Huawei Certification Cloud Service Training Courses

HCIA - Cloud Service

Learning Guide

Version: V3.0

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions

 and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.

All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: http://e.huawei.com

Huawei Certification System

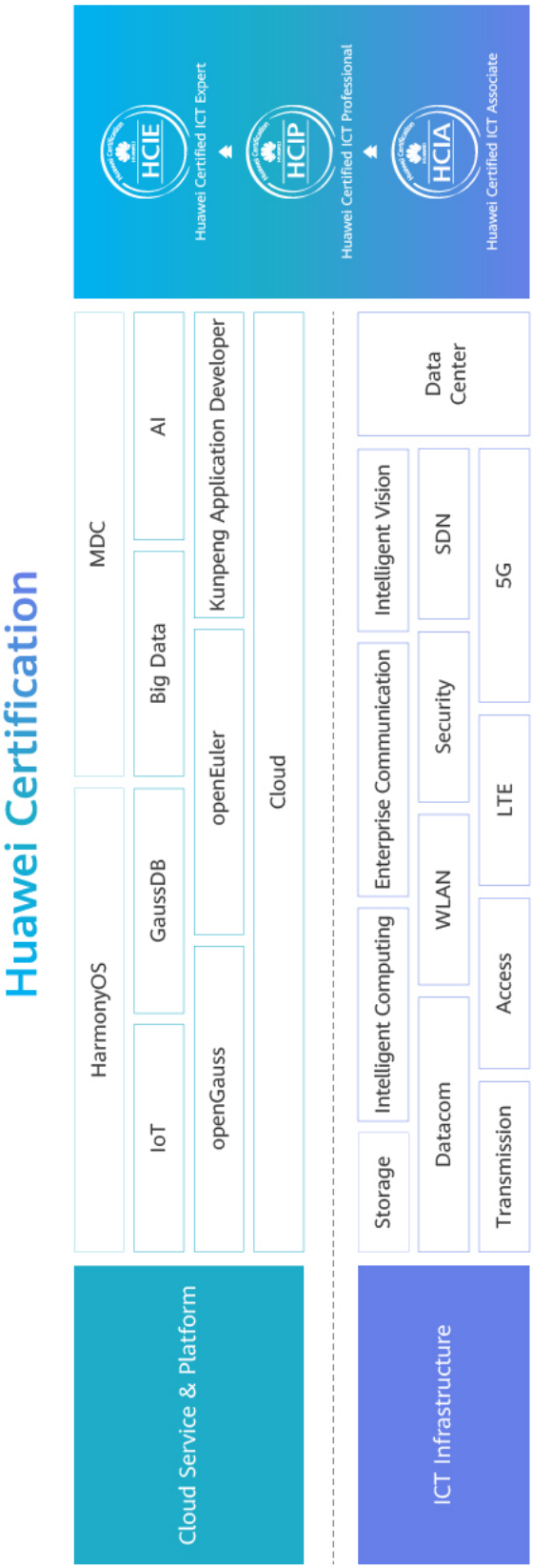
The Huawei certification system is a platform for shared growth, part of a thriving partner ecosystem. There are two types of certification: one for ICT architectures and applications, and one for cloud services and platforms. There are three levels of certification available:

- Huawei Certified ICT Associate (HCIA)
- Huawei Certified ICT Professional (HCIP)
- Huawei Certified ICT Expert (HCIE)

Huawei certification courses cover the entire ICT domain, with a focus on how today's architecture generates cloud-pipe-device synergy. The courses present the latest developments of all essential ICT aspects to foster a thriving ICT talent ecosystem for the digital age.

Huawei Certified ICT Associate - Cloud Service (HCIA - Cloud Service) is designed for frontline engineers at Huawei local offices and representative offices, and enthusiasts of Huawei Cloud services. An HCIA - Cloud Service engineer can create an IT architecture using common compute, storage, and network cloud services. After completing this course, you will understand the key concepts of cloud computing services. You will know how to use and manage cloud services to construct a cloud infrastructure.

Huawei certification helps you unlock opportunities to advance your career and take one more step towards the top of the industry.



Contents

1 Cloud Computing Basics

1.1 Introduction to Cloud Computing

1.1.1 Background

Information and communications technologies have developed rapidly, evolving from the personal computers to mobile Internet, and now to the Internet of Everything (IoE). With IoE, more and more devices are constantly being connected to the Internet, and a massive amount of data is generated every day, which is straining the limits of traditional IT infrastructures. The following are the main pain points for enterprises relying on traditional IT infrastructures:

- Device procurement takes a long time, which, in turn, slows down rollout of new business systems.
- Traditional hardware devices are isolated from each other, and reliability mainly depends on software.
- Hardware management is too complex as they have to manage different types of devices provided by different vendors.
- The performance of a single device is limited.

The device utilization is low, and the total cost of ownership (TCO) is high. A traditional IT architecture cannot meet requirements for rapid development of enterprise services. So is there any new IT architecture that can address these issues? The answer is cloud migration. Enterprise managers are embracing cloud migration. A cloud-based IT architecture has been gradually replacing traditional IT architectures. See Figure 1-1 for some details.

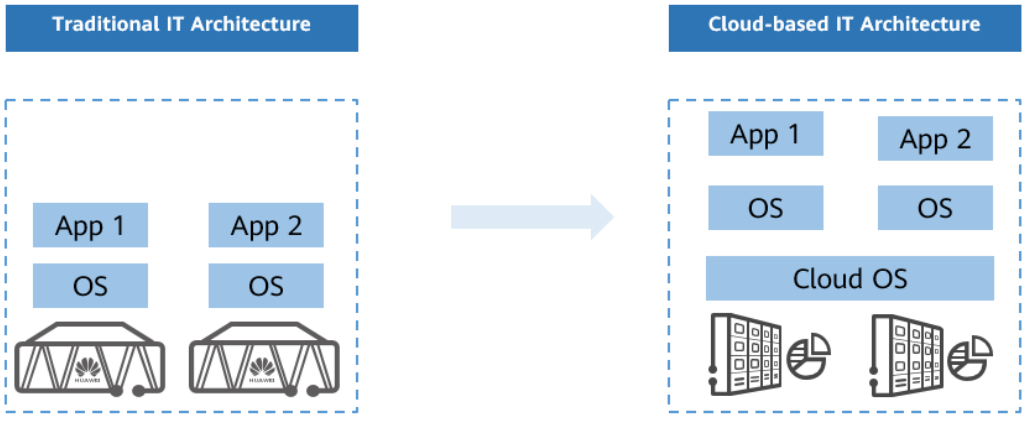


Figure 1-1 Cloud-based IT architecture

On the left, in Figure 1-1, is a traditional IT architecture, which is also known as a siloed architecture. In this architecture, each project or application in an enterprise has its own independent infrastructure (independent hardware and operating system). In the early days, physical servers were used in this way. At that time, physical servers were classified as website servers, mail servers, or other types of servers, according to the types of applications deployed on them. However, services expanded fast, and problems with this type of architecture were gradually exposed. This type of architecture is not reliable enough and single points of failure (SPOF) may occur. Only one main application runs on each device, resulting in low resource utilization. Applications are physically isolated, making it difficult for them to share data. With the increasing popularity of cloud computing technologies, more and more enterprises are moving towards cloud-based architecture. The distributed architecture, with resources deployed in clusters, greatly improves the application reliability. Cloud computing makes resource sharing easier and enables on-demand resource usage, which massively improves utilization and simplifies O&M.

1.1.2 What Is Cloud Computing?

Since the emergence of cloud computing, it has been defined in a variety of ways. The National Institute of Standards and Technology (NIST) defines cloud computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (such as networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This definition is widely accepted.

The key points in the definition are as follows:

- Cloud computing is a model, not a specific technology.
- With cloud computing, users can easily access IT resources such as networks, servers, storage, applications, and services.
- Cloud computing enables ubiquitous access to resources connected to a network.
- Resources can be quickly provisioned and released for elastic scaling. On-demand self-service enables minimal service provider interaction.

In addition to these points, we can, instead of considering just "cloud computing", we can consider the "cloud" and "computing" parts separately. The cloud part refers to networks and the Internet in general. The cloud includes the Internet and all the underlying infrastructure that the Internet runs on. Computing refers to a range of compute services (functions and resources) provided by a powerful computer. Cloud computing means delivering the powerful compute resources over the Internet.

While everyone may seem to have their own definitions of what cloud computing is, one thing about it is undeniable: It is everywhere around us.

So now let's look at the world of cloud computing that surrounds us.

1.1.3 Cloud Computing Around Us

Now our daily life benefits from cloud computing.

Baidu Wangpan is a cloud storage service from Baidu. It makes it easy to back up, synchronize, and share photos, videos, and documents. Baidu Wangpan is a popular

service throughout China. Without cloud computing, we would need to manually copy our files to other hard disks to back up, synchronize, and share files. With cloud computing, you can easily back up, synchronize, and share files over the Internet. All it takes is an app like Baidu Wangpan installed either on your phone or PC. With cloud computing, resources are shared. Data shared on the cloud can be shared and downloaded much more easily, and there are various techniques to ensure data on the cloud is automatically synchronized.

Enterprises and governments are both embracing cloud computing and migrating their services to the cloud. For example, in China, there is an individual income tax app that uses cloud computing technology to make tax declaration easier and faster. Now people can declare their income taxes from their phones. There is no need to physically visit their local tax authority. In this case, the migration to the cloud has broken down data silos and optimized tax process. Now, people can enjoy many online services provided by online service halls.

Nearly every province and city in China has moved some portion of their public facing services to the cloud. Figure 1-2 illustrates how an applicant can submit an application and supporting materials online. When the relevant government agencies receive the online application, they will share the data and work with each other to process the application.

Cloud migration can reduce government expenditures, generate revenue for cloud service providers, and provide the citizenry with more convenient services.

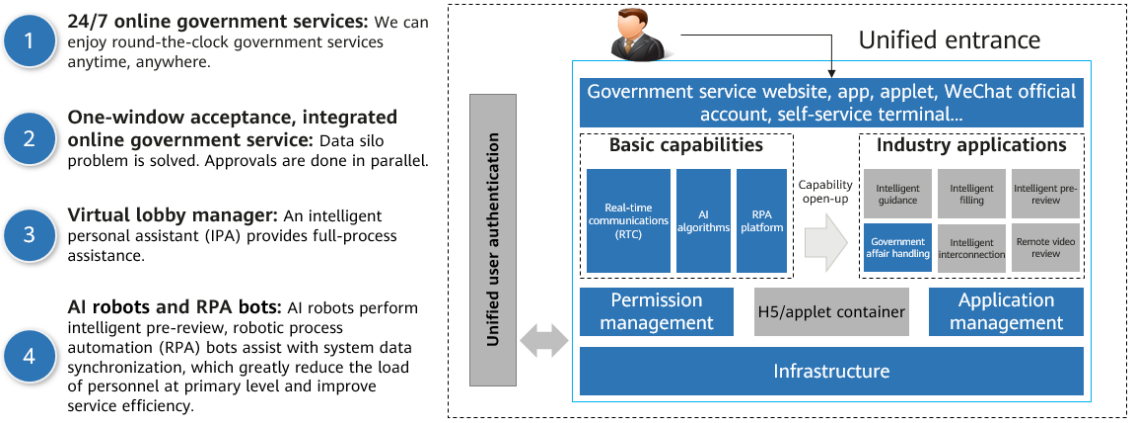


Figure 1-2 Online government services

1.1.4 Cloud Computing Models

Currently, there are no internationally recognized models for cloud computing, but there is an industry consensus that cloud computing can be classified based on deployment and operation models. Now, let's look at the deployment models first. For the most part, there are public, private, hybrid, and industry clouds.

In this course, we will examine the first three: public cloud, private cloud, and hybrid cloud.

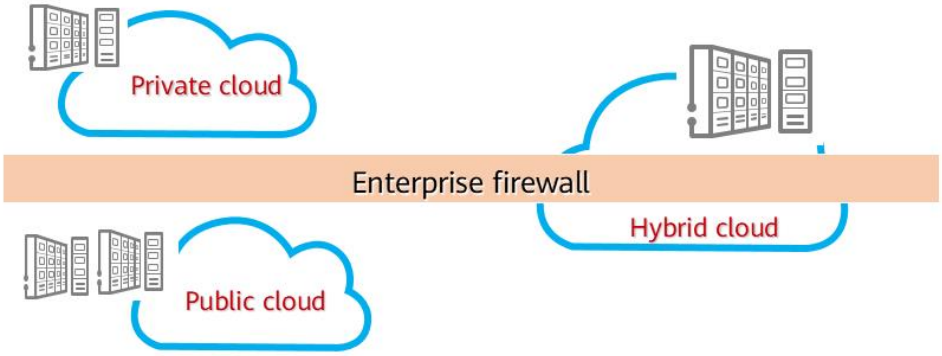


Figure 1-3 Deployment models of cloud computing

- Public cloud

  Public cloud was the first type of cloud to come out and is the most well-known deployment model. Currently, public cloud can provide users with many services. Users can access IT services from anywhere they have an Internet connection.

  Public clouds are usually built and managed by cloud service providers. The service provider handles all the infrastructure. The end users just purchase the computing resources or services and leave the O&M to the service provider. Public cloud resources are available for anyone with an Internet connection.

- Private cloud

  Private clouds are typically deployed within an organization. All of the data on a private cloud is stored in the enterprises private data center. Attempts to access such data are controlled by firewalls configured for maximum protection. A private cloud can be deployed based on the existing architecture of an organization. To save money, expensive hardware devices can be re-used for a private cloud. However, creating a private cloud to make sure existing equipment does not go to waste is a double-edged sword. A private cloud ensures security and lets you reuse existing devices, but, as time goes by, the devices will eventually break down or get overtaken by newer technology, and replacing them can be expensive. Also, this type of private cloud still does not make it easy to share data among users or enterprises.

  In recent years, a new type of private cloud has been developed, where you can purchase dedicated cloud services on a public cloud and then migrate just key services. This option gives you dedicated, high-performance compute and storage resources that are extremely reliable and secure.

- Hybrid cloud

  Hybrid cloud is a flexible cloud computing model consisting of some combination of public, private, and industry clouds. User services can be switched between these clouds as required. For security and ease of control, not all the enterprise information is placed on the public cloud. Most enterprises tend to adopt a hybrid cloud model. Many choose a combination of public and private clouds. Public cloud resources can be billed on a pay-per-user basis, which can result in tremendous savings for enterprises with demand that can fluctuate suddenly, for example, retail tends to see massive spikes in demand during holidays, and tends to be fairly seasonal. A hybrid cloud can also provide you with better options for disaster recovery. If a disaster

occurs on services deployed on a private cloud, they often can be temporarily transferred to a public cloud. This is a highly cost-effective approach. Another approach is to deploy services on a public cloud, and use another public cloud for disaster recovery.

A hybrid cloud allows you to take advantage of both public and private clouds. It lets you flexibly move applications across multiple clouds. A hybrid-cloud is cost-effective.

Of course, the hybrid cloud model has some disadvantages too. The maintenance and security of a hybrid cloud can be complex. In addition, because a hybrid cloud is a combination of different cloud platforms, data, and applications, integration can be a challenge. When developing a hybrid cloud, you may face compatibility issues between infrastructures.

So that's the main deployment models. There are also different operation models, like infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS), just to name a few.

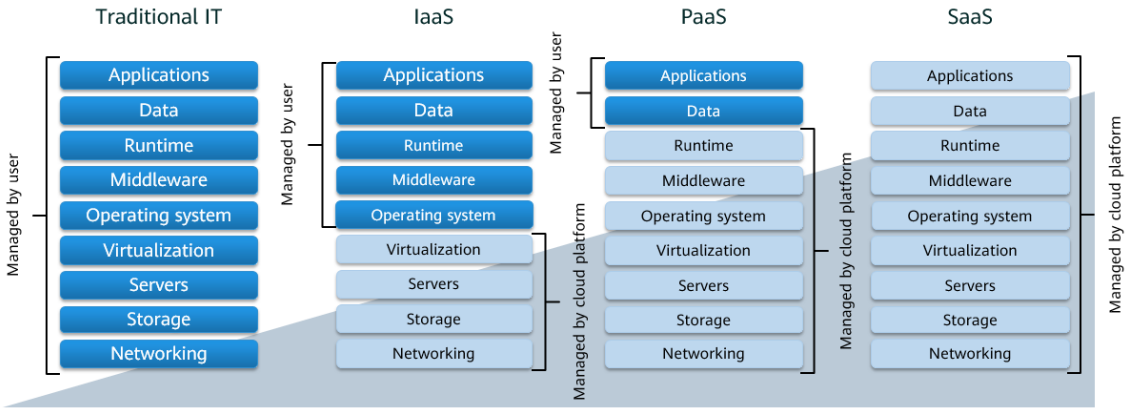


Figure 1-4 Cloud computing operation models

From Figure 1-4, you can see that all models have the same hierarchical architecture. Users mainly just interact with the applications. Data is generated during your use of the applications. An application can run only after the lowest-layer hardware, the operating system (OS) running on the hardware, middleware running on the OS, and operating environment of the application are all in place. The architecture of cloud computing can be divided into three layers: software, infrastructure, and platform.

Applications and data comprise the software layer. Hardware resources (servers, storage, and networking resources) and virtualization are all infrastructure. The OSs, middleware, and runtime environments are part of the platform layer.

In an IaaS model, cloud service providers provide the infrastructure layer, and users are responsible for other layers. With PaaS, cloud service providers are responsible for the infrastructure and platform layers, and users are responsible for software layer. In an SaaS model, cloud service providers are responsible for all three layers.

We can use an example to illustrate these models. For a popular triple-A game, for example, *Sekiro: Shadows Die Twice*, you need a certain OS, graphics card and so on. Figure 1-5 shows the specs described by the Activision, the game publisher.

System Requirements			
Minimum Requirements		**Recommended Requirements**	
OS	Windows 7 64-bit | Windows 8 64-bit | Windows 10 64-bit	OS	Windows 7 64-bit | Windows 8 64-bit | Windows 10 64-bit
CPU	Intel Core i3-2100 | AMD FX-6300	CPU	Intel Core i5-2500K | AMD Ryzen 5 1400
RAM	4 GB	RAM	8 GB
Storage	25 GB available space	Storage	25 GB available space
Graphics	NVIDIA GeForce GTX 760 | AMD Radeon HD 7950	Graphics	NVIDIA GeForce GTX 970 | AMD Radeon RX 570

Figure 1-5 Configuration requirements of *Sekiro: Shadows Die Twice*

In this figure, we can see the hardware requirements for the game. If you bought the PC yourself, then you are also responsible for the OS and for meeting the hardware requirements and installing the software yourself. That's how traditional IT architecture works. In an IaaS model, there is no hardware to deal with. You just buy a cloud server with whatever specifications you need, install the OS from an image, and then download and install your game. But you still might run into errors like this one:

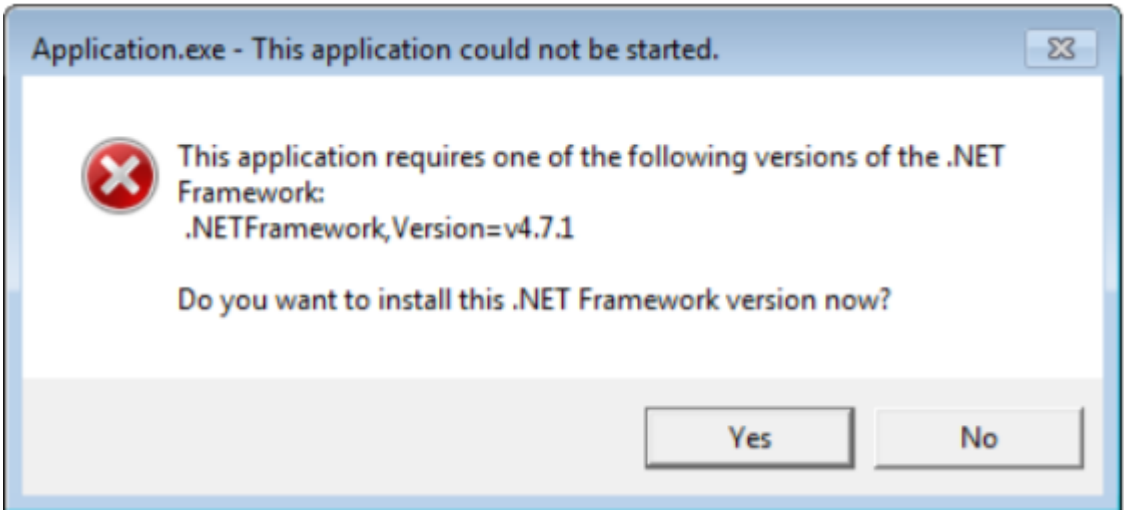


Figure 1-6 .NET Framework initialization error

The error occurs because the .NET Framework was not installed. In an IaaS model, runtime environments, like .NET are still your own responsibility.

In a PaaS model, the service provider takes care of the runtime too, so if you purchase a cloud server with the OS and .NET Framework already installed, that is PaaS.

And finally, if you purchase a cloud server with the OS, .NET Framework, and game software all preinstalled, and all you need to do is enter your username and password to play the game, that is SaaS.

1.1.5 Cloud Computing Characteristics

More and more enterprises and individuals are choosing cloud computing. Cloud computing is just better suited to the requirements of today's enterprises. So let's examine some of the characteristics and advantages of cloud computing driving its growing popularity.

1.1.5.1 Self-Service Resources

What's the first thing that comes to mind when someone says something is "self-service"? Shopping in a supermarket is a good example. In a supermarket, you browse the products and pick what you want based on your specific requirements. You can compare product descriptions, prices, and brands, and determine which one to purchase based on how well priced they are or other factors. Selecting items off the shelf is a self-service experience. Similarly, you can browse and select from different apps and services available on Huawei Cloud without needing any sales personnel to assist you.

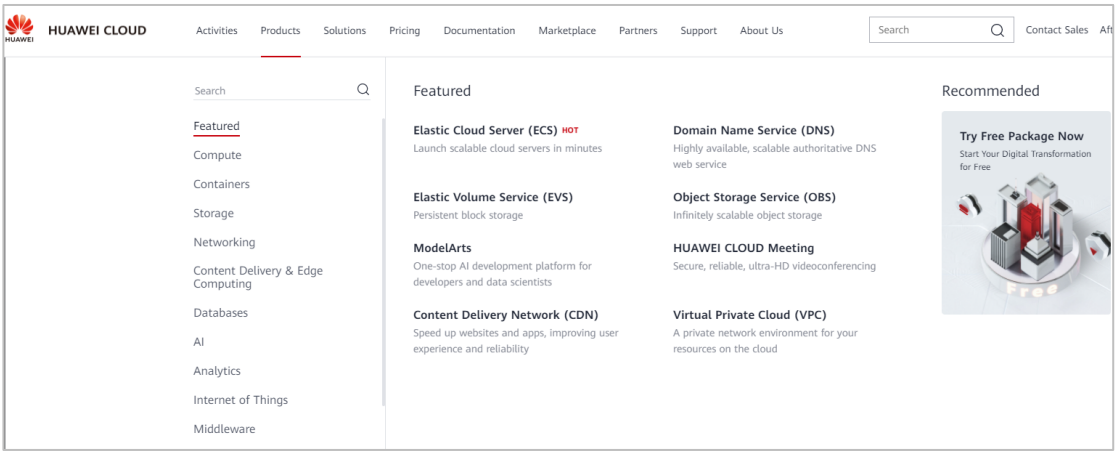


Figure 1-7 Featured cloud services

One of the prerequisites to selecting your own resources is knowing what your requirements actually are and being able to identify which products can meet those requirements. A supermarket offers an enormous variety of products, and a cloud service provider provides a wide range of cloud products, as shown in Figure 1-7. You need to know which product can suit your needs before placing an order.

1.1.5.2 Broad Network Access

As cloud computing is essentially providing computing power over the Internet, network access is inseparable from cloud services.

Now, almost everyone has access to the Internet from a PC, tablet, or phone, which means convenient, easy access to cloud computing for just about anyone.

As long as you have a network connection, you can access resource pools offered by cloud service providers through the network. You can take advantage of cloud services anytime and anywhere.

1.1.5.3 Resource Pooling

Resource pooling is another prerequisite for on-demand self-service. Resource pooling allows us to not only group similar products together, but also to sell them in more precise amounts. In a supermarket, products are located throughout the store based on what type of product they are, to make it easier for customers to find what they want. But this is a type of resource classification, not resource pooling. What is resource pooling?

With resource pooling, you do not just group products by type. You also break up the resources into the smallest possible unit.

Take a box of pasta, for example. You might have only enough pasta left one night for one person but you have friends coming over. You do not need another whole box of pasta, but what you have left in your pantry is not enough for dinner tonight. Unfortunately, the smallest unit for pasta is a whole box. Resource pooling would be if, instead of boxes of pasta, you could buy pasta by the weight. You could buy exactly as much or as little as you needed. A cafeteria is another good example. In a cafeteria, there might be fruit juices on tap, sorted by flavor. You can walk up to the juice bar and take as much of whichever flavor you need. You do not have to buy a whole bottle if you just want a taste.

Another function of resource pooling is to obscure the differences between different resources. It is like if the soda gun at a bar had both Pepsi and Coke, but the menu just offered "cola". In cloud computing, resources that can be pooled include compute, storage, and network resources. Compute resources include CPUs and memory. If CPUs are pooled, their smallest unit is core, and CPU vendors such as AMD and Intel are not displayed.

1.1.5.4 Rapid Deployment and Elastic Scaling

Service demand may fluctuate, so to ensure stable services during peak hours, enterprises scale out capacity, adding servers. When the traffic drops back down again later, they can easily scale back down, removing the extra servers. This is called rapid elastic scaling.

In cloud computing, you can scale resources manually or automatically (using preset policies). You can add or remove servers, or you can scale up or down the capacity of a single server.

