Elastic Compute Service

Product Introduction

Product Introduction

What is ECS?

Elastic Compute Service (ECS) is a type of computing service that features elastic processing capabilities. Its management mode is simpler and more efficient than that of physical servers. You can create instances, change the operating system, or release any number of ECS instances at any time according to business needs without upfront investment.

An ECS instance is a virtual computing environment, including CPU, memory, and other basic computing components. It is the actual operating entity offered by Alibaba Cloud. An ECS instance is the core concept of ECS. Other resources, such as disks, IPs, 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 an ECS instance, for example, to select the instance type, disks, the operating system, bandwidth, security groups and other resources.

Advantages

ECS has the following advantages that traditional servers and virtual hosts cannot match:

- **Stability**: The service availability is up to 99.95%. The data reliability is up to 99.9999999%. ECS supports automatic downtime migration, data snapshot backup and rollback, and system performance alarms.
- **Disaster recovery backup**: Each data segment is saved in multiple copies so that, if a single segment is physically damaged, the data can be quickly restored.
- **Security**: ECS supports the setting of security group rules, Anti-DDoS, multi-user isolation, and password cracking defense.
- **Multiline access**: ECS is based on the optimal routing algorithm of the Border Gateway Protocol (BGP). Multiline BGP machine rooms ensure smooth and balanced access throughout the country. Backbone machine rooms ensure high output bandwidth and dedicated bandwidth.
- **Low cost**: Large one-time investment is not required. Pay-As-You-Go and flexible payment options let you flexibly cope with business changes.
- **Controllability**: As an ECS user, you have the permissions of a super administrator. That means, you can completely control the operating system of ECS instances, resolve system problems on your own through the management terminal, and perform operations such as environment deployment and software installation.
- **Ease of use**: Various operating systems and applications are supported. Using images, you can simply deploy the same image by one click. You can quickly replicate the environment to multiple ECS instances for easy scaling. You can also create ECS instances in batches through custom images and disk snapshots.
- **API**: Through API invocation management of ECS, you can configure the access to one or multiple servers with the security group feature, making the development more convenient.

Features

ECS provides the following features:

- Allows instance creation in 13 regions, some of which provide multiple zones.
- Supports multiple instance generations, three instance type families, and dozens of instance types (ranging from 1-core 1GB to 16-core 128GB), meeting various demands.
- Provides three types of data storage disks (Basic Cloud Disks, Ultra Cloud Disks and SSD Cloud Disks) and I/O-optimized instances.
- Supports two types of IP addresses (public and private IP addresses), allowing for Intranet interconnection and Internet access.
- Supports two network types (Classic Network and VPC) to manage your network in different dimensions.

- Supports multiple Windows and Linux operating systems.
- Allows free activation of Alibaba Cloud Security products and provides network monitoring.
- Provides various image resources, including public images, custom images and shared images, allowing you to quickly deploy operating systems and applications without installation.
- Provides multiple management methods, including the console, management terminals and APIs, so that you will have complete control.
- Provides flexible payment methods (Subscription and Pay-As-You-Go).

Application scenarios

ECS is widely applied. It not only can be used independently as a simple Web server but also can be used with other Alibaba Cloud products (such as OSS and CDN) to provide advanced solutions.

The following are typical application scenarios of ECS:

Official corporate websites and simple Web applications

In the initial stage, such a website has a low traffic volume and only requires a single low-configuration ECS instance 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 resource insufficiency in case of traffic peak on the low-configuration ECS instance.

Multimedia and large-traffic APPs or websites

When ECS is used with OSS, static images, videos, and downloaded packages can be stored on the OSS to reduce storage fees. In addition, ECS can be used with CDN or Server Load Balancer to greatly shorten 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 and higher data reliability. Alternatively, multiple lower-configuration I/O-optimized ECS instances can be used with Server Load Balancer to deliver a high-availability architecture.

APPs or websites with a large traffic fluctuation

Some applications may encounter a huge traffic fluctuation within a short period. When ECS is used with Auto Scaling, the number of ECS instances automatically increases when the traffic rises, and automatically decreases when the traffic drops. This meets the resource requirements when the traffic volume reaches its peak value and reduces the cost. If ECS is used with Server Load Balancer, a high availability architecture is delivered.

Instances

Overview

An ECS instance is a virtual computing environment, including CPU, memory, operating system, bandwidth, disks, and other basic computing components. It can be understood as an independent virtual machine. An ECS instance is the core concept of the ECS. Other resources, such as disks, IPs, images, and snapshots, can only be used in conjunction with an ECS instance.

Instance generations

ECS instances are divided into different instance generations based on different underlying hardware. Currently the following instance generations are supported:

Instance Generation I

- Intel Xeon CPU
- DDR3 memory
- Optional I/O optimization

Instance Generation II

- Haswell CPU enables users to get larger instance types. In addition, some new instruction sets are included, doubling the integer and floating-point operation performance
- DDR4 memory enables faster memory access
- All are I/O-optimized instances that are used with SSD cloud disks for higher I/O performance

Instance Generation III

- Broadwell CPU enables users to get larger instance types.
- DDR4 memory enables higher memory bandwidth and faster memory access.
- All are I/O-optimized instances that are used with SSD cloud disks for higher I/O performance.

