ALIBABA CLOUD

Alibaba Cloud

Elastic Compute Service
Product Introduction

Document Version: 20220629

Alibaba Cloud

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.

2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company or individual in any form or by any means without the prior written consent of Alibaba Cloud.

3. The content of this document may be changed because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.

4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides this document based on the "status quo", "being defective", and "existing functions" of its products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, special, or punitive damages, including lost profits arising from the use or trust in this document (even if Alibaba Cloud has been notified of the possibility of such a loss).

5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.

6. Please directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
Danger	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	**Danger:** Resetting will result in the loss of user configuration data.
Warning	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	**Warning:** Restarting will cause business interruption. About 10 minutes are required to restart an instance.
Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	**Notice:** If the weight is set to 0, the server no longer receives new requests.
Note	A note indicates supplemental instructions, best practices, tips, and other content.	**Note:** You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

1.What is ECS?

Elastic Compute Service (ECS) is a high-performance, stable, reliable, and scalable IaaS-level service provided by Alibaba Cloud. ECS eliminates the need for upfront investments in IT hardware and allows you to scale computing resources on demand. This makes ECS instances more convenient and efficient than physical servers. ECS provides a variety of instance types that suit various business needs and helps boost business growth.

Why ECS?

ECS provides the following benefits:

- You do not have to purchase hardware or construct data centers.
- Instances can be delivered within minutes, which enables rapid deployment and reduces time to market.
- You can connect ECS instances to data centers and Border Gateway Protocol (BGP) computer rooms around the world.
- You can scale resources up or down based on your actual business needs at a transparent and clear cost.
- x86 architecture-based ECS instances, ECS bare metal instances, and heterogeneous computing ECS instances such as GPU-accelerated and FPGA-accelerated instances are provided.
- You can use ECS to access other Alibaba Cloud services over the internal network to reduce Internet traffic costs.
- A host of security solutions such as virtual firewalls, role-based permission control, internal network isolation, virus protection, and traffic throttling are provided.
- ECS comes with a performance monitoring framework and an active O&M system.
- Industry-standard API that improves ease of use and compatibility

For more information about the benefits of ECS, see Benefits and Scenarios.

Architecture

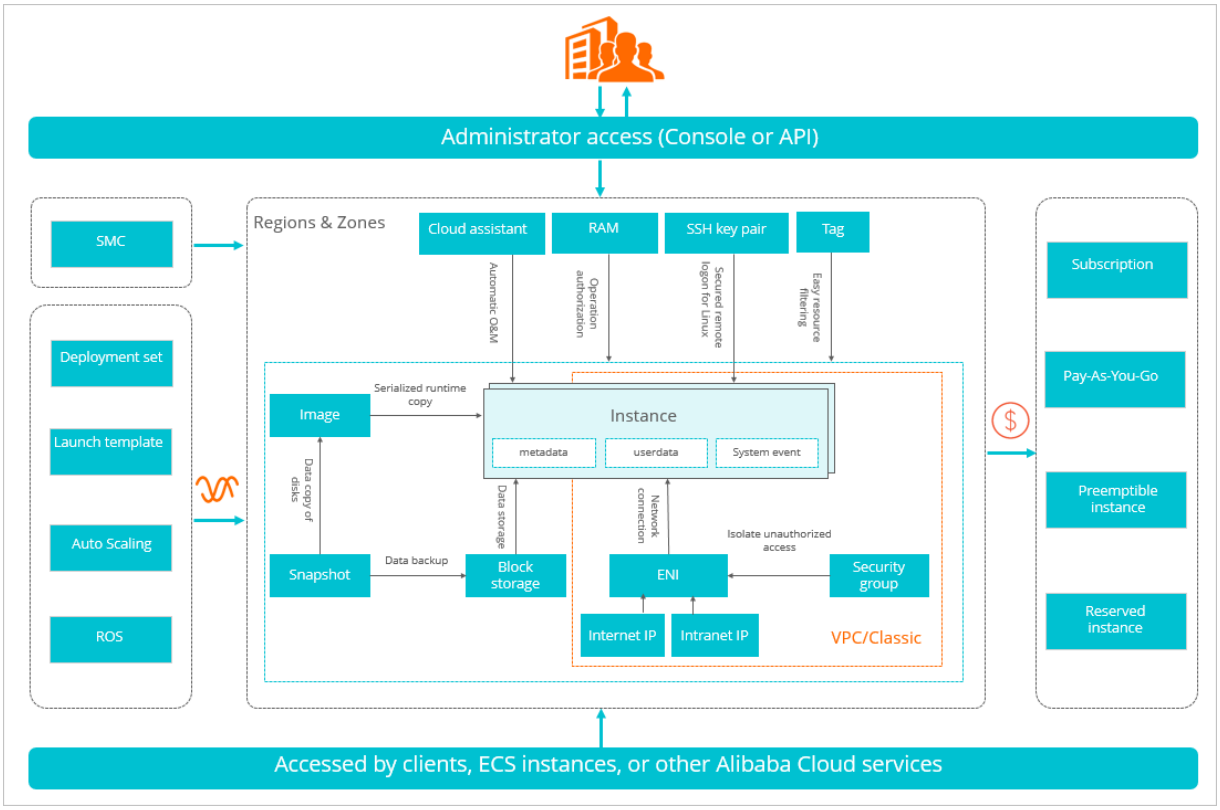
ECS comprises the following major components:

- Instance: An ECS instance is a virtual server that includes basic components such as CPU, memory, operating system (OS), network configurations, and disks. The computing performance, memory specifications, and applicable scenarios of an instance are determined by its instance type. Each instance type has particular specifications, including the number of vCPUs, memory capacity, and network performance.
- Image: Images provide OSs, initialization data of applications, and pre-installed software for instances. Multiple Linux distributions and Windows Server operating systems are supported.
- Elastic Block Storage: Elastic Block Storage (EBS) devices offer high performance and reduce latency. ECS comes with distributed storage architecture-based disks, , and physical storage-based local disks.
- Snapshot: A snapshot is a point-in-time capture of data on a disk or . Snapshots are often used to back up and restore data or to create custom images.
- Security group: A security group is a mutually accessible group of instances located within the same region that have the same security requirements. A security group works as a virtual firewall for the ECS instances inside it.
- Network:

- Virtual Private Cloud (VPC): A VPC is a logically isolated private cloud network. You can configure private CIDR blocks, route tables, and gateways for VPCs.
- Classic network: All instances of the classic network type are built on a shared infrastructure network that is planned and managed in a centralized manner by Alibaba Cloud.

For more information, visit the Elastic Compute Service page.

The following figure shows the architecture of ECS components. For more information about the functional components in the figure, see the ECS documentation.



Billing

An ECS instance includes computing resources (vCPUs and memory), image, EBS devices, public bandwidth, and snapshots. You are charged for these resources.

The following billing methods are supported:

- Subscription: You pay for resources upfront and use them over a period of time.
- Pay-as-you-go: You pay for resources after you use them. Resources can be purchased and released as needed.
- Preemptible instance: You can bid for available computing resources to create preemptible instances. Preemptible instances offer discounts compared with pay-as-you-go instances. However, preemptible instances can be reclaimed.
- Reserved instance: Reserved instances are discount coupons that are used together with pay-as-you-go instances. When you purchase a reserved instance, you make a commitment to use instances that have specified configurations such as instance type, region, and zone to receive discounted billing. Reserved instances are applied to offset the bills of computing resources.
- Savings plan: Savings plans are discount plans that are used together with pay-as-you-go instances. When you purchase a savings plan, you make a commitment to use a consistent amount (measured in

USD/hour) of resources to receive discounted billing. Saving plans are applied to offset the bills of computing resources and system disks.

- Storage capacity unit (SCU): SCUs are storage resource plans provided for use with pay-as-you-go storage resources. When you purchase an SCU, you make a commitment to use storage resources of specific capacity to receive discounted billing. SCUs are applied to offset the bills of various storage resources such as EBS devices, Apsara File Storage NAS file systems, and OSS buckets.

For more information about the billing methods of ECS instances, see Billing overview and the Pricing tab of the Elastic Compute Service product page.

Management tools

After you register an Alibaba Cloud account, you can use one of the following methods provided by Alibaba Cloud to create, use, or release ECS instances in a region:

- ECS console: a web service page used to manage ECS instances. For more information about the operations that you can perform in the ECS console, see Quick reference.
- ECS API: a Remote Procedure Call (RPC) API that supports GET and POST requests. For more information, see API Reference. The following developer tools can be used to call ECS API operations:
  - Alibaba Cloud CLI: a flexible and scalable management tool based on Alibaba Cloud APIs. You can use CLI to encapsulate Alibaba Cloud native APIs to develop custom features.
  - OpenAPI Explorer: allows you to retrieve API operations, call API operations, and dynamically generate SDK sample code.
  - Alibaba Cloud SDK: provides SDKs for a variety of programming languages such as Java, Python, and PHP.
- Resource Orchestration Service (ROS): automatically creates and configures Alibaba Cloud resources based on user-defined templates.
- Operation Orchestration Service (OOS): automatically manages and executes O&M tasks. You can define items such as execution tasks, sequence, and inputs and outputs in execution templates and use the templates to automate O&M tasks.
- Terraform: uses configuration files to call computing resources of Alibaba Cloud and other platforms that support Terraform. Terraform is an open source tool that implements version control.

Deployment suggestions

Before you purchase an ECS instance, consider the following factors:

- Region and zone

  A region represents an Alibaba Cloud data center. The region and zone determine the physical location of an ECS instance. After an instance is created, its metadata is established and its region cannot be changed. You can obtain metadata only of the ECS instances located within VPCs. Select a region and zone based on your geographical location, availability of Alibaba Cloud services, application availability requirements, and whether internal network communication is required. For example, if you want to access both ECS and ApsaraDB RDS over the internal network of Alibaba Cloud, the RDS instance and ECS instance must be within the same region. For more information, see Regions and zones.

- High availability

To ensure business consistency and continuity, we recommend that you use snapshots to back up data, and use multi-zone deployment, deployment sets, and Server Load Balancer (SLB) for disaster recovery.

- Network planning

  We recommend that you use VPC to plan your own private IP addresses. VPC supports all new features of instances and new instance types. VPC also supports business system isolation and multi-region system deployment. For more information, see What is a VPC?

- Security solutions
  - You can use ECS security groups free of charge to control inbound and outbound access policies and the port listening status of ECS instances. For more information, see Overview.
  - For applications deployed on ECS instances, Alibaba Cloud provides Anti-DDoS Basic and basic security services for free. For more information, see Anti-DDoS Basic and Basic security services.
    - Anti-DDoS Basic provides a basic defense capacity of up to 5 Gbit/s against DDoS attacks for free. By default, Anti-DDoS Basic is enabled. If you need advanced defense capacity to safeguard your ECS business, you can purchase an Anti-DDoS Pro or Anti-DDoS Premium instance. For more information, see What are Anti-DDoS Pro and Anti-DDoS Premium?
    - Basic security services for ECS are free of charge. Alibaba Cloud Security Center Basic Edition provides basic security hardening capabilities such as suspicious logon detection, vulnerability scan, and baseline check. If you want to upgrade your Security Center to the Anti-virus, Advanced, or Enterprise edition, you can purchase the edition. For more information, see What is Security Center?

