Elastic Compute Service

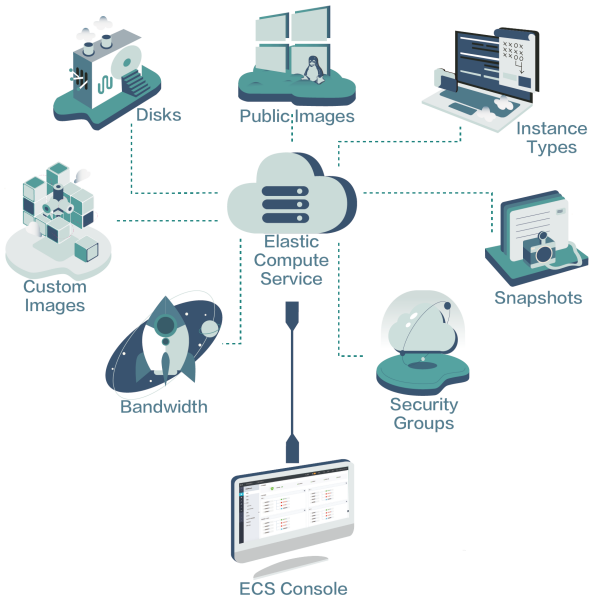
Product Introduction

Product Introduction

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 ECS console to configure the instance type, disks, operating system, and other affiliated resources for your ECS instance.



Advantages

ECS has the following advantages over traditional servers and virtual hosts:

Stability
ECS has 99.95% service availability and 99.9999999% data reliability. It also supports switchover, data snapshot backup and rollback, and system performance alarms.

Disaster recovery
Each data segment has multiple copies, which guarantees rapid restoration when one data segment is physically damaged.

Security
ECS supports security groups, 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 data centers ensure smooth and balanced access throughout the geographic region. Backbone data centers ensure high output bandwidth and dedicated bandwidth.

Low cost
Large one-time payments are not required. Flexible payment options and Pay-As-You-Go let you cope with business changes.

Controllability
As an ECS user, you have the permission of a super administrator. This allows you to completely control the operating system of ECS instances, resolve system problems using the management terminal, and perform operations such as environment deployment and software installation.

Ease of use
A variety of operating systems and applications are supported. Images can be deployed with the click of a button. You can quickly replicate the environment to multiple ECS instances for easy scaling. You can also create ECS instances in batches using custom images and disk snapshots.

API
API invocation management allows configuration of access to one or multiple servers with the security group feature, making development more convenient.

Features

ECS supports the following features:

Flexible instance configuration
Supports multiple instance generations, three instance type families, and dozens of instance types (ranging from 1-core 1 GiB to 56-core 480 GiB).

Multiple regions and zones
Allows instance creation in all regions, some of which have multiple zones.

Abundant image resources
Provides various image resources, including public images, custom images, and shared images, allowing quick operating system deployment and applications without installation.

Numerous operating systems
Supports multiple Windows and Linux operating systems.

Multiple storage methods
Provides three types of data storage disks (Basic Cloud Disks, Ultra Cloud Disks, and SSD Cloud Disks) and I/O-optimized instances.

Robust network and security

- Supports two network types (Classic Network and VPC), allowing network management in different dimensions.
- Supports two types of IP addresses (public and private IP addresses), allowing for Intranet interconnection and Internet access.
- Allows free activation of Alibaba Cloud Security products and provides network monitoring.

Convenient management
Provides multiple management methods, including the console, VNC, and APIs, ensuring complete control.

Flexible payment
Provides flexible payment methods (Subscription and Pay-As-You-Go).

Compared with Internet Data Centers (IDCs) and server vendors, Alibaba Cloud adopts more stringent IDC standards, server access standards, and O&M standards to ensure data reliability and high availability of cloud computing infrastructure and cloud servers.

In addition, each region of Alibaba Cloud consists of multiple zones. For higher availability, 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 higher-availability services in multiple regions and zones. Those services include disaster tolerance and backup, which are supported by Alibaba Cloud's mature solutions. Services can be switched smoothly within Alibaba Cloud's framework. For more information, refer to Alibaba Cloud's industry solutions. Alibaba Cloud's industry solutions support a variety of services, such as finance, E-commerce, and video services. Alibaba Cloud provides you with the following support services:

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

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 2007 and MTCS, which demand strict confidentiality of user data and user information as well as 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.

Alibaba Cloud VPC can connect your IDC through a leased line to build a hybrid cloud architecture. You can build 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 based on 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 need.

Secure: VPC features traffic isolation and attack isolation 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 as well as industries and enterprises not engaged in cloud computing.

Regions

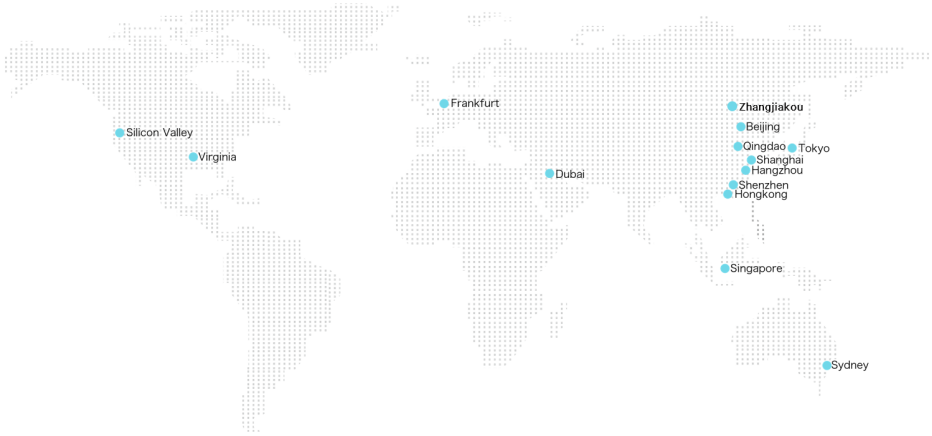
The following table lists the regions, corresponding cities, and Region IDs.