This feature inexpensively ensures the stability of services and applications. When an enterprise is in its infancy, it can purchase a small amount of resources, and then purchase more as the business grows. During special events, all resources can be used for key services as needed, and then later, idle resources can be reallocated for other purposes. Anytime existing resources cannot keep up with demand, you can purchase additional resources on-demand, and then release the additional resources again when demand drops back down. Cloud computing makes adding or removing resources extremely convenient.

1.1.5.5 Service Metrics

Services in cloud computing are measured by time-in-use, by resource quotas, or by traffic carried. These metrics enable auto scaling based on service volume. They allow for much more precise resource allocation.

Cloud computing allows for a clear view of how you are using any services you have purchased. Precise metrics allow for precise purchasing.

It should be noted that metering is not billing, but clear metering makes billing easier.

Most cloud computing services are billable, but not all. Some services are free. For example, Auto Scaling (AS) itself is a free service, although you will still be billed for the additional resources that are added based on an AS policy.

1.2 Cloud Computing Technologies

The main categories of cloud computing services are compute, networking, and storage.

Compute includes virtualization and container technologies. Networking products include traditional networks and virtual network technologies. Storage includes block storage, file storage, and object storage. Let's start with a look at compute technologies.

1.2.1 Compute

1.2.1.1 Virtualization

Virtualization is an actual technology. Virtualization is the process of creation of multiple independent virtual servers from a single physical server. Generally, an application can run only after it is installed on an OS, and a physical server can run only one OS at a time. But by dividing the physical resources into multiple virtual servers, each with its own OS, you can run multiple OSs on the same physical server.

Virtualization is the transformation of physical devices into folders or files so that software can be decoupled from the hardware it runs on.

A physical server is tangible. We can check items against a device list or physical configuration list. For example, in Figure 1-8, we can see the CPUs, memory, hard disks, and network interface card (NIC).

Figure 1-8 Physical server components

Virtualization transforms a physical server into folders or files that contain virtual machine (VM) configuration information and user data.

When multiple programs are running on the same OS of a physical server, there can be application conflicts and performance problems. Only running a single program on each server would address this problem but could end up wasting a tremendous amount of resources.

With virtualization, multiple VMs can run on a physical server, and each VM runs its own OS, which improves hardware utilization. Virtualization also frees applications from being shackled to a single server by allowing dynamic VM migration within a cluster without interrupting services and affecting user experience. Dynamic VM migration brings the added benefits of high availability (HA), dynamic resource scheduling (DRS), and distributed power management (DPM). It also enables service mobility, server consolidation, and fault tolerance for enterprise data centers, reducing the cost of operations and management.

Virtualization involves the following four features:

- Partitioning: Multiple VMs can run on the same physical server at the same time. The virtual machine monitor (VMM) has the capability of allocating underlying server resources to VMs. We call this partitioning.
- Isolation: The VMs on a server are logically isolated from each other. If a VM on a server crashes or is infected with a virus, the other VMs on the server are not affected.
- Encapsulation: A VM exists as a type of file. You can migrate a VM by just moving the file, copy and paste style.
- Independence: VMs are independent from the hardware. A VM can run without any modifications after it is migrated to a new physical server.

1.2.1.2 Containers

Containers use a lightweight OS virtualization technology. This technology allows an OS to be divided into independent units, each running on the same kernel, but independently, so they do not interfere with each other. These independent units are called containers.

With containers, applications can run almost anywhere the same way. Developers can create and test containers on their own computers and run them on VMs or other computers without any modifications.

Like virtual machines, containers also use virtualization.

A container consists of two parts:

- An application
- Running environment of the application, for example, libraries required by the application

A container is a lightweight, independent, and executable software package. A container contains an application and an environment for the application to run in. Both Linux and Windows applications can run in containers regardless of the environment. Containers isolate applications from their surrounding environments, so you can ignore differences between development and test environments. Teams running different software on the same infrastructure will face far fewer conflicts. One big difference between VMs and containers is that VMs have OSs installed, while containers do not.

So what problems have containers solved?

The modern software architecture is very complex, and configuring the software environment has become a big challenge for software development. Developers often say, "It runs fine on *my* computer," referring to the fact that it may not run on other computers. This is because most programs are dependent on various OS settings, and on various libraries and other components. For program written in Java, for example, you need to configure the engine, dependencies, and environment variables for the program to run properly.

You may say, that is what a VM is for, and while it is true that a VM can address this issue, VMs also have a number of downsides.

1. They use more resources than containers.

   The VMs themselves use up valuable memory and disk space that could otherwise be allocated to the applications. You can end up allocating hundreds of MB of memory to run an application that only needs 1 MB.

2. There are too many steps involved.

   A VM comes with an OS, so you cannot skip system-level steps, such as user login.

3. They start up slower than containers.

   Applications cannot run until the OS boots up, which may take several minutes.

Containers resolve all of these problems. They have following advantages:

1. They start up faster.

   Containers run like a process running directly on the lower level system resources. They do not need a VM in the middle. Starting a container is like starting a process

on the local host. There is no need to launch an entire OS, so the container can launch much faster.

2. They are much smaller.

   A container only needs a few components to run. A VM contains an entire OS. Containers are much smaller than VMs.

3. They are much lighter weight.

   A container only uses the resources it needs to a specific application, unlike a VM, which has to occupy all the resources required for an entire OS, regardless of what the application in the VM needs. Containers can also share OS resources, but resources allocated to a VM are occupied exclusively by that VM.

1.2.1.3 Comparison Between Containers and VMs

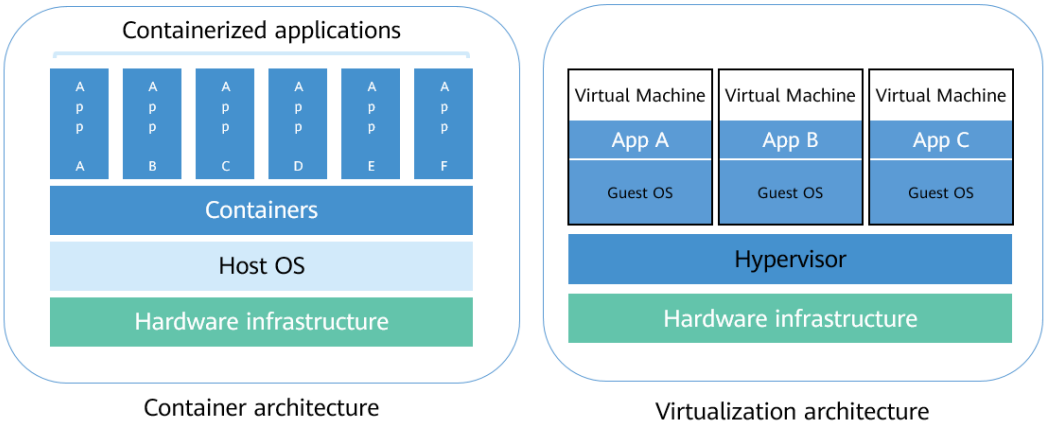


Figure 1-9 Comparison between container and VM architectures

Containers and VMs have similar advantages in terms of resource isolation and allocation, but have different architectures. Containers virtualize system resources while VMs virtualize the hardware. Containers are more portable and efficient.

A hypervisor sits on the software layer, between a physical server and an OS. It allows multiple OSs and applications to share the same physical hardware. As shown in the figure above, there is no virtualization layer in the container architecture. This is what makes containers lightweight. The lack of the intervening hypervisor, means applications running in containers have better performance than they would in VMs.

Virtualization technology gives us VMs. Container technology gives us containers. Enterprise applications can be deployed in VMs or containers. How do we decide which one to use? We need to understand the differences between the two.

Feature	Container	VM
Startup speed	Seconds	Minutes
Virtualization type	OS virtualization	Hardware virtualization
OS	Containers share the host OS	Each VM is installed with an OS
Security	Process-level isolation, which is less secure	Complete isolation, which is more secure
Isolation policy	Namespace, CGroups	Hypervisor
Image size	KB to MB	GB to TB
Performance	Equivalent to physical machine	Limited performance
Per-machine capacity	More than one thousand containers	Dozens of VMs

Figure 1-10 Container vs VM

Figure 1-10 compares containers and VMs from multiple perspectives.

- Startup speeds

  The architectures of the two explain their difference in startup speeds. A VM is essentially the same as a personal computer, complete with its own OS, so starting up a VM is like starting up an actual computer, and no applications can be run on a VM until after the OS is up and running. A container does not have its own OS. It uses the host OS. Launching containers is more like launching a handful of system processes in a system. These processes are started when applications are started, so containers can be launched very quickly. So that is why a container takes only seconds to launch but a VM can take several minutes.

- Virtualization type

  Containers use OS-level virtualization and share host kernel resources. VMs use hardware-level virtualization. Each VM has an independent OS. Containers are isolated at the process level, but VMs are isolated at the kernel level. Therefore, VMs are more secure.

- Image size

  First, let's establish what an image is. An image is a VM template, but what is a VM template? A VM template is a VM and OS, along with various programs packaged into a single file with a standard file format. You can create a VM directly from the file. Container images are similar to VM images. You can use images to create, deploy, and start containers. A VM image contains an independent system and applications, and the size is measured in GB to TB. A container image contains only a specific application, so it is measured in just KB or MB.

- Performance

  Container technology is a form of lightweight virtualization. Containers do not involve a hypervisor layer. Running an application in a container is similar to running it on a physical machine. Applications running on a VM cannot be started until after the VM where they are deployed is up and running. In addition, the virtualization layer required for VMs reduces performance. This is why containers deliver better performance than VMs.

- Per-machine capacity

  Compared with VMs, containers are lightweight and the images are small, so a single physical machine can run far more containers than VMs.

1.2.2 Network

A network is made up of various network devices. Traditional IT systems use physical network devices to control traffic. For example, switches connect hosts via network cables or optical fibers. In cloud computing, in addition to physical network devices, many network devices are virtualized and run on servers. These network devices are not connected by physical cables, but by routes in route tables.

1.2.2.1 Traditional Network

Network devices are essential for enterprise infrastructure. Let's examine some basic concepts related to traditional networks.

1.2.2.1.1 Concepts

- Broadcast and Unicast

  Both broadcast and unicast are communication modes on the network. In addition to broadcast and unicast, there is another mode called multicast, which is not discussed in this course.

  Broadcast is when there is just one sender, but the information is sent to all connected receivers. When two devices communicate with each other for the first time, the sender transmits a broadcast packet that will reach all possible devices on the network. All the devices check the content of the packet. If a device finds itself is the intended receiver, it will send a unicast packet to the sender. Otherwise, the device discards the broadcast packet.

  A lot of applications use broadcast, for example, Dynamic Host Configuration Protocol (DHCP), which uses different broadcast addresses in different scenarios. For example, if the network is 192.168.1.0/24, the broadcast address is 192.168.1.255. However, the broadcast address used by the DHCP client to search for a DHCP server is 255.255.255.255.

  Unicast is when there is just one sender, and one receiver. In half-duplex mode, each device can both send and receive, but not at the same time. When one device is sending, the other can only receive, and vice versa. In full-duplex mode, both devices can send and receive simultaneously.

  Most data on the network is transmitted in unicast mode, for example, connecting to email, web, or game servers before you send emails, view web pages, or play online games.

  When a device broadcasts a packet, the only useful information in the packet might be the source address and destination address. When a device unicasts a packet, the packet is likely to contain more useful information. If there are too many broadcast packets on a network and the bandwidth is fixed, the network gets congested. Broadcast packets can be received by all reachable destinations. However, broadcast packets can be spoofed, which can result in information leaks or even network

breakdowns. However, broadcast is necessary before a unicast connection can be established. To mitigate network congestion and security issues, network engineers divide a broadcast domain into multiple smaller ones and then use routes and a default gateway to enable communication between broadcast domains.

- Route and Default Gateway

  Back before we all had cellphones to make long-distance calls, you used to have to enter an area code before dialing the other party's phone number. To learn the area code, you might have needed to look it up in the yellow pages or some other directory. When making a long-distance call, the area code was like the route, and the yellow pages, or whatever book you used to look it up, was like the route table.

  If there are a large number of broadcast domains, the route table contains a lot of routes. Each time a device communicates with another device, the device needs to search for routes, which adversely affects network communication efficiency. To solve this problem, a default gateway is used. When the default gateway receives a communication request, it checks if its own route table contains a route with the destination you need. If there is such a route, the default gateway forwards the request based on that route. If there is no such a route, the default gateway returns a message indicating that the destination is unreachable.

  The default gateway is used if there is no route appropriate for a given destination address.

- VLAN

  To divide a broadcast domain into more manageable pieces, you can use a Virtual Local Area Network (VLAN). A VLAN is a subdivision of a physical LAN into separate logical broadcast domains. All hosts in a VLAN can communicate with each other, while hosts in different VLANs cannot. The following figure shows the details.

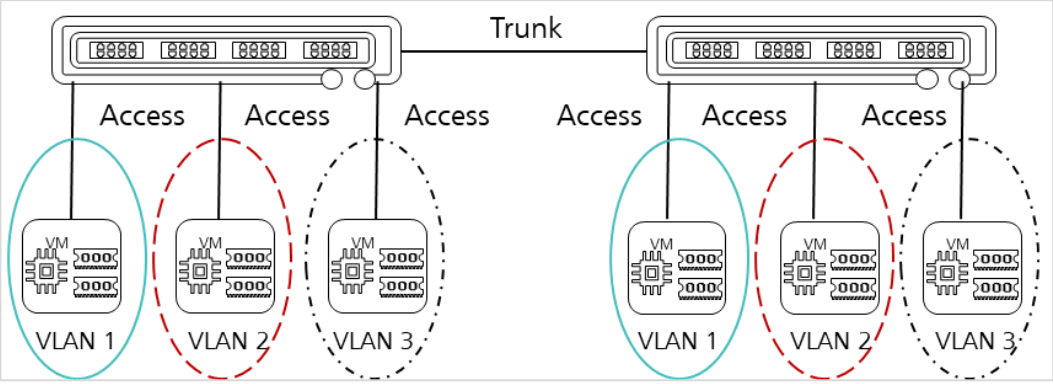


Figure 1-11 VLAN benefits

VLAN brings the following benefits:

- A broadcast domain is restricted to a single VLAN to save bandwidth and improve network processing.
- Packets in different VLANs are isolated from each other during transmission. Users in one VLAN cannot directly communicate with users in other VLANs.
- If one VLAN is faulty, other VLANs are not affected.

- VLAN can be used to divide users into different working groups. Users in the same working group do not have to be limited to a certain fixed physical range, so network construction and maintenance are more convenient and flexible.

A 4-byte 802.1Q tag is inserted in an Ethernet data frame to identify which VLAN the frame belongs to. The following figure shows the details.

Figure 1-12 Traditional Ethernet data frame

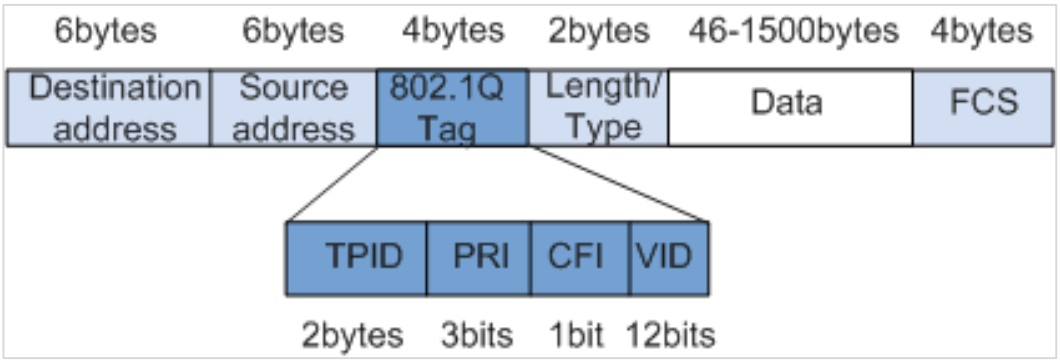


Figure 1-13 Tagged Ethernet data frame

A packet sent by a switch that supports the 802.1Q protocol includes a VLAN ID to identify which VLAN the packet belongs to. The following are the two types of Ethernet frames in a VLAN:

- Tagged frame: frame with a 4-byte 802.1Q tag
- Untagged frame: frame without a 4-byte 802.1Q tag

The operating system or switch port can add a tag to a data frame. Generally, the switch adds or removes tags for frames. There are access links and trunk links:

- Access link: connects a host to a switch. Generally, a host does not know which VLAN it belongs to, and host hardware cannot distinguish frames with VLAN tags. Hosts send and receive only untagged frames.
- Trunk link: connects a switch to another switch or to a router. Data of different VLANs is transmitted along a trunk link. The two ends of a trunk link must be able to distinguish frames with VLAN tags. Only tagged frames are transmitted along trunk links.

  After the 802.1Q defines VLAN frames, some ports on the device can identify VLAN frames, while others cannot. Depending on whether VLAN frames can be identified, ports can be classified as either access ports or trunk ports:

- An access port on a switch connects to the port on a host and can only connect to an access link. Only the VLAN whose ID is the same as the default VLAN ID is allowed on the access port. Ethernet frames sent from the access port are untagged.

- A trunk port on a switch connects to another switch and can only connect to a trunk link. Multiple tagged VLAN frames are allowed on the trunk port.

  Each port can be configured with a default VLAN with a port default VLAN ID (PVID). The meaning of the default VLAN varies according to the port type. The default VLAN ID of almost all switches is 1.

1.2.2.1.2 Network Devices

In virtualization, workloads are deployed on VMs, and VMs run on physical servers. Before connecting VMs to a network, the physical servers need to be connected to the network first. To do so, devices such as routers, layer-3 switches, layer-2 switches, and server NICs are required.

Before talking about the physical devices, we need to understand the seven-layer OSI model.

OSI is a universal protocol stack. It divides the process of communication into seven layers: application, presentation, session, transport, network, data link, and physical.

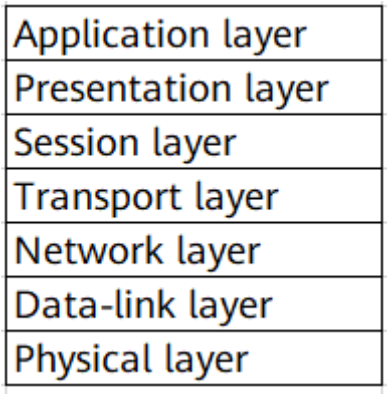


Figure 1-14 OSI protocol stack

Routers work at the network layer (layer 3), and VLANs work at the data link layer (layer 2). If a device has routing functions that let you check route tables, it is a layer-3 device. If a device can only create VLANs, it is a layer-2 device. A hub is a physical layer (layer 1) device. It can function as a switch but it cannot divide VLANs, so it does not qualify as a layer-2 device.

Routers and layer-3 switches work at the network layer. Layer-2 switches work at the data link layer. Physical server NICs and network cables and optical fibers that connect NICs working at the physical layer.

- Routers and Layer-3 Switches

  Both routers and layer-3 switches work at the network layer. Layer-3 switches include routing, and many routers configured with an additional switching board can create VLANs, but routers and layer-3 switches are not interchangeable.

  First, their primary functions are different. A switch uses ASIC for high-speed data exchange, while a router maintains a route table for routing traffic to different destinations and can separate broadcast domains. Even if a router can perform switching or if a switch can do routing, those are supplementary capabilities. Their key functions remain unchanged.

Second, switches are mainly used for LANs. Routers are mainly used for WANs. On a LAN, data tends to be exchanged frequently over a large number of just a few types of interfaces. Switches enable fast data forwarding and have two types of interfaces, including Ethernet cable interfaces (RJ45) and optical fiber interfaces. Each switch has many interfaces, which is what you need for a LAN. On a WAN, in contrast, there are many different types of networks and interfaces. Routers can provide powerful routing functions not only for LANs using the same protocol but also for LANs and WANs using different protocols. Routers can select the best routes, balance loads, back up links, and exchange routing information with other networks.

Third, a Layer 3 switch can deliver better data exchange performance than a router. Technically speaking, a router uses a software engine with a micro-processor to forward packets while a layer-3 switch uses hardware-based forwarding. After a layer-3 switch forwards the first packet of a data flow, it maps the MAC address to an IP address. That way, the next time data flows through the layer-3 switch, it can forward the packets without routing again. This eliminates latency caused by route selection and improves the efficiency of packet forwarding.

In addition, a layer-3 switch searches for routes for data flow and uses the ASIC cache technology for faster and less expensive forwarding. The router uses the longest match rule to forward packets. This complex process usually uses software and has a low forwarding efficiency. Therefore, in terms of performance, a layer-3 switch is better than a router and is applied to the LAN with frequent data exchange. With a powerful routing function and low packet forwarding efficiency, a router is applied to the connection of different types of networks without frequent data exchange, such as the connection between the LAN and Internet. If a router, especially a high-end router, is used on a LAN, not only its powerful routing functions are wasted, it also cannot meet the communication requirements of the LAN.

In cloud computing and virtualization, routers are usually deployed at the egress of an enterprise or institution network to connect to the Internet. To enable communications between an enterprise or institution network and the Internet, routers perform route forwarding and network address translation (NAT).

- Access Switches

  After servers are connected to the network, network traffic is classified as service, management, and storage traffic based on the usage. When services are accessed, service traffic is generated. If service data is stored on a professional storage device instead of a local server, then when the server accesses that storage device, storage traffic is generated. Management traffic refers to the traffic generated when users manage servers, VMs, and storage devices. Almost every physical device is configured with an independent management interface. If management traffic and service traffic are carried on different physical lines and interfaces, this is out-of-band management. If management traffic and service traffic are carried on a same physical channel, this is in-band management.

  In a cloud computing data center, a high-end layer-3 switch is used as the core of the entire network during network design. The default gateways of networks that generate traffic are configured on the switch. In this way, all traffic generated across broadcast domains will pass through the switch.

- Physical NIC

  A physical server uses its own physical NIC to connect to a network. A VM communicates with a network through various types of network ports.

1.2.2.2 Virtual Network

As cloud computing and virtualization have become more popular, layer-2 switches have had to be deployed on servers to connect to VMs. This has driven the need for virtual switches.

Cloud computing and virtualization have advantages. They are growing more and more mainstream. However, any new technology tends to introduce new challenges. Virtualization has evolved to the point where services are typically not carried on a physical server any more. A physical server used to have at least one physical cable connection to a switch, and services running on the server would share the same physical cable connection. Now, multiple VMs run on one physical server and use one cable connection to carry multiple types of traffic. The means new challenges include figuring out how to manage all the different types of traffic and how to monitor all their different statuses. To do this, we need to understand the virtual network architecture.

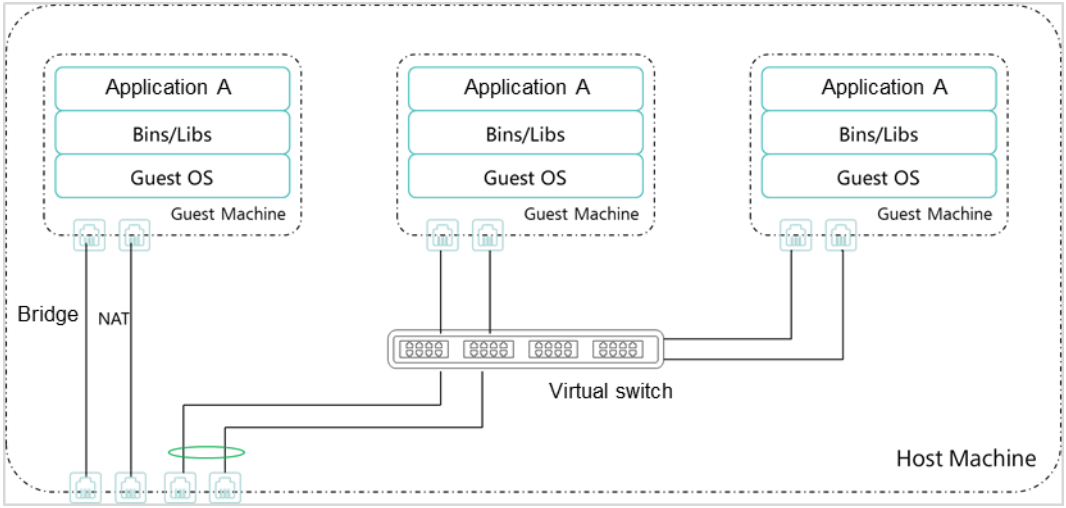


Figure 1-15 Virtual network architecture

The preceding figure shows a common virtual network architecture. In a personal or small-scale virtualization system, VMs are connected to physical NICs using bridges or NAT. In a large-scale virtualization system, VMs are connected to physical networks using virtual switches.

The VM network bridge and NAT in the preceding figure use bridge technology. You can use network bridges to interconnect multiple network ports so that packets received on a network port will be replicated to other network ports. In a virtualization system, the OS is responsible for interconnecting all network ports. The following figure shows bridge-based interconnection in a Linux system.

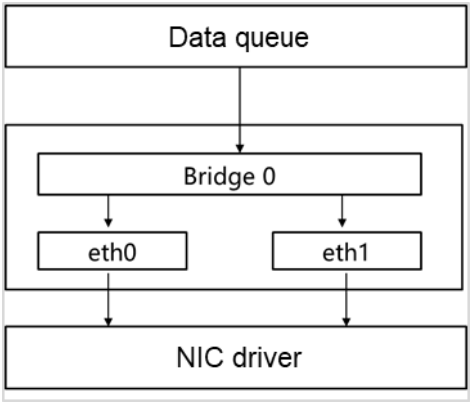


Figure 1-16 Bridge-based interconnection in a Linux system

Bridge 0 (network bridge device) is bound to eth0 and eth1. The upper-layer network protocol stack only knows Bridge 0. It does not need to know any other bridge details because the bridge is performed at the data link layer. When the upper-layer network protocol stack needs to transmit packets, it sends the packets to Bridge 0, and Bridge 0 determines whether the packets will be forwarded to eth0, eth1, or both. Similarly, packets received on eth0 or eth1 are sent to Bridge 0, and the Bridge 0 determines whether to forward, discard, or send the packets to the upper-layer network protocol stack.

