Alibaba Cloud
Elastic Compute Service

Product Introduction

Issue: 20181119

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 due to product version upgrades, adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall 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 the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults" and "as available" basis. 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 bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, 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 content of the Alibaba Cloud website, including but not limited to works, products, images, archives, information, materials, website architecture, website graphic layout, and webpage design, 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 the Alibaba Cloud website, product programs, or content 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 contact Alibaba Cloud directly if you discover any errors in this document.

Generic conventions

Table -1: Style conventions

Style	Description	Example
	This warning information 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.
	This warning information 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 restore business.
	This indicates warning information, supplementary instructions, and other content that the user must understand.	 **Note:** Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	 **Note:** You can use **Ctrl** + **A** to select all files.
>	Multi-level menu cascade.	**Settings** > **Network** > **Set network type**
Bold	It is used for buttons, menus, page names, and other UI elements.	Click **OK**.
`Courier font`	It is used for commands.	Run the `cd /d C:/windows` command to enter the Windows system folder.
Italics	It is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	It indicates that it is a optional value, and only one item can be selected.	`ipconfig` *[-all|-t]*
{} or {a|b}	It indicates that it is a required value, and only one item can be selected.	`swich` *{stand | slave}*

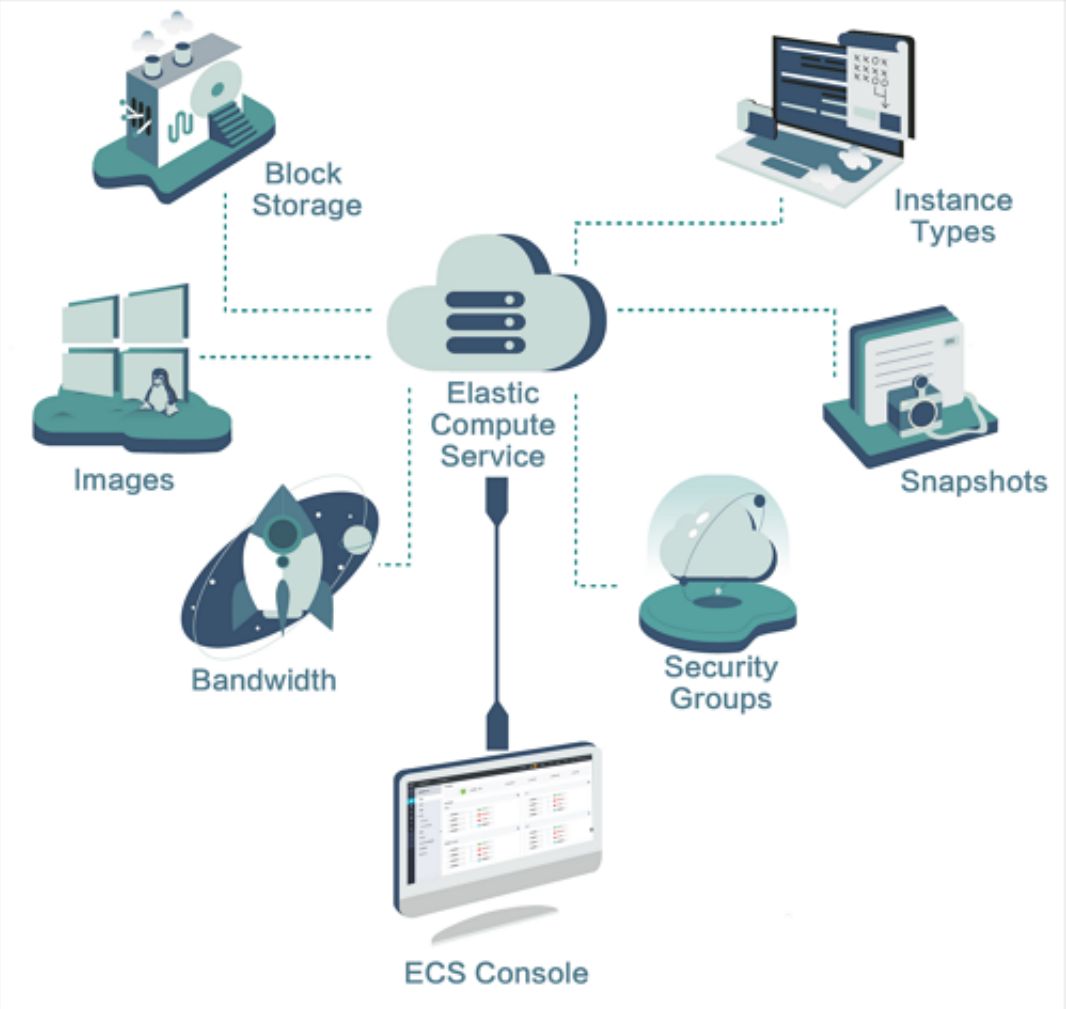
Contents

1 What is ECS?

This article gives a brief introduction to what ECS is, and the resources and services that it involves.

Elastic Compute Service (ECS) is a type of computing service that features elastic processing capabilities. ECS has a simpler and more efficient management mode than physical servers. You can create instances, change the OS, and add or release any number of ECS instances at any time to fit your business needs. An ECS instance is a virtual computing environment that includes CPU, memory, and other basic computing components. An instance is the core component of ECS and is the actual operating entity offered by Alibaba Cloud. Other resources, such as disks, images, and snapshots, can only be used in conjunction with an ECS instance.

The following figure illustrates the concept of an ECS instance. You can use the *ECS console* to configure the instance type, disks, OS, and other affiliated resources.



Basic concepts

It is helpful to understand the following concepts before you use ECS:

- *Region and zone*: A physical location where a data center is located.
- *ECS instance*: A virtual computing environment that includes the CPU, memory, OS, bandwidth, disks, and other basic computing components.
- *Instance types*: The specifications of an ECS instance, including the number of vCPU cores, memory, and networking performance. The instance type of an ECS instance determines its compute capability.
- *Images*: A running environment template for ECS instances. It generally includes an OS and preinstalled software.
- *Block storage*: Block level storage products for your ECS, including *Cloud disks and Shared Block Storage* based on the distributed storage architecture and *local disks* located on the physical server that an ECS instance is hosted on.
- *Snapshots*: A copy of the data on an elastic block storage device as it was at a specific point in time.
- *Network types*:
  - Virtual Private Cloud (VPC): A private network established in Alibaba Cloud. VPCs are logically isolated from other virtual networks in Alibaba Cloud. For more information, see *What is VPC*.
  - Classic network: A network majorly deployed in the public infrastructure of Alibaba Cloud.
- *Security group*: A logical group of instances that are in the same region and have the same security requirements and mutual trust. A security group works as a virtual firewall for the ECS instances inside it.

Operations

Alibaba Cloud provides an intuitive operation interface for you to manage your ECS instances.

- You can log on to the ECS console to operate ECS instances. For more information, see *User Guide*.
- You can use API to manage your ECS instances. For more information, see *API References*.
- You can also use Alibaba Cloud CLI to call API to manage ECS instances. For more information, see *Alibaba Cloud Command Line Interface*.

- You can use the open source tool Terraform to provision and manage ECS resources. Terraform provides a simple mechanism for deploying and versioning configuration files to Alibaba Cloud and other supported clouds. For more information, see *What is Terraform?*.

ECS pricing and billing

ECS supports both Subscription and Pay-As-You-Go billing methods. For more information, see *Billing methods*.

For pricing details, see the *Pricing* page.

Learning path

You can use the *ECS Learning Path* as a mentor to learn ECS basics or add to your knowledge.

Related services

The following services are frequently used together with ECS:

- *Alibaba Cloud Marketplace* is an online market. You can purchase software infrastructure, developer tools, and business software provided by third-party partners. If you have software you want to sell, you can become a marketplace service provider. For more information, see *Marketplace*.
- Auto Scaling enables you to dynamically scale your computing capacity up or down to meet the workload of your ECS instances according to scaling policies you specify. It also reduces the need of manual provision. For more information, see *What is Auto Scaling*.
- Container Service enables you to manage the life cycle of containerized applications by using Docker and Kubernetes. For more information, see *What is Container Service*.
- Server Load Balancer distributes the incoming traffic among multiple ECS instances according to the configured forwarding rules. For more information, see *What is Server Load Balancer*.
- CloudMonitor manages ECS instances, system disks, Internet bandwidth, and other resources. For more information, see *What is CloudMonitor*.
- Server Guard (Server Security) provides real-time awareness and defense against intrusion events, which safeguards the security of your ECS instances. For more information, see *What is Server Guard*.
- Anti-DDoS Basic prevents and mitigates DDoS attacks by routing traffic away from your infrastructure. Alibaba Cloud Anti-DDoS Pro safeguards your ECS instances under high volume DDoS attacks. For more information, see *What is Anti-DDoS Basic* and *What is Anti-DDoS Pro*.

- Alibaba Cloud SDK enables you to access Alibaba Cloud services and to manage your applications by using the programming language of your choice. For more information, see *Developer Resources*. You can use *OpenAPI Explorer* to debug ECS API and generate the SDK Demo.

2 Benefits of ECS

Compared with Internet Data Centers (IDCs) and server vendors, ECS has benefits in terms of availability, security, and elasticity.

Availability

Alibaba Cloud adopts more stringent IDC standards, server access standards, and O&M standards to guarantee data reliability and high availability of cloud computing infrastructure and cloud servers.

In addition, each Alibaba Cloud region consists of multiple zones. For greater fault tolerance, you can build active/standby or active/active services in multiple zones. For a finance-oriented solution with three IDCs in two regions, you can build fault tolerant systems in multiple regions and zones. Those services include disaster tolerance and backup, which are supported by the mature solutions built by Alibaba Cloud.

Switching between services is smooth within the Alibaba Cloud framework. For more information, see *E-Commerce Solutions*. Alibaba Cloud industry solutions support a variety of services, such as finance, E-commerce, and video services.

Alibaba Cloud provides you with the following support services:

- Products and services for availability improvement, including cloud servers, Server Load Balancer, multi-backup databases, and Data Transmission Services (DTS).
- Industry partners and ecosystem partners that help you build a more advanced and stable architecture and guarantee service continuity.
- Diverse training services that enable you to connect with high availability from the business end to the underlying basic service end.

Security

For cloud computing users, security and stability are priorities. Alibaba Cloud has passed a host of international information security certifications, including ISO 27001 and MTCS, which demand strict confidentiality of user data and user information, as well as user privacy protection. We recommend that you use ECS in an *Alibaba Cloud Virtual Private Cloud (VPC)*.

- Alibaba Cloud VPC offers more business possibilities.

  You only need to perform a simple configuration to connect your business environment to global IDCs, making your business more flexible, stable, and extensible.

- Alibaba Cloud VPC can connect to your IDC

  through a leased line to build a hybrid cloud architecture. You can build a more flexible business with robust networking derived from Alibaba Cloud's various hybrid cloud solutions and network products. A superior business ecosystem is made possible with Alibaba Cloud's ecosystem.
- Alibaba Cloud VPC is more stable and secure.

  Stable: After you build your business on VPC, you can update your network architecture and obtain new network functions on a daily basis as the network infrastructure evolves constantly, allowing your business to run steadily. You can divide, configure, and manage your network on VPC according to your needs.

  Secure: VPC features traffic isolation and attack isolation to protect your services from attack traffic on the Internet. By building your business on VPC, the first line of defense is established.

VPC provides a stable, secure, fast-deliverable, self-managed, and controllable network environment. VPC hybrid cloud brings the technical advantages of cloud computing to traditional industries, in addition to industries and enterprises that are not engaged in cloud computing.

Elasticity

Elasticity is a key benefit of cloud computing. Elasticity is a key benefit of cloud computing. By using Alibaba Cloud, you can have all the IT resources necessary to get an IT company of medium size provisioned within minutes. The available resources and capacities can meet the requirements of most companies, allowing their applications built on the cloud to handle a huge volume of transactions without problems.

- Elastic computing
  - Vertical scaling involves modifying the configurations of a server.

    After you purchase ECS or storage capacity of Alibaba Cloud, you can configure your server with great flexibility based on your actual transaction volume, whereas it may be difficult to change configurations in the traditional IDC model. For more information about vertical scaling, see *Change configurations*.
  - Horizontal scaling

    Horizontal scaling allows the re-division of resources between applications. For example, at peak hours for game or live video streaming apps, in the traditional IDC model, your hands may be tied if additional resources are required when you are already at full capacity. Cloud

computing uses elasticity to provide additional resources to you over that period. When the period ends, you can release unnecessary resources to reduce your business costs. By using both horizontal scaling and auto-scaling that Alibaba Cloud provides, you can determine how and when you scale your resources or apply your scaling based on business loads. For more information about horizontal scaling, see *Auto Scaling*.

- Elastic storage

  Alibaba Cloud provides elastic storage. In the traditional IDC model, if more storage space is required, you can only add servers, but the number of servers that you can add is limited. In the cloud computing model, however, the sky is the limit. Order as much storage space as you need to meet business demand. For more information about elastic storage, see *Resize a disk*.

- Elastic network

  Alibaba Cloud features elastic network as well. When you purchase Alibaba Virtual Private Cloud (VPC), you can set network configurations to be the same as those of data centers. In addition, VPC has the following benefits: interconnection between data centers, separate secure domains in data centers, and flexible network configurations and planning within the VPC For more information about elastic networks, see *Virtual Private Cloud*.

Alibaba Cloud incorporates elasticity in computing, storage, network, and business architecture design. By using Alibaba Cloud, you can build your business portfolio in any way you want.

Comparison between ECS and traditional IDCs

The table lists the benefits of ECS compared with traditional IDCs.

Item	ECS	Traditional IDCs
Equipment rooms	Provides independently developed DC powered servers with low PUE.	Provides traditional AC powered servers with high PUE.
	Provides backbone equipment rooms with high outbound bandwidth and dedicated bandwidth.	Provides equipment rooms with various quality levels and shared bandwidth primarily, difficult for users to choose.
	Provides multiline BGP equipment rooms, enabling smooth and balanced access throughout the country.	Provides equipment rooms with single or dual line primarily.

Item	ECS	Traditional IDCs
Ease of operation	Provides built-in mainstream OSs, including activated Windows OS.	Purchases and installs OS manually.
	Switches OS online.	Reinstalls OSs manually.
	Provides a Web-based console for online management.	Users must manage and maintain network manually.
	Provides mobile phone verification for password setting, increasing data security.	Has difficulty in resetting passwords, and exposes high risk of password cracking.
Disaster recovery and backup	Each data segment has multiple copies. When one copy is corrupted, the data can be quickly restored.	Users must build disaster recovery environment by themselves, and use traditiona l storage devices.
	Users can customize automatic snapshot policies to create automatic snapshots for data recovery.	Users must restore all corrupted data manually.
	Faults can be recovered fast and automatically.	Faults cannot be recovered automatically.
Security and reliability	Effectively prevents MAC spoofing and ARP attacks.	Fails to prevent MAC spoofing and ARP attacks.
	Effectively defends against DDoS attacks by using black holes and cleaning traffic.	Needs additional costs for devices for traffic cleaning and black hole shielding systems.
	Provides additional services, such as port scanning, Trojan scanning, and vulnerability scanning.	Typically encountered problems such as vulnerability , Trojan, and port scanning.
Flexible scalability	Activates cloud servers on demand and upgrades configurations online.	Needs a long time for server delivery.
	Adjusts outbound bandwidth as required.	One-off purchase of outbound bandwidth, unable to adjust.
	Combines with Server Load Balancer online, enabling	Uses hardware-based server load balancing, which is

Item	ECS	Traditional IDCs
	scaling up applications quickly and easily.	expensive and extremely difficult to set up.
Cost effectiveness	Low cost.	High cost.
	Small up front investment.	Large up front investment, possible waste of resources.
	Purchases on demand and pay as you go, meeting requirements for constant business changes.	Purchases up front to meet configuration requirements for peak hours.

3 Scenarios

ECS is a highly flexible solution. It can be used independently as a simple web server, or used with other Alibaba Cloud products, such as Object Storage Service (OSS) and Content Delivery Network (CDN), to provide advanced solutions.

ECS can be used in the following applications.

Official corporate websites and simple web applications

During the initial stage, corporate websites have low traffic volumes and require only low-configuration ECS instances to run applications, databases, storage files, and other resources. As your business expands, you can upgrade the ECS configuration and increase the number of ECS instances at any time. You no longer need to worry about insufficient resources during peak traffic.

Multimedia and large-traffic apps or websites

ECS can be used with OSS to store static images, videos, and downloaded packages, reducing storage fees. In addition, ECS can be used with CDN or Server Load Balancer to greatly reduce user access waiting time, reduce bandwidth fees, and improve availability.

Databases

A high-configuration I/O-optimized ECS instance can be used with an SSD cloud disk to support high I/O concurrency with higher data reliability. Alternatively, multiple lower-configuration I/O-optimized ECS instances can be used with Server Load Balancer to deliver a highly available architecture.

Apps or websites with large traffic fluctuations

Some applications may encounter large traffic fluctuations within a short period. When ECS is used with Auto Scaling, the number of ECS instances is automatically adjusted based on traffic. This feature allows you to meet resource requirements while maintaining a low cost. ECS can be used with Server Load Balancer to deliver a high availability architecture.

4 Instance type families

This article introduces the available ECS instance type families.

An ECS instance is the minimal unit that can provide computing capabilities and services for your business.

ECS instances are categorized into specification types, which are called type families, based on the business scenarios they can be applied to. You may select multiple type families for one business scenario. Each type family contains multiple *ECS instance types* with different CPU and memory specifications, including the CPU model and clock speed. Besides the instance type, you must also define a block storage, an image, and the network service when you create an instance.

Note:
The availability of instance type families and their types varies from region to region. Go to the *purchase page* to check the available instance types.

Alibaba Cloud ECS provides two kinds of instance type families: enterprise-level instance type families and entry-level instance type families. Type families for enterprise-level computing offer stable performance and dedicated resources, while entry-level type families are ideal for small and mid-sized websites, or individual customers. For the differences, see *Enterprise-level instances and entry-level instances FAQ*.

Note:

- If you are using sn1, sn2, t1, s1, s2, s3, m1, m2, c1, c2, c4, ce4, cm4, n1, n2, or e3, see *Phased-out instance types*.
- Upgrading instance types is supported within or between certain instance type families. For such families and corresponding upgrade rules, see *Instance type families that support upgrading instance types*.
- Upgrading instance types is not supported within or between the following instance type families: d1, d1ne, i1, i2, i2g, ga1, gn5, f1, f2, f3, ebmc4, ebmg5, sccg5, and scch5.

Alibaba Cloud ECS instances are categorized into the following type families:

- Type families for enterprise-level computing on the x86-architecture:
  - *g5, general-purpose type family*
  - *sn2ne, general-purpose type family with enhanced network performance*

- *ic5, intensive compute instance type family*
- *c5, compute instance type family*
- *sn1ne, compute optimized type family with enhanced network performance*
- *r5, memory instance type family*
- *re4, memory optimized type family with enhanced performance*
- *se1ne, memory optimized type family with enhanced network performance*
- *se1, memory optimized type family*
- *d1ne, big data type family with enhanced network performance*
- *d1, big data type family*
- *i2, type family with local SSD disks*
- *i2g, type family with local SSD disks*
- *i1, type family with local SSD disks*
- *hfc5, compute optimized type family with high clock speed*
- *hfg5, general-purpose type family with high clock speed*

- Type families for enterprise-level heterogeneous computing:
  - *gn6v, compute optimized type family with GPU*
  - *gn5, compute optimized type family with GPU*
  - *gn5i, compute optimized type family with GPU*
  - *gn4, compute optimized type family with GPU*
  - *ga1, visualization compute optimized type family with GPU*
  - *f1, compute optimized type family with FPGA*
  - *f2, compute optimized type family with FPGA*
  - *f3, compute optimized type family with FPGA*
- ECS Bare Metal Instance type families and Super Computing Cluster (SCC) instance type families:
  - *ebmhfg5, ECS Bare Metal Instance type family with high clock speed*
  - *ebmc4, computing ECS Bare Metal Instance type family*
  - *ebmg5, general-purpose ECS Bare Metal Instance type family*
  - *scch5, Super Computing Cluster (SCC) instance type family with high clock speed*
  - *sccg5, geneneral-purpose Super Computing Cluster (SCC) instance type family*
- Type families for entry-level computing on the x86-architecture:

- *t5, burstable instances*
- *xn4/n4/mn4/e4, type families of previous generations for entry-level users, computing on the x86-architecture*

g5, general-purpose type family

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU to memory ratio = 1:4
- Ultra high packet forwarding rate
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Scenarios where a large volume of packets are received and transmitted, such as the re-transmission of telecommunication information
  - Enterprise-level applications of various types and sizes
  - Medium and small database systems, cache, and search clusters
  - Data analysis and computing
  - Computing clusters and data processing reliant on memory

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.g5. large	2	8.0	N/A	1.0	300	2	2
ecs.g5. xlarge	4	16.0	N/A	1.5	500	2	3
ecs.g5. 2xlarge	8	32.0	N/A	2.5	800	2	4

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.g5. 3xlarge	12	48.0	N/A	4.0	900	4	6
ecs.g5. 4xlarge	16	64.0	N/A	5.0	1,000	4	8
ecs.g5. 6xlarge	24	96.0	N/A	7.5	1,500	6	8
ecs.g5. 8xlarge	32	128.0	N/A	10.0	2,000	8	8
ecs.g5. 16xlarge	64	256.0	N/A	20.0	4,000	16	8

Click *here* to view other instance type families.

sn2ne, general-purpose type family with enhanced network performance

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU to memory ratio = 1:4
- Ultra high packet forwarding rate
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) or Platinum 8163 (Skylake) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Scenarios where a large volume of packets are received and transmitted, such as the re-transmission of telecommunication information
  - Enterprise-level applications of various types and sizes
  - Medium and small database systems, cache, and search clusters
  - Data analysis and computing
  - Computing clusters and data processing depending on memory

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.sn2ne .large	2	8.0	N/A	1.0	300	2	2
ecs.sn2ne .xlarge	4	16.0	N/A	1.5	500	2	3
ecs.sn2ne .2xlarge	8	32.0	N/A	2.0	1,000	4	4
ecs.sn2ne .3xlarge	12	48.0	N/A	2.5	1,300	4	6
ecs.sn2ne .4xlarge	16	64.0	N/A	3.0	1,600	4	8
ecs.sn2ne .6xlarge	24	96.0	N/A	4.5	2,000	6	8
ecs.sn2ne .8xlarge	32	128.0	N/A	6.0	2,500	8	8
ecs.sn2ne .14xlarge	56	224.0	N/A	10.0	4,500	14	8

Click *here* to view other instance type families.

ic5, intensive compute instance type family

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU to memory ratio = 1:1
- Ultra high packet forwarding rate
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Web front-end servers

- Data analysis, batch compute, and video coding
- Scenarios where a large volume of packets are received and transmitted, such as the re-transmission of telecommunication information
- Massively Multiplayer Online (MMO) game front-ends

Instance types

Instance type	Vcpu	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.ic5. large	2	2.0	N/A	1.0	300	2	2
ecs.ic5. xlarge	4	4.0	N/A	1.5	500	2	3
ecs.ic5. 2xlarge	8	8.0	N/A	2.5	800	2	4
ecs.ic5. 3xlarge	12	12.0	N/A	4.0	900	4	6
ecs.ic5. 4xlarge	16	16.0	N/A	5.0	1,000	4	8

Click *here* to view other instance type families.

c5, compute instance type family

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU to memory ratio = 1:2
- Ultra high packet forwarding rate
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Scenarios where a large volume of packets are received and transmitted, such as the re-transmission of telecommunication information

