

Alibaba Cloud Elastic Compute Service

Product Introduction

Issue: 20180812

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 <code>cd /d C:/windows</code> command to enter the Windows system folder.
<i>Italics</i>	It is used for parameters and variables.	<code>bae log list --instanceid Instance_ID</code>
[] or [a b]	It indicates that it is a optional value, and only one item can be selected.	<code>ipconfig [-all -t]</code>
{ } or {a b}	It indicates that it is a required value, and only one item can be selected.	<code>swich {stand / slave}</code>

Contents

Legal disclaimer.....	I
Generic conventions.....	I
1 What is ECS.....	1
2 Benefits of ECS.....	5
3 Scenarios.....	10
4 Instance type families.....	11
5 Instances.....	48
5.1 What are ECS instances.....	48
5.2 ECS instance life cycle.....	48
5.3 Spot instances.....	51
5.4 ECS Bare Metal Instance and Super Computing Clusters.....	57
5.5 Burstable instances.....	62
5.6 Launch templates.....	70
6 Block storage.....	72
6.1 What is block storage?.....	72
6.2 Storage parameters and performance test.....	73
6.3 Elastic block storage.....	79
6.4 Triplicate technology.....	82
6.5 ECS disk encryption.....	84
6.6 Local disks.....	87
8 Images.....	91
9 Snapshots.....	95
9.1 What are ECS snapshots.....	95
9.2 Incremental snapshot mechanism.....	96
9.3 ECS Snapshot 2.0.....	98
9.4 ECS Snapshot 2.0 vs. traditional storage products.....	100
9.5 Scenarios.....	101
10 Cloud assistant.....	102
10.1 Cloud assistant.....	102

1 What is ECS

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 operating system, 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, operating system, and other affiliated resources for your ECS instance.



Basic concepts

Know the basic concepts before you proceed to use ECS:

- **Region and zone**: A physical location where a data center is located.
- **Instance**: A virtual computing environment that includes CPU, memory, operating system, bandwidth, disks, and other basic computing components.
- **Instance type**: The specification of an ECS instance, including the number of vCPU cores, memory, networking performance. An instance type determines the compute capability of an ECS instance.
- **Images**: A running environment template for ECS instances. It generally includes an operating system and preinstalled software.
- **Block storage**: Block level storage products for your ECS, including elastic 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 data on an elastic block storage device at a given time point.
- **Network types**: Alibaba Cloud provides two network types, including
 - 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 groups**: A logical group that groups instances in the same region with the same security requirements and mutual trust. A security group works as a virtual firewall for the ECS instances inside it.
- **SSH key pairs**: A secure authentication method to remotely log on to Linux instances. The public key is placed in a Linux instance, and you can use the private key to log on to the instance by using SSH commands or related tools. Besides, you can use a [password](#) to log on to a Linux instance.
- **IP address**: When an ECS instance is created, a private IP address is assigned to it for [intranet communication](#). If the instance needs Internet access, a public IP address is assigned.
- **EIP address (EIP)**: A public IP address resource that you can purchase and possess independently. You can bind an EIP address to a VPC-Connected ECS instance.
- **ECS console**: The Web application for managing ECS instances.
- **ECS console**: The Web application for managing ECS instances.

Related services

Alibaba Cloud marketplace is an online market. You can purchase software infrastructure, developer tools, and business software provided by third-party partners. You can become a marketplace service provider.

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, and to reduce the need of manual provision. For more information, see [What is Auto Scaling](#).

Container Service enables you to manage the lifecycle 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 [Introduction to 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 to Alibaba Cloud services and to manage your applications based on the language of your choice. For more information, see [Developer Resources](#). You can use [OpenAPI Explorer](#) to debug ECS API and generate the SDK Demo.

Operations

Alibaba Cloud offers a Web Services page to help you manage your cloud server ECS. 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](#).

Pricing and billing

ECS supports both Subscription and Pay-As-You-Go as billing methods. For more information, see [Billing methods](#) .

For the price information, see the [Pricing](#) page.

2 Benefits of ECS

Compared with Internet Data Centers (IDCs) and server vendors, the ECS has the benefits in the following aspects: Availability, Security, 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 region of Alibaba Cloud 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](#). Both three centres, e-commerce and video services, etc, you can find the corresponding industry solutions in ALI cloud.

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

Users of cloud computing are most concerned about security and stability. Alibaba Cloud has recently passed a host of international information security certifications, including ISO 27001 and MTCS, which demand strict confidentiality of user data and user information and user privacy protection. [Alibaba Cloud VPC](#) is the prime choice for providing your cloud computing services.

- **Alibaba Cloud VPC offers more business possibilities.** You only need to perform simple configuration to connect your business environment to global IDCs, making your business more flexible, stable, and extensible.

- **For the original self-built IDC computer room, there will be no problems.** 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 the robust networking derived from Alibaba Cloud's various hybrid cloud solutions and network products. A superior business ecosystem is 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 endless attack traffic on the Internet. After you build your business on VPC, the first line of defense is established.

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

Elasticity

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

Elastic computing

Vertical scaling involves modifying the configurations of a server. It is difficult to make configurations in the traditional IDC model. After you purchase ECS or storage capacity of Alibaba Cloud, you can configure your server with great flexibility based on your actual transaction volume. For more information about vertical scaling, see [Change configurations](#).

Horizontal scaling For example, at peak hours for game or live video streaming apps, in the traditional IDC model, your hands may be tied when the request for additional resources arises. Cloud computing now leverages elasticity to tide you over that period. When the period ends, you release unnecessary resources to reduce your business cost. 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](#).

Horizontal scaling For example, at peak hours for game or live video streaming apps, in the traditional IDC model, your hands may be tied when the request for additional resources arises. Cloud computing now leverages elasticity to tide you over that period. When the period ends, you release unnecessary resources to reduce your business cost. 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 has elastic storage. When more storage space is required, in the traditional IDC model, you can only add servers, but the number of servers that you can add is limited. However, in the cloud computing model, the sky is the limit. Order as you want to guarantee the sufficient storage space. For more information about elastic storage, see [Resize a disk](#).

Elastic network

Alibaba Cloud features elastic network as well. When you purchase the Alibaba Virtual Private Cloud (VPC), you can have the network configurations the same as those of data centers. In addition, you can have the following benefits: Interconnection between data centers , Separate secure domains in data centers , Flexible network configurations and planning within the VPC. For more information about elastic network, see [Virtual Private Cloud](#).

Elasticity of Alibaba Cloud is a combination of elastic computing, storage, network, and the elasticity to redesign business architecture. By using Alibaba Cloud, you can work out your business portfolio in whatever way you want.

Comparison between ECS and traditional IDCs

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

	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	Provides equipment rooms with various quality levels and

	ECS	Traditional IDCs
	bandwidth and dedicated bandwidth.	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.
Ease of operation	Provides built-in mainstream operating systems, including activated Windows OS.	Purchases and installs operating system manually.
	Switches operating systems online.	Reinstalls operating systems manually.
	Provides a Web-based console for online management.	Manages and maintains 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	Users can customize automatic snapshot policies to create automatic snapshots for data recovery.	Users must restore all corrupted data manually.
	User-Defined snapshots	Faults cannot be recovered automatically.
	Faults can be recovered fast and automatically.	Failed to prevent MAC spoofing and ARP attacks.
Security and reliability	Effectively prevents MAC spoofing and ARP attacks.	Failed to prevent MAC spoofing and ARP attacks.
	Effectively defend 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 long time for server delivery.

	ECS	Traditional IDCs
	Adjusts outbound bandwidth whenever required.	One-off purchase of outbound bandwidth, unable to adjust.
	Combines with Server Load Balancer online, enabling scaling up applications quickly and easily.	Uses hardware-based server load balancing, which is expensive and extremely difficult to set up.
Cost effectiveness	Costs low.	Costs high.
	Invests a little up front.	Invests a lot up front, causing serious 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

You can learn more about the available ECS instance type families, including the features, available instance types, and applicable scenarios.

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

ECS instances are categorized into multiple specification types, which are called type families, based on the business and scenarios. You may select various type families for one business scenario. Each type family contains multiple instance types with different CPU and memory specifications. *ECS instance type* includes the basic specifications of instances, including CPU model and clock speed. However, the attributes of a block storage, an image, and the network service of an ECS instance must also be defined simultaneously for the specific service type of the instance to be determined.

**Note:**

The availability of instance type families and their types varies according to the regions and the amount of resources. Go to the [purchase page](#) to check the available instance types.

Alibaba Cloud ECS instances are categorized into the following type families according to business scenarios: enterprise-level instance type families and entry-level instance type families. Type families for enterprise-level computing feature stable performance and dedicated resources. For enterprise-level instances, each vCPU core is supported by one Intel Xeon CPU core through hyper-threading. For the difference, please refer to [Enterprise-level instances and entry-level instances FAQ](#).

**Note:**

If you are using sn1, sn2, t1, s1, s2, s3, m1, m2, c1, c2, n1, n2, or e3, see [Phased-out instance types](#).

Alibaba Cloud ECS instances are categorized into the following type families according to system structures and business scenarios:

- Type families for enterprise-level computing on the x86-architecture, including:
 - [g5, general-purpose type family](#)
 - [sn2ne, general-purpose type family with enhanced network performance](#)
 - [c5, compute instance type family](#)

- *sn1ne*, compute optimized type family with enhanced network performance
- *r5*, memory instance type family
- *se1ne*, memory optimized type family with enhanced network performance
- *se1*, memory optimized type family
- *d1*, big data type family
- *d1*, big data type family
- *i2*, 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
- *c4*, *cm4*, and *ce4*, compute optimized type family with high clock speed
- Type families for enterprise-level heterogeneous computing, including:
 - *gn5*, compute optimized type family with GPU
 - *gn4*, compute optimized type family with GPU
 - *f1*, compute optimized type family with FPGA
 - *f2*, compute optimized type family with FPGA
- ECS Bare Metal Instance type families and Super Computing Cluster (SCC) instance type families, including:
 - *ebmg5*, general-purpose ECS Bare Metal Instance type family
 - *ebmg4*, general-purpose ECS Bare Metal Instance type family (Coming soon)
 - *ebmhfg5*, ECS Bare Metal Instance type family with high clock speed
 - *ebmhfg4*, ECS Bare Metal Instance type family with high clock speed (Coming soon)
 - *ebmhfg4*, ECS Bare Metal Instance type family with high clock speed (Coming soon)
 - *sccg5*, general-purpose Super Computing Cluster (SCC) instance type family (Coming soon)
 - *scch5*, Super Computing Cluster (SCC) instance type family with high clock speed (Coming soon)
- Type families for entry-level users, computing on the x86-architecture, including:
 - *t5*, burstable instances
 - Type families of previous generations for entry-level users, *xn4/n4/mn4/e4*

g5, general-purpose type family

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU : Memory = 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 of receiving and transmitting a large volume of packets, such as the re-transmission of telecommunication services
 - 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.g5.large	2	8.0	N/A	1.0	30	2	2
ecs.g5.xlarge	4	16.0	N/A	1.5	50	2	3
ecs.g5.2xlarge	8	32.0	N/A	2.5	80	2	4
ecs.g5.4xlarge	16	64.0	N/A	5.0	100	4	8
ecs.g5.6xlarge	24	96.0	N/A	7.5	150	6	8

Instance type	vCPU	Memory (GiB)	Local disks (GiB) *	Bandwidth (Gbit/s)	Packet forwarding rate (Thousand pps) **	NIC queues ***	ENIs ****
ecs.g5.8xlarge	32	128.0	N/A	10.0	200	8	8
ecs.g5.16xlarge	64	256.0	N/A	20.0	400	16	8

[Back to Contents](#) 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 : Memory = 1:4
- Ultra high packet forwarding rate
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) or Platinum 8163 (skylake)
- Higher computing specifications matching higher network performance
- Ideal for:
 - Scenarios that require receiving and transmitting a large volume of packets, such as the re-transmission of telecommunication services
 - 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	30	2	2
ecs.sn2ne.xlarge	4	16.0	N/A	1.5	50	2	3
ecs.sn2ne.2xlarge	8	32.0	N/A	2.0	100	4	4
ecs.sn2ne.4xlarge	16	64.0	N/A	3.0	160	4	8
ecs.sn2ne.8xlarge	32	128.0	N/A	6.0	250	8	8
ecs.sn2ne.14xlarge	56	224.0	N/A	10.0	450	14	8

**Note:**

You can change the configurations of an sn2ne to any instance type in the sn2, sn2ne, sn1, sn1ne, se1, and se1ne instance type family.