A bridge has the following functions:

- MAC learning: Initially, the network bridge does not know any of the MAC addresses and ports of connected services. It just forwards data like a HUB. However, when transmitting data, the network bridge learns the MAC addresses and their associated ports, and stores this information in a Content Addressable Memory (CAM) table.
- Packet forwarding: When transmitting a packet, the network bridge obtains its destination MAC address and checks the CAM table to see which port the packet will be sent through.

Each VM has a virtual NIC. A Linux OS will have a TAP device that runs in user mode, but the TUN/TAP device driver and virtual NIC driver run in kernel mode. Data sent from a VM towards the bridge first passes through the TUN/TAP character device file. This part is performed in user mode. Then, in kernel mode, the data passes through the TUN/TAP device driver and the virtual NIC driver, and is processed through the TCP/IP protocol stack. This happens in kernel mode. Finally, the data goes through the virtual NIC tap and on to the bridge, all in user mode again. The tap is now directly connected to Bridge0, and eth0 and eth1 from the preceding figure will act as the NIC tap of the VM. The following figure illustrates the process.

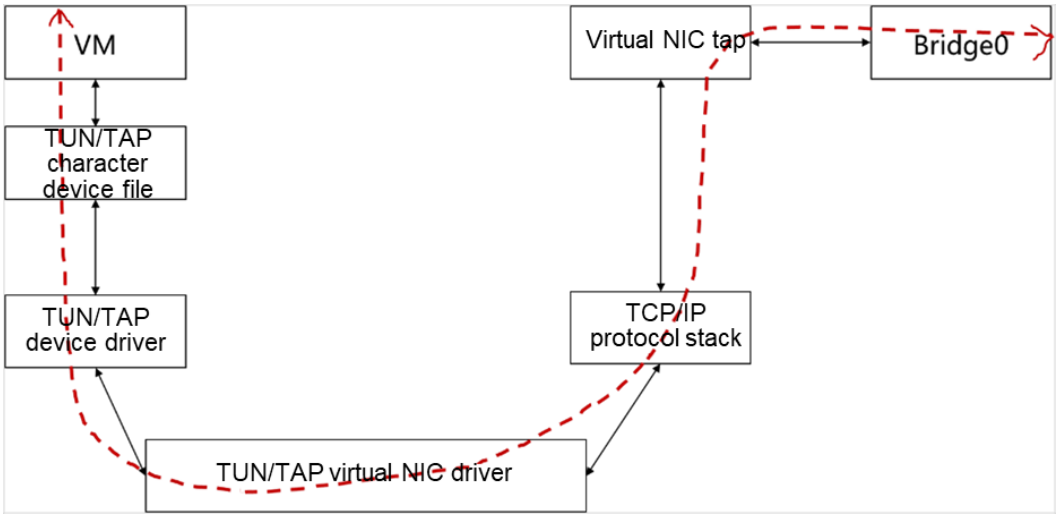


Figure 1-17 Traffic over a network bridge

VMs can communicate with external networks using bridges and NAT. When a bridge is used, the bridge functions as a switch, and the virtual NIC is connected to a port of the switch. If NAT is used, the network bridge functions as a router, and the virtual NIC is connected to a port of the router.

The bridge and NAT are suitable for personal or small-scale systems. If a bridge is used, the statuses of virtual NICs cannot be viewed and the traffic on virtual NICs cannot be monitored. A bridge only supports GRE tunneling. It does not support software-defined networking (SDN), which is currently a popular option. That is why in a large-scale system, a virtual switch is used for VMs to communicate. A virtual switch is like an upgraded network bridge. It provides all the same benefits but without the drawbacks.

Now, let's talk about virtual switches. The functions of a virtual switch are the same as those of bridge and NAT. Virtual switches enable traffic from a VM to be transmitted out across the network. The virtual switches are simulated by using the virtualization technology. Common virtual switches include OVS and EVS.

- Open vSwitch (OVS): An OVS is an open-source software-based virtual switch. It supports multiple standards and protocols with additional support for the OpenFlow protocol, and can be integrated with multiple open-source virtualization platforms. The OVS can be used to transmit traffic between VMs and for communication between VMs and the external network.
- Enhance vSwitch (EVS): An EVS is an enhanced OpenFlow-standard virtual switch that provides the same capabilities as an OVS, but with better I/O performance, thanks to the Intel Data Plane Development Kit (DPDK). DPDK allows the NIC to offload packet processing to user-mode processes, which significantly improves I/O performance.

The OVS and EVS are mainly different in terms of how they process the traffic. On an OVS, the traffic is received and sent in kernel mode, while on an EVS, the traffic is processed in user mode.

1.2.3 Storage

When it comes to storage, we may think of storage media commonly found around us, such as easy-to-carry USB drives and hard disks used in computers. We all know that storage is used for data storage and access. Enterprises, however, usually purchase dedicated devices to support their applications because they have a large amount of data to store and work with. In the following sections, we will learn how they store and access data. Mainstream storage is classified into three types: block storage, file storage, and object storage.

1.2.3.1 Three Storage Types

1.2.3.1.1 Block Storage (SAN)

Block storage commonly uses an architecture that connects storage devices and application servers over a network. This network is used only for data access between servers and storage devices. When there is an access request, data can be transmitted quickly between servers and backend storage devices as needed. From a client's perspective, block storage functions in much the same way as disks. One can format a disk with any file system and then mount it. A major difference between block storage and file storage is that block storage only provides storage space, leaving the rest of the work, such as file system formatting and management, to the client.

Block storage uses evenly sized blocks to store structured data. In block storage, data is stored without any metadata, which makes block storage useful when applications need to strictly control data structure. The most common usage is for databases. Databases can read and write structured data faster with raw block devices.

Currently, block storage is usually deployed in Fibre Channel Storage Area Network (FC SAN) and Internet Protocol Storage Area Network (IP SAN) based on the protocols and connectors used. FC SAN uses the Fibre Channel protocol to transmit data between servers (hosts) and storage devices, whereas, IP SAN uses the IP protocol for communication. FC technology can meet the growing needs for high-speed data transfer between servers and large-capacity storage systems. With the FC protocol, data can be transferred at a faster rate and with low protocol overheads, while still maintaining certain network scalability.

Block storage has the following advantages:

- Offers long-distance data transfer with a high bandwidth and a low transmission bit error rate.
- Based on SAN architecture and massive addressable devices, multiple servers can access a storage system over the network at the same time, eliminating the need for purchasing storage devices for every single server. This reduces the heterogeneity of storage devices and improves storage resource utilization.
- Protocol-based data transmission can be handled by the host bus adapter (HBA), occupying less CPU resources.

In a traditional block storage environment, data is transmitted over the fibre channel via block I/Os. To leverage the advantages of FC SAN, enterprises need to purchase additional FC components, such as HBAs and switches. Enterprises usually have an IP network-based architecture. As technologies evolve, block I/Os now can be transmitted over the IP network, which is called IP SAN. With IP SAN, legacy infrastructure can be reused, which is far more economical than investing in a brand new SAN environment. In addition, many remote and disaster recovery solutions are also developed based on the IP network, allowing users to expand the physical scope of their storage infrastructure.

Internet SCSI (iSCSI), Fibre Channel over IP (FCIP), and Fibre Channel over Ethernet (FCoE) are the major IP SAN protocols.

- iSCSI encapsulates SCSI I/Os into IP packets and transmits them over TCP/IP. iSCSI is widely used to connect servers and storage devices because it is cost-effective and easy to implement, especially in environments without FC SAN.
- FCIP allows FCIP entities, such as FCIP gateways, to implement FC switching over IP networks. FCIP combines the advantages of FC SAN and the more mature, widely-used IP infrastructure. This gives enterprises a better way to use existing investments and technologies for data protection, storage, and migration.
- FCoE achieves I/O consolidation. Usually, one server in a data center is equipped with two to four NICs and HBAs for redundancy. If there are hundreds of servers in a data center, the numerous adapters, cables, and switches required make the environment complex and difficult to manage and expand. FCoE achieves I/O consolidation via FCoE switches and Converged Network Adapters (CNA). CNAs replace the NICs and HBAs on the servers and consolidate IP traffic and FC traffic. In this way, servers no longer need various network adapters and independent networks, thus the requirement of NICs, cables, and switches is reduced. This massively lowers costs and management overheads.

Block storage is a high-performance network storage, but data cannot be shared between hosts in block storage. Some enterprise workloads may require data or file sharing between different types of clients, and block storage cannot do this.

1.2.3.1.2 File Storage (NAS)

File storage provides file-based, client-side access over the TCP/IP protocol. In file storage, data is transferred via file I/Os in the local area network (LAN). A file I/O is a high-level request for accessing a specific file. For example, a client can access a file by specifying the file name, location, or other attributes. The Network Attached Storage (NAS) system records the locations of files on disks and converts the client's file I/Os to block-level I/Os to obtain data.

File storage is a commonly used type of storage for desktop users. When you open and close a document on your computer, you use a file system. Clients can access file systems in file storage for file upload and download. Protocols used for file sharing between clients and storage include CIFS (SMB) and NFS. In addition to file sharing, file storage also provides file management functions, such as reliability maintenance and file access control. Although there are differences in managing file storage and local files, file storage can basically be seen as a directory. One can use file storage in almost the same way as in using local files.

Because NAS access requires the conversion of file system format, it is not suitable for applications using blocks, especially database applications that require raw devices.

File Storage has the following advantages:

- Comprehensive information access: Local directories and files can be accessed by users on other computers over LAN. Multiple end users can collaborate with each other on the same files, such as project documents and source code files.
- Good flexibility: NAS is compatible with both Linux and Windows clients.
- Low cost: NAS uses common and low-cost Ethernet components.

1.2.3.1.3 Object Storage

Users who frequently access the Internet and use mobile devices often need object storage technology. The core of object storage is to separate the data path from the control path. Object storage does not provide access to original blocks or files, but to the entire object data via system-specific APIs. You can access objects using HTTP/REST-based uniform resource locators (URLs), like you access websites using browsers. Object storage abstracts storage locations as URLs so that storage capacity can be expanded in a way that is independent of the underlying storage mechanism. This makes object storage an ideal way to build large-scale systems with high concurrency.

As the system grows, object storage can still provide a single namespace. This way, applications or users do not need to worry about which storage system they are using. By using object storage, you do not need to manage multiple storage volumes like using a file system. This greatly reduces O&M workloads.

Object storage has many advantages in processing unstructured data over traditional storage and delivers the advantages of both SAN and NAS.

It can distribute object requests to large-scale storage cluster servers. This enables an inexpensive, reliable, and scalable storage system for massive amounts of data. It is independent of platforms or locations, offering scalability, security, and data sharing:

- Security: data consistency and content authenticity. Object storage uses special algorithms to generate objects with strong encryption. Requests in object storage are verified in storage devices instead of using external verification mechanisms.
- Platform-independent design: Objects are abstract containers for data (including metadata and attributes). This allows objects to be shared between heterogeneous platforms, either locally or remotely, making object storage the best choice in cloud computing.
- Scalability: The flat address space used enables object storage to store a large amount of data without compromising performance. Both storage and OSD nodes can scale independently in terms of performance and capacity.

OSD intelligently manages and protects objects. Its protection and replication capabilities can be self-healed, enabling data redundancy at a low cost. If one or more nodes in a distributed object storage system fail, data can still be accessed. In such cases, three data nodes concurrently transfer data, making the transfer fast. As the number of data node servers increase, read and write speed up accordingly. In this way, performance is improved.

1.2.3.1.4 Summary of Block Storage, File Storage, and Object Storage

In block storage, file systems reside on top of application servers, and applications directly access blocks. The FC protocol is usually used for data transfer, and it has a higher transmission efficiency than the TCP/IP protocol used in file storage. The header of each protocol data unit (PDU) in TCP/IP is two times larger than the header of a data frame in FC. In addition, the maximum length of an FC data frame is larger than that in Ethernet. But data cannot be shared between hosts in block storage. Some enterprise workloads may require data or file sharing between different types of clients, and block storage cannot do this. In addition, block storage is complex and costly because additional components, such FC components and HBAs, need to be purchased.

File systems are deployed on file storage devices, and users access specific files, for example, opening, reading from, writing to, or closing a file. File storage maps file operations to disk operations, and users do not need to know the exact disk block where the file resides. Data is exchanged between users and file storage over the Ethernet in a LAN. File storage is easy to manage and supports comprehensive information access. One can share files by simply connecting file storage devices to a LAN. This makes file sharing and collaboration more efficient. But file storage is not suitable for applications that demand block devices, especially databases systems. This is because file storage requires the conversion of file system format and users access specific files instead of data.

Object storage uses a content addressing system to simplify storage management, ensuring that the stored content is unique. It offers terabyte to petabyte scalability for static data. When a data object is stored, the system converts the binary content of the stored data to a unique identifier. The content address is not a simple mapping of the directory, file name, or data type of the stored data. OBS ensures content reliability with globally unique, location-independent identifiers and high scalability. It is good at storing non-transactional data, especially static data and is applicable to archives, backups, massive file sharing, scientific and research data, and digital media.

1.2.3.2 Development of Enterprise Storage

In the early phase of enterprise storage, the physical disks used by enterprise storage are equipped on servers, and disks, CPUs, and memory are connected to the server main board. As the enterprise workloads grow and ICT technologies develop, the limitations and shortcomings of this architecture gradually emerge.

- Systems experience performance bottlenecks.
- Limited disk slots on servers fail to meat large capacity requirements.
- Data stored on individual disks has low reliability.
- Storage space utilization is low.
- Data is scattered in local storage systems.

Some storage technologies have been developed to cope with these shortcomings, among which Direct Attached Storage (DAS) was the first. DAS directly connects external storage devices to application servers over SCSI or FC interfaces, making storage devices part of the entire server structure. In this case, data and operating systems are often not separated. There are also some storage techniques used in DAS. For example, Just a

Bunch Of Disks (JBOD), which logically connects several physical disks in series to increase capacity but does not provide data protection. Redundant Arrays of Independent Disks (RAID) eliminates capacity limits and improves reliability at the same time. As the technologies evolve, new storage architectures, such as NAS and SAN, emerge.

NAS uses TCP/IP, ATM, and FDDI to establish the storage network, and connects storage systems and servers using switches. It integrates storage devices, network interfaces, and Ethernet technologies and allows for direct data access over Ethernet. This separates storage from file servers. SAN establishes a private network for data storage and connects storage arrays and servers through switches. SAN features fast data transmission, high flexibility, and reduced network complexity. It eliminates performance bottlenecks of the traditional architecture and massively improves the backup and disaster recovery efficiency of remote systems.

Today, most of the enterprises still use SAN and NAS storage architectures, which are regarded as traditional, centralized storage in the industry. As the Internet continues to rapidly develop, various applications and a growing number of users drive exponential growth of data, which adds significant pressure on local storage systems. To relieve local storage systems of this burden, distributed storage and distributed file systems have been introduced. Many enterprises are now trying distributed storage. Distributed storage virtualizes all disk space on each host of an enterprise to a virtual storage device. Data is distributed in the storage system, which improves system reliability, availability, and access efficiency.

Currently, common distributed storage products include Ceph (open source), HDFS, FusionStorage (Huawei), and vSAN (VMware).

2 Huawei Cloud

2.1 About Huawei Cloud

Huawei Cloud is the cumulation of thirty years of ICT infrastructure expertise. It synergizes cloud and AI technologies to provide powerful public cloud services while ensuring neutrality, security, and trustworthiness.

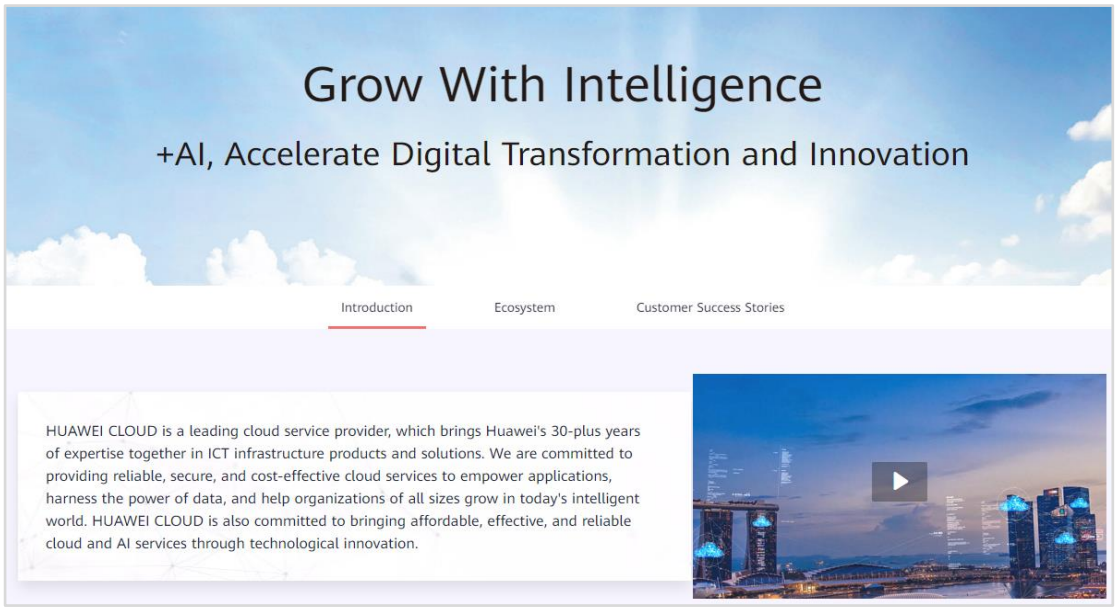


Figure 2-1 Huawei Cloud

In the previous chapter, we mentioned that public clouds are available for all users. You purchase any service you want on a public cloud portal simply with an official website account. All cloud service providers aim to minimize access latency when choosing their regions. In China, a place with mild winters and cool summers is ideal for a data center site to save on electricity. Guizhou, a province in southwestern China, is one such place. However, when a data center is deployed in Guizhou, the use of public cloud services in more distant regions would be prone to high latency due to the distance. As a result, public cloud vendors build data centers in regions that can ensure fast access from major cities. In addition to fast and stable access regardless of geographic location, diverse cloud services are essential for vendors to stay relevant. As one of the leading cloud service providers in the world, Huawei Cloud boasts the following competitive advantages:

- Wide node coverage

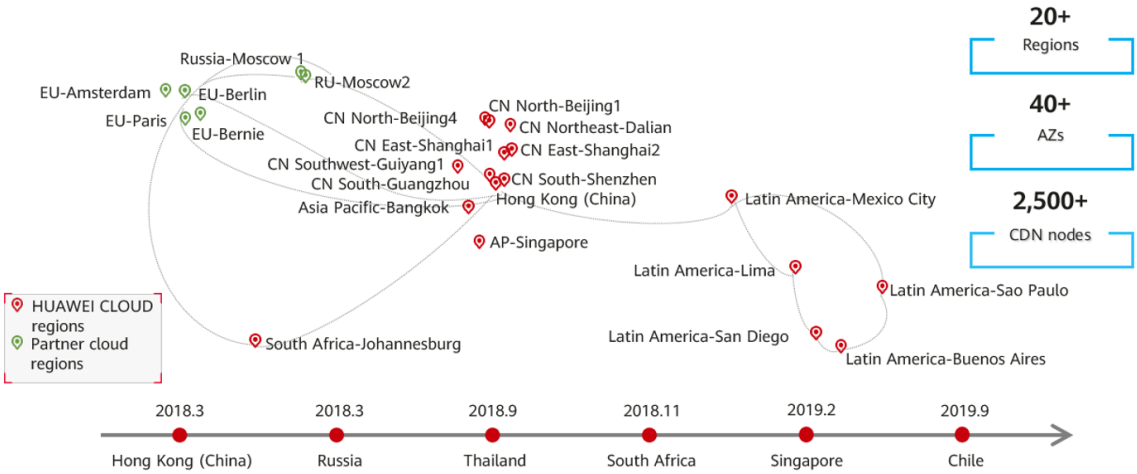


Figure 2-2 Huawei Cloud nodes around the world

In China, Huawei Cloud provides cloud services for high-speed interconnection nationwide on a *2 + 7 + N* architecture, where 2 indicates the two national data centers in Ulanqab and Guiyang. The national data center consists of three availability zones (AZs) that are 30 to 50 km apart. 7 indicates the major regional centers of Huawei Cloud, including CN North, CN East, CN South, and Hong Kong (China). *N* indicates the satellite nodes of Huawei Cloud. Each satellite node also functions as the e-Government cloud for the local government. Currently, these nodes are deployed in regions such as Ulanqab, Xiangyang, Yuxi, and Karamay.

Huawei Cloud is a brand with an international footprint. It has services in countries such as Singapore, Chile, Brazil, Mexico, and Peru, and joins hands with partners to run 45 AZs in 23 geographic regions. This helps Chinese businesses build and expand their presence outside China, and organizations of other countries embrace the Chinese market. Huawei Cloud also serves the Asia Pacific with local service teams in more than 10 countries and regions. In Latin America, Huawei Cloud is the cloud service provider with the most data centers.

- A massive number of services

  More than 200 cloud services have gone live in Huawei Cloud Marketplace, along with over 2,800 apps and 6,000+ partners. These cloud services cover almost all essential areas, including computing, storage, networking, software development, database, and security.

- Innovative chips

  Public clouds are powered by cloud data centers, which require intensive hardware investment. Huawei has invested in R&D from the beginning, and now uses the expertise gained to support both hardware and software for its cloud services. Huawei Cloud data centers differentiate from the competition with cutting-edge chips, the cornerstone of the IT industry. Any breakthrough made in chips would be impossible without long-term R&D investment, and Huawei has invested two decades into next-generation cloud data center chips for computing, networking,

storage, and security. This chip portfolio turbocharges the cloud services of Huawei in the Cloud 2.0 era.

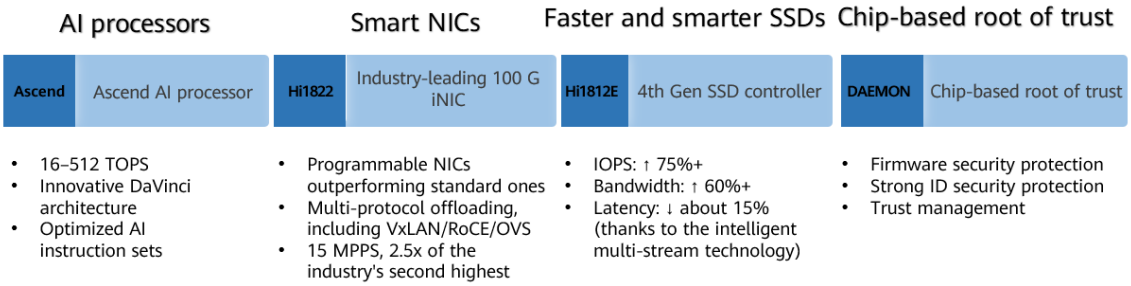


Figure 2-3 Overview of Huawei chips

- Intelligent infrastructure

  Huawei Cloud's user base is so large that even bigger cloud infrastructure is insufficient. It needs an ultra-large automated cloud infrastructure capable of global coordination and self-service. Such infrastructure is innovated from the ground up and runs on the brand-new QingTian architecture and Alkaid cloud OS.

  Huawei started on QingTian architecture in 2012, built a software-hardware collaboration system in 2014, and applied QingTian to Huawei Cloud in 2017. QingTian has over 500 patents and a proven track record of serving more than 100,000 nodes. Its emerging technologies include shared storage on bare metal servers (BMSs), 40 Gbit/s BMSs, and forwarding of tens of millions of packets per second.

  QingTian features software-hardware synergy on the data plane and an intelligent cloud brain on the management plane. The data plane is fueled by lean data centers and virtualization, diversified computing, QingTian cards, and ultra-fast engines to support all-resource offloading and acceleration. The intelligent cloud brain is a distributed cloud OS with all-domain resource scheduling for cloud-edge-device collaboration and governance.

  This brain is Huawei Cloud Alkaid, named after the seventh star of the Big Dipper that assisted ancient Chinese people in judging the change of seasons. Of equal critical assistance is the role the cloud OS plays in IT infrastructure. A futureproof cloud OS allocates, deploys, coordinates, and supplies resources of the entire cloud throughout its lifecycle. This is why Huawei Cloud defines its next-generation cloud OS as a cloud brain, an entity that can self-learn and self-upgrade.

  What makes Alkaid so special?

  Alkaid is a cloud brain designed for a future world where 5G, cloud, and AI will permeate every part of our lives. It accompanies businesses on their cloud journey in this digital age.

  (1) All-domain resource scheduling

  Distributed technologies are nothing new to cloud vendors, but not all of them excel when scheduling distributed cloud resources. Alkaid is powered by cloud-edge collaboration of over 180 cloud services and over 40 operators. All-domain resource scheduling contributes to streamlining the cloud, edge, and device for access at 5 ms latencies to Huawei Cloud everywhere. As a result, customers

enjoy 60% lower bandwidth costs and 10 times higher deployment efficiency. This feature makes Huawei Cloud the ideal option when latency is a dealbreaker – autonomous driving, AR/VR, and industrial Internet.

(2) Dynamic negotiation and governance

Some services are deterministic, but their computation involves many variables. IoV applications are a classic example, where latency can come from network transmission, but also device-cloud interaction, when every bit of data passes through wide area networks (such as 5G), edge sites, and data centers. This data is then processed by applications, optimized kernels, and more than 1,000 microservices such as hardware and software acceleration engines and high-speed storage. Despite being this long and complex, such processes require accuracy and stability but not at the cost of latency. To do this, dynamic negotiation and governance of the cloud OS is key for autonomous driving, real-time risk management, and industrial control. Alkaid coordinates more than 1,000 microservices to virtually eliminate service jitter, slash latency to milliseconds, and guarantee deterministic QoS.