Regions in Mainland China

Region	China North 1	China North 2	China North 3	China East 1	China East 2	China South 1
City	Qingdao	Beijing	Zhangjiakou	Hangzhou	Shanghai	Shenzhen
RegionId	cn-qingdao	cn-beijing	cn-zhangjiakou	cn-hangzhou	cn-shanghai	cn-shenzhen

International regions

Region	Hong Kong	Asia Pacific SE 1	Asia Pacific SE 2	Asia Pacific NE 1	US West 1	US East 1	Germany 1	Middle East 1
City	Hong Kong	Singapore	Sydney	Tokyo	Silicon Valley	Virginia	Frankfurt	Dubai
RegionId	cn-hongkong	ap-southeast-1	ap-southeast-2	ap-northeast-1	us-west-1	us-east-1	eu-central-1	me-east-1



Zones

Zones are physical areas with independent power grids and networks in one region. The network latency for ECS instances within the same zone is shorter.

Intranet communication can take place between ECS instances in different zones of the same region, and fault isolation can be performed between zones. Whether ECS instances can be deployed in the same zone depends on the requirements for disaster recovery capabilities and network latency.

- If your applications require high disaster recovery capabilities, we suggest you deploy your

ECS instances in different zones of the same region;
- If your applications require low network latency between instances, we suggest you create your ECS instances in the same zone.

How to select a region

Regions in mainland China

Regions in mainland China include China East 1, China East 2, China North 1, China North 2, China North 3, and China South 1.

They offer BGP backbone network lines covering all provinces and municipalities in mainland China and providing stable and fast access within mainland China.

They are similar to each other in terms of infrastructure, BGP network quality, service quality, ECS operation, and configuration. In general cases, we recommend you select a data center closest to your end users to further speed up user access.

International regions

Alibaba Cloud’ s international regions are data centers outside mainland China. They offer access at international bandwidth, targeting areas outside mainland China. Access to these regions from mainland China may experience high latency. Therefore, they are not recommended for access from mainland China.

Asia Pacific

Hong Kong

The data center in Hong Kong offers access at international bandwidth, covering Hong Kong and Southeast Asia.
If you have business operation in Hong Kong and Southeast Asia, you can select this region.

Asia Pacific SE 1

The data center in Asia Pacific SE 1 is provided by Alibaba Cloud’ s partner, SingTel, a dominant operator in Southeast Asia. With highly reliable business expertise and maturity, the company is well positioned to serve users across the region.
If you have business operation in Hong Kong and Southeast Asia, you can select this region.

Asia Pacific SE 2

The data center in Asia Pacific SE 2 is located in Sydney.

If you have business operation in Australia, select the Asia Pacific SE 2 region.

Asia Pacific NE 1

The data center in Asia Pacific NE 1 is located in Tokyo, Japan.
If you have business operation in Japan, Northeast Asia, and South Korea, select the Asia Pacific NE 1 region.

North America and South America

US West 1

The data center in US West 1 is located in Silicon Valley. It is directly connected to the backbone networks of multiple American operators through BGP lines. In addition to the United States, this data center extends its reach to South America and Continental Europe.
If you have business operation in America and Continental Europe, select this region.

US East 1

The data center in US East 1 is located in Virginia of the United States.
If you have business operation in America and Continental Europe, select this region.

Germany 1

The data center in Germany 1 is located in Frankfurt.
If you have business operation in Continental Europe, select the Germany 1 region.

Middle East 1

The data center in Middle East 1 is located in Dubai.
If you have business operation in Middle East, select the Middle East 1 region.

Intranet communication between Alibaba Cloud products across regions

Intranet communication between Alibaba Cloud products that are not in the same region is not supported, which means:

- ECS instances in different regions cannot communicate with each other on the intranet.
- ECS instances and other products in different regions, such as ApsaraDB for RDS and OSS instances, cannot communicate with each other on the intranet.
- Server Load Balancer cannot be deployed for ECS instances in various regions.

About business license record filing

If you need to file your business license for record, pay attention to the following points:

- If your company is located in Beijing, select the **China North 2** region for the ECS instance you bought.
- If your company is located in Guangdong, select the **China South 1** region for the ECS instance you bought.

Notice: In mainland China, the approval requirements for business record filing vary depending on provincial communication management bureaus. For details, refer to the information published on the business record filing website of your local bureau.

Application 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 OSS and CDN, to provide advanced solutions.

ECS can be used in applications such as:

Official corporate websites and simple web applications
In 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 high-availability 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.

Instances

Overview

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.

An ECS instance is the minimal unit that can provide computing services for your business. It provides computing capabilities at a certain specification.

> 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.

Overview

Instance type families

ECS instances are categorized into multiple specification types, which are also called type families, based on the business and usage scenarios. In the same business scenario, you can select various type families.

In the same type family, there are multiple types based on the CPU and memory specifications.

Instance types

We define two basic attributes for an ECS instance type: the specifications of the CPU and the memory, including CPU model and clock speed. However, the attributes of a disk, an image, and the network service of an ECS instance must be defined at the same time, for the specific service form of the instance to be determined.

Instance type families and instance types

According to the release history and the business scenarios, Alibaba Cloud ECS instances are categorized into the following type families:

- The latest type families for various business scenarios, including:
  - Type families for enterprise-class computing on the x86-architecture, including:
    - sn2, general purpose type family
    - sn2ne, general purpose type family with enhanced network performance
    - sn1, compute optimized type family
    - sn1ne, compute optimized type family with enhanced network performance
    - se1, memory optimized type family
    - se1ne, memory optimized type family with enhanced network performance
    - d1, big data type family
    - d1ne, big data type family with enhanced network performance
    - i1, type family with ephemeral disks
    - c4, compute optimized type family with high clock speed
  - Type families for enterprise-class heterogeneous computing, including:
    - gn5, compute optimized type family with GPU
    - gn4, compute optimized type family with GPU
    - ga1, visualization compute type family with GPU
    - f1, compute optimized type family with FPGA
  - Type families for beginners, computing on the x86-architecture
- Type families of previous generations for beginners, computing on the x86-architecture

The latest type families

All the ECS instances of the latest type families are I/O-optimized. They support the following disk types:

- SSD cloud disks
- Ultra cloud disks

Instances of the latest type families are categorized into the following type families based on the business scenarios.