2.Features

This topic provides an overview of Elastic Compute Service (ECS) features to help you better understand and use ECS.

Instance

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

ECS provides a variety of instance families for different use scenarios. Each instance family consists of multiple instance types that comprise various combinations of vCPU, memory, network, and storage capacities and offer you the flexibility to choose an appropriate mix of resources based on your business requirements.

- For information about the billing of instances, see Instance types.
- For more information about instances, see Overview.

Image

ECS images provide the information required to create ECS instances. This information includes OS data, initialization data of applications, and pre-installed software. You must select an image when you create an ECS instance. An image is a copy of data from one or more disks.

ECS images are classified into public images, custom images, shared images, community images, and Alibaba Cloud Marketplace images based on image sources. You can select images based on multiple elements such as regions, image types, OSs, and built-in software.

- For information about the billing of images, see Images.
- For more information about images, see Image overview.

Elastic Block Storage (EBS)

EBS is a high-performance, low-latency block storage service provided by Alibaba Cloud and provides block-level storage devices for use with ECS. EBS devices support random reads and writes to meet your data storage requirements. EBS devices can be used in the same manner as physical hard disks. You can partition and format EBS devices and create file systems for them.

Alibaba Cloud provides a variety of EBS devices, including cloud disks that are based on a distributed storage architecture and local disks that are based on the hard disks of physical machines. ECS supports multiple disk categories that offer different performance, including enhanced SSD (ESSD), standard SSD, and ultra disk.

- For information about the billing of EBS devices, see EBS devices.
- For more information about EBS, see Elastic Block Storage devices.

Snapshot

Snapshots are point-in-time backups of disks and can be used to back up data. When a disk fails, you can use a snapshot of this disk to roll back the disk. This way, the disk can be restored to the state it was in when the snapshot was created. You can use snapshots to back up critical disk data on a regular basis and eliminate the risk of data loss caused by accidental operations, attacks, or viruses.

- For information about the billing of snapshots, see Snapshots.
- For more information about snapshots, see Snapshot overview.

Security group

A security group acts as a virtual firewall to control the inbound and outbound traffic of ECS instances to improve security. Rules can be added to security groups to allow or deny traffic to or from the security groups. For more information about security groups, see Overview.

Network

ECS supports two network types: classic network and Virtual Private Cloud (VPC). The classic network is a shared basic network that implements Layer 3 isolation. A VPC is a private network that is logically isolated on the cloud and implements Layer 2 isolation for increased security and flexibility. IP addresses are assigned to provide access to ECS instances and allow the instances to communicate with other Alibaba Cloud services.

- You are charged only for outbound traffic to the Internet. For information about the billing of Internet usage, see Public bandwidth.
- For more information about networks, see Network types.

Deployment and elasticity

You can create launch templates, deployment sets, and auto provisioning groups to allow for fast and easy provisioning and management of ECS instances.

- A launch template stores persisted ECS instance configurations so that you do not need to make them each time you create ECS instances. For more information, see 实例启动模板概述.
- A deployment set is a policy that controls the distribution of ECS instances. You can use a deployment set to distribute your instances across different physical servers to ensure high service availability and implement underlying disaster recovery. For more information, see Overview.
- Auto Provisioning is a service that enables fast provisioning of ECS instances. You only need to make simple configurations to automate the creation of instances that use different billing methods (pay-as-you-go and preemptible instance) across instance types and zones. This improves the efficiency when you batch create a large number of instances. For more information, see Overview.
- Terraform allows you to use a simple template language to define, preview, and deploy cloud infrastructure in Alibaba Cloud. For more information, see Terraform overview.

Tag and resource group

Tags allow enterprises and individuals to identify and categorize their ECS resources and simplify the search and management of the resources. For more information, see Overview.

Resource groups are used to group your resources by usage, permission, and region, so that you can manage the resources in a hierarchical manner based on users and projects. For more information, see Resource groups.

O&M and monitoring

ECS provides comprehensive diagnostics and monitoring of instance OS states, network states, disk states, and security group rules to help you identify and resolve common issues in a timely manner. For more information, see Identify and troubleshoot instance issues.

ECS supports Cloud Assistant. Cloud Assistant is a native automated O&M tool developed for ECS. It allows you to batch maintain ECS instances and batch execute scripts on and send files to ECS instances in a password-free, logon-free manner without the use of jumper servers. These scripts can be shell, PowerShell, or bat scripts. For more information, see Overview.

3.Benefits

Compared with Internet Data Centers (IDCs) and common cloud servers, Elastic Compute Service (ECS) has the benefits of high availability, security, and elasticity.

High availability

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

Each Alibaba Cloud region has multiple zones. You can create active/standby or active/active ECS instances in multiple zones to achieve higher availability. You can build fault tolerant systems across multiple regions and zones to implement a financial-grade solution that spans three data centers across two regions. Alibaba Cloud provides mature solutions for fault tolerant services such as disaster recovery.

The Alibaba Cloud framework allows you to seamlessly switch between services. For more information about industry solutions, see Solutions by industry. Alibaba Cloud industry solutions support a variety of services, such as finance, e-commerce, and video services.

Alibaba Cloud provides you with the following support services:

- Products and services for high availability, such as Elastic Compute Service (ECS), Server Load Balancer (SLB), ApsaraDB Relational Database Service (RDS), and Data Transmission Service (DTS).
- Industry partners and ecosystem partners that help you build a more advanced and stable architecture and ensure service continuity.
- Diverse training services that help you achieve high availability from businesses to underlying services.

Security

Alibaba Cloud has passed a host of international information security certifications, such as ISO 27001 and MTCS. These certifications require to keep user data, profile, and privacy strictly confidential.