[Back to Contents](#) View other instance type families.

c5, compute instance type family

Features

- I/O-optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU : Memory = 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 that require receiving and transmitting a large volume of packets, such as the re-transmission of telecommunication services
- 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	30	2	2
ecs.c5.xlarge	4	8.0	N/A	1.5	50	2	3
ecs.c5.2xlarge	8	16.0	N/A	2.5	80	2	4
ecs.c5.4xlarge	16	32.0	N/A	5.0	100	4	8
ecs.c5.6xlarge	24	48.0	N/A	7.5	150	6	8
ecs.c5.8xlarge	32	64.0	N/A	10.0	200	8	8
ecs.c5.16xlarge	64	128.0	N/A	20.0	400	16	8

[Back to Contents](#) 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 : Memory = 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 that require receiving and transmitting a large volume of packets, such as the re-transmission of telecommunication services
 - 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	30	2	2
ecs.sn1ne.xlarge	4	8.0	N/A	1.5	50	2	3
ecs.sn1ne.2xlarge	8	16.0	N/A	2.0	100	4	4
ecs.sn1ne.4xlarge	16	32.0	N/A	3.0	160	4	8
ecs.sn1ne.8xlarge	32	64.0	N/A	6.0	250	8	8



Note:

You can change the configurations of an sn1ne instance to any instance type in the sn2, sn2ne, sn1, sn1ne, se1, and se1ne instance type family.

[Back to Contents](#) 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 that require receiving and transmitting a large volume of packets, such as re-transmission of telecommunication services
 - 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	30	2	2
ecs.r5.xlarge	4	32.0	N/A	1.5	50	2	3
ecs.r5.2xlarge	8	64.0	N/A	2.5	80	2	4
ecs.r5.4xlarge	16	128.0	N/A	5.0	100	4	8
ecs.r5.6xlarge	24	192.0	N/A	7.5	150	6	8
ecs.r5.8xlarge	32	256.0	N/A	10.0	200	8	8
ecs.r5.16xlarge	64	512.0	N/A	20.0	400	16	8

[Back to Contents](#) 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 : Memory = 1:8
- 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 that require receiving and transmitting a large volume of packets, such as the re-transmission of telecommunication services
 - High-performance databases, memory-based databases
 - Data analysis and mining, and distributed memory cache
 - Hadoop, Spark, and other enterprise-level applications with large memory requirements

Instance type

Instance type	vCPU	Memory (GB)	Local disks (GB) *	Bandwidth (Gbit/s)	Packet forwarding rate (Thousand pps) **	NIC queues ***	ENIs ****
ecs.se1ne.large	2	16.0	N/A	1.0	30	2	2
ecs.se1ne.xlarge	4	32.0	N/A	1.5	50	2	3
ecs.se1ne.2xlarge	8	64.0	N/A	2.0	100	4	4
ecs.se1ne.4xlarge	16	128.0	N/A	3.0	160	4	8
ecs.se1ne.8xlarge	32	256.0	N/A	6.0	250	8	8

Instance type	vCPU	Memory (GB)	Local disks (GB) *	Bandwidth (Gbit/s)	Packet forwarding rate (Thousand pps) **	NIC queues ***	ENIs ****
ecs.se1ne.14xlarge	56	480.0	N/A	10.0	450	14	8

**Note:**

You can change the configurations of an se1ne instance to any instance type in the sn2, sn2ne, sn1, sn1ne, se1, and se1ne instance type family.

[Back to Contents](#) View other instance type families.

se1, memory optimized type family**Features**

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

Instance type

Instance type	vCPU	Memory (GB)	Local disks (GB) *	Bandwidth (Gbit/s)	Packet forwarding rate (Thousand pps) **	NIC queues ***	ENIs ****
ecs.se1.large	2	16.0	N/A	0.5	10	1	2

Instance type	vCPU	Memory (GB)	Local disks (GB) *	Bandwidth (Gbit/s)	Packet forwarding rate (Thousand pps) **	NIC queues ***	ENIs ****
ECS.se1.xlarge	4	32.0	N/A	0.8	20	1	3
ecs.se1.2xlarge	8	64.0	N/A	1.5	40	1	4
ecs.se1.4xlarge	16	128.0	N/A	3.0	50	2	8
ecs.se1.8xlarge	32	256.0	N/A	6.0	80	3	8
ecs.se1.14xlarge	56	480.0	N/A	10.0	120	4	8

**Note:**

You can change the configurations of an se1ne instance to any instance type in the sn2, sn2ne, sn1, sn1ne, se1, and se1ne instance type family.

[Back to Contents](#) 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 : Memory = 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 in the Internet and finance industries, to store and compute massive data
- Elasticsearch, logs, and so on

Instance type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Thousands pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs.d1ne.2xlarge	8	32.0	4*5500	6.0	100	4	4
ecs.d1ne.4xlarge	16	64.0	8*5500	12.0	160	4	8
ecs.d1ne.6xlarge	24	96.0	12*5500	16.0	200	6	8
ecs.d1ne.8xlarge	32	128.0	16*5500	20.0	250	8	8
ecs.d1ne.14xlarge	56	224.0	28*5500	35.0	450	14	8



Note:

- You cannot change configurations of d1ne instances.
- For more information of d1ne type families, see [FAQ on d1 and d1ne](#).

[Back to Contents](#) 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 : Memory = 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 in the Internet and finance industries, to store and compute massive data
 - Elasticsearch, logs, and so on

Instance type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs.d1.2xlarge	8	32.0	4*5500	3.0	30	1	4
ecs.d1.4xlarge	16	64.0	8*5500	6.0	60	2	8
ecs.d1.6xlarge	24	96.0	12*5500	8.0	80	2	8
ecs.d1-c8d3.8xlarge	32	128.0	12*5500	10.0	100	4	8
ecs.d1.8xlarge	32	128.0	16*5500	10.0	100	4	8
ecs.d1-c14d3.14xlarge	56	160.0	12*5500	17.0	180	6	8
ecs.d1.14xlarge	56	224.0	28*5500	17.0	180	6	8



Note:

- You cannot change configurations of d1ne instances.

- For more information of d1ne type families, see [FAQ on d1 and d1ne](#) .

[Back to Contents](#) 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 : Memory = 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 type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) **	NIC queues ***	ENIs ****
ecs.i2.xlarge	4	32.0	1*894	1.0	50	2	3
ecs.i2.2xlarge	8	64.0	1*1788	2.0	100	2	4
ecs.i2.4xlarge	16	128.0	2*1788	3.0	150	4	8
ecs.i2.8xlarge	32	256.0	4*1788	6.0	200	8	8
ecs.i2.16xlarge	64	512.0	8*1788	10.0	400	16	8

**Note:**

You cannot change configurations of i2 instances.

[Back to Contents](#) 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 : Memory = 1:4, designed for big data scenarios
- 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 type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) **	NIC queues ***	ENIs ****
ecs.i1.xlarge	4	16.0	2 * 104	0.8	20	1	3
ecs.i1.2xlarge	8	32.0	2 * 208	1.5	40	1	4
ecs.i1.4xlarge	16	64.0	2*416	3.0	50	2	8
ecs.i1-c5d1.4xlarge	16	64.0	2 * 1456	3.0	40	2	8

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs.i1.8xlarge	32	128.0	2 * 832	6.0	80	3	8
ecs.i1-c10d1.8xlarge	32	128.0	2 * 1456	6.0	80	3	8
ecs.i1.14xlarge	56	224.0	2 * 1456	10.0	120	4	8

**Note:**

You cannot change configurations of i1 instances.

[Back to Contents](#) 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
- Steady computing performance
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors
- vCPU : Memory = 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 type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs.hfc5.large	2	4.0	N/A	1.0	30	2	2
ecs.hfc5.xlarge	4	8.0	N/A	1.5	50	2	3
ecs.hfc5.2xlarge	8	16.0	None	N/A	100	2	4
ecs.hfc5.4xlarge	16	32.0	N/A	3.0	160	4	8
ecs.hfc5.6xlarge	24	48.0	N/A	4.5	200	6	8
ecs.hfc5.8xlarge	32	64.0	N/A	6.0	250	8	8

**Note:**

You can change the configurations of an hfg5 instance to any instance type in the hfc5 and hfg5 instance type families.

[Back to Contents](#) 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
- Steady computing performance
- 3.1 GHz Intel Xeon Gold 6149 (Skylake) processors
- vCPU : Memory = 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 type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs.hfg5.large	2	8.0	N/A	1.0	30	2	2
ecs.hfg5.xlarge	4	16.0	N/A	1.5	50	2	3
ecs.hfg5.2xlarge	8	32.0	N/A	2.0	100	2	4
ecs.hfg5.4xlarge	16	64.0	N/A	3.0	160	4	8
ecs.hfg5.6xlarge	24	96.0	N/A	4.5	200	6	8
ecs.hfg5.8xlarge	32	128.0	N/A	6.0	250	8	8
ecs.hfg5.14xlarge	56	160.0	N/A	10.0	400	14	8



Note:

You can change the configurations of an hfg5 instance to any instance type in the hfc5 and hfg5 instance type families.

[Back to Contents](#) View other instance type families.

c4, cm4, and ce4, compute optimized type family with high clock speed

Features

- — I/O optimized

- Supports SSD Cloud Disks and Ultra Cloud Disks
- Steady computing performance
- 3.2 GHz Intel Xeon E5-2667 v4 (Broadwell) processors
- 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 type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs.c4.xlarge	4	8.0	N/A	1.5	20	1	3
ecs.c4.2xlarge	8	16.0	N/A	3.0	40	1	4
ECS.	16	32.0	N/A	6.0	80	2	8
ecs.cm4.xlarge	4	16.0	N/A	1.5	20	1	3
ecs.cm4.2xlarge	8	32.0	N/A	3.0	40	1	4
ecs.cm4.4xlarge	16	64.0	N/A	6.0	80	2	8
ecs.cm4.6xlarge	24	96.0	N/A	10.0	120	4	8
ecs.ce4.xlarge	4	32.0	N/A	1.5	20	1	3

[Back to Contents](#) 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 type

Instance type	vCPU	Memory (GB)	Local disks (GB) *	GPU	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) **	NIC queues ***	ENIs ****
ecs.gn5-c4g1.xlarge	4	30.0	440	1 * NVIDIA P100	3.0	30	1	3
ecs.gn5-c8g1.2xlarge	8	60.0	440	1 * NVIDIA P100	3.0	40	1	4
ecs.gn5-c4g1.2xlarge	8	60.0	880	2 * NVIDIA P100	5.0	100	2	4

Instance type	vCPU	Memory (GB)	Local disks (GB) *	GPU	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousands pps) **	NIC queues ***	ENIs ****
ecs.gn5-c8g1.4xlarge	16	120.0	880	2 * NVIDIA P100	5.0	100	4	8
ecs.gn5-c28g1.7xlarge	28	112.0	440	1 * NVIDIA P100	5.0	100	8	8
ecs.gn5-c8g1.8xlarge	32	240.0	1760	4 * NVIDIA P100	10.0	200	8	8
ecs.gn5-c28g1.14xlarge	56	224.0	880	2 * NVIDIA P100	10.0	200	14	8
ecs.gn5-c8g1.14xlarge	54.	480.0	3520	NVIDIA P100 8 *	25.0	400	14	8

**Note:**

- See [#unique_24](#).
- You cannot change configurations of gn5 instances.

[Back to Contents](#) 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
- Use Scenarios
 - 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 type

Instance type	vCPU	Memory (GB)	Local disks (GB)	GPU	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousands pps)	NIC queues	ENIs ****
ecs.gn4-c4g1.xlarge	4	30.0	N/A	1 * NVIDIA M40	3.0	30	1	3
ecs.gn4-c8g1.2xlarge	8	60.0	N/A	1 * NVIDIA M40	3.0	40	1	4
ecs.gn4.8xlarge	32	48.0	N/A	1 * NVIDIA M40	6.0	80	3	8
ecs.gn4-c4g1.2xlarge	8	60.0	N/A	2 * NVIDIA M40	5.0	50	1	4
ecs.gn4-c8g1.4xlarge	16	60.0	N/A	2 * NVIDIA M40	5.0	50	1	8
ecs.gn4.14xlarge	56	96.0	N/A	2 * NVIDIA M40	10.0	120	4	8

**Note:**

- See [#unique_24](#).
- You can change the configurations of a gn4 instance within the gn4 family.