Type families for enterprise-class computing on the x86-architecture

sn2, general purpose type family

Features

- vCPU : Memory = 1:4

- Stable computing performance
- 2.5 GHz Intel Xeon, E5-2682 v4 (Broadwell), or E5-2680 v3 (Haswell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
    - Enterprise-class 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)	Ephemeral storage (GiB)	Intranet bandwidth (Gbit/s)	Packet forwarding rate (10 thousand PPS)	NIC queues
ecs.sn2.medium	2	8	N/A	0.5	5	1
ecs.sn2.large	4	16	N/A	0.8	10	1
ecs.sn2.xlarge	8	32	N/A	1.5	20	1
ecs.sn2.3xlarge	16	64	N/A	3	40***	2
ecs.sn2.7xlarge	32	128	N/A	6	80**	3
ecs.sn2.13xlarge	56	224	N/A	10	120*	4

* Testing conditions: Four queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
** Testing conditions: Three queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
*** Testing conditions: Two queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

> You can change the configurations of an instance between any two type families of sn2, sn2ne, sn1, sn1ne, se1, and se1ne, and within the same instance type family.

sn2ne, general purpose type family with enhanced network performance

Features

- vCPU : Memory = 1:4

- Ultra high packet forwarding rate
- Stable computing performance
- 2.5 GHz Intel Xeon, E5-2682 v4 (Broadwell), or E5-2680 v3 (Haswell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
    - Scenarios of receiving and transmitting a large volume of packets, such as video bullet screen and retransmission of telecommunication services
    - Enterprise-class 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)	Ephemeral storage (GiB)	Intranet bandwidth (Gbit/s)	Packet forwarding rate (10 thousand PPS)	NIC queues
ecs.sn2ne.large	2	8	N/A	0.5	12	2
ecs.sn2ne.xlarge	4	16	N/A	0.8	30	2
ecs.sn2ne.2xlarge	8	32	N/A	1.5	100*	4
ecs.sn2ne.4xlarge	16	64	N/A	3	160*	4
ecs.sn2ne.8xlarge	32	128	N/A	6	250*	8
ecs.sn2ne.14xlarge	56	224	N/A	10	450*	14

* Testing conditions: No more than (vCPU core number/4) queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

sn1, compute optimized type family

Features

- vCPU : Memory = 1:2
- Stable computing performance
- 2.5 GHz Intel Xeon, E5-2682 v4 (Broadwell), or E5-2680 v3 (Haswell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)

- Ideal for:
  - Web front-end servers
  - Front ends of Massively Multiplayer Online (MMO) games
  - Data analysis, batch compute, and video coding
  - High performance science and engineering applications

Instance types

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)	Intranet bandwidth (Gbit/s)	Packet forwarding rate (10 thousand PPS)	NIC queues
ecs.sn1.medium	2	4	N/A	0.5	5	1
ecs.sn1.large	4	8	N/A	0.8	10	1
ecs.sn1.xlarge	8	16	N/A	1.5	20	1
ecs.sn1.3xlarge	16	32	N/A	3	40*	2
ecs.sn1.7xlarge	32	64	N/A	6	80**	3

* Testing conditions: Two queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
** Testing conditions: Three queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

> You can change the configurations of an instance between any two type families of sn2, sn2ne, sn1, sn1ne, se1, and se1ne, and within the same instance type family.

sn1ne, compute optimized type family with enhanced network performance

Features

- vCPU : Memory = 1:2
- Ultra high packet forwarding rate
- Stable computing performance
- 2.5 GHz Intel Xeon, E5-2682 v4 (Broadwell), or E5-2680 v3 (Haswell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
  - Scenarios of receiving and transmitting a large volume of packets, such as video bullet screen and retransmission of telecommunication services

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

Instance types

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)	Intranet bandwidth (Gbit/s)	Packet forwarding rate (10 thousand PPS)	NIC queues
ecs.sn1ne.large	2	4	N/A	0.5	12	2
ecs.sn1ne.xlarge	4	8	N/A	0.8	30	2
ecs.sn1ne.2xlarge	8	16	N/A	1.5	100*	4
ecs.sn1ne.4xlarge	16	32	N/A	3	160*	4
ecs.sn1ne.8xlarge	32	64	N/A	6	250*	8

* Testing conditions: No more than (vCPU core number/4) queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

> You can change the configurations of an instance between any two type families of sn2, sn2ne, sn1, sn1ne, se1, and se1ne, and within the same instance type family.

se1, memory optimized type family

Features

- vCPU : Memory = 1:8
- Stable computing performance
- 2.5 GHz Intel Xeon, E5-2682 v4 (Broadwell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
  - High performance databases and memory databases
  - Data analysis and mining, and distributed memory cache
  - Hadoop, Spark, and other enterprise-class applications that require large volume of memory

Instance types

Instance	vCPU	Memory	Ephemer	Intranet	Packet	NIC

type		(GiB)	al storage (GiB)	bandwidth (Gbit/s)	forwarding rate (10 thousand PPS)	queues
ecs.se1.large	2	16	N/A	0.5	5	1
ecs.se1.xlarge	4	32	N/A	0.8	10	1
ecs.se1.2xlarge	8	64	N/A	1.5	20	1
ecs.se1.4xlarge	16	128	N/A	3	40***	2
ecs.se1.8xlarge	32	256	N/A	6	80**	3
ecs.se1.14xlarge	56	480	N/A	10	120*	4

* Testing conditions: Four queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
** Testing conditions: Three queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
*** Testing conditions: Two queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

> You can change the configurations of an instance between any two type families of sn2, sn2ne, sn1, sn1ne, se1, and se1ne, and within the same instance type family.

se1ne, memory optimized type family with enhanced network performance

Features

- vCPU : Memory = 1:8
- Ultra high packet receive and forwarding rate
- Stable computing performance
- 2.5 GHz Intel Xeon, E5-2682 v4 (Broadwell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
    - Scenarios of receiving and transmitting a large volume of packets, such as video bullet screen and retransmission of telecommunication services
    - High performance databases and memory databases
    - Data analysis and mining, and distributed memory cache
    - Hadoop, Spark, and other enterprise-class applications that require large volume of memory

Instance types

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)	Intranet bandwidth (Gbit/s)	Packet forwarding rate (10 thousand PPS)	NIC queues
ecs.se1ne.large	2	16	N/A	0.5	12	2
ecs.se1ne.xlarge	4	32	N/A	0.8	30	2
ecs.se1ne.2xlarge	8	64	N/A	1.5	100*	2
ecs.se1ne.4xlarge	16	128	N/A	3	160*	4
ecs.se1ne.8xlarge	32	256	N/A	6	250*	8
ecs.se1ne.14xlarge	56	480	N/A	10	450*	14

* Testing conditions: No more than (vCPU core number/4) queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

You can change the configurations of an instance between any two type families of sn2, sn2ne, sn1, sn1ne, se1, and se1ne, and within the same instance type family.