We recommend that you use ECS in a virtual private cloud (VPC). VPCs provide a stable, secure, controllable network environment that can be delivered in a short period of time. The capability and architecture of VPC hybrid cloud bring the technical advantages of cloud computing to enterprises in traditional industries that have not implemented cloud computing. For more information, see Alibaba Cloud Virtual Private Cloud.

- Breadth of network products

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

- Interconnection with your IDC

  You can use Express Connect to connect Alibaba Cloud VPC to your IDC to build a hybrid cloud architecture. You can use a variety of hybrid cloud architectures to provide network services and robust networking.

- Stability

  After you build a VPC, you can update your network architecture and obtain new network features on a daily basis to constantly evolve your network infrastructure and ensure that your business is always running steadily.

- Security

VPC features traffic isolation and attack isolation to protect your services from cyber attacks. After you build your business in a VPC, the first line of defense is established.

Elasticity

Elasticity is a key benefit of cloud computing. Alibaba Cloud is capable of providing IT resources required by a medium-sized Internet enterprise within a few minutes. In this way, most enterprises that build business on the cloud can process huge business volumes.

Alibaba Cloud provides elastic computing, storage, networking, and business architecture planning and allows you to combine your businesses.

- Elastic computing
  - Vertical scaling

    Vertical scaling is the process where the configurations of an ECS instance are modified. After you purchase an ECS instance or storage capacity from Alibaba Cloud, you can configure the instance based on your transaction volume, whereas it may be difficult to change the configurations of a server in a traditional IDC. For more information about vertical scaling, see Overview of instance upgrade and downgrade.
  - Horizontal scaling

    Horizontal scaling allows the re-division of resources between applications. A traditional IDC may not be able to immediately provide sufficient resources for online gaming or live video streaming applications during peak hours. The elasticity of cloud computing makes it possible to provide the resources required during peak hours. When the load returns to normal levels, you can release unnecessary resources to reduce operation costs. The combination of ECS vertical and horizontal elasticity enables you to scale resources up and down by specific quantities as scheduled or against business load. For more information about horizontal scaling, see 什么是弹性伸缩Auto Scaling.
- Elastic storage

  In a traditional IDC, you must add servers to increase the storage space. However, the number of servers that you can add is limited. Alibaba Cloud provides unlimited storage capacity and allows you to order as much storage space based on the business requirements. For more information about elastic storage, see Resize a disk.
- Elastic network

  Alibaba Cloud VPCs are flexible to scale. You can configure Alibaba Cloud VPCs in the same ways as you would do with IDCs. In addition, VPCs have the following benefits: interconnection between data centers, separate secure domains in data centers, and flexible network configurations and planning within a VPC. For more information about elastic networks, see Virtual Private Cloud.

Comparison between ECS and traditional IDCs

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

Item	ECS	Traditional IDC
	Provides DC-powered servers that are developed by Alibaba Cloud and have low power usage effectiveness (PUE).	Provides traditional AC-powered servers with high PUE.

Item	ECS	Traditional IDC
Equipment room deployment	Provides backbone equipment rooms with high outbound bandwidth and dedicated bandwidth.	Provides equipment rooms with various quality levels and shared bandwidth primarily, difficult for users to choose from.
	Provides multi-line Border Gateway Protocol (BGP) computer rooms, which enables smooth and balanced access among regions.	Provides equipment rooms with a single or dual line primarily.
Ease of operation	Provides mainstream OSs, including activated Windows OS.	Requires users to purchase and install OSs manually.
	Easily switches between OSs online.	OSs have to be manually reinstalled.
	Provides a Web-based console for online management.	Users must manually perform management and maintenance operations.
	Provides mobile phone verification for password setting, increasing data security.	Brings difficulty in resetting passwords, and exposes high risk of password cracking.
Disaster recovery and backup	Stores three copies of each piece of data. When one copy is corrupted, the data can be quickly restored.	Users must build a disaster recovery environment by themselves, and use traditional storage devices.
	Users can customize automatic snapshot policies to create automatic snapshots for data recovery.	Does not support automatic recovery because the snapshot function is not provided.
	Hardware failures can be recovered quickly and automatically.	Users must restore corrupted data manually.
Security and reliability	Effectively prevents MAC spoofing and ARP attacks.	Fails to prevent MAC spoofing and ARP attacks.
	Effectively defends against DDoS attacks by using black holes and traffic scrubbing.	Needs additional costs for devices for traffic scrubbing and black hole shielding systems.
	Provides additional services, such as port scanning, trojan scanning, and vulnerability scanning.	Typically encounters problems such as port scanning, trojan scanning, and vulnerability scanning.
Flexible scalability	Activates cloud servers on demand and upgrades configurations online.	Needs a long time for server delivery.
	Adjusts outbound bandwidth as required.	Requires one-off purchase of outbound bandwidth that cannot be adjusted.
	Combines with SLB online, which makes it easy to scale up applications.	Uses hardware-based server load balancing, which is expensive and difficult to set up.

Item	ECS	Traditional IDC
Cost effectiveness	Low cost.	High cost.
	Small up-front investment.	Large up-front investment, causing serious waste of resources.
	Provides pay-as-you-go and flexible payment options to allow you to flexibly respond to business changes.	Requires users to purchase up front to meet configuration requirements during peak hours.

4.Scenarios

ECS is a highly flexible solution that can be used independently as a simple web or application server, or used together with other Alibaba Cloud services to deliver advanced solutions.

> **Note** This topic describes some typical scenarios of ECS and how to take advantage of the benefits of cloud computing while using ECS.