(3) Multi-objective optimization

The vision of a smarter world materializes with new tech. Intelligent cloud brains like Alkaid use resource profiling and predictive algorithms to provide customers with more accurate recommendations, such as billing policies, computing specifications, and scaling suggestions. Huawei Cloud also uses internal platform data and self-learning algorithms for self-tuning and better customer service. The in-house A-DNN algorithm and the industry-leading Atlas 900 AI computing cluster power Alkaid in finding approximate solutions to optimization problems faster.

(4) Diversified computing power

Huawei Cloud has the industry's most abundant computing resources, including Kunpeng, Ascend, x86, GPU, and FPGA, to give customers the flexible usage and deployment of computing power they need. It is estimated that featured mobile apps running on Kunpeng reduce costs by 5 times, boost concurrent and graphics rendering by 10 times, and drive encryption and decryption performance by 20 times. Alkaid tailors the business model to each commercial phase and holding period for 60% lower costs.

(5) Trustworthy full-stack technologies

Huawei Cloud is unique in the industry for its independent full-stack products and technologies. This means that it offers underlying chips (Kunpeng and Ascend), OSs, middleware, and databases and integrated devices (servers and storage devices). It also provides O&M support with over 80,000 troubleshooting issues.

These offerings are not just full coverage, but also fully reliable. Huawei's reputation is honed from 30 years of successful partnerships with major enterprises, and is showcased in the tech: the Kunpeng-powered stack for Huawei Cloud. This includes Alkaid's ability to predict when and where failures occur more accurately and self-heal, for 70% fewer hardware failures. Huawei

Cloud also provides Disaster Recovery Institute International (DRII) level-6 certification, meaning 0 RPO and data loss at the customer service layer.

In addition, Huawei Cloud has obtained more than 50 compliance certificates worldwide and armed all commercial cloud services with security features to safeguard cloud computing for organizations of all sizes.

2.2 Application Scenarios

Huawei Cloud classifies its solutions in four ways:

- By enterprise type

  This solution portfolio aims to help businesses, either start-up or mature companies, be cloud-native and thrive in the market with cloud-based marketing, management, and business expansion. The cloud services are designed for specific needs: cloud migration, start-up operations, website building, enterprise management, global operations, and logistics management.

- By industry

  This solution portfolio provides services tailored to each industry customer and their respective needs, such as financing, gaming, energy, and healthcare.

- By use case

  This solution portfolio pre-integrates products and capabilities, such as enhanced security, edge-cloud synergy, timely DR and backup, for the cloud migration of traditional ICT services.

- By organization type

  This solution portfolio makes the cloud journey of enterprises, public welfare and non-profit organizations, and HMS ecosystem partners go smoothly.

  For details, visit https://www.huaweicloud.com/intl/en-us/.

  Different enterprises have different understandings of the cloud. This is why Huawei has spent years developing solutions tailored to the cloud migration of different customers.

2.3 Delivery Modes

Huawei Cloud provides three delivery modes: public cloud, HUAWEI CLOUD Stack, and edge cloud.

- Public cloud

  This mode allows you to purchase cloud services on Huawei Cloud and use them over the Internet. You can select the quantity, specifications, and types as needed. Public cloud is a cost-effective and convenient option, suited to both individuals and enterprises.

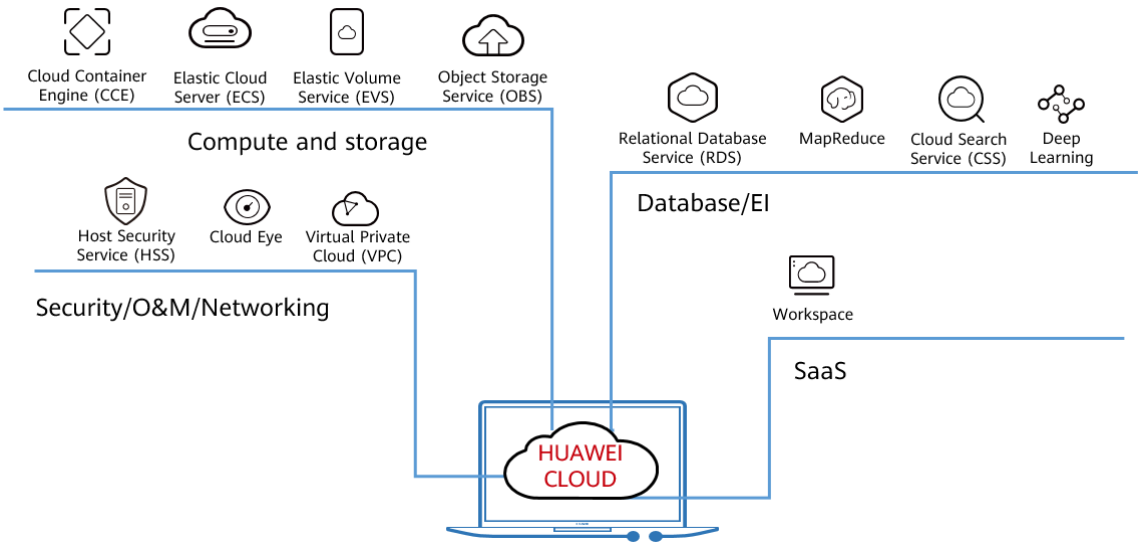


Figure 2-4 Public cloud

- HUAWEI CLOUD Stack

  HUAWEI CLOUD Stack is a cloud infrastructure designed for government organizations and enterprises that is deployed in customers' local data centers. Such hybrid delivery combines the rapid innovation of public clouds with the effective management of private clouds. This mode can also fit the organizational architectures and service procedures of businesses, and is a single cloud from the perspective of users. It is an ideal delivery mode for medium- and large-sized organizations that require on-premises data storage or physical isolation of devices.

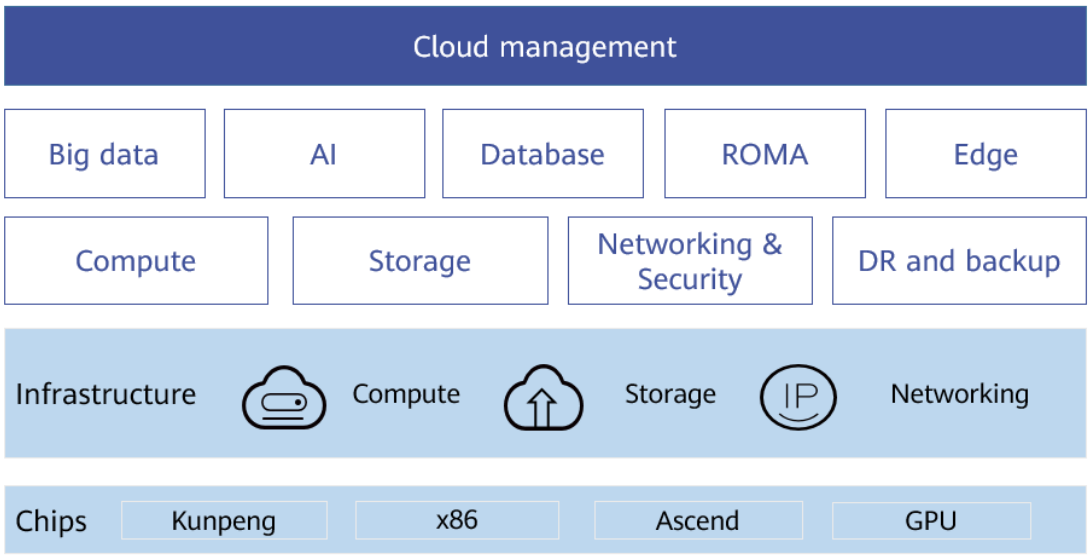


Figure 2-5 Hybrid architecture

- Edge cloud

  Edge clouds are deployed on high-quality nodes of large carriers in major provinces and cities in the Chinese mainland. Users can deploy latency-sensitive services, such as interactive entertainment, online education, and media creation, on the nearest

node for deterministic latency of less than 10 ms and better experience through global smart management and scheduling. An edge cloud node can be built near the user when the cloud data center is overwhelmed by a massive amount of ingested data. This node can pre-process and then send data to the cloud.

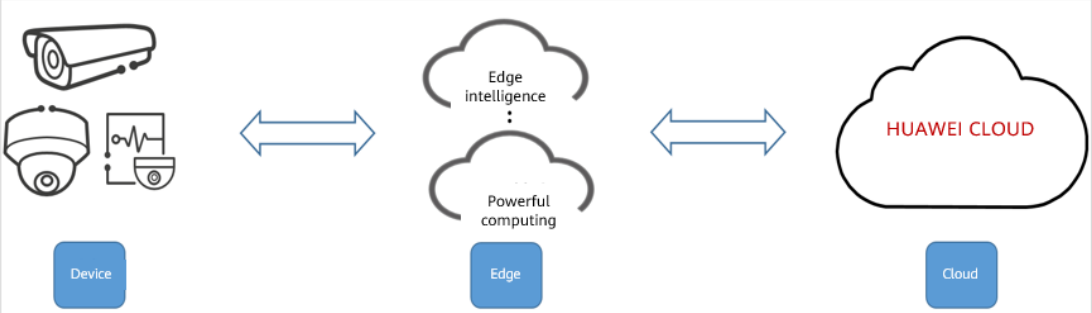


Figure 2-6 Edge cloud

2.4 Technical Support

A user-friendly public cloud system should provide not only easy-to-use services but also timely technical support. Huawei Cloud offers multiple support plans and professional technical services to provide users with a simple and secure cloud journey.

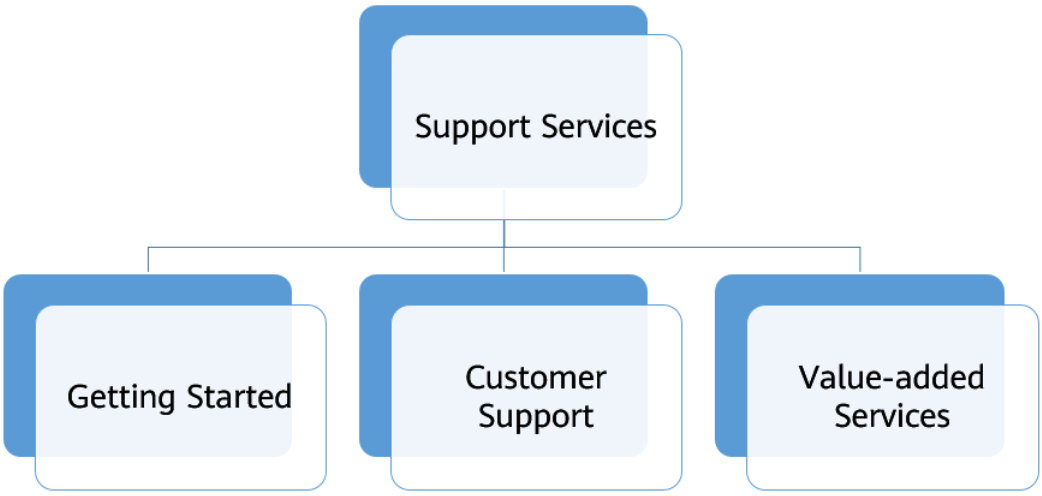


Figure 2-7 Support and service systems

Huawei Cloud's support services can be divided into three service systems:

- Getting Started

  Gives you a full picture of Huawei Cloud services through infographics, introductory courses, best practices in typical scenarios, and expert guides.

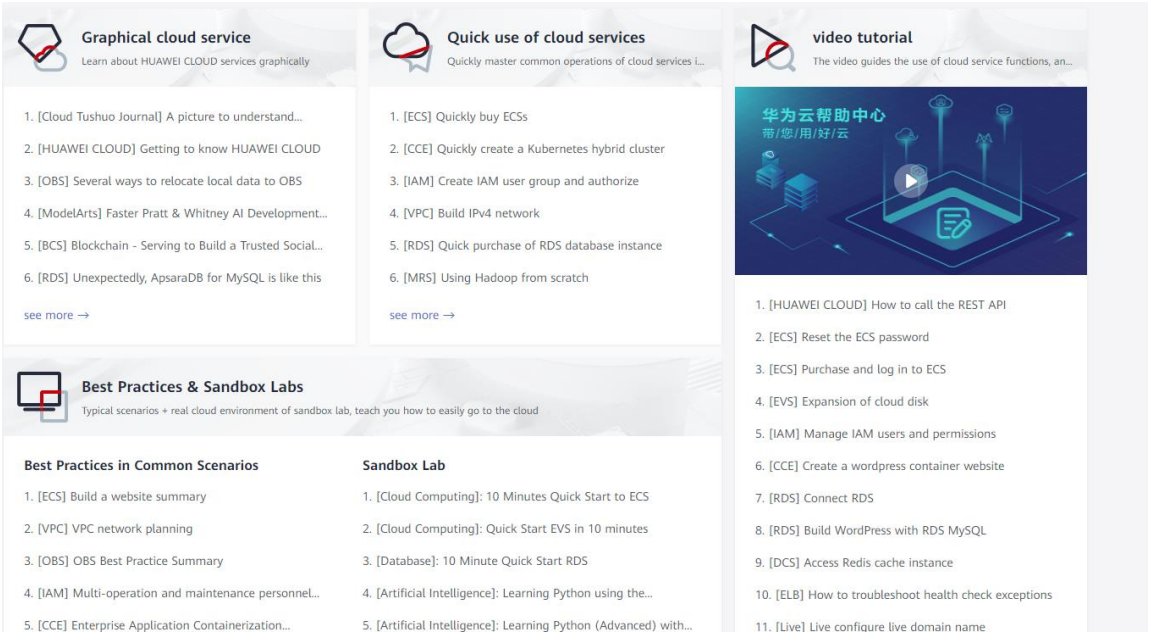


Figure 2-8 Huawei Cloud Support Services (1)

- Customer Support

  Provides timely consulting and technical support to meet customer needs around the clock.

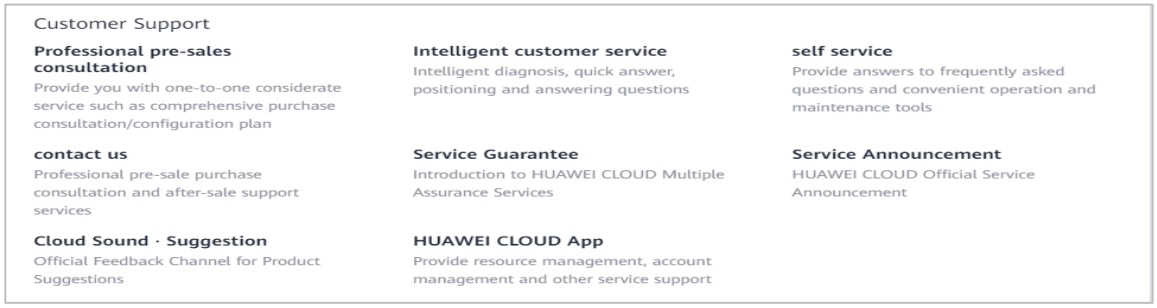


Figure 2-9 Huawei Cloud Support Services (2)

- Value-added Services

  In addition to customer support, Huawei Cloud offers a range of value-added services, including professional scenario-specific support, as well as training and certification on full-stack services.

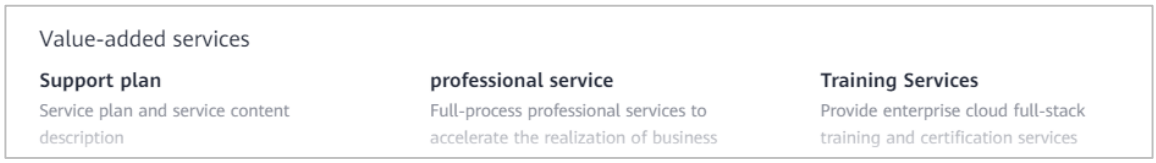


Figure 2-10 Huawei Cloud Support Services (3)

2.5 Huawei Cloud Ecosystem

In a natural ecosystem, each organism is vital and plays an important role. This is also true of the current cloud industry. The resource demands of applications at the IaaS, PaaS, and SaaS layers are so diversified that no cloud service provider alone can provide all these resources. To ensure resource supply, we need to build a collaborative ecosystem that combines the strength of different parties for a refined cloud platform and shared success.

An enterprise will embrace the ecosystem once it adopts the platform-based strategy. So too will Huawei Cloud, which is constantly fostering an ecosystem for shared innovation and success. Huawei Cloud serves as a solid foundation and teams up with partners to catalyze the intelligent transformation of industries. The Huawei Cloud ecosystem has the following features:

- Shared innovation: Huawei Cloud builds the application, data, and AI enablement platforms to help its ecosystem partners migrate applications to the cloud and implement SaaS.
- Shared capabilities: To empower industry applications across cloud, device, and edge, the QingTian architecture of Huawei Cloud streamlines the public, hybrid, and edge clouds for a unified application ecosystem, where innovative capabilities can be shared across industries, scenarios, and deployment modes.
- Shared success: Huawei Cloud and its partners are committed to serving more customers with excellent software for shared success in the digital age.

After years of development, the Huawei Cloud ecosystem has achieved outstanding results, with 1.8 million developers, 10,000+ consulting partners and 7,000+ technology partners already on board. More than 4,000 applications are available in Huawei Cloud Marketplace, attracting over 100,000 paying users who help to generate an annual revenue of over CNY1 billion. We are looking forward to welcoming even more users for greater success.

The partner system of Huawei Cloud consists of consulting partners and technology partners:

- Consulting partners are professional service companies that help customers design, deploy, build, migrate, and manage their workloads and applications. They can be consulting firms, resellers, or service integrators (SIs).
- Technology partners are commercial software companies providing software solutions that are either hosted on or integrated with Huawei Cloud. They can be independent software vendors (ISVs) or providers of SaaS, PaaS, developer tools, management and security services.

We provide comprehensive support to nurture the growth of partners, including training, technical certification, solution release, and business opportunities.

Another important platform of the Huawei Cloud ecosystem is Huawei Cloud Marketplace, where Huawei Cloud joins hands with partners to provide high-quality and easy-to-use software, services, and solutions for cloud computing and big data workloads. Customers seeking fast cloud migration and service roll-out can find the products and services they need on this transaction and delivery platform.

Huawei Cloud Marketplace provides product catalogs by industry and basic capability so that businesses can quickly find the solutions tailored to them.

The sales and service support of Huawei Cloud Marketplace around the world enables service providers to find more product sales sources and improve service experience for higher business profits.

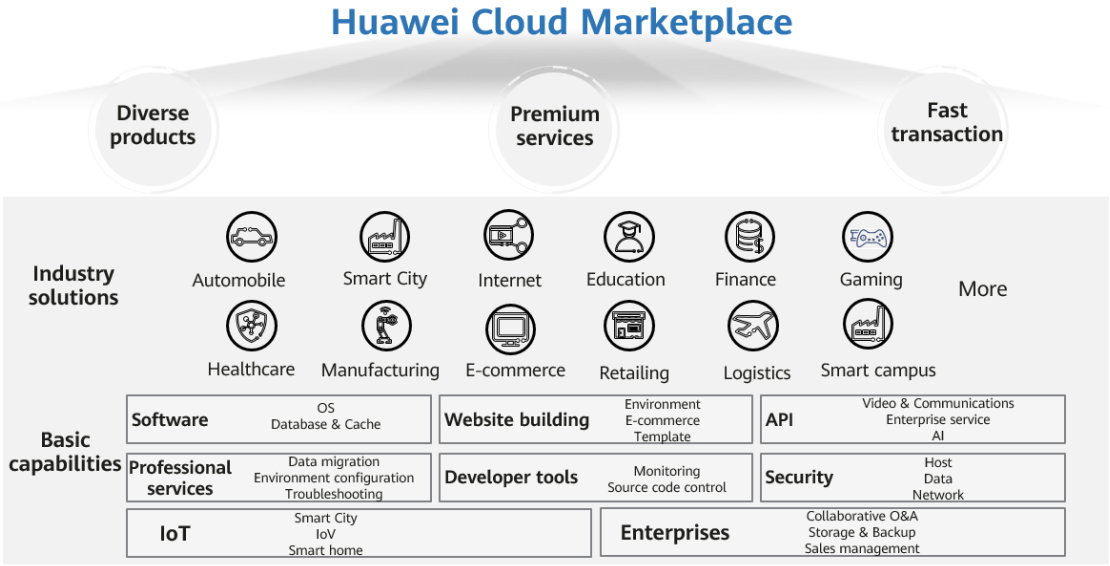


Figure 2-11 Huawei Cloud Marketplace

Increasing the number of developers and partners is vital to the Huawei Cloud ecosystem. To fulfill the commitment of building an industry-leading application development platform, Huawei Cloud provides simplified tools and templates that improve development efficiency. We also leverage industry know-how and asset models to offer application, data, and AI enablement services that can meet the ever-changing market demands.

Huawei Cloud Marketplace is an application distribution platform for government organizations and enterprises. The past two years have seen Huawei Cloud Marketplace grow exponentially, with more than 100,000 orders and an annual transaction volume of over CNY1 billion (including 30 partners with annual sales of over CNY10 million each).

AppGallery is Huawei's device application distribution platform, and is now the world's third largest application market with 530 million active users and a total of 384.4 billion applications distributed. Huawei Cloud works with AppGallery to build a one-stop solution for mobile applications and provide more innovative technologies and resources for the emerging applications integrated with the HMS Core, which offers a rich array of open device and cloud capabilities that facilitate efficient development, fast growth, and flexible monetization.

2.6 Quick Start

This section describes some key concepts to be familiar with when using Huawei Cloud.

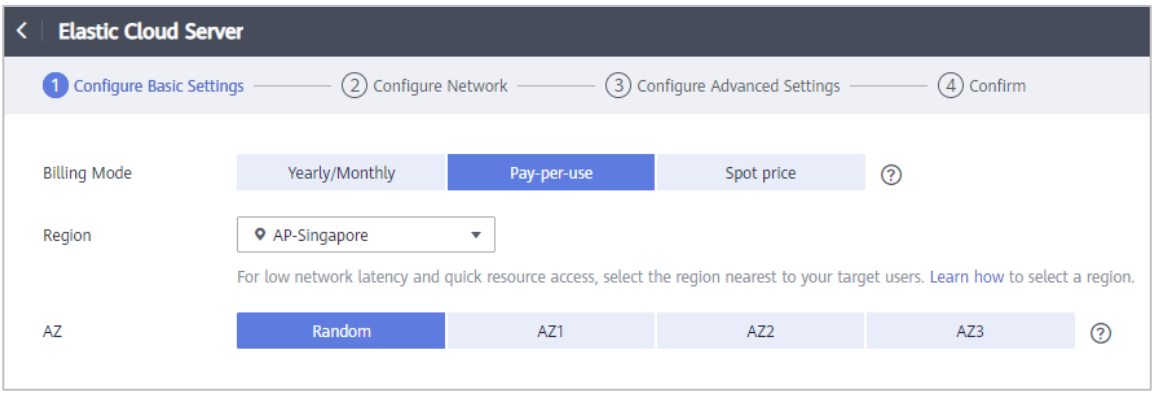


Figure 2-12 Buying an ECS instance

When buying an ECS instance on Huawei Cloud, you need to select the billing mode, region, and AZ. The billing modes of Huawei Cloud are as follows:

- Pay-per-use

  Pay-per-use is a postpaid billing mode in which you are billed based on the ECS instance type and usage duration. You can create or delete an ECS instance at any time.

- Yearly/Monthly

  Yearly/Monthly is a prepaid billing mode in which you specify the ECS instance type and usage duration in advance and are billed accordingly. This mode is more cost-effective than pay-per-use when the usage duration is predictable.

- Spot price

  Spot price is a postpaid billing mode in which your ECS instance is billed based on the usage duration. Because spot price allows you to take advantage of unused ECS capacity, the price is much lower than that of a pay-per-use ECS instance with the same specifications.

  Now let's move on to the concepts of region and AZ.

- Region

  Regions are classified by geographic location and network latency. Generally, a region is a city in which the public cloud data center is built. Public services such as Elastic Cloud Server (ECS), Elastic Volume Service (EVS), Virtual Private Cloud (VPC), Elastic IP (EIP), and Image Management Service (IMS) are shared within a region. A region can be either universal or dedicated. A universal region provides universal cloud services for public users, while a dedicated region provides only one service type or serves specific users.

- Availability Zone (AZ)

  An AZ is one or more physical data centers within a region. It has independent cooling, fire extinguishing, moisture-proof, and electricity facilities. Within an AZ, compute, networking, storage, and other resources are logically divided into multiple clusters. AZs within a region are connected using high-speed optical fibers for high availability.

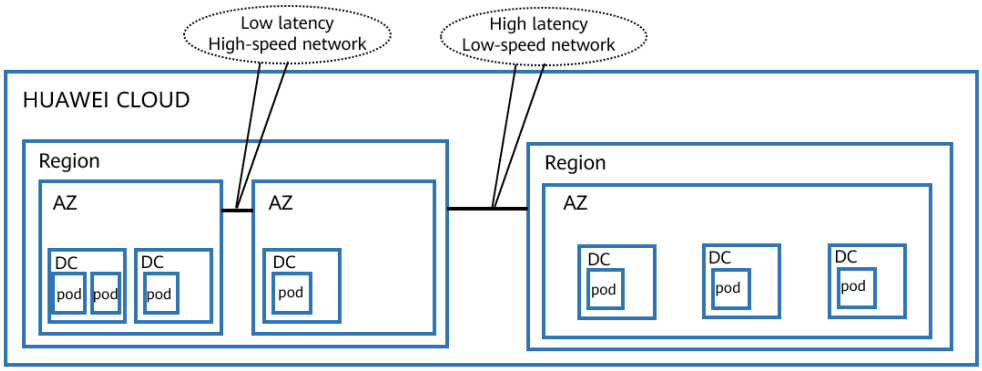


Figure 2-13 Region and AZ

Huawei Cloud services span countries and regions worldwide. You can select a region and AZ as required.

- Selecting a region

  You are advised to select the nearest region for lower network latency and faster access.

- Selecting an AZ

  AZs in a region can be regarded as equipment rooms in a city. The number of AZs needed depends on the DR and network latency requirements. You can deploy resources in different AZs of a region for timely DR, or in one AZ when the application is not latency-sensitive.

