code and data against malware attacks.

> **Note** The Enclave feature is in invitational preview. If you want to use this feature, go to the Enclave product page.

The Enclave feature is applicable to business in industries such as finance, Internet, and healthcare that require strong protection for sensitive and confidential data.

How Enclave works

Computing resources (including vCPUs and memory) are split within an ECS instance (the primary VM) and an Enclave VM (EVM) is created as a trusted execution environment. The security of the EVM is ensured in the following aspects:

- The underlying virtualization technology provides security isolation. The EVM is isolated from the primary VM and other ECS instances.

- The EVM runs an independent, customized, and trusted operating system. The EVM has no persistent storage, interactive connections, or external network channels, and allows communication with the primary VM only by using a local secure channel (based on vsock) to ensure a minimal attack surface. You can put applications that involve confidential data into the EVM for running, and make secure calls to interact with the applications that run within the primary VM.

The security provided by the Enclave feature is implemented in multiple aspects. At the underlying layer, the third-generation SHENLONG architecture that uses Trusted Platform Module (TPM) or Trust Cryptography Module (TCM) chips provides vTPM or vTCM devices for the EVM to enhance security and trusted capabilities. At the upper layer, highly compatible SDKs are provided so that you can build an Enclave environment for use in a quick manner. To verify the trusted capabilities, you can verify the code running in the confidential execution environment, such as by using SDKs. Confidential applications can generate attestation materials (including the platform, application information, and signatures) at runtime, and then verify the attestation materials by using the remote attestation server (with reference to KMS). When the primary VM splits resources to the EVM and the EVM starts to run, the underlying layer performs resource access isolation to ensure that the primary VM cannot access these split vCPU or memory resources. This ensures the normal operation and privacy of the EVM.

Limits

- Only g7, c7, and r7 instance families support the Enclave feature.
- You can create only one enclave for each ECS instance.
- Before you use an enclave, you must reserve at least one processor core and a portion of the memory for the primary VM. The remaining processor and memory resources can be flexibly allocated to the enclave. If Hyper-Threading is enabled, two hyperthreads that belong to one physical core are reserved. Therefore, an ECS instance that has the Enclave feature enabled must have at least four vCPUs.

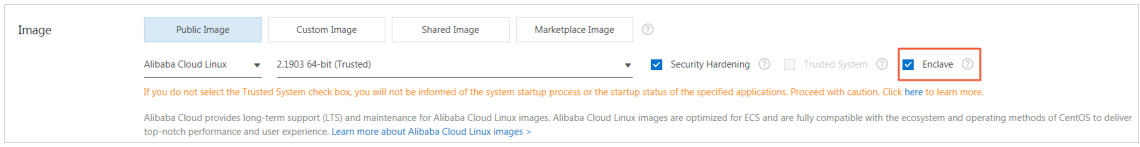
For information about other general limits, see 使用限制.

Use an enclave by means of a toolset

1. Install the Enclave Runtime toolset.

   The Enclave Runtime toolset manages the lifecycle of enclaves on the primary VM, including the startup and termination of enclaves. You can use one of the following methods to install the Enclave Runtime toolset:

   - When you create an ECS instance, select **Enclave**. The Enclave Runtime toolset is automatically installed.

     > **Note** When you create an ECS instance in an Enclave confidential computing environment by calling API operations, you can call only the RunInstances operation. You cannot set the confidential computing mode parameter (`SecurityOptions.ConfidentialComputingMode`) when you call the CreateInstance operation.

   - After the ECS instance is created, run the following commands to install the Enclave Runtime toolset on the primary VM:

```
sudo rpmkeys --import http://mirrors.aliyun.com/epel/RPM-GPG-KEY-EPEL-7
sudo yum install -y alinux-release-experimentals
sudo yum install -y https://enclave.oss-cn-hangzhou.aliyuncs.com/de-platform-runtime-0.1.0-1.2.al7.x86_64.rpm
```

After the Enclave Runtime toolset is installed, the local service attempts to automatically start the enclave. By default, the enclave image is stored in */usr/local/share/dragonfly/image.bin*. You can modify the */etc/enclave.conf* configuration file to change the storage path. The configuration file also provides additional configuration options, including the vCPU and memory resources allocated to the enclave.

2. After the Enclave Runtime toolset is manually installed for the first time, run the following command to download the enclave image and save it to your computer:

```
wget -O /usr/local/share/dragonfly/image.bin \
  https://enclave-cn-shenzhen.oss-cn-shenzhen.aliyuncs.com/download/linux/enclave_image/x86_64/0.1.0/image-0.1.0.bin
```

3. Run the **systemctl** commands to perform operations on the enclave:

```
systemctl status de_platform_service # View the running status.
systemctl start de_platform_service # Start the service.
systemctl restart de_platform_service # Restart the service.
systemctl stop de_platform_service # Stop the service.
```

Use an enclave by means of SDKs

Alibaba Cloud Enclave provides SDKs for you to develop your own applications on ECS instances that have Enclave enabled. Alibaba Cloud Enclave also provides a set of API definitions and code libraries that are compatible with SGX SDKs. If you already have an SGX application, you can run the application on a platform that has Enclave enabled only with a small amount of migration work.

1. Prepare the following Dockerfile file in the development environment:

```
FROM registry.cn-hangzhou.aliyuncs.com/alinux/aliyunlinux
RUN rpmkeys --import http://mirrors.aliyun.com/epel/RPM-GPG-KEY-EPEL-7 \
    && yum install -y alinux-release-experimentals \
    && yum install -y devtoolset-9 wget openssl-devel zlib-devel patch git cmake3 \
      https://enclave.oss-cn-hangzhou.aliyuncs.com/de-platform-runtime-0.1.0-1.2.al7.x86_64.rpm \
      https://enclave.oss-cn-hangzhou.aliyuncs.com/teesdk-0.1.0-1.1.al7.x86_64.rpm \
    && yum clean all -y \
    && wget -O /devtoolset9_enable.sh \
      https://enclave.oss-cn-hangzhou.aliyuncs.com/devtoolset9_enable.sh \
    && chmod +x /devtoolset9_enable.sh
WORKDIR /opt/app-root/src
ENTRYPOINT ["/devtoolset9_enable.sh"]
```

2. Run the following command to use Docker to create an enclave image:

```
docker build -t deenclave/sdk-builder .
```

After the command is run, an image named *deenclave/sdk-builder* is created. You can use this image to build DE Enclave applications. Alibaba Cloud provides the following SDK examples for your reference.

SDK example 1: Start a container to build an application

Alibaba Cloud provides an SDK sample program in */opt/alibaba/teesdk/desdk/examples/SampleMath*. This sample program receives two externally entered plane coordinate points, calculates the straight-line distance between the two points within the enclave, and sends the calculation results to the console. Perform the following operations:

1. Run the following command to start a container by using the created *deenclave/sdk-builder* image:

   ```
   docker run -it \
     -v /opt/alibaba/teesdk/desdk/examples/SampleMath:/opt/app-root/src:z \
     deenclave/sdk-builder
   ```

2. Run the following command to configure the environment variables within the started container instance:

   ```
   source /opt/alibaba/teesdk/desdk/environment
   ```

3. Run the **cmake** command to build the sample code within the started container:

   ```
   cmake3 -B build && \
   cmake3 --build build
   ```

   The application is located in */opt/app-root/src/build/SampleMath/App/app*, and the enclave is located in */opt/app-root/src/build/SampleMath/Enclave/enclave.signed.so*.

4. Upload the application to the primary VM.

5. Start the enclave on the primary VM, run the sample code SampleMath, and then check the execution result.

   ```
   [root@AliYun ~]# ./app
   A(3,4) -> B(1,8) -> 4.47214
   A(6,9) -> B(6,2) -> 7
   A(3,3) -> B(7,5) -> 4.47214
   ```

SDK example 2: Generate attestation materials through Attestation

Attestation is an authentication process that allows you to ensure that images, operating systems, and application code running in an enclave are not modified or tampered with. You can call API operations provided by the SDK in your Enclave application code to generate attestation materials and upload the attestation materials to the remote attestation server for verification. Then, the remote attestation server returns the verification results.

Alibaba Cloud provides the following sample programs. You can build and run these sample programs in the same way as you run SampleMath in SDK example 1.

- The sample program for generating attestation materials is located in */opt/alibaba/teesdk/desdk/examples/QuoteGenerationSample*.
- The sample program for verifying attestation materials is located in */opt/alibaba/teesdk/desdk/examples/QuoteVerificationSample*.

10.5. Replace UIO drivers with VFIO drivers

This topic describes how to replace a Userspace I/O (UIO) driver with a Virtual Function I/O (VFIO) driver to solve the runtime exceptions of Data Plane Development Kit (DPDK) for an Elastic Compute Service (ECS) instance.

Prerequisites

- Huge pages are configured for the instance.
- DPDK is installed on the instance.
- You are connected to the instance. For more information, see Guidelines on instance connection.

Context

If DPDK applications are deployed on your ECS instances of a sixth-generation (such as g6, c6, and r6) instance type or later, exceptions may occur when you run the applications. For example, the igb_uio port to which the network interface controllers (NICs) are bound may not be detected when you use Pktgen-DPDK to test the packet forwarding rate of an instance, and the following error message is reported:

```
 EAL: eal_parse_sysfs_value(): cannot open sysfs value
/sys/bus/pci/devices/0000:00:06.0/uio/uio0/portio/port0/start
```

You can replace UIO drivers with VFIO drivers to solve these problems. Operations performed on ECS bare metal instances are different from those performed on instances that are not ECS bare metal instances. For more information, see Operations performed on an instance that is not an ECS bare metal instance and Operations performed on an ECS bare metal instance.

Operations performed on an instance that is not an ECS bare metal instance

1. Run the following command to check the configurations of GRand Unified Bootloader (GRUB):

   ```
   cat /proc/cmdline
   ```

   Check whether the configurations of GRUB contain `intel_iommu=on` and do not contain `iommu=pt`.

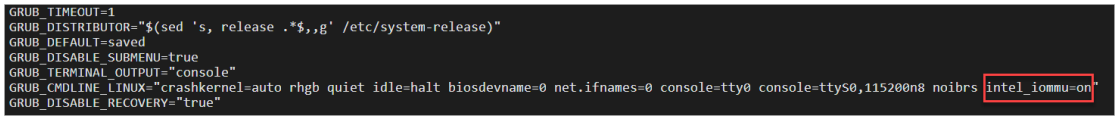
2. (Optional)If the configurations of GRUB do not contain `intel_iommu=on`, manually add intel_iommu=on.

   i. Run the following command to open the configuration file of GRUB:

   ```
   vim /etc/default/grub
   ```

   ii. Switch to the edit mode, add `intel_iommu=on` to the line of GRUB_CMDLINE_LINUX, and then save the configuration file.

   The following figure shows an example of the modified configuration file.

   iii. Run the following command to apply the modified configurations:

   ```
   grub2-mkconfig -o /boot/grub2/grub.cfg
   ```

   iv. Restart the instance and connect to the instance.

3. Run the following commands to install the VFIO and VFIO-PCI drivers:

```
modprobe vfio && \
modprobe vfio-pci
```

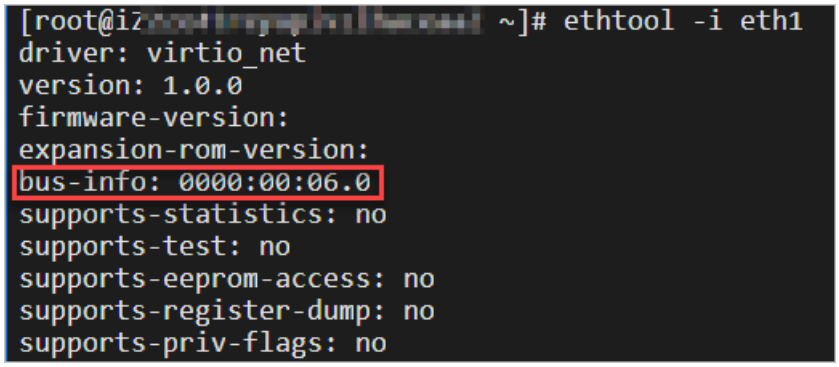
4. Run the following command to configure noiommu_mode:

```
echo 1 > /sys/module/vfio/parameters/enable_unsafe_noiommu_mode
```

5. Run the following command to view and record the bus-info value of the NIC bound to the instance:

```
ethtool -i ethX
```

Replace *ethX* with the ID of the NIC bound to the instance. In this example, eth1 is used.

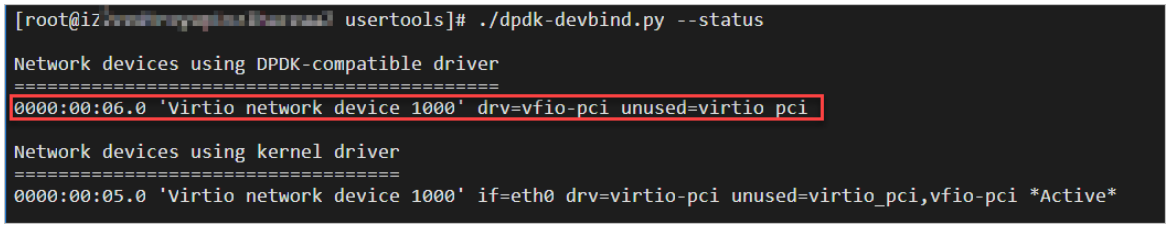


6. Switch to the *usertools* directory under the installation directory of DPDK and run the following command to bind the NIC to the VFIO-PCI driver:

```
./dpdk-devbind.py -b vfio-pci 0000:00:06.0
```

> **Note** In this example, `0000:00:06.0` is the bus-info value of eth1. In actual scenarios, replace it with the bus-info value of your NIC.

Run the `./dpdk-devbind.py --status` command to view the binding state of the NIC. The following command output indicates that the NIC whose bus-info value is `0000:00:06.0` has been bound to the VFIO-PIC driver.



7. Switch to the *build/app* directory under the installation directory of DPDK and run one of the following commands to start DPDK.
   - If the version of DPDK is 18.02 or later, run the following command:

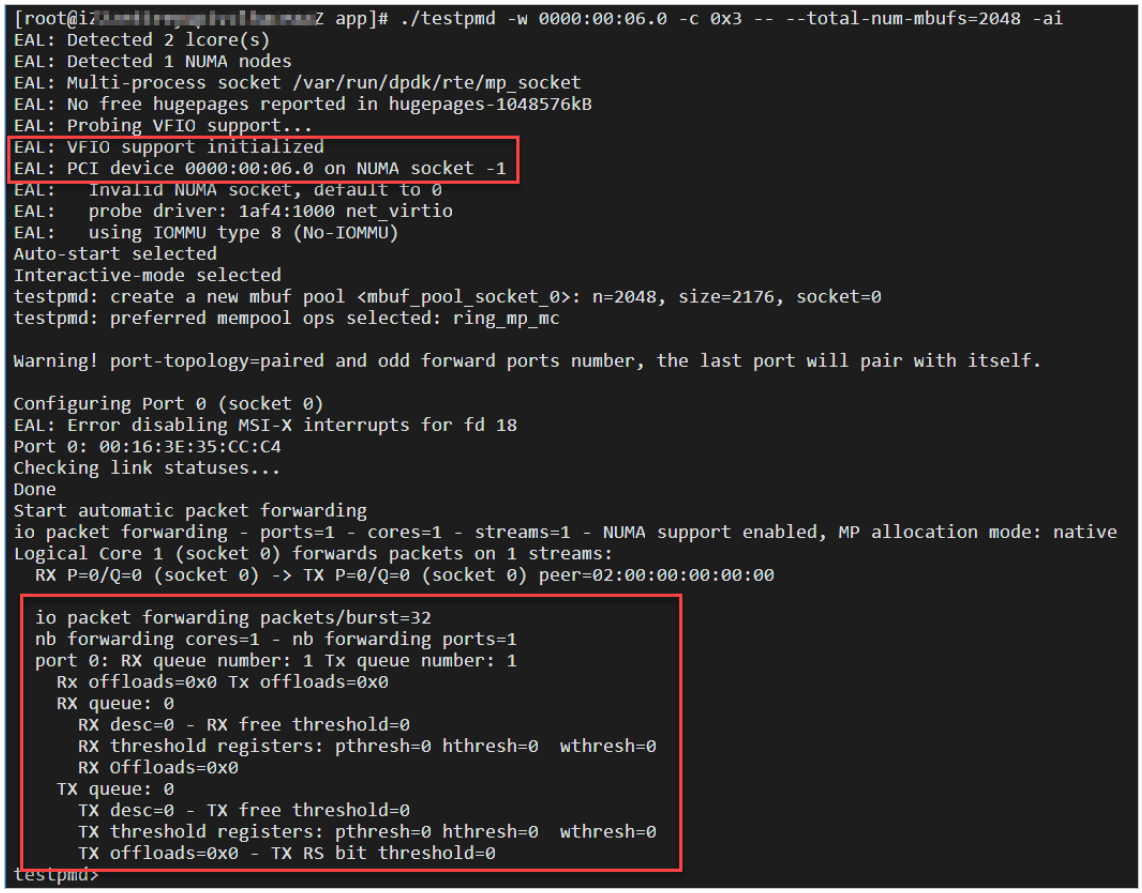
```
./testpmd -w 0000:00:06.0 -c 0x3 -- --total-num-mbufs=2048 -ai
```

   - If the version of DPDK is earlier than 18.02, run the following command:

```
./testpmd -w 0000:00:06.0 -c 0x3 -- --total-num-mbufs=2048 --disable-hw-vlan -ai
```

> **Note** `-w` specifies the bus-info value of an NIC. In this example, -w is set to 0000:00:06.0. `total-num-mbufs` specifies the huge page size. In this example, total-num-mbufs is set to 2048. In actual scenarios, replace them based on your needs.

The following figure shows the sample command output after DPDK is run.



Operations performed on an ECS bare metal instance

1. Run the following command to check the configurations of GRUB:

   ```
   cat /proc/cmdline
   ```

   Check whether the configurations of GRUB contain `intel_iommu=on` and `iommu=pt`.