Generation I, II, and III instance types cannot be upgraded or downgraded to be compatible with each other.

Instance family types

ECS instance types are divided by a certain standard into several instance type families based on their configurations and application scenarios.

Instance generation III type families

The type families included in instance generation III are:

- Compact type XN4
- General type N4
- Balanced type MN4
- Memory type E4

For more information, see the following table.

Type Family	Description	Features	Application Scenarios
Compact type XN4	The latest compact ECS instance types, with a new generation of CPU and memory providing higher computing power, applicable to scenarios with high computing requirements and compact memory configurations.	- The CPU/memory ratio is 1:1. - Intel Xeon E5-2682 v4 (Broadwell) processor - 2.5 GHz clock speed - I/O-optimized instances.	- Small Web applications, and small and medium databases; - Massively Multiplayer Online Game (MMO) front ends, data analysis, and computing; - Medium/large Web servers (high concurrency) - High-precision coding/decoding,

			rendering, and gene computing with a CPU.
General type N4	The latest general ECS instance types, with a new generation of CPU and memory providing higher computing power, applicable to general computing scenarios.	- The CPU/memory ratio is 1:2. - Intel Xeon E5-2682 v4 (Broadwell) processor - 2.5 GHz clock speed - I/O-optimized instances	- Small Web applications, and small and medium databases; - Massively Multiplayer Online Game (MMO) front ends, data analysis, and computing; - Medium/large Web servers (high concurrency); - High-precision coding/decoding, rendering, and gene computing with a CPU.
Balanced type MN4	The latest balanced ECS instance types, with a new generation of CPU and memory providing higher computing power, applicable to	- The CPU/memory ratio is 1:4. - Intel Xeon E5-2682 v4	- Small Web applications, and small and medium databases;

	scenarios requiring enhanced balance between computing and memory.	(Broadwell) processor - 2.5 GHz clock speed - Enhanced balance of computing capability - I/O-optimized instances	- Massively Multiplayer Online Game (MMO) front ends, data analysis, and computing; - Medium/large Web servers (high concurrency); - High-precision coding/decoding, rendering, and gene computing with a CPU; - Hadoop clusters, massive data processing.
Memory type E4	The latest memory-intensive ECS instance types, with a new generation of CPU and memory providing higher computing power, applicable to computing services requiring massive memory operations and search, and applications requiring massive I/O operations.	- The CPU/memory ratio is 1:8. - Intel Xeon E5-2682 v4 (Broadwell) processor - 2.5 GHz clock speed - Wider range of	- Applications that need lots of memory, like Cache/Redis, search, and in-memory databases; - Databases with high

		memory capacity options - I/O-optimized instances	I/O, for example, Oracle and MongoDB; - Hadoop clusters, massive data processing.

Instance generation II

The type families included in Instance Generation II are:

- General type N1
- General type N2
- Memory type E3
- Compute Optimized type SN1
- Compute Optimized type SN2

For more information, see the following table.

Type Family	Description	Features	Application Scenarios
General type N1	The latest general ECS instance types, with a new generation of CPU and memory providing higher computing power and faster I/O capabilities. Among N1 instances, the small instance types are suitable for entry-level users and the large ones are suitable for use in certain specialized fields such as gaming, rendering, genetics, encoding/decoding, deep learning and more.	- Intel Xeon E5-2680 v3 (Haswell) processor - 2.5GHZ clock speed - I/O-optimized instances	- Small Web applications, and small and medium databases; - Massively Multiplayer Online Game (MMO) front ends, data analysis, and computing; - Medium/large Web servers

			(high concurrency) - High-precision coding/decoding, rendering, and gene computing with a CPU
General type N2	The latest general ECS instance types, with a new generation of CPU and memory providing higher computing power and faster I/O capabilities. Compared to N1, instances of the N2 type family have a greater memory capacity and wider range of options.	- Intel Xeon E5-2680 v3 (Haswell) processor - 2.5GHZ clock speed - Wider range of memory capacity options - I/O-optimized instances	- Small Web applications, and small and medium databases; - Massively Multiplayer Online Game (MMO) front ends, data analysis, and computing; - Medium/large Web servers (high concurrency); - High-precision coding/decoding, rendering, and gene computing with a CPU;