[Back to Contents](#) 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 : Memory = 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
 - Finance analysis
 - Picture transcoding
 - Computational workloads, such as real-time video processing and security

Instance type

Instance type	vCPU	Memory (GB)	Local disks (GB)	FPGA	Network bandwidth capability (Out/into) (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	40	4	4

Instance type	vCPU	Memory (GB)	Local disks (GB)	FPGA	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps)	NIC queues	ENIs ****
ecs.f1-c28f1.7xlarge	28	112.0	N/A	Intel ARRIA 10 GX 1150	5.0	200	8	8

**Note:**

You cannot change configurations of f1 instances.

[Back to Contents](#) 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 : Memory = 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
 - Finance analysis
 - Picture transcoding
 - Computational workloads, such as real-time video processing and security

Instance type

Instance type	vCPU	Memory (GB)	Local disks (GB)	FPGA	Network bandwidth capability (Out/into) (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	80	4	4
ecs.f2-c8f1.4xlarge	16	120.0	N/A	2 * Xilinx Kintex UltraScale XCKU115	5.0	100	4	8
ecs.f2-c28f1.7xlarge	28	112.0	N/A	Xilinx Kintex UltraScale XCKU115	5.0	100	8	8
ecs.f2-c28f1.14xlarge	56	224.0	N/A	2 * Xilinx Kintex UltraScale XCKU115	10.0	200	14	8

**Note:**

You cannot change configurations of f2 instances.

[Back to Contents](#) 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 : Memory = 1:4
- 2.5 GHz Intel Xeon Platinum 8163 (Skylake) processors, 96-core vCPU, up to 2.9 GHz Turbo Boost
- High network performance: 4.5 million pps packet forwarding rate
- Supports VPC networks
- Ideal for:
 - Deployment of OpenStack, ZStack, and other private cloud services
 - Deployment of Docker containers and other services
 - Scenarios that require receiving and transmitting a large volume of packets, such as re-transmission of telecommunication services
 - Enterprise-level applications of various types and sizes
 - Medium and large database systems, caches, and search clusters
 - Data analysis and computing
 - Computing clusters and data processing depending on memory

Instance type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps)	NIC queues	ENIs
ecs.ebmg5.24xlarge	96	384.0	N/A	10.0	450	8	32



Note:

For more information about ECS Bare Metal Instance, see [ECS Bare Metal Instance and Super Computing Clusters](#).

[Back to Contents](#) View other instance type families.

ebmg4, general-purpose ECS Bare Metal Instance type family (Coming soon)

Features

- I/O optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU : Memory = 1:4
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors, up to 2.9 GHz Turbo Boost
- High network performance: 4 million pps packet forwarding rate
- Supports VPC networks
- Ideal for:
 - Deployment of OpenStack, ZStack, and other private cloud services
 - Deployment of Docker containers and other services
 - Scenarios that require receiving and transmitting a large volume of packets, such as re-transmission of telecommunication services
 - Enterprise-level applications of various types and sizes
 - Medium and large database systems, caches, and search clusters
 - Data analysis and computing
 - Computing clusters and data processing depending on memory

Instance type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs.ebmg4.8xlarge	32	128.0	N/A	10.0	400	8	12



Note:

For more information about ECS Bare Metal Instance, see [ECS Bare Metal Instance and Super Computing Clusters](#).

[Back to Contents](#) 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 : Memory = 1:4
- 3.7 GHz Intel Xeon E3-1240v6 (Skylake) processors, 8-core vCPU, up to 4.1 GHz Turbo Boost
- High network performance: 2 million pps packet forwarding rate
- Supports VPC networks
- Ideal for:
 - Gaming or financial applications featuring low latency and high performance (Supports Intel SGX)
 - Scenarios that require receiving and transmitting a large volume of packets, such as re-transmission of telecommunication services
 - High performance databases and in-memory databases

Instance type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) **	NIC queues ***	ENIs ****
ecs.ebmhfg5.2xlarge	8	32.0	N/A	6.0	200	8	6



Note:

For more information about ECS Bare Metal Instance, see [ECS Bare Metal Instance and Super Computing Clusters](#).

[Back to Contents](#) View other instance type families.

ebmhfg4, ECS Bare Metal Instance type family with high clock speed (Coming soon)

Features

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

- vCPU : Memory = 1:4
- 3.2 GHz Intel Xeon E5-2667 v4 (Broadwell) processors, up to 3.5 GHz Turbo Boost
- High network performance: 4 million pps packet forwarding rate
- Supports VPC networks
- Ideal for:
 - Gaming or financial applications featuring low latency and high performance
 - Scenarios that require receiving and transmitting a large volume of packets, such as re-transmission of telecommunication services
 - High performance databases and in-memory databases
 - Data analysis and mining, and distributed memory cache

Instance type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) **	NIC queues ***	ENIs ****
ecs.ebmhfg4.4xlarge	16	64.0	N/A	10.0	400	8	12



Note:

For more information about ECS Bare Metal Instance, see [ECS Bare Metal Instance and Super Computing Clusters](#).

[Back to Contents](#) View other instance type families.

ebmhfg4, ECS Bare Metal Instance type family with high clock speed (Coming soon)

Features

- I/O optimized
- Supports SSD Cloud Disks and Ultra Cloud Disks
- vCPU : Memory = 1:2
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors, up to 2.9 GHz Turbo Boost

- High network performance: 4 million pps packet forwarding rate
- Supports VPC networks
- Ideal for:
 - Scenarios that require receiving and transmitting a large volume of packets, such as re-transmission of telecommunication services
 - Enterprise-level applications of various types and sizes
 - Medium and large database systems, caches, and search clusters
 - Data analysis and computing

Instance type

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs. ebmc4. 8xlarge	32	64.0	N/A	10.0	400	8	12



Note:

For more information about ECS Bare Metal Instance, see [ECS Bare Metal Instance and Super Computing Clusters](#).

[Back to Contents](#) View other instance type families.

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

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 : Memory = 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 type

Instance type	vCPU	Memory (GB)	GPU	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousands pps) **	RoCE (Inbound / Outbound) (Gbit/s)	NIC queues ***	ENIs ****
ecs.sccg5.24xlarge	96	384.0	N/A	10.0	450	46	12	1



Note:

For more information about SCC, see [ECS Bare Metal Instance and Super Computing Clusters](#).

[Back to Contents](#) View other instance type families.

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

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 : Memory = 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 type

Instance type	vCPU	Memory (GB)	GPU	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousands pps) **	RoCE (Inbound / Outbound) (Gbit/s)	NIC queues ***	ENIs ****
ecs.scch5.16xlarge	64	192.0	N/A	10.0	450	46	12	1



Note:

For more information about SCC, see [ECS Bare Metal Instance and Super Computing Clusters](#).

[Back to Contents](#) View other instance type families.

t5, burstable instances

Features

- 2.5 GHz Intel Xeon processors
- The latest DDR4 memory
- No fixed ratio of CPU to memory
- A vcpu that suddenly increases speed, maintains basic performance, and is limited by vcpu points.
- Balance computing, memory, and network resources
- Supports VPC networks
- Ideal for:
 - Front ends of Web applications
 - Light load applications and microservices
 - Applications for development or testing environments

Instance type

Type family	vCPU	Memory (GB)	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.	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	1728	15%	2
ecs.t5-c1m2.2xlarge	8	16.0	72	1728	15%	2
ecs.t5-c1m4.2xlarge	8	32.0	72	1728	15%	2
ecs.t5-c1m1.4xlarge	16	16.0	144	1728	15%	2

Type family	vCPU	Memory (GB)	CPU credits/hour	Max CPU credit balance	Avg baseline CPU performance	ENIs ****
ecs.t5-c1m2.4xlarge	16	32.0	144	1728	15%	2

**Note:**

For more information about t5 instances, see [Burstable instances](#).

[Back to Contents](#) View other instance type families.

Type families of previous generations for entry-level users, xn4/n4/mn4/e4

Features

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

Instance type families	Features	vCPU : Memory	Ideal for:
xn4	Shared Basic Instance	1:1	<ul style="list-style-type: none"> • Front ends of Web applications • Light load applications and microservices • Applications for development or testing environments
n4	Shared Compute Instance	1:2	<ul style="list-style-type: none"> • Websites and Web applications • Development environment, building servers, code repositories, microservices, and

Instance type families	Features	vCPU : Memory	Ideal for:
			testing and staging environment <ul style="list-style-type: none"> Lightweight enterprise applications
mn4	Shared Standard Instance	1:4	<ul style="list-style-type: none"> Websites and Web applications Lightweight databases and cache Integrated applications and lightweight enterprise services
e4	Shared Memory Instance	1:8	<ul style="list-style-type: none"> Applications that require large volume of memory Lightweight databases and cache

**Note:**

You can change the configurations of an instance between any two type families of xn4, n4, mn4, and e4, and within the same instance type family.

xn4

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) **	NIC queues ***	ENIs ****
ecs.xn4.small	1	1.0	N/A	0.5	5	1	1

n4

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs.n4.small	1	2.0	N/A	0.5	5	1	1
ecs.n4.large	2	4.0	N/A	0.5	10	1	1
ecs.n4.xlarge	4	8.0	N/A	0.8	15	1	2
ecs.n4.2xlarge	8	16.0	N/A	1.2	30	1	2
ecs.n4.4xlarge	16	32.0	N/A	2.5	40	1	2
ecs.n4.8xlarge	32	64.0	N/A	5.0	50	1	2

mn4

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs.mn4.small	1	4.0	N/A	0.5	5	1	1
ecs.mn4.large	2	8.0	N/A	0.5	10	1	1
ecs.mn4.xlarge	4	16.0	N/A	0.8	15	1	2
ecs.mn4.2xlarge	8	32.0	N/A	1.2	30	1	2

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs.mn4.4xlarge	16	64.0	N/A	2.5	40	1	2

e4

Instance type	vCPU	Memory (GB)	Local disks (GB)	Network bandwidth capability (Out/into) (Gbit/s)	Packet forwarding rate (Thousand pps) ^{**}	NIC queues ^{***}	ENIs ^{****}
ecs.e4.small	1	8.0	N/A	0.5	5	1	1

[Back to Contents](#) 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. In some cases, such as when the computing resources of an instance, including CPU and memory, are released, or an instance is in the downtime migration, data on the local disks is erased. For more information, see [Local disks](#).

** The maximum packet forwarding rate of inbound or 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 2 or more vCPU cores supports elastic network interfaces. An entry-level instance with 4 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.