Official websites and lightweight web applications

A new website has low traffic and requires only a low-configuration ECS instance to run web applications such as Apache and NGINX, host databases, and store files. As your website develops, you can upgrade the ECS instance configurations or add ECS instances at any time to provision sufficient resources for handling traffic spikes.

Multimedia and high-concurrency applications or websites

ECS can be used with Object Storage Service (OSS) to store static images, videos, or downloaded packages to reduce storage costs. In addition, ECS can work with Alibaba Cloud Content Delivery Network (CDN) and Server Load Balancer (SLB) to shorten waiting time, reduce public bandwidth costs, and improve service availability. For more information, see What is OSS?, What is Alibaba Cloud CDN?, and What is SLB?

Databases with high I/O requirements

ECS supports databases with high I/O requirements, such as OLTP and NoSQL databases. A high-configuration I/O optimized ECS instance can be used with ESSDs to achieve high I/O concurrency and higher data reliability. Alternatively, multiple lower-configuration I/O optimized ECS instances can be used with SLB to deliver a high availability architecture. For more information, see ESSDs and What is SLB?

Applications and websites with sharp traffic fluctuations

Some applications may experience sharp traffic fluctuations within a short period of time. When ECS is used with Auto Scaling, the number of ECS instances is automatically adjusted based on traffic. This way, you can meet the changing resource requirements at a low cost. ECS can also work with SLB to deliver a high availability architecture. For more information, see What is Auto Scaling? and What is SLB?

Big data and real-time online and offline analysis

ECS provides big data instance families that support Hadoop distributed computing, log processing, and large data warehouses. Big data instance families adopt a local storage architecture, which helps deliver better network performance for Hadoop and Spark clusters while providing abundant storage space and higher storage performance. For more information, see Big data instance families.

AI applications such as machine learning and deep learning

ECS allows you to use GPU-accelerated compute optimized instances to build AI applications based on frameworks such as TensorFlow. Such instances have lower computing capacity requirements for clients and are suitable for scenarios such as image processing and real-time online rendering for cloud gaming and AR/VR applications. For more information, see GPU-accelerated compute optimized instance families.

More cases

For more scenarios of ECS, see Alibaba Cloud solutions.

5.Terms

This topic describes the terms used in Elastic Compute Service (ECS).

ECS terms

Term	Description
ECS instance	An ECS instance is a virtual server that includes basic components such as vCPUs, memory, an operating system (OS), network configurations, and disks.
ECS instance type	Instance types define the basic attributes such as computing capacity, storage capacity, and networking performance of ECS instances. Instance types must be used together with images, Elastic Block Storage (EBS) devices, and network resources to create ECS instances that serve different purposes.
image	Images contain information that is necessary to run ECS instances, such as OSs and initialization data of applications.
public image	Public images are base images provided by Alibaba Cloud. Public images are licensed and include Windows Server OS images and mainstream Linux OS images.
• Alibaba Cloud Linux 3 • Alibaba Cloud Linux 2	Alibaba Cloud Linux 2 and 3 are OSs provided by Alibaba Cloud. They offer a safe, stable, and high-performance customized environment for applications on ECS instances and are optimized for the infrastructure of Alibaba Cloud.
custom image	You can create or import custom images. Custom images contain the initial system environment, application environment, and related software configurations. This eliminates the need for repeated manual configurations.
Elastic Block Storage device	EBS devices offer high performance and reduce latency. You can partition and format EBS devices and create file systems on the devices to meet the data storage requirements of your business.
disk	Disks are block-level EBS devices that use a triplicate mechanism to ensure 99.9999999% data durability for ECS instances.
local disk	Local disks are located on the same physical server as the ECS instance to which the disks are attached. Local disks are cost-effective and provide high storage I/O. However, the durability of data stored on local disks is determined by the reliability of the associated physical server, which increases the risks of single points of failure (SPOFs).
snapshot	A snapshot is a point-in-time backup of a disk and is used to back up or restore the disk.
security group	A security group is a virtual firewall that is used to control the inbound and outbound traffic of ECS instances in the security group.
SSH key pair	An SSH key pair is a secure and convenient authentication method provided by Alibaba Cloud for instance logons. An SSH key pair consists of a public key and a private key. You can use SSH key pairs to log on to only Linux instances.

Term	Description
Instance RAM role	Instance Resource Access Management (RAM) roles enable ECS instances to assume roles with specific access permissions. An instance can access the APIs of specified Alibaba Cloud services and manage specified Alibaba Cloud resources based on a Security Token Service (STS) temporary credential to ensure high security.
virtual private cloud (VPC)	A VPC is a private network established on Alibaba Cloud. VPCs are logically isolated from each other based on tunnels. You have full control over your VPCs. For example, you can specify CIDR blocks and configure route tables and gateways for your VPCs.
elastic network interface (ENI)	An ENI is an independent virtual network interface that can be bound to or unbound from an ECS instance to implement the flexible scaling and migration of services.
launch template	A launch template contains configuration information that you can use to create ECS instances and eliminates repeated manual configurations.
deployment set	Deployment sets support the high availability strategy. After you apply a high availability strategy to a deployment set, all the instances within the deployment set are distributed across different physical servers to ensure business availability and disaster recovery capabilities at the underlying layer.
dedicated host	A dedicated host is a cloud host whose physical resources are exclusively reserved for a single tenant. Dedicated hosts meet strict security and compliance requirements and support Bring Your Own License (BYOL) when you migrate services to the cloud.
auto provisioning group	Auto provisioning groups support quick deployment of instance clusters across instance types and zones. Auto provisioning groups can create preemptible instances and pay-as-you-go instances by using a combination of provisioning policies to provide high stability at low cost.
tag	Each tag consists of a key and a value. You can add tags to resources that have identical characteristics, such as resources that belong to the same organization and resources that serve the same purpose. You can use tags to search for and manage resources in an efficient manner.
resource group	Resource groups allow you to manage resources across services and regions based on your business requirements and manage the permissions of resource groups.
Cloud Assistant	Cloud Assistant is an automated O&M tool provided by Alibaba Cloud. Cloud Assistant allows you to perform operations such as running commands in ECS instances and sending files to ECS instances without logging on to the ECS instances.
system event	System events are scheduled or unexpected O&M events that affect the running status of ECS instances and require the restart, stop, or release of ECS instances. For system events, ECS sends you notifications that contain information such as solutions and event cycles so that you can back up data and make preparations in a timely manner.