			- Hadoop clusters, massive data processing.
Memory type E3	The latest memory-intensive ECS instance types, with a new generation of CPU and memory providing higher computing power and a wider range of memory options. These types are suitable for computing services that require large volumes of memory operations and queries as well as applications that require large volumes of I/O operations.	- Intel Xeon E5-2680 v3 (Haswell) processor - 2.5GHZ clock speed - Memory-intensive - I/O-optimized instances	- Applications that need lots of memory, like Cache/Redis, search, and in-memory databases; - Databases with high I/O, for example, Oracle and MongoDB; - Hadoop clusters, massive data processing.
Compute Optimized type SN1	The new generation of exclusive ECS instance types is designed with a new generation of CPU and memory providing higher computing power and faster I/O capabilities. The fixed performance computing capability is optimized.	- Intel Xeon E5-2680 v3 (Haswell) processor - 2.5 GHz clock speed - I/O-optimized instances - The fixed performance computing capability is	- Medium/large Web servers (high concurrency); - Massively Multiplayer Online Game (MMO) front ends, data analysis,

			and computing; - High-precision coding/dec oding, rendering, gene computing and other fixed performanc e computing with a CPU.
		optimized.	
Compute Optimized type SN2	The new generation of ECS instance types is designed with a new generation of CPU and memory providing higher computing power and faster I/O capabilities. The fixed performance computing capability is optimized. Compared to SN1 instances, instances of the SN2 type family have a greater memory capacity and wider range of options.	- Intel Xeon E5-2680 v3 (Haswell) processor - 2.5 GHz clock speed - I/O-optimized instances - The fixed performanc e computing capability is optimized.	- Medium/lar ge Web servers (high concurrency); - Massively Multiplayer Online Game (MMO) front ends, data analysis, and computing; - High-precision coding/dec oding, rendering, gene computing and other fixed performanc

			e computing with a CPU.

Instance generation I

All Generation I instance types are legacy instance types. They are still grouped using the previous method (i.e. by the number of cores - 1, 2, 4, 8 and 16 cores) and not sensitive to type families.

Instance types

An instance is the minimal unit that can provide computing services for your business. It uses a certain specification to provide you with the associated computing capabilities.

An ECS instance type defines two basic properties: the instance CPU and memory configurations (including CPU model, clock speed, and so on). However, only when used in conjunction with a disk category and image and network types will the instance type uniquely identify the specific service form of the instance.

Instance life cycle

The instance life cycle begins with creation (purchase) and ends with final release (the expiration of the yearly/monthly instance subscription; suspension due to unpaid fees or voluntary release of the user in the case of payment based on amount).

Inherent instance statuses

During this life cycle, there are several inherent instance statuses, as listed below:

Status	Status Property	Explanation	Corresponding API Status
Preparing	Intermediate Status	After an instance is created, this is the status it is in before running. If the instance remains in this status for a long time, it indicates that the instance has encountered an exception.	Pending
Created	Stable Status	An instance is in this	Stopped

		status when it has been created and is awaiting start up. If the instance remains in this status for a long time, it indicates that the instance has encountered an exception.	
Starting	Intermediate Status	An instance is in this status after it is restarted, started up, etc. in the console or through API and before it is running. If the instance remains in this status for a long time, it indicates that the instance has encountered an exception.	Starting
Running	Stable Status	The instance is operating properly. In this status, the instance can accommodate your business.	Running
Stopping	Intermediate Status	An instance is in this status after the stop operation is performed in the console or through API and before it actually stops. If the instance remains in this status for a long time, it indicates that the instance has encountered an exception.	Stopping
Stopped	Stable Status	The instance has been stopped properly. In this status, the instance cannot accommodate external services.	Stopped
Re-initializing	Intermediate Status	An instance is in this status after the system disk and/or data disk is re-initialized in the console or through API and before it is running. If the instance remains in this status for a long time, it indicates that the instance has encountered an exception.	Stopped

Replacing System Disk	Intermediate Status	An instance is in this status after the operating system is replaced or another such operation is performed in the console or through API and before it is running. If the instance remains in this status for a long time, it indicates that the instance has encountered an exception.	Stopped
Expired	Stable Status	The yearly/monthly instance subscription has expired because you have not promptly renewed it; the pay-as-you-go instances has expired because they are in arrearage. **Note:** After expiration, both of the yearly/monthly and pay-as-you-go instances will continue to run 15 days, and the data will be retained for extra 15 days. Then the instances will be released and the data will be removed permanently .	Stopped

API status chart

This table describes the correspondence relationships between console statuses and API statuses. The API status chart is shown below:
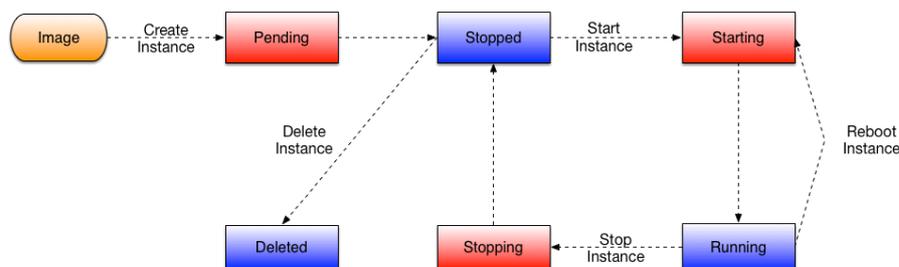
Disks

Overview

An ECS disk can be used separately or in combination to meet the requirements of different application scenarios. After reading this chapter, users can select proper data storage options as required.