- Web front-end servers
- Massively Multiplayer Online (MMO) game front-ends
- Data analysis, batch compute, and video coding
- High-performance science and engineering applications

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.c5. large	2	4.0	N/A	1.0	300	2	2
ecs.c5. xlarge	4	8.0	N/A	1.5	500	2	3
ecs.c5. 2xlarge	8	16.0	N/A	2.5	800	2	4
ecs.c5. 3xlarge	12	24.0	N/A	4.0	900	4	6
ecs.c5. 4xlarge	16	32.0	N/A	5.0	1,000	4	8
ecs.c5. 6xlarge	24	48.0	N/A	7.5	1,500	6	8
ecs.c5. 8xlarge	32	64.0	N/A	10.0	2,000	8	8
ecs.c5. 16xlarge	64	128.0	N/A	20.0	4,000	16	8

Click *here* to view other instance type families.

sn1ne, compute optimized type family with enhanced network performance

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU to memory ratio = 1:2

- Ultra high packet forwarding rate
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) or Platinum 8163 (Skylake) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Scenarios where a large volume of packets are received and transmitted, such as the re-transmission of telecommunication information
  - Web front-end servers
  - Massively Multiplayer Online (MMO) game front-ends
  - Data analysis, batch compute, and video coding
  - High-performance science and engineering applications

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.sn1ne .large	2	4.0	N/A	1.0	300	2	2
ecs.sn1ne .xlarge	4	8.0	N/A	1.5	500	2	3
ecs.sn1ne .2xlarge	8	16.0	N/A	2.0	1,000	4	4
ecs.sn1ne .3xlarge	12	24.0	N/A	2.5	1,300	4	6
ecs.sn1ne .4xlarge	16	32.0	N/A	3.0	1,600	4	8
ecs.sn1ne .6xlarge	24	48.0	N/A	4.5	2,000	6	8
ecs.sn1ne .8xlarge	32	64.0	N/A	6.0	2,500	8	8

Click *here* to view other instance type families.

r5, memory instance type family

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- Ultra high packet forwarding rate
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Scenarios where a large volume of packets are received and transmitted, such as the re-transmission of telecommunication information
  - High-performance databases and high memory databases
  - Data analysis and mining, and distributed memory cache
  - Hadoop, Spark, and other enterprise-level applications with large memory requirements

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.r5. large	2	16.0	N/A	1.0	300	2	2
ecs.r5. xlarge	4	32.0	N/A	1.5	500	2	3
ecs.r5. 2xlarge	8	64.0	N/A	2.5	800	2	4
ecs.r5. 3xlarge	12	96.0	N/A	4.0	900	4	6
ecs.r5. 4xlarge	16	128.0	N/A	5.0	1,000	4	8
ecs.r5. 6xlarge	24	192.0	N/A	7.5	1,500	6	8
ecs.r5. 8xlarge	32	256.0	N/A	10.0	2,000	8	8

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.r5. 16xlarge	64	512.0	N/A	20.0	4,000	16	8

Click *here* to view other instance type families.

re4, memory optimized instance type family with enhanced performance

Features

- Supports SSD Cloud Disks and Ultra Cloud Disks
- I/O-optimized
- Optimized for high-performance databases, high memory databases, and other memory-intensive enterprise applications
- 2.2 GHz Intel Xeon E7 8880 v4 (Broadwell) processors, up to 2.4 GHz Turbo Boot
- vCPU to memory ratio = 1:12, up to 1920.0 GiB memory
- ecs.re4.20xlarge and ecs.re4.40xlarge have been certified by SAP HANA
- Ideal for:
  - High-performance databases and high memory databases (for example, SAP HANA)
  - Memory intensive applications
  - Big Data processing engines, such as Apache spark or Presto

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.re4. 20xlarge	80	960.0	N/A	15.0	2,000	16	8
ecs.re4. 40xlarge	160	1920.0	N/A	30.0	4,500	16	8

Click *here* to view other instance type families.

se1ne, memory optimized type family with enhanced network performance

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU to memory ratio = 1:8
- Ultra high packet receive and forwarding rate
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) or Platinum 8163 (Skylake) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Scenarios where a large volume of packets are received and transmitted, such as the re-transmission of telecommunication information
  - High-performance databases and large memory databases
  - Data analysis and mining, and distributed memory cache
  - Hadoop, Spark, and other enterprise-level applications with large memory requirements

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.se1ne .large	2	16.0	N/A	1.0	300	2	2
ecs.se1ne .xlarge	4	32.0	N/A	1.5	500	2	3
ecs.se1ne .2xlarge	8	64.0	N/A	2.0	1,000	4	4
ecs.se1ne .3xlarge	12	96.0	N/A	2.5	1,300	4	6
ecs.se1ne .4xlarge	16	128.0	N/A	3.0	1,600	4	8

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.se1ne .6xlarge	24	192.0	N/A	4.5	2,000	6	8
ecs.se1ne .8xlarge	32	256.0	N/A	6.0	2,500	8	8
ecs.se1ne .14xlarge	56	480.0	N/A	10.0	4,500	14	8

Click *here* to view other instance type families.

se1, memory optimized type family

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU to memory ratio = 1:8
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - High-performance databases and large memory databases
  - Data analysis and mining, and distributed memory cache
  - Hadoop, Spark, and other enterprise-level applications with large memory requirements

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.se1. large	2	16.0	N/A	0.5	100	1	2

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.se1.xlarge	4	32.0	N/A	0.8	200	1	3
ecs.se1.2xlarge	8	64.0	N/A	1.5	400	1	4
ecs.se1.4xlarge	16	128.0	N/A	3.0	500	2	8
ecs.se1.8xlarge	32	256.0	N/A	6.0	800	3	8
ecs.se1.14xlarge	56	480.0	N/A	10.0	1,200	4	8

Click *here* to view other instance type families.

d1ne, big data type family with enhanced network performance

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- High-volume local SATA HDD disks with high I/O throughput and up to 35 Gbit/s of bandwidth for a single instance
- vCPU to memory ratio = 1:4, designed for big data scenarios
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Hadoop MapReduce, HDFS, Hive, HBase, and so on
  - Spark in-memory computing, MLlib, and so on
  - Enterprises that require big data computing and storage analysis, such as those in the Internet and finance industries, to store and compute massive volumes of data
  - Elasticsearch, logs, and so on

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.d1ne.2xlarge	8	32.0	4 * 5500	6.0	1,000	4	4
ecs.d1ne.4xlarge	16	64.0	8 * 5500	12.0	1,600	4	8
ecs.d1ne.6xlarge	24	96.0	12 * 5500	16.0	2,000	6	8
ecs.d1ne-c8d3.8xlarge	32	128.0	12 * 5500	20.0	2,000	6	8
ecs.d1ne.8xlarge	32	128.0	16 * 5500	20.0	2,500	8	8
ecs.d1ne-c14d3.14xlarge	56	160.0	12 * 5500	35.0	4,500	14	8
ecs.d1ne.14xlarge	56	224.0	28 * 5500	35.0	4,500	14	8

Note:

- You cannot change configurations of d1ne instances.
- For more information about d1ne type families, see *FAQ on d1 and d1ne*.

Click *here* to view other instance type families.

d1, big data type family

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- High-volume local SATA HDD disks with high I/O throughput and up to 17 Gbit/s of bandwidth for a single instance
- vCPU to memory ratio = 1:4, designed for big data scenarios

- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Hadoop MapReduce, HDFS, Hive, and HBase
  - Spark in-memory computing and MLlib
  - Enterprises that require big data computing and storage analysis, such as those in the Internet and finance industries, to store and compute massive volumes of data
  - Elasticsearch and logs

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.d1.2xlarge	8	32.0	4 * 5500	3.0	300	1	4
ecs.d1.3xlarge	12	48.0	16 * 5500	4.0	400	1	6
ecs.d1.4xlarge	16	64.0	8 * 5500	6.0	600	2	8
ecs.d1.6xlarge	24	96.0	12 * 5500	8.0	800	2	8
ecs.d1-c8d3.8xlarge	32	128.0	12 * 5500	10.0	1,000	4	8
ecs.d1.8xlarge	32	128.0	16 * 5500	10.0	1,000	4	8
ecs.d1-c14d3.14xlarge	56	160.0	12 * 5500	17.0	1,800	6	8
ecs.d1.14xlarge	56	224.0	28 * 5500	17.0	1,800	6	8

Note:

For more information about d1 type families, see *FAQ on d1 and d1ne* .

Click *here* to view other instance type families.

i2, type family with local SSD disks

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- High-performance local NVMe SSD disks with high IOPS, high I/O throughput, and low latency.
- vCPU to memory ratio = 1:8, designed for high-performance databases
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - OLTP and high-performance relational databases
  - NoSQL databases, such as Cassandra and MongoDB
  - Search applications, such as Elasticsearch

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.i2.xlarge	4	32.0	1 * 894	1.0	500	2	3
ecs.i2.2xlarge	8	64.0	1 * 1788	2.0	1,000	2	4
ecs.i2.4xlarge	16	128.0	2 * 1788	3.0	1,500	4	8
ecs.i2.8xlarge	32	256.0	4 * 1788	6.0	2,000	8	8
ecs.i2.16xlarge	64	512.0	8 * 1788	10.0	4,000	16	8

Click *here* to view other instance type families.

i2g, type family with local SSD disks

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- High-performance local NVMe SSD disks with high IOPS, high I/O throughput, and low latency.
- vCPU to memory ratio = 1:4, designed for high-performance databases
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - OLTP and high-performance relational databases
  - NoSQL databases, such as Cassandra and MongoDB
  - Search applications, such as Elasticsearch

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.i2g. 2xlarge	8	32.0	1 * 894	2.0	1,000	2	4
ecs.i2g. 4xlarge	16	64.0	1 * 1788	3.0	1,500	4	8
ecs.i2g. 8xlarge	32	128.0	2 * 1788	6.0	2,000	8	8
ecs.i2g. 16xlarge	64	256.0	4 * 1788	10.0	4,000	16	8

Click *here* to view other instance type families.

i1, type family with local SSD disks

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks

- High-performance local NVMe SSD disks with high IOPS, high I/O throughput, and low latency
- vCPU to memory ratio = 1:4, designed for big data scenarios
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - OLTP and high-performance relational databases
  - NoSQL databases, such as Cassandra and MongoDB
  - Search applications, such as Elasticsearch

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.i1.xlarge	4	16.0	2 * 104	0.8	200	1	3
ecs.i1.2xlarge	8	32.0	2 * 208	1.5	400	1	4
ecs.i1.3xlarge	12	48.0	2 * 312	2.0	400	1	6
ecs.i1.4xlarge	16	64.0	2 * 416	3.0	500	2	8
ecs.i1-c5d1.4xlarge	16	64.0	2 * 1456	3.0	400	2	8
ecs.i1.6xlarge	24	96.0	2 * 624	4.5	600	2	8
ecs.i1.8xlarge	32	128.0	2 * 832	6.0	800	3	8
ecs.i1-c10d1.8xlarge	32	128.0	2 * 1456	6.0	800	3	8
ecs.i1.14xlarge	56	224.0	2 * 1456	10.0	1,200	4	8

Click *here* to view other instance type families.

hfc5, compute optimized type family with high clock speed

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- Stable performance
- 3.1 GHz Intel Xeon Gold 6149 (Skylake) processors
- vCPU to memory ratio = 1:2
- Higher computing specifications matching higher network performance
- Ideal for:
  - High-performance Web front-end servers
  - High-performance science and engineering applications
  - Massively Multiplayer Online (MMO) games and video coding

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.hfc5. large	2	4.0	N/A	1.0	300	2	2
ecs.hfc5. xlarge	4	8.0	N/A	1.5	500	2	3
ecs.hfc5. 2xlarge	8	16.0	N/A	2.0	1,000	2	4
ecs.hfc5. 3xlarge	12	24.0	N/A	2.5	1,300	4	6
ecs.hfc5. 4xlarge	16	32.0	N/A	3.0	1,600	4	8
ecs.hfc5. 6xlarge	24	48.0	N/A	4.5	2,000	6	8

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.hfc5. 8xlarge	32	64.0	N/A	6.0	2,500	8	8

Click *here* to view other instance type families.

hfg5, general-purpose type family with high clock speed

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- Stable performance
- 3.1 GHz Intel Xeon Gold 6149 (Skylake) processors
- vCPU to memory ratio = 1:4, except for the 56 vCPU instance type
- Higher computing specifications matching higher network performance
- Ideal for:
  - High-performance Web front-end servers
  - High-performance science and engineering applications
  - Massively Multiplayer Online (MMO) games and video coding

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.hfg5. large	2	8.0	N/A	1.0	300	2	2
ecs.hfg5. xlarge	4	16.0	N/A	1.5	500	2	3

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.hfg5. 2xlarge	8	32.0	N/A	2.0	1,000	2	4.
ecs.hfg5. 3xlarge	12	48.0	N/A	2.5	1,300	4	6
ecs.hfg5. 4xlarge	16	64.0	N/A	3.0	1,600	4	8
ecs.hfg5. 6xlarge	24	96.0	N/A	4.5	2,000	6	8
ecs.hfg5. 8xlarge	32	128.0	N/A	6.0	2,500	8	8
ecs.hfg5. 14xlarge	56	160.0	N/A	10.0	4,000	14	8

Click *here* to view other instance type families.

gn6v, compute optimized type family with GPUs

Features

- I/O-optimized
- Supports SSD Cloud Disk and Ultra Cloud Disk
- NVIDIA V100 GPU processors
- vCPU to memory ratio = 1:4
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Deep learning, autonomous vehicles, voice recognition, and other AI applications
  - Scientific computing, computational finance, genomics, and environmental analysis

Instance types

Instance types	vCPU	Memory (GiB)	Local disks (GiB)*	GPU	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.gn6v-c8g1.2xlarge	8	32.0	N/A	1 * NVIDIA V100	2.5	800	4	4
ecs.gn6v-c8g1.8xlarge	32	128.0	N/A	4 * NVIDIA V100	10.0	2,000	8	8
ecs.gn6v-c8g1.16xlarge	64	256.0	N/A	8 * NVIDIA V100	20.0	2,500	16	8

Note:

For more information, see *Create a compute optimized instance with GPUs*.

Click *here* to view other instance type families.

gn5, compute optimized type family with GPU

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- NVIDIA P100 GPU processors
- No fixed ratio of vCPU to memory
- High-performance local NVMe SSD disks
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Deep learning
  - Scientific computing, such as computational fluid dynamics, computational finance, genomics, and environmental analysis
  - High-performance computing, rendering, multi-media coding and decoding, and other server-side GPU compute workloads

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	GPU	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.gn5-c4g1.xlarge	4	30.0	440	1 * NVIDIA P100	3.0	300	1	3
ecs.gn5-c8g1.2xlarge	8	60.0	440	1 * NVIDIA P100	3.0	400	1	4
ecs.gn5-c4g1.2xlarge	8	60.0	880	2 * NVIDIA P100	5.0	1,000	2	4
ecs.gn5-c8g1.4xlarge	16	120.0	880	2 * NVIDIA P100	5.0	1,000	4	8
ecs.gn5-c28g1.7xlarge	28	112.0	440	1 * NVIDIA P100	5.0	1,000	8	8
ecs.gn5-c8g1.8xlarge	32	240.0	1760	4 * NVIDIA P100	10.0	2,000	8	8
ecs.gn5-c28g1.14xlarge	56	224.0	880	2 * NVIDIA P100	10.0	2,000	14	8
ecs.gn5-c8g1.14xlarge	54	480.0	3520	8 * NVIDIA P100	25.0	4,000	14	8

Note:

For more information, see *Create a compute optimized instance with GPUs*.

Click *here* to view other instance type families.

gn5i, compute optimized type family with GPU

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- NVIDIA P4 GPU processors
- vCPU to memory ratio = 1:4
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Deep learning
  - Multi-media coding and decoding, and other server-side GPU compute workloads

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	GPU	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.gn5i-c2g1.large	2	8.0	N/A	1 * NVIDIA P4	1.0	100	2	2
ecs.gn5i-c4g1.xlarge	4	16.0	N/A	1 * NVIDIA P4	1.5	200	2	3
ecs.gn5i-c8g1.2xlarge	8	32.0	N/A	1 * NVIDIA P4	2.0	400	4	4
ecs.gn5i-c16g1.4xlarge	16	64.0	N/A	1 * NVIDIA P4	3.0	800	4	8
ecs.gn5i-c16g1.8xlarge	32	128.0	N/A	2 * NVIDIA P4	6.0	1,200	8	8

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	GPU	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.gn5i-c28g1.14xlarge	56	224.0	N/A	2 * NVIDIA P4	10.0	2,000	14	8

Note:
For more information, see *Create a compute optimized instance with GPUs*.

Click *here* to view other instance type families.

gn4, compute optimized type family with GPU

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- NVIDIA M40 GPU processors
- No fixed ratio of CPU to memory
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Deep learning
  - Scientific computing, such as computational fluid dynamics, computational finance, genomics, and environmental analysis
  - High-performance computing, rendering, multi-media coding and decoding, and other server-side GPU compute workloads

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	GPU	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.gn4-c4g1.xlarge	4	30.0	N/A	1 * NVIDIA M40	3.0	300	1	3
ecs.gn4-c8g1.2xlarge	8	30.0	N/A	1 * NVIDIA M40	3.0	400	1	4
ecs.gn4.8xlarge	32	48.0	N/A	1 * NVIDIA M40	6.0	800	3	8
ecs.gn4-c4g1.2xlarge	8	60.0	N/A	2 * NVIDIA M40	5.0	500	1	4
ecs.gn4-c8g1.4xlarge	16	60.0	N/A	2 * NVIDIA M40	5.0	500	1	8
ecs.gn4.14xlarge	56	96.0	N/A	2 * NVIDIA M40	10.0	1,200	4	8

Note:

For more information, see *Create a compute optimized instance with GPUs*.

Click *here* to view other instance type families.

ga1, visualization compute type family with GPU

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- AMD S7150 GPU processors
- vCPU to memory ratio = 1:2.5
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors

- High-performance local NVMe SSD disks
- Higher computing specifications matching higher network performance
- Ideal for:
  - Rendering, multimedia coding and decoding
  - Machine learning, high-performance computing, and high-performance databases
  - Other server-end business scenarios that require powerful concurrent floating-point compute capabilities

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	GPU	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.ga1.xlarge	4	10.0	1 * 87	0.25 * AMD S7150	1.0	200	1	3
ecs.ga1.2xlarge	8	20.0	1 * 175	0.5 * AMD S7150	1.5	300	1	4
ecs.ga1.4xlarge	16	40.0	1 * 350	1 * AMD S7150	3.0	500	2	8
ecs.ga1.8xlarge	32	80.0	1 * 700	2 * AMD S7150	6.0	800	3	8
ecs.ga1.14xlarge	56	160.0	1 * 1400	4 * AMD S7150	10.0	1,200	4	8

Note:
For more information, see *Create an instance of ga1*.

Click *here* to view other instance type families.

f1, compute optimized type family with FPGA

Features

- I/O-optimized

- Supports SSD Cloud Disks and Ultra Cloud Disks
- Intel ARRIA 10 GX 1150 FPGA
- vCPU to memory ratio = 1:7.5
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Deep learning and reasoning
  - Genomics research
  - Financial analysis
  - Picture transcoding
  - Computational workloads, such as real-time video processing and security

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	FPGA	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.f1-c8f1.2xlarge	8	60.0	N/A	Intel ARRIA 10 GX 1150	3.0	400	4	4
ecs.f2-c8f1.4xlarge	16	120.0	N/A	2 * Intel ARRIA 10 GX 1150	5.0	1,000	4	8
ecs.f1-c28f1.7xlarge	28	112.0	N/A	Intel ARRIA 10 GX 1150	5.0	2,000	8	8
ecs.f2-c28f1.14xlarge	56	224.0	N/A	2 * Intel ARRIA 10 GX 1150	10.0	2,000	14	8

Click *here* to view other instance type families.

f2, compute optimized type family with FPGA

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- Xilinx Kintex UltraScale XCKU115
- vCPU to memory ratio = 1:7.5
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Deep learning and reasoning
  - Genomics research
  - Financial analysis
  - Picture transcoding
  - Computational workloads, such as real-time video processing and security

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	FPGA	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.f2-c8f1.2xlarge	8	60.0	N/A	Xilinx Kintex UltraScale XCKU115	2.0	800	4	4
ecs.f2-c8f1.4xlarge	16	120.0	N/A	2 * Xilinx Kintex UltraScale XCKU115	5.0	1,000	4	8
ecs.f2-c28f1.7xlarge	28	112.0	N/A	Xilinx Kintex UltraScale	5.0	1,000	8	8

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	FPGA	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
				XCKU115				
ecs.f2-c28f1.14xlarge	56	224.0	N/A	2 * Xilinx Kintex UltraScale XCKU115	10.0	2,000	14	8

Click *here* to view other instance type families.

f3, compute optimized type family with FPGA

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- Xilinx 16nm Virtex UltraScale + VU9P
- vCPU to memory ratio = 1:4
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors
- Higher computing specifications matching higher network performance
- Ideal for:
  - Deep learning and reasoning
  - Genomics research
  - Speeding up database access
  - Picture transcoding, such as converting JPEG to WebP
  - Real-time video processing, such as H.265 video compression

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	FPGA	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.f3-c16f1.4xlarge	16	64.0	N/A	1 * Xilinx VU9P	5.0	1,000	4	8
ecs.f3-c16f1.8xlarge	32	128.0	N/A	2 * Xilinx VU9P	10.0	2,000	8	8
ecs.f3-c16f1.16xlarge	64	256.0	N/A	4 * Xilinx VU9P	20.0	2,500	16	8

Click *here* to view other instance type families.

ebmhfg5, ECS Bare Metal Instance type family with high clock speed

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU to memory ratio = 1:4
- 3.7 GHz Intel Xeon E3-1240v6 (Skylake) processors, 8-core vCPU, up to 4.1 GHz Turbo Boot
- High network performance: 2 million pps packet forwarding rate
- Supports VPC network only
- Provides dedicated hardware resources and physical isolation
- Supports Intel SGX
- Ideal for:
  - Workloads that require direct access to physical resources, or scenarios where binding a license to the hardware is required
  - Gaming or financial applications featuring high performance
  - High-performance Web servers
  - Enterprise-level applications, such as high-performance databases

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENISs*****
ecs. ebmhfg5. 2xlarge	8	32.0	N/A	6.0	2,000	8	6

Note:
For more information about ECS Bare Metal Instance, see *ECS Bare Metal Instance and Super Computing Clusters*.

Click *here* to view other instance type families.

ebmc4, computing ECS Bare Metal Instance type family

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU to memory ratio = 1:2
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors, up to 2.9 GHz Turbo Boot
- High network performance: 4 million pps packet forwarding rate
- Supports VPC network only
- Provides dedicated hardware resources and physical isolation
- Ideal for:
  - Scenarios where a large volume of packets are received and transmitted, such as the re-transmission of telecommunication information
  - Third-party virtualization (includes but is not limited to Xen and KVM), and AnyStack (includes but is not limited to OpenStack and ZStack)
  - Containers (includes but is not limited to Docker, Clear Container, and Pouch)
  - Enterprise-level applications, such as medium and large databases
  - Video coding

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs. ebmc4. 8xlarge	32	64.0	N/A	10.0	4,000	8	12

Note:
For more information about ECS Bare Metal Instance, see *ECS Bare Metal Instance and Super Computing Clusters*.