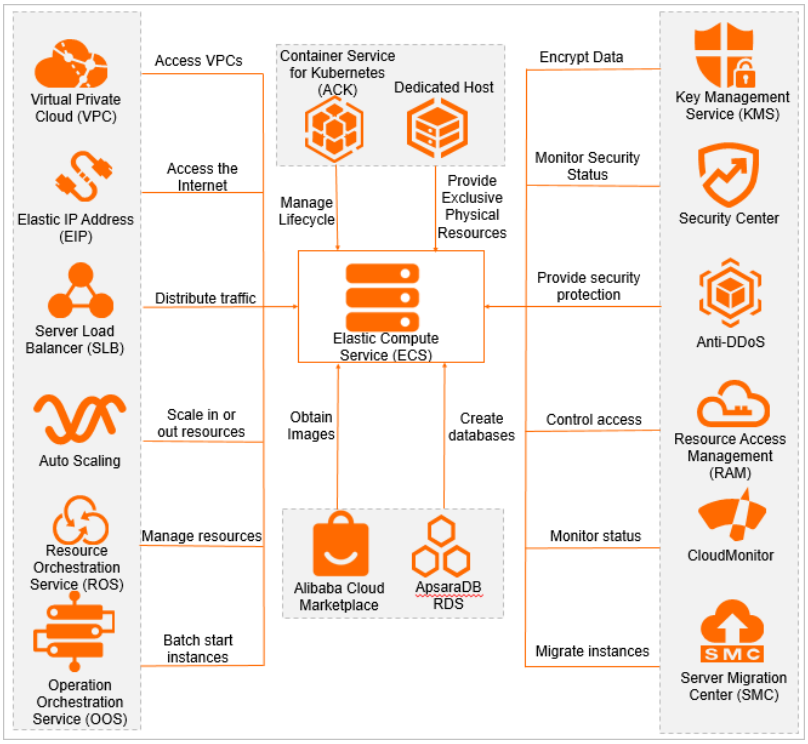
Billing terms

Term	Description
subscription	In this billing method, you pay for resources upfront and use them over a specified period of time.
pay-as-you-go	In this billing method, you pay for resources after you use them. Resources can be purchased and released as needed.
preemptible instance	In this billing method, you can bid for available computing resources to create preemptible instances. Preemptible instances are discounted compared with pay-as-you-go instances. However, preemptible instances can be reclaimed.
reserved instance	Reserved instances are discounted coupons that are used together with pay-as-you-go instances. When you purchase a reserved instance, you make a commitment to use instances that have specified configurations such as instance type, region, and zone to receive discounted billing. Reserved instances are applied to offset the bills of computing resources.
savings plan	Savings plans are discounted plans that are used together with pay-as-you-go instances. When you purchase a savings plan, you make a commitment to use a consistent amount (measured in USD/hour) of resources to receive discounted billing. Savings plans are applied to offset the bills of computing resources and system disks.
Resource Assurance	After an elasticity assurance or a capacity reservation is created, the system generates a private pool to reserve resources for a specific number of instances that have specific attributes. During the validity period of the elasticity assurance or capacity reservation, you always have access to the resources reserved in the private pool when you want to create instances.
storage capacity unit (SCU)	SCUs are storage resource plans that are used together with pay-as-you-go storage resources. When you purchase an SCU, you make a commitment to use storage resources of specific capacity to receive discounted billing. SCUs are applied to offset the bills of various storage resources such as EBS devices, Apsara File Storage NAS file systems, and Object Storage Service (OSS) buckets.
pay-by-bandwidth	Pay-by-bandwidth is a billing method for network usage. You are charged based on the specified bandwidth.
pay-by-traffic	Pay-by-bandwidth is a billing method for network usage. You are charged for the actual amount of traffic that you use. You must configure a maximum bandwidth value to prevent unexpected fees caused by traffic bursts.

6.Supported Alibaba Cloud services

This topic describes the relationships between Elastic Compute Service (ECS) and other Alibaba Cloud services.

The following figure shows the relationships between ECS and other Alibaba Cloud services.



The following table describes the relationships between ECS and other Alibaba Cloud services.

Service	Relationship	References
Virtual Private Cloud	Virtual Private Cloud (VPC) provides an isolated, stable, secure, fast-deliverable, and controllable network environment for ECS. You have complete control over IP addresses and network topology within a VPC. VPCs are suitable for users who have high network security requirements.	IP addresses of ECS instances in VPCs
Elastic IP Address	Elastic IP addresses (EIPs) are NAT IP addresses that are located in the public gateway of Alibaba Cloud. By means of NAT gateways, EIPs are mapped to the primary elastic network interfaces (ENIs) of the ECS instances that are associated with the EIPs. You can associate EIPs with ECS instances that are located in VPCs to enable the instances to communicate over the Internet.	• Elastic IP addresses • Associate or disassociate an EIP
Server Load Balancer	Server Load Balancer (SLB) distributes traffic across multiple ECS instances.	Architecture