Cloud disks provide block-level data storage for ECS instances, using a triplicate, distributed system. It provides 99.9999999% data reliability assurance for the ECS instances.

In addition, cloud disks can be classified into Basic Cloud Disks, Ultra Cloud Disks, and SSD Cloud Disks by performance.

- Basic Cloud Disks are for least I/O-intensive application scenarios and provide an I/O performance of several hundred IOPS for ECS instances.
- Ultra Cloud Disks are for medium I/O load application scenarios and provide a storage performance of up to 3,000 random IOPS for ECS instances.
- SSD Cloud Disks are for I/O-intensive applications and provide stable, high random IOPS performance.

For detailed instructions on attaching a disk, please refer to Attach data disks.

Disk comparison

Block Storage	SSD Cloud Disk	Ultra Cloud Disk	Basic Cloud Disk
Maximum capacity	32768 GB	32768 GB	2000 GB
Maximum IOPS	20000	3000	Several hundred
Maximum throughput	256 Mbps	80 Mbps	30 Mbps
Performance Calculation Formula	IOPS=min{30*capacity,20000} Throughput=min{50+0.5*capacity,256}Mbps	IOPS=min{1000+6*capacity,3000} Throughput=min{50+0.1*capacity,80}Mbps	N/A
Access Latency	0.5 - 2 ms	1 - 3 ms	5 - 10 ms
Data reliability	99.9999999%	99.9999999%	99.9999999%
API Name	cloud_ssd	cloud_efficiency	cloud
Price*	US$0.2140/GB/mo	US$0.0920/GB/mo	US$0.0580/GB/mo

Typical Application Scenarios	- I/O-intensive applications - Medium/Large relational databases - NoSQL databases	- Medium/Small databases - Large-scale development and testing - Web server logs	Infrequent access or low-I/O applications

*The prices shown are for the Hong Kong region. For more information, see ECS Price.

For more information about ephemeral SSD disks, see ephemeral SSD disks.

Methods to test disk performance

- Test random writing 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 reading 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 writing throughput ：
   fio -direct=1 -iodepth=64 -rw=randwrite -ioengine=libaio -bs=64k -size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=iotest -name=Write_PPS_Testing
- Test reading throughput ：
   fio -direct=1 -iodepth=64 -rw=randread -ioengine=libaio -bs=64k -size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=iotest -name=Read_PPS_Testing

Descriptions of fio parameters：

Parameter	Description
-direct=1	Ignore I/O cache when testing. Data is written directly.
-rw=randwrite	Read and write policies. Available options: randread (random read), randwrite(random write), read(sequential read), write(sequential write) and 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 each time, and returns only after the kernel is completed. In this case, the iodepth is always less than 1

	for a single job, but can be resolved by multiple concurrent jobs. Usually 16 - 32 concurrent jobs can fill up the iodepth. Asynchronous method uses libaio to submit a batch of I/O request each time, thus reduces 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.
-size=1G	The size of the testing file is 1G.
-numjobs=1	The number of testing jobs is 1.
-runtime=1000	Testing time is 1000 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 of showing the testing results. Group_reporting means sums up statistics of each job, instead of showing statistics by different jobs.
-filename=iotest	The output path and name of test files. Remember to delete the related files to free up space.
-name=Rand_Write_Testing	The name of the testing task.

Features and application scenarios of disks

Cloud disks are divided into SSD cloud disks, ultra cloud disks and Basic Cloud Disks. The following gives a description of the features and application scenarios of these types of disks separately.

- SSD Cloud Disks
- Ultra Cloud Disks
- Basic Cloud Disks

SSD Cloud Disks

Product features

SSD cloud disks use a distributed, triplicate mechanism to provide high-performance storage with stable, high random I/O and high data reliability. They provide the following features:

- High performance: They provide a maximum of 20,000 random read/write IOPS and a maximum throughput of 256Mbps.
- IOPS: Each GB of capacity provides 30 random IOPS and a maximum of 20,000 random

read/write IOPS. For example, a 100GB SSD cloud disk can provide 3,000 IOPS, and a 334GB SSD cloud disk can provide 10,020 IOPS.

Note: Expected IOPS performance can be achieved only when the SSD cloud disk is attached to an I/O-optimized instance. An SSD cloud disk attached to a non I/O-optimized instance cannot achieve the expected IOPS performance.

- Throughput: SSD cloud disk throughput=min {50+0.5xdisk_size, 256} Mbps; initially 50Mbps, it increases by 0.5Mbps for each GB, up to 256Mbps.
- Performance baseline:
  - When the block size is 4KB/8KB, the IOPS can reach up to 20,000;
  - When the block size is 16KB, the IOPS can reach up to approximately 16,300, and the throughput reaches the maximum value of 256Mbps;
  - When the block size is 32KB, the IOPS can reach up to approximately 8,150, and the throughput reaches the maximum value of 256Mbps;
  - When the block size is 64KB, the IOPS can reach up to approximately 4,100;
  - And so on.
- Data reliability: SSD cloud disks adopt a distributed, triplicate mechanism to provide 99.9999999% data reliability.
- Maximum capacity: A single SSD cloud disk provides up to 32,768GB storage space.
- Independent attaching: SSD cloud disks can be attached to any ECS instance in the same zone.