d1, big data type family

Features

- High-volume ephemeral SATA HDD disks with high I/O throughput and a maximum of 20 Gbit/s of intranet bandwidth for a single instance
- vCPU : Memory = 1:4, designed for big data scenarios
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
    - Hadoop MapReduce, HDFS, Hive, HBase, and so on
    - Spark in-memory computing, MLib, and so on
    - For those enterprises that require big data computing and storage analysis, such as enterprises in Internet and finance industries, to store and compute massive data
    - ElasticSearch, logs, and so on

Instance types

Instance	vCPU	Memory	Ephemer	Intranet	Packet	NIC

type		(GiB)	al storage (GiB)	bandwidth (Gbit/s)	forwarding rate (10 thousand PPS)	queues
ecs.d1.2xlarge	8	32	4 * 5587	3	30	1
ecs.d1.4xlarge	16	64	8 * 5587	6	60***	2
ecs.d1.6xlarge	24	96	12 * 5587	8	80***	2
ecs.d1.8xlarge	32	128	16 * 5587	10	100**	4
ecs.d1.14xlarge	56	224	28 * 5587	17	140*	6

* Testing conditions: Six queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
** Testing conditions: Four queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
*** Testing conditions: Two queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

> You can change the configurations of an instance between d1 and d1ne.
>
> For more information of d1 type family, see FAQ on d1 and d1ne.

d1ne, big data type family with enhanced network performance

Features

- High-volume ephemeral SATA HDD disks with high I/O throughput and a maximum of 40 Gbit/s of intranet bandwidth for a single instance
- vCPU : Memory = 1:4, designed for big data scenarios
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
  - Hadoop MapReduce, HDFS, Hive, HBase, and so on
  - Spark in-memory computing, MLib, and so on
  - For those enterprises that require big data computing and storage analysis, such as enterprises in Internet and finance industries, to store and compute massive data
  - ElasticSearch, logs, and so on

Instance types

Instance	vCPU	Memory	Ephemer	Intranet	Packet	NIC

type		(GiB)	al storage (GiB)	bandwidth (Gbit/s)	forwarding rate (10 thousand PPS)	queues
ecs.d1ne.2xlarge	8	32	4 * 5587	6	80*	2
ecs.d1ne.4xlarge	16	64	8 * 5587	12	160*	4
ecs.d1ne.6xlarge	24	96	12 * 5587	16	200*	6
ecs.d1ne.8xlarge	32	128	16 * 5587	20	250*	8
ecs.d1ne.14xlarge	56	224	28 * 5587	40	450*	14

* Testing conditions: No more than (vCPU core number/4) queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

You can change the configurations of an instance between d1 and d1ne.

For more information of d1 type family, see FAQ on d1 and d1ne.

i1, type family with ephemeral disks

Features

- High-performance ephemeral NVMe SSD disks: supporting high IOPS and I/O throughput and low latency.
- vCPU : Memory = 1:4, designed for high performance databases
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
  - OLTP and high performance relational databases
  - NoSQL databases, such as Cassandra and MongoDB
  - Search applications, such as Elastic Search

Instance types

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)	Intranet bandwidth (Gbit/s)	Packet forwarding rate (10 thousand PPS)	NIC queues
ecs.i1.xlarge	4	16	2 * 104	0.8	10	1

ecs.i1.2xlarge	8	32	2 * 208	1.5	20	1
ecs.i1.4xlarge	16	64	2 * 416	3	40***	2
ecs.i1-c5d1.4xlarge	16	64	2 * 1456	3	40***	2
ecs.i1-c15d2.6xlarge	24	96	2 * 1456	4.5	60***	2
ecs.i1.8xlarge	32	128	2 * 832	6	80**	3
ecs.i1-c10d1.8xlarge	32	128	2 * 1456	6	80**	3
ecs.i1.14xlarge	56	224	2 * 1456	10	120*	4

* Testing conditions: Four queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
** Testing conditions: Three queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
*** Testing conditions: Two queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

> You cannot change configurations of i1 instances.

c4, compute optimized type family with high clock speed

Features

- Stable computing performance
- 3.2 GHz Intel Xeon E5-2667 v4 (Broadwell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
  - High performance Web front-end servers
  - High performance science and engineering applications
  - Massively Multiplayer Online (MMO) games and video coding

Instance types

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)	Intranet bandwidth (Gbit/s)	Packet forwarding rate (10 thousand	NIC queues

					PPS)	
ecs.c4.xlarge	4	8	N/A	1.5	20	1
ecs.c4.2xlarge	8	16	N/A	3	40	1
ecs.c4.4xlarge	16	32	N/A	6	80**	2
ecs.cm4.xlarge	4	16	N/A	1.5	20	1
ecs.cm4.2xlarge	8	32	N/A	3	40	1
ecs.cm4.4xlarge	16	64	N/A	6	80**	2
ecs.cm4.6xlarge	24	96	N/A	10	120*	4
ecs.ce4.xlarge	4	32	N/A	1.5	20	1

* Testing conditions: Four queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
** Testing conditions: Two queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

You can change the configurations of an instance within c4.

Type families for enterprise-class heterogeneous computing

gn5, compute optimized type family with GPU

Features

- NVIDIA P100 GPU processors
- No fixed ratio of CPU to memory
- High performance ephemeral SSD disks
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
    - Deep learning
    - Scientific computing, such as computational fluid dynamics, computational finance, genomics, and environmental analysis
    - High performance computing, rendering, multi-media coding and decoding, and other server-side GPU compute workloads

Instance types

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)	GPU	Intranet bandwidth (Gbit/s)	Packet forwarding rate (10 thousand PPS)	NIC queues
ecs.gn5-c4g1.xlarge	4	30	440	1 * NVIDIA P100	3	30	1
ecs.gn5-c8g1.2xlarge	8	60	440	1 * NVIDIA P100	3	30	1
ecs.gn5-c4g1.2xlarge	8	60	880	2 * NVIDIA P100	5	100**	2
ecs.gn5-c8g1.4xlarge	16	120	880	2 * NVIDIA P100	5	100**	2
ecs.gn5-c8g1.8xlarge	32	240	1760	4 * NVIDIA P100	10	200*	8
ecs.gn5-c8g1.14xlarge	56	480	3520	8 * NVIDIA P100	25	400*	14

* Testing conditions: No more than (vCPU core number/4) queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
** Testing conditions: Two queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