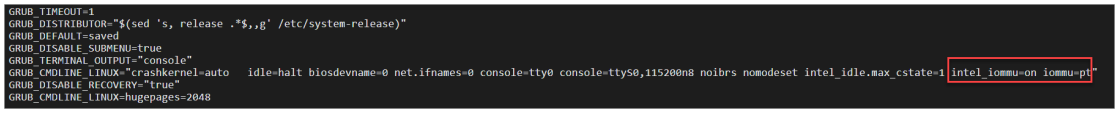
2. (Optional)If the configurations of GRUB do not contain `intel_iommu=on` and `iommu=pt`, manually add intel_iommu=on and iommu=pt.

   i. Run the following command to open the configuration file of GRUB:

   ```
   vim /etc/default/grub
   ```

ii. Switch to the edit mode, add `intel_iommu=on` and `iommu=pt` to the line of GRUB_CMDLINE_LINUX, and then save the configuration file.

The following figure shows an example of the modified configuration file.



iii. Run the following command to apply the modified configurations:

```
grub2-mkconfig -o /boot/grub2/grub.cfg
```

iv. Restart the instance and connect to the instance.

3. Run the following commands to install the VFIO and VFIO-PCI drivers:

```
modprobe vfio && \
modprobe vfio-pci
```

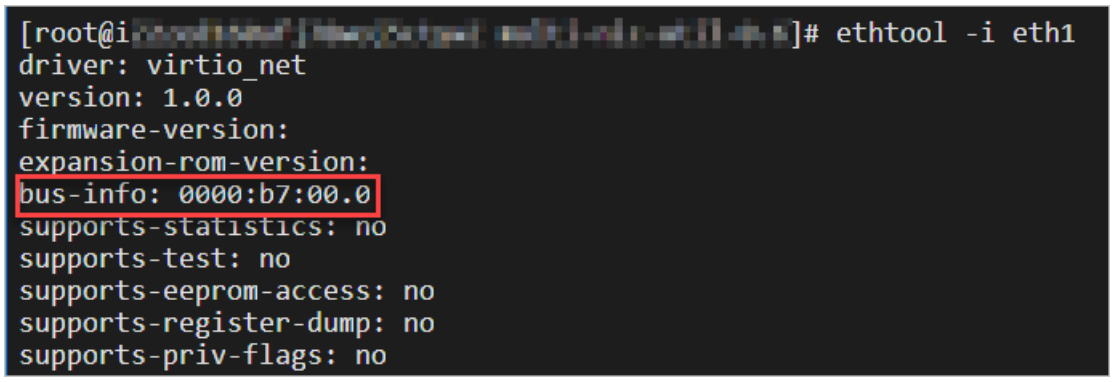
4. Run the following command to configure noiommu_mode:

```
echo 1 > /sys/module/vfio/parameters/enable_unsafe_noiommu_mode
```

5. Run the following command to view and record the bus-info value of the NIC bound to the instance:

```
ethtool -i ethX
```

Replace *ethX* with the ID of the NIC bound to the instance. In this example, eth1 is used.



6. Switch to the *usertools* directory under the installation directory of DPDK and run the following command to bind the NIC to the VFIO-PCI driver:

```
./dpdk-devbind.py -b vfio-pci 0000:b7:00.0
```

> **Note** In this example, `0000:b7:00.0` is the bus-info value of eth1. In actual scenarios, replace it with the bus-info value of your NIC.

Run the `./dpdk-devbind.py --status` command to view the binding state of the NIC. The following command output indicates that the NIC whose bus-info value is `0000:b7:00.0` has been bound to the VFIO-PIC driver.

```
[root@i                                    s]# ./dpdk-devbind.py --status

Network devices using DPDK-compatible driver
============================================
0000:b7:00.0 'Virtio network device 1000' drv=vfio-pci unused=virtio_pci

Network devices using kernel driver
===================================
0000:b6:00.0 'Virtio network device 1000' if=eth0 drv=virtio-pci unused=virtio_pci,vfio-pci *Active*
```

7. Switch to the *build/app* directory under the installation directory of DPDK and run one of the following commands to start DPDK.
   - If the version of DPDK is 18.02 or later, run the following command:

     ```
     ./testpmd -w 0000:b7:00.0 -c 0x3 -- --total-num-mbufs=2048 -ai
     ```

   - If the version of DPDK is earlier than 18.02, run the following command:

     ```
     ./testpmd -w 0000:b7:00.0 -c 0x3 -- --total-num-mbufs=2048 --disable-hw-vlan -ai
     ```

     > **Note** `-w` specifies the bus-info value of an NIC. In this example, -w is set to 0000:b7:00.0. `total-num-mbufs` specifies the huge page size. In this example, total-num-mbufs is set to 2048. In actual scenarios, replace them based on your needs.

     The following figure shows the sample command output after DPDK is run.

```
[root@i                                    app]# ./testpmd -w 0000:b7:00.0 -c 0x3 -- --total-num-mbufs=2048 -ai
EAL: Detected 104 lcore(s)
EAL: Detected 2 NUMA nodes
EAL: Multi-process socket /var/run/dpdk/rte/mp_socket
EAL: No free hugepages reported in hugepages-1048576kB
EAL: Probing VFIO support...
EAL: VFIO support initialized
EAL: PCI device 0000:b7:00.0 on NUMA socket 1
EAL:   probe driver: 1af4:1000 net_virtio
EAL:   using IOMMU type 8 (No-IOMMU)
Auto-start selected
Interactive-mode selected
testpmd: create a new mbuf pool <mbuf_pool_socket_0>: n=2048, size=2176, socket=0
testpmd: preferred mempool ops selected: ring_mp_mc
testpmd: create a new mbuf pool <mbuf_pool_socket_1>: n=2048, size=2176, socket=1
testpmd: preferred mempool ops selected: ring_mp_mc

Warning! port-topology=paired and odd forward ports number, the last port will pair with itself.

Configuring Port 0 (socket 1)
EAL: Error disabling MSI-X interrupts for fd 19
Port 0: 00:16:3E:30:CF:C7
Checking link statuses...
Done
Start automatic packet forwarding
io packet forwarding - ports=1 - cores=1 - streams=1 - NUMA support enabled, MP allocation mode: native
Logical Core 1 (socket 0) forwards packets on 1 streams:
  RX P=0/Q=0 (socket 1) -> TX P=0/Q=0 (socket 1) peer=02:00:00:00:00:00

  io packet forwarding packets/burst=32
  nb forwarding cores=1 - nb forwarding ports=1
  port 0: RX queue number: 1 Tx queue number: 1
    Rx offloads=0x0 Tx offloads=0x0
    RX queue: 0
      RX desc=0 - RX free threshold=0
      RX threshold registers: pthresh=0 hthresh=0  wthresh=0
      RX Offloads=0x0
    TX queue: 0
      TX desc=0 - TX free threshold=0
      TX threshold registers: pthresh=0 hthresh=0  wthresh=0
```

10.6. Manage software on Linux instances

10.6.1. Add a software repository

In most cases, software packages for Linux are stored in software repositories. After you add a software repository, you can use the package management tool provided by Linux to search for, install, and update software applications in the repository. This topic describes how to add software repositories to different Linux distributions. In these examples, Alibaba Cloud software repositories are used.

Context

All users (even those without Alibaba Cloud accounts) can use Alibaba Cloud software repositories free of charge. You can go to the Alibaba Open Source Image Site to obtain the software repository of your Linux distribution.

Add a software repository to a CentOS instance

In this example, an Elastic Compute Service (ECS) instance that runs a CentOS 7 operating system is used. Operations may vary based on the version of your operating system.

> **Note** CentOS 6 and CentOS 8 have reached their end of life (EOL). In accordance with Linux community rules, all content has been removed from the CentOS 6 and CentOS 8 repository addresses. If you continue to use the default repository address of CentOS 6 or CentOS 8, an error is reported. We recommend that you change the repository address of CentOS 6 or CentOS 8. For more information, see Change the CentOS 6 source address and Change CentOS 8 repository addresses.

1. Connect to the instance.

   For more information, see Connection methods.

2. Run the following command to back up the original software repository:

   ```
   sudo mv /etc/yum.repos.d/CentOS-Base.repo /etc/yum.repos.d/CentOS-Base.repo.backup
   ```

3. Run one of the following commands to add the Alibaba Cloud CentOS 7 software repository to the instance:

   ```
   sudo wget -O /etc/yum.repos.d/CentOS-Base.repo http://mirrors.aliyun.com/repo/Centos-7.repo
   ```

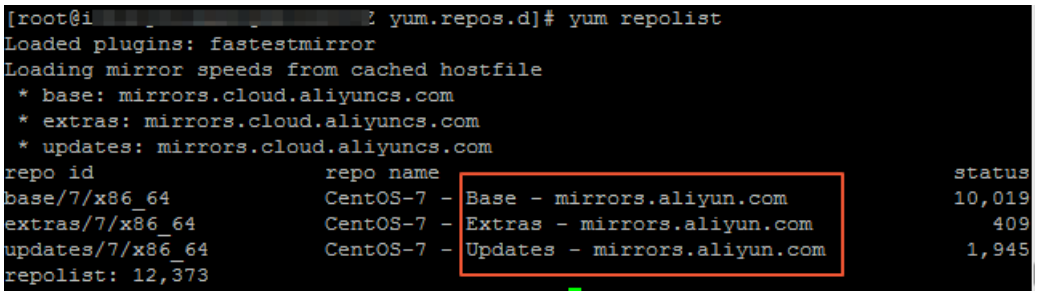
   ```
   sudo curl -o /etc/yum.repos.d/CentOS-Base.repo http://mirrors.aliyun.com/repo/Centos-7.repo
   ```

> **Note**
>
> - To add an Alibaba Cloud software repository to an instance that runs another version of CentOS, go to the Alibaba Open Source Image Site and click `centos`.
> - To add a software repository that is not from Alibaba Cloud, replace `http://mirrors.aliyun.com/repo/Centos-7.repo` in the preceding commands with the URL of the software repository that you want to add.

4. Run the following command to generate a local cache for fast search and installation of software:

```
sudo yum clean all && sudo yum makecache
```

5. Run the `sudo yum repolist` command to check whether the software repository is added. The following command output indicates that the Alibaba Cloud software repository is added to the CentOS 7 instance.



Add a software repository to an Ubuntu instance

In this example, an ECS instance that runs an Ubuntu 18.04 operating system is used. Operations may vary based on the version of your operating system.

1. Connect to the instance.

   For more information, see Connection methods.

2. Run the following command to back up the original software repository:

```
sudo cp /etc/apt/sources.list /etc/apt/sources.list.bakup
```

3. Run the `sudo vim /etc/apt/sources.list` command to open the sources.list file and add the following information to the file.

   For information about how to add information to files, see Use the Vim editor.

```
deb http://mirrors.cloud.aliyuncs.com/ubuntu/ bionic main restricted universe multivers
e
deb-src http://mirrors.cloud.aliyuncs.com/ubuntu/ bionic main restricted universe multi
verse
deb http://mirrors.cloud.aliyuncs.com/ubuntu/ bionic-security main restricted universe
multiverse
deb-src http://mirrors.cloud.aliyuncs.com/ubuntu/ bionic-security main restricted unive
rse multiverse
deb http://mirrors.cloud.aliyuncs.com/ubuntu/ bionic-updates main restricted universe m
ultiverse
deb-src http://mirrors.cloud.aliyuncs.com/ubuntu/ bionic-updates main restricted univer
se multiverse
deb http://mirrors.cloud.aliyuncs.com/ubuntu/ bionic-proposed main restricted universe
multiverse
deb-src http://mirrors.cloud.aliyuncs.com/ubuntu/ bionic-proposed main restricted unive
rse multiverse
deb http://mirrors.cloud.aliyuncs.com/ubuntu/ bionic-backports main restricted universe
multiverse
deb-src http://mirrors.cloud.aliyuncs.com/ubuntu/ bionic-backports main restricted univ
erse multiverse
```

> **Note**
>
> - To add an Alibaba Cloud software repository to an instance that runs another version of Ubuntu, go to the Alibaba Open Source Image Site and click `ubuntu`. To add an Alibaba Cloud software repository to an ECS instance and reduce data transfer costs, replace `http://mirrors.aliyun.com/ubuntu/` on the ubuntu page of the Alibaba Open Source Image Site with `http://mirrors.cloud.aliyuncs.com/ubuntu`.
> - To add a software repository that is not from Alibaba Cloud, replace the added information with the information of the software repository that you want to add.

4. Run the `sudo apt-get update` command to update software package information in the software repository.

Add a software repository to a Debian instance

In this example, an ECS instance that runs a Debian 8.9 operating system is used. Operations may vary based on the version of your operating system.

1. Connect to the instance.

   For more information, see Connection methods.

2. Run the following command to back up the original software repository:

   ```
   sudo cp /etc/apt/sources.list /etc/apt/sources.list.bakup
   ```

3. Run the `sudo vim /etc/apt/sources.list` command to open the sources.list file and add the following information to the file.

   For information about how to add information to files, see Use the Vim editor.

```
deb http://mirrors.cloud.aliyuncs.com/debian/ jessie main non-free contrib
deb http://mirrors.cloud.aliyuncs.com/debian/ jessie-proposed-updates main non-free con
trib
deb-src http://mirrors.cloud.aliyuncs.com/debian/ jessie main non-free contrib
deb-src http://mirrors.cloud.aliyuncs.com/debian/ jessie-proposed-updates main non-free
contrib
```

> **Note**
>
> - To add an Alibaba Cloud software repository to an instance that runs another version of Debian, go to the Alibaba Open Source Image Site and click `debian`. To add an Alibaba Cloud software repository to an ECS instance and reduce data transfer costs, replace `http://mirrors.aliyun.com/debian/` on the debian page of the Alibaba Open Source Image Site with `http://mirrors.cloud.aliyuncs.com/debian/`.
> - To add a software repository that is not from Alibaba Cloud, replace the added information with the information of the software repository that you want to add.

4. Run the `sudo apt-get update` command to update software package information in the software repository.

Add a software repository to a Fedora instance

1. Connect to the instance.

   For more information, see Connection methods.

2. Run the following commands to back up the original software repository:

```
sudo mv /etc/yum.repos.d/fedora.repo /etc/yum.repos.d/fedora.repo.backup
sudo mv /etc/yum.repos.d/fedora-updates.repo /etc/yum.repos.d/fedora-updates.repo.backu
p
```

3. Run one of the following commands to add the Fedora software repository of Alibaba Cloud to the instance:

```
sudo wget -O /etc/yum.repos.d/fedora.repo http://mirrors.aliyun.com/repo/fedora.repo
```

```
sudo curl -o /etc/yum.repos.d/fedora.repo http://mirrors.aliyun.com/repo/fedora.repo
```

> **Note** To add a Fedora software repository that is not from Alibaba Cloud, replace `http://mirrors.aliyun.com/repo/fedora.repo` in the preceding commands with the URL of the software repository that you want to add.

4. Run one of the following commands to add the fedora-updates software repository of Alibaba Cloud to the instance:

```
sudo wget -O /etc/yum.repos.d/fedora-updates.repo http://mirrors.aliyun.com/repo/fedo
ra-updates.repo
```

```
sudo curl -o /etc/yum.repos.d/fedora-updates.repo http://mirrors.aliyun.com/repo/fedo
ra-updates.repo
```

> **Note** To add a fedora-updates software repository that is not from Alibaba Cloud, replace `http://mirrors.aliyun.com/repo/fedora-updates.repo` in the preceding commands with the URL of the software repository that you want to add.

5. Run the following command to generate a local cache:

```
sudo yum clean all && sudo yum makecache
```

What's next

After you add a software repository, you can install software packages. For more information, see Install software packages.

10.6.2. Install software packages

This topic describes how to install software packages in different Linux distributions by using Apache HTTP Server in Alibaba Cloud software repositories. After you add a software repository, you can use the package management tool provided by Linux to install packages based on your needs.

Prerequisites

A software repository that contains the software packages to be updated is added before you install the software packages. In this example, you must add an Alibaba Cloud software repository before you proceed. For more information about how to add a software repository, see Add a software repository.

Install a software package in CentOS

1. Connect to a Linux instance. For more information, see Connection methods.
2. Run the following command to install a software package:

```
yum install <package> # Replace <package> with the software package that you want to install.
```

In this example, run the following command to install Apache HTTP Server:

```
yum install httpd
```

Install a software package in Debian or Ubuntu

1. Connect to a Linux instance. For more information, see Connection methods.
2. Run the following commands to install a software package:

```
# apt-get update
# apt-get install <package> # Replace <package> with the software package that you want to install.
```

In this example, run the following commands to install Apache HTTP Server:

```
# apt-get update
# apt-get install apache2
```

Install a software package in OpenSUSE

1. Connect to a Linux instance. For more information, see Connection methods.
2. Run the following command to install a software package:

```
zypper install <package> # Replace <package> with the software package that you want to install.
```

In this example, run the following command to install Apache HTTP Server:

```
zypper install apache2
```

10.6.3. Update software

This topic describes how to update a specific software application or all applications in different Linux distributions. Software packages in Linux distributions are constantly updated to add new features, fix bugs, and provide security updates. You can update software to the latest version based on your business requirements.

Prerequisites

A software repository that contains software package updates is added before you update software applications. For more information about how to add a software repository, see Add a software repository.

Update software applications in CentOS

1. Connect to a Linux instance. For more information, see Connection methods.
2. Update software applications.
   - Run the following command to update a single software application:

```
yum update <package> # Replace <package> with the software application that you want to update.
```

For example, if you want to update Apache HTTP Server, run the following command:

```
yum update httpd
```

   - Run the following command to update all software applications in the system.

> **Notice** If you run this command, the update process may update the kernel of the operating system and cause some issues. For example, the system may not start or certain software applications may be incompatible with the updated kernel. We recommend that you configure the system to skip kernel updates before you run this command.

```
yum update
```

Update software applications in Ubuntu or Debian

1. Connect to a Linux instance. For more information, see Connection methods.
2. Run the following command to obtain a list of software package updates:

```
apt-get update
```

3. Update software applications.
   - Run the following command to update a software application:

     ```
     apt-get install <package> # Replace <package> with the software application that you want to update.
     ```

     For example, if you want to update Python, run the following command:

     ```
     apt-get install python
     ```

   - Run the following command to update all software applications in the system:

     > **Note** If you run this command, the update process may update the kernel of the operating system and cause some issues. For example, the system may not start or certain software applications may be incompatible with the updated kernel. We recommend that you configure the system to skip kernel updates before you run this command.