Click *here* to view other instance type families.

ebmg5, general-purpose ECS Bare Metal Instance type family

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU to memory ratio = 1:4
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors, 96-core vCPU, up to 2.7 GHz Turbo Boot
- High network performance: 4 million pps packet forwarding rate
- Supports VPC network only
- Provides dedicated hardware resources and physical isolation
- Ideal for:
  - Workloads that require direct access to physical resources, or scenarios where binding a license to the hardware is required
  - Third-party virtualization (includes but is not limited to Xen and KVM), and AnyStack (includes but is not limited to OpenStack and ZStack)
  - Containers (includes but is not limited to Docker, Clear Container, and Pouch)
  - Enterprise-level applications, such as medium and large databases
  - Video encoding

Instance types

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs. ebmg5. 24xlarge	96	384.0	N/A	10.0	4,000	8	32

Note:
For more information about ECS Bare Metal Instance, see *ECS Bare Metal Instance and Super Computing Clusters*.

Click *here* to view other instance type families.

scch5, Super Computing Cluster (SCC) instance type family with high clock speed

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- Supports both RoCE and VPC networks, of which RoCE is dedicated to RDMA communication
- With all features of ECS Bare Metal Instance
- 3.1 GHz Intel Xeon Gold 6149 (Skylake) processors
- vCPU to memory ratio = 1:3
- Ideal for:
  - Large-scale machine learning applications
  - Large-scale high-performance scientific and engineering applications
  - Large-scale data analysis, batch computing, video encoding

Instance types

Instance type	vCPU	Memory (GiB)	GPU	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	RoCE (Inbound / Outbound) (Gbit/s)	NIC queues****	ENIs*****
ecs.scch5.16xlarge	64	192.0	N/A	10.0	4,500	46	8	32

Note:
For more information about SCC, see *ECS Bare Metal Instance and Super Computing Clusters*.

Click *here* to view other instance type families.

sccg5, geneneral-purpose Super Computing Cluster (SCC) instance type family

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- Supports both RoCE and VPC networks, of which RoCE is dedicated to RDMA communication
- With all features of ECS Bare Metal Instance
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors
- vCPU to memory ratio = 1:4
- Ideal for:
  - Large-scale machine learning applications
  - Large-scale high-performance scientific and engineering applications
  - Large-scale data analysis, batch computing, video encoding

Instance types

Instance type	vCPU	Memory (GiB)	GPU	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	RoCE (Inbound / Outbound) (Gbit/s)	NIC queues****	ENIs*****
ecs.sccg5.24xlarge	96	384.0	N/A	10.0	4,500	46	8	32

Note:
For more information about SCC, see *ECS Bare Metal Instance and Super Computing Clusters*.

Click *here* to view other instance type families.

t5, burstable instances

Features

- 2.5 GHz Intel Xeon processors
- The latest DDR4 memory
- No fixed ratio of vCPU to memory
- Baseline CPU performance, burstable, but restricted by accumulated CPU credits
- Resource balance among compute, memory, and networks
- Supports VPC network only
- Ideal for:
  - Web application servers
  - Lightweight web servers
  - Development and testing environments

Instance types

Instance type	vCPU	Memory (GiB)	CPU credits/ hour	Max CPU credit balance	Avg baseline CPU performance	ENIs*****
ecs.t5-lc2m1.nano	1	0.5	6	144	10%	1
ecs.t5-lc1m1.small	1	1.0	6	144	10%	1
ecs.t5-lc1m2.small	1	2.0	6	144	10%	1
ecs.t5-lc1m2.large	2	4.0	12	288	10%	1
ecs.t5-lc1m4.large	2	8.0	12	288	10%	1
ecs.t5-c1m1.large	2	2.0	18	432	15%	1
ecs.t5-c1m2.large	2	4.0	18	432	15%	1
ecs.t5-c1m4.large	2	8.0	18	432	15%	1
ecs.t5-c1m1.xlarge	4	4.0	36	864	15%	2
ecs.t5-c1m2.xlarge	4	8.0	36	864	15%	2
ecs.t5-c1m4.xlarge	4	16.0	36	864	15%	2
ecs.t5-c1m1.2xlarge	8	8.0	72	1,728	15%	2
ecs.t5-c1m2.2xlarge	8	16.0	72	1,728	15%	2
ecs.t5-c1m4.2xlarge	8	32.0	72	1,728	15%	2
ecs.t5-c1m1.4xlarge	16	16.0	144	3,456	15%	2

Instance type	vCPU	Memory (GiB)	CPU credits/ hour	Max CPU credit balance	Avg baseline CPU performance	ENIs*****
ecs.t5-c1m2 .4xlarge	16	32.0	144	3,456	15%	2

Note:
For more information about t5, see *Basic concepts*.

Click *here* to view other instance type families.

xn4/n4/mn4/e4, type families of previous generations for entry-level users, computing on the x86 -architecture

Features

- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- The latest DDR4 memory
- No fixed ratio of CPU to memory

Type family	Features	vCPU to memory ratio	Ideal for
xn4	Compact entry-level instances	1:1	• Front ends of Web applications • Light load applications and microservices • Applications for development or testing environments
n4	General entry-level instances	1:2	• Websites and Web applications • Development environment, building servers, code repositories,

Type family	Features	vCPU to memory ratio	Ideal for
			microservices, and testing and staging environment • Lightweight enterprise applications
mn4	Balanced entry-level instances	1:4	• Websites and Web applications • Lightweight databases and caches • Integrated applications and lightweight enterprise services
e4	Memory entry-level instances	1:8	• Applications that require large volume of memory • Lightweight databases and cache

xn4

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.xn4.small	1	1.0	N/A	0.5	50	1	1

n4

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.n4. small	1	2.0	N/A	0.5	50	1	1
ecs.n4. large	2	4.0	N/A	0.5	100	1	1
ecs.n4. xlarge	4	8.0	N/A	0.8	150	1	2
ecs.n4. 2xlarge	8	16.0	N/A	1.2	300	1	2
ecs.n4. 4xlarge	16	32.0	N/A	2.5	400	1	2
ecs.n4. 8xlarge	32	64.0	N/A	5.0	500	1	2

mn4

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.mn4. small	1	4.0	N/A	0.5	50	1	1
ecs.mn4. large	2	8.0	N/A	0.5	100	1	1
ecs.mn4. xlarge	4	16.0	N/A	0.8	150	1	2
ecs.mn4. 2xlarge	8	32.0	N/A	1.2	300	1	2
ecs.mn4. 4xlarge	16	64.0	N/A	2.5	400	1	2

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.mn4. 8xlarge	32	128.0	N/A	5	500	2	8

e4

Instance type	vCPU	Memory (GiB)	Local disks (GiB)*	Bandwidth (Gbit/s)**	Packet forwarding rate (Thousand pps)***	NIC queues****	ENIs*****
ecs.e4. small	1	8.0	N/A	0.5	50	1	1
ecs.ce4. xlarge	2	16.0	N/A	0.5	100	1	1
ecs.ce4. xlarge	4	32.0	N/A	0.8	150	1	2
ecs.e4. 2xlarge	8	64.0	N/A	1.2	300	1	3
ecs.ce4. xlarge	16	128.0	N/A	2.5	400	1	8

Click *here* to view other instance type families.

* *Cache disks*, or *Local disks*, are the disks located on the physical servers (host machines) that ECS instances are hosted on. They provide temporary block level storage for instances. Block storage capacity is measured in binary units. In some cases, such as when the computing resources of an instance, including CPU and memory, are released, or an instance is inactive while migration occurs, data on the local disks is erased. For more information, see *Local disks*.

** The maximum bandwidth of inbound and outbound traffic.

*** The maximum packet forwarding rate of inbound and outbound traffic. For more information about packet forwarding rate testing, see *Test network performance*.

**** The maximum number of NIC queues that an instance type supports. If your instance is running CentOS 7.3, the maximum number of NIC queues is used by default.

***** An enterprise-level instance with two or more vCPU cores supports elastic network interfaces. An entry-level instance with four or more vCPU cores supports elastic network interfaces. For more information about elastic network interfaces, see *Elastic network interfaces*.

5 Instances

5.1 What are ECS instances

An ECS instance is a virtual computing environment that includes CPU, memory, operating system, bandwidth, disks, and other basic computing components.

An ECS instance is an independent virtual machine, and is the core element of ECS. Other resources, such as disks, IPs, images, and snapshots can only be used in conjunction with an ECS instance.

5.2 ECS instance life cycle

The life cycle of an ECS instance begins when it is created and ends when it is released.

Instance status

During this process, an ECS instance may undergo several status changes, as explained in the following table.

Status	Status attribute	Description	Corresponding API status	Viewable in the console
Preparing	Intermediate	After an instance is created, it remains in this status before running. If an instance is in this status for a long time, an exception occurs.	Pending	No
Starting	Intermediate	An instance is in this status when it is either *started* or *restarted* in the console or by using an API before it is running. If an instance is in this status for	Starting	Yes

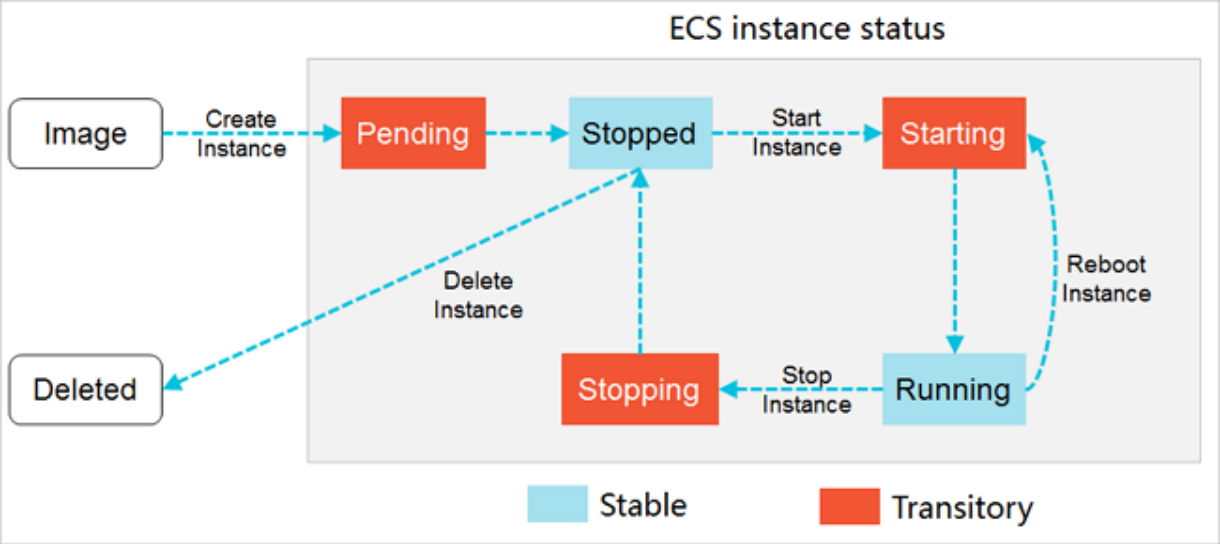
Status	Status attribute	Description	Corresponding API status	Viewable in the console
		a long time, an exception occurs.		
Running	Stable	The instance is operating normally and can accommodate your business needs.	Running	Yes
Stopping	Intermediate	An instance is in this status after the stop operation is performed in the console or when using an API but before the instance actually stops. If an instance is in this status for a long time, an exception occurs.	Stopping	Yes
Stopped	Stable	The instance has been stopped properly. In this status, the instance cannot accommodate external services.	Stopped	Yes
Expired	Stable	A yearly or monthly subscribed instance is in this status if it expires because it has not been timely renewed. A Pay-As-You-Go instance is in this status only when	Stopped	Yes

Status	Status attribute	Description	Corresponding API status	Viewable in the console
		you have an overdue payment. After an ECS instance expires, it continues running for 15 days, and the data on its disks is retained for an additional 15 days, after which the instance will be released and the data will be permanently removed. In this status, the instance cannot accommodate external services.		
Expiring	Stable	A Subscription instance is in this status for 15 days before it expires. After it is *renewed*, the instance is in the **Running**status.	Stopped	Yes
Locked	Stable	An instance is in this status because of an overdue account or security risks. To unlock the instance, *open a ticket*.	Stopped	Yes
Release pending	Stable	A Subscription instance is in this status after you	Stopped	Yes

Status	Status attribute	Description	Corresponding API status	Viewable in the console
		apply for a refund before it expires.		

API status changes

The following figure illustrates API status changes of an instance within its life cycle.

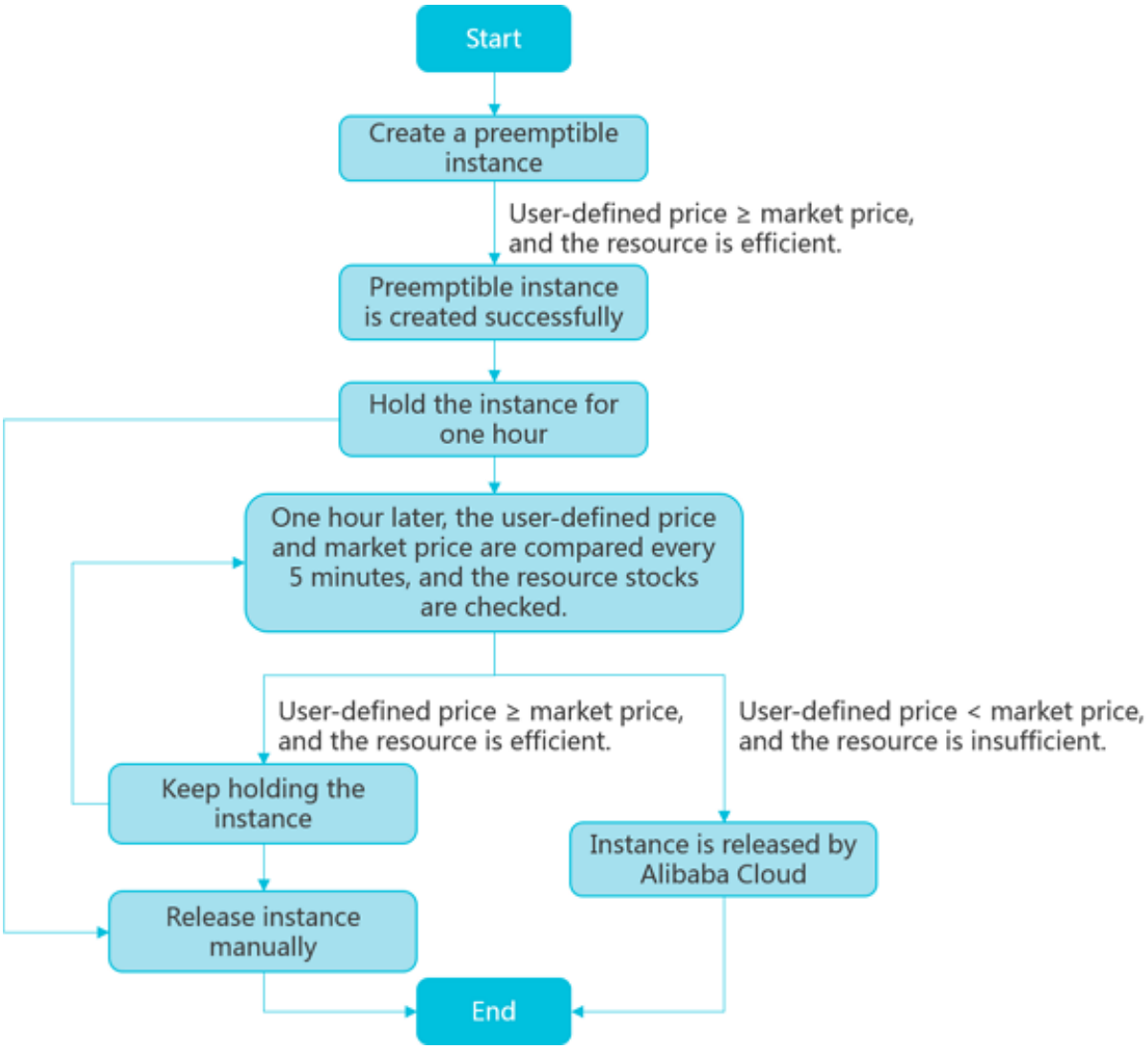


5.3 Preemptible instance

Preemptible instances are on-demand instances. They are designed to reduce your ECS costs in some cases. When you create a preemptible instance, you can set a maximum price per hour to bid for a specified instance type. If your bid is higher than or equal to the current market price, your instance is created. A preemptible instance is held without interruption for at least one hour after it is created. After one hour, your bid is compared with the market price every five minutes. When the market price exceeds your bid or the resource stock is insufficient, the instance is automatically released. The following figure shows the life cycle of a preemptible instance.

Note:

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



Scenarios

Preemptible instances are ideal for stateless applications, such as scalable Web services and applications for rendering figures, big data analysis, and massively parallel computing. Applicatio ns requiring higher level of distribution, scalability, and fault tolerance capability benefit from preemptible instances with respect to costs and throughput.

You can deploy the following businesses on preemptible instances:

- Real-time analysis
- Big data
- Geological surveys
- Image coding and media coding
- Scientific computing
- Scalable Web sites and Web crawlers

- Image and media coding
- Testing

Preemptible instances are not suitable for stateful applications, such as databases, because it is difficult to store application states if the instance is released because of a failed bid or other reasons.

Bidding modes

You can bid for a preemptible instance only one time in either of the following bidding modes:

- *SpotWithPriceLimit*
- *SpotAsPriceGo*

SpotWithPriceLimit

In this mode, you must set the highest price you want to pay for a specified instance type. When creating a preemptible instance by using *RunInstances*, you can bid in this mode.

Currently, the maximum bid of a preemptible instance is the price of a Pay-As-You-Go instance of the same configuration. When creating a preemptible instance, you can set a price according to the market price history, business features, and the estimated future price fluctuation. When the market price is lower than or equal to your bid, and the resource stock is sufficient, the instance continues to run. If your estimated quote is accurate, you can hold the instance after the one hour *guaranteed duration*. Otherwise, your instance can get automatically released at any time.

SpotAsPriceGo

When creating a preemptible instance by using *RunInstances*, you can create a preemptible instance with the SpotAsPriceGo bidding mode by setting `SpotStrategy` to `SpotAsPriceGo`, which means you always set the real-time market price as the bidding price until the instance is released because of stock shortage.

Guaranteed duration

When a preemptible instance is created, it has a guaranteed duration of one hour, namely, the first hour after it is run. During this period, the instance is not released because of stock shortage, and you can run services on the instance as usual. Beyond the guaranteed duration, the market price and stock is checked every five minutes. If the market price at any given point in time is higher than your bid or the instance type stock is insufficient, your preemptible instance will be automatically released.

Price and billing

Preemptible instance price and billing considerations:

- **Price**

  The preemptible instance price applies to the instance type only, including vCPU and memory, but not to system disks, data disks, or network bandwidth. The prices for system disks, data disks, or network bandwidth are the same as for *Pay-As-You-Go* instances.

- **Billing cycle**

  Preemptible instances are billed on an hourly basis during their life cycles. You are billed for the entire hour even if your usage is less than an hour.

- **Billing duration**

  Instances are billed according to the actual period of use. The actual period of use is the duration from instance creation to instance release. After an instance is released, it is no longer billed. If you stop the instance by using *StopInstance* or in the *ECS console*, the instance continues to be billed.

- **Market price**

  During creation of a preemptible instance, it runs when your bid is higher than the current market price and the relevant demand and supply conditions are satisfied. The final price you pay for your instance type is based on the current market price.

The actual market price of a preemptible instance fluctuates according to the changes in the demand and supply of a given instance type, and you can take advantage of these price fluctuatio ns. If you purchase preemptible instance types at the right time, the computing costs are reduced, whereas your throughput is increased for the period the instance is held.

Quota

For more information about the preemptible instance quota, see *Limits*.

Create a preemptible instance

You can purchase a preemptible instance by using the *RunInstances* interface.

After a preemptible instance is created, it can be used in exactly the same way as a Pay-As-You-Go instance. You can also use it with other cloud products, such as cloud disks or EIP addresses.

Stop a preemptible instance

You can stop a preemptible instance in the *ECS console* or by using the *StopInstance* interface. The VPC-Connected preemptible instances support the *No fees for stopped instances (VPC-Connected)* feature.

The network type and the bidding mode of a preemptible instance determine whether it can start after it is stopped, as displayed in the following table.

Network type + Bidding mode	Stop instance	Start instance
VPC + SpotWithPriceLimit	Keep Instance, Fees Apply	During the guaranteed duration, the instance can be started successfully. After the guaranteed duration: • If your bid is not lower than the market price and the resource stock is sufficient, the instance can be started successfully. • If your bid is lower than the market price or the resource stock is insufficient, the instance cannot be started successfully.
Classic + SpotWithPriceLimit	N/A	
VPC + SpotAsPriceGo	Keep Instance, Fees Apply	During the guaranteed duration, the instance can be started successfully. After the guaranteed duration: • If the resource stock is sufficient, the instance can be started successfully. • If the resource stock is insufficient, the instance cannot be started.
Classic + SpotAsPriceGo	N/A	
VPC + SpotWithPriceLimit	Stop Instance, No Fees	During the guaranteed duration, the instance can be started successfully only if the resource stock is sufficient. After the guaranteed duration:

Network type + Bidding mode	Stop instance	Start instance
		• If your bid is not lower than the market price and the resource stock is sufficient, the instance can be started successfully. • If your bid is lower than the market price or the resource stock is insufficie nt, the instance cannot be started successfully.
VPC + SpotAsPriceGo	Stop Instance, No Fees	During the guaranteed duration, the instance can be started successfully only if the resource stock is sufficient. After the guaranteed duration: • If the resource stock is sufficient, the instance can be started successfully. • If the resource stock is insufficient, the instance cannot be started.

Release a preemptible instance

When the guaranteed period ends, we automatically release your preemptible instance because of changes in the market price or short resource stock. Additionally, you can independently *release the instance*.

When a preemptible instance is released because of market price or changes in the demand and supply of resources, the instance's status changes to **Pending Release**. Then, the instance is released in about five minutes. You can use *instance metadata* or the `OperationLocks` information returned by calling the *DescribeInstances* interface to check if an instance is in the **Pending Release** status.

Note:
Although you can check if a preemptible instance is in the **Pending Release** status by using the API and save a small amount of data while the instance is in this status, we recommend that

you design your applications so work can be properly resumed if the preemptible instance is immediately released. When you release the instance manually, you can test whether or not your application functions normally if a preemptible instance is immediately recovered.

Generally, we release preemptible instances in the order of bidding price, from low to high. If multiple preemptible instances have the same bidding price, they are randomly released.

Best practices

When using a preemptible instance, consider the following:

- Set an appropriate bidding price. In other words, you must quote a competitive price to meet your business budget and hedge against the future market price fluctuations. By using this price, your preemptible instance can be created. In addition, the price must meet your expectations based on your own business assessment.
- The image must have all the software configurations that your applications need, assuring that you can run your business immediately after the instance is created. Additionally, you can use *User-defined data* to run commands at startup.
- Store your business data on storage products that are independent from preemptible instances , such as cloud disks that are not set to release together with instances, OSS, or RDS.
- Split your tasks by using grids, Hadoop, queuing-based architecture, or check points, to facilitate store computing results frequently.
- Use the release notification to monitor the status of a preemptible instance. You can use *metadata* to check the instance status every minute. The metadata of an instance is updated five minutes before it is released automatically.
- Test your applications in advance, to make sure that they can handle events such as accidental release of an instance. To test the applications: Run the applications on a Pay-As-You-Go instance, release the instance, and then check how the applications can handle the release.

For more information, see *FAQ about preemptible instances*.