Application scenarios

SSD cloud disks have stable, high random I/O performance and high data reliability. They are applicable to the following scenarios:

- PostgreSQL, MySQL, Oracle, SQL Server, and other medium/large relational database applications.
- Medium/large development and testing environments with high data reliability requirements.

Ultra cloud disks

Product features

Ultra cloud disks adopt the hybrid media of SSD and HDD as the storage media. They provide the following features:

- I/O performance: Provides up to 3,000 random read/write IOPS and a maximum throughput of 80Mbps.
- Data reliability: Basic cloud disks adopt a distributed, triplicate mechanism to provide 99.9999999% data reliability.
- Performance criteria:

- IOPS: Initially 1,000 IOPS, it increases by 6 IOPS for each GB, up to 3,000 IPOS;
- Throughput: Initially 50Mbps, it increases by 0.1Mbps for each GB, up to 80Mbps; For example, a 250GB ultra cloud disk features 2,500 random read/write IOPS and a throughput of 75Mbps.
- Maximum capacity: A single ultra cloud disk provides up to 32768GB storage space.
- Independent attaching: Ultra cloud disks can be attached to any ECS instance in the same zone.

Application scenarios

- MySQL, SQL Server, PostgreSQL, and other small/medium relational database applications.
- Medium/large development and testing environments with high data reliability requirements and intermediate performance requirements.

Basic cloud disks

Product features

Basic cloud disks adopt HDDs as the storage medium and use a distributed, triplicate mechanism to provide high data reliability. They provide the following features:

- I/O performance: They provide a random read/write IOPS performance of several hundreds and a maximum throughput of 30-40Mbps.
- Data reliability: Basic cloud disks adopt a distributed, triplicate mechanism to provide 99.9999999% data reliability.
- Maximum capacity: A single basic cloud disk provides up to 2,000GB storage space.
- Independent attaching: Basic cloud disks can be attached to any ECS instance in the same zone.

Application scenarios

- This option is applicable to application scenarios in which data is not frequently accessed or that have low I/O loads. If an application requires higher I/O performance, it is recommended using an SSD cloud disk.
- Application environments that require low costs and have random I/O reading/writing.

Introduction to triplicate technology

The Alibaba Cloud Distributed File System provides stable, efficient and reliable random data access capabilities for ECS.

Chunks

The reading and writing operations of ECS users to the virtual disks are eventually translated into the reading and writing operations to the files on the Alibaba Cloud data storage platform. Alibaba Cloud provides a flat linear storage space, inside which the linear addresses are divided into slices. Each slice is called a chunk. Alibaba Cloud creates three copies for each chunk and stores these copies on different nodes according to a certain strategy to ensure the reliability of user data.

Principle of triplicate technology

The Alibaba Cloud data storage system consists of three roles: Master, Chunk Server and Client.

The writing operation of an ECS user, going through several conversions, is finally executed by the Client. A brief introduction to the procedures is as follows:

1. The Client calculates out the chunk corresponding to this writing operation.
2. The Client sends a request to this chunk for the storage location of the three copies.
3. The Client sends writing requests to the three Chunk Servers according to the results returned from the Master.
4. If all the three copies are written successfully, the Client returns a message to the user indicating that the operation is successful; otherwise, the Client returns a message to the user indicating that the operation fails.

The distribution strategy of the Master is decided on the basis of an overall consideration of the disk using conditions of all the Chunk Servers in the cluster, the distribution of the Chunk Servers under different kinds of switch racks, the power supply, and the instrument load to ensure that all the copies of a Chunk are distributed on different Chunk Servers on different racks. This can effectively prevent data unavailability caused by the failure of a Chunk Server or rack.

Data protection mechanism

When some data nodes are corrupted or some hard drives on a certain data node fail, the number of valid copies of some Chunks in the cluster will be less than 3. Once this occurs, the Master initiates the copy mechanism that copies data between Chunk Servers to make the number of valid copies of all the Chunks in the cluster reach 3.

In sum, for the data on the cloud disk, all the user operations, adding data, modifying data or deleting data, will all be synchronized to the three copies. This mode can ensure the reliability and consistency of user data.

While, the data losses caused by virus infection, personal erroneous deleting or hacker invasion in ECS instances need to be solved by technical solutions, such as creating copies and snapshots. There is not any single technology that can solve all the problems. You need to select a proper data protection measure according to the actual situation in order to build up a solid protection for your

business data.

Ephemeral SSDs

Ephemeral SSDs

Ephemeral SSD disks are no longer on sale. The information here is only for your reference.

Ephemeral SSDs use the local storage of the physical machines where the instances are located. This type of storage provides block-level data access capabilities to instances and has low latency, high random IOPS, and high throughput I/O capability.