  Another important service is Identity and Access Management (IAM). This service allows you to grant users permissions to access resources in your Huawei Cloud account. You can use your own Huawei Cloud account to create IAM users and assign permissions to them.

  Enterprise Management is another service for managing users and access permissions. However, Enterprise Management also allows accounting and application management, and supports more fine-grained authorization for resource usage. In IAM, you can create IAM projects, while in Enterprise Management you can create enterprise projects.

- IAM projects

  An IAM project groups and physically isolates compute, storage, and networking resources in the same region. All your resources are managed by project, which can be a department or a project group. Resources cannot be transferred between IAM projects, but can only be deleted and then created or purchased again. You can create multiple projects using one account.

- Enterprise projects

  An enterprise project groups and manages resources across regions. An enterprise project can contain resources of the same department or project group in multiple regions. Resources can be transferred between enterprise projects.

  If you have enabled Enterprise Management, you cannot create IAM projects and can only manage existing projects. Due to their higher flexibility, enterprise projects will be used preferentially in the future.

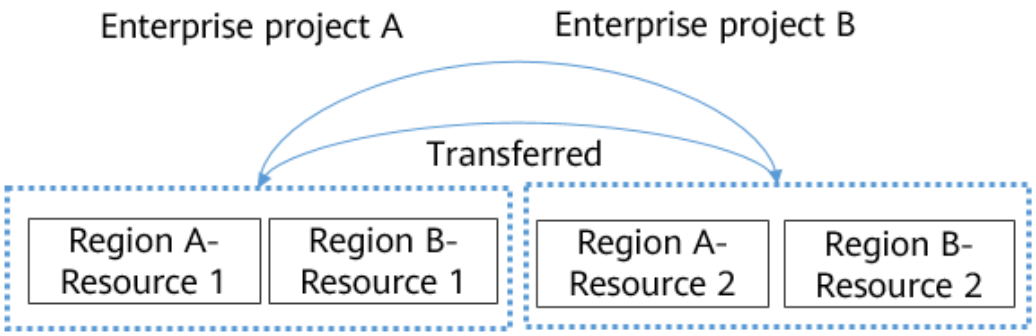


Figure 2-14 Enterprise project

Now that we have given an overview of Huawei Cloud, we will dive into the major Huawei Cloud services in the following chapters to understand their respective positioning, advantages, and usage.

3 Compute Cloud Services

Compute resources are essential to the development of enterprise service systems. For cloud computing, compute services are the most important cloud services.

In this section, we will learn about the compute services offered by Huawei Cloud, including Elastic Cloud Server (ECS), Bare Metal Server (BMS), Image Management Service (IMS), Auto Scaling (AS), Cloud Container Engine (CCE), GPU Accelerated Cloud Server (GACS), Cloud Phone (CPH), and Dedicated Host (DeH). The first five of these cloud services will be introduced in detail.

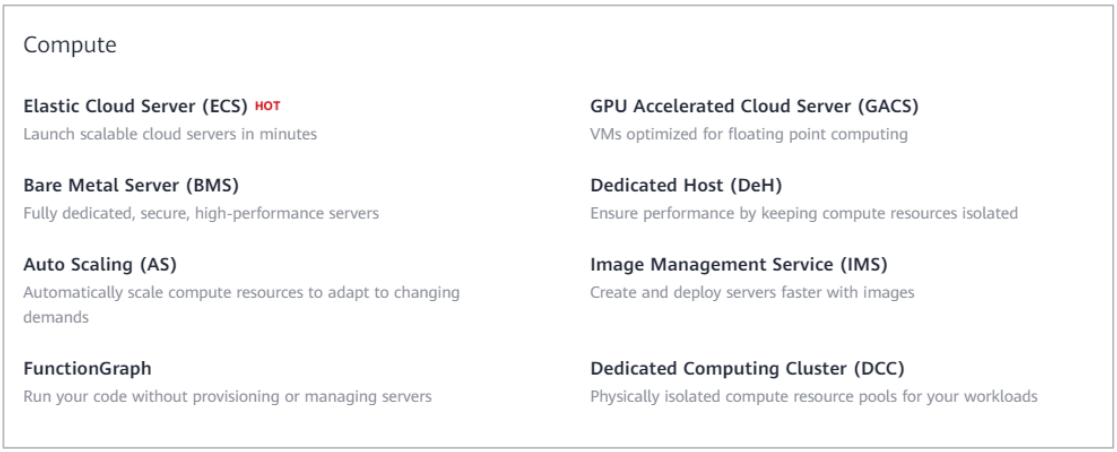


Figure 3-1 Compute cloud service overview

3.1 ECS

3.1.1 What Is ECS?

An Elastic Cloud Server (ECS) is a basic computing unit that consists of vCPUs, memory, OS, and EVS disks. After creating an ECS, you can use it on the cloud in the same way you would use your local PC or physical server.

3.1.2 Architecture

ECS is a cloud service product that interworks with a variety of other cloud service products. Let's take a look at the architecture of an ECS.

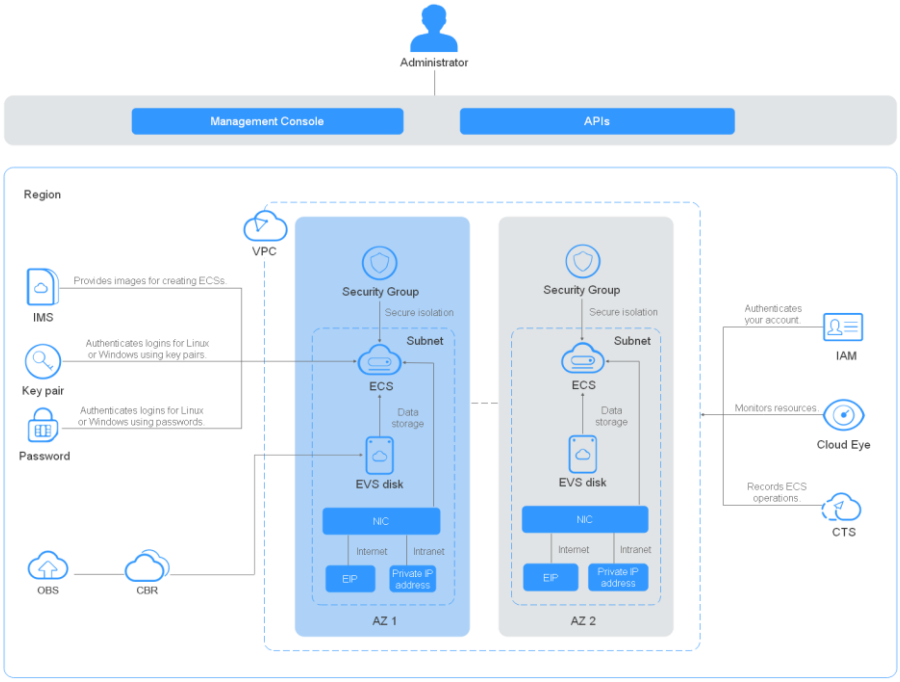


Figure 3-2 ECS architecture

- ECSs can be deployed in different availability zones (AZs). If ECSs in some AZs become faulty, ECSs in other AZs in the same region can still work properly.
- With Virtual Private Cloud (VPC), you can configure subnets and security groups to establish a dedicated network environment for your ECSs. You can also bind elastic IP addresses (EIPs) (with bandwidth assigned) to your ECSs to communicate with the Internet.
- With Image Management Service (IMS), you can install OSs on ECSs or use private images to batch create ECSs.
- Elastic Volume Service (EVS) provides storage space and Volume Backup Service (VBS) provides data backup and restoration for ECSs.
- Cloud Eye monitors the resource usage of ECSs.
- Cloud Backup and Recovery (CBR) backs up data for EVS disks and ECSs, and uses snapshot backups to restore EVS disks and ECSs when necessary.

3.1.3 Advantages

ECS has the following advantages.

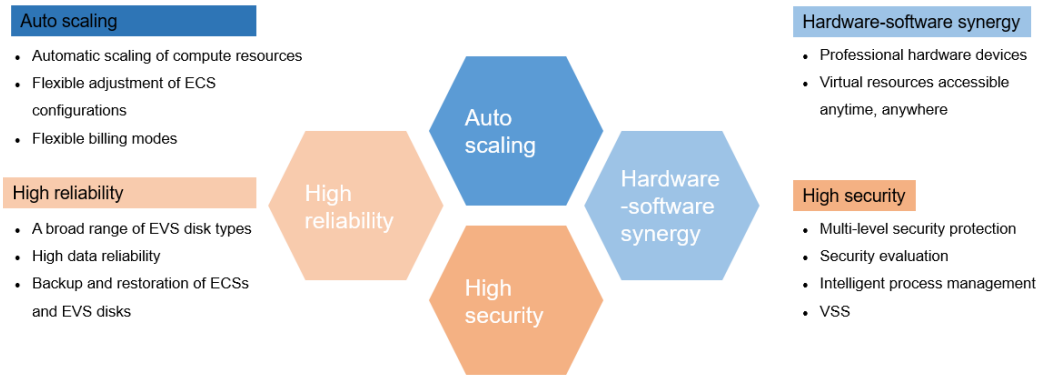


Figure 3-3 ECS advantages

- High reliability

  Huawei Cloud provides a broad range of EVS disk types for you to choose from. With the distributed architecture, EVS delivers high I/O throughput and ensures that data can be rapidly migrated and restored if any data replica is unavailable, preventing data loss caused by a single hardware fault. In addition, Huawei Cloud supports backup and restoration of ECSs and EVS disks. You can set automatic backup policies to back up in-service ECSs and EVS disks. You can also configure policies on the management console or use an API to back up the data of ECSs and EVS disks at a specified time.

- High security

  Huawei Cloud provides multiple security services for ECSs, such as Web Application Firewall (WAF) and Vulnerability Scan Service (VSS). Huawei Cloud also evaluates the cloud environment security to help you quickly identify security vulnerabilities and threats, and provides security configuration check and recommendations to reduce or eliminate losses from network viruses or attacks. Intelligent process management automatically prohibits the execution of unauthorized programs based on a customized whitelist, thereby ensuring ECS security.

- Hardware-software synergy

  ECSs are deployed on professional hardware devices that allow in-depth virtualization optimization, delivering best-in-class virtual server performance. You can obtain scalable, dedicated resources from the virtualized resource pool anytime, anywhere, providing a reliable, secure, flexible, and efficient environment for your applications. You can use your ECS the way you would use your local PC or physical server.

- Auto scaling

  You can configure Auto Scaling (AS) policies based on service traffic to adjust the number of ECSs in an AS group to ensure that the service is running properly. You can also configure periodic or scheduled AS policies based on your business expectations and operating plans.

3.1.4 How to Buy an ECS

Let's see how to purchase an ECS.

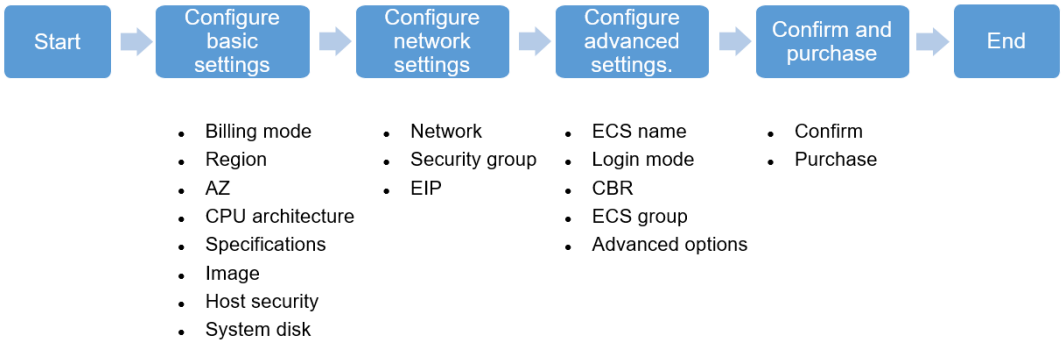


Figure 3-4 ECS purchase process

Below are the steps required to purchase an ECS:

- Preparations

  Register for a Huawei Cloud account and top up that account. Fees will be deducted when you purchase an ECS.

- Basic settings

  Select a billing mode, region, and AZ.

  - **Billing Mode**: An ECS can be billed on a pay-per-use, yearly/monthly, or spot price basis. For yearly/monthly subscriptions, the longer you subscribe, the more you save.
  - **Region** and **AZ**: ECSs in different regions cannot communicate with each other over an intranet. For low network latency and quick resource access, select the region nearest to your target users.
  - **CPU Architecture**: x86 uses Complex Instruction Set Computing (CISC). Kunpeng uses Reduced Instruction Set Computing (RISC).
  - **Specifications** and **Image**: Select specifications and an image for ECS based on service requirements.

  In brief, you can determine the operating system, specifications, and EVS disk type of your ECS.

- Network settings

  Configure a subnet, security group, and extension NIC.

  - **Subnet**: A subnet is a network used to manage ECSs. It provides IP address management and DNS resolutions for ECSs in it. The IP addresses of all ECSs in a subnet belong to the subnet. You can use DHCP to randomly assign an IP address in the subnet or manually specify an IP address in the subnet to the ECS.
  - **Security Group**: A security group is a collection of access control rules for ECSs that have the same security requirements and that are mutually trusted. It enhances security for ECSs. Select a suitable security group for the ECS NIC.
  - (Optional) **Extension NIC**: You can add additional NICs to the ECS.

- Advanced settings

  In previous settings, you have configured compute, storage, and network resources for the ECS. Next, configure some advanced settings, including the ECS name, login mode, and cloud backup and recovery (CBR).

  - **ECS Name**: You can customize ECS names in compliance with naming rules. If you intend to purchase multiple ECSs at a time, the system automatically adds a hyphen followed by a four-digit incremental number to the end of each ECS.
  - **Login Mode**: **Key pair** allows you to use a key pair for login authentication. **Password** allows you to use a username and its initial password for login authentication. For Linux ECSs, the initial password is the root password. For Windows ECSs, the initial password is the Administrator password.
  - **Cloud Backup and Recovery**: With CBR, you can back up data for ECSs and EVS disks, and use backups to restore ECSs and EVS disks when necessary.
  - **ECS Group (Optional)**: An ECS group allows ECSs within the group to be automatically allocated to different hosts. To improve service reliability, select an ECS group.
  - **Advanced Options**: You can configure other advanced and optional settings.

  After you complete the configurations, your selections are displayed for you to check and confirm. Read and select the check box of the agreement and disclaimer, and click **Submit**.

3.1.5 How to Access an ECS

After purchasing an ECS, you can log in to the ECS to deploy applications on it. There are several ways to log in to an ECS.

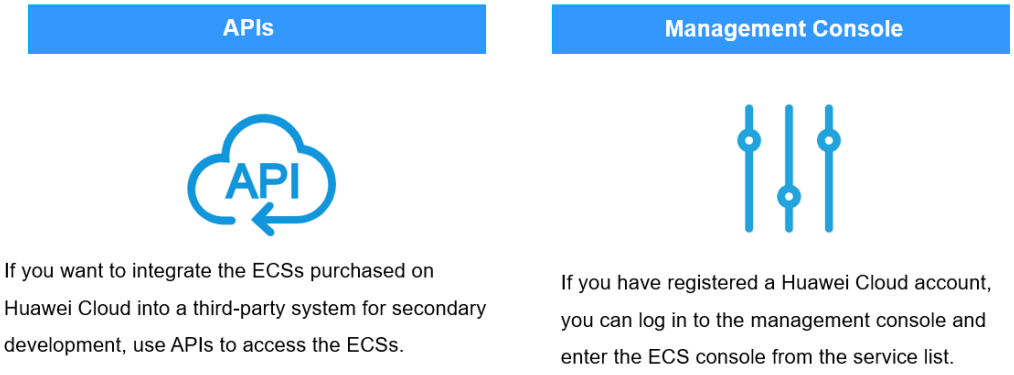


Figure 3-5 ECS access

- APIs

  If you want to integrate the ECSs purchased on Huawei Cloud into a third-party system for secondary development, you can use APIs to access the ECSs.

- Management console

If you have registered a Huawei Cloud account, you can log in to the Huawei Cloud official website, enter the ECS console from the service list of the management console, and remotely log in to ECSs.

- SDKs

  You can also log in to ECSs using SDKs.

  The method for logging in to an ECS depends on the operating system. For example, the methods for logging in to a Windows ECS and a Linux ECS are different.

3.1.5.1 Logging In to a Windows ECS

Before logging in to a Windows ECS, note the following constraints:

- You can only log in to a running ECS.
- The username for logging in to a Windows ECS is **Administrator**.
- To log in to a key-pair-authenticated ECS, use the password obtaining function provided by the management console to decrypt the key file used during ECS creation.

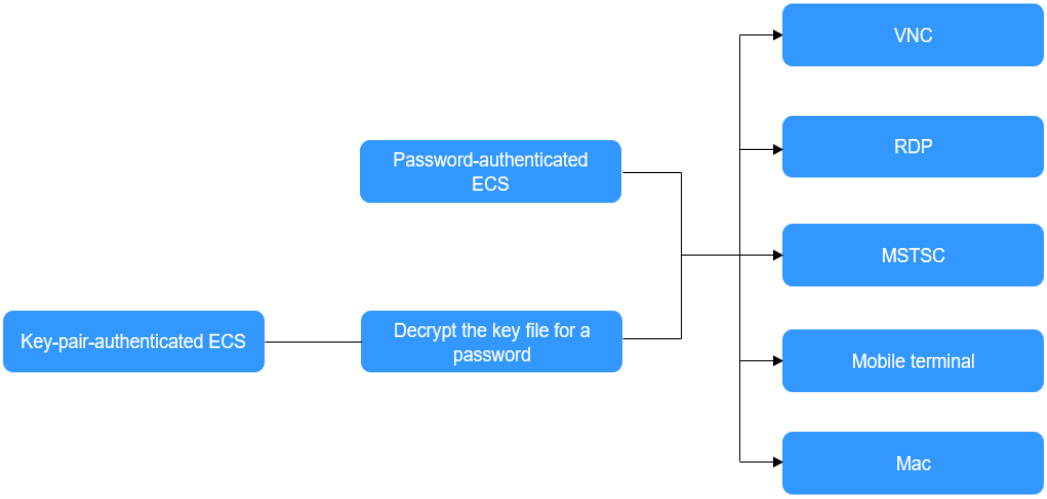


Figure 3-6 Logging in to a Windows ECS

As shown in Figure 3-6, you can log in to a Windows ECS in any of the following ways:

- Using VNC on the management console: The login username is **Administrator**. In this login mode, you do not need to bind an EIP to the ECS.
- Using the RDP file provided on the management console: The login username is **Administrator**. The ECS must have an EIP bound and port 3389 must be opened. This is the recommended method for logging in to Windows ECSs.
- Using MSTSC: The login username is **Administrator**. The ECS must have an EIP bound and port 3389 must be opened.
- From a mobile terminal: Install a remote connection tool, such as Microsoft Remote Desktop, on your mobile terminal before logging in to the ECS. The login username is **Administrator**. The ECS must have an EIP bound and port 3389 must be opened.

- From a Mac: Install a remote connection tool, such as Microsoft Remote Desktop for Mac before logging in to the ECS. The login username is **Administrator**, the ECS must have an EIP bound, and port 3389 must be opened.

3.1.5.2 Logging In to a Linux ECS

Logging in to a Linux ECS has the following constraints:

- You can only log in to a running ECS.
- The username for logging in to a Linux ECS is **root**.

You can log in to a password-authenticated ECS in any of the following ways:

- Using VNC on the management console: The login username is **root**. In this login mode, you do not need to bind an EIP to the ECS.
- Using an SSH password: The login username is **root**. The ECS must have an EIP bound and port 22 must be opened.
- From a mobile terminal: Use an SSH client, such as Termius or JuiceSSH to log in. The login username is **root**, the ECS must have an EIP bound, and port 22 must be opened.

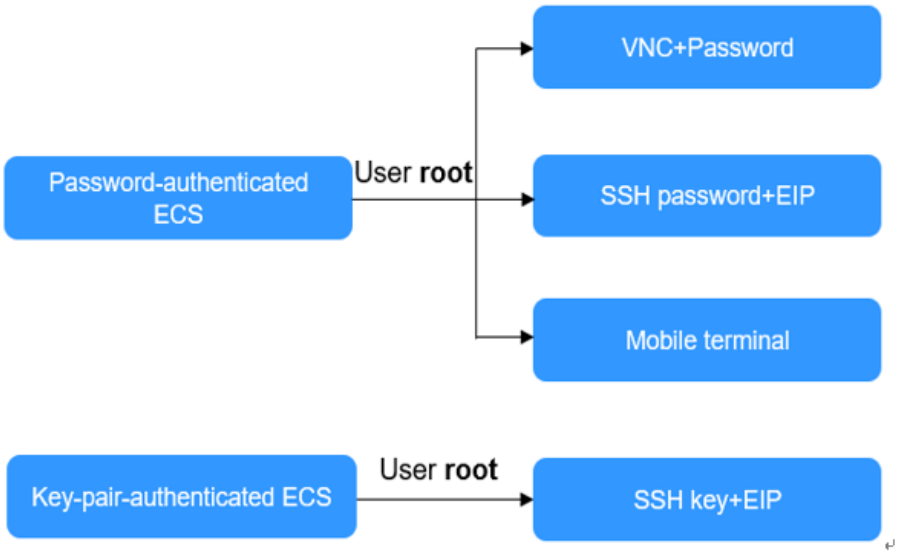


Figure 3-7 Logging in to a Linux ECS

Next, let's see how to use a Windows ECS or Linux ECS after logging in.

3.1.6 How to Use an ECS

3.1.6.1 Reinstalling or Changing an ECS OS

If the OS of an ECS fails to start or requires optimization, reinstall or change the OS.

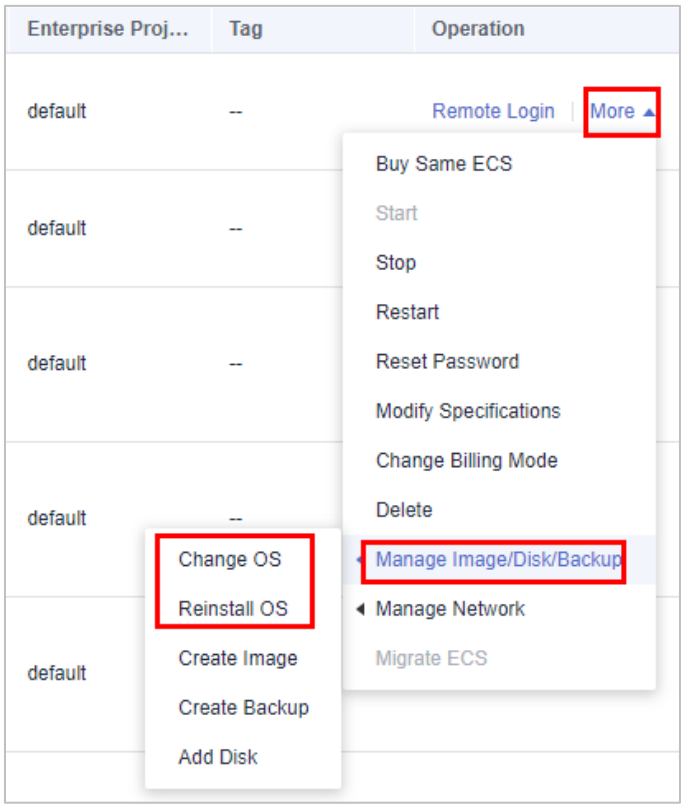


Figure 3-8 Reinstalling or changing an ECS OS

As shown in Figure 3-8, you can easily reinstall or change the OS of an ECS. Note the following when reinstalling the OS:

- After the OS is reinstalled, the IP and MAC addresses of the ECS remain unchanged.
- Reinstalling the OS clears the data in all partitions of the EVS system disk, including the system partition. Therefore, back up data before reinstalling the OS.
- Reinstalling the OS does not affect data disks.
- Do not perform any operations on the ECS immediately after its OS is reinstalled. Wait for several minutes until the system successfully injects the password or key. Otherwise, the injection may fail, and the ECS cannot be logged in to.
- For a Windows ECS, if you set a new password during OS reinstallation, the locally stored RDP file will become invalid, and you will need to download an RDP file again to log in to the ECS.

3.1.6.2 Modifying Specifications

If the specifications of an existing ECS cannot meet service requirements, you can modify the ECS specifications as needed, such as increasing the number of vCPUs and scaling up memory.

Note the following when modifying ECS specifications:

- Sold-out vCPUs and memory cannot be selected.
- ECS specifications (vCPU or memory) reduction degrades ECS performance.

- Certain ECS types do not support specifications modification.
- When the disk status is **Expanding**, the specifications of the ECS where the disk is attached cannot be modified.
- Before modifying the specifications of a Windows ECS, modify the SAN policy to prevent disks from going offline after the specifications are modified.
- If you want to modify the specifications of an ECS billed on a yearly/monthly basis, select the target specification, pay for the shortage or get a refund, and restart the ECS.
- If you want to modify the specifications of a pay-per-use ECS, there is no need to make an upfront payment.

3.1.6.3 Resetting a Password

An ECS password is important for logins. Keep your password secure. If you forget the password or the password expires, you can reset the password. The prerequisites for resetting the password of an ECS are as follows:

- You have installed password reset plug-ins before the password expires or is forgotten.
  - For ECSs that are created using public images, the password reset plug-ins are installed by default.
  - If the ECS is created using a private image and does not have the password reset plug-ins installed, see the solutions provided on the Huawei Cloud website.
- Do not delete the CloudResetPwdAgent and CloudResetPwdUpdateAgent process. Otherwise, one-click password reset will become unavailable.
- ECSs created using SUSE 11 SP4 must have 4 GB or larger memory.
- DHCP must be enabled in the VPC to which the ECS belongs.
- The ECS network connectivity is normal.

You can reset the password of an ECS only when the ECS meets the preceding prerequisites.