For more information about using APIs to create preemptible instances, see *Use APIs to manage preemptible instances*.

5.4 ECS Bare Metal Instance and Super Computing Clusters

ECS Bare Metal (EBM) Instance is a new type of computing product that features the elasticity of virtual machines and the performance and characteristics of physical machines. As a product completely and independently developed by Alibaba Cloud, EBM Instance is based on the next

-generation of virtualization technology. Compared with the previous generation, not only is the common virtual cloud server supported, but also is nested virtualization. The resource elasticity of common cloud servers is retained, while nested virtualization technology creates a user experience comparable to physical machines.

Super Computing Clusters (SCCs) are based on EBM Instances. With the help of the high-speed interconnectivity of RDMA (Remote Direct Memory Access) technology, SCCs greatly enhance network performance and increase the acceleration ratio of large-scale clusters. SCCs have all the advantages of EBM Instances and offer high-quality network performance featuring high bandwidth and low latency.

Advantages

EBM Instances

EBM Instances create value for customers through technological innovation. EBM Instances have the following advantages:

- **Exclusive computing resources**

  As a cloud-based elastic computing product, the EBM Instances exceed the performance and isolation of contemporary physical machines and enable exclusive computing resources without virtualization performance overheads and feature loss. EBM Instances support 8, 16, 32, and 96 CPU cores and ultrahigh frequency. An EBM Instance with 8 cores, for example, supports an ultrahigh frequency of up to 3.7 to 4.1 GHz, providing better performance and responsiveness for gaming and financial industries than similar products.

- **Encrypted compute**

  For security, the EBM Instances use a chip-level trusted execution environment (Intel® SGX) in addition to physical server isolation. This allows the instances to compute only the encrypted data in a safe and trusted environment, and provides improved security for the customer data on the cloud. This chip-level hardware security protection provides a safe box for the data of cloud users and allows users to control all the data encryption and key protection procedures. For more information, see *Intel SGX*.

- **Any Stack on Alibaba Cloud**

  EBM Instances combine the performance strengths and features of physical machines and *the ease-of-use and cost-effectiveness of cloud servers*. They can effectively meet your demands for high-performance computing and help you build new hybrid clouds. Due to their flexibility, elasticity, and other strengths, EBM Instances allow you to deploy any stack on them, such

as Xen, KVM, and VMWare. As a result, offline private clouds can be seamlessly migrated to Alibaba Cloud without the performance overhead issues that may arise because of nested virtualization. This facilitates a new approach for you to move businesses onto the cloud.

- **Heterogeneous instruction set processor support**

  The virtualization 2.0 technology used by EBM Instances is independently developed by Alibaba Cloud. It can zero-cost support ARM and other instruction set processors.

SCC

SCCs are based on EBM Instance, and were released by Alibaba Cloud to meet the demands of applications such as high performance computing, artificial intelligence, machine learning, scientific or engineering computing, data analysis, and audio and video processing. In the clusters, nodes are connected by Remote Direct Memory Access (RDMA) networks featuring high bandwidth and low latency, guaranteeing the highly parallel efficiency demanded by applications that require high-performance computing. Meanwhile, the RoCE (RDMA over Convergent Ethernet) rivals an Infiniband network in terms of connection speed, and supports more extensive Ethernet-based applications. The combination of SCCs built on EBM Instance and other Alibaba Cloud computing products such as ECS and GPU servers provides *the Alibaba Cloud elastic high-performance computing (E-HPC) platform* with ultimate high performance parallel computing resources, making supercomputing on the cloud a reality.

Features

EBM Instances and SCC have the following features:

- CPU specifications:
  - EBM Instances: Supports 8 cores, 16 cores, 32 cores, and 96 cores, and supports high clock speed.
  - SCC: Supports 64 cores and 96 cores, and provides support for high clock speed.
- Memory specifications:
  - EBM Instances: Supports 32 GiB to 768 GiB memory. For better computing performance, the ratio of CPU to memory is 1:2 or 1:4.
  - SCC: The ratio of CPU to memory is 1:3 or 1:4.
- Storage specifications: Supports starting from the virtual machine image and cloud disk to deliver instances in minutes.
- Network configurations:

  - Supports Virtual Private Cloud (VPC) networks, maintaining interoperability with ECS, GPU cloud servers, and other cloud products. Delivers performance and stability comparable to physical machine networks.
  - (Only for SCC) Supports RDMA communication through high-speed RoCE networks.
- Images: Supports images of Alibaba Cloud ECS.
- Security settings: Maintains the same security policies and flexibility as existing cloud server ECS instances.

The following table compares EBM Instance or SCC, physical servers, and virtual servers.

Feature type	Features	EBM Instances/ SCC	Physical servers	Virtual servers
Automated O&M	Delivery in minutes	Y	N	Y
Computing	Zero performance loss	Y	Y	N
	Zero feature loss	Y	Y	N
	Zero resource competition	Y	Y	N
Storage	Fully compatible with ECS cloud disks	Y	N	Y
	Start from cloud disks (system disks)	Y	N	Y
	System disk can be quickly reset	Y	N	Y
	Uses ECS images	Y	N	Y
	Supports cold migration between physical and virtual servers	Y	N	Y
	Requires no installation of operating system	Y	N	Y

Feature type	Features	EBM Instances/ SCC	Physical servers	Virtual servers
	Discards local RAID, and provides stronger protection of data on cloud disks	Y	N	Y
Network	Fully compatible with the ECS VPC networks	Y	N	Y
	Fully compatible with the ECS classic networks	Y	N	Y
	Free of bottlenecks for communications between physical and virtual server clusters in the VPC	Y	N	Y
Management	Fully compatible with the existing ECS management system	Y	N	Y
	Consistent user experience on VNC and other features with that of virtual servers	Y	N	Y
	Guaranteed OOB network security	Y	N	N/A

Instance type families

The type families of EBM Instances include:

- ebmg5, general purpose EBM Instance type family
- ebmhfg5, high frequency EBM Instance type family
- ebmc4, compute EBM Instance type family

The type families of SCC include:

- scch5, Super Computing Cluster (SCC) instance type family with high clock speed
- sccg5, geneneral-purpose Super Computing Cluster (SCC) instance type family

For more information, see *EBM Instance type families* and *SCC Instance type families*.

Billing methods

EBM Instances and SCC instances support Subscription and Pay-As-You-Go. For more information about billing methods, see *Billing method comparison*.

Related operations

You can *create an EBM instance* or *create an SCC server instance* in the console.

For more information, see *FAQs about EBM Instances* and *SCC FAQ*.

5.5 Launch templates

A launch template helps you quickly create an ECS instance. A template contains configurations that you can use to create instances for various scenarios with specific requirements.

A template can include any configurations except passwords. It can include key pairs, RAM roles, instance type, and network configurations.

You can create multiple versions of each template. Each version can contain different configurat ions. You can then create an instance using any version of the template.

Console operations

- *Create a template*
- *Create multiple versions in one template*
- *Change the default version*
- *Use a launch template*
- *Delete a template or version*

API operations

- *CreateLaunchTemplate*
- *CreateLaunchTemplateVersion*
- *DescribeLaunchTemplates*
- *DescribeLaunchTemplateVersions*
- *ModifyLaunchTemplateDefaultVersion*

- *DeleteLaunchTemplate*
- *DeleteLaunchTemplateVersion*

5.6 Burstable instances

5.6.1 Basic concepts

Burstable instances (also called t5 instances) provide a baseline level of CPU performance with the ability to burst above the baseline. Each t5 instance provides a baseline CPU performance and earns CPU credits at a specified rate based on instance types. A t5 instance consumes CPU credits to meet service requirements once it is started. When the required performance is higher than the baseline, the instance consumes more CPU credits to seamlessly increase CPU performance without affecting the environment or applications on the instance.

t5 instances come with two running modes: *Standard* and *Unlimited*.

Concepts

- **Baseline CPU performance**

  Each t5 instance type provides a baseline level of CPU performance,

  - which means that each vCPU core has a usage limit under normal workloads. For example, when the ecs.t5-lc1m2.small standard instance runs under normal workloads, the maximum CPU usage is 10%. More credits will be consumed to burst above that baseline. After the credits are used up, the maximum CPU usage is 10%.
  - By contrast, t5 unlimited instances are not restricted by the baseline and can maintain high CPU performance for any time period. However, fees are charged for excess credits.

- **CPU credits**

  Each t5 instance earns CPU credits at a fixed rate based on the baseline CPU performance . One CPU credit represents the computing performance, which is related to the number of vCPU cores, CPU usage, and running time. For example:

  - One CPU credit = One vCPU core running at 100% usage for 1 minute
  - One CPU credit = One vCPU core running at 50% usage for 2 minutes
  - One CPU credit = Two vCPU cores running at 25% usage for 2 minutes

  To support a vCPU core running at 100% usage for an hour, 60 CPU credits are required.

- **Initial CPU credits**

Every time you create a t5 instance, each vCPU core of the instance is immediately allocated 30 CPU credits, which are called initial CPU credits. Initial CPU credits are allocated only upon instance creation. Additionally, they are consumed first when the instance starts to spend credits.

- **CPU credit acquisition rate**

  t5 instances acquire CPU credits on a per-minute basis. The CPU credit acquisition rate indicates CPU credits acquired by a t5 instance per unit time (minute). It is determined by the baseline CPU performance. The calculation formula is as follows:

  ```
  CPU credit acquisition rate = Baseline CPU performance x number of vCPUs
  ```

  Example: Use the ecs.t5-c1m2.xlarge instance as an example. Its average baseline CPU performance is 15%, so the CPU credit acquisition rate is 0.6 CPU credit per minute, that is, 36 CPU credits per hour.

- **CPU credit consumption**

  A t5 instance consumes CPU credits once it is started. Initial CPU credits are consumed first. The formula for calculating the consumed CPU credits per minute is as follows:

  ```
  CPU credits consumed per minute = One CPU credit x actual CPU performance
  ```

  Example: Use the ecs.t5-lc1m2.small instance as an example. It consumes 0.2 CPU credit when it runs at 20% CPU usage for one minute.

- **Accrued CPU credits**

  When the CPU usage of a t5 instance is lower than the baseline CPU performance, the instance accrues CPU credits because the CPU credit consumption rate is lower than the CPU credit acquisition rate. Otherwise, the instance consumes CPU credits overall. The CPU credit accrual rate is determined by the difference between the actual CPU load and the baseline performance. It can be calculated using the following formula:

  ```
  CPU credits accrued per minute = One CPU credit x (baseline CPU performance - actual CPU performance)
  ```

  You can *view the accrued and consumed CPU credits* on the ECS console.

- **Max. CPU credit balance**

  CPU credits increase when the CPU credit acquisition rate is higher than the CPU credit consumption rate. The accrued credits do not expire on a running instance. However, there is

an upper limit for the credits that can be accrued by an instance, namely, the maximum CPU credit balance. The upper limit varies with instance types.

Taking ecs.t5-lc2m1.nano as an example, the maximum CPU credit balance is 144. When the CPU credit balance reaches 144, accrual pauses. When the balance is lower than 144, accrual restarts.

The initial credits are not included in the credit balance.

How stopping an instance impacts CPU credits

After you stop a t5 through the *view CPU utilization and CPU credits* feature or the *StopInstance API*, CPU credits change according to the billing method and network type, as shown in the following table.

Network type	Billing method	How CPU credits change after the instance is stopped
Classic network	Subscription or Pay-As-You-Go	The existing CPU credits are valid, and the credit accrual continues.
VPC	Subscription	
	Pay-As-You-Go (with the *no fees for stopped VPC instances* function disabled)	
	Pay-As-You-Go (with *no fees for stopped VPC instances* function enabled)	CPU credits accrued before the stoppage become invalid. The instance acquires initial CPU credits again after it is restarted.

The instance continues to accrue CPU credits after it is restarted.

If a Pay-As-You-Go instance has overdue payment or a Subscription instance expires, its CPU credits remain valid, but no new CPU credits will be accrued. After you *reactivate* or *renew* the instance, it automatically accrues CPU credits.

Instance types

t5 instances use Intel Xeon processors. The following table lists instance types. In this table:

- **CPU credits/hour** indicates the total CPU credits allocated to all vCPU cores of a t5 instance per hour.

- **Average baseline CPU performance** indicates the average baseline CPU performance of each vCPU core of a t5 instance.

Instance type	vCPU	Average baseline CPU performance	Initial CPU credits	CPU credits/ hour	Max. CPU credit balance	Memory (GiB)
ecs.t5-lc2m1.nano	1	10%	30	6	144	0.5
ecs.t5-lc1m1.small	1	10%	30	6	144	1.0
ecs.t5-lc1m2.small	1	10%	30	6	144	2.0
ecs.t5-lc1m2.large	2	10%	60	12	288	4.0
ecs.t5-lc1m4.large	2	10%	60	12	288	8.0
ecs.t5-c1m1.large	2	15%	60	18	432	2.0
ecs.t5-c1m2.large	2	15%	60	18	432	4.0
ecs.t5-c1m4.large	2	15%	60	18	432	8.0
ecs.t5-c1m1.xlarge	4	15%	120	36	864	4.0
ecs.t5-c1m2.xlarge	4	15%	120	36	864	8.0
ecs.t5-c1m4.xlarge	4	15%	120	36	864	16.0
ecs.t5-c1m1.2xlarge	8	15%	240	72	1,728	8.0
ecs.t5-c1m2.2xlarge	8	15%	240	72	1,728	16.0
ecs.t5-c1m4.2xlarge	8	15%	240	72	1,728	32.0

Instance type	vCPU	Average baseline CPU performance	Initial CPU credits	CPU credits/ hour	Max. CPU credit balance	Memory (GiB)
ecs.t5-c1m1.4xlarge	16	15%	480	144	3,456	16.0
ecs.t5-c1m2.4xlarge	16	15%	480	144	3,456	32.0

Examples

- The following uses the ecs.t5-c1m1.xlarge instance as an example.
  - For each vCPU core, the average baseline performance is 15%. Therefore, the total baseline performance of the instance is 60% (4 vCPU x 15%). Details are as follows:
    - When the instance uses only one vCPU core, this core provides the baseline performance of 60%.
    - When the instance uses two vCPU cores, each core is allocated the baseline performance of 30%.
    - When the instance uses three vCPU cores, each core is allocated the baseline performance of 20%.
    - When the instance uses all four vCPU cores, each core is allocated the baseline performance of 15%.

    Note:
    When the business needs arise, CPU credits are consumed to improve the CPU performance. The performance of each vCPU core can increase to 100%.
  - An instance acquires 36 CPU credits per hour, which means that each vCPU core acquires nine CPU credits per hour.
- The following uses the ecs.t5-c1m2.4xlarge instance as an example.
  - For each vCPU core, the average baseline performance is 15%. Therefore, the total baseline computing performance of the instance is 240% (16 vCPU x 15%). Details are as follows:

- When the instance uses only one vCPU core, this core provides the baseline performanc e of 100%.
- When the instance uses two vCPU cores, each core is allocated the baseline performanc e of 100%.
- When the instance uses three vCPU cores, each core is allocated the baseline performance of 80%.
- When the instance uses all 16 vCPU cores, each core is allocated the baseline performance of 15%.

> **Note:**
> When the business needs arise, CPU credits are consumed to improve the CPU performance. The performance of each vCPU core can increase to 100%.

— An instance acquires 144 CPU credits per hour, which means that each vCPU core acquires nine CPU credits per hour.

Billing methods

t5 instances support both the Pay-As-You-Go and Subscription billing methods. For differences between the billing methods, see *billing method comparison*.

5.6.2 t5 standard instances

t5 standard instances are ideal for scenarios where you do not usually, but occasionally require high CPU performance, such as lightweight Web servers, development and testing environments, and low or mid-performance databases.

If the instance has accrued few credits, its performance gradually declines to the baseline level within 15 minutes, so that the instance performance does not drop dramatically when the accrued CPU credits are used up. When the accrued CPU credits are used up, the actual CPU performanc e of the t5 instance cannot be higher than the baseline CPU performance.

Fees

There is no additional fee except the cost of creating an instance.

Examples

Take a t5 standard instance of the ecs.t5-lc1m2.small type as an example. The following describes how its CPU credits change:

1. When the instance is created, 30 initial CPU credits are allocated to it. That is, the total CPU credits are 30 before it is started. After it is started, CPU credits accrue at the rate of 0.1 credits per minute. Meanwhile, the credits are consumed during its running.
2. During the first minute, if the CPU usage is 5%, 0.05 initial CPU credits are consumed, while 0.1 CPU credits are allocated. Therefore, 0.05 CPU credits are accrued.
3. After the instance has started for N minutes, if the CPU usage is 50%, 0.5 CPU credits are consumed while 0.1 CPU credits are allocated within one minute. Therefore, 0.4 CPU credits are consumed during this one minute.
4. When the accrued CPU credits are used up, the maximum CPU usage is 10%.

5.6.3 t5 unlimited instances

t5 unlimited instances can maintain high CPU performance for any period of time, without being limited to the baseline CPU performance.

Concepts

In addition to the *basic concepts*, you also need to understand the following concepts before using t5 unlimited instances:

- **Advance credits**

  The credits that are used in advance but should be obtained within the next 24 hours.

- **Excess credits**

  After the credits for the next 24 hours are used up, additional credits incur fees that are billed by the hour.

When a t5 unlimited instance runs out of its CPU credit balance, advance credits are used first to address the requirement of high CPU performance. When the CPU usage is lower than the baseline, the earned CPU credits are used to pay down (offset) the advance credits.

Billing rules

- **Fees are not charged in the following cases:**
  - The hourly t5 instance price automatically covers all interim spikes in usage if the average CPU utilization of the instance is at or below the baseline over a 24-hour period or the instance lifetime, whichever is shorter. You do not have to pay additional fees.
  - An instance earns a maximum number of credits in a 24-hour period. For example, a t5-lc1m1.small instance can earn a maximum of 144 credits in a 24-hour period. When the advance credits are less than that maximum, no additional fees are charged.

- **Fees are charged in the following cases:**
  - If the consumed advance credits exceed the maximum credits (that is, excess credits generated), fees are charged at the end of the time period.
  - If advance credits are used and the instance is stopped or released before the advance credits are cleared, a one-off fee is charged for the advance credits.
  - If excess credits are used after advance credits are used up, additional fees are charged.
  - If a t5 instance is changed from unlimited mode to standard mode, the fees for advance credits are charged immediately, and the accrued credits remain unchanged.

    The fees charged are shown in the following table:

Region	Windows instance (USD/credit)	Linux instance (USD/credit)
Mainland China	0.0008	0.0008
Other regions	0.0016	0.0008

Examples

Use a t5-lc1m1.small unlimited instance (purchased in US West 1 region) as an example. The following describes how its CPU credits change.

1. After a Linux instance is created, it is allocated 30 initial CPU credits. When the instance starts, it is able to spend 144 credits in advance, which are the maximum of CPU credits for the next 24 hours. Therefore, there are 174 CPU credits when the instance starts.
2. After the instance starts, assuming the CPU utilization is 50%, the instance consumes 0.5 initial CPU credits per minute, while 0.1 CPU credits are allocated in the meantime. As a result, CPU credits continue to decrease.
3. After the instance has been running for N minutes, assuming the accrued CPU credits are used up, the advance credits are spent to maintain high CPU performance.
4. After the instance has been running for N + X minutes, assuming the 144 advance credits are used up, excess credits are used to maintain high CPU performance.
5. After the instance has been running for N + X + Y minutes, assuming 50 excess credits are used and the CPU utilization drops to 5% (below the baseline), the instance begins to earn 0.1 credits per minute, which are used to pay down (offset) the consumed advance credits. When the advance credits are restored to 144, credits begin to accrue to the instance (0.1 credits per minute).

Consumption details

At the end of the N + X + Y minutes, if excess credits are no longer used, fees are charged.

During the above time period, the Linux instance uses 50 excess credits. The additional fee is: 0.0016 USD/credit x 50 credits = 0.08 USD.

5.6.4 Manage t5 instances

You can use the console or the API to create t5 instances and change their instance types.

Create an instance

You can create a t5 instance following *creating an instance by using the wizard*. When creating a t5 instance, consider the following settings:

- Network type: Only VPC is supported.
- Image: 512 MiB is the minimum memory size for a t5 instance, and you can select only the Linux or Windows Server 1709 operating system. Operating systems that require a minimum memory of 1 GiB, such as Windows Server 2016, are not supported. For more information about selecting images, see *how to select a system image*.
- t5 instance type: Select the **Enable Unlimited Mode for T5 Instances** check box to create a t5 instance without performance constraints. If it is not selected, a t5 standard instance is created. You can change the mode after creating a t5 instance.

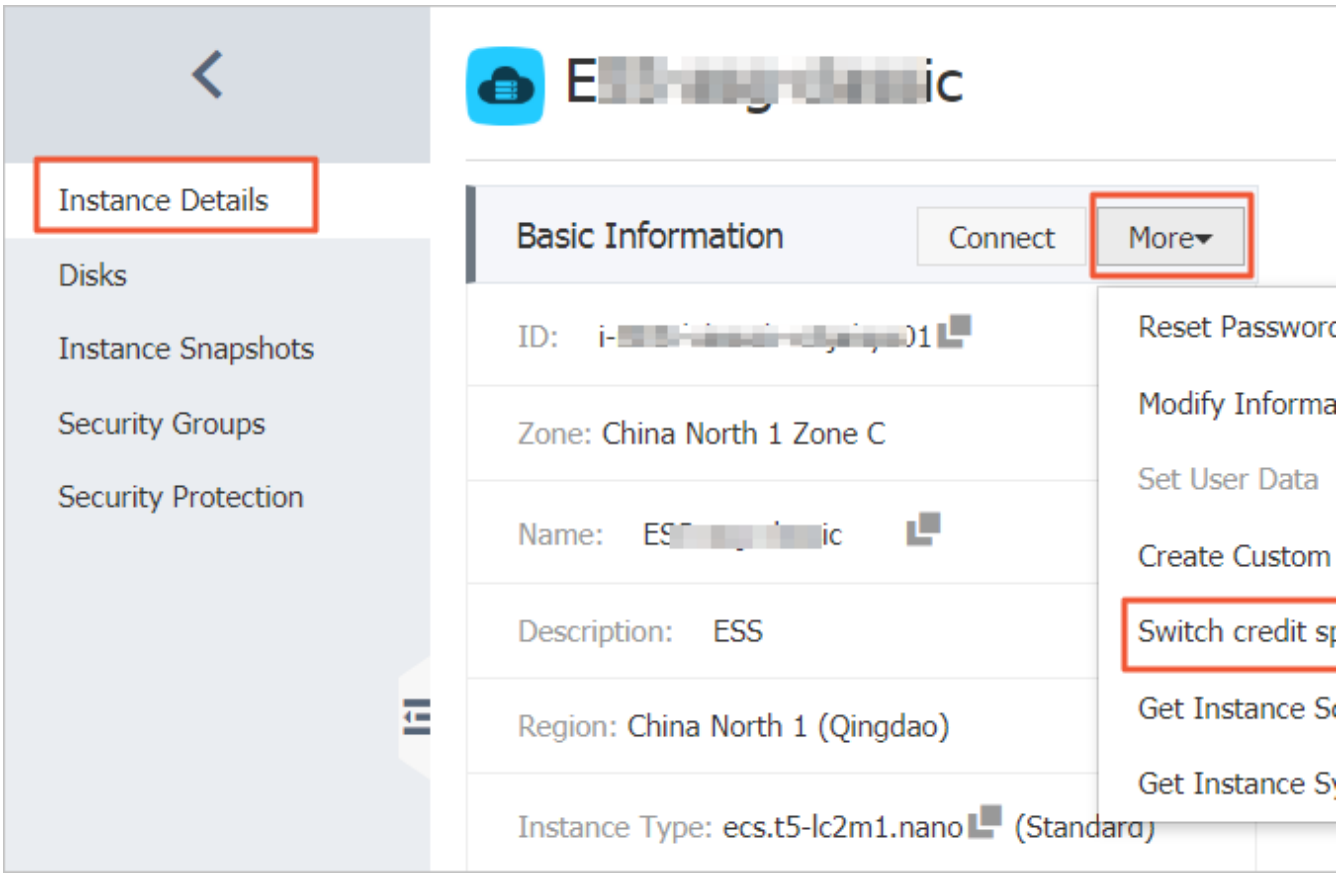
Change the running mode of a t5 instance

Within the lifecycle of a t5 instance, you can change its running mode between **Standard** and U**nlimited** in real time through the console or the API *ModifyInstanceAttribute*.