You can change the configurations of an instance within gn5.

gn4, compute optimized type family with GPU

Features

- NVIDIA M40 GPU processors
- No fixed ratio of CPU to memory
- High performance ephemeral SSD disks
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors

- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
  - Deep learning
  - Scientific computing, such as computational fluid dynamics, computational finance, genomics, and environmental analysis
- High performance computing, rendering, multi-media coding and decoding, and other server-side GPU compute workloads

Instance types

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)	GPU	Intranet bandwidth (Gbit/s)	Packet forwarding rate (10 thousand PPS)	NIC queues
ecs.gn4-c4g1.xlarge	4	30	N/A	1 * NVIDIA M40	3	40	1
ecs.gn4-c8g1.2xlarge	8	30	N/A	1 * NVIDIA M40	3	40	1
ecs.gn4.8xlarge	32	48	N/A	1 * NVIDIA M40	5	50**	3
ecs.gn4-c4g1.2xlarge	8	60	N/A	2 * NVIDIA M40	5	50	1
ecs.gn4-c8g1.4xlarge	16	60	N/A	2 * NVIDIA M40	5	50***	2
ecs.gn4.14xlarge	56	96	N/A	2 * NVIDIA M40	10	120*	4

* Testing conditions: Four queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
** Testing conditions: Three queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
*** Testing conditions: Two queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

You can change the configurations of an instance within gn4.

ga1, visualization compute type family with GPU

Features

- AMD S7150 GPU processors
- vCPU : Memory = 1:2.5
- High performance ephemeral SSD disks
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
  - Rendering, multimedia coding and decoding
  - Machine learning, high-performance computing, and high performance databases
  - Other server-end business scenarios that require powerful concurrent floating-point compute capabilities

Instance types

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)	GPU	Intranet bandwidth (Gbit/s)	Packet forwarding rate (10 thousand PPS)	NIC queues
ecs.ga1.2xlarge	8	20	1 * 175	0.5 * AMD S7150	1.5	15	1
ecs.ga1.4xlarge	16	40	1 * 350	1 * AMD S7150	3	40***	2
ecs.ga1.8xlarge	32	80	1 * 700	2 * AMD S7150	6	80**	3
ecs.ga1.14xlarge	56	160	1 * 1400	4 * AMD S7150	10	120*	4

* Testing conditions: Four queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
** Testing conditions: Three queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.
*** Testing conditions: Two queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

You can change the configurations of an instance within ga1.

f1, compute optimized type family with FPGA

Features

- Intel Arria 10 GX 1150 FPGA
- vCPU : Memory = 1:7.5
- 2.5 GHz Intel Xeon E5-2682 v4 (Broadwell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- Ideal for:
  - Deep learning and reasoning
  - Genomics research and finance analysis
  - Computational workloads, such as real-time video processing and security

Instance types

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)	FPGA	Intranet bandwidth (Gbit/s)	Packet forwarding rate (10 thousand PPS)	NIC queues
ecs.f1-c8f1.2xlarge	8	60	440	Intel Arria 10 GX 1150	3	40	1
ecs.f1-c8f1.4xlarge	16	120	880	Intel Arria 10 GX 1150 * 2	5	70*	2

* Testing conditions: Two queues for NICs are enabled and CentOS 7.3 is used. If you want to adjust multiple-queue for NIC, restart the instance.

You cannot change configurations of f1 instances.

Type families for beginners, computing on the x86-architecture

Features

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

Instance types

Type family	Features	vCPU : Memory	Idea for
xn4	Compact shared instances	1:1	- Front ends of Web applications - Light load applications and microservices - Applications for development or testing environments
n4	General shared instances	1:2	- Websites and Web applications - Development environment, building servers, code repositories, microservices, and testing and staging environment - Lightweight enterprise applications
mn4	Balanced shared instances	1:4	- Websites and Web applications - Lightweight

			databases and cache - Integrated applications and lightweight enterprise services
e4	Memory shared instances	1:8	- Applications that require large volume of memory - Lightweight databases and cache

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 (GiB)	Ephemeral storage (GiB)
ecs.xn4.small	1	1	N/A

n4

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)
ecs.n4.small	1	2	N/A
ecs.n4.large	2	4	N/A
ecs.n4.xlarge	4	8	N/A
ecs.n4.2xlarge	8	16	N/A
ecs.n4.4xlarge	16	32	N/A
ecs.n4.8xlarge	32	64	N/A

mn4

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)
ecs.mn4.small	1	4	N/A
ecs.mn4.large	2	8	N/A
ecs.mn4.xlarge	4	16	N/A
ecs.mn4.2xlarge	8	32	N/A
ecs.mn4.4xlarge	16	64	N/A
ecs.mn4.8xlarge	32	128	N/A

e4

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)
ecs.e4.small	1	8	N/A

Type families of previous generations for beginners, computing on the x86-architecture

Features

- 2.5 GHz Intel Xeon E5-2680 v3 (Haswell) processors
- The network performance of an instance matching the computing type (the more advanced the computing type, the more powerful the network performance)
- I/O-optimized
- Supporting the following disk types:
    - SSD cloud disks
    - Ultra cloud disks

Instance types

Type family	Features	vCPU : Memory	Idea for
n1	General shared instances	1:2	- Small and medium-sized web servers - Batch processing - Distributed analysis - Advertisem

			ent services
n2	Balanced shared instances	1:4	- Medium-sized Web servers - Batch processing - Distributed analysis - Advertisement services - Hadoop clusters
e3	Memory shared instances	1:8	- Cache, Redis - Search - Memory databases - Databases with high I/O, for example, Oracle and MongoDB - Hadoop clusters - Computing scenarios that involve massive data processing

n1

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)
ecs.n1.tiny	1	1	N/A
ecs.n1.small	1	2	N/A
ecs.n1.medium	2	4	N/A

ecs.n1.large	4	8	N/A
ecs.n1.xlarge	8	16	N/A
ecs.n1.3xlarge	16	32	N/A
ecs.n1.7xlarge	32	64	N/A

n2

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)
ecs.n2.small	1	4	N/A
ecs.n2.medium	2	8	N/A
ecs.n2.large	4	16	N/A
ecs.n2.xlarge	8	32	N/A
ecs.n2.3xlarge	16	64	N/A
ecs.n2.7xlarge	32	128	N/A

e3

Instance type	vCPU	Memory (GiB)	Ephemeral storage (GiB)
ecs.e3.small	1	8	N/A
ecs.e3.medium	2	16	N/A
ecs.e3.large	4	32	N/A
ecs.e3.xlarge	8	64	N/A
ecs.e3.3xlarge	16	128	N/A

> You can change the configurations among the three shared instance type families (n1, n2, and e3), and within the same instance type family.