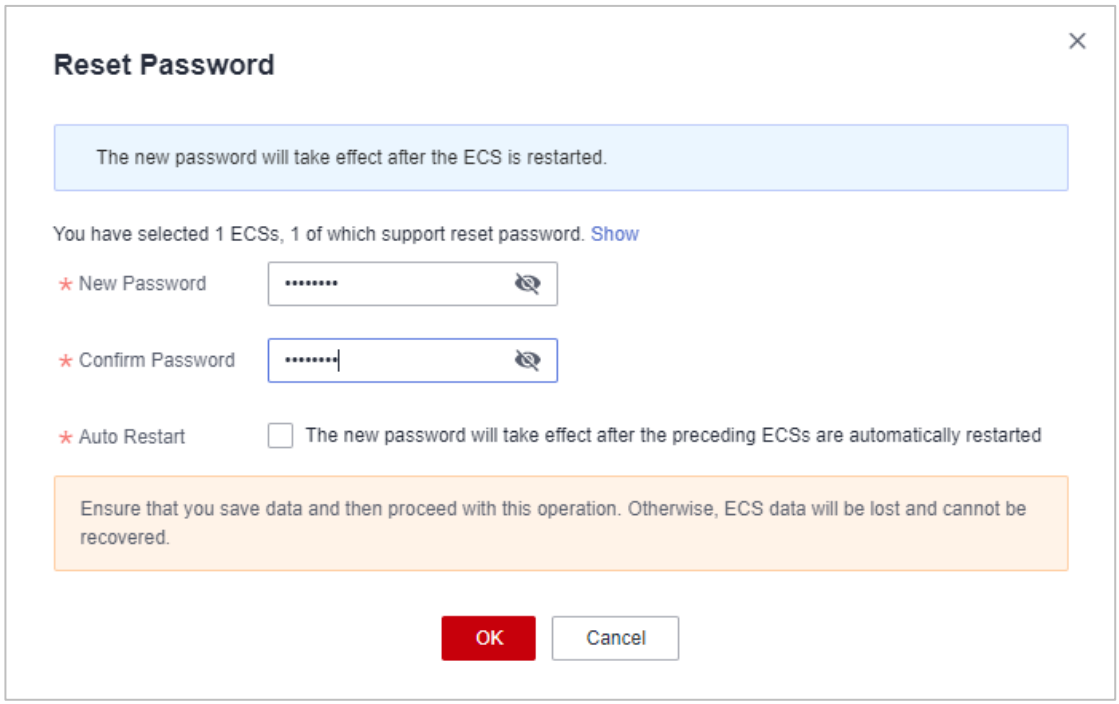


Figure 3-9 Resetting a password

3.1.7 Application Scenarios

ECSs are widely used to replace traditional hardware servers. Next, let's take a look at some typical application scenarios.

- Web applications

  If an enterprise wants to deploy website development and test environment, small-scale database applications, or web servers, it is recommended that general computing ECSs work with Elastic Load Balance (ELB) and Virtual Private Cloud (VPC). General computing ECSs provide a balance of compute, memory, and network resources and a baseline level of vCPU performance with the ability to burst above the baseline. This type of ECS is suitable for general workloads, such as web servers, enterprise R&D, and small-scale databases.

- Enterprise e-commerce

  E-commerce enterprises have the following pain points:

  - Sudden traffic surges: The traffic can surge to hundreds of times its normal levels during promotions, flash sales, and sweepstakes. Servers become overloaded and e-commerce platforms may even crash.
  - Poor user experience: Massive amounts of static data, such as product pictures and videos, is usually stored on servers, which is costly, time-consuming, and loads slowly. Users in different network environments may experience delayed access to such data, resulting in poor user experience.
  - Lack of data support for business decision-making: Due to a lack of big data platforms and analysis tools, existing users, commodities, and transaction data cannot be effectively analyzed. As a result, investments in promoting e-commerce websites are high but repeat purchase rates remain low.

- Security not guaranteed: E-commerce enterprises have to deal with risks in various processes, such as traffic diversion, registration and login, browsing and comparison, coupon obtaining, ordering, payment, delivery, and evaluation. These risks may come from credential stuffing, promotion exploitation, scalpers, web page tampering, DDoS attacks, data breaches, and Trojans.

To sum up, e-commerce enterprises feature large data volumes and data process requests, and require large memory and fast data exchange and processing, such as precision marketing, e-commerce, and mobile apps. Memory-optimized ECSs can provide a large amount of memory, ultra-high I/O EVS disks, and bandwidths and are suitable for the e-commerce industry.

- Graphics rendering

  Graphics rendering and engineering drawing, for example, requires high graphic and video quality, large memory size, massive data processing, high I/O concurrency, fast data exchange and processing, and high GPU performance. GPU-accelerated G1 ECSs are good choices for graphics rendering because they are developed based on NVIDIA Tesla M60 hardware virtualization, support DirectX and OpenGL, and provide up to 1 GiB of GPU memory and 4096 x 2160 resolution.

- Data analysis

  Data analysis requires high I/O capabilities and fast data exchange and processing to process a large amount of data. Example scenarios include Hadoop distributed computing, large-scale parallel data processing, and log processing. Disk-intensive ECSs are recommended for data analysis because they are delivered with local disks for high storage bandwidth and IOPS. In addition, local disks are more cost-effective in massive data storage scenarios. Disk-intensive ECSs have the following characteristics: use local disks to provide higher sequential read/write performance and lower latency, thereby improving file read/write performance; provide powerful and stable computing capabilities to ensure efficient processing of computing jobs; provide higher intranet performance, including high intranet bandwidth and Packet per Second (PPS), to meet the requirements for data interaction between ECSs during peak hours.

- High-performance computing

  High-performance computing scenarios cover a host of industries such as scientific computing, genetic engineering, games and animation, and biopharmaceuticals. Each vCPU of a high-performance computing ECS corresponds to the hyper-thread of an Intel® Xeon® Scalable processor core. High-performance computing ECSs provide massive parallel computing resources and high-performance infrastructure services to fulfill the requirements of high-performance computing and massive storage and to ensure efficient rendering.

3.2 BMS

3.2.1 What Is BMS?

Bare Metal Server (BMS) combines the scalability of VMs with the high performance of physical servers. It provides dedicated servers on the cloud, delivering the performance and security required by core databases, critical applications, high-performance computing (HPC), and big data.

The difference is that BMSs can be easily configured and purchased on the cloud platform, but traditional physical servers can only be configured and purchased in person.

BMSs support automatic provisioning, automatic O&M, VPC connection, and interconnection with shared storage. You can provision and use BMSs as easily as ECSs and enjoy excellent computing, storage, and network performance of physical servers.

3.2.2 Architecture

BMS works together with other cloud services to provide compute, storage, network, and imaging.

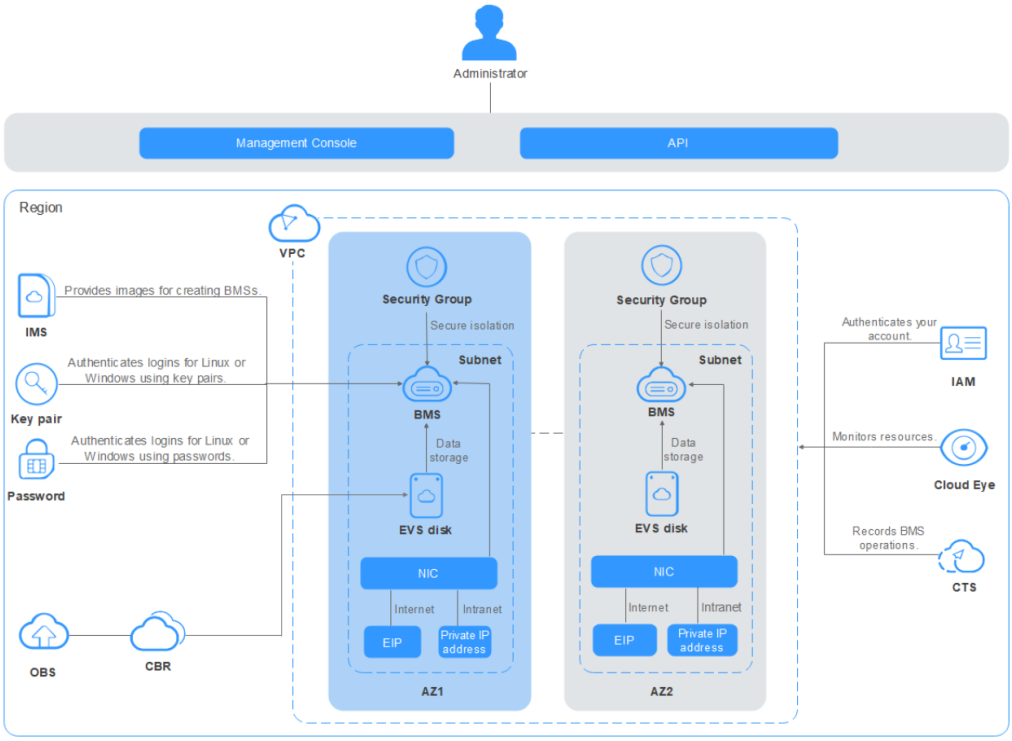


Figure 3-10 BMS architecture

- BMSs are deployed in multiple availability zones (AZs) connected with each other through an internal network. If an AZ becomes faulty, other AZs in the same region will not be affected.
- With the Virtual Private Cloud (VPC) service, you can build a dedicated network for BMS, configure subnets and security groups, and allow resources deployed in the VPC to communicate with the Internet through an EIP (with bandwidth assigned).

- With the Image Management Service (IMS), you can install OSs on BMSs or create BMSs using private images for rapid service deployment.
- The Elastic Volume Service (EVS) provides storage, and Volume Backup Service (VBS) provides data backup and restoration.
- Cloud Eye is a key tool to monitor BMS performance, reliability, and availability. Using Cloud Eye, you can monitor BMS resource usage in real time.
- Cloud Backup and Recovery (CBR) backs up data for EVS disks and BMSs, and uses snapshot backups to restore the EVS disks and BMSs when necessary.

3.2.3 Advantages

BMSs feature high performance just like traditional physical servers do. In addition, BMSs have some other advantages.

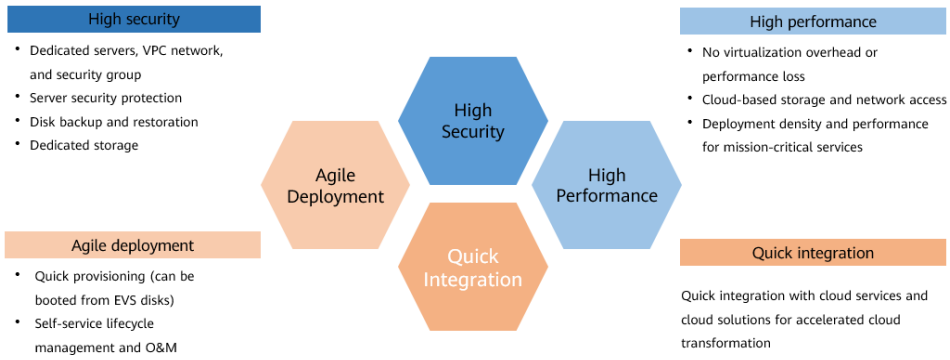


Figure 3-11 BMS advantages

- High security

  BMS allows you to use dedicated compute resources, add servers to VPCs and security groups for network isolation, and integrate related components for server security. BMSs running on the QingTian architecture can use EVS disks, which can be backed up for restoration. BMS interconnects with Dedicated Storage Service (DSS) to ensure the data security and reliability required by enterprise services.

- High performance

  BMS has no virtualization overhead, allowing compute resources to be fully dedicated to running services. Running on QingTian, an architecture from Huawei that is designed with hardware-software synergy in mind, BMS supports high-bandwidth, low-latency storage and networks on the cloud, meeting the deployment density and performance requirements of mission-critical services such as enterprise databases, big data, containers, HPC, and AI.

- Agile deployment

  The hardware-based acceleration provided by the QingTian architecture enables EVS disks to be used as system disks. The required BMSs can be provisioned within minutes when you submit your order. This greatly improves the deployment efficiency. You can manage your BMSs throughout their lifecycle from the management console or using open APIs with SDKs.

- Quick integration

  Within a given VPC, cloud services and cloud solutions (such as databases, big data applications, containers, HPC, and AI solutions) can be quickly integrated to run on BMSs, accelerating cloud transformation.

3.2.4 Differences Between BMSs, ECSs, and Physical Servers

A lack of flexibility is the main problem with physical servers. Although cloud computing is super popular right now, some enterprises may still choose physical servers for absolute best possible performance. The only reason is that physical servers do not have any performance loss caused by virtualization overhead.

However, it takes a long time to deploy physical servers, the O&M is complex, and the architecture cannot be reconstructed easily. When physical servers break down, it takes a lot of time, effort, and money to fix them.

When enterprises choose to avoid VMs (ECSs), it is typically because VMs are not able to provide the performance required by their core databases. Additionally, they do not want to adjust their core applications to adapt to VM deployment. These enterprises are faced with a dilemma.

BMS is designed to address this dilemma. It provides physical servers exclusive to a particular enterprise's use, so they do not have to compromise on performance or resource isolation.

Meanwhile, it delivers cloud capabilities such as online delivery, automatic O&M, VPC interconnection, and interconnection with shared storage. You can provision and use BMSs as easily as ECSs and enjoy excellent computing, storage, and network performance of physical servers.

BMS can also offer services that ECSs cannot provide due to various architecture restrictions, such as virtualization services, high-performance computing services, services that have high requirements on I/O performance, and services that have high requirements on core data control and resource isolation. In addition, Huawei Cloud provides O&M for BMSs, which helps keep your costs down.

Item	BMS	ECS	Physical Server
Physical resources	Exclusive	Shared	Exclusive
Application scenarios	Mission-critical applications or services that require high performance	General-purpose and specific services	Traditional services
Provisioning	Flexible	Flexible	Inflexible
Advanced features	Automatic provisioning, automatic O&M, VPC interconnection, and interconnection with shared storage	Automatic provisioning, automatic O&M, VPC interconnection, and interconnection with shared storage	Traditional features

Figure 3-12 Comparisons between BMSs, ECSs, and physical servers

3.2.5 How to Buy a BMS

The procedure for buying a BMS is as follows.

Figure 3-13 Buying a BMS

You can:

- Create a common BMS.
- Create a BMS that can be quickly provisioned.
- Create a BMS running on Dedicated Cloud (DeC).

If you want to create a BMS that has the same OS and applications as an existing BMS, you can create a private image using the existing BMS and then use the image to create a desired BMS.

Step 1 Configure basic settings.

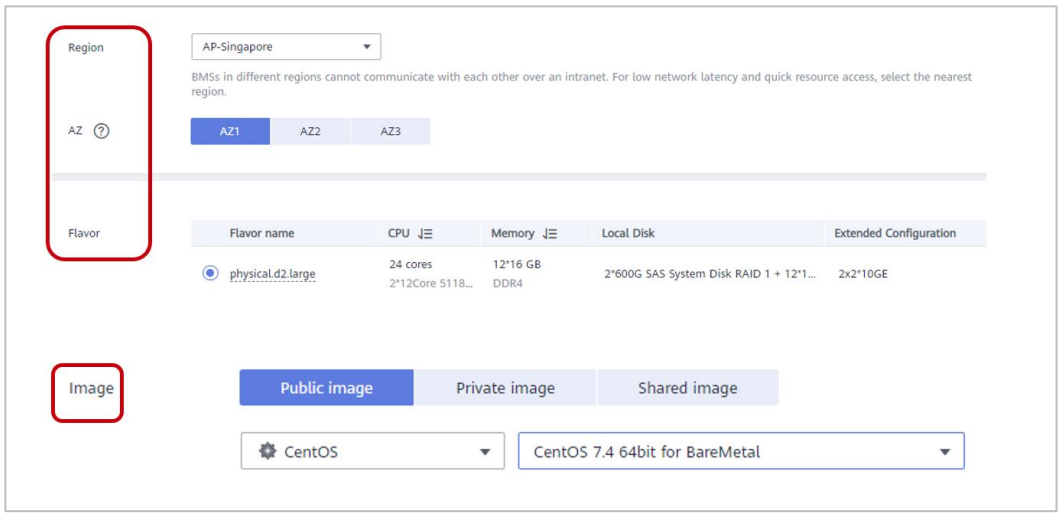


Figure 3-14 Basic settings

You need to set the following parameters:

- **Region** and **AZ**: BMSs in different regions cannot communicate with each other over a VPC. For low network latency and quick resource access, select the region nearest to your target users.
- **Flavor** and **Image**: Select a flavor and image based on service requirements.

Step 2 Configure the network.

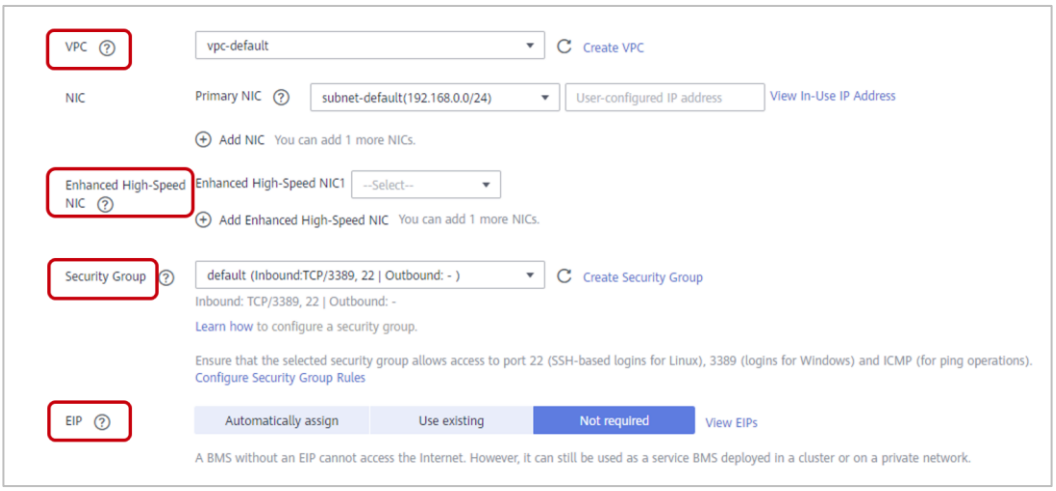


Figure 3-15 Network configuration

When you use VPC for the first time, the system automatically creates a VPC for you, including a security group and NIC. The default subnet segment is 192.168.1.0/24 and the subnet gateway is 192.168.1.1. Dynamic Host Configuration Protocol (DHCP) is enabled for the subnet.

Security groups are used to control access to BMSs. You can define different access control rules for a security group, and these rules take effect for all BMSs added to this security group. When creating a BMS, you can only select a single security group, but after the BMS is created, you can associate it with additional groups.

Five types of networks are available for BMS: VPC, high-speed network, enhanced high-speed network, user-defined VLAN, and InfiniBand network. They are isolated from each other.

- Virtual Private Cloud (VPC)

  A VPC is a logically isolated, configurable, and manageable virtual network. It helps to improve the security of BMSs in the cloud system and simplifies network deployment. You can configure security groups, VPNs, IP address segments, and bandwidth in a VPC. In this way, you can easily manage and configure internal networks and make secure and quick network changes. You can also customize access rules to control BMS access within a security group and across security groups to enhance BMS security.

- High-speed network

  A high-speed network is an internal network connecting BMSs. It provides high bandwidth for connecting BMSs in the same AZ. If you want to deploy services requiring high throughput and low latency, you can create high-speed networks. Currently, the BMS service supports high-speed networks with up to 10 Gbit/s of bandwidth.

- Enhanced high-speed network

  Enhanced high-speed networks use upgraded hardware and software and provide performance superior to high-speed networks.

  An enhanced high-speed network has the following advantages over a high-speed network:

- The bandwidth is at least 10 Gbit/s.
- The number of network planes can be customized and up to 4,000 subnets are supported.
- VMs on a BMS can access the Internet.

- User-defined VLAN

  You can use any 10GE Ethernet NICs that are not being used by the system to configure user-defined VLANs. QinQ technology is used to isolate networks and provide additional physical planes and bandwidths. You can create VLANs to isolate network traffic. Two NICs must be bonded for high availability. User-defined VLANs in different AZs cannot communicate with each other.

- IB network

  An IB network features low latency and high bandwidth and is good for High Performance Computing (HPC) projects. It uses 100 Gbit/s Mellanox IB NICs, dedicated IB switches, and controller software UFM for network communication and management, and uses the Partition Key to isolate IB networks of different users (similar to VLANs in an Ethernet).

Step 3 Configure advanced settings.

Configure the BMS name, login mode, and advanced settings.

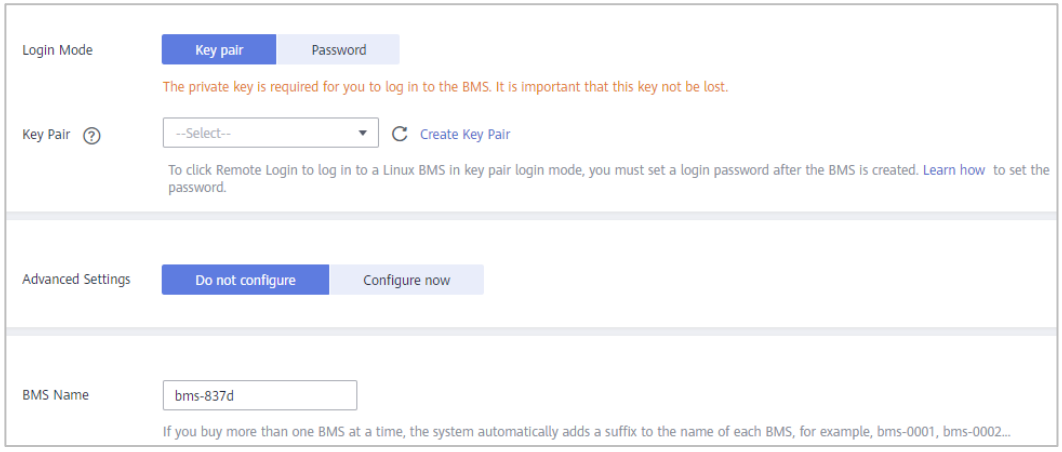


Figure 3-16 Advanced settings

You can choose to use a key pair or password for remote login authentication. For a Linux BMS, you are advised to choose key pair authentication. You can create a key pair and download the private key for remote login authentication. For security purposes, a private key can be downloaded only once, so take care not to lose your downloaded private keys. You can also import the public keys of your existing key pairs to Huawei Cloud, and then use the corresponding private keys to authenticate remote logins.

3.2.6 How to Use a BMS

3.2.6.1 Reinstalling the OS

If the OS of a BMS fails to start, gets infected by a virus, or requires operation improvements, reinstall the OS.

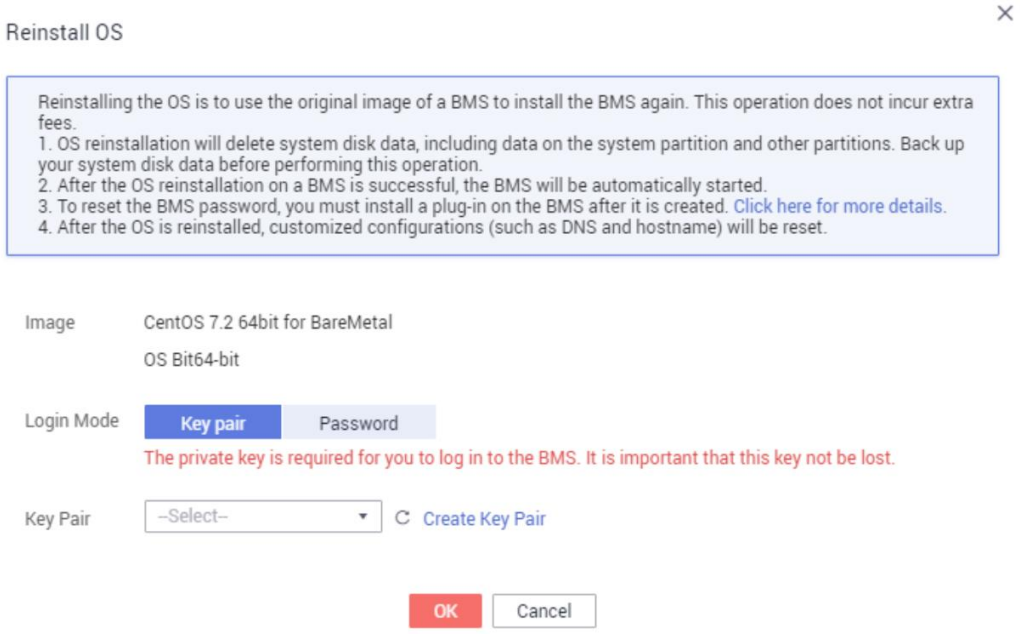


Figure 3-17 Reinstalling the OS

Precautions for reinstalling a BMS OS:

- Reinstalling the OS will interrupt services running on the BMS.
- Reinstalling the OS destroys all of the data on all of the partitions of the system disk. Back up data before performing this operation.
- Do not stop or restart the BMS during the reinstallation, or the reinstallation may fail.
- After the OS is reinstalled, custom configurations that had previously existed, such as a DNS and hostname, will be gone. These customizations will have to be repeated.

3.2.7 Application Scenarios

- Securities and finance

  Financial and security industries have high compliance requirements, and some customers have strict data security requirements. BMSs meet requirements for exclusive, dedicated resource usage, data isolation, as well as operation monitoring and tracking.

- High-performance computing

In certain scenarios, such as supercomputing centers and DNA sequencing, a large amount of data needs to be processed. They have high computing performance, stability, and timeliness requirements. BMSs can meet all of these requirements.

- Core databases

  Some critical database services cannot be deployed on VMs and must be deployed on physical servers that feature dedicated resources, isolated networks, and assured performance. BMSs are dedicated for each individual user, meeting the isolation and performance requirements.