Note:
The t5 instance must be in a **Running** state (`Running`) before you can change its mode.

Follow the steps below to change the running mode in the console:

1. Log on to the *ECS console*.
2. In the left-side navigation pane, click **Instances**.
3. Select a region.
4. In the list of instances, find the target instance, and click the instance ID.
5. In the **Basic Information** part of the **Instance details** page, click **More**, and then select **Switch credit specification mode** to change to the "Unlimited" mode. You can select **Switch credit specification mode** again to change back to the "Standard" mode.



The following table describes how the running mode of a t5 instance changes according to the operations or its payment status:

Operation/status	Results
Stop the instance	The default mode is "Standard" after it is started again.
Reboot the instance	The running mode remains unchanged after the reboot.
Payment failed	The running mode changes to "Standard". The original mode is restored after the payment is successful.

View CPU utilization and CPU credits

In the ECS console, you can view the CPU information about an instance such as CPU utilization, consumed CPU credits, accrued CPU credits, excess CPU credits, and advance CPU credits.

1. Log on to the *ECS console*.
2. In the left-side navigation pane, click **Instances**.
3. Select a region.
4. In the instance list, find the target instance. Then click the instance ID or click **Manage** in the **Actions** column.

5. View the metrics in the **Monitoring Information** area of the **Instance Details** page.

You can also view the CPU usage after connecting to the instance remotely:

- Windows: *Connect to the instance*, and then view the CPU usage in the **Task Manager**.
- Linux: *Connect to the instance*, and then run the `top` command to view the CPU usage.

Change the instance type

In the ECS console, if you find that the CPU usage remains at the baseline performance level for a long period of time or rarely exceeds the baseline level, your instance type may not meet your needs or may exceed your needs. In either case, you may consider changing the instance type.

You can change the instance type based on the billing method:

- Subscription instances: You can change the instance type by *upgrading or downgrading instance configurations*.
- Pay-As-You-Go instances: You can change the configurations by *changing the instance type*.

For the target instance type families, see *instance type families that support upgrading instance types*.

6 Block storage

6.1 What is block storage?

Overview

Block storage is a high-performance, low latency block storage service for Alibaba Cloud ECS. Similar to a hard disk, you can format block storage and create a file system on it to easily meet the data storage needs of your business.

Alibaba Cloud provides a variety of block-level storage products based on a distributed storage architecture and local disks located on the physical servers where ECS instances are hosted. Specifically, the storage products are as follows:

- *Cloud Disk*, which is a block-level data storage product provided by Alibaba Cloud for ECS, uses a *multiple distributed system*, and features low latency, high performance, persistence, high reliability, and more. Cloud disks can be created, resized, and released at any time.
- Shared block storage is a block-level data storage device that supports simultaneous read and write access to multiple ECS instances. Similar to the cloud disk, shared block storage uses a *multiple distributed system*. It supports simultaneous access to multiple instances, and features low latency, high performance, and high reliability. Shared Block Storage applies to shared access scenarios for block storage devices under a shared everything architecture.
- Local disks are the disks attached to the physical servers (host machines) on which ECS instances are hosted. They are designed for business scenarios requiring high storage I/O performance and massive storage cost performance. Local disks provide local storage and access for instances, and features low latency, high random IOPS, high throughput, and cost-effective performance.

For more information about the performance of block-level storage products, see *Storage parameters and performance test*.

Block storage, OSS and NAS

Currently, Alibaba Cloud provides three types of data storage products: block storage, *Object Storage Service (OSS)*, and *Network Attached Storage (NAS)*.

the following three types of data storage products:

- Block storage: A high-performance and low-latency block-level storage device for ECS. It supports random reads and writes. You can format block storage and create a file system on

it as you would with a hard disk. , thereby enabling block storage to meet the data needs of numerous business scenarios.

- OSS: A huge storage space designed for storing massive amounts of unstructured data on the Internet, including images, audio, and video. You can access the data stored in OSS anytime , anywhere, by using APIs. Generally, OSS is applicable to business scenarios as website construction, separation of dynamic and static resources, and CDN acceleration.
- NAS: A storage space designed to store massive amounts of unstructured data that can be accessed by using standard file access protocols , such as the Network File System (NFS) protocol for Linux, and the Common Internet File System (CIFS) protocol for Windows. You can set permissions to allow different clients to access the same file at the same time. NAS is suitable for business scenarios such as file sharing across departments, non-linear file editing, high-performance computing, and containerization (such as with Docker).

6.2 Storage parameters and performance test

This document describes the performance index of block storage, performance testing methods, and how to interpret the testing results.

Performance index of block storage

The main index for measuring storage performance include IOPS, throughput, and latency.

- **IOPS**

  IOPS stands for Input/Output Operations per Second, which means the number of write or read operations that can be performed each second. Transaction-intensive applications, such as database applications, are sensitive to IOPS.

  The following table lists common performance characteristics that are measured.

IOPS performance characteristics	Description	
Total IOPS	The total number of I/O operations per second	
Random read IOPS	The average number of random read I/O operations per second	Random access to locations on storage devices
Random write IOPS	The average number of random write I/O operations per second	

IOPS performance characteristics	Description	
Sequential read IOPS	The average number of sequential read I/O operations per second	Sequential access to locations on storage devices
Sequential write IOPS	The average number of sequential write I/O operations per second	

- **Throughput**

  Throughput measures the data size successfully transferred per second.

  Applications that require mass read or write operations (such as Hadoop offline computing applications) are sensitive to throughput.

- **Latency**

  Latency is the period that is needed to complete an I/O request.

  For latency-sensitive applications (such as databases) in which high latency may lead to performance reduction or error reports in applications, we recommend that you use SSD disks, SSD Shared Block Storage, or local SSD disks.

  For throughput-sensitive applications (such as Hadoop offline computing) that are less sensitive to latency, we recommend that you use ECS instances with local HDD disks, such as instances of the d1 or d1ne instance type family.

Performance

This section describes the performance of various block storage products.

Block Storage capacity is measured in binary units, such as kibibyte (KiB), mebibyte (MiB), gibibyte (GiB), or Tebibyte (TiB).

Note:
1 KiB is *1,024 bytes*. 1MiB is 1,024 KiB. 1GiB is 1,024 MiB. 1TiB is 1,024 GiB.

- **Cloud disks**

  The following table lists the features and typical scenarios of different types of cloud disks.

Parameter	ESSD Cloud Disk	SSD Cloud Disk	Ultra Cloud Disk	Basic Cloud Disk
Capacity of a single disk	32,768 GiB	32,768 GiB	32,768 GiB	2,000 GiB
Max. IOPS	1,000,000	25,000*	5,000	Several hundreds
Max. throughput	4,000 MBps	300 MBps*	140 MBps	30−40 MBps
Formulas to calculate performance of a single disk**	IOPS = min{1800 + 50 x capacity, 1,000,000}	IOPS = min{1800 + 30 x capacity, 25,000}	IOPS = min{1800 + 8 x capacity, 5 ,000}	N/A
	Throughput = min {120 + 0.5 x capacity, 4,000} MBps	Throughput = min{120 + 0.5 x capacity, 300} MBps	Throughput = min{100 + 0.15 x capacity, 140} MBps	N/A
Data reliability	99.9999999%	99.9999999%	99.9999999%	99.9999999%
API name	cloud_essd	cloud_ssd	cloud_efficiency	cloud
Scenarios	• OLTP databases : relational databases such as MySQL, PostgreSQL , Oracle, and SQL Server • NoSQL databases: non-relational databases such as MongoDB, HBase, and Cassandra • ElasticSea rch distribute d logs: Elasticsearch , Logstash and Kibana	• Large and medium-sized relational databases , such as MySQL, SQL Server, PostgreSQL, and Oracle • Large or medium-sized developmen t or testing applications that require high data reliability	• Small or medium-sized relational databases , such as MySQL, SQL Server, and PostgreSQL • Large or medium-sized developmen t or testing applications that require high data reliability and medium performance	• Applicatio ns with infrequent access or low I/O load. If higher I/O performance is needed, we recommend that you use SSD disks. • Applications that require low costs and random read and write I/O operations

Parameter	ESSD Cloud Disk	SSD Cloud Disk	Ultra Cloud Disk	Basic Cloud Disk
	(ELK) log analysis			

* The performance of an SSD Cloud Disk varies with the data block size. Smaller data blocks result in lower throughput and higher IOPS, as shown in the following table. An SSD Cloud Disk can achieve the expected performance only when it is attached to an I/O-optimized instance. In other words, an SSD Cloud Disk cannot achieve the expected performance if it is not attached to an I/O-optimized instance.

Data block size	Maximum IOPS	Throughput
4 KiB	About 25,000	Far smaller than 300 MBps
16 KiB	About 17,200	Close to 300 MBps
32 KiB	About 9,600	
64 KiB	About 4,800	

** An SSD Cloud Disk is taken as an example to describe the performance of a single disk:

- The maximum IOPS: The baseline is 1,800 IOPS. It increases by 30 IOPS per GiB of storage. The maximum IOPS is 25,000.
- The maximum throughput: The baseline is 120 MBps. It increases by 0.5 MBps per GiB of storage. The maximum throughput is 300 MBps.

The random write latency varies with the disk categories as follows:

- ESSD disks: 0.1−0.2 ms
- SSD disks: 0.5−2 ms
- Ultra Cloud Disks: 1−3 ms
- Basic Cloud Disks: 5−10 ms

- **Shared Block Storage**

  The following table lists the features and typical scenarios of different types of Shared Block Storage.

Parameter	SSD Shared Bock Storage	Ultra Shared Block Storage
Capacity	• Singe disk: 32,768 GiB • Single instance: 128 TiB	• Singe disk: 32,768 GiB • Single instance: 128 TiB

Parameter	SSD Shared Bock Storage	Ultra Shared Block Storage
Maximum random read/write IOPS*	30,000	5,000
Maximum sequential read/ write throughput*	512 MBps	160 MBps
Formulas to calculate performance of a single disk**	IOPS = min{1600 + 40 x capacity, 30,000}	IOPS = min{1000 + 6 x capacity, 5,000}
	Throughput = min{100 + 0.5 x capacity, 512} MBps	Throughput = min{50 + 0.15 x capacity, 160} MBps
Scenarios	• Oracle RAC • SQL Server • Failover cluster • High-availability architecture of servers	• High-availability architecture of servers • High-availability architecture of development and testing databases

* The maximum IOPS and throughput listed in the preceding table are the maximum performance of a bare shared block storage device that is attached to two or more instances at the same time during stress tests.

** An SSD Shared Block Storage is used as an example to describe the performance of a single disk:

- The maximum IOPS: The baseline is 1,600 IOPS. It increases by 40 IOPS per GiB of storage. The maximum IOPS is 30,000.
- The maximum throughput: The baseline is 100 MBps. It increases by 0.5 MBps per GiB of storage. The maximum throughput is 512 MBps.

The latency varies with the shared block storage categories as follows:

- SSD Shared Block Storage: 0.5−2 ms
- Ultra Shared Block Storage: 1−3 ms

- **Local disks**

  For the performance of local disks, see *Local disks*.

Test disk performance

Depending on the OS that an instance is running, the following tools are recommended to test disk performance:

- For Linux, DD, fio, or sysbench is recommended.
- For Windows, fio or Iometer is recommended.

Note:
The disk benchmark tested by different tools varies with different operating systems. The performance parameters in this article are the results tested by fio with a Linux instance, and are used as the index reference of block storage product performance.

This section describes how to test disk performance, taking the fio tool used with a Linux instance as an example. Before you test the disk, verify that the disk is 4 KiB aligned.

Warning:
You can test bare disks to obtain more accurate performance data, but the structure of the file system will be damaged. Make sure that you back up your data before testing. We recommend that you use a new ECS instance without data to test the disks to avoid data loss.

- Test random write IOPS:

```
fio -direct=1 -iodepth=128 -rw=randwrite -ioengine=libaio -bs=4k -
size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=iotest -
name=Rand_Write_Testing
```

- Test random read IOPS:

```
fio -direct=1 -iodepth=128 -rw=randread -ioengine=libaio -bs=4k -
size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=iotest -
name=Rand_Read_Testing
```

- Test write throughput:

```
fio -direct=1 -iodepth=64 -rw=write -ioengine=libaio -bs=1024k -size
=1G -numjobs=1 -runtime=1000 -group_reporting -filename=iotest -name
=Write_PPS_Testing
```

- Test read throughput:

```
fio -direct=1 -iodepth=64 -rw=read -ioengine=libaio -bs=1024k -size=
1G -numjobs=1 -runtime=1000 -group_reporting -filename=iotest -name=
Read_PPS_Testing
```

The command for testing random read IOPS is used as an example to describe the meaning of the parameters of a fio command, as shown in the following table.

Parameter	Meaning
-direct=1	Ignore I/O buffer when testing. Data is written directly.

Parameter	Meaning
-iodepth=128	Indicates that when you use AIO, the maximum number of I/O issues at the same time is 128.
-rw=randwrite	Indicates that the read and write policy is random write. Other options include: • randread (random read) • read (sequential read) • write (sequential write) • randrw (random read and write)
-ioengine=libaio	Use libaio as the testing method (Linux AIO, Asynchronous I/O). Usually there are two ways for an application to use I/O: • Synchronous Synchronous I/O only sends out one I/O request at a time, and returns only after the kernel is completed. In this case, the iodepth is always less than 1 for a single job, but can be resolved by multiple concurrent jobs. Usually 16−32 concurrent jobs can fill up the iodepth. • Asynchronous The asynchronous method uses libaio to submit a batch of I/O requests each time, thus reducing interaction times and making interactions more effective.
-bs=4k	Indicates the size of each block for one I/O is 4 KiB. If not specified, the default value 4 KiB is used. When IOPS is tested, we recommend that you set `bs` to a small value, for example, such as 4k in this example command. When throughput is tested, we recommend that you set `bs` to a large value, such as 1024k in this example command.
-size=1G	Indicates the size of the testing file is 1 GiB.

Parameter	Meaning
-numjobs=1	The number of testing jobs is 1.
-runtime=1000	Testing time is 1,000 seconds. If not specified, the test will write data of the file whose size is specified by `-size` block by block, with the data block size specified by `-bs`.
-group_reporting	The display mode for showing the testing results. Group_reporting means the statistics of each job are summed up, instead of all statistics of each job being shown.
-filename=iotest	The output path and name of the test files, for example, iotest. You can test bare disks to obtain more accurate performance data, but the test causes damage to the structure of the file system. Make sure that you back up your data before testing.
-name=Rand_Write_Testing	The name of the testing task.

6.3 Cloud disks and Shared Block Storage

Cloud disks and Shared Block Storage are block-level data storage products provided by Alibaba Cloud for ECS that features low latency, high performance, persistence, and high reliability. They use a *triplicate distributed system* to provide 99.9999999% data reliability for ECS instances. Cloud disks and Shared Block Storage can automatically copy your data within the target zone to help you prevent unexpected hardware faults from causing data unavailability or service disruption. Just like what you do with a hard disk, you can partition and format the cloud disks and Shared Block Storage attached to an ECS instance, create a file system, and store data on them.

You can expand the cloud disks and Shared Block Storage as needed at any time. For more information, see *Linux _ Resize a data disk* and *increase system disk size*. You can also create snapshots to back up data for the cloud disks and Shared Block Storage. For more information about snapshots, see *what are ECS snapshots*.

Cloud disks and Shared Block Storage differ in whether they can be simultaneously attached to multiple ECS instances and perform read and write operations. Details are as follows:

- Cloud disks can be attached to only one ECS instance in the same zone of the same region.
- Shared Block Storage devices can be mounted to a maximum of eight ECS instances in the same zone of the same region.

Note:
Shared Block Storage is currently in public beta phase. You can *open a ticket* to submit your application for beta testing.

Cloud disks

- **Performance-based category**
  - ESSD: An ultra-high-performance cloud product based on the next generation distributed block storage architecture. ESSD combines 25 GE networks with RDMA technology, offering the capability of up to 1 million random read/write operations and a shorter single-link latency. ESSD is currently in public beta phase. For more information, see *FAQ about ESSD cloud disks*.
  - SSD cloud disks: high-performance disks with stable and high random I/O performance and high data reliability
  - Ultra cloud disks: with high cost performance, medium random I/O performance, and high data reliability
  - Basic cloud disks: with high data reliability and general random I/O performance
- **Function-based category**
  - System disks: have the same life cycle as the ECS instance to which it is mounted. A system disk is created and released at the same time as the instance. Shared access is not allowed. The available size range of a single system disk varies according to the image, as follows:
    - Linux (excluding CoreOS) and FreeBSD: 20−500 GiB
    - CoreOS: 30−500 GiB
    - Windows: 40−500 GiB
  - Data disks: can be *created separately* or at the same time as ECS instances. A data disk created with an ECS instance has the same life cycle as the instance, and is created and released along with the instance. Data disks created separately can be *released independently* or at the same time as the corresponding ECS instances. Shared access is not allowed. The performance of data disks depends on the cloud disk type. For more information, see *storage parameters and performance test*.

When used as data disks, up to 16 cloud disks can be attached to one ECS instance.

Shared Block Storage

Shared Block Storage is a block-level data storage service with strong concurrency, high performance, and high reliability. It supports concurrent reads from and writes to multiple ECS instances, and provides data reliability of up to 99.9999999%. Shared Block Storage can be mounted to a maximum of 8 ECS instances.

Shared Block Storage can only be used as data disks and can only be created separately. Shared access is allowed. You can set the Shared Block Storage device to be released when the ECS instances are released.

Shared Block Storage can be divided into:

- SSD Shared Block Storage, which uses SSD as the storage medium to provide stable and high-performance storage with enhanced random I/O and data reliability.
- Ultra Shared Block Storage, which uses the hybrid media of SSD and HDD as the storage media.

When used as data disks, Shared Block Storage allows up to 16 data disks to be attached to each ECS instance.

For more information, see *FAQ about Shared Block Storage*.

Billing

Shared Block Storage is currently in public beta phase free of charge.

The billing method of a cloud disk depends on how it is created:

- Cloud disks created with Subscription instances are billed before the service is ready for use. For more information, see *Subscription*.
- Cloud disks created at the same time as Pay-As-You-Go instances, or created separately, are billed on a Pay-As-You-Go basis. For more information, see *Pay-As-You-Go*.

You can change the billing method of the cloud disk, as shown in the following table.

Conversion of billing methods	Feature	Effective time	Suitable for
Subscription -> Pay-As-You-Go	*Renew for configuration downgrade*	Effective from the next billing cycle	Subscription cloud disks mounted to Subscription instances. The billing method of

Conversion of billing methods	Feature	Effective time	Suitable for
			the system disk cannot be changed.
Pay-As-You-Go -> Subscription	*Upgrade configurations*	Effective immediately	Pay-As-You-Go data disks mounted to Subscription instances. The billing method of the system disk cannot be changed.
	Switch from Pay-As-You-Go to Subscription billing		System disks and data disks mounted to Pay-As-You-Go instances.

Related operations

You can perform the following operations on cloud disks:

- If *a cloud disk or Shared Block Storage device is created separately from a data disk*, you must *attach a cloud disk* in the ECS console, and then connect to the ECS instance to *partition and format the data disk*.
- If you want to encrypt the data on a cloud disk, *encrypt the disk*.
- If your system disk capacity is insufficient, you can *increase system disk size*.
- If you want to expand the data disk capacity, you can *resize the data disk*.
- If you want to change the OS, you can *change the system disk*.
- If you want to back up the data of a cloud disk or Shared Block Storage device, you can *manually create snapshots for the cloud disk or Shared Block Storage* or *apply an automatic snapshot policy to it* to automatically create snapshots on schedule.
- If you want to use the OS and data environment information of one instance on another instance, you can *create a customized image using the system disk snapshots of the latter instance*.
- If you want to restore a cloud disk or Shared Block Storage device to the status when the snapshot is created, you can *roll back a cloud disk* using its snapshot.
- If you want to restore a cloud disk to its status at the time of creation, you can *reinitialize a cloud disk*.
- If you do not need a cloud disk or Shared Block Storage device, you can *detach a cloud disk* and *release a cloud disk*.

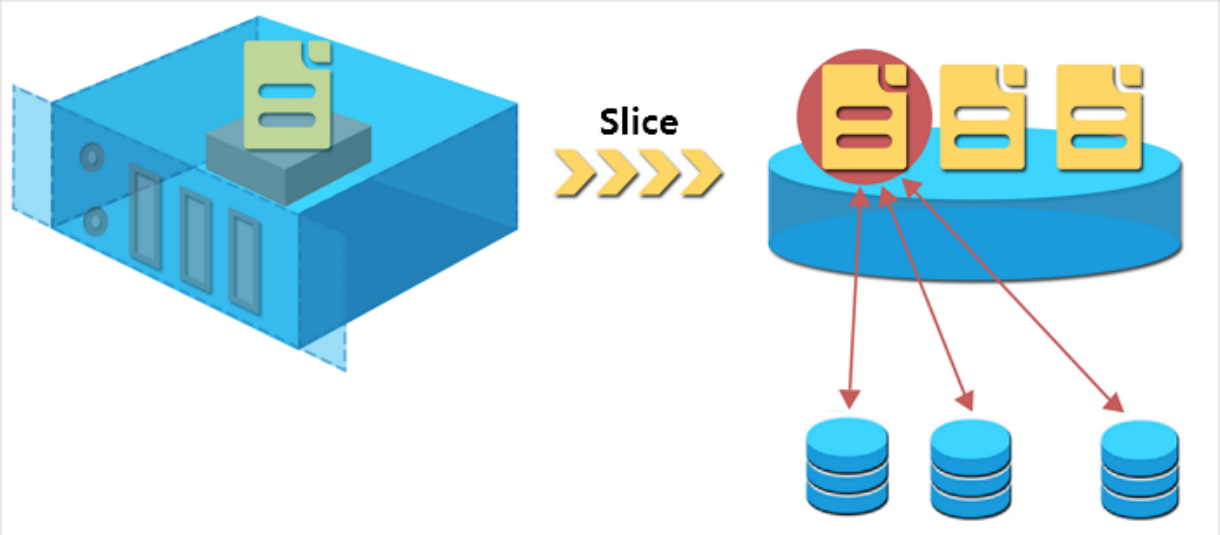
- If you no longer need a Subscription billed cloud disk, you can *convert the billing methods of cloud disks*, and then *detach a cloud disk* and *release a cloud disk*.

For more information about operations on cloud disks, see *cloud disks* in *Cite LeftUser GuideCite Right*.

6.4 Triplicate technology

The Alibaba Cloud Distributed File System provides stable and efficient data access and reliability for ECS. Triplicate technology, that is, the process of making and distributing three copies of data, is the principle concept implemented in the Alibaba Cloud Distributed File System.