> **Note** If you run this command, the update process may update the kernel of the operating system and cause some issues. For example, the system may not start or certain software applications may be incompatible with the updated kernel. We recommend that you configure the system to skip kernel updates before you run this command.

     ```
     apt-get upgrade
     ```

Update software applications in OpenSUSE

1. Connect to a Linux instance. For more information, see Connection methods.
2. Run the following command to obtain a list of software packages that need to be updated:

   ```
   zypper list-updates
   ```

3. Update software applications.
   - Run the following command to update a software application:

     ```
     zypper update <package> # Replace <package> with the software application that you want to update.
     ```

     For example, if you want to update Python, run the following command:

     ```
     zypper update python
     ```

   - Run the following command to update all software applications in the system:

     > **Note** If you run this command, the update process may update the kernel of the operating system and cause some issues. For example, the system may not start or certain software applications may be incompatible with the updated kernel. We recommend that you configure the system to skip kernel updates before you run this command.

> **Note** If you run this command, the update process may update the kernel of the operating system and cause some issues. For example, the system may not start or certain software applications may be incompatible with the updated kernel. We recommend that you configure the system to skip kernel updates before you run this command.

```
zypper update
```

Result

After a software application is updated, you can check the software version. If the latest version number is displayed, the software application is updated.

10.7. Configure time

10.7.1. Alibaba Cloud NTP server

This topic describes Alibaba Cloud Network Time Protocol (NTP) servers. Alibaba Cloud provides internal and public NTP servers to synchronize the local time of Elastic Compute Service (ECS) instances in networks.

Internal and public NTP servers

NTP is used to synchronize computer time in a network.

The consistency of time and time zones in ECS is crucial because it can affect task execution results. For example, when you update a database or analyze logs, the time sequence significantly impacts the results. When you run business on ECS instances, you must standardize the time zones of all involved instances to prevent issues such as logical confusions and network request errors. You can use NTP servers to synchronize the local time of all your ECS instances in a network.

ECS provides high-precision NTP servers for your use. The `ntp.cloud.aliyuncs.com` server offers a distributed NTP service that uses Stratum 1 servers. Stratum 1 servers are suitable for industries such as finance, communication, scientific research, and astronomy that require precise timing. The NTP service is also used to synchronize the local time between ECS instances and other cloud services. The following table describes the domain names of Alibaba Cloud NTP servers in various networks. These domain names support only IPv4.

Classic network (internal network)	Virtual private cloud (VPC) (internal network)	Internet
-	ntp.cloud.aliyuncs.com	ntp.aliyun.com
ntp1.cloud.aliyuncs.com	ntp7.cloud.aliyuncs.com	ntp1.aliyun.com
ntp2.cloud.aliyuncs.com	ntp8.cloud.aliyuncs.com	ntp2.aliyun.com
ntp3.cloud.aliyuncs.com	ntp9.cloud.aliyuncs.com	ntp3.aliyun.com
ntp4.cloud.aliyuncs.com	ntp10.cloud.aliyuncs.com	ntp4.aliyun.com
ntp5.cloud.aliyuncs.com	ntp11.cloud.aliyuncs.com	ntp5.aliyun.com
ntp6.cloud.aliyuncs.com	ntp12.cloud.aliyuncs.com	ntp6.aliyun.com

Classic network (internal network)	Virtual private cloud (VPC) (internal network)	Internet

-	-	ntp7.aliyun.com

Other public services

The following table describes other public services provided by Alibaba Cloud.

Public service	Description
Public Domain Name System (DNS): 223.5.5.5/223.6.6.6	Domain name: `http://www.alidns.com`
Website of public images: `https://developer.aliyun.com/mirror`	Update frequency: Images are updated from 02:00 to 04:00 every day (UTC+8). The images contain a wide collection of Linux distributions and open source software.

References

- Configure the NTP service for Windows instances
- Configure the NTP service for ECS instances that run CentOS 6

10.7.2. Configure the NTP service for Windows instances

This topic describes how to enable and configure the NTP service for a Windows instance to ensure that the local system time is precisely synchronized.

Context

By default, ECS instances in all Alibaba Cloud regions use UTC+8. You can configure or change time zones for your instances.

Windows Server 2012 R2 Datacenter Edition 64-bit is used in this topic to demonstrate how to use the NTP service to synchronize the local system time for Windows instances.

Enable the NTP service

By default, the Windows Time service is enabled on Windows Server operating systems. The NTP service must be enabled for Windows instances to synchronize the local system time. Perform the following operations to check and enable the NTP service:

1. Connect to the Windows ECS instance. For more information, see Connection methodsGuidelines on instance connection.
2. Click the  icon and open the **Run** dialog box. Run the `services.msc` command in the Run dialog box.
3. In the **Services** dialog box, find and double-click **Windows Time**.

4. In the **Windows Time Properties (Local Computer)** dialog box, perform the following operations:
   i. Set **Startup type** to **Automatic**.
   ii. Make sure that the value of **Service status** is **Running**. Otherwise, click **Start**.
   iii. Click **Apply** and then click **OK**.

Modify the default NTP server address

By default, Windows Server operating systems use the Microsoft NTP server (time.windows.com), but errors may occur while the operating systems are synchronizing with the Microsoft NTP server. When you use a Windows instance, you can replace the default NTP server with an internal NTP server provided by Alibaba Cloud. Perform the following operations to modify the default NTP server address:

1. Connect to the Windows ECS instance. For more information, see Connection methodsGuidelines on instance connection.
2. In the notification area of the taskbar, click the date and time, and then click **Change date and time settings**.
3. In the **Date and Time** dialog box, click the **Internet Time** tab and then click **Change settings**.
4. In the **Internet Time Settings** dialog box, select **Synchronize with an Internet time server**, enter the address of an Alibaba Cloud internal NTP server, and then click **Update now**. For more information, see Alibaba Cloud NTP server.

Modify the NTP synchronization interval

The default NTP synchronization interval is five minutes. You can modify the interval. Perform the following operations to modify the NTP synchronization interval:

1. Connect to the Windows ECS instance. For more information, see Connection methodsGuidelines on instance connection.
2. Click the  icon and open the **Run** dialog box. Run the `regedit` command in the Run dialog box.
3. In the left-side navigation pane of **Registry Editor**, choose **HKEY_LOCAL_MACHINE > SYSTEM > CurrentControlSet > Services > W32Time > TimeProviders > NtpClient** and double-click **SpecialPollInterval**.
4. In the **Edit DWORD (32-bit) Value** dialog box, select **Decimal** in the **Base** section and enter a value in the **Value data** field. The entered value is the new synchronization interval. Unit: seconds.
5. Click **OK**.
6. Restart the instance for the changes to take effect.

   You can restart the instance for the changes to take effect. If you cannot restart the instance due to business requirements, you can restart the NTP service for the changes to take effect. Perform the following operations:

   i. Click the  icon and open the **Run** dialog box. Run the `services.msc` command in the Run dialog box.
   ii. In the **Services** dialog box, find and double-click **Windows Time**.
   iii. In the **Windows Time Properties (Local Computer)** dialog box, click **Stop**.

iv. Click **Start** after the service enters the **Stopped** state.

v. Click **OK**.

Related information

- Alibaba Cloud NTP server
- Configure the NTP service for ECS instances that run CentOS 6
- Configure chrony for Linux instances (CentOS 7)
- Configure chrony for Linux instances (Alibaba Cloud Linux 2)

10.7.3. Configure chrony for Linux instances (Alibaba Cloud Linux 2)

This topic describes how to modify the time zone of a Linux instance and how to enable, configure, and use chrony to ensure that the local system time of the instance is synchronized precisely with the standard time. An instance that runs an Alibaba Cloud Linux 2.1903 LTS 64-bit operating system is used in the examples.

Prerequisites

An inbound rule is added to a security group of the ECS instance to allow traffic on UDP port 123. For more information, see Add a security group rule.

Context

Alibaba Cloud Linux 2 uses chrony to synchronize local system time with the standard time. chrony consists of the following core programs:

- chronyd is a daemon process that runs in the background. chronyd is used to adjust the system clock that runs in the kernel to synchronize with the NTP clock server. chronyd can determine accurate statistics for the difference between the UTC time and the local system time and adjust the system time accordingly.
- chronyc provides a user interface to monitor the performance of chronyd and to change various operating parameters in chronyd. chronyc can run on a server controlled by chronyd or a server not controlled by chronyd.

For more information, visit Chrony.

Change the time zone of a Linux instance

1. Connect to the Linux instance. For more information, see Connect to a Linux instance by using a password or key.
2. Run the following command to view the time zone list:

   ```
   ls /usr/share/zoneinfo/<Name of the time zone folder>
   ```

   For example, you can run the following command to view the `Hong_Kong` time zone in the list:

   ```
   ls /usr/share/zoneinfo/Asia
   ```

3. Run the following command to change the time zone:

```
ln -sf /usr/share/zoneinfo/Asia/Hong-Kong /etc/localtime
```

4. Run the following command to update the real-time clock (RTC):

```
hwclock -w
```

5. Run the following command to view the time zone:

```
timedatectl status
```

A command output similar to the following one indicates that the time zone is changed to `Hong_Kong`:

```
      Local time: -- 2020-09-14 08:00:04 UTC
  Universal time:-- 2020-09-14 08:00:04 UTC
        RTC time: -- 2020-09-14 08:00:04
       Time zone: Asia/Hong-Kong (UTC, +0000)
```

Enable chrony

1. Connect to the Linux instance. For more information, see Connect to a Linux instance by using a password or key.
2. Run the following commands to start the chronyd service and configure it to run on system startup:

```
systemctl start chronyd.service
systemctl enable chronyd.service
```

3. Run the following command to view the time synchronization status of the instance to check whether the service is started:

```
chronyc tracking
```

4. Run the following command to view the list of servers that have chrony enabled:

```
chronyc -n sources -v
```

Configure chrony

1. Connect to the Linux instance. For more information, see Connect to a Linux instance by using a password or key.
2. Run the following command to open the configuration file of chrony:

```
vim /etc/chrony.conf
```

3. Find `server <NTP server> minpoll 4 maxpoll 10 iburst` and press the /key to edit the file. Add `#` at the beginnings of sentences that contain the information of NTP servers that you want to hide.
4. Add a row of NTP server information in the `server <Required NTP server> minpoll 4 maxpoll 10 iburst` format. Then, press the *Esc* key and enter `:wq` to save the file and exit.

   For more information about NTP servers, see Alibaba Cloud NTP server.
5. Run the following commands to start the chronyd service and configure it to run on system startup:

```
systemctl start chronyd.service
systemctl enable chronyd.service
```

6. Run the following command to view the list of servers that have chrony enabled:

```
chronyc -n sources -v
```

Manually synchronize the clock by using chrony

1. Run the following command to access chrony:

```
chronyc
```

2. Run the following command to synchronize the clock in chrony:

```
makestep
```

> **Note** You can run the **help** command to obtain instructions for common chrony commands.

10.7.4. Configure chrony for Linux instances (CentOS 7)

This topic describes how to change the time zone of an Elastic Compute Service (ECS) Linux instance and how to enable, configure, and use chrony to ensure that the local system time of the instance is synchronized with the UTC (Coordinated Universal Time) time. In this topic, an instance that runs a CentOS 7.8 operating system is used.

Prerequisites

An inbound rule is added to a security group of the ECS instance to allow traffic on UDP port 123. For more information, see Add a security group rule.

Context

By default, ECS instances in all Alibaba Cloud regions use UTC+8. You can configure or change time zones for your instances.

CentOS 7 instances use chrony to synchronize the local system time with the UTC time. Compared with Network Time Protocol (NTP), chrony can synchronize the system clock more quickly and accurately on CentOS 6 instances and minimize time and frequency differences. chrony consists of the following core programs:

- chronyd is a daemon process that runs in the background. chronyd is used to adjust the system clock that runs in the kernel to synchronize with the NTP clock server. chronyd can determine accurate statistics for the difference between the UTC time and the local system time and adjust the system time accordingly.
- chronyc provides a user interface to monitor the performance of chronyd and to change various operating parameters in chronyd. chronyc can run on a server controlled by chronyd or a server not controlled by chronyd.

For more information, see Chrony.

Change the time zone of a Linux instance

1. Connect to the Linux instance. For more information, see Connect to a Linux instance by using a password or key.
2. Run the following command to view the time zone list:

   ```
   ls /usr/share/zoneinfo/<Name of the time zone folder>
   ```

   For example, you can run the following command to view the `Hong_Kong` time zone in the list:

   ```
   ls /usr/share/zoneinfo/Asia
   ```

3. Run the following command to change the time zone:

   ```
   ln -sf /usr/share/zoneinfo/Asia/Hong-Kong /etc/localtime
   ```

4. Run the following command to update the real-time clock (RTC):

   ```
   hwclock -w
   ```

5. Run the following command to view the time zone:

   ```
   timedatectl status
   ```

   A command output similar to the following one indicates that the time zone is changed to `Hong_Kong`:

   ```
         Local time: -- 2020-09-14 08:00:04 UTC
     Universal time:-- 2020-09-14 08:00:04 UTC
           RTC time: -- 2020-09-14 08:00:04
          Time zone: Asia/Hong-Kong (UTC, +0000)
   ```

Enable chrony

1. Connect to the Linux instance. For more information, see Connect to a Linux instance by using a password or key.
2. Run the following commands to start the chronyd service and configure it to run on system startup:

   ```
   systemctl start chronyd.service
   systemctl enable chronyd.service
   ```

3. Run the following command to view the time synchronization status of the instance to check whether the service is started:

   ```
   chronyc tracking
   ```

4. Run the following command to view the list of servers that have chrony enabled:

   ```
   chronyc -n sources -v
   ```

Configure chrony

1. Connect to the Linux instance. For more information, see Connect to a Linux instance by using a password or key.

2. Run the following command to open the configuration file of chrony:

```
vim /etc/chrony.conf
```

3. Find `server <NTP server> minpoll 4 maxpoll 10 iburst` and press the /key to edit the file. Add `#` at the beginnings of sentences that contain the information of NTP servers that you want to hide.
4. Add a row of NTP server information in the `server <Required NTP server> minpoll 4 maxpoll 10 iburst` format. Then, press the *Esc* key and enter `:wq` to save the file and exit.

   For more information about NTP servers, see Alibaba Cloud NTP server.
5. Run the following commands to start the chronyd service and configure it to run on system startup:

```
systemctl start chronyd.service
systemctl enable chronyd.service
```

6. Run the following command to view the list of servers that have chrony enabled:

```
chronyc -n sources -v
```

Manually synchronize the clock by using chrony

1. Run the following command to access chrony:

```
chronyc
```

2. Run the following command to synchronize the clock in chrony:

```
makestep
```

> **Note** You can run the **help** command to obtain instructions for common chrony commands.

10.7.5. Configure the NTP service for ECS instances that run CentOS 6

This topic describes how to change the time zone of a Linux instance, and how to enable and configure the Network Time Protocol (NTP) service to ensure that the local time for ECS instances is precisely synchronized. CentOS 6.5 is used in this topic.

Prerequisites

The NTP service uses UDP port 123 for communication. Before you configure the NTP service, make sure that UDP port 123 is enabled. You can run the `netstat -nupl` command to check whether UDP port 123 is enabled. For more information about how to allow traffic on UDP port 123, see Add security group rules.

Context

By default, ECS instances in all Alibaba Cloud regions use UTC+8. You can configure or change time zones for your instances.

The NTP service ensures that the local time for ECS instances is synchronized with the standard time. In Linux, you can run the **ntpdate** or **ntpd** command to synchronize the system clock to an NTP server. This topic describes the standard and custom NTP service configurations. You can choose one based on your requirements. For more information, see the "Internal and public NTP servers" section in Alibaba Cloud NTP server.

- `ntpdate` performs a one-time-only update to the system clock. For newly purchased instances, you can use `ntpdate` to synchronize time.
- `ntpd` adjusts the system clock in small steps. For instances with running workloads, we recommend that you use `ntpd` to synchronize time.

Modify the time zone of a Linux instance

1. Connect to the Linux instance. For more information, see Connection methodsGuidelines on instance connection.

   > **Note** You must open and edit the time zone configuration file as a root user. The `sudo` command is used in this example.

2. Run the `sudo rm /etc/localtime` command to delete the local time URL in the system.
3. Run the `sudo vi /etc/sysconfig/clock` command to use vim to open and edit the */etc/sysconfig/clock* configuration file.
4. Enter `i` to add a time zone city. For example, you can add `Zone=Asia/Shanghai`, press the Esc key to exit the edit mode, and then enter `:wq` to save and exit.

   You can run the `ls /usr/share/zoneinfo` command to query the list of time zones. `Shanghai` is included in the list of time zones.
5. Run the `sudo ln -sf /usr/share/zoneinfo/Asia/Shanghai /etc/localtime` command to update the time zone change.
6. Run the `hwclock -w` command to update the real-time clock (RTC).
7. Run the `sudo reboot` command to restart the instance.
8. Run the `date -R` command to check whether the time zone change takes effect. If the change does not take effect, repeat the preceding operations.

Enable the standard NTP service

1. Connect to the Linux instance. For more information, see Connection methodsGuidelines on instance connection.
2. Run the `sudo service ntpd start` command to enable the NTP service.
3. Run the `chkconfig ntpd on` command to enable the NTP service to run upon startup.
4. Run the `ntpstat` command to check whether the NTP service is enabled.
5. (Optional)Run the `ntpq -p` command to view the list of NTP peers and run the `sudo chkconfig --list ntpd` command to view the run level for the NTP service.