If you are using t1, t2, s1, s2, s3, m1, m2, c1, or c2, see Generation I instance types.

Instance life cycle

The life cycle of an instance begins when you create the instance. The life cycle ends when the instance is released, either after a monthly or yearly subscription expires, when you manually release a Pay-As-You-Go instance, or because of an outstanding payment.

Instance statuses

The life cycle of an instance involves several inherent instance statuses, as listed in the following table:

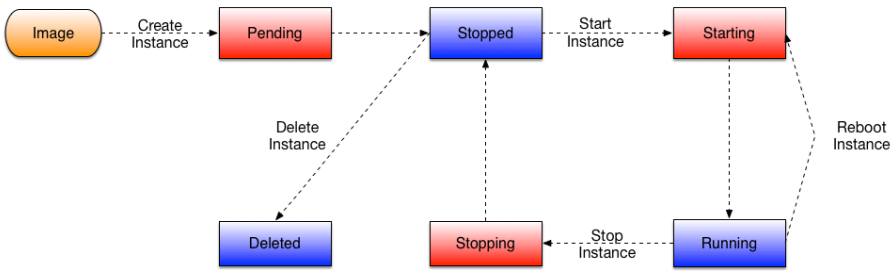
Status	Status property	Description	Corresponding API status
Preparing*	Intermediate status	After an instance is created, it remains in this status before running.	Pending
Created*	Stable status	An instance is in this status when it has been created and is awaiting start up.	Stopped
Starting*	Intermediate status	An instance is in this status after it is started or restarted in the console or using API until it is running.	Starting
Running	Stable status	The instance is operating properly and can accommodate your business needs.	Running
Stopping*	Intermediate status	An instance is in this status after the stop operation is performed in the console or through API but before it actually stops.	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 using API until it is running.	Stopped
Replacing System Disk	Intermediate status	An instance is in this status after the operating system is replaced or another	Stopped

		such operation is performed in the console or using API until it is running.	
Expired	Stable status	The yearly/monthly instance subscription has expired because it has not been properly renewed. The Pay-As-You-Go instances have expired because they are in arrears. **Note:** After expiration, both the yearly/monthly and Pay-As-You-Go instances will continue running for 15 days, and data will be retained for an extra 15 days, after which the instances will be released and the data will be removed permanently.	Stopped

* If an instance remains in the Preparing, Created, Starting, Stopping, Re-initializing, or Replacing System Disk status for a long time, it has encountered an exception.

API status chart

This flowchart describes the corresponding relationships between console statuses and API statuses. The API status chart is shown below.



Disks

Overview

An ECS disk can be used jointly or separately to meet the requirements of different application scenarios. ECS disks are categorized into ephemeral SSD disks and cloud disks. Compared with ephemeral SSD disks, cloud disks are more reliable as they use a triplicate distributed system to provide block-level data storage for ECS instances, ensuring 99.9999999% data reliability. Cloud disks are categorized as one of the following:

SSD cloud disks
Ideal for I/O-intensive applications, and provide stable and high random IOPS performance.

Ultra cloud disks
Ideal for medium I/O load application scenarios and provide a storage performance of up to 3,000 random IOPS for ECS instances.

Basic cloud disks
Ideal for least I/O-intensive application scenarios and provide an I/O performance of several hundred IOPS for ECS instances.

Note: For detailed instructions on attaching a disk, refer to Attach a data disk from Elastic Compute Service – User Guide.

Disk comparison

The following table describes the features and typical application scenarios for different types of cloud disks.

Item	SSD cloud disk	Ultra cloud disk	Basic cloud disk
Maximum capacity	32768 GB	32768 GB	2000 GB
Maximum IOPS	20000	3000	Several hundreds
Maximum throughput	300 MBps	80 MBps	30 MBps
Performance calculation formula	IOPS = min{1200+30*capacity, 20000} Throughput = min{80 +	IOPS = min{1000+6*capacity, 3000} Throughput = min {50+0.1*capacity,	N/A

	0.5*capacity, 300} MBps	80} MBps	
Access latency	0.5 ms–2 ms	1 ms–3 ms	5 ms–10 ms
Data reliability	99.9999999%	99.9999999%	99.9999999%
API name	cloud_ssd	cloud_efficiency	cloud
Price*	$0.15 USD/GB/month	$0.08 USD/GB/month	$0.05 USD/GB/month
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

* Prices shown are for the US West region. For more information, see ECS Price at https://www.alibabacloud.com/product/ecs#pricing.

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=/dev/[device] -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=/dev/[device] -name=Rand_Read_Testing
- Test writing throughput ：
  fio -direct=1 -iodepth=64 -rw=write -ioengine=libaio -bs=64k -size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=/dev/[device] -name=Write_PPS_Testing
- Test reading throughput ：
  fio -direct=1 -iodepth=64 -rw=read -ioengine=libaio -bs=64k -size=1G -numjobs=1 -runtime=1000 -group_reporting -filename=/dev/[device] -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 requests 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=/dev/[device]	The output path and name of the test files or device. Testing naked disks can obtain more accurate performance data, but will damage the file structure. Make sure that you back up your data before testing.
-name=Rand_Write_Testing	The name of the testing task.

Disk categories and application scenarios

SSD cloud disks

Product features

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

High random I/O performance

The maximum random read/write IOPS is 20,000. The base is 1200 IOPS, and each GB of

capacity provides 30 random IOPS. For example, a 100 GB SSD cloud disk can provide 4,200 IOPS, and a 334 GB SSD cloud disk can provide 11,220 IOPS.

High throughput

The maximum throughput is 300 MBps. The throughput of an SSD cloud disk can be determined using the equation min {80 + 0.5*disk_size, 300} MBps.

High data reliability

SSD cloud disks adopt a distributed, triplicate mechanism to provide 99.9999999% data reliability.

Large storage capacity

A single SSD cloud disk provides up to 32,768 GB storage space.

Independent attaching

SSD cloud disks can be attached to any ECS instance in the same zone.

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.