When you perform read and write operations on cloud disks, the operations are translated into the corresponding processes on the files stored in Alibaba Cloud data storage system. The Distributed File System of Alibaba Cloud uses a flat design in which a linear address space is divided into slices, also called chunks. Each chunk has three copies stored on different server nodes on different racks. This guarantees data reliability.



How triplicate technology works

Triplicate technology involves three key components: Master, Chunk Server, and Client. To demonstrate how triplicate technology works, in this example, the write operation of an ECS user undergoes several conversions before being executed by the Client. The process is as follows:

1. The Client determines the location of a chunk corresponding to a write operation.
2. The Client sends a request to the Master to query the storage locations (that is, the Chunk Servers) of the three copies of the chunk.

3. The Client sends write requests to the corresponding three Chunk Servers according to the results returned from the Master.
4. The Client returns a message that indicates whether the operation was successful.

This strategy guarantees that all the copies of a chunk are distributed on different Chunk Servers on different racks, effectively reducing the potential of total data loss caused by failure of a Chunk Server or a rack.

Data protection

If a system failure occurs because of a corrupted node or hard drive failure, some chunks may lose one or more of the three valid chunk copies associated with them. If this occurs and triplicate technology is enabled, the Master replicates data between Chunk Servers to replace the missing chunk copies across different nodes.

To summarize, all your operations (additions, modifications, or deletions) on cloud disk data are synchronized to the three chunk copies at the bottom layer. This mode ensures the reliability and consistency of your data.

Furthermore, we recommend you implement appropriate backup strategies, *snapshots*, and other precautionary actions to restore and protect your data and guarantee its availability against other types of failures, such as viruses, human error, or malicious activity on your account. No single technology can solve all the problems, so you must choose appropriate data protection measures to establish a solid defense line for your valuable business data.

6.5 ECS disk encryption

ECS disks in this article refer to **cloud disks** and **Shared Block Storage devices**. They are referred to as **ECS disks** in the following contents, unless otherwise specified.

What is ECS disk encryption?

The ECS disk encryption feature allows you to encrypt new ECS disks so that you can meet encryption needs for scenarios such as certification requirements and business security. The ECS disk encryption feature means you do not have to create, maintain, or protect your own key management infrastructure, nor change any of your existing applications or maintenance processes. In addition, no extra encryption or decryption operations are required, making ECS disk encryption operations invisible to your applications or other operations.

Encryption and decryption processes hardly degrade ECS disk performance. For information on the performance testing method, see *storage parameters and performance test*.

After an encrypted ECS disk is created and attached to an ECS instance, you can encrypt data that is:

- Stored directly on the ECS disk.
- Transmitted between the ECS disk and the instance. However, data in the instance operating system is not encrypted.
- Created from the encrypted ECS disk, such as snapshots. These snapshots are called encrypted snapshots.

Encryption and decryption are performed on the host that runs the ECS instance, so the data transmitted from the ECS instance to the cloud disk is encrypted.

ECS disk encryption supports all available cloud disks (Basic Cloud Disks, Ultra Cloud Disks, SSD Cloud Disks, and ESSDs) and shared block storage (Ultra Shared Block Storage and SSD Shared Block Storage).

ECS disk encryption supports all available instance types and is supported in all regions.

ECS disk encryption dependencies

ECS disk encryption is dependent on the Key Management Service (KMS), which must be in the same region. However, you do not need to perform any additional operations in the KMS console to activate ECS disk encryption.

The first time you use the ECS disk encryption function (such as when you are creating ECS instances or ECS disks), you must first authorize and activate KMS. Otherwise, you cannot create encrypted ECS disks or instances with encrypted disks.

If you use an API or the CLI to use the ECS disk encryption function, such as *CreateInstance* or *CreateDisk*, you must first activate KMS on the Alibaba Cloud console.

The first time you encrypt a disk in a target region, Alibaba Cloud automatically creates a Customer Master Key (CMK) in the KMS region, exclusively for ECS. The CMK cannot be deleted . You can query the CMK in the KMS console.

Key management for ECS disk encryption

ECS disk encryption handles key management for you. Each new ECS disk is encrypted by using a unique 256-bit key (derived from the CMK). This key is also associated with all snapshots created from this ECS disk and any ECS disks subsequently created from these snapshots. These keys are protected by the key management infrastructure of Alibaba Cloud provided by KMS. This approach implements strong logical and physical security controls to prevent unauthorized access. Your data and the associated keys are encrypted based on the industry standard AES-256 algorithm.

You cannot change the CMK associated with encrypted ECS disks and snapshots.

The key management infrastructure of Alibaba Cloud conforms to the recommendations in (NIST) 800-57 and uses cryptographic algorithms that comply with the (FIPS) 140-2 standard.

Each Alibaba Cloud account has a unique CMK in each region. This key is separate from the data and is stored in a system protected by strict physical and logical security controls. Each encrypted disk and its snapshots use an encryption key that is unique to the specific disk. The encryption key is created from and encrypted by the CMK for the current user in the current region. The disk encryption key is only used in the memory of the host that runs your ECS instance. The key is never stored in plaintext in any permanent storage media (such as an ECS disk).

Fees

The ECS disk encryption features incur no additional fees.

The CMK that ECS creates for you in each region is a service key. It does not consume your master key quota in a given region, meaning no additional fees are incurred.

Note:

No additional fees are charged for any **read/write** operations on a disk, such as mounting/umounting, partitioning, and formatting. However, if you perform operations on a disk in the ECS console or by using APIs, KMS APIs are called and such calls consume the KMS API quota in the current region.

These operations include:

- Creating encrypted disks by calling *CreateInstance* or *CreateDisk*.
- Attaching an encrypted disk to an instance by calling *AttachDisk*.
- Detaching an encrypted disk from an instance by calling *DetachDisk*.
- Creating a snapshot by calling *CreateSnapshot*.
- Restoring a disk by calling *ResetDisk*.
- Re-initializing a disk by calling *ReInitDisk*.

Create an encrypted ECS disk

Currently, only cloud disks can be encrypted. You can create an encrypted cloud disk in the following ways:

- Create a cloud disk as a data disk when creating an ECS instance or :
  - Check Encrypted to create a encrypted blank cloud disk.
  - Select an encrypted screenshot to create a cloud disk.
- When using APIs or the CLI:
  - Set the parameter `DataDisk.n.Encrypted` (*CreateInstance*) or `Encrypted` (*CreateDisk*) to `true`.
  - Specify the `SnapshotId` parameter of the encrypted snapshot in CreateInstance or CreateDisk.

Convert unencrypted data to encrypted data

You cannot directly convert an **unencrypted disk** to an **encrypted disk**, or perform the converse operation.

You cannot convert a snapshot created from an **unencrypted disk** to an **encrypted snapshot**, or perform the converse operation.

Therefore, if you must switch the existing data from status **unencrypted** to **encrypted**, we recommend that you use the `rsync` command in a Linux instance or the `robocopy` command in a Windows instance to copy data from an **unencrypted** disk to a (new) **encrypted** disk.

Therefore, if you must switch the existing data from status **encrypted** to **unencrypted**, we recommend that you use the `rsync` command in a Linux instance or the `robocopy` command in a Windows instance to copy data from an **encrypted disk** to a (new) **unencrypted disk** .

Limits

ECS disk encryption has the following limits:

- You can only encrypt ECS disks, not local disks or ephemeral disks.
- You can only encrypt data disks, not system disks.
- You cannot directly convert existing unencrypted disks into encrypted disks.
- You cannot convert encrypted disks into unencrypted disks.
- You cannot convert unencrypted snapshots to encrypted snapshots.
- You cannot convert encrypted snapshots to unencrypted snapshots.
- You cannot share images created from encrypted snapshots.
- You cannot copy images created from encrypted snapshots across regions.
- You cannot export images created from encrypted snapshots.
- You cannot define CMKs for each region. They are generated by the system.
- The ECS system creates CMKs for each region. You cannot delete these keys, and you do not incur fees from them.
- After a cloud disk is encrypted, you cannot change the CMK used for encryption and decryption .

6.6 Local disks

Local disks are the disks attached to the physical servers (host machines) on which ECS instances are hosted. They are designed for business scenarios requiring high storage I/O performance. Local disks provide local storage and access for instances, and feature low latency, high random IOPS, high throughput, and cost-effective performance.

Because a local disk is attached to a single physical server, the data reliability depends on the reliability of the physical server, which may create single points of failure in your architecture. We recommend that you implement data redundancy at the application layer to guarantee data availability.

Warning:

Using a local disk for data storage carries the risk of data loss (for example, if the host machine is down). Therefore, we recommend you never use a local disk to store any data that requires long-

term persistence. If no data reliability architecture is available for your application, we strongly recommend that you build your ECS with *cloud disks or Shared Block Storage devices*.

This document details information about local disks and instances that support local disks. If you are using a previous generation local SSD disk, see *local SSD disks in the previous generation disks*.

Disk types

Currently, Alibaba Cloud provides two types of local disks:

- Local NVMe SSD: This disk is used together with instances of the following type families: i2, i1 , and gn5. The instance type families i1 and i2 apply to the following scenarios:
  - Online games, e-businesses, livestreaming, and other industries that provide online businesses and have low latency and high I/O performance requirements on block level storage for I/O-intensive applications.
  - Business scenarios that have high requirements on the storage I/O performance and availability of the application layer, such as NoSQL non-relational databases, MPP data warehouses, and distributed file systems.
- Local SATA HDD: This disk is used together with instances of the d1ne and d1 type families. It is applicable to businesses that require big data computing and storage analysis for massive data storage and offline computing business scenarios. It fully meets the needs of distributed computing business models (such as those built on the Hadoop framework) across instance storage performance, capacity, and intranet bandwidth.

Performance of local NVMe SSD

The following table lists the performance of local NVMe SSD of an i1 ECS instance.

Parameters	Local NVMe SSD
Maximum capacity	Single disk: 1,456 GiB Total: 2,912 GiB
Maximum IOPS	Single disk: 240,000 Total: 480,000
Maximum throughput	Read throughput per disk: 2 GBps Total read throughput: 4 GBps Write throughput per disk: 1.2 GBps Total write throughput: 2.4 GBps
Single-disk performance *	Write performance:

Parameters	Local NVMe SSD
	• Single-disk IOPS: IOPS = min{165 * capacity, 240,000} • Single disk throughput: Throughput = min{0.85 * capacity, 1,200} MBps Read performance: • Single disk IOPS: IOPS = min{165 * capacity, 240,000} • Single disk throughput: Throughput = min{1.4 * capacity, 2,000} MBps
Access latency	Microsecond-level

* Single disk performance calculations are as follows:

- Write IOPS for a single local NVMe SSD: 165 IOPS for each GiB, up to 240,000 IOPS.
- Write throughput for a single local NVMe SSD: 0.85 MBps for each GiB, up to 1,200 Mbit/s.

Performance of local SATA HDD

The following table lists the performance of local SATA HDD of a d1ne or d1 ECS instance.

Parameters	Local SATA HDD
Maximum capacity	Single disk: 5,500 GiB Total capacity per instance: 154,000 GiB
Maximum throughput	Single disk: 190 MBps Total throughput per instance: 5,320 MBps
Access latency	Millisecond-level

Billing

Local disks are charged according to the instances to which they are attached. For more information about instance billing methods, see *Subscription* and *Pay-As-You-Go*.

Lifecycle

A local disk has the same lifecycle as the instance that it is attached to. This means that:

- You can create a local disk only when creating an instance that has local storage. The capacity of a local disk is determined by the ECS instance type. You cannot increase or decrease it.
- When the instance is released, the local disk is released with it.

Instance operations

The following table details how operations on an instance that has local storage affect the state of the data on the local disk.

Operation	State of the data on the local disk	Result
Restart within the operating system/restart or force restart in the ECS console	Retained	Both the storage volumes and data on the local disk are retained.
Shutdown within the operating system/stop or force stop in the ECS console	Retained	Both the storage volumes and data on the local disk are retained.
Release in the ECS console	Erased	The storage volumes on the local disk are erased and the data on it is not retained.
Downtime migration	Erased	The storage volumes on the local disk are erased and the data on it is not retained.
Out-of-service (before the computing resources of an instance is released)	Retained	Both the storage volumes and data on the local disk are retained.
Out-of-service (after the computing resources of an instance is released)	Erased	The storage volumes on the local disk are erased and the data on it is not retained.

Related operations

If your ECS instance is attached with local disks, you must connect to the instance to *format the disk*. Unlike cloud disks, you cannot perform the following operations on local disks:

- Independently create an empty local disk or create a local disk from a snapshot.
- Attach a local disk in the ECS console.
- Detach and release a local disk.
- Increase the size of a local disk.
- Re-initialize a local disk.
- Create a snapshot for a local disk and use the snapshot to roll back the local disk.

7 Network and security

7.1 Network types

Alibaba Cloud offers Virtual Private Cloud (VPC) networks and classic networks.

Virtual Private Cloud (VPC)

VPCs are isolated networks established in Alibaba Cloud and logically isolated from each other . You can customize the topology and IP addresses in a VPC. We recommend VPC if you are skilled in network management and have higher network security requirements.

For more information about VPC, see *Virtual Private Cloud* documentation.

Classic network

A classic network is majorly deployed in the public infrastructure of Alibaba Cloud, which is responsible for its planning and management. We recommend classic networks if your business requirements are high in terms of network usability.

Note:
If you did not purchase an ECS instance before 17:00 (UTC+8) on June 14, 2017, you cannot choose the Classic network option.

VPC vs. Classic networks

The following table lists all the functional differences between the VPCs and classic networks.

Items	VPC	Classic network
Two-layer logic isolation	Supported	Not supported
Custom private network blocks	Supported	Not supported
Private IP addresses	Unique within one VPC. Replicable between VPCs.	Unique in the global Classic network
Communicate within or between private networks	Able to communicate within a VPC, but isolated between VPCs	Able to communicate in one region and under one account
Tunneling	Supported	Not supported
Custom router	Supported	Not supported
Routing table	Supported	Not supported
Switches	Supported	Not supported

Items	VPC	Classic network
SDN	Supported	Not supported
Self-built NAT gateway	Supported	Not supported
Self-built VPN	Supported	Not supported

7.2 Intranet

Currently, Alibaba Cloud servers communicate over an intranet. They use a gigabit of shared bandwidth for non I/O optimized instances, and 10 gigabits of shared bandwidth for I/O optimized instances, with no special restrictions. However, because this is a shared network, the bandwidth may fluctuate.

If you have to transmit data between two ECS instances in the same region, use an intranet connection. Intranet connections can also be used to connect any combination of ECS, RDS, SLB , and OSS if they are deployed in the same region. The Internet speed of these products is based on a gigabit of shared bandwidth.

The network types, owners, regions, and security groups affect the intranet communication of ECS instances. See the following table for details.

Network type	Owners	Regions	Security groups	How to enable intranet communication
VPC, same VPC	One account or different accounts	Same	Same	Enabled by default.
			Different	Authorize security groups for each other.
VPC, different VPCs	One account or different accounts	Same	Either the same or different	Use Express Connect. For more information, see *Application scenarios from Product Introduction to Express Connect*.
		Different	Different	
Classic	One account	Same	Same	Enabled by default.

Network type	Owners	Regions	Security groups	How to enable intranet communication
	Different accounts		Either the same or different	Authorize security groups for each other. For more information, see *Scenarios of security groups*.

Private IP addresses are used for intranet communication. You cannot *change the private IP address* of an instance of the Classic network type, but you can change the private IP address of a VPC-Connected ECS instance. Private and public addresses of ECS instances do not support virtual IP (VIP) configuration.

By default, instances of different network types cannot communicate with one another in one intranet. VPC provides the *ClassicLink* function, which allows you to link an ECS instance in the classic network to cloud resources in a VPC through the intranet.

7.3 IP addresses of a classic network-connected ECS instance

IP addresses are mainly used for remote access to your instance or to the services deployed on your instance. Currently, for ECS instances of the classic network type, IP addresses are distributed in a unified way and divided into public and private IP addresses.

Intranet IP addresses

Each classic network-connected ECS instance is assigned an intranet IP address.

Scenarios

Intranet IP addresses can be used in the following scenarios:

- Load balancing
- Mutual intranet access between ECS instances
- Mutual intranet access between ECS instances and other cloud services, such as OSS and RDS

Communication traffic through intranet IP addresses within an intranet is free of charge. For more information, see *Intranet*.

Modify an intranet IP address

Once a classic network-connected ECS instance is created, you cannot change its intranet IP address.

> **Note:**
> Do not change an intranet IP address within a guest operating system. Otherwise, communication within an intranet is interrupted.

Public IP addresses

If you purchase bandwidth for Internet access, a public IP address is assigned to your classic network-connected ECS instance. You cannot change the public IP address once it is assigned.

Scenarios

A public IP address is used in the following scenarios:

- Mutual access between an ECS instance and the Internet
- Mutual Internet access between ECS instances and other Alibaba Cloud services

Assign a public IP address

When you create an ECS instance, a public IP address is assigned to it if **Assign public IP** is selected.

For a Subscription instance with no public IP address, you can use the *Upgrade Configuration* or the *Renew for Configuration Downgrade* feature to purchase public network bandwidth.

> **Note:**
> - For a Pay-As-You-Go classic network-connected ECS instance with no public IP address, you cannot assign a public IP address after the instance is created.
> - For a classic network-connected ECS instance, you cannot unbind or release its public IP address once the IP address is assigned. If you set the bandwidth to 0 Mbit/s when renewing an instance for configuration downgrade, in the next purchase cycle, the public IP address is retained, but the instance cannot access the Internet.

Billing

You are billed for usage of Internet outbound traffic only. For more information, see *Billing of network bandwidth*.

Multicast and broadcast

Intranet IP addresses cannot be used for multicast or broadcast.

7.4 IP addresses of VPC-Connected ECS instances

Each VPC-Connected ECS instance can communicate within an intranet by using a private IP address or over the Internet by using a public IP address.

Private IP addresses

Each VPC-Connected ECS instance is assigned a private IP address when it is created. That address is determined by the VPC and the CIDR block of the VSwitch to which the instance is connected.

Scenarios

A private IP address can be used in the following scenarios:

- Load balancing
- Communication among ECS instances within an intranet
- Communication between an ECS instance and other cloud products (such as OSS and RDS) within an intranet

For more information, see *Intranet*.

Modify a private IP address

To meet your business needs, you can modify the private IP address of a VPC-Connected ECS instance in the ECS console. For more information, see *Change the private IP of an ECS instance*.

Public IP addresses

VPC-Connected ECS instances support two public IP address types:

- NatPublicIp, which is assigned to a VPC-Connected ECS instance, can be released only, and cannot be unbound from the instance.
- Elastic public IP (EIP). For more information, see *What is an EIP address*.

When a VPC-Connected ECS instance accesses the Internet, its public IP address is mapped to its private IP address through network address translation (NAT).

You cannot find a network interface for Internet access by running commands within the operating system.

Scenarios

NatPublicIp and EIP are applicable to different scenarios:

- NatPublicIp: If you want to assign a public IP address to a VPC-Connected ECS instance when creating the instance and do not want to retain the public IP address when the instance is released, you can use a NatPublicIp address.
- EIP: If you want to keep a public IP address and bind it to any of your VPC-Connected ECS instances in the same region, you can use an EIP address.

Obtain a public IP address

- NatPublicIp: When creating a VPC-Connected ECS instance, if you select **Assign a public IP**, a NatPublicIp is assigned to the instance when it is created.
- EIP: You can apply for an EIP address and bind it to a VPC-Connected ECS instance. In this case, do not assign a NatPublicIp to an instance. For more information, see *Apply for an EIP address*.

Release a public IP address

- NatPublicIp: When a NatPublicIp address is assigned to an instance, you can only release the IP address, but cannot unbind it. Only a NatPublicIp address that is assigned to a Subscription instance can be released. For more information, see *Renew for configuration downgrade*.
- EIP: If you do not need an EIP address, unbind it from a VPC-Connected ECS instance and release it in the EIP console. For more information, see *Unbind and release an EIP address*.

Billing

You are billed for outbound Internet traffic usage only. For more information, see *Billing of network bandwidth*.

7.5 Multi-queue for NICs

A single CPU is not sufficient for handling network interruptions. Therefore, you should route NIC interruptions in the ECS instances to different CPUs. Results of network PPS and bandwidth tests show that a solution that uses two queues instead of one queue can enhance network performance by 50% to 100%. A solution that uses four queues can bring further significant increases in network performance.

ECS instance types supporting multi-queue

See *Instance type families* to find instance types supporting multi-queue and the number of queues that are supported.

Images supporting multi-queue

The following public images officially provided by Alibaba Cloud support multi-queue:

> **Note:**
> Whether an image supports multi-queue is not related to the memory address width of the operating system.

- CentOS 6.8/6.9/7.2/7.3/7.4
- Ubuntu 14.04/16.04
- Debian 8.9
- SUSE Linux Enterprise Server 12 SP1
- Windows 2012 R2 and Windows 2016: You may be invited to test this feature in the future.

The SUSE Linux Enterprise Server 12 SP2 edition will be available soon.

Configure multi-queue support for NICs on a Linux ECS instance

We recommend that you use one of the latest Linux distributions, such as CentOS 7.2, to configure multi-queue for the NICs.

Here we take CentOS 7.2 as an example to illustrate how to configure multi-queue for the NIC. In this example, we want to configure two queues, and the NIC name is eth0.

- To check whether the NIC supports multi-queue, run the command: `ethtool -l eth0`.
- To enable multi-queue for the NIC, run the command: `ethtool -L eth0 combined 2`.
- If you are using more than one NIC, configure each NIC.

```
    [root@localhost ~]# ethtool -l eth0
    Channel parameters for eth0:
    Pre-set maximums:
    RX: 0
    TX: 0
    Other: 0
    Combined: 2 # This line indicates that a maximum of two queues
can be configured
    Current hardware settings:
    RX: 0
    TX: 0
    Other: 0
    Combined: 1 #It indicates that one queue is currently taking
effect
```

```
    [root@localhost ~]# ethtool -L eth0 combined 2 # It sets eth0 to
use two queues currently
```

- We recommend that you enable the irqbalance service so that the system can automatically adjust the allocation of the NIC interrupts on multiple CPU cores. Run the command: `systemctl start irqbalance` (this feature is enabled by default in CentOS 7.2).
- If the network performance is not improved as expected after the multi-queue feature is enabled, you can enable the RPS feature. See the following Shell script.

```
#!/bin/bash
cpu_num=$(grep -c processor /proc/cpuinfo)
quotient=$((cpu_num/8))
if [ $quotient -gt 2 ]; then
    quotient=2
elif [ $quotient -lt 1 ]; then
    quotient=1
fi
for i in $(seq $quotient)
do
    cpuset="${cpuset}f"
done
for rps_file in $(ls /sys/class/net/eth*/queues/rx-*/rps_cpus)
do
    echo $cpuset > $rps_file
done
```

Configure multi-queue support for NICs on a Windows ECS instance

> **Note:**
> We are inviting Windows users to test the performance improvement. Windows systems see improved network performance after using multi-queue for NICs, but the improvement is not as much as for Linux systems.

If you are using a Windows instance, you must install the driver to use the multi-queue feature for NICs.

To install the driver for Windows systems, follow these steps:

1. *Open a ticket* to request and download the driver installation package.
2. Unzip the driver installation package. For Windows 2012/2016 systems, use the driver in the Win8/amd64 folder.
3. Upgrade the NIC driver:
   a. Select **Device Manager** > **Network adapters**.
   b. Right click **Red Hat VirtIO Ethernet Adapter** and select **Update Driver**.

c. Select the Win8/admin64 directory of the driver directory that you have unzipped, and update the driver.