Service	Relationship	References
CloudMonitor	CloudMonitor develops monitoring solutions for ECS instances, system disks, and public bandwidth.	• View the monitoring information of an instance • View the monitoring data of a disk
Resource Access Management	Resource Access Management (RAM) allows you to manage the identities and permissions of Alibaba Cloud services and users to implement access control over ECS resources.	RAM overview
Anti-DDoS	Anti-DDoS Origin Basic is a service that protects ECS instances from DDoS attacks to ensure system stability. If inbound traffic to an instance exceeds the maximum traffic rate allowed by the instance type, Alibaba Cloud Security throttles the traffic.	Anti-DDoS Basic
Security Center	Alibaba Cloud Security Center provides ECS with basic security services such as unusual logon detection, vulnerability scan, and baseline check. You can check the security status of your ECS instances in the ECS console or in the Security Center console.	Basic security services
Key Management Service	The ECS disk encryption feature uses Key Management Service (KMS) to encrypt data stored in ECS instances.	• Encrypt a system disk • Encrypt a data disk
Auto Scaling	Auto Scaling (ESS) adjusts the number of ECS instances based on business and policy changes. • ESS adds ECS instances to ensure computing capabilities as business requirements increase. • ESS removes ECS instances to reduce costs as business requirements decrease.	• Create a scaling group based on an existing ECS instance • Create a scaling configuration (ECS)
Resource Orchestration Service	You can create a stack template by using the Resource Orchestration Service (ROS) console or by calling ROS API operations. Then, you can use the template to quickly create and manage ECS resources.	ROS overview
Operation Orchestration Service	You can use the ACS-ECS-BulkyStartInstances public template of Operation Orchestration Service (OOS) to start multiple ECS instances at a time.	Overview
ApsaraDB RDS	ECS instances work with ApsaraDB RDS instances that reside within the same region as the ECS instances to build a typical service access architecture. This reduces network latency and Internet traffic fees and ensures the optimal performance of ApsaraDB RDS instances.	How do I access an RDS instance from an ECS instance over the intranet?

Service	Relationship	References
Container Service for Kubernetes	Container Service for Kubernetes (ACK) uses Docker containers to manage application lifecycles on groups of ECS instances.	• What is Container Service for Kubernetes? • Create an ECI from a specified ECS instance type
Dedicated Host	You can deploy ECS instances on a dedicated host and gain exclusive access to its physical resources. Dedicated Host allows you to redeploy your business or migrate your business to the cloud at minimal costs. ECS instances deployed on dedicated hosts meet strict regulatory compliance requirements.	• Dedicated hosts • Create one or more ECS instances on a dedicated host
Server Migration Center	You can use Server Migration Center (SMC) to migrate ECS instances within an Alibaba Cloud account or across Alibaba Cloud accounts.	Migrate servers
Alibaba Cloud Marketplace	Alibaba Cloud Marketplace is a platform where third-party partners provide various software and services such as software infrastructure, business software, website construction, hosted O&M, cloud security, data, API operations, and solutions for ECS.	Alibaba Cloud Marketplace images

7.Regions and zones

This topic provides a complete list of regions and zones of Alibaba Cloud.

Region

Alibaba Cloud's data centers in China share similar attributes in regards to infrastructure, BGP network quality, service quality, ECS operation, and configuration. After a resource is created, you cannot change the region of the resource. The following table describes the information about all regions of Alibaba Cloud, including the region IDs and the cities where the regions reside.

> **Note** Available regions vary with services. You must consider regions available for different services.

- Regions in the Chinese mainland

Region	City	Region ID	Number of zones
China (Qingdao)	Qingdao	cn-qingdao	2
China (Beijing)	Beijing	cn-beijing	12
China (Zhangjiakou)	Zhangjiakou	cn-zhangjiakou	3
China (Hohhot)	Hohhot	cn-huhehaote	2
China (Ulanqab)	Ulanqab	cn-wulanchabu	3
China (Hangzhou)	Hangzhou	cn-hangzhou	8
China (Shanghai)	Shanghai	cn-shanghai	11
China (Shenzhen)	Shenzhen	cn-shenzhen	6
China (Heyuan)	Heyuan	cn-heyuan	2
China (Guangzhou)	Guangzhou	cn-guangzhou	2
China (Chengdu)	Chengdu	cn-chengdu	2
China (Nanjing)	Nanjing(China (Nanjing) is one of the local regions and is in invitational preview.)	cn-nanjing	1

- Regions outside the Chinese mainland

Country	City	Region ID	Number of zones
China	China (Hong Kong)	cn-hongkong	3
Singapore	Singapore (Singapore)	ap-southeast-1	3

Country	City	Region ID	Number of zones
Australia	Australia (Sydney)	ap-southeast-2	2
Malaysia	Malaysia (Kuala Lumpur)	ap-southeast-3	2
Indonesia	Indonesia (Jakarta)	ap-southeast-5	3
Philippines	Philippines (Manila)	ap-southeast-6	1
Thailand	Thailand (Bangkok)	ap-southeast-7	1
India	India (Mumbai)	ap-south-1	2
Japan	Japan (Tokyo)	ap-northeast-1	2
South Korea	South Korea (Seoul)	ap-northeast-2	1
US	US (Silicon Valley)	us-west-1	2
US	US (Virginia)	us-east-1	2
Germany	Germany (Frankfurt)	eu-central-1	2
UK	UK (London)	eu-west-1	2
UAE	UAE (Dubai)	me-east-1	1

When you select a region, you must consider the following factors:

- Geographical locations

  Select a region based on the geographical location of you and your users.