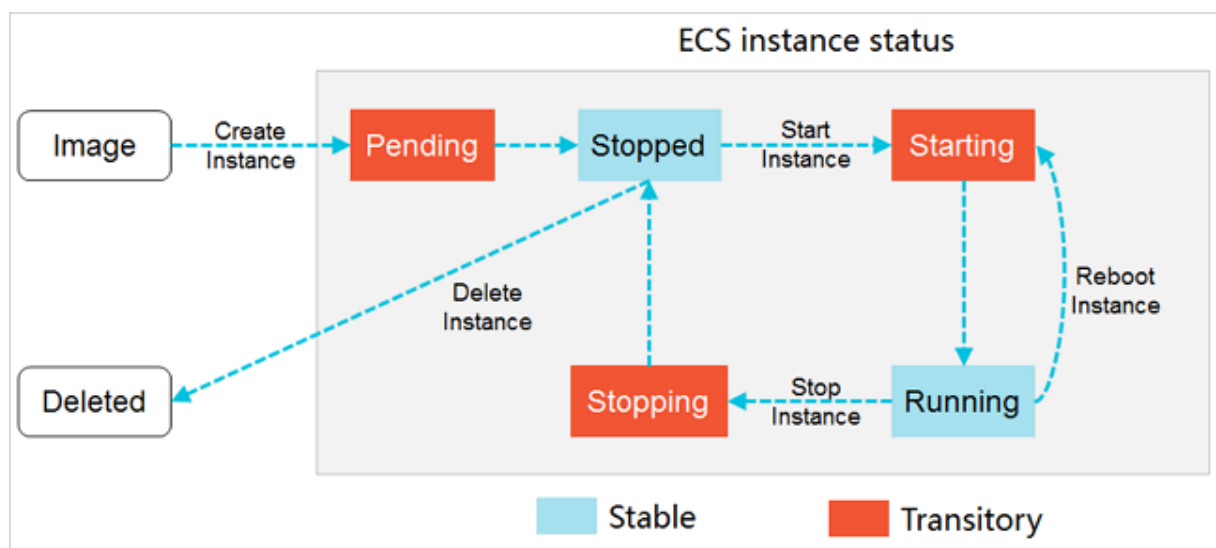
Status	Status attribute	Description	Corresponding API status	Viewable in the console
Preparing	Intermediate	After an instance is created, it remains in this status before running. If an instance is in this status for a long time, an exception occurs.	Pending	No
Starting	Intermediate	An instance is in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

Status	Status attribute	Description	Corresponding API status	Viewable in the console
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 you have an overdue payment	Stopped	Yes

Status	Status attribute	Description	Corresponding API status	Viewable in the console
		. 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 <i>renewed</i> , 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, <i>open a ticket</i> .	Stopped	Yes
Release pending	Stable	A Subscription instance is in this status after you apply for a refund before it expires.	Stopped	Yes

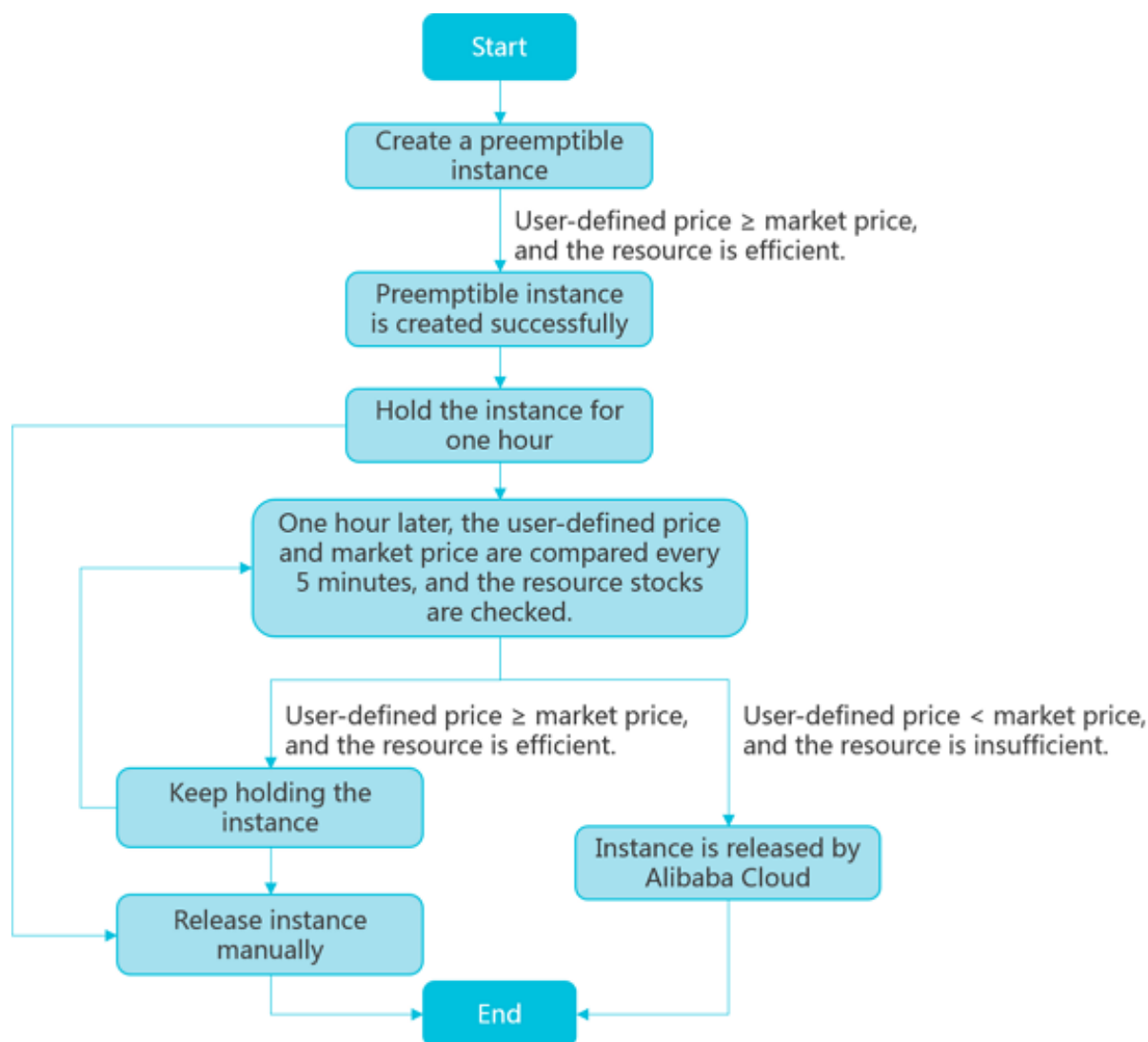
API status changes

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



5.3 Spot instances

Spot instances are a type of on-demand instances. They are designed to reduce your ECS costs in some cases. When you create a spot instance, you can set a maximum rate per hour for bidding a specified instance type. When your bid is higher than or equal to the current market price, your instance runs. You can hold a spot instance without interruption for at least one hour. 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 spot instance.



Scenarios

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

You can deploy the following businesses on spot instances:

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

- Image and media coding
- Testing

Spot 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 some other reasons.

Bidding modes

The spot instance supports a one-time bid request. You can bid for a spot 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 using [RunInstances](#) API to create a spot instance, you can bid in this mode.

Currently, the maximum bid of a spot instance is the price of a Pay-As-You-Go instance of the same configuration. When creating a spot 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, [Guaranteed duration](#) you can hold the instance even after one hour. Otherwise, your instance gets automatically released at any time.

SpotAsPriceGo

By setting `SpotStrategy` to `SpotAsPriceGo`, you can create a spot instance with the `SpotAsPriceGo` bidding mode, 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

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

Price and billing

Spot instance price and billing considerations:

- **Price**

The spot 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**

Spot 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 use by using API or on the ECS console, and the instance continues to charge.

- **Market price**

During creation of a spot 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 spot instance fluctuates according to the changes in the demand and supply of a given instance type. Therefore, you can take full advantage of the price fluctuations of spot instances. If you purchase spot instance types at the right time, the computing costs are reduced but your business throughput for this period is increased.

Quota

For more information about the spot instance quota, see #####.

Create a spot instance

You can purchase a spot instance by using the [RunInstances](#) interface.

After a spot instance is created, it can be used exactly 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 spot instance

You can stop a spot instance in the [ECS console](#) or by using the [StopInstance](#) interface. The VPC-Connected spot instances support the ##### feature.

he network type and the bidding mode of a spot 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: <ul style="list-style-type: none"> 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.
Classic + SpotWithPriceLimit	N/A	
VPC + SpotAsPriceGo	Keep Instance, Fees Apply	During the guaranteed duration, the instance can be started successfully. After the guaranteed duration: <ul style="list-style-type: none"> 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: <ul style="list-style-type: none"> If your bid is not lower than the market price and the resource stock is sufficient, the instance can be started successfully.

Network type + Bidding mode	Stop instance	Start instance
		<ul style="list-style-type: none"> If your bid is lower than the market price or the resource stock is insufficient, the instance cannot be started successfully.
VPC + SpotAsPriceGo	Stop Instance, No Fees	<p>During the guaranteed duration, the instance can be started successfully only if the resource stock is sufficient. After the guaranteed duration:</p> <ul style="list-style-type: none"> 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 spot instance

When the guaranteed period ends, we automatically release your spot instance because of changes in the market price or short resource stock. Additionally, you can independently [release the instance](#).

When a spot instance is released because of market price or changes in the demand and supply of resources, the instance enters the **Pending Release** status. Then, the instance is released in about five minutes. You can use ##### 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 spot 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 spot instance is immediately released. When you release the instance manually, you can test whether or not your application functions normally if a spot instance is immediately recovered.

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

Best practices

When using a spot instance, consider the following:

- Set a proper 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 spot 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 `#####` to run commands upon instance startup.
- It is recommended that you use storage media that is not affected by the spot instance release to save your important data. Store your business data on storage products that are independent from spot 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 spot instance. You can use `#####` 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 spot instances](#)

For more information about using APIs to create spot instances, see [Using APIs to manage spot 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 both elasticity of virtual machines and performance and characteristics of physical machines. As a product completely and independently developed by Alibaba Cloud, EBM Instances are based on the next-generation virtualization technology. Compared with the previous generation of virtualization technology, the next-generation virtualization technology with an innovative approach, not only supports the common virtual cloud server but also completely supports the nested virtualization

technology. It retains the resource elasticity of common cloud servers and adopts nested virtualization technology such that it keeps the user experience of physical machines intact.

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

Advantages

EBM Instances

EBM Instances realize the value of customers by way of technological innovation. Specifically, EBM Instances have the following advantages:

- **Exclusive computing resources**

As a cloud-based elastic computing product, the EBM Instances outshine the performance and isolation of contemporary physical machine and enables exclusive computing resources without virtualization performance overheads and feature loss. EBM Instances support 8, 16, 32, and 96 CPU cores and ultrahigh frequency. Considering an EBM Instance with 8 cores as an example, it 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 the physical server isolation. This allows 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.

- **Any Stack on Alibaba Cloud**

An EBM Instance combines the performance strengths and complete features of physical machines and the ease-of-use and cost-effectiveness of cloud servers. It can effectively meet your demands for high-performance computing and help you build new hybrid clouds. Thanks to the flexibility, elasticity, and all the other strengths it inherits from both virtual and physical machines, it is powered with re-virtualization ability. As a result, offline private clouds can be seamlessly migrated to Alibaba Cloud without any issues regarding performance overhead 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 completely developed by Alibaba Cloud. It can zero-cost support ARM and other instruction set processors.

SCC

Alibaba Cloud also released Super Computing Clusters based on the EBM Instance to meet the demands for high performance computing, artificial intelligence, machine learning, scientific or engineering computing, data analysis, audio and video processing, and so on. In the cluster, nodes are connected by Remote Direct Memory Access (RDMA) networks featuring high bandwidth and low latency, guaranteeing the high parallel efficiency posed by applications that need high-performance computing. Meanwhile, the RoCE (RDMA over Convergent Ethernet) may rival an Infiniband network in terms of connection speed, and supports more extensive Ethernet-based applications. The combination of the SCC built on the EBM Instance and other Alibaba Cloud computing products such as the ECS and GPU servers provides the with ultimate high performance parallel computing resources, which makes 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 provide support for high clock speed.
- Memory specifications:
 - EBM Instances: Supports 32 GiB to 768 GiB memory. To provide 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: To deliver instances in minutes, supports starting from the virtual machine image and cloud disk.
- 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. Here, Y indicates “Support”, N indicates “Not Support”, and N/A indicates no data available or not applicable.

Features	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

Features	Features	EBM Instances/ SCC	Physical servers	Virtual servers
	Requires no installation of operating system	Y	N	Y
	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 family

The type families of EBM Instances include:

- General purpose EBM Instance type families, including ebmg5 and ebmg4
- High frequency EBM Instance type families, including ebmhfg5 and ebmhfg4
- Compute EBM Instance type families, including ebmc4

The type families of SCC include scch5 and sccg5.

For more information, see [Instance type families](#).

Billing methods

Currently, EBM Instances and SCC instances are billed on a monthly basis only. For more information about billing methods, see [#####](#).

Related operations

You can [#####](#) in the console [##### SCC ##](#).

For more information, see [FAQs about EBM Instances](#).

5.5 Burstable instances

Burstable instances (also called t5 instances) can handle sudden rise in requirements of CPU performance. Each t5 instance provides a baseline CPU performance. When your t5 instance is running, it accumulates and consumes CPU credits. The instance type determines the rate at which CPU credits are distributed. t5 instances seamlessly increase your CPU performance, without affecting the instance environment or applications.

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

How t5 instances work

Basic concepts

Before you use t5 instances, you must know the following concepts:

- **Baseline CPU performance**

The instance type of any t5 instance determines its baseline CPU performance, which means each vCPU core of an instance has a maximum usage for normal workloads. For example, when an ecs.t5-lc1m2.small instance is used for normal workloads, the maximum CPU usage is 10%.