Configure the custom NTP service

1. Connect to the Linux instance. For more information, see Connection methodsGuidelines on instance connection.

2. Run the `sudo vi /etc/ntp.conf` command to use vim to open and edit the configuration file of the NTP service.
3. Find the `server ntp server iburst` information and enter `i` to start to edit the file. For NTP servers that you do not need, you can add a number sign (`#`) at the beginning of the lines to hide the servers.
4. Add a new line of NTP server information in the following format: `server <the NTP server that you want to add> iburst`. After you edit the file, press the Esc key and enter `:wq` to save the file and exit.
5. Run the `sudo service ntpd start` command to enable the custom NTP service.
6. Run the `chkconfig ntpd on` command to enable the custom NTP service to run upon startup.
7. Run the `ntpstat` command to check whether the NTP service is enabled.

Related information

- Alibaba Cloud NTP server
- Configure the NTP service for Windows instances

11.Instance FAQ

This topic provides answers to frequently asked questions about Elastic Compute Service (ECS) instances.

- FAQ about purchasing instances
  - How do I check which instance resources are available for purchase within a specific region or zone?
  - What do I do if resources are sold out when I attempt to purchase an instance within a specific region or zone?
  - How do I select an instance type that is suitable for my business?
  - How long does it take to create an ECS instance?
  - I paid for an instance but no instance was created. Why?
  - Why does a purchased instance have a memory size different from that defined in the instance type?
- FAQ about enterprise-level instances
  - What are enterprise-level instances? What are shared instances?
  - What are the differences between enterprise-level and shared instances?
  - Which instance families are enterprise-level instance families? Which instance families are shared instance families?
  - In what business scenarios do I need to purchase enterprise-level instances?
  - How is the network performance of enterprise-level instances?
  - Which disk categories do enterprise-level instances support?
  - Which image types do enterprise-level instances support?
  - What are the limits on changing the instance types of enterprise-level instances?
  - Can I change a shared instance into an enterprise-level instance?
- FAQ about persistent memory-optimized instances
  - What characteristics do persistent memory-optimized instances have?
  - What requirements do persistent memory-optimized instances have for operating systems?
  - After I purchase a persistent memory-optimized instance, how do I configure its persistent memory as memory?
  - When persistent memory is being used as memory on persistent memory-optimized instances, can I deploy and run Redis applications on the instances?
  - I already have a Redis cluster that is based on regular memory and want to migrate business from this cluster to persistent memory-optimized instances. How do I do this? What items do I need to take note of?
  - Can I deploy and run a parameter server (PS) architecture on persistent memory-optimized instances whose persistent memory is used as memory?
  - After I purchase a persistent memory-optimized instance, how do I configure its persistent memory as local disks?
  - To which applications are local disks that can deliver higher performance applicable?
  - Can I deploy Redis or MySQL applications on persistent memory-optimized instances whose persistent memory is used as local disks? Do the applications need to be modified in the same manner as they do when persistent memory is used as memory?

- How is the performance of persistent memory used as local disks, compared with local non-volatile memory express (NVMe) SSDs and cloud disks?
- How reliable is persistent memory?

- FAQ about GPU-accelerated instances

  Elastic GPU Service FAQ

- FAQ about ECS bare metal instances
  - What are the differences between ECS bare metal instances, traditional cloud hosts (virtual machines), and traditional physical machines?
  - How is the network performance of ECS bare metal instances?
  - Which disk categories do ECS bare metal instances support? How many data disks can be attached to a single ECS bare metal instance?
  - Can the configurations of ECS bare metal instances be upgraded or downgraded? Do the instances support failover?

- FAQ about Super Computing Cluster (SCC)
  - How do I create an SCC instance?
  - How are SCC instances billed?
  - How do I create an SCC by using E-HPC?
  - How do I use SCC RDMA?

- FAQ about preemptible instances
  - FAQ about purchasing preemptible instances
    - Which instance types support preemptible instances?
    - In which regions can I create preemptible instances?
    - When I attempt to purchase an ECS instance, the Preemptible Instance option is unavailable on the instance buy page. Why?
    - How do I bid for a preemptible instance?
    - What is the relationship between the user-defined maximum hourly price and the spot price of a preemptible instance?
    - Can I view the spot price of an instance type when I purchase a preemptible instance?
    - Can I view the price history of a preemptible instance type? How?
    - How many preemptible instances can be purchased by a single account?
    - How do I increase my vCPU-based quotas?
  - FAQ about changing preemptible instances
    - Can preemptible instances be converted into subscription instances?
    - Can the instance type of a preemptible instance be changed?
    - A preemptible instance does not meet my requirements. How do I upgrade it to a higher-specification instance type at a low cost?

- FAQ about releasing preemptible instances
  - If I select Use Automatic Bid (SpotAsPriceGo) when I create a preemptible instance, will the created instance be automatically released due to insufficient resources?
  - If I set the maximum hourly price to the pay-as-you-go price for the selected instance type when I create a preemptible instance, may the instance be automatically released?
  - Can I cancel or reschedule the automatic release of a preemptible instance?
  - I do not have overdue payments in my account. Why have my preemptible instances been released?
  - How are release rates on the instance buy page calculated? Why are most of them in the range of 0% to 3%?
  - Am I notified when my preemptible instances are about to be released? How am I notified?
  - Can the data of a preemptible instance be automatically retained when the instance is released?
  - How do I retain the public IP address of a preemptible instance when the instance is released?
- FAQ about preemptible instance charges
  - How is a preemptible instance billed if its lifespan is shorter than 1 hour?
  - To what resources are the prices of preemptible instances applicable?
  - Why do preemptible instances of the same instance type and region vary widely in spot prices across different zones?
  - May the spot price of a preemptible instance exceed the pay-as-you-go price?
  - For the first hour after a preemptible instance is created, am I charged at a price that fluctuates with the spot price?
  - Do I continue to be charged for preemptible instances after they are stopped?
- FAQ about preemptible instances without a protection period
  - Which is more cost-effective: a preemptible instance with a protection period or a preemptible instance without a protection period?
  - Is the release rate of preemptible instances without a protection period higher than that of preemptible instances with a protection period?
  - Are preemptible instances without a protection period prioritized for release over those with a protection period?
  - If I have created a preemptible instance without a protection period, do subsequently created preemptible instances also not have a protection period?
  - If a preemptible instance does not have a protection period, do I still receive a notification 5 minutes in advance before the instance is released?
  - Are fewer resources provided for preemptible instances with a protection period than for those without a protection period?
  - Can I convert between preemptible instances with and without a protection period?

- FAQ about reserved instances
  - What are reserved instances?
  - Do reserved instances provide reserved resources?
  - What operating systems do reserved instances support?
  - Which instance families do reserved instances support?
  - Can reserved instances be applied to preemptible instances?

- FAQ about Virtual Network Computing (VNC)
  - Does a VNC management terminal allow multiple users to log on simultaneously?
  - What do I do if I forget the connection password?
  - Why am I unable to connect to a VNC management terminal even after I reset my connection password?
  - I was prompted with an authentication failure when I attempted to connect to a VNC management terminal. What do I do?
  - What do I do if a black screen appears while I am connected to a VNC management terminal?
  - What do I do if a VNC management terminal cannot be accessed?
  - Why am I unable to use Internet Explorer 8.0 to access a VNC management terminal?
  - When I use Firefox to access a VNC management terminal, an error message is returned indicating that a secure connection cannot be established. What do I do?
  - How do I connect to a Linux instance?
  - What are the default username and password used to connect to the operating system of an ECS instance?
  - How do I adjust the desktop resolution of a Windows instance?
  - Why do two cursors appear after I log on to a Windows instance by using a VNC management terminal?
- FAQ about using third-party software
  - What do I do if a Data Plane Development Kit (DPDK) application cannot be deployed on an ECS instance?

- FAQ about upgrading and downgrading instance configurations
  - Can I upgrade the instance types and other configurations of subscription instances?
  - Can I upgrade the instance types and other configurations of pay-as-you-go instances?
  - How long does it take to upgrade the configurations of an instance?
  - How is the fee for an instance configuration upgrade calculated?
  - Are my cloud service configurations affected if I upgrade the configurations of ECS instances?
  - How do I upgrade ECS resources?
  - I have upgraded the configurations of an instance but no changes have taken effect. Why?
  - After I place an order to upgrade the configurations of an instance, can I cancel the order to restore the instance to its original configurations?
- FAQ about managing instances
  - What do I do if I cannot access a website that runs on an ECS instance?
  - My ECS instance was stuck in the Starting state, and AliyunService was disabled or deleted. What do I do?
  - How do I use f1 instances?
  - How do I use FTP tools in macOS to upload files?
  - How do I apply for an ICP filing for my domain name after I purchase an ECS instance?
  - An ECS instance cannot load the kernel to start. What do I do?
  - How do I change the logon password from within an instance?
  - Why am I unable to add sound or video cards to ECS instances?

- What do I do if an order cannot be placed to change the billing method of an instance from pay-as-you-go to subscription?
- How long after an order is paid does it take to change the billing method of an instance from pay-as-you-go to subscription?
- What do I do if the billing method of an instance cannot be changed from pay-as-you-go to subscription?
- When I change the billing method of an instance from pay-as-you-go to subscription, does the billing method for network usage of the instance also change?
- I have an unpaid order to change the billing method of an instance from pay-as-you-go to subscription. If I upgrade the configurations of the instance, is the order still valid?
- What do I do if the billing method of an instance cannot be changed from subscription to pay-as-you-go?
- When I attempt to change the billing method of a disk in an instance, an error message is returned indicating that I have already changed the billing method three times. What does this mean?
- Why am I unable to change a pay-as-you-go instance into a subscription one?
- How do I view the expiration time of a subscription instance?

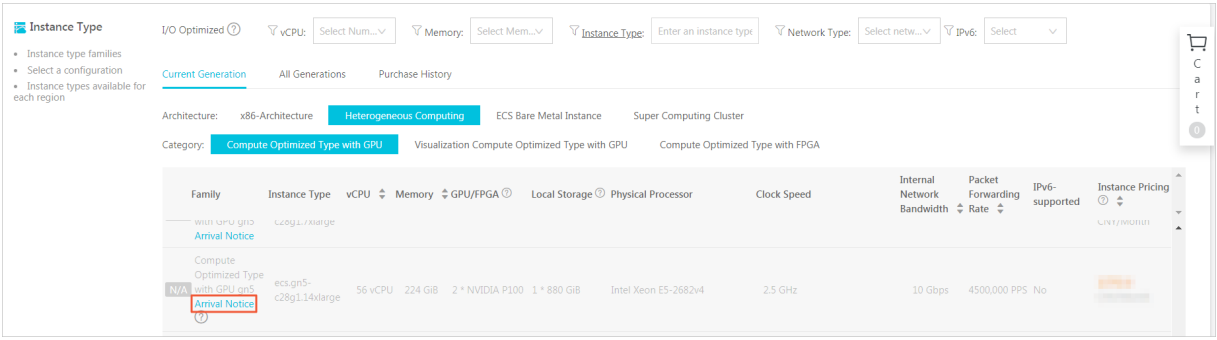
How do I check which instance resources are available for purchase within a specific region or zone?

What do I do if resources are sold out when I attempt to purchase an instance within a specific region or zone?

If a specific instance type is sold out when you attempt to purchase an instance of that instance type within a specific region or zone, take one of the following measures:

- Select another region.
- Select another zone.
- Select another instance type.

Instance resources are dynamic. Alibaba Cloud replenishes as appropriate the resources that are insufficient. If the resources that you want to purchase are still unavailable after you take all of the preceding measures, try again later. You can also use the arrival notification feature to receive notifications when the resources are replenished.



How do I select an instance type that is suitable for my business?

Consider the following factors when you select an instance type. For more information about how to select an instance type, see Best practices for instance type selection.

- Your business needs

- Your website type
- The average number of page views per day on your website
- The size of your homepage
- The data capacity of your website

How long does it take to create an ECS instance?

It takes a minute or two to create an ECS instance.

- You can connect to a Linux instance immediately after it is created. For more information, see Connect to an ECS instance.
- After a Windows instance is created, you must use Sysprep to initialize the operating system before you can connect to the instance. Do not restart the instance while its operating system is being initialized. After the operating system is initialized, you can connect to the instance. For more information, see Connect to an ECS instance. The amount of time required to initialize the operating system is determined based on whether the instance is I/O optimized.
  - For an I/O optimized Windows instance, it takes 2 to 3 minutes to initialize the operating system.
  - For a non-I/O optimized Windows instance, it takes 10 minutes to initialize the operating system.

> **Note** If an error occurs while the instance is being created, submit a ticket.

I paid for an instance but no instance was created. Why?

If resources within the specified zone are insufficient to create an instance of your selected instance type, the instance cannot be created. A refund is automatically credited to your account. If you do not receive a refund within half an hour, submit a ticket.

Why does a purchased instance have a memory size different from that defined in the instance type?

The memory size defined in each instance type is the total memory size, which includes the amount of memory occupied by the system, such as the BIOS reserved memory and memory overheads of running the kernel and the hypervisor. As a result, the size of available memory on your instance is smaller than the memory size defined in the instance type. Different instance families may use different technology stacks, which causes the percentage of memory occupied by the system to vary slightly.

What are enterprise-level instances? What are shared instances?

Enterprise-level instances belong to a series of instance families released by Alibaba Cloud in September 2016. Enterprise-level instances provide high performance, consistent computing power, and balanced network performance. Enterprise-level instances have exclusive and consistent computing, storage, and network resources, and are suitable for enterprise scenarios that require high business stability.

Shared instances belong to a series of instance families that are intended for individuals or small and medium-sized websites. Shared instances share resources, in contrast with enterprise-level instances that each have their own resources exclusively. As a result, shared instances do not provide consistent computing performance, but cost less.

What are the differences between enterprise-level and shared instances?

Enterprise-level instances use a CPU-bound scheduling scheme. Each vCPU is bound to a CPU hyperthread. Instances do not compete for CPU resources and do provide consistent computing performance as specified in the service-level agreement (SLA).

Shared instances use a CPU-unbound scheduling scheme. Each vCPU is randomly allocated to an idle CPU hyperthread. vCPUs of different instances compete for CPU resources, which causes computing performance to fluctuate when traffic loads are heavy. Shared instances can guarantee availability as specified in the SLA but not performance.

Which instance families are enterprise-level instance families? Which instance families are shared instance families?

Among the instance families that are available for purchase, t6, t5, s6, n4, mn4, xn4, and e4 are shared instance families, and the rest are enterprise-level instance families.

In what business scenarios do I need to purchase enterprise-level instances?

For business scenarios to which different enterprise-level instances are applicable, see Instance family and Best practices for instance type selection.

How is the network performance of enterprise-level instances?

The network performance of enterprise-level instances depends on their specifications. The higher their specifications are, the higher network performance the instances can deliver. For more information about the network performance of different instance specifications, see Instance family.

Which disk categories do enterprise-level instances support?

For information about the disk categories that enterprise-level instances support, see Disk categories.

Which image types do enterprise-level instances support?

For information about the public images that enterprise-level instances support, see Overview.

You can also import custom images. For more information, see Import custom images.

What are the limits on changing the instance types of enterprise-level instances?

For information about the limits on changing the instance types of enterprise-level instances, see Instance families that support instance type changes.

Can I change a shared instance into an enterprise-level instance?

Yes, you can change a shared instance into an enterprise-level instance. For more information, see Instance families that support instance type changes.

What characteristics do persistent memory-optimized instances have?

Persistent memory-optimized instances use high-capacity persistent memory, and provide slower access but higher data durability than the instances that use regular memory. When persistent memory-optimized instances are stopped or restarted, data in their persistent memory is not lost. Persistent memory can be used as memory or local disks.

- When persistent memory is used as memory, you can move some data from regular memory to persistent memory, such as non-hot data that does not require high-speed storage access. Persistent memory offers large capacity at a low price per GiB and can help reduce the total cost of ownership (TCO) per GiB of memory.
- When persistent memory is used as local disks, it functions like local SSDs to deliver ultra-high performance and a read/write latency as low as 400 nanoseconds. You can use persistent memory for core application databases that require consistent response time. You can also replace cache disks with persistent memory to achieve higher IOPS, higher bandwidth, and lower latency and improve cluster-wide business performance.

What requirements do persistent memory-optimized instances have for operating systems?

The following image versions can be used on persistent memory-optimized instances:

- Alibaba Cloud Linux 2
- CentOS 7.6 or later
- Ubuntu 18.10 or later
- SUSE Linux 12 SP4 or later

Alibaba Cloud provides support for Alibaba Cloud Linux 2. Alibaba Cloud Linux 2 integrates tools for scenarios in which persistent memory is applicable to work out of the box. In some scenarios such as those in which Redis applications are used, Alibaba Cloud Linux 2 outperforms community-supported Linux distributions by 20% in terms of performance.

After I purchase a persistent memory-optimized instance, how do I configure its persistent memory as memory?

You can use tools to configure the persistent memory as memory. For more information, see Configure persistent memory usage.

When persistent memory is being used as memory on persistent memory-optimized instances, can I deploy and run Redis applications on the instances?

You can significantly reduce the TCO per GiB of memory for Redis applications by running them on persistent memory-optimized instances. To ensure performance, you must modify the Redis applications by stratifying their data and storing the data in different types of memory. You can store non-hot data in persistent memory and hot data in regular memory.

To reduce your modification costs, Alibaba Cloud provides re6p instance types exclusively for Redis applications. You can run several commands to deploy Redis applications on re6p instances. For more information, see Deploy Redis applications on persistent memory-optimized instances.

> **Note** When you purchase persistent memory-optimized instances, select instance types named in the ecs.re6p-redis.<nx>large format.

I already have a Redis cluster that is based on regular memory and want to migrate business from this cluster to persistent memory-optimized instances. How do I do this? What items do I need to take note of?

You must ensure business consistency and data reliability while you are migrating business from the Redis cluster to persistent memory-optimized instances. We recommend that you first purchase a single persistent memory-optimized instance and test it with a small amount of required business to check whether the basic performance and capacity model of the instance meet your business requirements. If the basic performance and capacity model of the instance meet your business requirements, you can scale out to more persistent memory-optimized instances and take over business from the entire Redis cluster.

Can I deploy and run a parameter server (PS) architecture on persistent memory-optimized instances whose persistent memory is used as memory?

Server nodes in a PS architecture store all training parameters of training clusters, traditionally in their memory. These parameters consume a large amount of memory and are expensive to store. When you run a PS architecture on persistent memory-optimized instances whose persistent memory is used as memory, we recommend that you store all parameters in persistent memory and only hash tables in regular memory. This way, you can significantly reduce the TCO of training clusters.