Pay attention to the following issues when using ephemeral SSDs:

- The storage space provided by ephemeral SSDs of servers has potential single point of failure (SPOF) risks. It is recommended to implement data redundancy at the application layer to ensure data availability.
- Users cannot upgrade or downgrade the CPU, memory and ephemeral SSDs after ordering.

Parameter

Parameter	Ephemeral SSD
Maximum capacity	800 GB
Maximum IOPS	12000 random read/write
Maximum throughput	250 - 300Mbps
Performance Calculation Formula	N/A
Access Latency	0.5 - 2 ms
Data reliability	Only physical machine reliability, no SLA
API Name	ephemeral_ssd
Price*	0.8 RMB/GB/Month
Typical Application Scenarios	Distributed applications such as Hadoop and NoSQL, which feature ultra-high reliability and require low-latency and high I/O storage.

Product features

Ephemeral SSDs use the ephemeral SSD of the instance's physical server as storage space. They provide the following features:

- Low latency: Under normal conditions, they provide an access latency in microseconds.
- High random I/O: They provide a maximum performance of 12,000 random IOPS.
- High throughput: They provide a maximum performance of 300 MB/s I/O throughput.
- Limited data reliability: Since the ephemeral SSD uses a single physical server, the data reliability is determined by the reliability of the physical server; there are likely SPOF risks. It is recommended to implement data redundancy at the application layer to ensure data availability.
- Up to 800 GB capacity: A single ephemeral SSD provides a maximum storage space of 800 GB.
- Attaching/detaching not supported: Since the ephemeral SSD uses the local disk of a physical server, it does not support independent attaching/detaching.
- Users cannot upgrade or downgrade the CPU, memory and ephemeral SSDs after ordering.

Application scenarios

An ephemeral SSD features low latency, high random I/O, and high throughput. It is applicable to distributed, I/O-intensive applications with redundancy and scenarios that do not require high data reliability, for example:

- Distributed applications: NoSQL and MPP data warehouse, distributed file systems and other I/O-intensive applications. These applications have their own distributed data redundancy. Ephemeral SSDs can provide low latency, high random I/O, and high throughput I/O performance.
- Logs for large online applications: Large online applications can produce a large amount of log data and require high-performance storage. At the same time, the log data does not require highly reliable storage.
- As a swap partition of an instance: When the memory required by an application exceeds the memory actually allocated, a swap space can be used in Linux. When the swap space is enabled, the Linux system can frequently swap in-use memory pages from the physical memory to the swap space (no matter whether it is a dedicated partition of the existing file system or a swap file). In addition, it can free up space for memory pages that require a high access speed.

Network and security

Intranet

At present, Alibaba Cloud servers communicate via intranet, using a gigabit of shared bandwidth with

no special restrictions. Because this is a shared network, we cannot guarantee that the bandwidth speed will not fluctuate.

If you need to transmit data between two ECS instances in the same region, we suggest using an intranet connection. In addition, the intranet speed of RDS, Server Load Balancer, and OSS instances is also based on a gigabit shared bandwidth environment. Intranet connections can also be used to connect these products.

At present, you can use a direct intranet connection to link Server Load Balancer, RDS, and OSS with ECS instances in the same region.

For ECS instances in the intranet:

- By default, intranet communication is only used for instances of the same account in the same region.
- An intranet connection can also be used for instances of the same account and region, but different zones, even if the intranet IPs are in different network segments.
- For intranet communication between instances in the same region but of different accounts, you can use security groups. For more information, see Security Group Applications.
- The intranet IP addresses of instances cannot be modified or changed.
- Intranet and Internet addresses of instances do not support VIP (virtual IP) configuration.

Intranet DNS addresses

It often happens that customers modify their servers' DNS configurations and then want to change the configuration back, but forget the original address. During parameter setting, refer to Alibaba Cloud's intranet DNS addresses in supported regions, shown below:

North China 1:

- 10.202.72.116
- 10.202.72.118

East China 1:

- 10.143.22.116
- 10.143.22.118

East China 2 :

- 100.100.2.136
- 100.100.2.138

Hong Kong:

- 10.143.22.116
- 10.143.22.118

US East 1 :

    - 10.143.22.116
    - 10.143.22.118

North China 2:

    - 10.202.72.116
    - 10.202.72.118

South China 1:

    - 100.100.2.138
    - 100.100.2.136

Asia-Pacific:

    - 100.100.2.136
    - 100.100.2.138

IP Addresses for Classic Network

IP addresses are an important means for users to access ECS instances and for ECS instances to provide external services. Currently, classic IP addresses are uniformly distributed by Alibaba Cloud. They are divided into public IP addresses and private IP addresses.

Private IP addresses

An instance will be allocated with a private network card and bound to a specific private IP address. Private IP addresses are required and cannot be modified.

If you independently change a private IP address in an operating system, communication in the private network will be interrupted.

Communication traffic through private IPs between instances in the same region is free.