  - Regions in the Chinese mainland

    In the Chinese mainland, we recommend that you select a region that is the closest to the geographical location of your users to speed up access. However, in terms of network infrastructures, Border Gateway Protocol (BGP) network quality, quality of service (QoS), and usage of and configurations on Elastic Compute Service (ECS) instances, Alibaba Cloud regions in the Chinese mainland are almost the same. BGP networks ensure fast access to all regions in the Chinese mainland.

  - Regions outside the Chinese mainland

    Bandwidths provided in regions outside the Chinese mainland are applicable to users in those regions. If you are located in the Chinese mainland, we recommend that you do not select regions outside the Chinese mainland because you may experience high latency if you select those regions.

- Connection between Alibaba Cloud services

  If you use multiple Alibaba Cloud services together, take note of the following items:

  - ECS instances, ApsaraDB RDS instances, and Object Service Storage (OSS) buckets that are created in different regions cannot communicate with each other over internal networks.

  - Server Load Balancer (SLB) cannot balance requests from ECS instances deployed in different regions. ECS instances that you purchased in different regions cannot be deployed on the same SLB instance.

- Resource price

  The prices of resources may vary with regions. For more information, see the Pricing page.

- ICP license filing

  When you select a region, you must consider the special requirements of some regions. For example, if you purchase an ECS instance in a region in the Chinese mainland and use the instance as a web server, you must apply for an ICP license.

  If you want to apply for an ICP license, take note of the following items:

  - If you want to apply for an ICP license for services in Beijing, select the **China (Beijing)** region.
  - If you want to apply for an ICP license for services in Guangdong, select the **China (Shenzhen)** region.

  > **Note** The approval requirements for ICP licenses vary from province to province. For information about the latest requirements, visit the ICP license application website of the local communications administration.

Zone

A zone is a physical area within a region that has its own independent power grids and networks. The network latency between instances within the same zone is lower than that between different zones.

Zones within the same region can access each other over the internal network. When a zone is down, other zones are not affected. We recommend that you choose a deployment method based on your business requirements for disaster recovery and network latency.

- If your application requires high disaster recovery capabilities, we recommend that you deploy instances in different zones within the same region.
- If your application requires low latency, we recommend that you deploy instances within the same zone.

8.Usage notes

To ensure that Elastic Compute Service (ECS) instances can run normally, take note of the following items before you use the instances.

Operation notes

- After you create an ECS instance, only you have the administrator permissions on the instance and can log on to the instance.
- Do not resell or sublicense bandwidth assigned to your ECS instances without authorization. Failure to comply with the note causes your instances to be stopped, locked, and released.
- Do not use your ECS instances for malicious, fraudulent, or illegal activities, such as click farming or fraudulent transactions on e-commerce websites such as Taobao. Failure to comply with the note results in suspension or termination of your account.
- Do not uninstall relevant hardware drivers.
- Do not modify the media access control (MAC) addresses of network interface controllers (NICs) unless necessary.
- Do not enable SELinux.
- ECS instances with 4 GiB or larger memory must use 64-bit operating systems. A 32-bit operating system can address a maximum of 4 GiB of memory. Alibaba Cloud allows you to create instances that run the following 64-bit operating systems. The operating system versions on the instance buy page prevail.
  - Alibaba Cloud Linux 64-bit
  - CoreOS 64-bit
  - CentOS 64-bit
  - Debian 64-bit
  - FreeBSD 64-bit
  - openSUSE 64-bit
  - SUSE Linux 64-bit
  - Ubuntu 64-bit
  - Red Hat 64-bit
  - Windows 64-bit
- To ensure service continuity and prevent failover-induced service unavailability, we recommend that you configure relevant software to automatically start on instance startup. If service applications are connected to databases, we recommend that you enable auto-reconnection for these service applications.
- We recommend that you do not upgrade the kernel or operating system of an instance. To upgrade the kernel, see How do I prevent Linux instance startup failures after the kernel is upgraded?
-

Notes on Windows instances

- Do not stop the built-in AliyunService or shutdownmon.exe process. Otherwise, ECS instances may fail to be stopped or restarted in the ECS console.
- Do not modify the hostname of the domain controller.

- We recommend that you do not create custom images by using a virtual machine that acts as a domain controller.
- Do not rename, delete, or disable the administrator account. Otherwise, the use of the server may be affected.
- If your instance uses basic disks, we recommend that you do not use the virtual memory. If your instance uses ultra disks, standard SSDs, or enhanced SSDs (ESSDs), you can use the virtual memory.
- Proceed with caution when you use the administrator account to resize disks, manage spanned volumes or the registry, and update the system. Failure to comply with the note can result in data loss.
- A 32-bit Windows operating system supports up to four vCPUs.
- Make sure that a minimum of 2 GiB memory is available when you build a website or deploy a web environment on a Windows instance. Instance types that have only one vCPU and one GiB of memory do not support MySQL databases.
- For information about usage notes on images and operating systems, see Select an image.

Notes on Linux instances

- Do not modify the content of the default */etc/issue* file on Linux instances. If you modify the /etc/issue file on a Linux instance and then use this instance to create a custom image, the operating system distribution of the image cannot be recognized. This causes instances created from the image to fail to start.
- Do not modify the permissions of the directories in the root partition, especially */etc*, */sbin*, */bin*, */boot*, */dev*, */usr*, and */lib*. Improper modifications of permissions may cause errors.
- Do not rename, delete, or disable the root account.
- Do not compile the Linux kernel or perform operations on it.
- If your instance uses basic disks, we recommend that you do not use swap partitions. If your instance uses ultra disks, standard SSDs, or ESSDs, you can use swap partitions.
- Proceed with caution when you use the root account to run the fio, mkfs, or fsck command or resize disks. Failure to comply with the note can result in data loss.
- For information about usage notes on images and operating systems, see Select an image.

Limits

For information about limits on ECS, see 使用限制.