You can modify the applications to suit your business requirements,or submit a ticket to contact Alibaba Cloud for technical support.

After I purchase a persistent memory-optimized instance, how do I configure its persistent memory as local disks?

You can use tools to configure the persistent memory as local disks. For more information, see Configure persistent memory usage.

To which applications are local disks that can deliver higher performance applicable?

To optimize performance or costs for I/O-intensive applications, you can select persistent memory-optimized instances. Persistent memory-optimized instances can address common issues such as the following ones:

- The latency of single SQL queries is high, and SQL queries require a more consistent response time.
- It takes longer than required to load resources on frontend game servers, heavily loaded databases, and heavily loaded web servers.
- You purchased large cloud disks or local disks for their high IOPS and bandwidth but now have unused capacity that causes unnecessary costs.

You can use persistent memory-optimized instances in typical I/O-intensive scenarios such as the following ones:

Can I deploy Redis or MySQL applications on persistent memory-optimized instances whose persistent memory is used as local disks? Do the applications need to be modified in the same manner as they do when persistent memory is used as memory?

Yes, you can deploy Redis and MySQL applications on persistent memory-optimized instances whose persistent memory is used as local disks. You do not need to modify these applications before they can recognize the persistent memory as standard SSDs.

How is the performance of persistent memory used as local disks, compared with local non-volatile memory express (NVMe) SSDs and cloud disks?

The following table describes the performance comparison between local NVMe SSDs, enhanced SSDs (ESSDs), and persistent memory that is used as local SSDs.

> **Note** The performance data in the following table is for reference only. Data in the results of your own tests prevails.

Metric	Persistent memory of 128 GiB	NVMe SSD of 1,788 GiB	ESSD of 800 GiB at performance level 1 (PL1)
Read bandwidth	8 to 10 GB/s	2 to 3 GB/s	0.2 to 0.3 GB/s
Read/write bandwidth	8 to 10 GB/s	1 to 2 GB/s	0.2 to 0.3 GB/s
Write bandwidth	2 to 3 GB/s	1 to 2 GB/s	0.2 to 0.3 GB/s
Read IOPS	1,000,000	500,000	20,000 to 30,000
Read/write IOPS	1,000,000	300,000	20,000 to 30,000
Write IOPS	1,000,000	300,000	20,000 to 30,000
Read latency	300 to 400 nanoseconds	100,000 nanoseconds	250,000 nanoseconds
Write latency	300 to 400 nanoseconds	20,000 nanoseconds	150,000 nanoseconds

You can perform the following steps to test the performance of the persistent memory when it is used as local disks:

1. Configure persistent memory as local disks and attach the disks to instances.

   For more information, see Configure persistent memory usage.

2. Use disk performance test tools to test the performance of the persistent memory as local disks.

   For information about how to use fio to test disk performance in Linux, see the "Commands used to test the performance of local disks" section in Test the performance of EBS devices.

How reliable is persistent memory?

The reliability of data stored in persistent memory depends on the reliability of persistent memory devices and the physical servers to which these devices are attached. Risks of single points of failure exist. To ensure the reliability of application data, we recommend that you implement data redundancy at the application layer and use cloud disks for long-term data storage.

When persistent memory-optimized instances are released, data stored in their persistent memory is automatically cleared. We recommend that you back up data before you release the instances. It takes longer to release persistent memory-optimized instances than to release other types of instances.

What are the differences between ECS bare metal instances, traditional cloud hosts (virtual machines), and traditional physical machines?

For information about the differences, see Overview.

How is the network performance of ECS bare metal instances?

The network performance of ECS bare metal instances depends on their specifications. The higher their specifications are, the higher network performance the instances can deliver. For more information about the network performance of different ECS Bare Metal Instance types, see Instance family.

Which disk categories do ECS bare metal instances support? How many data disks can be attached to a single ECS bare metal instance?

ECS bare metal instances support ultra disks, standard SSDs, and enhanced SSDs (ESSDs). Up to 16 data disks can be attached to a single ECS bare metal instance.

Can the configurations of ECS bare metal instances be upgraded or downgraded? Do the instances support failover?

The configurations of ECS bare metal instances cannot be upgraded or downgraded. The instances do support failover. When the physical machine that hosts an ECS bare metal instance fails, the instance is failed over. Data is retained in the data disks of the instance.

How do I create an SCC instance?

You can create an SCC instance in one of the following ways:

- If you want to use only the remote direct memory access (RDMA) feature, log on to the ECS console and create an SCC instance. For more information, see Create an SCC instance.
- If you want to use the HPC scheduler and the cluster resizing feature in addition to RDMA, log on to the E-HPC console, create an SCC, and then create an SCC instance.

How are SCC instances billed?

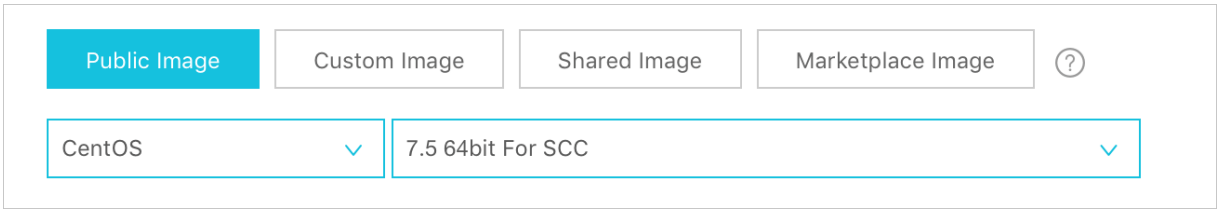
SCC instances can be billed on a weekly, monthly, or yearly subscription basis.

How do I create an SCC by using E-HPC?

You can log on to the E-HPC console or call the CreateCluster operation to create an SCC.

How do I use SCC RDMA?

When you create an SCC instance, select an image that is customized for SCC and supports RDMA RoCE drivers and OFED stacks. Then, you can use SCC RDMA by virtue of the IB verbs API or implement RDMA-based communication by using the MPI.



Which instance types support preemptible instances?

Instance types that support pay-as-you-go instances also support preemptible instances. If a preemptible instance of a specific instance type cannot be created due to insufficient resources, try another instance type.

In which regions can I create preemptible instances?

Preemptible instances can be created within all Alibaba Cloud regions. If preemptible instances cannot be created within a specific region due to insufficient resources, try another region.

When I attempt to purchase an ECS instance, the Preemptible Instance option is unavailable on the instance buy page. Why?

The availability of the Preemptible Instance option depends on your ECS usage. The Preemptible Instance option is available to only authorized users.

How do I bid for a preemptible instance?

To create a preemptible instance, you must specify a maximum hourly price to bid for spare resources of an instance type. If the spot price (current market price per hour) of the instance type does not exceed your specified maximum hourly price, the preemptible instance is created and billed based on the spot price. For more information, see Create a preemptible instance.

What is the relationship between the user-defined maximum hourly price and the spot price of a preemptible instance?

If the spot price does not exceed the user-defined maximum hourly price, the preemptible instance is created and billed based on the spot price. After a preemptible instance is created, it enters a protection period during which it cannot be automatically released due to insufficient resources or fluctuations in the spot price.

After the protection period ends, the system checks the spot price and resource availability of the instance type every 5 minutes. If the spot price exceeds the user-defined maximum hourly price or if resources for the instance type are insufficient, the preemptible instance that is in the Running state is released.

Can I view the spot price of an instance type when I purchase a preemptible instance?

Yes, when you create a preemptible instance in the ECS console, you can view the spot price and price history of each selected instance type. The total instance price includes fees for the instance type, storage, and bandwidth and is displayed in the lower part of the instance buy page. The instance type fee is the spot price of the selected instance type.

Can I view the price history of a preemptible instance type? How?

Yes, you can select an instance type to view its price history when you attempt to create a preemptible instance in the ECS console. You can also call the DescribeSpotPriceHistory operation to view the price history of a preemptible instance type.

How many preemptible instances can be purchased by a single account?

vCPU-based quotas apply to preemptible instances within each account, whereas instance-based quotas do not. When you create a preemptible instance, you can view the available vCPU-based quota after you select an instance type. For more information, see the "Instance limits" section in Limits.

How do I increase my vCPU-based quotas?

To increase your vCPU-based quotas, submit a ticket.

Can preemptible instances be converted into subscription instances?

No, preemptible instances cannot be converted into subscription instances.

Can the instance type of a preemptible instance be changed?

No, the instance types of preemptible instances cannot be changed.

A preemptible instance does not meet my requirements. How do I upgrade it to a higher-specification instance type at a low cost?

The instance types of preemptible instances cannot be changed. We recommend that you create a custom image based on the system disk of your preemptible instance and then use the custom image to create a preemptible instance of a higher-specification instance type. For more information, see Create a custom image from a snapshot and Create an ECS instance by using a custom image.

If I select Use Automatic Bid (SpotAsPriceGo) when I create a preemptible instance, will the created instance be automatically released due to insufficient resources?

When you select Use Automatic Bid (SpotAsPriceGo), the spot price is used as the bid price and can raise as high as the pay-as-you-go price for the selected instance type. This ensures that the created preemptible instance will not be automatically released due to fluctuations in the spot price. However, the instance may still be automatically released due to insufficient resources for the instance type.

If I set the maximum hourly price to the pay-as-you-go price for the selected instance type when I create a preemptible instance, may the instance be automatically released?

You can set the maximum hourly price to the pay-as-you-go price for your selected instance when you create a preemptible instance. This ensures that the created instance will not be automatically released due to fluctuations in the spot price. However, the instance may still be automatically released due to insufficient resources for the instance type.

Can I cancel or reschedule the automatic release of a preemptible instance?

Yes. Before your preemptible instance is released, you can cancel or reschedule the automatic release of the preemptible instance at anytime.

I do not have overdue payments in my account. Why have my preemptible instances been released?

After the protection period of a preemptible instance ends, if the spot price exceeds the user-defined maximum hourly price or if resources are insufficient, the instance is automatically released.

How are release rates on the instance buy page calculated? Why are most of them in the range of 0% to 3%?

Release rates vary based on zones and instance types. The single-day release rate of an instance type within a zone is calculated based on the following formula: Single-day number of released preemptible instances/Single-day total number of preemptible instances. Most release rates are in the range of 0% to 3% as expected. Alibaba Cloud continues our efforts to reduce release rates and ensure higher availability of your preemptible instances.

Am I notified when my preemptible instances are about to be released? How am I notified?

Yes, you are notified when your preemptible instances are about to be released. When your preemptible instance needs to be released due to a spot price change or insufficient resources, the instance first enters the To Be Released state and is automatically released in 5 minutes.

You can use CloudMonitor to subscribe to notifications for interrupted preemptible instances. For more information, see Configure event notifications.

To check whether an instance is in the To Be Released state, you can view the instance metadata, or call the DescribeInstances operation and view the returned OperationLocks data. For more information, see Overview of ECS instance metadata and DescribeInstances.

Can the data of a preemptible instance be automatically retained when the instance is released?

No, the data of a preemptible instance cannot be automatically retained when the instance is released. When a preemptible instance is no longer needed, we recommend that you create snapshots to back up your data and environment and then release the instance. You can purchase preemptible instances at anytime. For more information, see Create a snapshot of a disk.

How do I retain the public IP address of a preemptible instance when the instance is released?

When a preemptible instance is released, its public IP address is reclaimed. If you want to retain the IP address, we recommend that you use Elastic IP Address (EIP). For more information, see Elastic IP addresses.

You can convert the public IP address into an EIP. For more information, see Convert the public IP address of a VPC-type instance to an EIP and Convert the public IP address of an instance in the classic network into an EIP.

How is a preemptible instance billed if its lifespan is shorter than 1 hour?

If a preemptible instance is released shortly after it is created and its lifespan is shorter than 1 hour, the instance is billed based on its lifespan in seconds. The hourly price of the instance is the spot price that was in effect when the instance was created. The per-second price of the instance is calculated by using the following formula: Per-second price = Hourly price/3600. The fees for a preemptible instance from creation to release are accurate to two decimal places. Accrued costs of less than USD 0.01 for a preemptible instance are not charged.

To what resources are the prices of preemptible instances applicable?

The prices of preemptible instances are applicable to only instance types. You are charged for other resources such as system disks, data disks, and network bandwidth at the same prices as those of pay-as-you-go instances.

Why do preemptible instances of the same instance type and region vary widely in spot prices across different zones?

Price differences are caused by resource stock differences. Each zone corresponds to different data centers. Within each region, data centers for zones may vary in terms of construction conditions, capacity, sales, and deployment policies. Resource inventories for each instance type may also vary across different zones. The spot prices of preemptible instances fluctuate with supply and demand changes and therefore are different across zones.

May the spot price of a preemptible instance exceed the pay-as-you-go price?

No, the spot price of a preemptible instance can be up to but cannot exceed the pay-as-you-go price. You do not need to worry about being charged more for a preemptible instance than for a pay-as-you-go instance with the same configurations.

For the first hour after a preemptible instance is created, am I charged at a price that fluctuates with the spot price?

No, the hourly price of the first instance hour for a preemptible instance is set at the beginning of the hour and remains in effect for the whole hour.

Do I continue to be charged for preemptible instances after they are stopped?

Yes, you continue to be charged for your preemptible instances after they are stopped. When a preemptible instance is no longer needed, we recommend that you create snapshots to back up your data and environment and then release the instance. You can purchase preemptible instances at anytime.

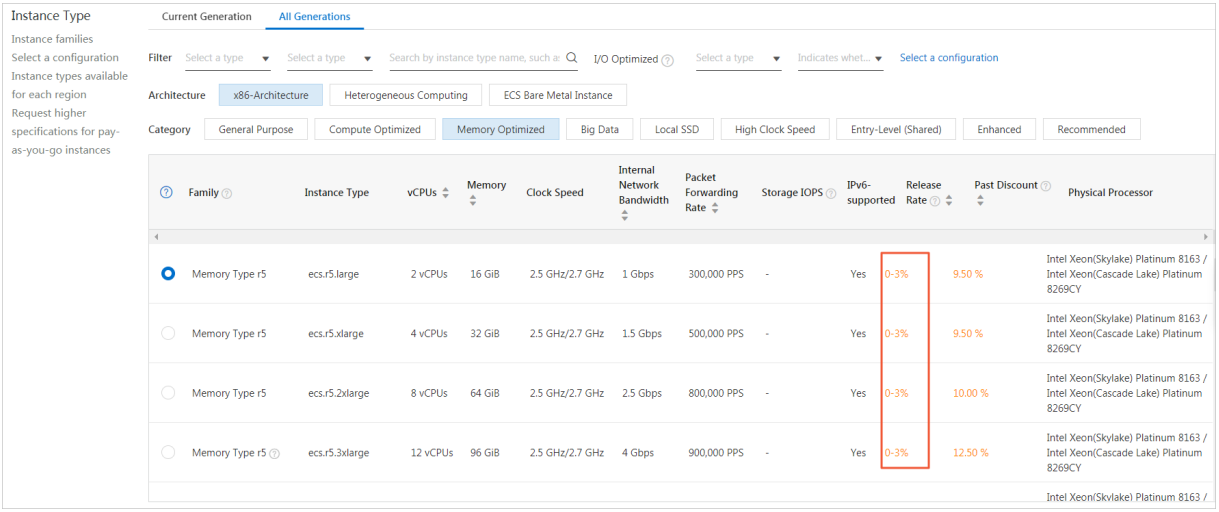
> **Note** You continue to be charged for preemptible instances after you stop them by using the ECS console or by calling the StopInstance operation.

Which is more cost-effective: a preemptible instance with a protection period or a preemptible instance without a protection period?

Preemptible instances without a protection period are more cost-effective and have a discount of about 10% off compared with preemptible instances with a protection period.

Is the release rate of preemptible instances without a protection period higher than that of preemptible instances with a protection period?

You can check the release rate of each instance type on the instance buy page in the ECS console regardless of whether you attempt to create preemptible instances with or without a protection period. For each instance type, the release rate is determined by the bidding policy and the supply-demand relationships of resources.



Are preemptible instances without a protection period prioritized for release over those with a protection period?

No, preemptible instances without a protection period are not prioritized for release over those with a protection period.

If I have created a preemptible instance without a protection period, do subsequently created preemptible instances also not have a protection period?

No, preemptible instances are created without a protection period only if you choose not to configure a protection period for them. By default, each preemptible instance is created with a protection period of 1 hour.

> **Note** You can configure a protection period for preemptible instances only when you create them.

If a preemptible instance does not have a protection period, do I still receive a notification 5 minutes in advance before the instance is released?

Yes, you still receive a notification 5 minutes in advance before a preemptible instance is released even if it does not have a protection period.

Are fewer resources provided for preemptible instances with a protection period than for those without a protection period?

No, the same number of resources are provided for preemptible instances with and without a protection period.

Can I convert between preemptible instances with and without a protection period?

No, you cannot convert between preemptible instances with and without a protection period. By default, each preemptible instance is created with a protection period of 1 hour. You can configure a protection period for preemptible instances only when you create them. The protection period settings of preemptible instances cannot be changed after the instances are created.

What are reserved instances?

A reserved instance is a discount coupon that can be automatically applied to one or more pay-as-you-go instances, excluding preemptible instances. Reserved instances can also be used to reserve instance resources. A combination of reserved instances and pay-as-you-go instances provides higher cost-effectiveness and a higher degree of flexibility as compared with subscription instances.

Do reserved instances provide reserved resources?

Zonal reserved instances provide reserved resources, but regional reserved instances do not.

What operating systems do reserved instances support?

Reserved instances support both Windows and Linux. For example, you can purchase a reserved Linux instance, and this reserved instance can be applied to pay-as-you-go Linux instances that match its attributes regardless of the image type (public images, custom images, shared images, or Alibaba Cloud Marketplace images).

To apply a reserved instance to pay-as-you-go instances created from Bring Your Own License (BYOL) images, submit a ticket.

Which instance families do reserved instances support?