- **CPU credits**

Each t5 instance obtains CPU credits at a fixed distribution rate, which is determined by the baseline CPU performance. A CPU credit is a measuring unit to calculate performance, which is determined by the number of vCPU core, CPU usage, and work time. For example:

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

If one vCPU core runs at 100% usage for one hour, it consumes 60 CPU credits.

- **Initial CPU credits**

Every time you create a t5 instance, 30 CPU credits are immediately allocated to the instance, which are called initial CPU credits. Instances are allocated with initial CPU credits only once and at the time of creation. When an instance begins to consume CPU credits, the initial CPU credits are used first. The initial CPU credits do not expire.

- **CPU credit distribution rate**

The CPU credit distribution rate is the number of CPU credits that a t5 instance obtains per minute. It is determined by the baseline CPU performance. You can use the following formula to determine the CPU credit distribution rate according to the baseline CPU performance.

```
CPU credit distribution rate = (60 CPU credits * Baseline CPU performance) / 60 minutes
```

Example: A t5 instance of the ecs.t5-1c1m2.small type provides a baseline CPU performance of 10%, so the CPU credit distribution rate is 0.1 CPU credits per minute, or six CPU credits per hour.

- **Expiration of CPU credits**

Once accumulated, the CPU credits are saved only for 24 hours. The credits are invalid after 24 hours. The initial CPU credits do not expire.

- **CPU credit consumption**

When a t5 instance is started, the instance consumes the CPU credits (first the initial CPU credits, then the accumulated CPU credits) to raise the CPU usage to meet your business

requirements. When you want to use one vCPU at a certain usage for one minute, the number of the consumed CPU credits can be calculated by using the following formula:

```
CPU credits consumed per minute = 1 CPU credit * Actual CPU usage
```

Example: A t5 instance of the ecs.t5-1c1m2.small type is used at 50% computing capability for 1 minute, it consumes 0.5 CPU credits.

- **CPU credit accumulation**

When the CPU usage of a t5 instance is lower than the baseline CPU performance, the instance accumulates CPU credits because the consumption speed is lower than the CPU credit distribution rate. Otherwise, the CPU credits are consumed. The accumulation speed is calculated by using the following formula:

```
CPU credit accumulation per minute = 1 CPU credit * (Baseline CPU  
performance - Actual CPU usage) - Expired CPU credits within the  
minute
```

When the distribution of CPU credits is larger than consumption, the CPU credits increase; otherwise, the CPU credits decrease.

You can view CPU accumulation and consumption on the ECS Management Console.

When the accumulated CPU credits are cleared, the actual CPU computing capability of the instance cannot be higher than the baseline CPU performance.

Example

Take a t5 instance of the ecs.t5-1c1m2.small type as an example to introduce how CPU credits are accumulated.

1. When the instance is created, 30 initial CPU credits are allocated to it. These are the total CPU credits it has before it is started. When it starts, it consumes CPU credits and is allocated with CPU credits at the 0.1 CPU credits per minute.
2. During the first minute, when it is initialized, if the actual CPU usage is 5%, the CPU credits change as follows: 0.05 initial CPU credits are consumed, 0.1 CPU credits are allocated, and no CPU credits have expired. Therefore, 0.05 CPU credits are accumulated during this one minute.
3. During the N minute after the instance starts, if the actual CPU usage is 50%, the initial CPU credits are out, and 0.1 CPU credits expire, the CPU credits change as follows: 0.5 CPU credits are consumed, and 0.1 CPU credits are allocated. Therefore, 0.5 CPU credits are consumed during this one minute, and no CPU credits are accumulated.

4. When the accumulated CPU credits are cleared, the maximum actual CPU usage is 10%.

CPU credits accumulation on a stopped t5 instance

After you stop a t5 instance in the [ECS console](#) or [StopInstance](#) by using the StopInstance interface, CPU credits changes vary according to the billing method and network type, as shown in the following table.

Network type	Instance billing method	CPU credits changes after stopping
Classic	Subscription or Pay-As-You-Go	When you start a stopped instance, CPU credits continue accumulating.
VPC	Subscription	
	Pay-As-You-Go with No fees for No fees for stopped instances (VPC-Connected) stopped instances (VPC-Connected) disabled	When you start a stopped instance, CPU credits continue accumulating.
	Pay-As-You-Go with No fees for No fees for stopped instances (VPC-Connected) stopped instances (VPC-Connected) enabled	

When you start a stopped instance, CPU credits continue accumulating.

If the instance runs out-of-service because of payment overdue or expiration, the CPU credits remain valid, but the CPU credit accumulation stops. After the instance is [Reactivate an instance](#) reactivated or [renewed](#), CPU credits start to accumulate automatically.

Instance types

t5 instances use Intel Xeon processors. The instance types are shown in the following table. In this table:

- **CPU credits/hour** is the total number of CPU credits allocated to all vCPU cores of a t5 instance per hour.
- **Average baseline CPU performance** is the average baseline CPU performance of each vCPU core for a t5 instance.

Instance types	vCPU	CPU credits/ hour	Avg baseline CPU performance	Memory (GB)
ecs.t5-lc2m1.nano	1	6	10%	0.5
ecs.t5-lc1m1.small	1	6	10%	1.0
ecs.t5-lc1m2.small	1	6	10%	2.0
ecs.t5-lc1m2.large	2	12	10%	4.0
ecs.t5-lc1m4.large	2	12	10%	8.0
ecs.t5-c1m1.large	2	18	15%	2.0
ecs.t5-c1m2.large	2	18	15%	4.0
ecs.t5-c1m4.large	2	18	15%	8.0
ECS.	4	36	15%	4.0
ecs.t5-c1m2.xlarge	4	36	15%	8.0
ecs.t5-c1m4.xlarge	4	36	15%	16.0
ecs.t5-c1m1.2xlarge	8	72	15%	8.0
ecs.t5-c1m2.2xlarge	8	72	15%	16.0
ecs.t5-c1m4.2xlarge	8	72	15%	32.0
ecs.t5-c1m1.4xlarge	16	144	15%	16.0
ecs.t5-c1m2.4xlarge	16	144	15%	32.0

Here, we use ecs.t5-c1m1.xlarge as an example to explain the t5 instance configuration:

- Each vCPU core has an average baseline computing performance of 15%. Therefore, the total baseline computing performance of a t5 instance of the ecs.t5-c1m1.xlarge type is 60%, which means:
 - If the instance only uses one vCPU core, this core has a baseline computing performance of 60%.
 - If the instance only uses two vCPU cores, each core is allocated with a baseline computing performance of 30%.
 - If the instance only uses three vCPU cores, each core is allocated with a baseline computing performance of 20%.
 - If the instance uses all four vCPU cores, each core is allocated with a baseline computing performance of 15%.
- One instance is allocated with 36 CPU credits per hour, which means that each vCPU core is allocated with nine CPU credits per hour.

Billing methods

t5 instances support the following billing methods: Pay-As-You-Go and Subscription. For more information on the billing methods, please see [#unique_56](#).

Create an instance

See [Create an ECS instance](#) to create a t5 instance. When creating a t5 instance, consider the following settings:

- Region: t5 instances are unavailable in the China North 3 (Zhangjiakou), Asia Pacific SE 3 (Kuala Lumpur), Asia Pacific NE 1 (Tokyo), US East 1 (Virginia), and Middle East 1 (Dubai) regions. For the detailed zones in other regions that supports t5 instances, see the ECS purchase page.
- Network type: Only VPC is supported.
- Image and instance type: The minimum t5 instance memory configuration of 512 MB only supports Linux. To create a Windows instance, the minimum memory is 1 GB. For more information on image selection, see [How to select a system image](#).

Manage t5 instances

View CPU usage

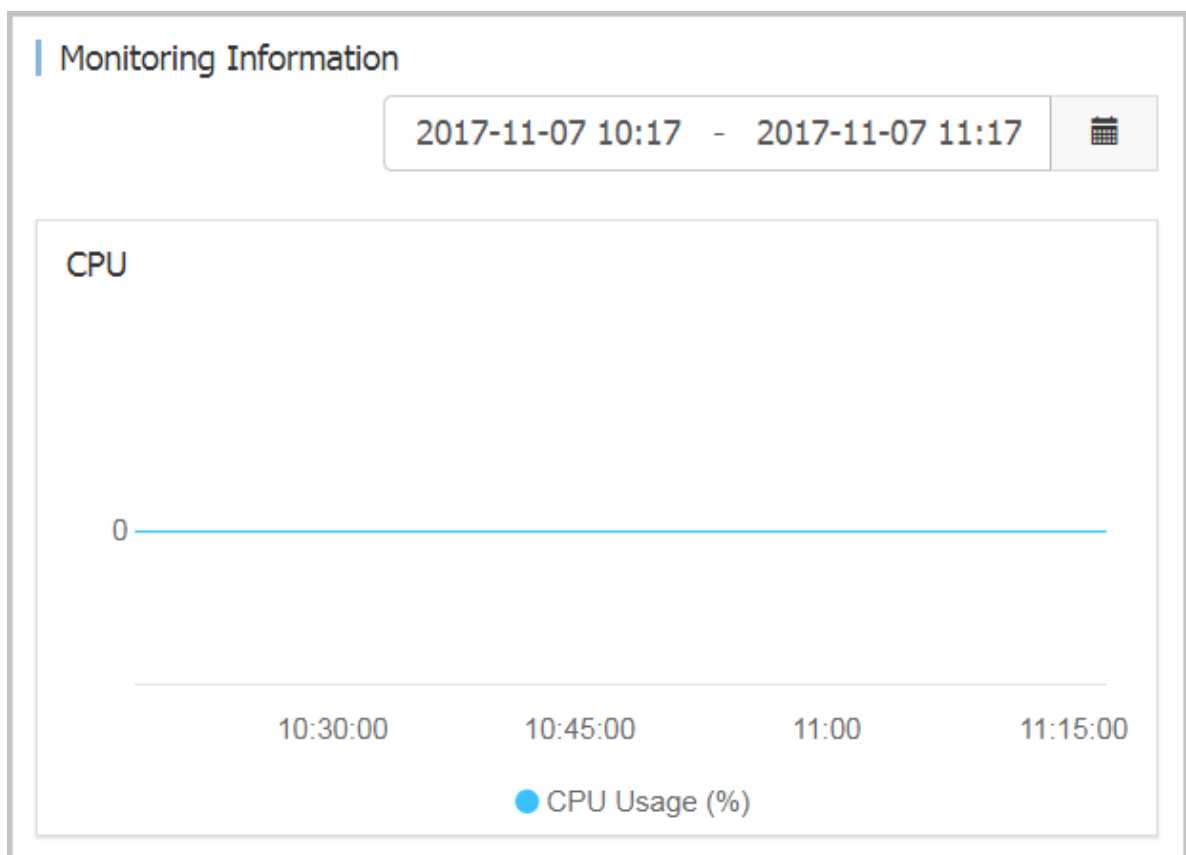
You can view CPU usage in any of the following ways:

- In the ECS console, go to the **Monitoring Information** section of the **Instance Details** page to view the instance CPU usage. You can also remotely connect to the ECS instance to view CPU usage.
- You can also remotely connect to the ECS instance to view CPU usage.

View CPU usage in the ECS console

To view CPU usage in the ECS console, follow these steps:

1. Log on to the [ECS console](#).
2. On the left-side navigation pane, click **Instances**.
3. Select a region.
4. Find a t5 instance, and click the instance ID or in the **Actions** column, click **Manage**.
5. In the **Monitoring Information** section, view CPU usage information.



Remotely connect to the instance to view CPU usage

The methods vary according to the operating system:

- Windows: [Connect to the instance](#) and view the information in the **Task Manager**.
- Linux: [Connect to the instance](#) and run the `top` command to view the CPU usage.

Change configurations

In the ECS console, if you see that the CPU usage is at the baseline level of CPU performance for an extended period or it never exceeds the baseline level, your current instance type is insufficient for your needs or exceeds your needs. In these cases, consider changing the instance type.