4. Recommended: Restart the Windows system after the driver is upgraded.

The multi-queue feature for NICs is now ready to use.

7.6 Elastic network interfaces

An Elastic Network Interface (ENI) is a virtual network interface that can be attached to an ECS instance in a VPC. By using ENIs, you can build high-availability clusters, implement failover at a lower cost, and achieve refined network management. The ENI feature is available in all regions.

Scenarios

ENIs can be used in the following scenarios:

- **Deploying a high-availability cluster**

  An ENI can meet the demands of a high-availability architecture for multiple network interfaces on a single instance.

- **Providing a low-cost failover solution**

  You can detach an ENI from a failed ECS instance and then attach it to another ECS instance to quickly redirect the failed instance’s traffic to a backup instance. This action recovers the service immediately.

- **Managing the network with refined controls**

  You can configure multiple ENIs for an instance. For example, you can use some ENIs for internal management and other ENIs for Internet business access, so as to isolate managerial data from business data. You can also configure precisely-targeted security group rules for each ENI based on the source IP address, protocols, ports, and more to achieve secured traffic control.

ENI types

ENIs are classified into two types:

- **Primary ENI**

  The ENI created by default upon the creation of an instance in a VPC is called the primary ENI. The life cycle of the primary ENI is the same as that of the instance and you are not allowed to remove the primary ENI from the instance.

- **Secondary ENI**

You can create a secondary ENI and attach it to an instance or detach it from the instance. Multiple private IPs are supported for each secondary ENI. The maximum number of ENIs that you can attach to one instance varies with the instance type. For more information, see *Instance type families*.

ENI attributes

The following table displays ENI attributes.

Attribute	Quantity
Primary private IP addresses	1
MAC address	1
Security group	Min. 1, and Max. 5
Description	1
ENI name	1

Limitations

ENIs have the following limitations:

- By default, one account can own up to 100 ENIs in one region. The quota increases with the membership level. If you require a higher quota, *open a ticket*.
- The ECS instance must be in the same zone of the same region as the ENI, but they do not have to be in the same VSwitch.
- The number of ENIs that can be attached to an ECS instance is determined by the instance type. For more information, see *Instance type families*.
- Only I/O optimized instance types support ENIs.
- Attaching multiple ENIs does not increase the instance bandwidth.

Note:

The instance bandwidth capability varies with the instance type.

Related operations

For images that cannot identify ENIs, you can log on to the instance to *configure the ENI*.

Console operations

You can complete the following operations in the ECS console:

- *Attach an ENI when creating an instance*
- *Create an ENI*
- *Delete an ENI*
- *Attach an ENI to an instance*: The instance must be in a **Stopped** or **Running** status.
- *Detach an ENI from an instance*: The instance must be in a **Stopped** or **Running** status.
- *Modify attributes of an ENI*: You can modify attributes of an ENI, including its name, security group, and description.
- When an ENI is attached to an instance, you can view the information of the ENI on the instance details page and the network interfaces page.

API operations

You can complete the following operations by using APIs:

- *Create an ENI*
- *Delete an ENI*
- *Query ENI list*
- *Attach an ENI to an instance*: The instance must be in a **Stopped** or **Running** status.
- *Detach an ENI from an instance*: The instance must be in a **Stopped** or **Running** status.
- *Modify attributes of an ENI*: You can modify attributes of an ENI, including its name, its security group, and its description.
- You can use the *DescribeInstances* interface to query the information of an ENI when the ENI is attached to an instance.

7.7 Security group

A security group is a virtual firewall that provides Stateful Packet Inspection (SPI). Security groups are used to set network access control for one or more ECS instances. As an important means of security isolation, security groups are used to divide security domains on the cloud.

A security group is a logical group that contains instances in the same region with the same security requirements and mutual trust. Each instance belongs to at least one security group , which must be specified at the time of creation. Instances in the same security group can communicate through the intranet network, but instances in different security groups cannot communicate by default. However, mutual access between two security groups can be authorized.

Security group restrictions

- There is a maximum limit for the number of security groups you can have for a region. The limit depends on your level of experience with Alibaba Cloud. For new users, the limit is 100 security groups. For more experienced users, the limit is higher. To raise the upper limit, you can *open a ticket*.
- Each Elastic Network Interface (ENI) of an instance can join to up to five security groups by default. You can *open a ticket* to raise the upper limit to a maximum of 16.
- Security groups have two network types: classic network and Virtual Private Cloud (VPC).
  - Classic network instances can join security groups on classic networks in the same region.

    A single security group on a classic network cannot contain more than 1,000 instances. If you require mutual intranet access between more than 1,000 instances, you can allocate them to different security groups and authorize mutual access.
  - VPC instances can join security groups on the same VPC.

    A single security group on a VPC cannot contain more than 2,000 private IP addresses (shared by the primary and secondary ENIs). If you require mutual intranet access between more than 2,000 private IP addresses, you can allocate the relevant instances to different security groups and authorize mutual access.
- Adjusting security groups will not affect the continuity of user service.
- Security groups are stateful. If an outbound packet is permitted, inbound packets correspond ing to this connection will also be permitted.

Security group rules

Security group rules can be set that permit or forbid ECS instances in a security group from accessing a public network or intranet in the inbound and outbound directions.

You can create or delete security group rules at any time. Once changes are made, the updated security group rules are automatically applied to ECS instances in the security group.

When setting security group rules, make sure they are concise. If you add an ECS instance to multiple security groups, hundreds of rules may apply to the instance, which may cause connection errors when you access the instance.

Security group rule restrictions :

- Each security group can have a maximum of 100 security group rules in total, including both inbound and outbound rules.

- Each ENI of an instance can have a maximum of 500 security group rules.

7.8 SSH key pairs

What is an SSH key pair?

An SSH key pair, or key pair for short, is a secure authentication method offered by Alibaba Cloud for remote log-on to your Linux instance. It is an alternative to authentication using a username and password.

The cryptography feature uses the**public key** to encrypt data, and then the local client uses the **private key** to decrypt the data. Together, the public and private keys are known as a key pair.

The Linux ECS instance stores the public key. You use the private key to connect to your instance by entering SSH commands or using other tools, and you no longer need to remember a username and password to log on. Username and password authentication is disabled by ECS once the SSH key pair is enabled to guarantee security.

Benefits

Compared with typical username and password authentication, SSH key pair has the following benefits:

High security

Using an SSH key pair to log on to a Linux instance is more secure and reliable.

- A key pair prevents brute force password-cracking attacks.
- It is impossible to deduce the private key even if the public key is maliciously acquired.

Ease of use

- You can remotely log on to the instance by configuring the key pair in the ECS console and on the local client. You do not have to enter the password every time you log on.
- We recommend this method if you maintain multiple ECS instances.

Limits

Using an SSH key pair has the following restrictions:

- Applies only to Linux instances.
- Alibaba Cloud only supports the creation of 2048-bit RSA key pairs.
  - Alibaba Cloud holds the public key of the key pair.
  - After the key pair is created, you must download and keep the private key for further use.

  - The private key is in the unencrypted PEM-encoded PKCS#8 format.
- An Alibaba Cloud account can have a maximum of 500 key pairs in a region.
- A Linux instance can be only bound to one SSH key pair. If a key pair has already been bound to your instance, the new key pair replaces the old one.
- During the lifecycle of a Linux instance, you can bind or unbind an SSH key pair at any time. After you bind or unbind a key pair, you must *restart the instance* for the change to take effect.
- All instances of any *instance type family*, except for the I/O optimized instances of Generation I, support SSH key pairs.

Create an SSH key pair

To create an SSH key pair, you can use either of the following methods:

- *Create an SSH key pair* in the ECS console.

  Note:
  Once you create a key pair in the ECS console, you must immediately download and keep the private key for further use. If SSH key pair authentication is enabled for an ECS instance, you cannot log on to the ECS instance without the private key of the key pair.
- Create an SSH key pair by using other key pair builders and *import it* to ECS.

  The following key types are supported:

  - rsa
  - dsa
  - ssh-rsa
  - ssh-dss
  - ecdsa
  - ssh-rsa-cert-v00@openssh.com
  - ssh-dss-cert-v00@openssh.com
  - ssh-rsa-cert-v01@openssh.com
  - ssh-dss-cert-v01@openssh.com
  - ecdsa-sha2-nistp256-cert-v01@openssh.com
  - ecdsa-sha2-nistp384-cert-v01@openssh.com
  - ecdsa-sha2-nistp521-cert-v01@openssh.com

If your key pair is generated by Alibaba Cloud, you must download the private key and keep it safe . When a key pair is bound to an ECS instance, you cannot log on to that ECS instance if you do not have the private key.

Related operations

- If you do not have an SSH key pair, you can *create an SSH key pair*.
- If you have created an SSH key pair by using another tool, you can *import an SSH key pair*.
- If you do not need a key pair, you can *delete an SSH key pair*.
- If you want to enable or disable SSH key pair authentication for logging on to a Linux ECS instance, you can *bind or unbind an SSH key pair*.
- You can allocate an SSH key pair when *creating an ECS instance*.
- You can *log on to an instance by using an SSH key pair*.

7.9 Anti-DDoS Basic

Anti-DDoS Basic is a free Distributed Denial of Service (DDoS) protection service that safeguards data and applications on your ECS instance. As a global service from Alibaba Cloud Security, Anti-DDoS Basic offers a mitigation capacity of 5 Gbit/s against common DDoS attacks. When the inbound traffic of an ECS instance exceeds its limits, which is determined by the ECS instance type, Alibaba Cloud Security enables throttling to maintain stable performance.

How Anti-DDoS Basic works

When Anti-DDoS Basic is enabled, Alibaba Cloud Security monitors the inbound traffic in real time. When massive or abnormal traffic involving DDoS attacks is monitored, Alibaba Cloud Security redirects the traffic, drops malicious traffic, and passes clean traffic back to the ECS instance. This process is called **flow cleaning**. For more information, see *How Anti-DDoS Basic works*.

Note:

If Anti-DDoS Basic is enabled for an ECS instance, when the inbound traffic from Internet is higher than 5 Gbit/s, to secure the global cluster, Alibaba Cloud Security triggers a black hole to receive such traffic. For more information, see *Alibaba Cloud black hole policies*.

Flow cleaning is triggered in the following situations:

- When specified attacks are identified in the inbound traffic.
- When the inbound traffic to an ECS instance exceeds the specified threshold.

Methods to clean the flow include filtering ICMP packets, limiting bit rate, and limiting the packet forwarding rate.

Therefore, when using Anti-DDoS Basic, you must set the following thresholds:

- BPS threshold: When the inbound traffic exceeds the BPS threshold, flow cleaning is triggered.
- PPS threshold: When the inbound packet forwarding rate exceeds the PPS threshold, flow cleaning is triggered.

Cleaning thresholds of each instance type

The configuration of each instance type determines its maximum flow cleaning threshold. The following table lists the cleaning thresholds of some *available* and *phased-out* instance types.

Instance type	Maximum BPS threshold (Mbit/s)	Maximum PPS threshold (PPS)
ecs.g5.16xlarge	20,000	4,000,000
ecs.g5.22xlarge	30,000	4,500,000
ecs.g5.2xlarge	2,500	800,000
ecs.g5.4xlarge	5,000	1,000,000
ecs.g5.6xlarge	7,500	1,500,000
ecs.g5.8xlarge	10,000	2,000,000
ecs.g5.large	1,000	300,000
ecs.g5.xlarge	1,500	500,000
ecs.sn2ne. 14xlarge	10,000	4,500,000
ecs.sn2ne. 2xlarge	2,000	1,000,000
ecs.sn2ne. 4xlarge	3,000	1,600,000
ecs.sn2ne. 8xlarge	6,000	2,500,000
ecs.sn2ne.large	1,000	300,000
ecs.sn2ne.xlarge	1,500	500,000
ecs.c5.16xlarge	20,000	4,000,000
ecs.c5.2xlarge	2,500	800,000
ecs.c5.4xlarge	5,000	1,000,000
ecs.c5.6xlarge	7,500	1,500,000
ecs.c5.8xlarge	10,000	2,000,000

Instance type	Maximum BPS threshold (Mbit/s)	Maximum PPS threshold (PPS)
ecs.c5.large	1,000	300,000
ecs.c5.xlarge	1,500	500,000
ecs.sn1ne. 2xlarge	2,000	1,000,000
ecs.sn1ne. 4xlarge	3,000	1,600,000
ecs.sn1ne. 8xlarge	6,000	2,500,000
ecs.sn1ne.large	1,000	300,000
ecs.sn1ne.xlarge	1,500	500,000
ecs.r5.16xlarge	20,000	4,000,000
ecs.r5.22xlarge	30,000	4,500,000
ecs.r5.2xlarge	2,500	800,000
ecs.r5.4xlarge	5,000	1,000,000
ecs.r5.6xlarge	7,500	1,500,000
ecs.r5.8xlarge	10,000	2,000,000
ecs.r5.large	1,000	300,000
ecs.r5.xlarge	1,500	500,000
ecs.re4.20xlarge	15,000	2,000,000
ecs.re4.40xlarge	30,000	4,000,000
ecs.se1ne. 14xlarge	10,000	4,500,000
ecs.se1ne. 2xlarge	2,000	1,000,000
ecs.se1ne. 4xlarge	3,000	1,600,000
ecs.se1ne. 8xlarge	6,000	2,500,000
ecs.se1ne.large	1,000	300,000
ecs.se1ne.xlarge	1,500	500,000
ecs.se1.14xlarge	10,000	1,200,000
ecs.se1.2xlarge	1,500	400,000
ecs.se1.4xlarge	3,000	500,000
ecs.se1.8xlarge	6,000	800,000
ecs.se1.large	500	100,000

Instance type	Maximum BPS threshold (Mbit/s)	Maximum PPS threshold (PPS)
ecs.d1ne. 2xlarge	6,000	1,000,000
ecs.d1ne. 4xlarge	12,000	1,600,000
ecs.d1ne. 6xlarge	16,000	2,000,000
ecs.d1ne. 8xlarge	20,000	2,500,000
ecs.d1ne. 14 x large	35,000	4,500,000
ecs.d1.2xlarge	3,000	300,000
ecs.d1.4xlarge	6,000	600,000
ecs.d1.6xlarge	8,000	800,000
ecs.d1.8xlarge	10,000	1,000,000
ecs.d1-c8d3.8xlarge	10,000	1,000,000
ecs.d1.14xlarge	17,000	1,800,000
ecs.d1-c14d3.14xlarge	17,000	1,400,000
ecs.i2.xlarge	1,000	500,000
ecs.i2.2xlarge	2,000	1,000,000
ecs.i2.4xlarge	3,000	1,500,000
ecs.i2.8xlarge	6,000	2,000,000
ecs.i2.16xlarge	10,000	4,000,000
ecs.i1.xlarge	800	200,000
ecs.i1.2xlarge	1,500	400,000
ecs.i1.4xlarge	3,000	500,000
ecs.i1-c10d1.8xlarge	6,000	800,000
ecs.i1-c5d1.4xlarge	3,000	400,000
ecs.i1.14xlarge	10,000	1,200,000
ecs.hfc5.large	1,000	300,000
ecs.hfc5.xlarge	1,500	500,000
ecs.hfc5.2xlarge	2,000	1,000,000
ecs.hfc5.4xlarge	3,000	1,600,000
ecs.hfc5.6xlarge	4,500	2,000,000

Instance type	Maximum BPS threshold (Mbit/s)	Maximum PPS threshold (PPS)
ecs.hfc5.8xlarge	6,000	2,500,000
ecs.hfg5.large	1,000	300,000
ecs.hfg5.xlarge	1,500	500,000
ecs.hfg5.2xlarge	2,000	1,000,000
ecs.hfg5.4xlarge	3,000	1,600,000
ecs.hfg5.6xlarge	4,500	2,000,000
ecs.hfg5.8xlarge	6,000	2,500,000
ecs.hfg5.14xlarge	10,000	4,000,000
ecs.c4.2xlarge	3,000	400,000
ecs.c4.4xlarge	6,000	800,000
ecs.c4.xlarge	1,500	200,000
ecs.ce4.xlarge	1,500	200,000
ecs.cm4.4xlarge	6,000	800,000
ecs.cm4.6xlarge	10,000	1,200,000
ecs.cm4.xlarge	1,500	200,000
ecs.gn5-c28g1.14xlarge	10,000	4,500,000
ecs.gn5-c4g1.xlarge	3,000	300,000
ecs.gn5-c4g1.2xlarge	5,000	1,000,000
ecs.gn5-c8g1.2xlarge	3,000	400,000
ecs.gn5-c8g1.4xlarge	5,000	1,000,000
ecs.gn5-c28g1.7xlarge	5,000	2,250,000
ecs.gn5-c8g1.8xlarge	10,000	2,000,000
ecs.gn5-c8g1.14xlarge	25,000	4,000,000
ecs.gn5i-c2g1.large	1,000	100,000
ecs.gn5i-c4g1.xlarge	1,500	200,000
ecs.gn5i-c8g1.2xlarge	2,000	400,000
ecs.gn5i-c16g1.4xlarge	3,000	800,000
ecs.gn5i-c28g1.14xlarge	10,000	2,000,000

Instance type	Maximum BPS threshold (Mbit/s)	Maximum PPS threshold (PPS)
ecs.gn4-c4g1.xlarge	3,000	300,000
ecs.gn4-c8g1.2xlarge	3,000	400,000
ecs.gn4-c4g1.2xlarge	5,000	500,000
ecs.gn4-c8g1.4xlarge	5,000	500,000
ecs.gn4.8xlarge	6,000	800,000
ecs.gn4.14xlarge	10,000	1,200,000
ecs.ga1.xlarge	1,000	200,000
ecs.ga1.2xlarge	1,500	300,000
ecs.ga1.4xlarge	3,000	500,000
ecs.ga1.8xlarge	6,000	800,000
ecs.ga1.14xlarge	10,000	1,200,000
ecs.f1-c28f1.7xlarge	5,000	2,000,000
ecs.f1-c8f1.2xlarge	2,000	800,000
ecs.f2-c28f1.14xlarge	10,000	2,000,000
ecs.f2-c28f1.7xlarge	5,000	1,000,000
ecs.f2-c8f1.2xlarge	2,000	400,000
ecs.f2-c8f1.4xlarge	5,000	1,000,000
ecs.t5-c1m1.2xlarge	1,200	400,000
ecs.t5-c1m1.large	500	100,000
ecs.t5-c1m1.xlarge	800	200,000
ecs.t5-c1m2.2xlarge	1,200	400,000
ecs.t5-c1m2.large	500	100,000
ecs.t5-c1m2.xlarge	800	200,000
ecs.t5-c1m4.2xlarge	1,200	400,000
ecs.t5-c1m4.large	500	100,000
ecs.t5-c1m4.xlarge	800	200,000
ecs.t5-lc1m1.small	200	60,000
ecs.t5-lc1m2.large	400	100,000

Instance type	Maximum BPS threshold (Mbit/s)	Maximum PPS threshold (PPS)
ecs.t5-lc1m2.small	200	60,000
ecs.t5-lc1m4.large	400	100,000
ecs.t5-lc2m1.nano	100	40,000
ecs.ebmg4.8xlarge	10,000	4,500,000
ecs.ebmg5.24xlarge	10,000	4,500,000
ecs.sccg5.24xlarge	10,000	4,500,000
ecs.xn4.small	500	50,000
ecs.mn4.small	500	50,000
ecs.mn4.large	500	100,000
ecs.mn4.xlarge	800	150,000
ecs.mn4.2xlarge	1,200	300,000
ecs.mn4.4xlarge	2,500	400,000
ecs.n4.small	500	50,000
ecs.n4.large	500	100,000
ecs.n4.xlarge	800	150,000
ecs.n4.2xlarge	1,200	300,000
ecs.n4.4xlarge	2,500	400,000
ecs.n4.8xlarge	5,000	500,000
ecs.e4.small	500	50,000
ecs.sn1.medium	500	100,000
ecs.sn1.large	800	200,000
ecs.sn1.xlarge	1,500	400,000
ecs.sn1.3xlarge	3,000	500,000
ecs.sn1.7xlarge	6,000	800,000
ecs.sn2.medium	500	100,000
ecs.sn2.large	800	200,000
ecs.sn2.xlarge	1,500	400,000
ecs.sn2.3xlarge	3,000	500,000

Instance type	Maximum BPS threshold (Mbit/s)	Maximum PPS threshold (PPS)
ecs.sn2.7xlarge	6,000	800,000
ecs.sn2.13xlarge	10,000	120,000

Related operations

By default, Anti-DDoS Basic is enabled for an ECS instance after it is created. You can do the following:

- Set a threshold for flow cleaning. After an ECS instance is created, the maximum threshold for the instance type is used for Anti-DDoS Basic by default. However, the maximum BPS threshold for some instance types is excessive for security. Therefore, you must set a threshold according to your business needs. For more information, see *DDoS basic protection configuration* in the Anti-DDoS Basic documentation.
- Cancel flow cleaning, which is not recommended. When the inbound traffic to an ECS instance exceeds the cleaning threshold, the traffic, including normal business traffic, is cleaned. To avoid business interruptions, you can cancel flow cleaning. For more information, see *How to cancel flow cleaning*Cite LeftCite Right.

  Warning:
  If you cancel flow cleaning, when the inbound traffic to an ECS instance exceeds 5 Gbit/s, all traffic is routed to a black hole.

8 Images

An image is a running environment template for ECS instances. It generally includes an operating system and preinstalled software. You can use an image to create an ECS instance or change the system disk of an ECS instance. It works as a file copy that includes data from one more multiple disks. These disks can be a single system disk, or the combination of the system disk and data disks.

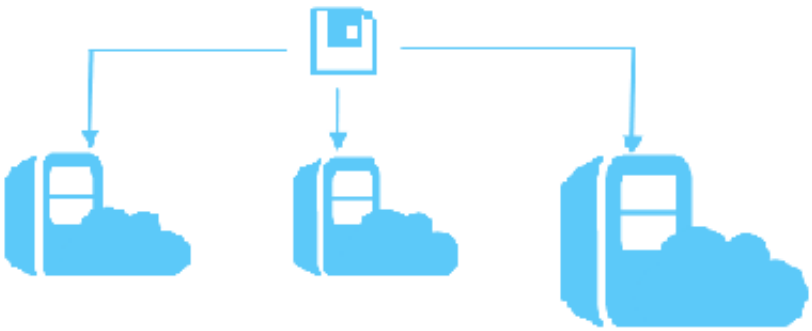


Image types

ECS provides a diverse types of images for you to easily access image resources.

Type	Description	Source
Public image	Public images officially provided by Alibaba Cloud support nearly all main Windows and Linux versions. These images are of high stability and are licensed. You can customize your application environment based on a public image.	Officially provided by Alibaba Cloud
Custom image	Custom images created based on your existing physical server, virtual machine, or cloud host. These images are flexible to meet your personalized needs.	• You can create it based on an existing instance. • You can also import one from the on-premises environment into the corresponding region.
Cloud Marketplace	Provided by third-party service providers (ISV, independent software) Vendor). The image of the Marketplace includes not only the operating system required for the application, but also the configuration environment. It saves you complicated deployment process and deploy the environment with one-click.	*Alibaba Cloud Marketplace*

Type	Description	Source
Shared image	Shared by other Alibaba Cloud users.	A custom image shared by other Alibaba Cloud users.