Reserved instances support the following instance families:

- General-purpose instance families: g7, g6e, g6, g5, g5ne, and sn2ne
- Compute-optimized instance families: c7, c6e, c6, c5, ic5, and sn1ne
- Memory-optimized instance families: r7, r6e, r6, r5, re6, re4, and se1ne
- Big data instance family: d2s
- Instance families with local SSDs: i3, i3g, i2, i2g, and i2gne
- Instance families with high clock speeds: hfg7, hfc7, hfr7, hfg6, hfc6, hfr6, hfg5, and hfc5
- GPU-accelerated compute-optimized instance families: gn7, gn6i, gn6e, gn6v, gn5, and gn5i
- ECS Bare Metal Instance families: ebmgn7, ebmgn6i, ebmgn6e, ebmg6, ebmc6, ebmr6, ebmhfg6, ebmhfc6, and ebmhfr6
- Burstable instance families: t6 and t5

Can reserved instances be applied to preemptible instances?

No, reserved instances cannot be applied to preemptible instances.

Can the instance families of reserved instances be changed?

No, the instance families of reserved instances cannot be changed.

To what scenarios are zonal reserved instances applicable?

We recommend that you purchase zonal reserved instances when you have clear requirements to reserve resources.

To what scenarios are regional reserved instances applicable?

We recommend that you purchase regional reserved instances if you want to have a higher degree of zone flexibility or instance size flexibility.

How is the zone flexibility of reserved instances applied?

Only regional reserved instances provide zone flexibility. Example:

Assume that you are running the following pay-as-you-go instance:

One ecs.c5.xlarge Linux instance in Qingdao Zone B. This instance is named C5PAYG-b.

Assume that you have purchased the following reserved instance:

One regional ecs.c5.xlarge reserved instance in the China (Qingdao) region. This instance is named C5RI.

C5RI is matched to C5PAYG-b.

You release C5PAYG-b and create another pay-as-you-go Linux instance of the same instance type named C5PAYG-c in Qingdao Zone C. C5RI is matched to C5PAYG-c.

How is the instance size flexibility of reserved instances applied?

Only regional reserved instances provide instance size flexibility. Example:

Assume that you have one regional ecs.g5.4xlarge reserved instance. It can be applied to one ecs.g5.4xlarge pay-as-you-go instance, two ecs.g5.2xlarge pay-as-you-go instances, or four ecs.g5.xlarge pay-as-you-go instances.

Assume that you have a one-year regional ecs.g5.xlarge reserved instance. It can be applied to offset the hourly bills of an ecs.g5.xlarge pay-as-you-go instance or 50% of the hourly bills of an ecs.g5.2xlarge pay-as-you-go instance for one year.

Do zonal reserved instances provide instance size flexibility?

No, zonal reserved instances do not provide instance size flexibility. A zonal reserved instance can be applied to only pay-as-you-go instances of the same instance type.

Do zonal reserved instances provide zone flexibility?

No, zonal reserved instances do not provide zone flexibility. A zonal reserved instance can be applied only to pay-as-you-go instances in the same zone as it.

Can a zonal reserved instance be changed into a regional one?

Yes, a zonal reserved instance can be changed into a regional one. The scope of a reserved instance can be changed in the following ways:

- From a zone to a region
- From a region to a zone

- From one zone to another within the same region for a zonal reserved instance

Can the scope of a reserved instance be changed from one region to another?

No, the scope of a reserved instance cannot be changed from one region to another. For example, assume that you have a zonal reserved instance scoped to Hangzhou Zone B. You can change the instance scope to another zone within the China (Hangzhou) region or change the instance into a regional reserved instance scoped to the China (Hangzhou) region. However, you cannot change the scope of the instance to a zone of another region or change the instance into a regional reserved instance scoped to another region.

Can reserved instances be used across accounts?

No, reserved instances cannot be used across accounts.

Can reserved instances be used to cover the storage and network usage charges of pay-as-you-go instances?

No, reserved instances cannot be used to cover the storage or network usage charges of pay-as-you-go instances. Reserved instances are applied to cover the charges for vCPUs and memory of pay-as-you-go instances. For pay-as-you-go Windows instances, reserved instances can also cover image charges.

Can I configure a reserved instance to be applied to a specific pay-as-you-go instance?

No, you cannot configure a reserved instance to be applied to a specific pay-as-you-go instance. When multiple pay-as-you-go instances match the attributes of a reserved instance, the reserved instance is applied based on the optimized matching scheme.

How are reserved instances billed?

Reserved instances are billed separately and support the All Upfront, Partial Upfront, and No Upfront payment options.

The term of a reserved instance begins at the time when the instance is purchased. You are charged based on the payment option that you selected regardless of whether the reserved instance is matched to pay-as-you-go instances. The All Upfront option is most cost-effective.

When does a reserved instance take effect after it is purchased?

A reserved instance takes effect and is billed starting on the hour when the reserved instance is purchased. The reserved instance expires at 00:00:00 on the day after its term end date. For example, assume that you purchase a reserved instance with a one-year term at 13:45:00 on February 26, 2019. The reserved instance takes effect and is billed starting from 13:00:00 on February 26, 2019. The reserved instance expires at 00:00:00 on February 27, 2020. If you already have eligible pay-as-you-go instances when you purchase a reserved instance, the reserved instance is applied to offset the bills generated by the pay-as-you-go instances starting from the hour of 13:00:00 to 14:00:00 on February 26, 2019 until the reserved instance expires.

After I modify, split, or merge a reserved instance, when does the operation take effect?

When reserved instances are modified, split, or merged, new reserved instances are generated and the original ones become invalid. The new reserved instances take effect and the original ones become invalid. Both occur on the top of the hours of the modification, splitting, or merging operations. For example, assume that you split one ecs.g5.2xlarge zonal reserved instance RI1 into two zonal ecs.g5.xlarge reserved instances RI2 and RI3 at 13:45:00 on February 26, 2019. At 13:00:00 on February 26, 2019, RI1 becomes invalid and RI2 and RI3 take effect. Starting from 13:00:00 on February 26, 2019, the instance type eligible for resource reservation and billing discounts is ecs.g5.xlarge, not ecs.g5.2xlarge anymore. If RI2 and RI3 are matched to pay-as-you-go instances immediately after they take effect, RI2 and RI3 are also applied to offset the hourly bills of ecs.g5.xlarge pay-as-you-go instances starting from 13:00:00 on February 26, 2019.

Why is the No Upfront payment option not displayed on the buy page?

The availability of this option depends on your ECS usage.

Can the payment option of a reserved instance be changed?

No, the payment options of reserved instances cannot be changed.

Can reserved instances be resold?

No, reserved instances cannot be resold.

Can I use reserved instances to cover the image fees of pay-as-you-go Windows instances?

Yes, reserved instances can be applied to cover the image fees of pay-as-you-go Windows instances. This is because reserved Windows instances already include Windows images at no additional cost.

Can reserved instances be applied to cover the image fees of pay-as-you-go Linux instances?

No, reserved instances cannot be applied to cover the image fees of pay-as-you-go Linux instances.

Are the consumption details of reserved instances refreshed every hour?

Yes, the consumption details of reserved instances are refreshed every hour.

Can a reserved instance be applied to more than one pay-as-you-go instance at a time?

Yes, a reserved instance can be applied to more than one pay-as-you-go instance at the same time. The reserved instance checks for eligible pay-as-you-go bills on an hourly basis and deducts fees based on its computing power.

> **Note** The computing power and term of each reserved instance are fixed. You cannot increase the computing power of a reserved instance by shortening its term.

For example, assume that you have a reserved instance with the following attributes:

- Instance type: c5.large

- Instances: 1 (indicating that the reserved instance can match one pay-as-you-go instance of the specified instance type)
- Term: one year

The following examples demonstrate how the reserved instance is applied based on the pay-as-you-go instances that exist:

- Assume that six c5.large pay-as-you-go instances exist for 1 hour each. Each of these pay-as-you-go instances consumes 1 hour of computing power equal to that delivered by the c5.large reserved instance every hour. The reserved instance is randomly applied to one of the pay-as-you-go instances. You cannot configure the reserved instance to be applied to all six pay-as-you-go instances by shortening the term of the reserved instance to two months.
- Assume that six c5.large pay-as-you-go instances exist for 10 minutes each. The six instances consume 10 minutes of computing power each and in total consume the amount of computing power that the c5.large reserved instance can deliver every hour. The reserved instance is applied to all six pay-as-you-go instances.
- Assume that six c5.large pay-as-you-go instances exist for 15 minutes each. The six instances consume 15 minutes of computing power each, in total exceeding the amount of computing power that the c5.large reserved instance can deliver every hour. The reserved instance is randomly applied to the pay-as-you-go instances to offset the charges for 1 hour of computing power.

When I attempted to connect to a Linux instance by using Workbench, the connection failed and I received a timeout prompt. Why?

This may be caused by one of the following issues. Troubleshoot these issues:

- The instance is not in the Running state.
- No rules in the security groups of the instance allow traffic on the specified port. For more information about security group rules, see Connect to a Linux instance by using a password or key.

When I attempted to connect to a Linux instance by using Workbench, the connection failed and I prompted that the connection was denied. Why?

This may be caused by one of the following issues. Troubleshoot these issues:

- The remote service that corresponds to SSH or RDP is not enabled on the instance.
- Ports used for remote connections, such as SSH port 22 and RDP port 3389, are not enabled on the instance.

When I attempted to connect to a Linux instance by using Workbench, the connection failed and I was prompted with a username or password error. Why?

The username or password that you entered is incorrect. Perform the following operations:

- Enter the correct username. The username of Linux instances is typically root.
- Enter the correct password. If you forget your password, you can reset it. For more information, see Reset the logon password of an instance.
- To use the root username to log on to the Linux instance, make sure that the */etc/ssh/sshd_config* file contains the `PermitRootLogin yes` setting. For more information, see Connect to a Linux

instance by using a password or key.

When I attempted to connect to a Windows instance by using Workbench, the connection failed and I received a timeout prompt. Why?

This may be caused by one of the following issues. Troubleshoot these issues:

- The instance is not in the Running state.
- No rules in the security groups of the instance allow traffic on the specified port. For more information about security group rules, see Connect to a Windows instance by using a password or key.
- The remote service that corresponds to SSH or RDP is not enabled on the instance.
- Ports used for remote connections, such as SSH port 22 and RDP port 3389, are not enabled on the instance.

When I attempted to connect to a Windows instance by using Workbench, the connection failed and I was prompted with a username or password error. Why?

The username or password that you entered is incorrect. Perform the following operations:

- Enter the correct username. The username of Windows instances is typically administrator.
- Enter the correct password. If you forget your password, you can reset it. For more information, see Reset the logon password of an instance.
- If you log on to the Windows instance as a non-administrator user, the user must belong to the Remote Desktop Users group.

Does a VNC management terminal allow multiple users to log on simultaneously?

No, a VNC management terminal allows a single user to log on at a time.

What do I do if I forget the connection password?

You can reset your connection password. For more information, see Change the VNC password.

Why am I unable to connect to a VNC management terminal even after I reset my connection password?

If the instance to which you are connecting is not I/O optimized, you must restart the instance by using the ECS console or by calling the RebootInstance operation for the new password to take effect.

> **Note** If you only restart the instance from within the instance, the new password does not take effect.

I was prompted with an authentication failure when I attempted to connect to a VNC management terminal. What do I do?

The authentication failure may occur due to an incorrect password. Perform the following troubleshooting operations:

1. Enter the correct connection password.
2. If you forget your password, you can reset it and try again. For more information, see Change the VNC password.

> **Note** If the instance to which you are connecting is not I/O optimized, you must restart the instance by using the ECS console or by calling the RebootInstance operation for the new password to take effect.

What do I do if a black screen appears while I am connected to a VNC management terminal?

A black screen indicates that the instance is in sleep mode. Perform the following operations based on your operating system:

- For a Linux instance, click your mouse or press any key to activate the instance and go to the logon page.
- For a Windows instance, choose **Send Remote Call** > **CTRL+ALT+DELETE** in the upper-left corner to go to the logon page.

What do I do if a VNC management terminal cannot be accessed?

You can use a browser to access the VNC management terminal for troubleshooting. For example, you can use Google Chrome to access the VNC management terminal and press the F12 key to open the developer tools panel. Then, click the **Console** tab and identify errors based on the information displayed.

Why am I unable to use Internet Explorer 8.0 to access a VNC management terminal?

VNC management terminals support Internet Explorer 10 and later.

We recommend that you use Google Chrome because it is most compatible with the Alibaba Cloud Management Console.

When I use Firefox to access a VNC management terminal, an error message is returned indicating that a secure connection cannot be established. What do I do?

This problem occurs if the encryption algorithm used by your version of Firefox is different from that used by the VNC management terminal.

We recommend that you use Google Chrome because it is most compatible with the Alibaba Cloud Management Console.

How do I connect to a Linux instance?

Linux instances support SSH for connection. You can use one of the following methods to connect to a Linux instance:

- Connect to a Linux instance by using password authentication (with VNC)
- Connect to a Linux instance by using a password or key (with Workbench)
- Connect to a Linux instance by using a password (with third-party client tools)
- Connect to a Linux instance by using an SSH key pair (with third-party client tools)

- Connect to a Linux instance from a mobile device (with third-party client tools)

What are the default username and password used to connect to the operating system of an ECS instance?

The default username varies per operating system type.

- For a Windows instance, the default username is `administrator`.
- For a Linux instance, the default username is `root`.

The password used to connect to the operating system of an instance is set by yourself when you create the instance. For more information, see Create an instance by using the wizard. If you forget the password, you can reset it. For more information, see Reset the logon password of an instance.

> **Note** This password is used to connect to the instance operating system, not to VNC management terminals.

How do I adjust the desktop resolution of a Windows instance?

You can use one of the following methods to adjust the desktop resolution of a Windows instance:

- Connect to the instance by using a VNC management terminal and adjust the desktop resolution on the instance.

  The following example demonstrates how to adjust the desktop resolution of an instance that runs a `Windows Server 2019 Datacenter 64-bit` operating system:

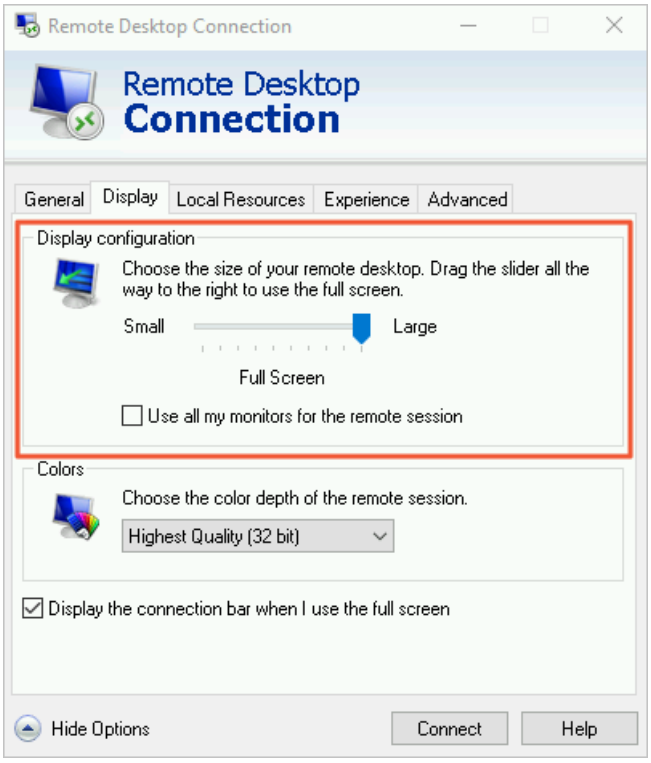
  i. Connect to the instance by using a VNC management terminal in the ECS console. For more information, see Connect to a Windows instance by using a password.

  ii. On the Windows desktop, right-click a blank area and select **Display settings**.

  iii. In the **Scale and layout** section, adjust the resolution.

- Use the Remote Desktop Connection (RDC) tool on your computer.

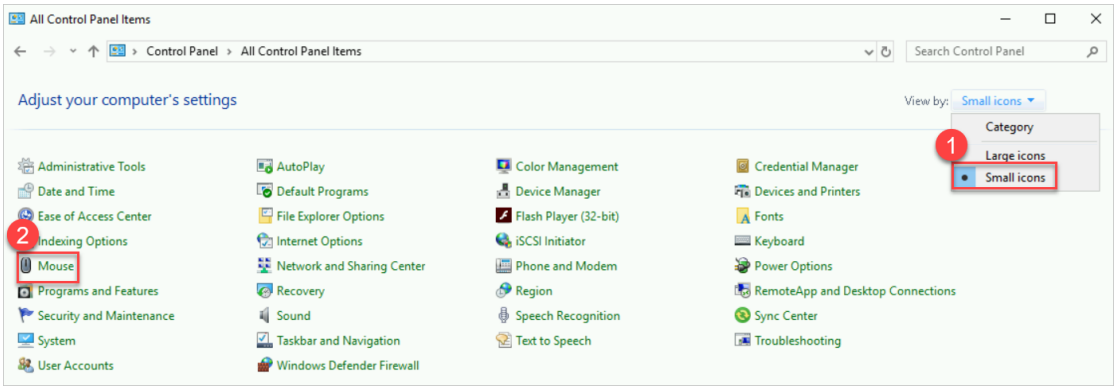
You cannot adjust the desktop resolution of a Windows instance after you connect to the instance from your computer by using RDC. You must adjust the display settings in RDC on your computer, and connect to the instance by using RDC to apply the new display settings to the instance.