Performance baselines

Block size	Maximum IOPS	Maximum throughput
4/8 KB	20,000	N/A
16 KB	17,200	256 MBps
32 KB	9,600	256 MBps
64 KB	4,800	N/A

Application Scenarios

SSD cloud disks have stable and 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 to large development and testing environments with high requirements for data

reliability.

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:

High random I/O performance

The maximum random read/write IOPS is 3,000. The random read/write IOPS is initially 1,000 and increases by 6 IOPS for each GB. For example, a 250 GB ultra cloud disk features 2,500 random read/write IOPS.

High throughput

The maximum throughput is 80 MBps. The throughput is initially 50 MBps and increases by 0.1 MBps for each GB. For example, a 250 GB ultra cloud disk features a throughput of 75 MBps.

High data reliability

Ultra cloud disks adopt a distributed, triplicate mechanism to provide 99.9999999% data reliability.

Large storage capacity

A single ultra cloud disk provides up to 32768 GB storage space.

Independent attaching

Ultra cloud disks can be attached to any ECS instance in the same zone.

Application Scenarios

Ultra cloud disks are applicable to the following scenarios:

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

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:

High random I/O performance

The maximum random read/write IOPS is of several hundreds.

High throughput

The maximum throughput is 30 MBps–40 MBps.

High data reliability

Disks adopt a distributed triplicate mechanism provides 99.9999999% data reliability.

Large storage capacity

A single basic cloud disk provides up to 2,000 GB storage space.

Independent attaching

Basic cloud disks can be attached to any ECS instance in the same zone.

Application Scenarios

Basic cloud disks are applicable to the following scenarios:

- Applicable to scenarios in which data is not frequently accessed, or which have low I/O loads. If an application requires higher I/O performance, using an SSD cloud disk is recommended.
- 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

When ECS users perform read and write operations onto virtual disks, the operations are translated into corresponding processes on the files stored in the Alibaba Cloud Distributed File System. Alibaba Cloud provides a flat storage space, in which the linear addresses are divided into chunks, also referred to as slices. Alibaba Cloud employs a certain strategy to create three copies for each chunk and stores these copies on different nodes, ensuring the reliability of user data.

Principles of triplicate technology

The Alibaba Cloud data storage system consists of three roles: Master, Chunk Server, and Client. The write operation of an ECS user goes through several conversions and is executed by the Client. The procedure is as follows:

1. The Client calculates the chunk corresponding to a given writing operation.
2. The Client sends a request to the Master for the storage location of the three copies of the chunk.
3. The Client sends writing requests to the 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.

The distribution strategy of the Master is decided based on an overall consideration of the following:

- Disk use conditions of all Chunk Servers in the cluster.
- Distribution of the Chunk Servers under different kinds of switch racks.
- The power supply.
- The instrument load.

This strategy ensures 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 three. If this occurs, the Master initiates the copy mechanism to copy data between Chunk Servers, making three valid copies of all Chunks in the cluster.

In sum, for the data on the cloud disk, all user operations and data addition or modification will be synchronized to the three copies. This mode ensures the reliability and consistency of user data.

To prevent data losses caused by virus infection or cyber-attacks, we recommend that you use the triplicate technology with other protection methods, such as taking snapshots.

Ephemeral SSDs

Ephemeral SSD disks use the local storage of physical machines at which the instances are located. This type of storage provides block-level data access capabilities to instances. It 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 SSD disks after purchasing.
- Since the ephemeral SSD uses the local disk of a physical server, it does not support independent attaching/detaching.

Product features

Ephemeral SSD disks provide the following features:

Low latency
Under normal conditions, the access latency is in microseconds.

High random I/O performance
The maximum random IOPS is 12,000.

High throughput
The maximum I/O throughput is 300 MBps.

Large storage capacity
A single ephemeral SSD provides a maximum storage space of 800 GB.

Application Scenarios

Ephemeral SSD disks are applicable to distributed, I/O-intensive applications with redundancy and scenarios that do not require high data reliability, such as the following:

Distributed applications

NoSQL and MPP data warehouse, distributed file systems, and other I/O-intensive applications have their own distributed data redundancy. Ephemeral SSD disks can provide low latency, high random I/O, and high throughput I/O performance.

Logs for large online applications

Large online applications can produce large amounts of log data and require high-performance storage. At the same time, the log data does not require highly reliable storage.

As the 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

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

If you need to transmit data between two ECS instances in the same region, you should use an intranet connection. Intranet connections can also be used to connect RDS, Server Load Balancer, and OSS instances. The internet speed of these instances is based on a gigabit shared bandwidth environment. At present, you can also use a direct intranet connection to link RDS, Server Load Balancer, and OSS instances with ECS instances in the same region.

For ECS instances in the intranet:

For instances of Classic network:

- Intranet communication is by default used only for instances in the same security group of the same account in the same region.

- An intranet communication can also be used for instances in the same security group of the same account and region but of different zones, even if the intranet IP addresses 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 Application scenarios of security group from ECS User Guide.

For instances of VPC network:

- Intranet communication is by default used only for instances in the same security group of the same account and same VPC network in the same region.
- An intranet communication can also be used for instances of the same account and region but of different VPC networks only if you use ExpressConnect to authorize their intranet communication. For more information, see Application scenarios from Product Introduction to ExpressConnect.

The intranet IP addresses of instances cannot be modified or changed.

Intranet and Internet addresses of instances do not support virtual IP (VIP) configuration.

Instances of different network types cannot communicate with each other in intranet.

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 and private IP addresses.

Private IP addresses

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

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

Communication traffic through private IP addresses 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

Each instance is by default configured with a public network interface card. Unlike private IP addresses, public IP addresses are optional. If you select a public network bandwidth greater than 0 Mbps when purchasing an instance, a public IP address will be allocated during creation of the instance.

Regardless of your selected billing method, you must select a public network bandwidth limit. The bandwidth limit you select will determine the limit of the outgoing bandwidth for the public network card.

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

A security group is a virtual firewall that provides stateful packet inspection (SPI). Security groups are used to set network access control for one or more ECSs. As an important means of security isolation, 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 up to five security groups.
- Each user can have up to 100 security groups.
- Adjusting security groups will not affect the continuity of user 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).
    - Classic Network type instances can join security groups on classic networks in the same region.
    - VPC type instances 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 inbound and outbound directions.

You can authorize or delete 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, make sure security group rules are simple. If you associate an ECS instance with multiple security groups, up to hundreds of rules may apply to the instance, which may cause connection errors when you access the instance.