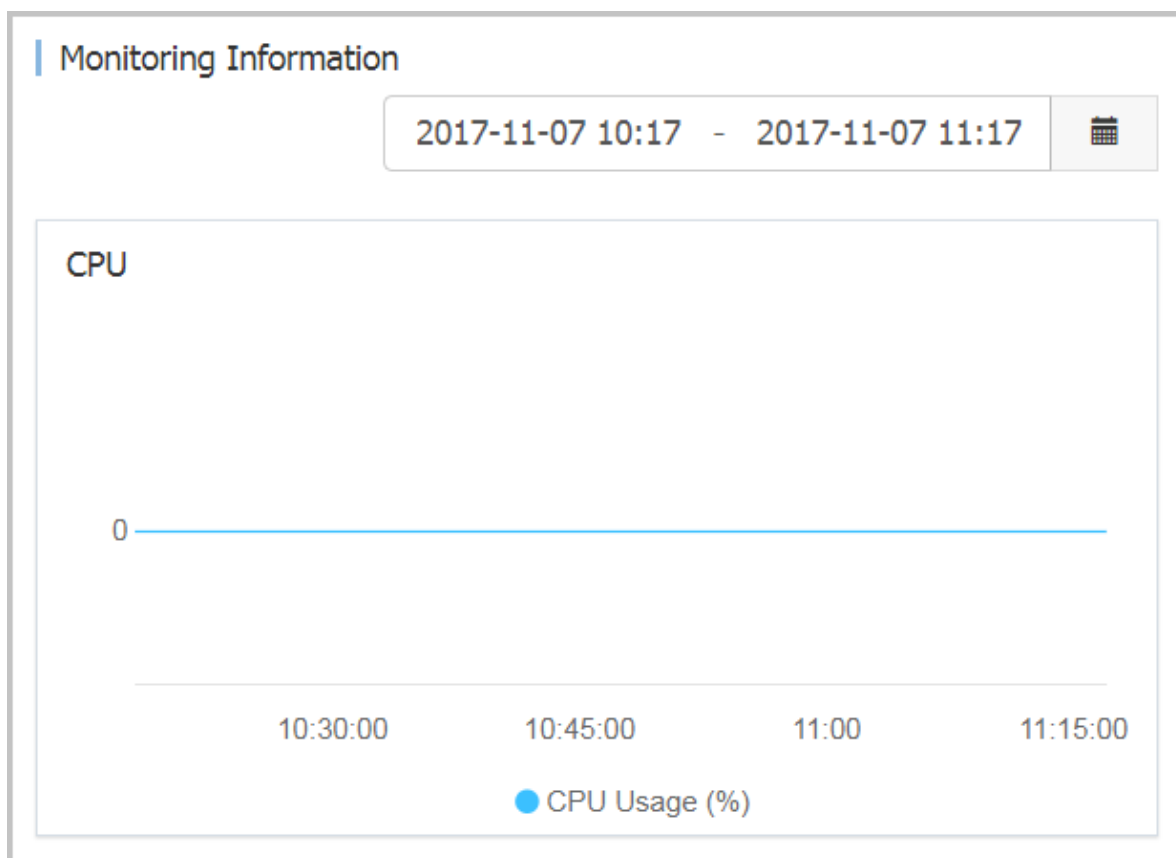
You can change the instance type based on the billing method:

- For a Subscription instance, you can [Change configurations](#) change the instance type. You can change the specification of an instance to any type in the t5 instance type family, any enterprise-level instance type families, or any type within the [xn4](#), [n4](#), [mn4](#), or [e4](#) type family.
- For a Pay-As-You-Go instance, you can change the instance type.

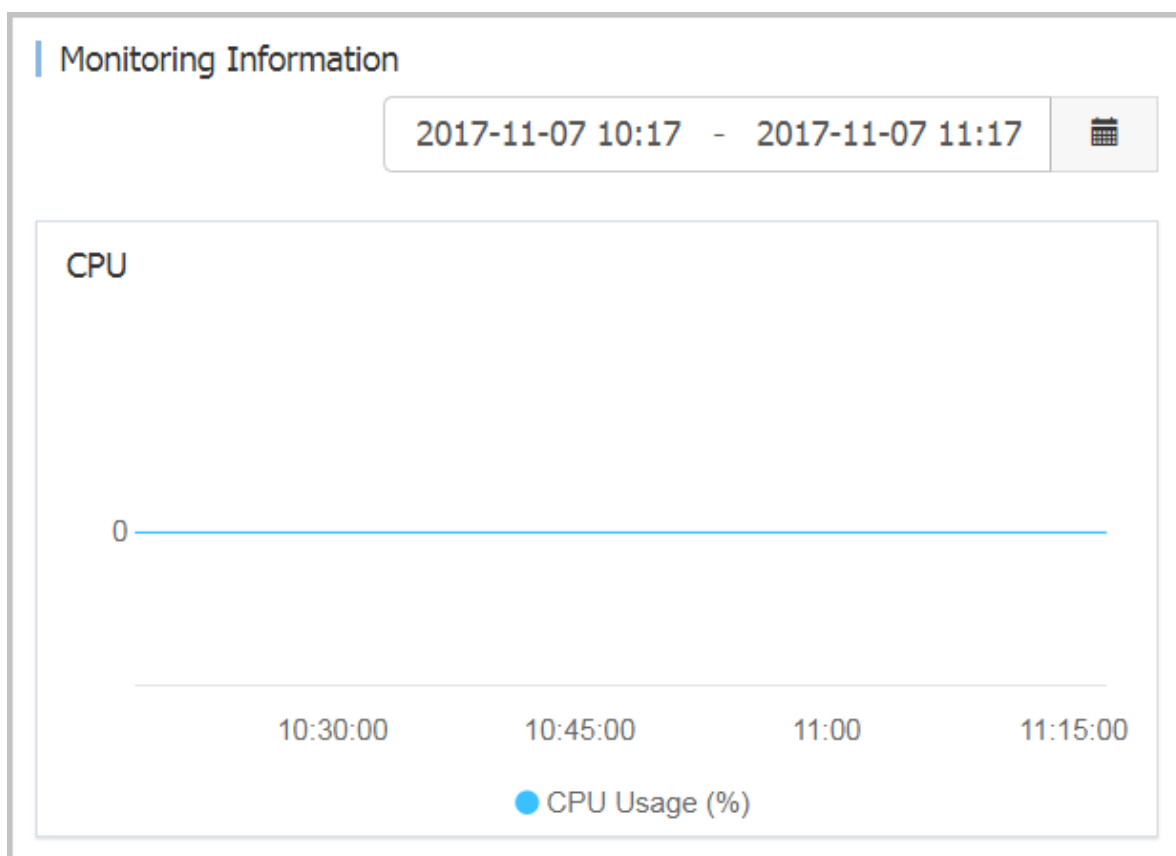
View CPU credits

Log on to the [ECS console](#) and go to the Instance Details page to view the accumulated and consumed CPU credits of a t5 instance.

- To view the accumulated CPU credits of a t5 instance:



- To view the consumed CPU credits of a t5 instance:



5.6 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 configurations. 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*](#)

6 Block storage

6.1 What is block storage?

Overview

Alibaba Cloud provides a wide variety of block-level storage products for ECS. These include elastic block storage based on the distributed storage architecture and local disks located on the physical servers where ECS instances are hosted. Specifically:

- [Elastic block storage](#), is a low-latency, persistent, and high-reliability random block-level data storage service provided by Alibaba Cloud to ECS users. It uses a [triplicate distributed system](#) to provide 99.9999999% data reliability for ECS instances. Elastic block storage can be created, released, and resized at any time.
- [Local disks](#) are the disks attached to the physical servers (host machines) that ECS instances are hosted on. They provide temporary block-level storage for instances, featuring low latency, high random IOPS, and high I/O throughput. They are designed for business scenarios requiring high storage I/O performance.

For more information about the performance of block-level storage products, see [Storage parameters and performance test](#).

Block storage, OSS, vs. NAS

Currently, Alibaba Cloud provides three types of data storage products, namely block storage, [Network Attached Storage \(NAS\)](#) and [Object Storage Service \(OSS\)](#).

The differences are as follows:

- Block storage: A high-performance and low-latency block-level storage device provided by Alibaba Cloud to ECS users. It supports random reads and writes. You can format the block storage and create a file system on it as you would with a hard disk. Block storage can be used for data storage in most common business scenarios.
- OSS: A huge storage space, which is suitable for storing massive unstructured data, including images, short videos, and audio files generated on the Internet. You can access the data stored in OSS anytime and anywhere by using APIs. Generally, OSS is used for business scenarios as Internet business website construction, separation of dynamic and static resources, and CDN acceleration.

- **NAS:** A storage space for massive unstructured data, like OSS. For NAS, you must access the data by using standard file access protocols, such as the Network File System (NFS) protocol for the Linux system, and the Common Internet File System (CIFS) protocol for the Windows system. You can set permissions to allow different clients to access the same file at the same time. NAS is applicable to business scenarios such as file sharing across departments, radio and television non-linear editing, high-performance computing, and Docker.

6.2 Storage parameters and performance test

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

Performance index of block storage

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

IOPS

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

The following table lists the most common performance characteristics measured: sequential operations and random operations.

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	
Sequential read IOPS	The average number of sequential read I/O operations per second	Sequential access to locations on storage devices contiguously
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 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 error reports in applications, we recommend that you use SSD Cloud disks, SSD Shared Block Storage, or local SSD disks.

For throughput-sensitive applications that are not sensitive to latency, such as I/O intensive applications, 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.

Cloud disks

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

Parameters	SSD Cloud Disks	Ultra Cloud Disks	Basic Cloud Disks
Capacity of a single disk	32,768 GiB	32,768 GiB	2,000 GiB
Maximum IOPS	20,000*	3,000	Several hundreds
Maximum throughput	300 MBps*	80 MBps	30 MBps–40 MBps
Formulas to calculate performance of a single disk**	$IOPS = \min\{1200 + 30 * capacity, 20000\}$	$IOPS = \min\{1000 + 6 * capacity, 3000\}$	N/A
	Throughput = $\min\{80 + 0.5 * capacity, 300\}$ MBps	Throughput = $\min\{50 + 0.1 * capacity, 80\}$ MBps	N/A
Data reliability	99.9999999%	99.9999999%	99.9999999%
API name	cloud_ssd	cloud_efficiency	cloud
Typical scenarios	<ul style="list-style-type: none"> Small or medium-sized relational databases, such as MySQL, SQL Server, 	<ul style="list-style-type: none"> Small or medium-sized relational databases, such as MySQL, 	<ul style="list-style-type: none"> Applications with infrequent access or low I/O operations

Parameters	SSD Cloud Disks	Ultra Cloud Disks	Basic Cloud Disks
	PostgreSQL, or Oracle <ul style="list-style-type: none"> Small or medium-sized development or testing applications that require high data reliability 	SQL Server, or PostgreSQL <ul style="list-style-type: none"> Small or medium-sized development or testing applications that require high data reliability and medium performance 	<ul style="list-style-type: none"> Applications that require low costs and random read and write I/O operations

The performance of an SSD Cloud Disk varies according to the size of the data blocks. 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.

Data block size	Maximum IOPS	Throughput
4 KiB or 8 KiB	About 20,000	Small, far smaller than 300 MBps
16 KiB	About 17,200	Close to 300 MBps
32 KiB	About 9,600	
64 KiB	About 4,800	

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

- The maximum IOPS: The baseline is 1,200 IOPS. It can increase by 30 IOPS per GiB of storage. The maximum IOPS is 20,000.
- The maximum throughput: The baseline is 80 MBps. It can increase by 0.5 MBps per GiB of storage. The maximum throughput is 300 MBps.

The latency varies according to the disk categories as follows:

- SSD Cloud 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.

Parameters	SSD Shared Block Storage	Ultra Shared Block Storage
Capacity	<ul style="list-style-type: none"> Single disk: 32,768 GiB All disks attached to one instance: Up to 128 TiB 	<ul style="list-style-type: none"> Single disk: 32,768 GiB All disks attached to one instance: Up to 128 TiB
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**	$IOPS = \min\{1600 + 40 * \text{capacity}, 30000\}$	$IOPS = \min\{1000 + 6 * \text{capacity}, 5000\}$
	$\text{Throughput} = \min\{100 + 0.5 * \text{capacity}, 512\} \text{ MBps}$	$\text{Throughput} = \min\{50 + 0.15 * \text{capacity}, 160\} \text{ MBps}$
Typical scenarios	<ul style="list-style-type: none"> Oracle RAC SQL Server Failover cluster High-availability of servers 	<ul style="list-style-type: none"> High-availability 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.