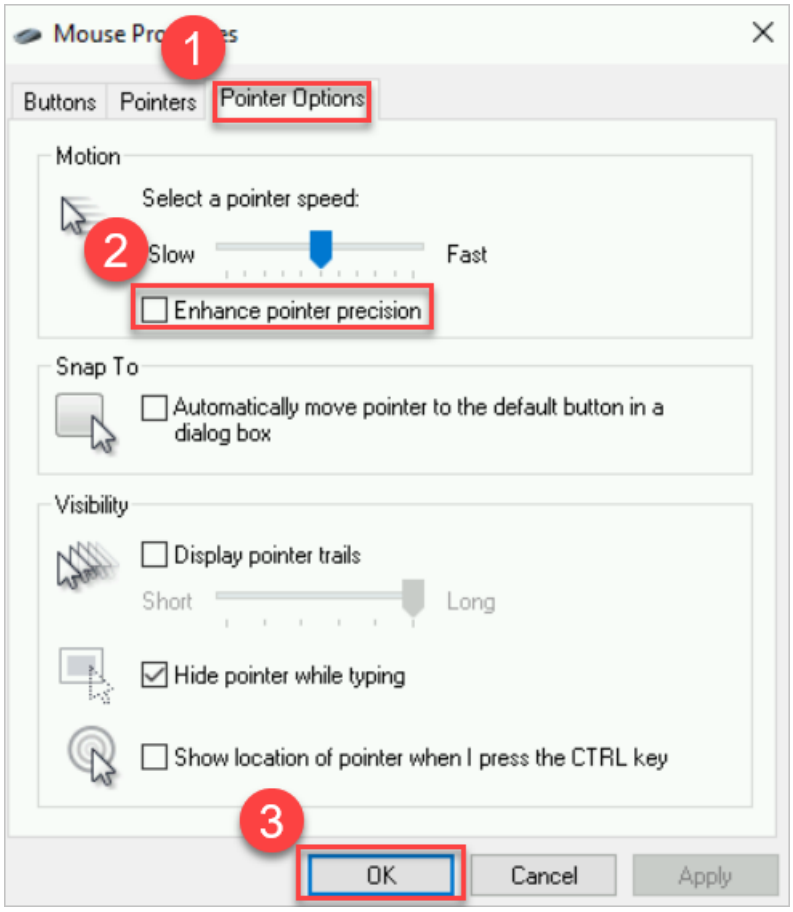
Why do two cursors appear after I log on to a Windows instance by using a VNC management terminal?

If two cursors appear when you use a VNC management terminal to log on to a Windows instance of the g7, c7, or r7 instance family, perform the following steps to modify mouse settings:

1. Modify mouse settings in Control Panel.
   i. Open Control Panel, set View by to **Small icons**. Then, click **Mouse**.



ii. In the **Mouse Properties** dialog box, click the **Pointer Options** tab, clear **Enhance pointer precision**, and then click **OK**.

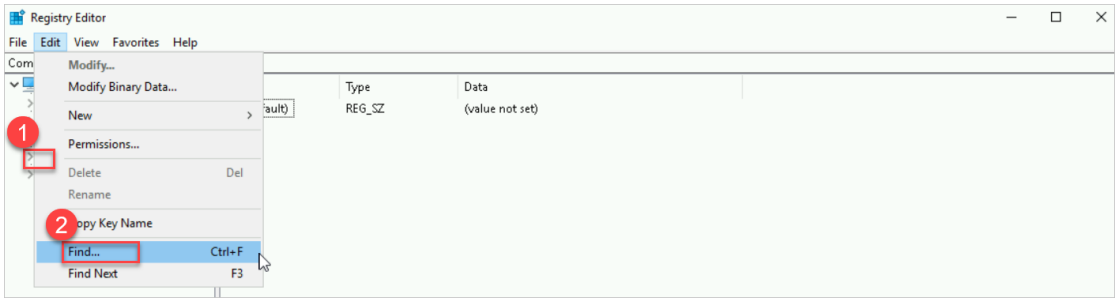


2. Modify mouse settings in the registry.

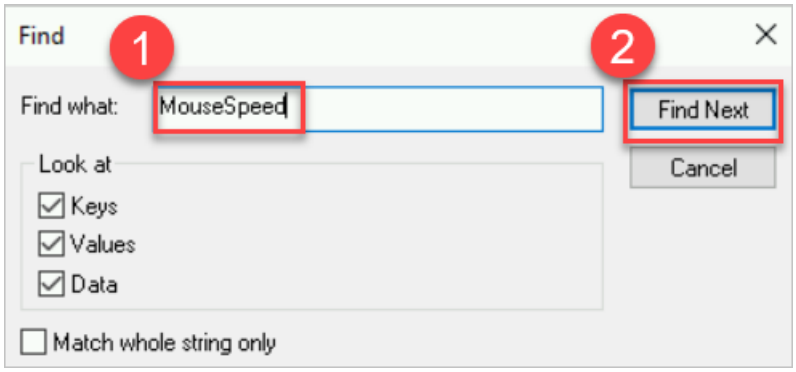
> **Note** If two cursors do not appear in the user logon window, skip this step.

i. Open Registry Editor.

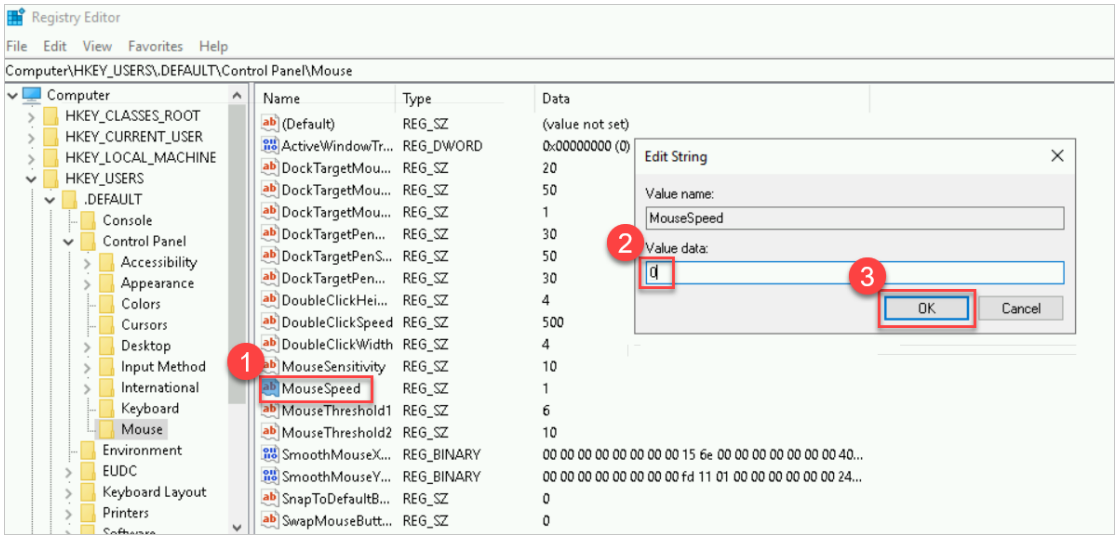
ii. In the left-side navigation pane, click **HKEY_USERS** and choose **Edit > Find** in the top navigation bar.



iii. In the Find dialog box, enter MouseSpeed and click **Find Next**.



iv. Double-click **MouseSpeed**. In the Edit String dialog box, set **Valid data** to **0** and click **OK**.



3. Use a VNC management terminal to log on to the Windows instance again.

What do I do if a Data Plane Development Kit (DPDK) application cannot be deployed on an ECS instance?

We recommend that you deploy DPDK applications on g5ne instances. For more information about the g5ne instance types, see Instance family.

If you deploy a DPDK application on an instance of a sixth-generation (such as g6, c6, or r6) instance family or later, an exception may occur when the DPDK application runs. For example, when you use Pktgen-DPDK to test the packet forwarding rate of an instance, the igb_uio port to which the network interface controllers (NICs) are bound may not be detected and the following error is reported:

```
 EAL: eal_parse_sysfs_value(): cannot open sysfs value
/sys/bus/pci/devices/0000:00:06.0/uio/uio0/portio/port0/start
```

We recommend that you use one of the following methods to solve the issue:

- Upgrade DPDK to DPDK Release 21.05 or later. This method is applicable to x86-based instances. For more information about DPDK versions, visit the official DPDK website.
- Replace UIO drivers with VFIO drivers. This method is applicable to both x86-based instances and ARM-based instances. For more information, see Replace UIO drivers with VFIO drivers.

Can I upgrade the instance types and other configurations of subscription instances?

Yes, you can upgrade the instance types and other configurations of subscription instances. For more information, see Upgrade the instance types of subscription instances.

Can I upgrade the instance types and other configurations of pay-as-you-go instances?

Yes, but you must stop pay-as-you-go instances before you can upgrade their instance types and other configurations. You can upgrade the instance types and other configurations of pay-as-you-go instances by following the instructions in Change the instance type of a pay-as-you-go instance or by calling the ModifyInstanceSpec operation.

How long does it take to upgrade the configurations of an instance?

- Subscription instances do not need to be stopped for their instance types to be upgraded. It takes about 15 minutes to upgrade the instance type of a subscription instance.
- Pay-as-you-go instances must be stopped for their instance types to be upgraded. It also takes about 15 minutes to upgrade the instance type of a pay-as-you-go instance.
- You can upgrade the bandwidths of instances without stopping the instances. The upgrade process takes about 5 minutes.

How is the fee for an instance configuration upgrade calculated?

The fee and its calculation method are displayed in the ECS console when you upgrade the instance type or other configurations of an instance. You can also view the billing details on the Account Overview page.

Are my cloud service configurations affected if I upgrade the configurations of ECS instances?

Pay-as-you-go instances must be stopped before their configurations can be upgraded. After you upgrade the configurations of a subscription instance, you must restart the instance for the new configurations to take effect. The upgrade operation interrupts the services that run on the instance for a short period of time. We recommend that you upgrade the configurations of instances during off-peak hours. Instances can seamlessly resume services after upgrades without the need to reconfigure environments.

How do I upgrade ECS resources?

For information about how to upgrade ECS resources, see Overview of instance configuration changes.

- ECS instances, except those that use local storage, can have their CPU and memory resources scaled and their bandwidths upgraded while the instances are running. After post-upgrade configurations of instances take effect, you can also downgrade the configurations of the instances.
- Up to 16 data disks can be attached to each ECS instance. You can extend cloud disks. Cloud disks cannot be shrunk after they are extended.
- The bandwidth of each ECS instance is measured in Mbit/s and can range from 0 Mbit/s to 200 Mbit/s. You can modify the bandwidth or change the billing method for network usage.

I have upgraded the configurations of an instance but no changes have taken effect. Why?

After you upgrade the configurations of an instance, you must restart the instance by using the ECS console or by calling an API operation for the new configurations to take effect.

After I place an order to upgrade the configurations of an instance, can I cancel the order to restore the instance to its original configurations?

No, after an order to upgrade the configurations of an instance takes effect, the order cannot be canceled and the configurations of the instance are upgraded. If you want to restore the instance to its original configurations, you can downgrade its configurations. You are charged for the configuration downgrade.

What do I do if I cannot access a website that runs on an ECS instance?

For information about how to resolve this problem,see You cannot access websites that run on ECS instances.

If the problem persists, submit a ticket.

My ECS instance was stuck in the Starting state, and AliyunService was disabled or deleted. What do I do?

Problem description: After an ECS instance was started, it remained in the Starting state for an extended period of time and then automatically stopped. You logged on to the instance and found that AliyunService was deleted or disabled in the system services.

Solution:

- If AliyunService was disabled, perform the following operations:
  i. Change the state of AliyunService to automatic.
  ii. Restart the instance.
- If AliyunService was deleted, perform the following operations:
  i. Run the following command to add AliyunService to the instance:

```
sc create AliyunService type= "own" start= "auto" binPath= "C:\Program Files\AliyunService\AliyunService.exe -d" tag= "no" DisplayName= "AliyunService"
```

> **Note** Make sure that you leave a space after each equal sign `(=)`.

  ii. Find the registry key *HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\services\AliyunService*, and change `c:\Program Files\AliyunService\AliyunService.exe -d` to `"c:\Program Files\AliyunService\AliyunService.exe" -d`.
  iii. Restart the instance.

How do I use f1 instances?

After you create an f1 instance, Alibaba Cloud shares an FPGA development image to you. Only CentOS 7u2 images are supported. The FPGA development image includes the complete Intel Quartus development suite and the f1 instance constraint files to provide a complete cloud development environment.

> Note

Basic procedure to use an f1 instance:

1. After the development phase is complete, you can generate an intermediate QAR file during the compilation stage and upload the file to an Object Storage Service (OSS) bucket. Currently, this file can be uploaded only to an OSS bucket within the China (Hangzhou) region. Then, you can call an API operation to register the QAR file information with Alibaba Cloud.

   We recommend that you use the free Intel Quartus development suite to complete development, compilation, and simulation operations on the cloud.

2. Alibaba Cloud verifies your registration request about the QAR file and sends you an email that includes an FPGA image ID if your request is determined as valid.
3. To deploy the image, you can call an API operation with the f1 instance ID and FPGA image ID specified to associate the instance with the image.

   You can initiate the association operation in all scenarios where the ECS API is available.

   - If the f1 instance has never been associated with an FPGA image, initiate the association operation.
   - If the f1 instance was previously associated with an FPGA image and had the image loaded, erase the FPGA image from the f1 instance before you initiate the association operation.

4. After you associate the FPGA image with the instance, call an API operation to load the image.

   You must initiate the load operation from the f1 instance. Then, the underlying service of Alibaba Cloud burns the associated FPGA image to the corresponding FPGA on the instance.

If you want to restore the f1 instance to its initial state, call an API operation to erase the burned FPGA image from the f1 instance.

For more information about how to manage f1 instances, see the following topics:

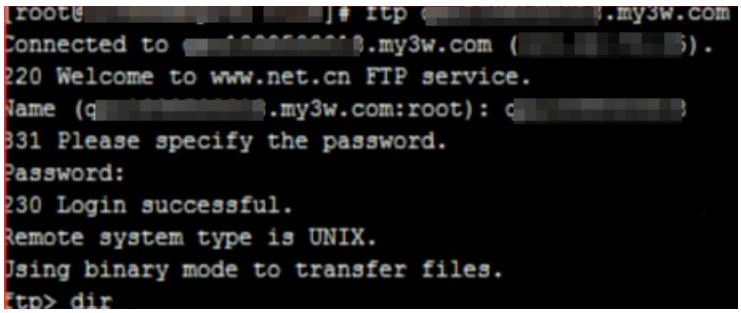
- Create an f1 instance
- Use OpenCL on an f1 instance
- Use RTL Compiler on an f1 instance

How do I use FTP tools in macOS to upload files?

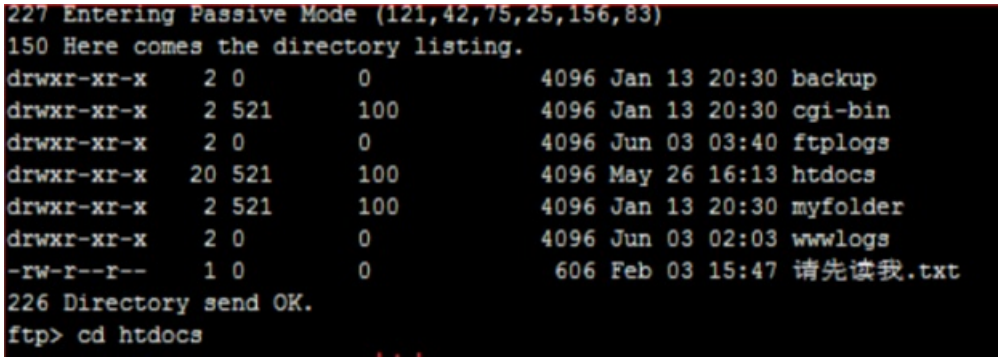
Method 1: Use the Terminal of macOS to upload files

Open the Terminal in macOS or iTerm2 for macOS. To download iTerm2, visit iTerm2. Make sure that you select the correct destination path.

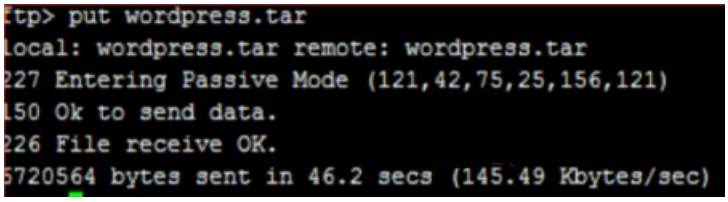
1. Connect to the FTP server.



2. Access the destination directory. In Windows, use the working directory as the destination directory. In Linux, switch to the *htdocs* directory and use it as the destination directory.



3. Run the **put** command to upload files.



Method 2: Use a third-party tool to upload files

1. Download Yummy FTP.
2. Install Yummy FTP.
3. Enter the server IP address, username, and password. Set Protocol to Standard (FTP). Set Port to 21 or a different port number that you are using, and leave the SSH key field unselected.
4. Click **Connect**.
5. In the right-side pane, select the destination directory. In Windows, use the current working directory. In Linux, select the htdocs directory. In the left-side pane, select files. Then, click the Upload icon to upload the files.

   If you attempt to install Yummy FTP and are prompted with a message similar to "Your security preferences only allow the installation of applications from the Mac App Store and authorized developers", perform the following steps:

   i. Choose **System Preference > Security and Privacy**.

   ii. Click the security lock in the lower-left corner of the window and enter the administrator password.

   iii. Set **Allow apps download from** to **Anywhere**.

   Then, you can use Yummy FTP to upload files.

If you have further questions, submit a ticket.

How do I apply for an ICP filing for my domain name after I purchase an ECS instance?

Before you apply for an ICP filing, make sure that your ECS instance meets the ICP filing requirements. For more information, see Prepare and check the instance and access information.

An ECS instance cannot load the kernel to start. What do I do?

Problem description: The system does not respond when you select an option from the GRUB menu on system startup. After you have mounted the LiveCD image to the ECS instance, you can log on to the instance and confirm that the file system permissions are correct and that message logs show no exceptions.

Cause: The system has been attacked by ransomware.

Solution: Back up your data and re-initialize the system.

How do I change the logon password from within an instance?

For information about how to change the logon password from within an instance, see Change the logon password of an instance by connecting to the instance.

Why am I unable to add sound or video cards to ECS instances?

Alibaba Cloud ECS instances are not multimedia servers and do not provide sound or video card components. Sound or video cards cannot be added to ECS instances.

Can I transfer the unused time of an ECS instance to another ECS instance?

No, the unused time of an ECS instance cannot be transferred to other ECS instances. If you want both high flexibility and cost-effectiveness, we recommend that you use a combination of reserved instances and pay-as-you-go instances. For more information, see Overview.