- Mobile apps

  Kunpeng-powered BMSs are fully compatible with the ARM instruction sets used by many terminals, and they provide a one-stop solution for the development, testing, launch, and usage phases of mobile apps, especially mobile phone games.

3.3 IMS

3.3.1 What Is IMS?

Image Management Service (IMS) allows you to manage the entire lifecycle of your images. You can create ECSs or BMSs from public, private, or shared images. You can also create a private image from a cloud server or an external image file to make it easier to migrate workloads to the cloud or in the cloud.

An image is a server or disk template that contains an operating system (OS), service data, and necessary application software, such as database software.

3.3.2 Image Types

Images can be public, private, shared, or Marketplace images. A public image is a standard image provided by the cloud platform. A private image is created by users. A shared image is a private image another user has shared with you. A Marketplace image is provided by service providers who have extensive experience configuring and maintaining cloud servers. All the Marketplace images are thoroughly tested and have been approved by Huawei Cloud before being published.

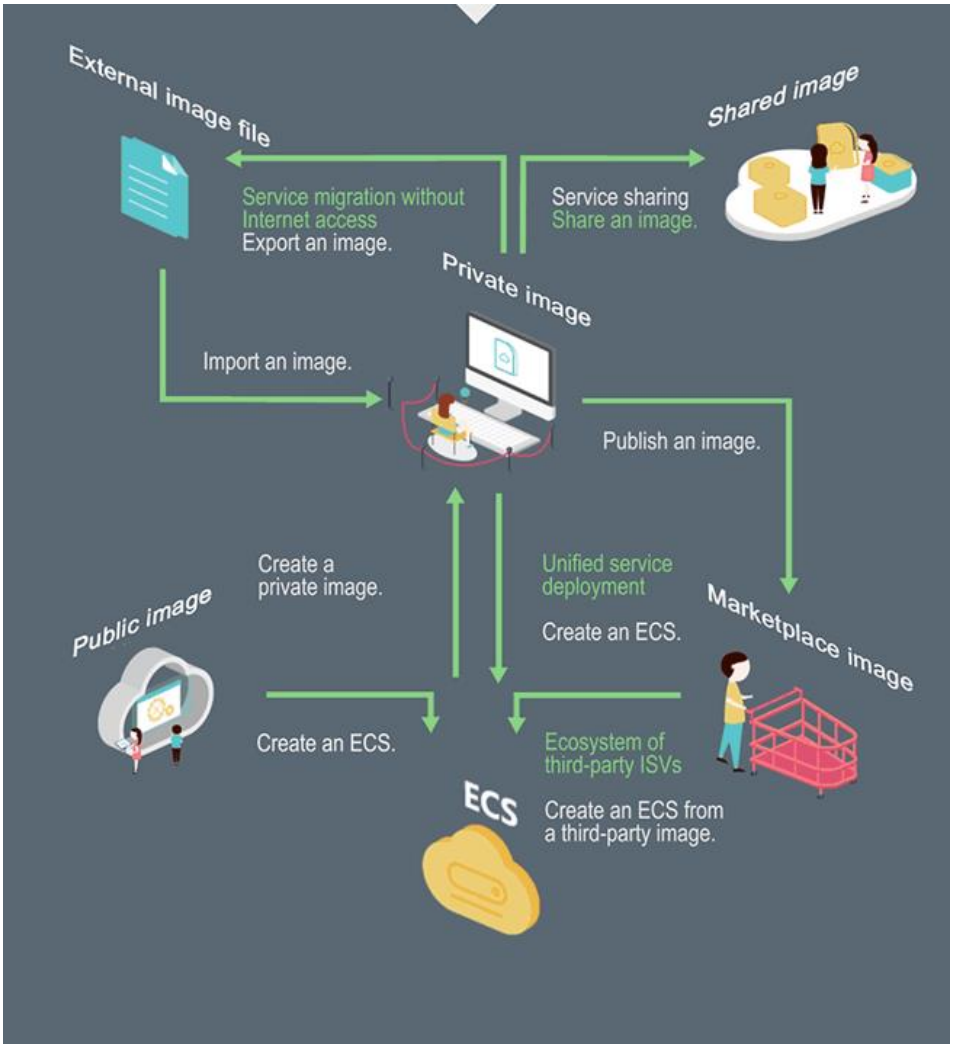


Figure 3-18 IMS architecture

- Public image: A public image is a standard image provided by the cloud platform and is available to all users. It contains an OS and various preinstalled public applications. If a public image does not contain the application environment or software you need, you can use a public image to create an ECS and then install the software you need. Public images include the following OSs to choose from: Windows, CentOS, Debian, openSUSE, Fedora, Ubuntu, EulerOS, and CoreOS.
- Private image: A private image is only available to the user who created it. It contains an OS, service data, preinstalled public applications, and custom applications that the image creator added. A private image can be a system disk image, data disk image, or full-ECS image.
  - A system disk image contains an OS and pre-installed software for various services. You can use a system disk image to create ECSs and migrate your services to the cloud.
  - A data disk image contains only service data. You can use a data disk image to create EVS disks and use them to migrate your service data to the cloud.
  - A full-ECS image contains an OS, pre-installed software, and service data.

- Shared image: A shared image is a private image another user has shared with you.
- Marketplace image: A Marketplace image is a third-party image published in the Marketplace. It has an OS, application environment, and software pre-installed. You can use these images to deploy websites and application development environments in just a few clicks. No additional configuration is required. Marketplace images are provided by service providers who have extensive experience configuring and maintaining cloud servers. All the images are thoroughly tested and have been approved by Huawei Cloud before being published.

3.3.3 Advantages

The following figure shows the IMS advantages.

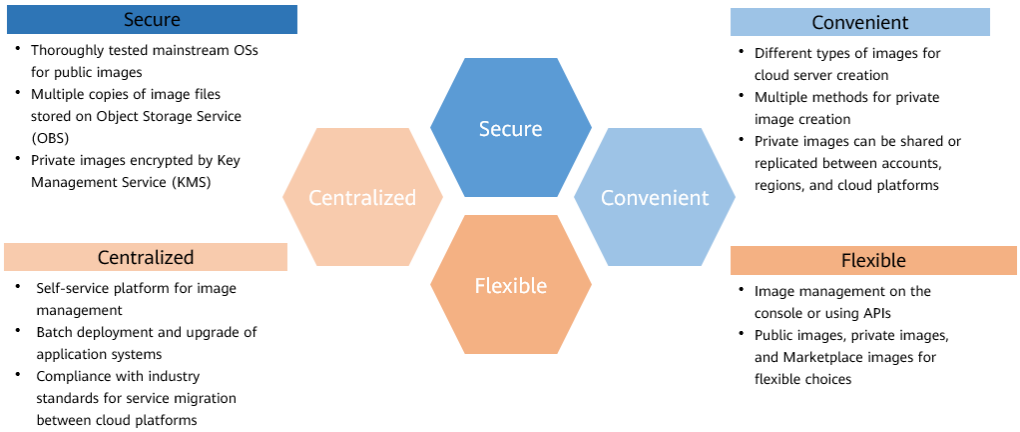


Figure 3-19 IMS advantages

- Convenient

  You can use different types of images to create cloud servers in a batch, simplifying service deployment.

  You can create private images from ECSs, BMSs, or external image files. When you create a private image, you can select the system or data disks of a cloud server or at its entirety. Private images can be transferred between accounts, regions, or cloud platforms through image sharing, replication, and export.

- Secure

  Public images use Huawei EulerOS and mainstream OSs such as Windows Server, Ubuntu, and CentOS. These OSs have been thoroughly tested to provide secure and stable services.

  Multiple copies of image files are stored on Object Storage Service (OBS), which provides excellent data reliability and durability. Private images can be encrypted for data security by using envelope encryption provided by Key Management Service (KMS).

- Flexible

  You can manage images through the management console or using APIs.

  You can use a public image to deploy a general-purpose environment, or use a private image or Marketplace image to deploy a custom environment.

You can use IMS to migrate servers to the cloud or in the cloud, and to back up server environments.

- Centralized

  IMS provides a self-service platform that simplifies image management and maintenance. IMS allows you to uniformly deploy and upgrade application systems, improving O&M efficiency and ensuring consistency of application environments. Public images comply with industry standards. Pre-installed components only include clean installs, and only kernels from well-known third-party vendors are used to make it easier to transfer images from or to other cloud platforms.

3.3.4 How to Create an Image

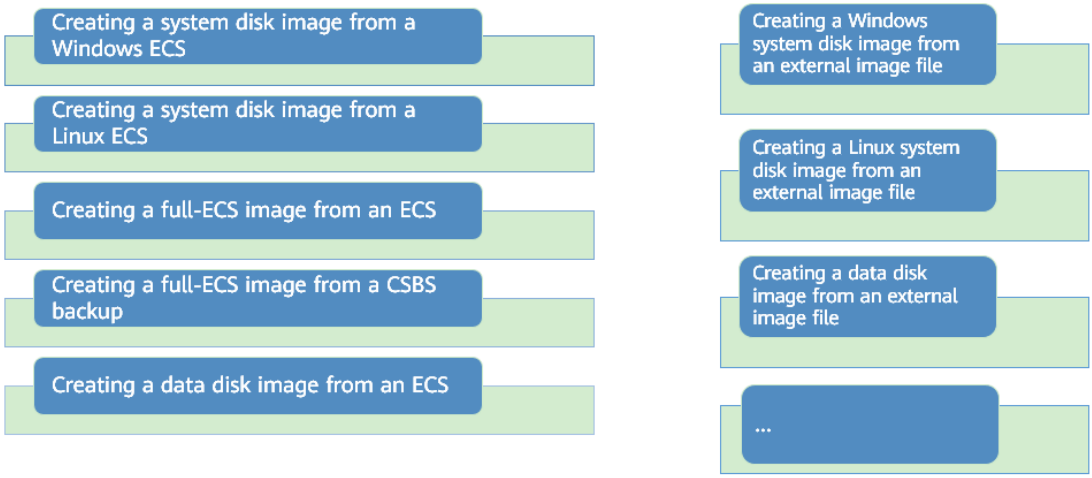


Figure 3-20 Creating an image

You can use an ECS or external image file to create an ECS private image. You can also:

- Use an ISO file to create an ECS system disk image.
- Use a CBR backup to create a full-ECS image.
- Use a BMS to create a system disk image.

The following process shows how to create a system disk image from a Windows ECS.

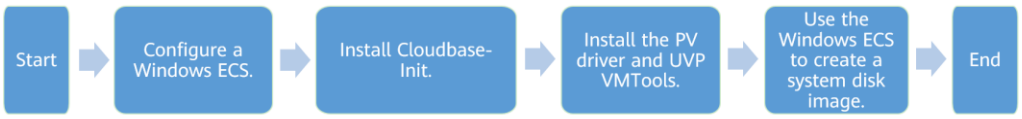


Figure 3-21 Process of creating an image

Prepare a Windows ECS and check whether the ECS NIC is configured to use DHCP.

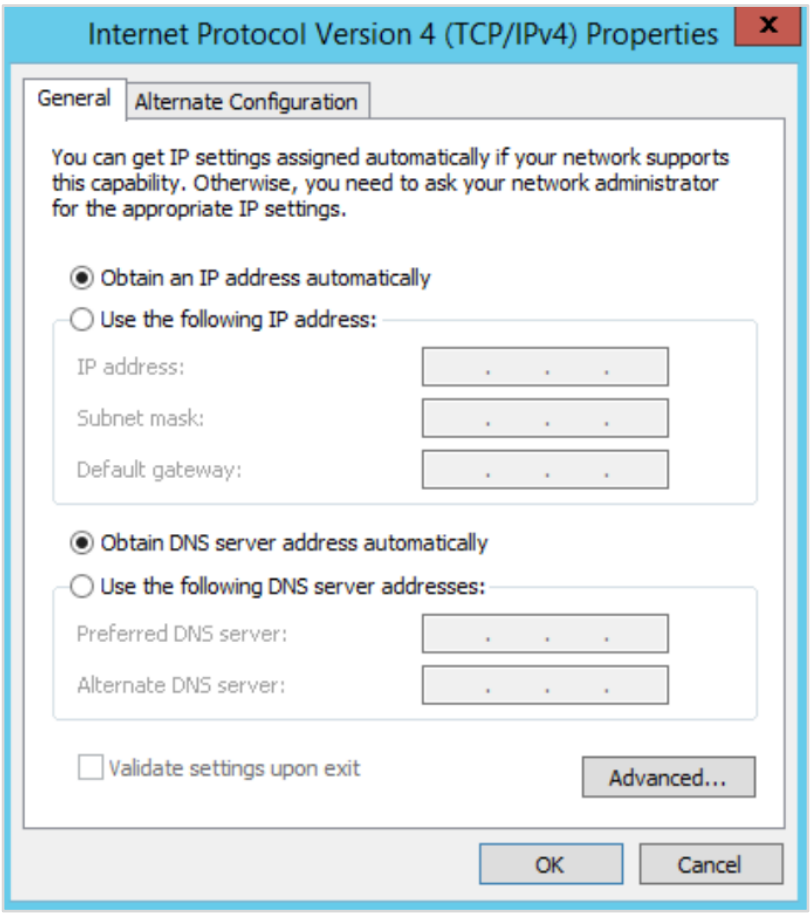


Figure 3-22 Obtaining IP addresses (Windows)

If the Windows ECS is using a static IP address, you will have to log in to the ECS and change the network settings to use DHCP.

- Log in to the Windows ECS. Choose **Start** > **Control Panel** > **Network and Internet** > **Network and Sharing Center** > *A connection with a static IP address* > **Properties** > **General**.
- On the **General** tab, select **Obtain an IP address automatically** and **Obtain DNS server address automatically**, and click **OK**.

To ensure that ECSs created from a private image are configurable, you are advised to install Cloud-init (Linux)/Cloudbase-init (Windows) on the ECS before using it to create a private image.

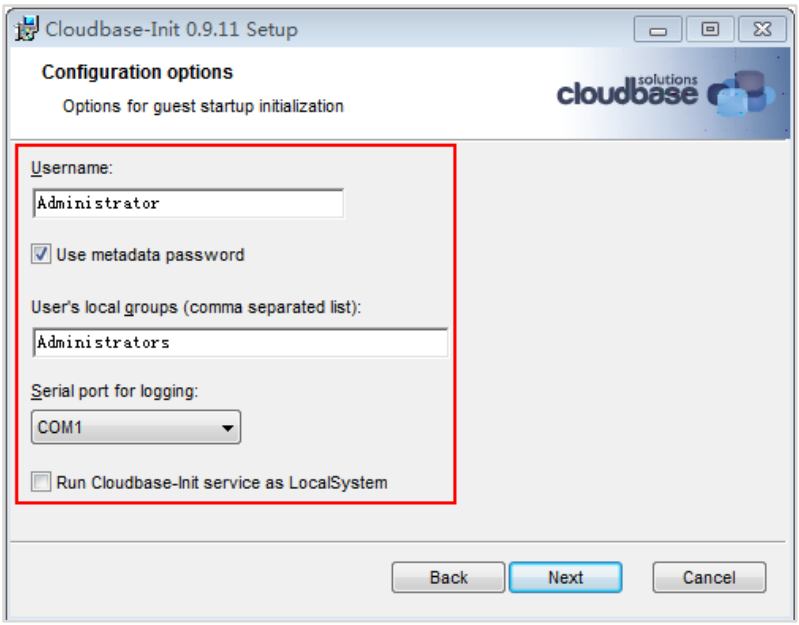


Figure 3-23 Installing Cloudbase-Init

To ensure that ECSs created from a private image support both Xen and KVM virtualization, install the PV driver and UVP VMTools on the ECS before using it to create a private image.

Then, use the ECS to create a Windows system disk image.

- On the **Image Management Service** page, click **Create Image**.
- In the **Image Type and Source** area, select **System disk image** for **Type**.
- By default, **ECS** is selected for **Source**. Select an ECS from the drop-down list.

3.3.5 How to Manage an Image

3.3.5.1 Modifying Image Information

You can modify the image name, description, minimum and maximum memory, NIC multi-queue, and SR-IOV driver.

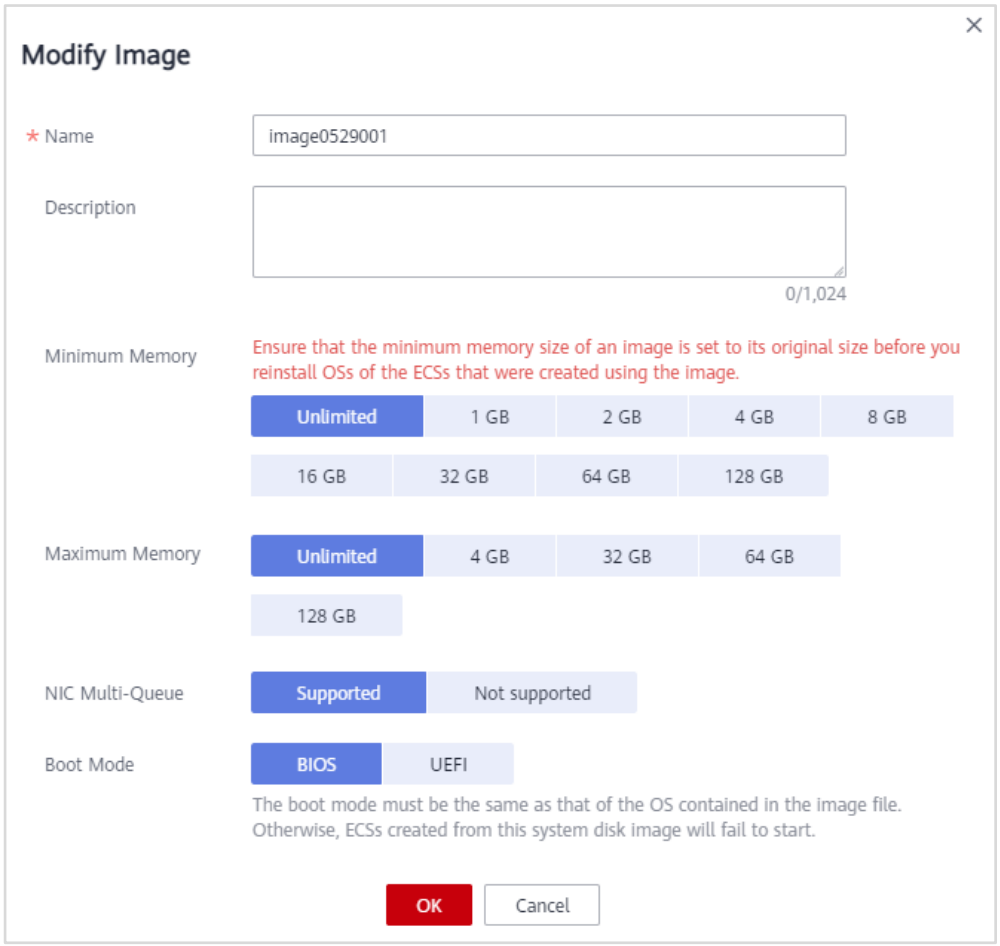


Figure 3-24 Modifying image information

- Only private images that are in the **Normal** state can be modified.
- NIC multi-queue enables multiple CPUs to process NIC interruptions for load balancing.
- After the SR-IOV driver is installed for an image, the network performance of ECSs created from the image will be greatly improved.

3.3.5.2 Deleting an Image

You can delete images that will be no longer used. Deleting an image does not affect the ECSs created from that image.

Note that:

- Deleted private images cannot be retrieved. Perform this operation only when absolutely necessary.
- After a private image is deleted, it cannot be used to create cloud servers or EVS disks.

- After a private image is deleted, ECSs created from the image can still be used and are still billed. However, the OS cannot be reinstalled for the ECSs and an ECS with the same configuration cannot be recreated.

Deleting the source image of a replicated image has no effect on the replicated image. Similarly, deleting a replicated image has no effect on its source.

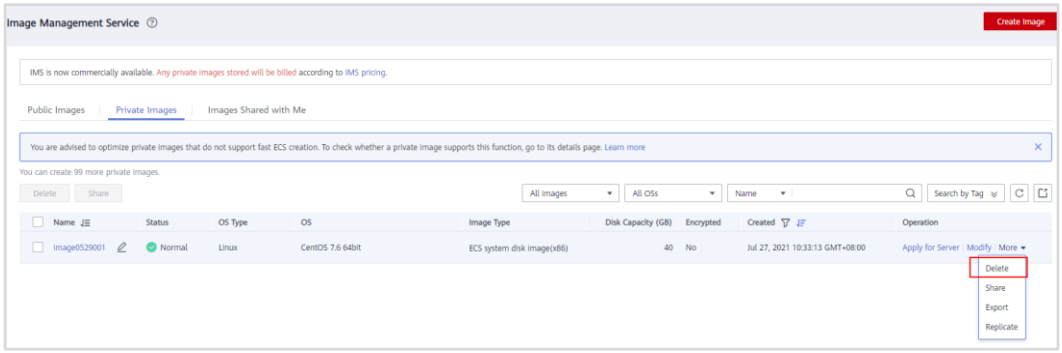


Figure 3-25 Deleting an image

3.3.5.3 Sharing an Image

You can share your private images with other accounts. These accounts can use your shared private images to quickly create ECSs or EVS disks.

- You can share your private images with others.
- You can share images, stop sharing images, and add or delete tenants that can use the shared images.
- The recipient can choose to accept or reject the shared images, or remove images they have previously accepted.

3.3.5.4 Encrypting an Image

You can create an encrypted image to securely store data by using envelope encryption provided by Key Management Service (KMS). Encrypted images can be created from external image files or encrypted cloud servers.

- Encrypted images cannot be shared with other users or published in the Marketplace.
- The system disk of an ECS created from an encrypted image is also encrypted, and its key is the same as the image key.
- If an ECS has an encrypted system disk, private images created from the ECS are also encrypted.

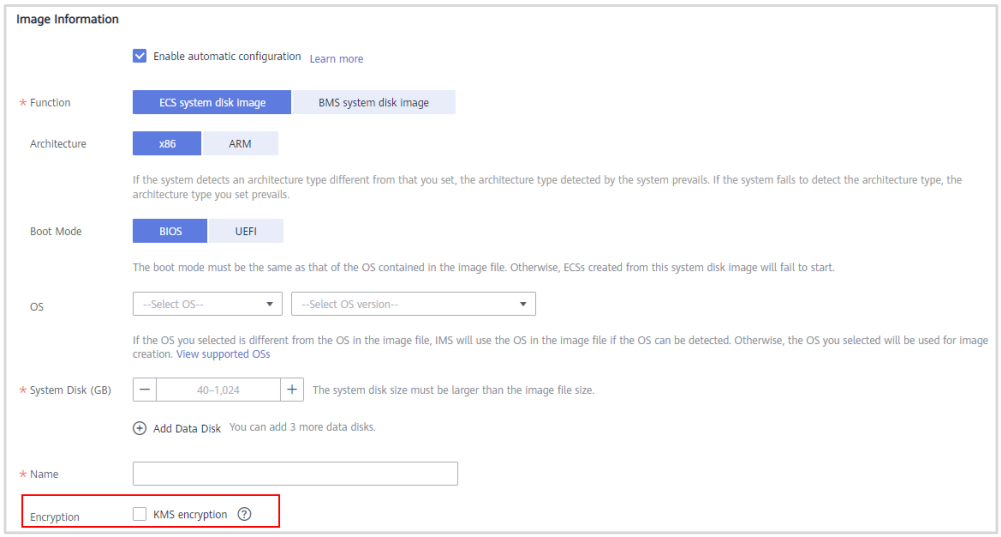


Figure 3-26 Encrypting an image

3.3.5.5 Replicating an Image Within a Region

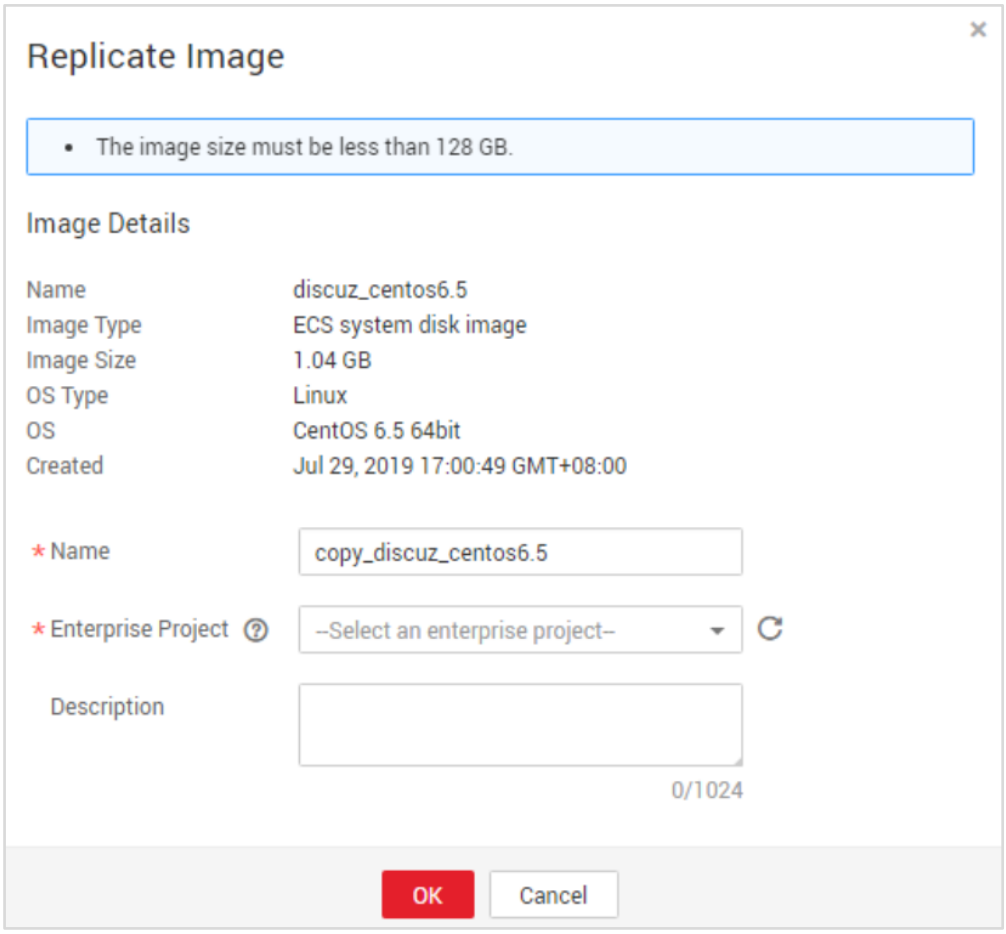


Figure 3-27 Replicating an image

By replicating images within a region, you can convert encrypted and unencrypted images into each other or enable some advanced features, for example, quick cloud server provisioning.