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

- The maximum IOPS: The baseline is 0 IOPS. It can increase by 40 IOPS per GiB of storage. The maximum IOPS is 30,000.
- The maximum throughput: The baseline is 50 MBps. It can increase by 0.5 MBps per GiB of storage. The maximum throughput is 512 MBps.

The latency varies according to 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

According to the OS on which an instance is running, you can use different tools to test disk performance:

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

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, you must make sure the disk is 4K aligned.

You can use fio to test the performance of a cloud disk.



Warning:

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. We recommend that you use a new ECS instance without data on the disks to test the disks by using fio.

- 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
```

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

Parameter	Description
-direct=1	Ignore I/O buffer when testing. Data is written directly.
-rw=randwrite	Read and write policies. Available options: <ul style="list-style-type: none"> • randread (random read) • randwrite(random write) • 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 and asynchronous. 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. The asynchronous method uses libaio to submit a batch of I/O requests each time , thus reducing interaction times, and makes interaction more effective.
-bs=4k	The size of each block for one I/O is 4k. If not specified, the default value 4k is used. When IOPS is tested, we recommend that you set the bs to a small value, such as 4k in this example command. When throughput is tested, we recommend that you set the bs to a large value , such as 1024k in the IOPS tests.
-size=1G	The size of the testing file is 1 GiB.
-numjobs=1	The number of testing jobs is 1.
-runtime=1000	Testing time is 1,000 seconds. If not specified, the test will go on with the value specified for -size, and write data in -bs each time.
-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.

Parameter	Description
-filename=iotest	The output path and name of the test files. 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 Elastic block storage

Elastic block storage is a low-latency, persistent, and high-reliability random block-level data storage service provided by Alibaba Cloud to ECS users. It uses a [triplicate distributed system](#) to provide 99.9999999% data reliability for ECS instances. Elastic block storage supports the automatic copying of your data within the zone. It prevents unexpected hardware faults from causing data unavailability and protects your service against the threat of component faults. As with a hard disk, you can partition the elastic block storage attached to an ECS instance, create a file system, and store data on it.

You can expand your elastic block storage as needed at any time. For more information, see [Linux _ Resize a data disk](#) or [Increase system disk size](#). You can also create snapshots to back up data for the elastic block storage. For more information about snapshots, see [Snapshots](#).

Based on whether it can be attached to multiple ECS instances, the elastic block storage can be divided into:

- Cloud disks: Can be attached to only one ECS instance in the same zone of the same region.
- Shared Block Storage: Can be attached to up to eight ECS instances in the same zone of the same region.

Cloud disks

Based on performance, cloud disks can be divided into:

- SSD Cloud Disk: Adopts SSD (Solid-state drive) as a storage medium to deliver stable and high-performance storage with high random I/O and high data reliability.
- Ultra Cloud Disk: Adopts the hybrid media of SSD and HDD (Hard disk drive) as a storage media.
- Basic Cloud Disk: Adopts HDD as a storage medium.

Cloud disks can be used as:

- **System disks:** Has the same lifecycle as the ECS instance it is attached to. It is created and released along with the instance. Shared access is not allowed. The available size range of a single system disk varies according to the image:
 - Linux (excluding CoreOS) and FreeBSD: 20–500 GiB
 - CoreOS: 30–500 GiB
 - Windows: 40–500 GiB
- **Data disks:** Can be [created separately](#) or jointly with ECS instances. Shared access is not allowed. The data disk created with an ECS instance has the same lifecycle as the instance, and is created and released along with the instance. Separately created data disks can be [released independently](#) or with ECS instances.

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 high level of concurrency, performance, and reliability. It supports concurrent reads and writes to multiple ECS instances, and provides data reliability of up to 99.9999999%. Shared Block Storage can be attached to a maximum of eight ECS instances. This service is currently in public beta, during which Shared Block Storage can be attached to a maximum of four 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 to be released when the ECS instances are released.

Based on different performance, the Shared Block Storage can be divided into:

- **SSD Shared Block Storage:** Adopts SSD as the storage medium to provide stable and high-performance storage with enhanced random I/O and data reliability.
- **Ultra Shared Block Storage:** Adopts 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, during which it is free of charge.

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

- Cloud disks created with Subscription (monthly or yearly subscription) instances are billed by upfront payment before the service is ready for use. For more information, see [Subscription](#).
- Cloud disks created jointly with 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](#).

Information about changing the billing method of a cloud disk is shown in the following table.

Conversion of billing methods	Features	Effective date	Suitable for
Subscription —> Pay-As-You-Go	Renew and downgrade configurations	Effective from the next billing cycle	Subscription cloud disks attached to Subscription instances . The billing method of the system disk cannot be changed.
Pay-As-You-Go —> Subscription	Upgrade configurations	Effective immediately	Pay-As-You-Go data disks attached to Subscription instances . The billing method of the system disk cannot be changed.
	Switch from Pay-As-You-Go to subscription		The system disks and data disks attached to the Pay-As-You-Go instances.

Related operations

You can perform the following operations on an elastic block storage:

- If [an elastic block storage device is created separately from a data disk](#), you must [attach the device to an instance](#) in the console, and then connect to the ECS instance to [partition and format the data disk](#).
- If you want to encrypt the data on elastic block storage, [encrypt the storage](#).
- If your data disk capacity is insufficient, you can [resize the data disk](#).
- If you want to change the operating system, you have to change the system disk.
- If you want to back up the data of the elastic block storage, you can [manually create snapshots for the elastic block storage](#) or [apply an automatic snapshot policy to it](#) to automatically create snapshots on schedule.

- If you want to use the operating system and data environment information of one instance on another instance, you can [create a custom image by using the system disk snapshots of the former](#).
- If you want to restore the elastic block storage to the status when the snapshot is created, you can [roll back the disk](#) by using its snapshot.
- If you want to restore the elastic block storage to its status at the time of creation, you can [reinitialize the disk](#).
- If you do not need the elastic block storage, you can [detach](#) and [release it](#).

For more information about the operations on cloud disks, see the [Cloud disks](#) section in the *Cite LeftUser GuideCite Right*.

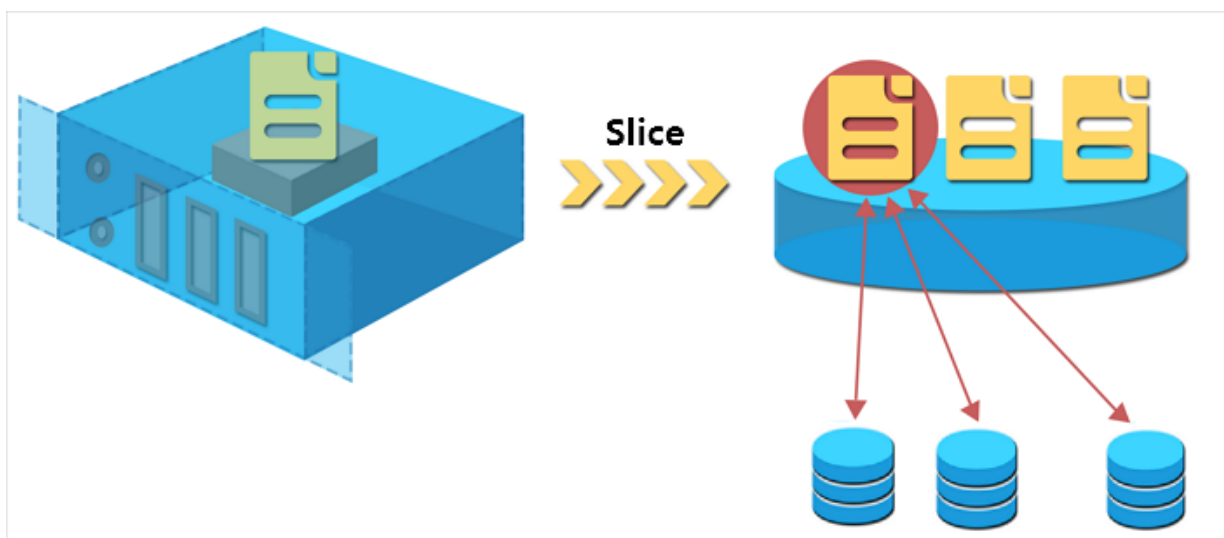
6.4 Triplicate technology

The Alibaba Cloud distributed File System provides stable and efficient data access and reliability for ECS. Triplicate technology, that is, the process of triple replication of data, is the principle concept designed by Alibaba Cloud and implemented in the Distributed File System.

Chunks

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.

For the data on the cloud disk, all user operations, including addition, modification, and deletion of data, are synchronized to the three copies. This mode guarantees the reliability and consistency of user data.



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 one of your write operations.
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 to the user indicating 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.

Triplicate technology is strongly recommended to be used in conjunction with other data protection means, such as regular data backups or [snapshots](#). Make sure that all appropriate actions are implemented to protect your data and guarantee its availability.



6.5 ECS disk encryption

What is ECs disk encryption?

When you want to encrypt the data stored on a disk due to business needs or certification requirements, you can use ECS disk encryption function to encrypt cloud disks and shared block storage (referred to collectively as **cloud disks**, unless otherwise specified). This secure encryption feature allows you to encrypt new cloud disks. 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 decryption operations are required, so the operation of the disk encryption function is practically invisible to your applications or your operations.

Encryption and decryption barely degrades cloud disk performance. For information on the performance testing method, see [#####](#).

After an encrypted cloud disk is created and attached to an ECS instance, the data in the following list can be encrypted:

- Data on the cloud disk.
- Data transmitted between the cloud disk and the instance. However, data in the instance operating system is not encrypted.
- All snapshots created from the encrypted cloud disk. 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, and SSD cloud disks) and shared block storage (ultra and SSD).

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

ECS disk encryption dependencies

The ECS disk encryption function depends on the Key Management Service (KMS) in the same region. However, you are not required to perform any additional operations in the KMS console, unless you want to perform separate KMS operations.KMS operational requirements.