Security group rule restrictions

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

Handling network interrupts with a single CPU is prone to bottlenecks. You can route NIC interrupts in the ECS instances to different CPUs for processing. In the network PPS and network bandwidth tests, the solution using two queues can improve the performance by 50% to 100% compared to that using only one queue. A solution that uses four queues provides much more significant performance enhancement.

ECS instance types supporting multi-queue

Refer to Instance generations and type families to find whether an instance type supports multi-queue and the number of queues.

Images supporting multi-queue

At present, among the public images officially provided by Alibaba Cloud, the ones shown in the following table support multi-queue. Whether an image supports multi-queue is not related with the memory address width of the operating system.

Image name	Support multi-queue?	Notes
Windows 2012 R2	Yes	Not available yet. You must be invited for test.

Windows 2016	Yes	Not available yet. You must be invited for test.
CentOS 7.2	Yes	None
CentOS 6.8	Yes	None
Ubuntu 16.04	Yes	None
Ubuntu 14.04	Yes	None
Debian 8.6	Yes	None
SUSE Linux Enterprise Server 12 SP1	Yes	None
OpenSUSE 13.1	Yes	None
CoreOS	Yes	None

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

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

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

- Check whether the NIC supports multi-queue. Run the command: ethtool -l eth0.
- Enable multi-queue for the NIC. Run the command: ethtool -L eth0 combined 2.

If you are using more than one NIC, make the configuration for each NIC.

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

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

We recommend that you enable the irqbalance service so that the system can automatically adjust the allocation of the NIC interrupts on multiple CPU cores. Run the command:

systemctl start irqbalance. This feature is enabled by default in CentOS 7.2.

If the network performance improvement is not up to your expectation when the multi-queue feature is enabled, you can enable the RPS feature. Refer to the following Shell script.

```
    #!/bin/bash
cpu_num=$(grep -c processor /proc/cpuinfo)
quotient=$((cpu_num/8))
if [ $quotient -gt 2 ]; then
quotient=2
elif [ $quotient -lt 1 ]; then
quotient=1
fi
for i in $(seq $quotient)
do
cpuset="${cpuset}f"
done

for rps_file in $(ls /sys/class/net/eth*/queues/rx-*/rps_cpus)
do
echo $cpuset > $rps_file
done
```

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

Note: At present, we are inviting Windows users to test the performance improvement.

Windows systems will see improved network performance after using multi-queue for NICs, but the effect is not as good as it is in the Linux system.

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

Use the following steps to install the driver for Windows systems.

Open a ticket to request and download the driver installation package.

Unzip the driver installation package. You will see several folders. For Windows 2012/2016 systems, use the driver under the Win8/amd64 folder.

Upgrade the NIC driver:

i. Select **Device Manager** > **Network adapters**.
ii. Right click **Red Hat VirtIO Ethernet Adapter** and select **Update Driver...**
iii. Select the Win8/admin64 directory of the driver directory that you just unzipped,

and update the driver.

After the driver upgrade is done, we recommend that you restart the Windows system.

Now you can start using the multi-queue feature for NICs.

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.

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 maintain a consistent environment and application deployment across multiple regions.

Snapshots

Overview

A snapshot is a copy of data on a disk at a certain point in time. Scheduled creation of disk snapshots ensures continuous operation of your business. Snapshot is a simple and efficient data protection method, and is recommended for the following scenarios:

Routine backup of system and data disks

You can back up business-critical data at regular intervals using snapshots to prevent data loss from misoperations, attacks, and viruses.

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

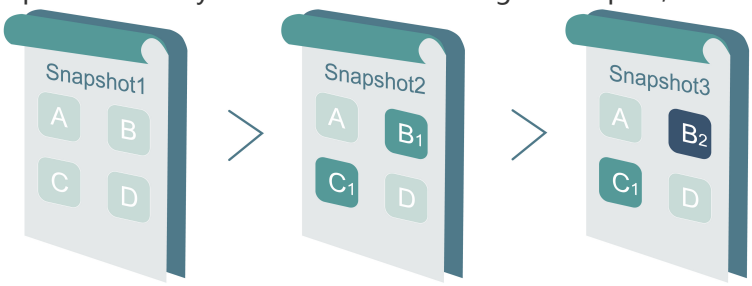
You can take snapshots of production data to provide close-to-real-time production data for data mining, report queries, and developing and testing applications. You can also take snapshots to reuse data on a disk as basic data for another disk.

Restoring data stored on cloud disks

Cloud disks provide a secure storage method to ensure that your stored content will not be lost. However, if the data stored on a cloud disk is incorrect due to an application error, or the data is maliciously tampered by hackers through an application vulnerability, snapshots ensure that your data can be restored to the desired state.

Incremental snapshot mechanism

Snapshots on Alibaba Cloud are taken using an incremental method. In this method, two snapshots are compared and only the data that has changed is copied, as shown in the following image:



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 will be copied to the snapshot. In this example:

- In Snapshot 1, all data on the disk is copied since it is the first disk snapshot.
- Snapshot 2 only copies the changed data blocks B1 and C1. Data blocks, A and D, are referenced from Snapshot 1.
- 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 will be deleted, but C1 will remain because blocks that are referenced by other snapshots cannot be deleted. When you roll back to Snapshot 3, block C1 will be recovered.

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 be deleted. In this way, when the disk is restored to the status at the time of Snapshot 3, data block C1 can also be restored.

Snapshot creation time varies depending on actual volume. For a frame of reference, it typically takes several minutes to manually create a 40 GB snapshot.

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.

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 IO. The features of ECS Snapshot 2.0 are described in the following table.

Feature	Original snapshot specifications	Snapshot 2.0 specifications	User benefit
Snapshot quota	(Number of disks)*6+6	64 snapshots for each disk	Longer protection circle Smaller protection granularity
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
Implementation principle	COW (Copy-on-write)	ROW (Redirect-on-write)	Mitigated performance impact of the snapshot task on business IO write

The implementation of ECS Snapshot 2.0 features is described in the following table.

Feature	Implementation
Snapshot quota	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. 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	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 snapshot will be deleted).
Implementation principle	The implementation principle is not made visible to users, allowing snapshots to be taken at any time of day without affecting user experience.

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.

Change history

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

Beijing data center went live.	April 2014
ECS API was officially launched.	April 2014
ECS’ s brand new user-defined image feature went live.	July 2013
The official website of Alibaba Cloud was successfully launched, and sales of ECS to external customers began.	July 2011