Private IP addresses can be used in the following scenarios:

    - Load balancing of the Server Load Balancer
    - Intranet mutual access between ECS instances
    - Intranet mutual access between an ECS instance and another cloud service (such as OSS and
      RDS)

Public IP addresses

By default, each instance is configured with a public network interface card. Different from private IP

addresses, public IP addresses are optional. If you select a public network bandwidth greater than 0 Mbps when purchasing an instance, you will be allocated a public IP address during creation of the instance.

Regardless of the fee method you select, when you purchase an instance, you must select a public network bandwidth limit. The outgoing bandwidth for the public network card will be limited based on the bandwidth limit you select.

Public network traffic will be charged.

Public IP addresses can be used in the following scenarios:

> - Mutual access between an ECS instance and the Internet
> - Mutual access between an ECS instance and another cloud service

Multicast and broadcast

ECS does not support multicast or broadcast.

Security groups

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

A security group is a virtual firewall that provides the feature stateful packet inspection. Security groups are used to set network access control for one or more ECSs. As an important security isolation means, security groups are used to divide security domains on the cloud.

Security group restrictions

> - A single security group cannot contain more than 1,000 instances. If you require Intranet mutual access between more than 1,000 instances, you can allocate them to different security groups and permit mutual access through mutual authorization.
> - Each instance can join a maximum of five security groups.
> - Each user can have a maximum of 100 security groups.
> - Adjusting security groups will not affect the continuity of a user's service.
> - Security groups are stateful. If an outbound packet is permitted, inbound packets corresponding to this connection will also be permitted.
> - Security groups have two network types: classic network and Virtual Private Cloud (VPC).

- Instances of the Classic Network type can join security groups on the classic networks in the same region.
- Instances of the VPC type can join security groups on the same VPC.

Security group rules

Security group rules can be set to permit or forbid ECS instances associated with security groups to access a public network or an intranet from the inbound and outbound directions.

You can authorize or cancel security group rules at any time. Security group rules you have changed will automatically apply to ECS instances associated with security groups.

When setting security group rules, please note the following restrictions:

- No rule in a security group can permit outbound access of an ECS instance while forbidding inbound access to the ECS instance, and vice versa.
- Security group rules must be simple. If you allocate multiple security groups to an instance, the instance may apply up to hundreds of rules. When you access the instance, the network may be disconnected.

Security group rule restrictions

Each security group can have a maximum of 100 security group rules.

Images

Overview

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.

ECS allows you to easily obtain an image in the following ways:

- Choosing a public image officially provided by Alibaba Cloud (multiple Windows and Linux versions are available)
- Creating a custom image based on an existing ECS instance
- Choosing an image shared by another Alibaba Cloud account

You can import an offline image file into an ECS cluster to generate a custom image.

You can also copy a custom image to another region to achieve consistent environment and application deployment across regions.

Snapshots

Overview

A snapshot is a copy of data on a disk at a certain time point.

When using a disk, you may encounter the following scenarios:

- When writing or storing data to disks, you may want to use data on a disk as basic data for another disk.
- Cloud disks (Basic Cloud Disk, Ultra Cloud Disk, and SSD Cloud Disk) provide a secure storage method to ensure your stored content will not be lost. However, if the data stored on a disk is incorrect due to an application error, or a hacker maliciously tampered with your data using an application vulnerability, you will need another mechanism to ensure your data can be restored to the desired state in case of problems.

Through snapshots, the above needs can be met easily and effectively.

Mechanism

Alibaba Cloud provides the snapshot mechanism. By creating a snapshot, you can retain a copy of the data on a disk at a certain time point. Scheduled creation of disk snapshots ensures the continuous operation of your business.

Incremental snapshot mechanism

Snapshots use an incremental method. Two snapshots are compared and only the data that has changed is copied, as shown in the image below:
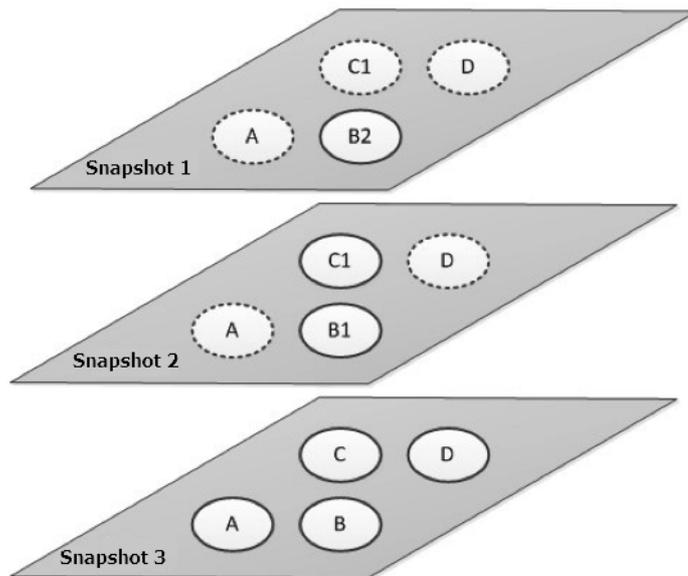
In the figure above, Snapshot 1, Snapshot 2, and Snapshot 3 are the first, second, and third snapshots of a disk. The file system checks the disk data by blocks. When a snapshot is created, only the blocks with changed data will be copied to the snapshot. In this example:

- Because Snapshot 1 is the first disk snapshot, it copies all the data on the disk.
- Snapshot 2 only copies the changed data blocks B1 and C1. Data blocks A and D reference A and D in Snapshot 1.
- In the same way, Snapshot 3 copies the changed data block B2. Data blocks A and D continue to reference Snapshot 1 and data block C1 references Snapshot 2.
- When the disk needs to be restored to the status at the time of Snapshot 3, you can perform snapshot rollback to copy data blocks A, B2, C1, and D to the disk.
- If Snapshot 2 is deleted, data block B1 in the snapshot will be deleted but data block C1 will not. In this way, when the disk is restored to the status at the time of Snapshot 3, data block C1 can also be restored.

It generally takes several minutes to manually create a 40 GB snapshot, depending on the actual data volume.

Snapshots are stored on the Object Storage Service (OSS), but they are invisible to users and will not be computed in the OSS space occupied by the users' buckets. Snapshot operations can only be performed through the ECS console or APIs.

Enhanced Snapshot 2.0 Product Specifications

Built on original basic snapshot features, ECS Snapshot 2.0 data backup service provides higher

snapshot quota and more flexible automatic task policies, further reducing its impact on business IO, detailed as below:

Features	Original snapshot specifications	Snapshot 2.0 specifications	User value	Examples
Snapshot quota	(Number of disks)*6+6	64 snapshots for each disk.	Longer protection circle and smaller protection granularity.	- A snapshot backup of a data disk for non-core businesses occurs at 00:00 every day, which retains backup data for over 2 months. - A snapshot backup of a data disk for core businesses occurs at a 4-

				hour interval, which retains backup data for over 10 days.
Automatic task policy	Hardcoded, triggered once every day, and unmodifiable.	The snapshot point in time, recurring days of week, and snapshot retention period can be customized. The disk quantity and related details associated with an automatic snapshot policy can be queried.	More flexible protection policy	- A user has 24 points in time available every day for taking snapshots and can take snapshots several times a day. - A user can choose any day from Monday to Sunday as the

		31		recurring day of week for taking 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 snapshot will

				be deleted).
Implementation principle	COW (Copy-On-Write)	ROW (Redirect-On-Write)	Mitigated performance impact of the snapshot task on business IO write	Invisible to users, supporting taking snapshots at any time.

Compare technical advantages

Compared with the snapshot feature of traditional storage products, Alibaba Cloud ECS Snapshot 2.0 has many advantages, detailed as below.

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 the storage device capacity upon first purchase, allowing only for meeting data protection needs for a few core services.
Scalability	Auto scaling, allowing the user to scale up and down according to their business scale with one click, and take effect within seconds.	Poor scalability, restrained by production and storage performance, available capacity, vendor support capabilities, among others. A scaling takes about 1-2 weeks.
Cost	Billed based on the actual amount of data changed in the user's business and the capacity used by snapshots.	Big upfront investment, involving software licenses, reserved space, and upgrade and maintenance expense - big waste of money.
Usability	24x7 online aftersales support.	Complex operations, greatly restrained by vendor support capabilities.

Application scenarios

As a simple and efficient data protection method, snapshot is recommended for the following

business scenarios:

Routine backup of system and data disks. You can back up business-critical data at regular intervals using snapshots to avoid data loss risks arisen from misoperations, attacks, viruses and others.

OS replacement. Before important operations such as upgrading application software or migrating business data, you need to create one or more snapshots. In case of any issues occurring during the upgrade or migration, you can restore timely to normal status using the snapshots.

Use 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.

Change history

- In November 2016, Germany data center went live.
- In November 2016, Generation III went live.
- In November 2016, Japan data center went live.
- In January 2016, the system disk was resized.
- In November 2015, Instance Generation II went live.
- In November 2015, the security group feature went live.
- In October 2015, West-USA Zone 1B went live.
- In September 2015, the image market was commercialized.
- In September 2015, Singapore data center went live.
- In September 2015, the efficient cloud disk went live.
- In August 2015, the tag grouping feature went live.
- In August 2015, the Virtual Private Cloud (VPC) went live.
- In June 2015, the image of Windows Server 2003 was deprecated.
- In May 2015, the shared image went live.
- In April 2015, disk resizing went live.
- In December 2014, the ephemeral SSD was officially commercialized.
- In October 2014, deployment of the Docker container application was allowed.
- In August 2014, Shenzhen data center went live.
- In August 2014, the independent cloud disk feature went live.
- In July 2014, the zone feature went live.
- In June 2014, the automatic snapshot feature went live.
- In May 2014, Hong Kong data center went live.
- In May 2014, the image market went live.

- In April 2014, Beijing data center went live.
- In April 2014, ECS API was officially launched.
- In July 2013, ECS's brand new user-defined image feature went live.
- In July 2011, the official website of Alibaba Cloud successfully went live and ECS went on sale
  to external customers.