Do ECS instances provide databases by default?

No, ECS instances do not provide databases by default. To use database services, perform one of the following operations:

- Deploy your own database.
- Purchase ApsaraDB RDS instances in the ApsaraDB RDS console.
- Use a database image provided in Alibaba Cloud Marketplace.

Can I build a database on an ECS instance?

Yes, you can install database software and configure a database environment on an ECS instance. You can also separately purchase ApsaraDB RDS instances in the ApsaraDB RDS console.

Do ECS instances support Oracle databases?

Yes, ECS instances support Oracle databases. Before you install an Oracle database, we recommend that you perform a performance stress test on the ECS instance to ensure that the instance can satisfy the read and write requirements of the database.

Are public and private IP addresses independent? Can I specify or add IP addresses?

In the classic network, public and private IP addresses are independent of each other. Private IP addresses in the classic network are used for communication between ECS instances and between ECS instances and OSS buckets or ApsaraDB RDS instances. If the public bandwidth of an ECS instance is 0 Mbit/s, no public IP address is assigned to the instance. In normal cases, public and private IP addresses in the classic network do not change. You cannot specify, select, or add IP addresses in the classic network.

In virtual private clouds (VPCs), NAT gateways map public IP addresses to private IP addresses. You can add IP addresses by specifying or automatically assigning secondary private IP addresses to elastic network interfaces (ENIs).

Can load balancing be implemented for a single ECS instance?

Both Linux and Windows ECS instances can be load-balanced. Make sure that the configurations of ECS instances used as web servers meet the requirements for website code to run. Load balancing can be implemented for as few as one ECS instance. However, we recommend you implement load balancing for two or more ECS instances within an account.

Can I change the region of an ECS instance that I purchased?

No, the regions of purchased ECS instances cannot be changed. To change the region of an ECS instance, you can use the `ACS-ECS-CloneInstancesAcrossRegion` public template provided by Operation Orchestration Service (OOS) to copy the instance to another region. The new and original instances have identical disk data but different IP addresses.

Can I adjust the partition size of a purchased disk?

For system security and stability purposes, system disks cannot be repartitioned on Windows or Linux instances. If you use a third-party tool to repartition system disks, unknown exceptions such as system failures and data loss may occur.

If you repartition a data disk, data may be lost. We recommend that you do not repartition data disks.

How do I replace the automatically assigned public IP address of an ECS instance with an EIP?

To replace the automatically assigned public IP address of an ECS instance with an EIP, make sure that the instance uses the pay-by-bandwidth billing method for network usage and that you have purchased an EIP. Then, use one of the following methods to replace the public IP address with the EIP:

- Method 1:
  i. Change the billing method for network usage from pay-by-bandwidth to pay-by-traffic. For more information, see Public bandwidth.
  ii. Convert the automatically assigned public IP address into an EIP. For more information, see Convert the static public IP address of an ECS instance in a VPC to an EIP.
  iii. Disassociate the EIP obtained in the previous step. For more information, see Disassociate an EIP from a cloud resource.
  iv. Associate the EIP that you purchased with the ECS instance. For more information, see Associate an EIP with an ECS instance.
- Method 2:

In the OOS console, use the `ACS-ECS-ConvertsPublicIPToNewEIPByInstanceId` template to replace the automatically assigned public IP address of the ECS instance with an EIP. To perform this operation in the China (Hangzhou) region, access the ACS-ECS-ConvertsPublicIPToNewEIPByInstanceId template in the OOS console and select China (Hangzhou) from the drop-down region list in the top navigation bar.

How do I view subscription ECS instances in all regions within my account?

You can go to the Renewal page to view subscription ECS instances within all regions within your account.

1. Log on to the ECS console.
2. In the top navigation bar, choose **Expenses > Renewal Management**.

When can I force stop an ECS instance? What are the consequences?

If an instance cannot be stopped by a proper shutdown procedure, you can force stop the instance. A forced stop is equivalent to a power outage and can result in loss of unsaved data.

Why am I unable to reactivate my ECS instance?

You may be unable to reactivate an ECS instance due to one of the following reasons:

- You have overdue payments in your account. Pay the outstanding bills and try again.
- The system is busy. Try again later.
- Resources of the specified instance type are sold out.

> Note

Why has an ECS instance with release protection enabled been automatically released from a scaling group?

Auto Scaling can automatically release an ECS instance created by a scale-out event even if you have enabled release protection for the instance by using the ECS console or by calling the ModifyInstanceAttribute operation.

To prevent the ECS instance from being automatically released, you must change its state to Protected in the Auto Scaling console. For more information, see Put an ECS instance into the Protected state.

How do I test the packet forwarding rate of an instance?

For information about how to test the packet forwarding rate of an instance, see Best practices for testing network performance. For information about how to test the packet forwarding rate of a g7, c7, or r7 instance, see Use Pktgen to test the network performance of ECS instances.

How do I migrate data between ECS instances?

Perform different operations to migrate data from one instance to another based on whether both instances are located within the same region and belong to the same account.

- To migrate instance data within the same region and the same account, use the image or snapshot service.

- Use the image service to migrate instance data.
  a. Create a custom image from the source instance.
  b. Use the custom image to create an instance or replace the image of an existing instance. For more information, see Create an ECS instance by using a custom image and Change the operating system.
- Use the snapshot service to migrate instance data.
  a. Create snapshots for the disks that are attached to the source instance. For more information, see Create a snapshot of a disk.
  b. Create disks from the snapshots. For more information, see Create a disk from a snapshot.
  c. Attach the new disks to the destination instance. For more information, see Attach a data disk.

- To migrate instance data across regions within the same account, perform the following steps:
  i. Create a custom image from the source instance.
  ii. Copy the custom image to the destination region. For more information, see Copy a custom image.
  iii. Use the shared image in the destination account to create an instance or replace the image of an existing instance. For more information, see Create an ECS instance by using a custom image and Change the operating system.
- To migrate instance data across accounts within the same region, perform the following steps:
  i. Create a custom image from the source instance.
  ii. Share the custom image to the destination account. For more information, see Share or unshare a custom image.
  iii. Use the shared image in the destination account to create an instance or replace the image of an existing instance. For more information, see Create an ECS instance by using a custom image and Change the operating system.
- To migrate instance data across regions and accounts, perform the following steps:
  i. Create a custom image from the source instance.
  ii. Copy the custom image to the destination region. For more information, see Copy a custom image.
  iii. Share the new image (image copy) to the destination account. For more information, see Share or unshare a custom image.
  iv. Use the shared image in the destination account to create an instance or replace the image of an existing instance. For more information, see Create an ECS instance by using a custom image and Change the operating system.

For more information, see Use the snapshot and image features to migrate instance data.

In addition, Server Migration Center (SMC) also supports the full migration, incremental migration, batch migration, and VPC-based migration features and allows you to migrate data between ECS instances in the SMC console. For more information, see Migrate between two ECS instances.

What is the AliVulfix process in an ECS instance?

The AliVulfix process is an Alibaba Cloud Security program that scans ECS instances for vulnerabilities.

How do I protect ECS instances against attacks?

ECS instances use Alibaba Cloud Security to defend against DDoS attacks. CloudMonitor deployed on ECS instances can automatically detect network attacks and scrub suspicious traffic. Alibaba Cloud implements blackhole filtering to protect ECS instances against high-volume attacks. To strengthen security protection, we recommend that you install security software and disable ports that are not commonly used.

What security services does Alibaba Cloud provide?

Alibaba Cloud Security is backed by the robust data analysis capabilities of the cloud computing platform of Alibaba Cloud to provide a comprehensive set of security services such as security vulnerability detection, website trojan detection, host intrusion detection, and anti-DDoS protection.

For information about more security services, visit the Alibaba Cloud Security Services page.

How do I handle mining programs or apply to unlock affected servers?

You can handle mining programs or apply to unlock affected servers in the Security Center or handle mining worms in the Cloud Firewall console. For more information, see Defend against mining programs and Best practices for handling mining programs.

If a server is locked because it is infected by a mining virus or attacked, you can apply to unlock the server. To apply to unlock servers, go to the Penalties page in the Security Center.

When you apply to unlock servers, take note of the following items:

- You can apply to unlock servers in your account only once.
- A server that has been unlocked is automatically checked after 3 days. If a mining program is detected on the server again, the server is locked and cannot be unlocked.
- After the server is unlocked, back up its data as soon as possible.

I have already renewed an expired Linux instance but I am still unable to access the website hosted on it. What do I do?

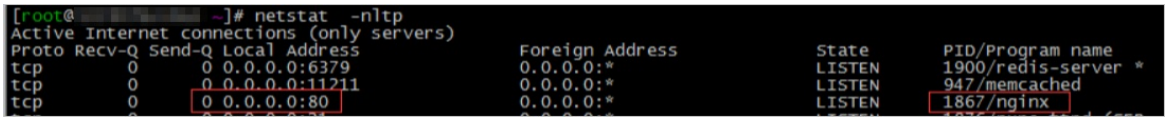
Problem description: A Linux ECS instance is in the Stopped state after it has expired. After you have renewed and restarted the instance, you still cannot access the website hosted on it.

Problem description: This may be because the website service has not been started.

Solution:

1. Connect to the instance and run the following command to check whether the website service has been started:

   ```
   # netstat -nltp //Check whether port 80 on the instance is being listened on.
   ```

2. If no information about port 80 is displayed in the command output, the website service has not been started. Run a command to manually start the website service and relevant services.

   In Linux, websites are typically developed based on PHP and MySQL.

   - In Apache, you need to only start the website service and MySQL.

```
#/etc/init.d/httpd start          //Start the website service. This command is applic
able to Apache.
#/etc/init.d/mysqld start         //Start MySQL.
```

- In NGINX, you must start the website service, PHP, and MySQL.

```
#/etc/init.d/nginx start          //Start the website service. This command is applic
able to NGINX.
#/etc/init.d/php-fpm start        //Start PHP.
#/etc/init.d/mysqld start         //Start MySQL.
```

3. Check again whether the website service has been started.

```
#netstat -nltp //Check whether port 80 on the instance is being listened on.
```

4. After the website service has been started, access the website again.

If the problem persists, submit a ticket.

How do I activate a Windows ECS instance within a VPC?

To activate a Windows ECS instance within a VPC, you must use a specific Key Management Service (KMS) domain name. For more information, see Activate the VPC-Connected Windows instances using KMS servers.

How do I query, partition, and format the disks of a Linux instance?

You can run the **df -h** command to check the capacity and usage of disks, and run the **fdisk -l** command to view disk information. For information about how to partition and format the disks of Linux instances, see Partition and format a data disk on a Linux instance.

How do I upload files to a Linux instance?

You can use the FTP service to upload files to a Linux instance.

How do I change the owner and owner group of directories and files on a Linux instance?

If the file or directory permissions are not correctly configured on the web server, a 403 error is reported when you access a website hosted on the instance. Before you adjust a file or directory, you must identify the identity under which the file or directory process is running.

You can run the **ps** and **grep** commands to query the identities under which processes are running.

You can run the **ls -l** command to query the owners and owner groups of files and directories.

To change the owners and groups, run the chown command. For example, you can run the **chown -R www.www /alidata/www/phpwind/** command to change the owners and groups of all files and directories under the */alidata/www/phpwind* directory to the www account.

How do I update software repositories for Linux instances?

You can use an automatic tool to update software repositories for Linux instances. For more information, see Use scripts to automatically update software repositories in Linux instances.

What limits apply to the transfer and change of public IP addresses of ECS instances?

The following limits apply to the transfer and change of public IP addresses of ECS instances:

- ECS instances in the classic network
  - You cannot transfer public IP addresses across accounts.
  - The public IP address of an instance can be changed within 6 hours after the instance is created, and can be changed up to three times. For more information, see Change the public IP address of an instance.
  - If Anti-DDoS Pro is deployed, you can change the IP address of an instance up to 10 times by using the Anti-DDoS Pro console. For more information, see Change the public IP address of an ECS origin server in Anti-DDoS Pro User Guide.
- ECS instances in VPCs
  - You cannot transfer public IP addresses or EIPs across accounts.
  - If no public IP address is assigned to your instance, you can associate an EIP with the instance. You can replace the public IP address of your instance with an EIP.
  - If you assign a public IP address to your instance while you create the instance, take note of the following limits:
    - The public IP address of an instance can be changed within 6 hours after the instance is created, and can be changed up to three times. For more information, see Change the public IP address of an instance.
    - You can convert this public IP address into an EIP and then replace the EIP. For more information, see Convert the public IP address of a VPC-type instance to an EIP.

If you have further questions, submit a ticket.

Can I access amazon.com from my ECS instance?

You can access amazon.com from your ECS instance if the instance can properly connect to the Internet.

Why am I unable to access a website hosted outside Chinese mainland after I log on to my ECS instance?

Websites hosted outside Chinese mainland can be accessed by your ECS instance only when the websites comply with the laws, regulations, and regulatory requirements of the country or region where your instance is located. Make sure that your ECS instance can properly connect to the Internet and that the website complies with the preceding laws, regulations, and regulatory requirements.

I cannot purchase more pay-as-you-go instances. What do I do?

If you have reached the maximum number of pay-as-you-go instances that you can purchase, you cannot purchase more pay-as-you-go instances. For more information, see the "Instance limits" section in Limits. You can log on to the ECS console and click Privileges on the Overview page to view your resource quotas. For more information, see View and increase instance quotas.

How can I view the resource quota?

For more information about how to view the limits and quotas of resources, see 使用限制.

After a pay-as-you-go instance is stopped manually or due to an overdue payment, am I still charged for it?

Stopped due to an overdue payment: When a payment becomes overdue in your account, your pay-as-you-go instance is automatically stopped and billing for the instance stops. Instances do not always remain in the Stopped state after they are stopped due to overdue payments. For more information, see Pay-as-you-go.

Manually stopped: You can use the ECS console or call the StopInstance operation to stop a running pay-as-you-go instance. When the instance is stopped, it enters the **Stopped** state. How stopped pay-as-you-go instances are billed depends on their network types.

- VPC: You can enable economical mode for pay-as-you-go instances within VPCs.
  - After economical mode is enabled, billing for pay-as-you-go instances begins when they are created, and billing for some instance resources stops when the instances enter the **Stopped** state and resumes when the instances are started. When a pay-as-you-go instance enters the Stopped state, economical mode stops the billing only for the vCPUs, memory, and public IP address of the instance. Other resources such as disks and EIP of the instance continue to be billed. For more information, see Economical mode.
  - After economical mode is disabled, billing for pay-as-you-go instances continues when they enter the **Stopped** state.
- Classic network: Pay-as-you-go instances in the classic network are billed regardless of whether they are in the **Stopped** state.

What do I do if an order cannot be placed to change the billing method of an instance from pay-as-you-go to subscription?

You may be unable to place the order due to one of the following reasons:

- The instance is in a state that does not support billing method changes. For example, you have an unpaid order for the instance.
- Billing method changes are not allowed due to an upcoming scheduled automatic release.
- Billing method changes are not allowed because instance information has been changed.
- A previous order to change the billing method of the instance has not been paid.

If one of the preceding errors is reported, adjust the instance accordingly.

How long after an order is paid does it take to change the billing method of an instance from pay-as-you-go to subscription?

The billing method of your instance is changed after the order is paid. It can take up to 4 seconds to change the billing method of 20 instances. After the change is complete, you can see that the billing method of your instance has been changed to **Subscription** in the ECS console.

What do I do if the billing method of an instance cannot be changed from pay-as-you-go to subscription?

Submit a ticket.

When I change the billing method of an instance from pay-as-you-go to subscription, does the billing method for network usage of the instance also change?

No, the billing method for network usage of the instance does not change. Only the billing method of instances and disks can be changed from pay-as-you-go to subscription. For information about how to change the billing method for network usage, see Overview of instance configuration changes.

I have an unpaid order to change the billing method of an instance from pay-as-you-go to subscription. If I upgrade the configurations of the instance, is the order still valid?

An order is created when you change the billing method of your instance from pay-as-you-go to subscription. You must pay for the order to complete the change. If you upgrade the configurations of the instance before the order is paid, the order payment cannot be completed because the instance components are different and the original order no longer matches. If you still want to change the billing method of your instance, you must cancel the unpaid order and place a new order.

What do I do if the billing method of an instance cannot be changed from subscription to pay-as-you-go?

You may be unable to change the billing method of an instance from subscription to pay-as-you-go due to one of the following reasons:

- The instance is in a state that does not support billing method changes. For example, you have an unpaid order for the instance.
- The instance is in the **Expired** state.
- Billing method changes are not allowed because instance information has been changed. For example, the bandwidth of the instance has been temporarily upgraded.

If one of the preceding errors is reported, adjust the instance accordingly. If the problem persists, submit a ticket.

When I attempt to change the billing method of a disk in an instance, an error message is returned indicating that I have already changed the billing method three times. What does this mean?

Each ECS instance can have its configurations downgraded up to three times. Downgrade operations include downgrades of instance specifications, bandwidth downgrades, and the change of the disk billing method from subscription to pay-as-you-go.

Why am I unable to change a pay-as-you-go instance into a subscription one?

The pay-as-you-go instance whose billing method you want to change must meet the following requirements:

- The instance type of the instance is not retired. For more information, see Retired instance types.
- The instance is not a preemptible instance.
- You do not have unpaid orders for the instance.

  If you have unpaid orders for the instance, you must pay for the orders or cancel the orders before you can change the billing method of the instance.
- The automatic release time is not set for the instance.

  If the automatic release time is set for the instance, you must cancel the automatic release of the instance before you change its billing method. For more information, see Disable automatic release.

- The instance is in the **Running** or **Stopped** state.

  Note: An order to change the billing method of an ECS instance must be placed when the ECS instance is in the Running or Stopped state. If the instance state changes before the payment completes, the order fails and the billing method does not change. You can go to the Billing Management console and pay for the order when the instance is in the Running or Stopped state again.

How do I view the expiration time of a subscription instance?

You can log on to the ECS console and go to the **Instances** page to view the expiration time of your subscription instance in the **Billing Method** column.

> **Note** If the **Billing Method** column is not displayed, click the icon in the upper-right corner. In the Column Filters dialog box, select **Billing Method** and click **OK**.