Public images

Alibaba Cloud provides authorized and certificated public images, which cover nearly all the trending and popular platforms. Note that the available public images are different when you select different instances. For the built-in service contained in the various public image releases, please go to the official website of the operating system provider for reference. The following images are offered by Alibaba Cloud ECS for public use:

Platform	Version of public images
Windows Server	• Windows Server 2008 R2 Enterprise Edition 64 bit Chinese Edition • Windows Server 2008 R2 Enterprise Edition 64 bit English Edition • Windows Server 2012 R2 Data Center Edition 64 bit Chinese Edition • Windows Server 2012 R2 Data Center Edition 64 bit English Edition • Windows Server 2016 R2 Data Center Edition 64 bit Chinese Edition • Windows Server 2016 R2 Data Center Edition 64 bit English Edition • Windows Server Version 1709 Data Center Edition 64 bit Chinese Edition • Windows Server Version 1709 Data Center Edition 64 bit English Edition
CentOS	• CentOS 6.8 64bit • CentOS 6.8 32bit • CentOS 6.9 64bit • CentOS 7.2 64bit • CentOS 7.3 64bit • CentOS 7.4 64bit • CentOS 7.5 64bit
Ubuntu	• Ubuntu 14.04 64bit • Ubuntu 14.04 32bit • Ubuntu 16.04 64bit • Ubuntu 16.04 32bit
Debian	• Debian 8.9 64bit • Debian 9.2 64bit • Debian 9.5 64bit
Red Hat	• Red Hat Enterprise Linux 7.5 64bit

Platform	Version of public images
	• Red Hat Enterprise Linux 7.4 64bit • Red Hat Enterprise Linux 6.9 64bit
SUSE Linux	• SUSE Linux Enterprise Server 11 SP4 64bit • SUSE Linux Enterprise Server 12 SP4 64bit
OpenSUSE	OpenSUSE 42.3 64bit
Aliyun Linux	Aliyun Linux 17.1 64bit
CoreOS	CoreOS 1465.8.0 64bit
FreeBSD	FreeBSD 11.1 64bit

Image format

Currently, ECS supports VHD, qcow2, and RAW. You must convert other formats before using them in ECS.

Billing details

Billing details of the images are as follows:

Type	Description
Public image	Only Windows Server and Red Hat Enterprise Linux public images involve billing calculation, which are included in the bill when an instance is created. The public Windows Sever images or Red Hat Enterprise Linux images are provided with certificated license and authorized support from Microsoft or Red Hat, you do have to purchase additional license. • Red Hat Enterprise Linux: Billing is related to the instance type. • Windows Server: Free of charge in Alibaba Cloud regions that are in mainland China. For international regions, public Windows Server images are charging services. Other public images are free of charge to use.
Custom image	Free. Potential costs include: • If you use a snapshot to create a custom image: — If the image used by the system disk snapshot comes from the Marketplace, the following cost may incur: the fees for the image, and the fees for snapshot capacity. — If the image used by the system disk snapshot does not come from the Marketplace, the following cost may incur: the fees for snapshot capacity.

Type	Description
	Currently, snapshot is commercialized. • If you use an instance to create a custom image, and the image is from the Marketplace, comply to the billing policies from the ISV.
Alibaba Cloud Marketplace	Subject to ISV policies.
Shared image	If the origin of the shared image is from the Marketplace, it is subject to the ISV policies.

For more information, see *Pricing overview*.

Limits

Custom images, marketplace images, and shared images vary depending on the region. For more information about regions and zones, see *Cite LeftGeneral ReferenceCite Right* *Regions and zones*.

Related operations

Console operations

- You can *create instances by using existing images*.
- You can change the system disk in any of the following ways:
  - *By using a public image*.
  - *By using other images other than public ones*.
- You can obtain custom images in the following ways:
  - *By creating a custom image by using a snapshot*.
  - *By creating a custom image by using a an instance*.
  - *By importing a custom image*.
- After creating custom images, you can perform the following operations:
  - *Copy your custom images to other regions*.
  - *Share your custom images with other Alibaba Cloud users*.
  - *Export custom images* to local testing environments or your private cloud environments.

API operations

You can view the *APIs about images* in the Developer Guide.

9 Snapshots

9.1 What are ECS snapshots

A snapshot is a copy of data on an elastic block storage device at a given time point. For more information about how snapshots are created, see Incremental snapshot mechanism.

> **Note:**
> Creation of a snapshot may reduce the I/O performance of a block storage device, generally by less than 10%, resulting in sharp decrease in I/O speed. We recommend that you create snapshots during off-peak business hours.

Features

Currently, Alibaba Cloud provides Snapshot 2.0 service. Compared with the former version, Snapshot 2.0 has better performance in capacity limit, scalability, cost, and usability. For more information, see ECS Snapshot 2.0 vs. traditional storage products.

> **Note:**
> The Snaphsot 2.0 service is currently online. Unless and otherwise specified, either “snapshot” or “snapshot service” in all ECS articles is assumed as Snaphsot 2.0 service.

Scenarios

The snapshot service meets your requirements, such as:

- Creating an elastic block storage device that has the data of an existing storage device. For example, by using the snapshot service, you can create a cloud disk from a snapshot.
- Restoring the data on an elastic block storage device. You can roll back the storage device from a snapshot. For example, when the data on an elastic block storage device is incorrect caused by an application error or the data is maliciously tampered by hackers by using an application vulnerability, you can use its snapshot to restore its data to the expected status.
- Creating multiple copies of production data. You can create a custom image from a snapshot of a system disk of an existing instance, and then create the image to create a new instance.

For more information, see Scenarios.

Classification

Snapshots are classified into two categories:

- Manual snapshots, which are created manually. You can create snapshots for an elastic block storage device at any time to back up data.
- Auto snapshots, which are created automatically according to the automatic snapshot policy applied to an elastic block storage device. You can create an automatic snapshot policy and apply it to the storage device. Then snapshots will be created automatically at the given time points.

Snapshot charges

Currently, the snapshot service is free of charge.

View the size of snapshots

A snapshot chain of an elastic block storage device is created once the first snapshot is created. You can view the total size of the snapshots of an elastic block storage device by using the **Snapshot Chain** feature in the ECS console.

Encryption

All the snapshots of encrypted cloud disks or shared block storage are encrypted. These snapshots are called encrypted snapshots. Encrypted snapshots cannot be converted to unencrypted snapshots, and vice versa. For more information, see ECS disk encryption.

Delete snapshots

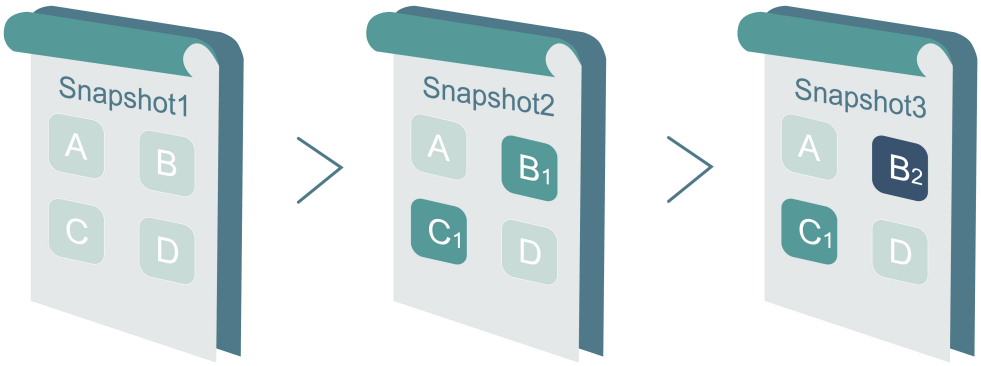
If your business no longer requires a snapshot of an elastic block storage device, you can delete the snapshot. If you have applied an automatic snapshot policy to the storage device, you can delete the automatic snapshot policy.

9.2 Incremental snapshot mechanism

Alibaba Cloud provides the snapshot feature. You can create snapshots as scheduled. Save disk data at a specific time to guranttee the availability of your business.

Incremental snapshot mechanism

In this method, two snapshots are compared and only the data that has changed is copied. See the following figure.



- In the preceding figure, Snapshot 1, Snapshot 2, and Snapshot 3 are the first, second, and third snapshots of a disk. The file system checks the disk data by blocks. When a snapshot is created, only the blocks with changed data are copied to the snapshot. In this example:
  1. In Snapshot 1, all data on the disk is copied because it is the first disk snapshot.
  2. Snapshot 2 only copies the changed data blocks B1 and C1. Data blocks, A and D, are referenced from Snapshot 1.
  3. Snapshot 3 copies the changed data block B2 but references data blocks, A, D, from Snapshot 1, and references C1 from Snapshot 2.
- When you roll back the disk to Snapshot 3, blocks A, B2, C1, and D are copied to the disk, to replicate Snapshot 3.
- When you delete Snapshot 2, block B1 is deleted, but block C1 is retained because blocks that are referenced by other snapshots cannot be deleted. When you roll back a disk to Snapshot 3 , block C1 is recovered.

Creation time

Snapshot creation time varies depending on actual volume to be copied. It takes long time to create the first snapshot of a disk, because the snapshot copies the global data. Then only the blocks with changed data are copied to a snapshot, which consumes shorter time.

Influence of snapshot creation

When snapshot creation is in progress, the performance of a disk is reduced.

Snapshot chains

A snapshot chain contains all snapshots of a disk. Each disk has one snapshot chain, and the snapshot chain ID is identical to the disk ID. A snapshot chain has the following information:

- Snapshot quantity: The number of existing snapshots of a disk.

- Snapshot capacity: The storage space that all the snapshots in the chain occupy.

  Note:
  The snapshot service charges according to the snapshot capacity. You can use the snapshot chain to check the snapshot capacity for each disk.

- Snapshot quota: Each disk has a maximum of 64 snapshots. Therefore, each chain can have up to 64 snapshots, including manual and automatic snapshots.

  Note:
  When the snapshot quota is exceeded, if more automatic snapshots are to be created, the automatic snapshots are deleted automatically in a chronological order; if you want to create more snapshots manually, delete unnecessary snapshots manually. For more information, see *Cite LeftApply an automatic snapshot policy to a disk and Delete a snapshotCite Right*

9.3 ECS Snapshot 2.0

Built on original basic snapshot features, ECS Snapshot 2.0 data backup service provides a higher snapshot quota and more flexible automatic task policies, further reducing its impact on business I/O. The features of ECS Snapshot 2.0 are described in the following table.

Note:
Disks in this topic refer to Elastic Block Storage. For more information, see *Cloud disks and Shared Block Storage*.

Feature	Original snapshot specifications	Snapshot 2.0 specifications	User benefit	Example
Snapshot quota	(Number of disks)*6+6	64 snapshots for each disk	Longer protection circle Smaller protection granularity	• Snapshot backup of a data disk for non-core businesses occurs at 00:00 every day. This backup data is retained for over 2 months.

Feature	Original snapshot specifications	Snapshot 2.0 specifications	User benefit	Example
				• Snapshot backup of a data disk for core businesses occurs every 4 hours. This backup data is retained for over 10 days.
Automatic task policy	Hardcoded, triggered once daily, and unmodifiable	Customizable weekly snapshot day, time of day , and snapshot retention period Query-able disk quantity and related details associated with an automatic snapshot policy	More flexible protection policy	• A user can take snapshots on the hour and for several times in a day. • A user can choose any day as the recurring day for taking weekly snapshots. • A user can specify the snapshot retention period or choose to retain it permanentl y. When the maximum number of automatic snapshots has been reached , the oldest automatic

Feature	Original snapshot specifications	Snapshot 2.0 specifications	User benefit	Example
				snapshot will be deleted.
Implementation principle	COW (Copy-on-write)	ROW (Redirect-on-write)	Mitigated performance impact of the snapshot task on business I/O write	The implementation principle is not made visible to users, allowing snapshots to be taken at any time of day without affecting user experience.

9.4 ECS Snapshot 2.0 vs. traditional storage products

Alibaba Cloud ECS Snapshot 2.0 has many advantages compared with the snapshot feature of traditional storage products, as described in the following table.

Comparison item	ECS Snapshot 2.0	Snapshot feature of traditional storage products
Capacity limit	Unlimited capacity, meeting data protection needs for extra-large businesses.	Capacity limited by initial storage device capacity, merely meeting data protection needs for a few core services.
Scalability	One-click auto scaling, allowing you to scale up and down according to their business scale, in mere seconds.	Poor scalability, restrained by factors such as production and storage performance, available capacity, and vendor support capabilities. Scaling typically takes 1 ~ 2 weeks.
Cost	Billed based on the actual amount of data changed in your business and snapshot size.	Large, inefficient upfront investment involving software licenses, reserved space, and upgrade and maintenance expenses.
Usability	24x7 online post-sales support.	Complex operations, greatly restrained by vendor support capabilities.

9.5 Scenarios

As a simple and efficient data protection method, snapshots are recommended for the following scenarios:

- Routine backup of system and data disks. You can back up business-critical data at regular intervals by using snapshots to avoid data loss caused by misoperations, attacks, viruses, and others.
- Before important operations such as replacing system disks, upgrading application software , or migrating business data, you must create one or more snapshots. In case that any issue occurs during an upgrade or migration, you can timely restore it to normal status by using the snapshots.
- Using of multiple copies of production data. You can take snapshots of production data to provide close-to-real-time production data for data mining, report queries, and developing and testing applications.

10 Cloud assistant

Cloud assistant is a lightweight and convenient maintenance ECS feature for automated and batched invocation of daily maintenance tasks.

By installing the cloud assistant client on ECS instances, you can run Bat/PowerShell (for Windows instances) scripts or Shell scripts (for Linux instances) on one or more running ECS instances in the ECS console or by calling APIs. The invocation is exclusive to individual instances to complete tasks rapidly. You can also set command invocation to the periodical mode to keep the ECS instance at a specific status or run the command as a daemon for ECS instances. Cloud assistant does not initiate any operations. All operations are within your controllable range.

Scenarios

You can use cloud assistant in the following scenarios.

- Install, uninstall, or update applications for ECS instances that are in the `Running` status.
- Update patches for ECS instances that are in the Running status.
- Add configuration for ECS instances that are in the Running status.
- Set daemon process for ECS instances that are in the Running status.
- Retrieve monitoring and log information for ECS instances that are in the Running status.
- Other maintenance tasks that must be completed by running scripts.

Terminology

Term	Common name	Description
Cloud assistant	Cloud assistant	A convenient feature provided by Alibaba Cloud ECS for automated and batched invocation of daily maintenance tasks.
Cloud assistant client	Client	The client program that is installed on ECS instance. All operations to ECS instances are performed by using the client.
Command	Command	The specific command and operation to be invoked on ECS instances, such as a shell script.
One-time invocation	Invocation	One or more ECs in the specified If a command is invoked on an instance or multiple instances only once then, it is called as one-time invocation (`Invocation`).
Periodical invocation	Timed Invocation	When you invoke a command on an instance or multiple instances, you can specify the invocation sequence/period to run the command process periodically.

Term	Common name	Description
Invocation status	InvokeStatus	The relationship among command invocation status. The invocation status can be divided into three levels: • **Overall invocation status:** The general invocation status for all the target ECS instances when invoking a daemon process. • **Instance invocation status:** The invocation status of command invocation that are batch processed on all the ECS instances. • **Invocation-record status:** The invocation status of a specific ECS instance when invoking a command.

Limits

Cloud assistant has the following limits:

- You must install and manage the cloud assistant as the administrator. Specifically, the Linux instance administrator is root and the Windows instance administrator is administrator.
- You must manage the cloud assistant as an the administrator.
- The size of the source Bat/PowerShell script or Shell script must be less than 16 KB.
- Requirements on the status of the target ECS instances:
  - ECS instances must be connected to the intranet.
  - The ECS instance must be in the `Running` status.
  - The network type of the ECS instance must be VPC.
- Other limits for using cloud assistant:

Supported images	Supported regions
• Windows Server 2008/2012/2016 • Ubuntu 12/14/16 • Centos 5/6/7 • Debian 7/8/9 • RedHat 5/6/7 • SuSE Linux Enterprise Server 11/12 • Opensuse • Aliyun Linux • Freebag • Coreos	• China East 1 (Hangzhou) • China East 2 (Shanghai) • China North 2 (Beijing) • China North 3 (Zhangjiakou) • China South 1 (Shenzhen) • Hong Kong • Asia Pacific SE 1 (Singapore) • Asia Pacific SE 2 (Sydney) • Asia Pacific SE 3 (Kuala Lumpur) • US East 1 (Virginia)

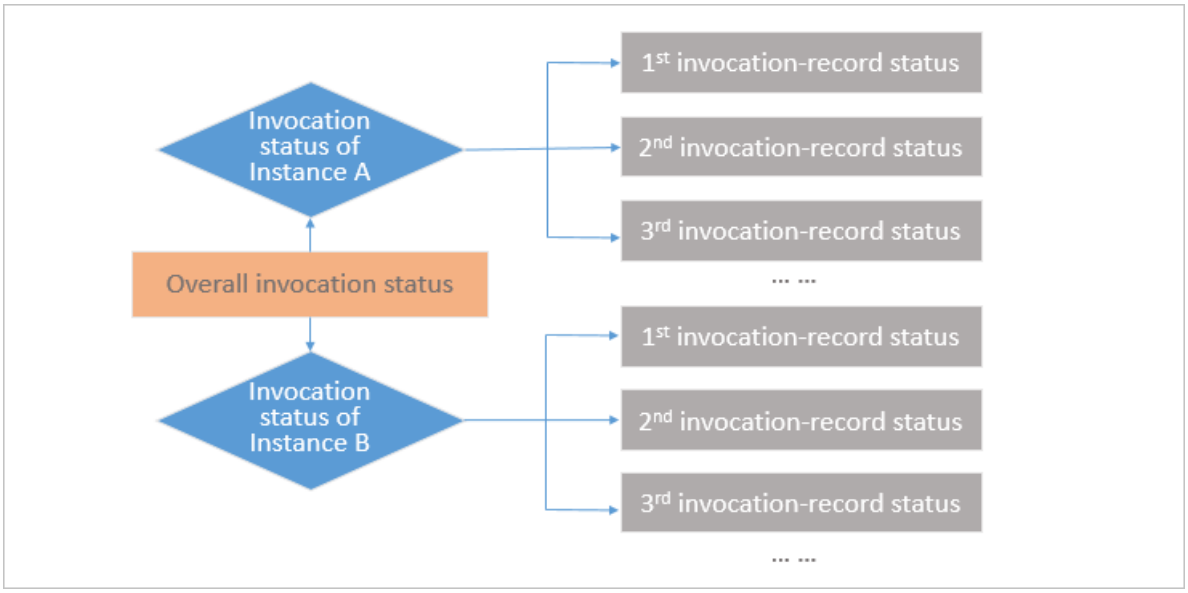
Supported images	Supported regions
	• Germany 1 (Frankfurt)

Billing details

Cloud assistant features are free of charge.

Invocation status

- Specifically, the invocation status of a command consists of `Running`, `Stopped`, `Finished`, and `Failed`.
- Generally, the invocation status of a command includes **overall invocation status** , **instance invocation status** , and **invocation-record status**. The relationships among various levels are shown in the following figure.

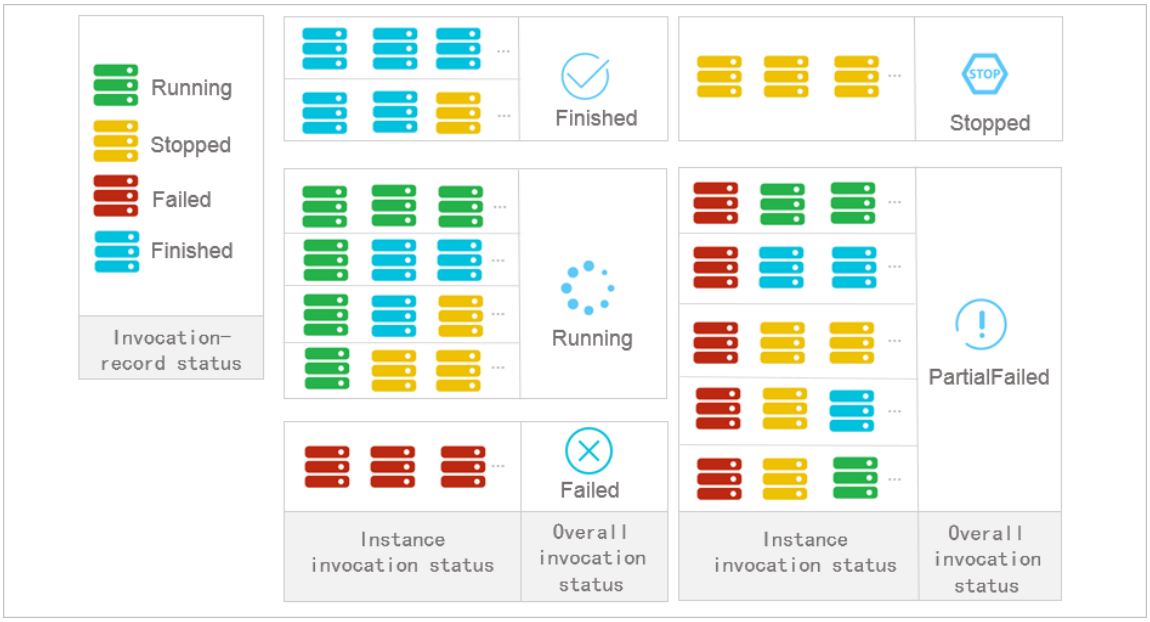


For one-time invocations

- **Overall invocation status**:
  - When the invocation status of all instances are `Finished`, the overall invocation status is displayed as `Finished`.
  - When the invocation status of some instances are `Finished` and those of some others are `Stopped`, the overall invocation status is displayed as. `Finished`
  - When the invocation status of all instances are `Failed`, the overall invocation status is displayed as `Failed`.
  - When the invocation status of all instances are `Stopped`, the overall invocation status is displayed as `Stopped`.

- When the invocation statuses of all or some instances are `Running`, the overall invocation status is displayed as `Running`.
- When the invocation statuses of some instances are `Failed`, the overall invocation status is displayed as `PartialFailed`.

  Take three ECS instances as an example. The following picture shows the relationships between the overall invocation status and the instance invocation status during a one-time invocation on multiple instances.



- **Instance invocation status**: The command is invoked only once in a one-time invocation, so the instance invocation status and the invocation-record status are identical.
- **Invocation-record status**:
  - `Running`: Indicates that the command is being executed.
  - `Stopped`: Indicates that the command invocation has been manually stopped by the user.
  - `Finished`: Indicates that the command invocation has been completed smoothly. But invocation completion does not indicate invocation success. You can confirm whether the invocation is successful based on the actual `Output`of the command process.
  - `Failed`: Indicates that the command process has timed out (`Timeout`) and failed.

For periodical invocations

- **Overall invocation status**: The overall invocation status is always `Running` unless you stop all the scheduled invocation for all instances.
- **Instance invocation status**: The instance invocation status is always `Running` unless you stop the current invocation.

- **Invocation-record status**:
  - `Running`: The command is being executed.
  - `Stopped`: You have stopped the command invocation.
  - `Finished`: The command invocation is complete. However, invocation completion does not guarantee invocation success. You can confirm whether the invocation is successful or not based on the actual `Output` of the command process.
  - `Failed`: The command process is timed out (`Timeout`) and fails.

How to use cloud assistant

You must install cloud assistant client on your ECS instance beforehand to use cloud assistant.

Currently the cloud assistant is not available on the console. You can use it by APIs. For more information, see *Auto manage instances*.

References

- *Cloud Assistant Client*
- APIs:
  - *CreateCommand*
  - *InvokeCommand*
  - *DescribeInvocations*
  - *DescribeInvocationResults*
  - *StopInvocation*
  - *ModifyCommand*
  - *DescribeCommands*
  - *DeleteCommand*