The first time you use the ECS disk encryption function when you are creating ECS instances or cloud disks, you must follow the prompts to authorize and activate KMS. Otherwise, you cannot create encrypted cloud 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 website.

The first time you encrypt a disk in a given region, Alibaba Cloud automatically creates a Customer Master Key (CMK) in the KMS region, exclusively for ECS. You cannot delete this CMK and can query it in the KMS console.

Key management for ECS disk encryption

The ECS disk encryption function handles key management for you. Each new cloud 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 cloud disk and any cloud disks subsequently created from these snapshots. The bit key (from the user master key) is encrypted. 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 by using an industry standard AES-256 algorithm.

You cannot change the CMK associated with encrypted cloud 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 for ECS product in each region. This key is separate from the data and stored in a system protected by strict physical and logical security controls. Each encrypted disk uses an encryption key unique to the specific disk and its snapshots. 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.

Fees

ECS does not charge any additional fees for the disk encryption function.

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 and no additional fees are incurred.



Note:

您对磁盘的任何读写操作（例如 `mount/umount`、分区、格式化等）都不会产生费用。凡是涉及磁盘本身的管理操作（见下面列表），无论是通过 ECS 管理控制台还是通过 API，均会以 API 的形式使用到密钥管理服务（KMS），将会记入到您在该地域的 KMS 服务 API 调用次数。

These operations include:

- Creating encrypted cloud disks by calling `CreateInstance` or `CreateDisk`.
- Attaching an encrypted cloud disk to an instance by calling `AttachDisk.Mount` (`attachdisk`).
- Detaching an encrypted cloud disk from an instance by calling `DetachDisk.Unload` (`detachable disk`).
- Creating a snapshot by calling `CreateSnapshot`.
- Restoring a cloud disk by calling `ResetDisk`.
- Re-initializing a cloud disk by calling `ReInitDisk`.

Create an encrypted cloud 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 purchasing an ECS instance:
 - Check `Encrypted` to create a encrypted blank cloud disk.
 - Select an encrypted screenshot to create a cloud disk. Select `encrypt snapshot` to create the cloud.
- 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 convert an **encrypted disk** to an **unencrypted disk**.

Likewise, you cannot convert a snapshot created from an **unencrypted disk** to an **encrypted snapshot** or convert a snapshot created from **encrypted disk** to an **unencrypted snapshot**.

Therefore, to convert data from **unencrypted** data to **encrypted** data, 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**.

To convert data from **encrypted** data to **unencrypted** data, 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 cloud disks, not local disks.
- You can only encrypt data disks, not system disks.
- You cannot directly convert existing unencrypted disks into encrypted disks.
- You cannot convert encrypted cloud disks into unencrypted cloud 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 choose your CMKs for each region, as these are generated by the system.
- The ECS system creates CMKs for each region. You cannot delete these keys, but no fees are incurred.
- After a cloud disk is encrypted, you cannot change the CMK used for encryption and decryption.

6.6 Local disks

Local disks are located on the physical servers (host machines) that ECS instances are hosted on. They provide temporary block level storage for instances, featuring low latency, high random IOPS, and high I/O throughput. They are designed for business scenarios requiring high storage I/O 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 cause single points of failure. We recommend that you implement data redundancy at the application layer to guarantee data availability.



Warning:

Using a local disk for data storage can carry a risk of data loss in some cases, such as when the host machine is down. Therefore, never store any business data that requires long-term

persistence on a local disk. If no data reliability architecture is available for your application, we strongly recommend that you build your ECS with [elastic block storage](#).

Categories

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](#), [gn5](#), and [ga1](#). The instance type families i1 and i2 apply to the following scenarios:
 - Online games, e-businesses, live videos, media, 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 the d1 type families](#). It is applicable to the Internet, finance, and other allied 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 represented by Hadoop in multiple aspects, such as 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: <ul style="list-style-type: none"> • Single disk IOPS: $IOPS = \min\{165 * \text{capacity}, 240000\}$

Parameters	Local NVMe SSD
	<ul style="list-style-type: none">Single disk throughput: $\text{Throughput} = \min\{0.85 * \text{capacity}, 1200\}$ MBps Read performance: <ul style="list-style-type: none">Single disk IOPS: $\text{IOPS} = \min\{165 * \text{capacity}, 240000\}$Single disk throughput: $\text{Throughput} = \min\{1.4 * \text{capacity}, 2000\}$ MBps
Access latency	In microseconds

* Explanation on single disk performance calculation formulas:

- 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	In milliseconds

Billing

Local disks charges are covered in the payment for 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:

- 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.

Operations on an instance affect the data on the local disk

The following table shows 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	Description
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.
Shut down 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 comes 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 creating an empty local disk or creating a local disk from a snapshot.
- Attaching a local disk in the ECS console.
- Detaching and releasing a local disk.
- Increasing the size of a local disk.
- Re-initializing a local disk.
- Creating a snapshot for a local disk and using the snapshot to roll back the local disk.

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.

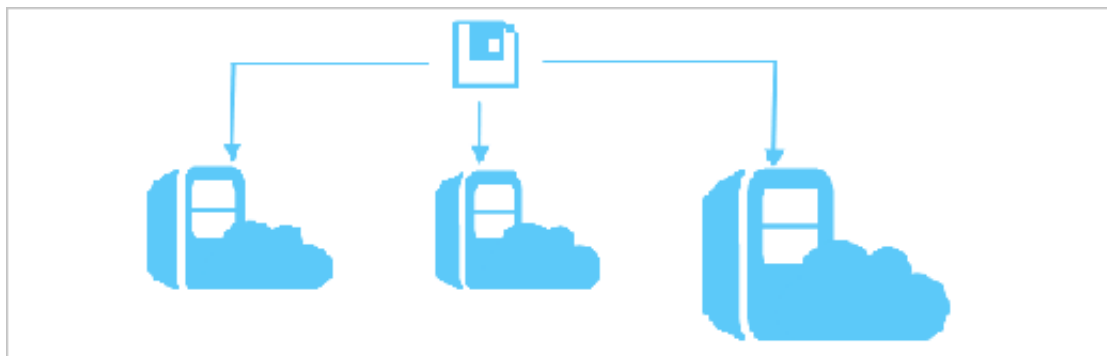


Image types

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

Type	Description	Source
Public image	<p>Public images officially provided by Alibaba Cloud support nearly all main Windows and Linux versions. Including:</p> <ul style="list-style-type: none">• Windows Server• Centos• Ubuntu• Debian• SuSE Linux• Opensuse• Aliyun Linux• Coreos• Freebag <p>These images are of high stability and are licensed. You can customize your application environment based on a public image.</p>	Officially provided by Alibaba Cloud

Type	Description	Source
Custom image	Custom mirrors created based on your existing physical server, virtual machine, or cloud host. These images are flexible to meet your personalized needs.	<ul style="list-style-type: none"> 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
Shared image	Shared by other Alibaba Cloud users.	A Custom image shared by other Alibaba Cloud users.

Image format

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

Fees

Details of the images are as follows:

Type	Fee description
Public image	Free
Custom image	<p>Free. Potential costs include:</p> <ul style="list-style-type: none"> If you use a snapshot to create a custom image: <ul style="list-style-type: none"> 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.

Type	Fee description
	<ul style="list-style-type: none"> — 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. <p>Current snapshot is commercialized.</p> <ul style="list-style-type: none"> • 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 [Cite Left ECS Pricing](#) [Cite Right](#).

Limits

Except for public images, custom images, Marketplace images, and shared images vary depending on the region. For more information about regions and zones, see [Cite Left General Reference](#) [Cite 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:
 - Using a public image
 - Using other images other than public ones
- You can obtain custom images in the following ways:
 - Creating a custom image by using a snapshot
 - Creating a custom image by using an instance
 - Importing a custom image
- You can copy your custom images to other regions.
- You can share your custom images with other Alibaba Cloud users.
- You can export custom images to local testing environments or your private cloud environments.

API operations

You can view the *.Cite LeftDevelopment GuideCite Right* for APIs about images.

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 Snapshsot 2.0 service is currently online. Unless and otherwise specified, either “snapshot” or “snapshot service” in all ECS articles is assumed as Snapshsot 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

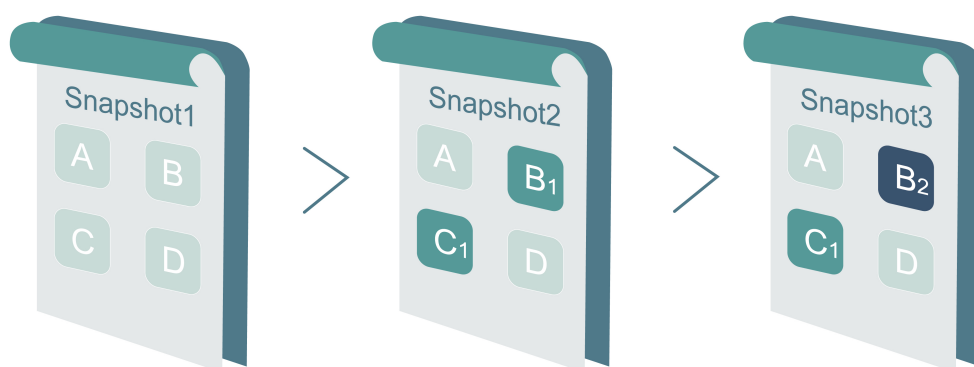
If your business no longer requires a snapshot of an elastic block storage device, you can delete the snapshot. If you have applied an automatic snapshot policy to the storage device, you can delete the automatic snapshot policy.

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 guarantee 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 [Elastic 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	<ul style="list-style-type: none">• 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
				<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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 permanently. 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

10.1 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

Term	Common name	Description
		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 (<i>Invocation</i>).
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.
Invocation status	InvokeStatus	<p>The relationship among command invocation status. The invocation status can be divided into three levels:</p> <ul style="list-style-type: none">• 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:

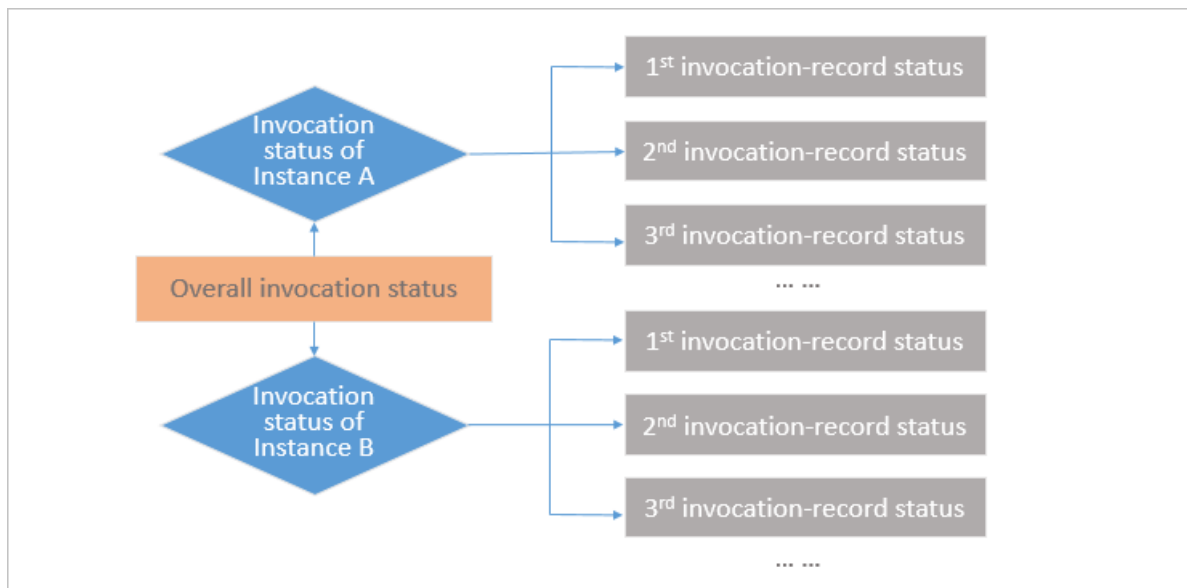
Supported images	Supported regions
<ul style="list-style-type: none">• Windows Server 2008/2012/2016• Ubuntu 12/14/16• Centos 5/6/7• Debian 7/8/9• RedHat 5/6/7• SuSE Linux Enterprise Server 11/12• Opensuse• Aliyun Linux• Freebag• Coreos	<ul style="list-style-type: none">• China East 1 (Hangzhou)• China 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)• 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.

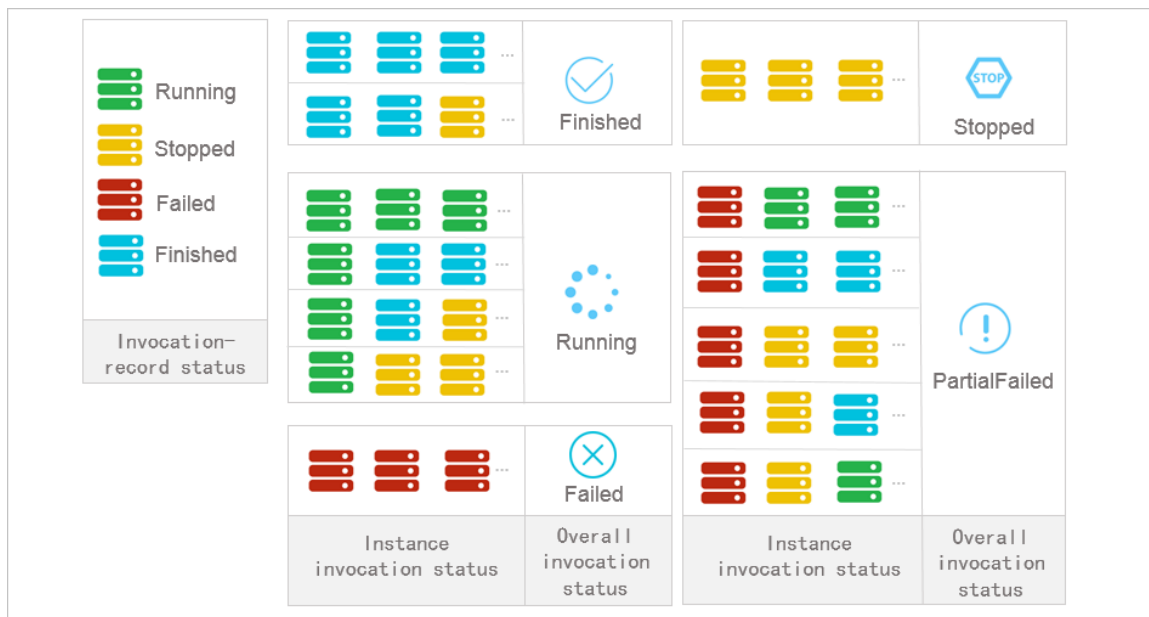


For one-time invocations

- **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](#)