You may need to replicate an image in the following scenarios:

- Creating an unencrypted version of an encrypted image

  Encrypted images cannot be shared with other users or published in the Marketplace. If you want to publish or share an encrypted image, you need to create an unencrypted version.

- Replicating an encrypted image

  The key used for encrypting an image cannot be changed directly. If you want to change the key of an encrypted image, you can replicate this image and encrypt the new image using a different key.

- Creating an encrypted version of an unencrypted image

  If you want to encrypt an unencrypted image, you can replicate the image and encrypt the new image using a key.

3.3.5.6 Exporting an Image

You can export an image if you want to store the image on specified storage devices or use the image on other cloud platforms.

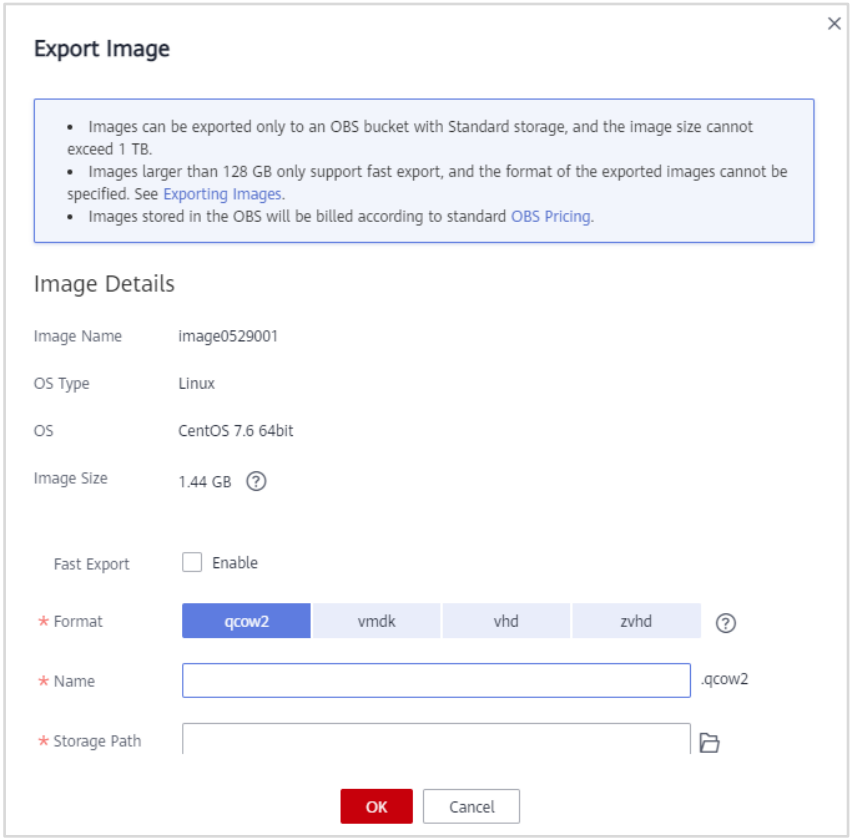


Figure 3-28 Exporting an image

You can export private images to OBS buckets in a specified format and then download the images from the buckets to specified storage devices. Images exported in different formats may vary in size. You will be charged for the OBS storage occupied by exported images.

3.3.6 Application Scenarios

Usually, IMS is used in the following scenarios:

- Migrating servers to the cloud or in the cloud

  You can import local images to the cloud platform and use the images to quickly create cloud servers for service migration to the cloud. A variety of image types can be imported, including VHD, VMDK, QCOW2, and RAW.

  You can also share or replicate images across regions to migrate ECSs between accounts and regions.

- Deploying a specific software environment

  Use shared or Marketplace images to quickly build custom software environments without having to manually configure environments or install any software. This is especially useful for Internet startups.

- Batch deploying software environments

  Prepare an ECS with an OS, the partition arrangement you prefer, and software installed to create a private image. You can use the image to create batch clones of your custom ECS.

- Backing up server environments

  Create an image from an ECS to back up the ECS. If the ECS breaks down due to software faults, you can use the image to restore the ECS.

3.4 AS

3.4.1 What Is AS?

Auto Scaling (AS) automatically adjusts Elastic Cloud Server (ECS) and bandwidth resources to keep up with changes in demand based on pre-configured AS policies. When service demands increase, AS scales out ECS or bandwidth resources to ensure stable service capabilities. When service demands decrease, AS scales in ECS or bandwidth resources to reduce costs.

3.4.2 Key Concepts

Before learning how to use AS, you need to understand a few basic concepts.

- AS group: An AS group consists of a collection of instances for use in the same scenario. It is the basis for enabling or disabling AS policies and performing scaling actions.

- AS configuration: An AS configuration is a template specifying specifications for the instances to be added to an AS group. The specifications include the ECS type, vCPUs, memory, image, disk, and login mode.
- AS policy: An AS policy can trigger scaling actions to adjust the number of instances in an AS group. An AS policy defines the conditions for triggering a scaling action and the operation that will be performed. When the triggering condition is met, the system automatically triggers a scaling action.
- Scaling action: A scaling action adds instances to or removes instances from an AS group. It ensures that the number of instances deployed for an application is the same as the expected number of instances by adding or removing instances when the triggering condition is met, which improves system stability.
- Cooldown period: To prevent an alarm policy from being repeatedly triggered for the same event, we use a cooldown period. The cooldown period specifies how long any alarm-triggered scaling action will be disallowed after a previous scaling action is complete. A cooldown period does not work for scheduled or periodic scaling actions, but a cooling period countdown starts after any scheduled or periodic scaling action is complete.

  For example, if you set the cooldown period to 300 seconds (5 minutes), and there is a scaling action scheduled for 10:32, but a previous scaling action was complete at 10:30, any alarm-triggered scaling actions will be denied during the cooldown period from 10:30 to 10:35, but the scheduled scaling action will still be triggered at 10:32. If the scheduled scaling action ends at 10:36, a new cooldown period starts at that time and ends at 10:41.
- Bandwidth scaling: AS automatically adjusts a bandwidth based on the configured bandwidth scaling policies. AS can only adjust the bandwidths of pay-per-use EIPs and shared bandwidths. It cannot adjust yearly/monthly bandwidths.

3.4.3 Architecture

AS allows you to scale ECS instances and bandwidths.

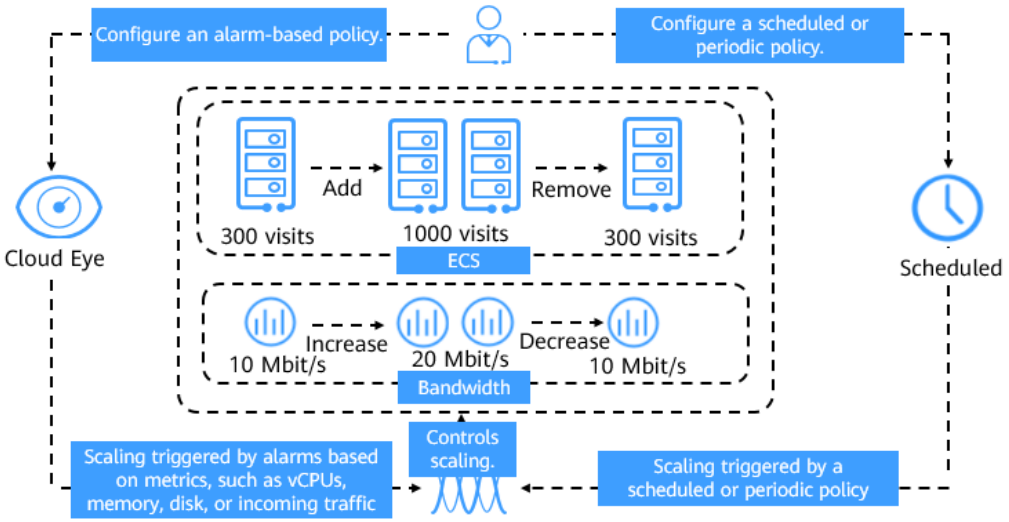


Figure 3-29 AS architecture

- Scaling control: You can configure AS policies, configure metric thresholds, and schedule when different scaling actions are taken. AS will trigger scaling actions on a repeating schedule, at a specific time, or when the configured thresholds are reached.
- Policy configuration: You can configure alarm-based, scheduled, and periodic policies as needed.
- Alarm-based: You can configure scaling actions to be taken when alarm metrics such as vCPU, memory, disk, and inbound traffic reaches the thresholds.
- Scheduled: You can schedule scaling actions to be taken at a specific time.
- Periodic: You can configure scaling actions to be taken at scheduled intervals, a specific time, or within a particular time range.

When Cloud Eye generates an alarm for a monitoring metric, for example, CPU usage, AS automatically increases or decreases the number of instances in the AS group or the bandwidth. When the configured triggering time arrives, a scaling action is triggered to increase or decrease the number of ECS instances or the bandwidth.

3.4.4 Advantages

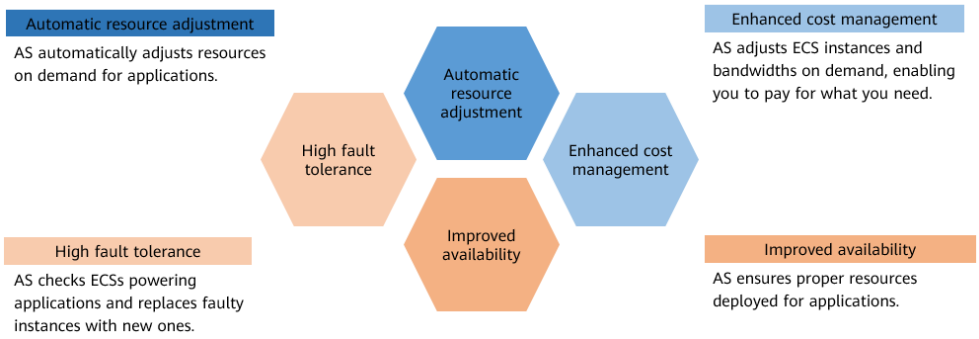


Figure 3-30 AS advantages

AS has following advantages:

- Automatic resource scaling: AS adds ECS instances and increases bandwidth for your applications when the access volume increases and removes unneeded resources when the access volume drops, ensuring system stability and availability.
- Enhanced cost management: AS enables you to use instances and bandwidths on demand by automatically adjusting resources for your applications, eliminating waste of resources and reducing costs.
- Improved availability: AS ensures that you always have the right amount of resources available to handle the fluctuating load of your applications. When working with ELB, AS automatically associates a load balancing listener with any instances newly added to an AS group. Then, ELB automatically distributes access traffic to all healthy instances in the AS group through the listener, which improves system availability.
- High fault tolerance: AS monitors the statuses of instances in an AS group, and replaces any unhealthy instances it detects with new ones.

3.4.5 How to Use AS

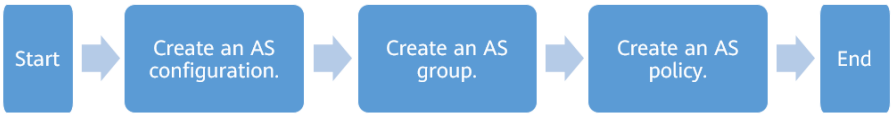


Figure 3-31 How to use AS

Step 1 Create an AS configuration. You can create an AS configuration based on an existing ECS instance or based on a new template:

- Based on an existing ECS instance

  You can use an existing ECS instance to quickly create an AS configuration. Then, the specifications of this instance, such as the vCPUs, memory, image, disk, and ECS type, will be applied to instances added to the AS group in scaling actions.

- Based on a new template

  If you have special requirements for the specifications of the ECS instances used for capacity expansion, specify the specifications in a template and use it to create the AS configuration. Then, the specifications will be applied to any instances added to the AS group in scaling actions.

Step 2 Create an AS group. When creating an AS group, you need to configure parameters such as **Multi-AZ Scaling Policy**, **Max. Instances**, **Expected Instances**, **Min. Instances**, **Instance Removal Policy**, and **Health Check Method**.

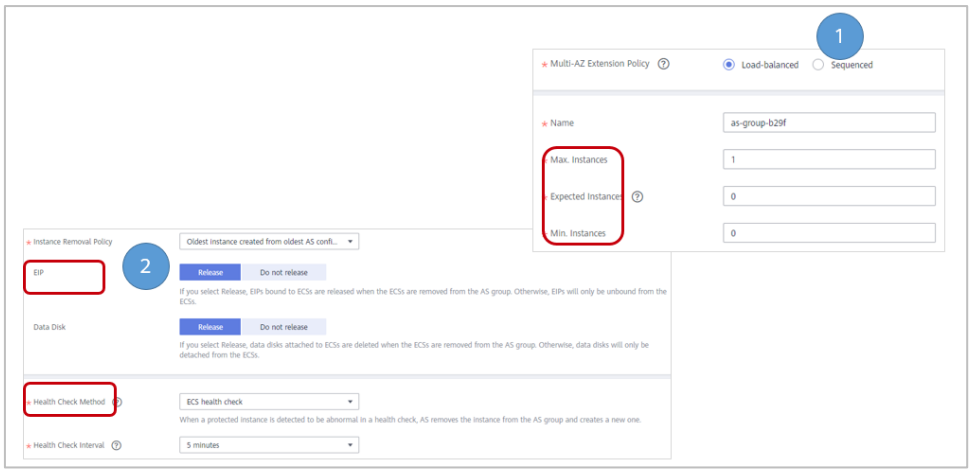


Figure 3-32 Main parameters for creating an AS group

- **Multi-AZ Scaling Policy**: Required only when two or more AZs are selected.
- **Max/Min. Instances**: Specifies the minimum or maximum number of ECS instances in an AS group.
- **Expected Instances**: Specifies the number of ECS instances that are expected to run in an AS group. This value has to be between the minimum and maximum numbers of instances.

- **Instance Removal Policy**: Controls which instances are first to be removed during scale-in actions. The instances that are not in the AZs used by an AS group will be removed first.

Step 3 Create an AS policy.

Main parameters include **Policy Type** and **Cooldown Period**.

If the service load is unpredictable, you can configure alarm-based AS policies. These policies are used to trigger scaling actions based on real-time monitoring data such as CPU usage to dynamically adjust the number of instances in the AS group. AS recounts the cooldown period after a scaling action is complete. During the cooldown period, scaling actions triggered by alarms will be denied. Scheduled and periodic scaling actions are not affected.

3.4.6 Application Scenarios

AS automatically scales resources to keep up with service demands based on pre-configured AS policies. With automatic resource scaling, you can enjoy reduced costs, improved availability, and high fault tolerance. AS is good for the following scenarios:

- Heavy-traffic forums: Service load changes of a heavy-traffic forum website are difficult to predict. AS can dynamically adjust the number of ECS instances based on monitored ECS metrics, such as vCPU and memory usage.
- E-commerce: Large-scale e-commerce promotions can attract visits that may break your website. AS can automatically add ECS instances and increase the bandwidth to ensure that promotions will go smoothly.
- Live streaming: A live streaming website might broadcast popular programs during certain times, for example, from 14:00 to 16:00 every day. AS can automatically add ECS instances and increase the bandwidth during this period to ensure smooth a viewing experience.

3.5 CCE

3.5.1 What Is CCE?

Cloud Container Engine (CCE) is a highly scalable, enterprise-class, managed Kubernetes service for you to run containers and applications. With CCE, you can easily deploy, manage, and scale containerized applications on Huawei Cloud.

3.5.2 Key Concepts

CCE supports native Kubernetes APIs and kubectl, and provides a graphical console that delivers an E2E user experience. Before using CCE, you are advised to understand related basic concepts.

- Cluster

  A cluster is a combination of cloud resources required for container running, such as cloud servers and load balancers. In a cluster, one or more cloud servers (also called

nodes) are deployed in the same subnet to provide compute capacity for container running.

- Node

  A node is a server (a VM or PM) on which containerized applications run. The node agent (kubelet) runs on each node to manage containers on the nodes. The number of nodes in a cluster can be scaled.

- Node Pool

  A node pool contains one node or a group of nodes with identical configuration in a cluster.

- Pod

  Pod is the smallest and simplest unit in Kubernetes that you create or deploy. A pod encapsulates an application container (or, in some cases, multiple containers), storage resources, a unique network IP address, and options that govern how the containers should run.

- Container

  A container is a running instance of a Docker image. Multiple containers can run on the same node. Containers are actually software processes. Unlike traditional software processes, containers have separate namespaces and do not run directly on a host.

- Workload

  A workload is an application running on Kubernetes. No matter how many components are there in your workload, you can run it in a group of Kubernetes pods. A workload is an abstract model of a group of pods in Kubernetes. Workloads classified in Kubernetes include Deployments, StatefulSets, DaemonSets, jobs, and cron jobs.

- Image

  Docker creates an industry standard for packaging containerized applications. Docker images are like templates that include everything needed to run containers, and are used to create Docker containers. In other words, a Docker image is a special file system that includes the required programs, libraries, resources, and configuration files to make an application run. It also contains parameters you can configure for your application, such as anonymous volumes, environment variables, and users. An image does not contain any dynamic data, and its content remains unchanged after being built. When deploying containerized applications, you can use images from Docker Hub, SoftWare Repository for Container (SWR), and your private image registries. For example, a Docker image can contain a complete Ubuntu operating system, in which only the required programs and dependencies are installed. Images become containers at runtime. That is, containers are created from images. Containers can be created, started, stopped, deleted, and suspended.

- Layer-7 load balancing (ingress)

  An ingress is a set of routing rules for requests entering a cluster. It provides Services with URLs, load balancing, SSL termination, and HTTP routing for external access to the cluster.

- Image repository

An image repository stores Docker images that can be used to deploy containerized services.

3.5.3 Architecture

CCE is deeply integrated with high-performance Huawei Cloud services, including compute (ECS/BMS), network (VPC/EIP/ELB), and storage (EVS/OBS/SFS) services. It supports heterogeneous computing architectures such as GPU, NPU, and Arm. By using multi-AZ and multi-region disaster recovery, CCE ensures high availability of Kubernetes clusters.

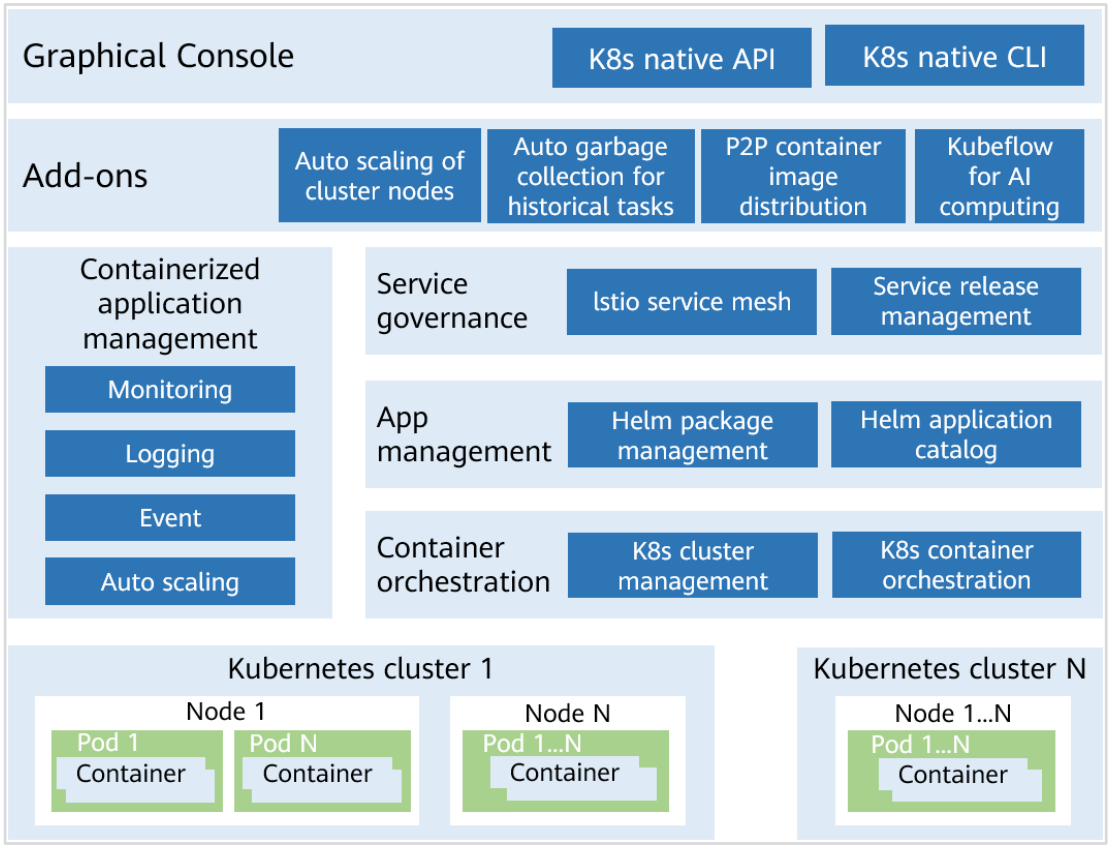


Figure 3-33 CCE architecture

You can create multiple Kubernetes clusters on Huawei Cloud CCE and use the clusters to manage your nodes. CCE allows you to manage containers via the console or the native Kubernetes CLI. Supported functions include container O&M, service governance, application management, and container orchestration. CCE also supports extensions and add-ons, such as auto node scaling add-ons and auto garbage collection of historical tasks.

3.5.4 Advantages

CCE has the following four advantages.

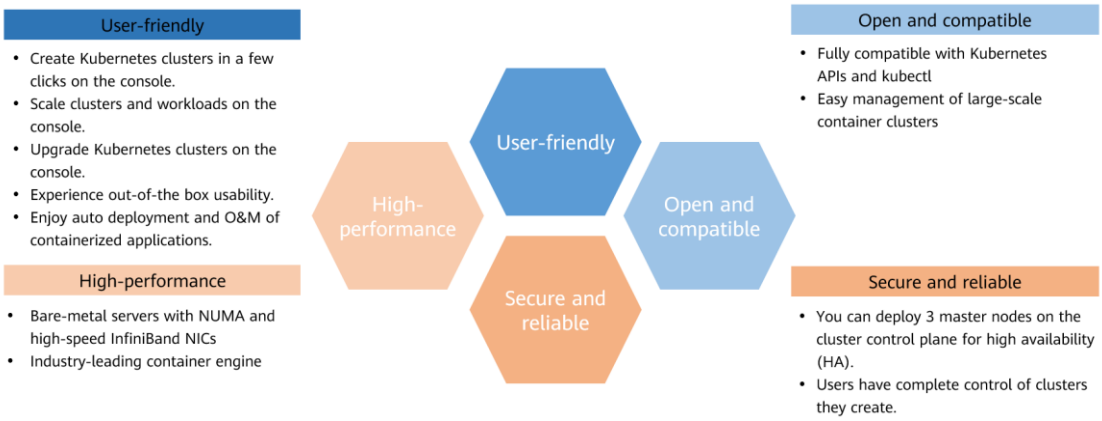


Figure 3-34 CCE advantages

- Easy to use
  - Creating a Kubernetes cluster is as easy as a few clicks on the console. You can create either VM nodes or bare-metal nodes, or both, in a cluster.
  - From auto deployment to O&M, you can manage your containerized applications all in one place throughout their lifecycle.
  - You can also scale your clusters and workloads in just a few clicks on the console. Auto scaling policies can be flexibly combined to deal with in-the-moment load spikes.
  - The console enables you to easily upgrade your clusters.
  - Application Service Mesh (ASM) and Helm charts are pre-integrated, delivering out-of-the-box usability.
- Robust performance
  - CCE draws on years of field experience in computing, network, storage, and heterogeneous infrastructure. You can concurrently launch containers at scale.
  - The bare-metal NUMA architecture and high-speed InfiniBand network cards yield three- to five-fold improvement in computing performance.
- Secure and reliable
  - Highly available: Each cluster has three master nodes, avoiding a single point of failure on the cluster control plane. Faults in one or two of the master nodes do not interrupt the whole cluster. CCE allows you to deploy nodes and workloads in a cluster across AZs. Such a multi-active architecture ensures service continuity against host faults, data center outages, and natural disasters.
  - Secure: Clusters are private and completely controlled by users with HUAWEI CLOUD accounts and Kubernetes RBAC capabilities deeply integrated. Users can set different RBAC permissions for sub-users on the GUI.
- Open and compatible
  - CCE streamlines deployment, resource scheduling, service discovery, and dynamic scaling of applications that run in Docker containers.

– CCE is built on Kubernetes and compatible with Kubernetes native APIs and kubectl (a command line tool). CCE provides full support for the most recent Kubernetes and Docker releases.

3.5.5 How to Use CCE

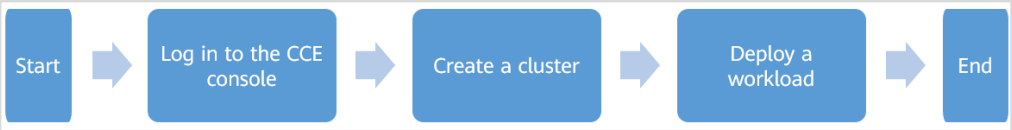


Figure 3-35 Using CCE

The following shows the steps of using CCE:

- Register a Huawei Cloud account and log in to the CCE console.
- Select a cluster type and create a cluster.
- Deploy a workload (application) using an existing or newly created image or orchestration template.

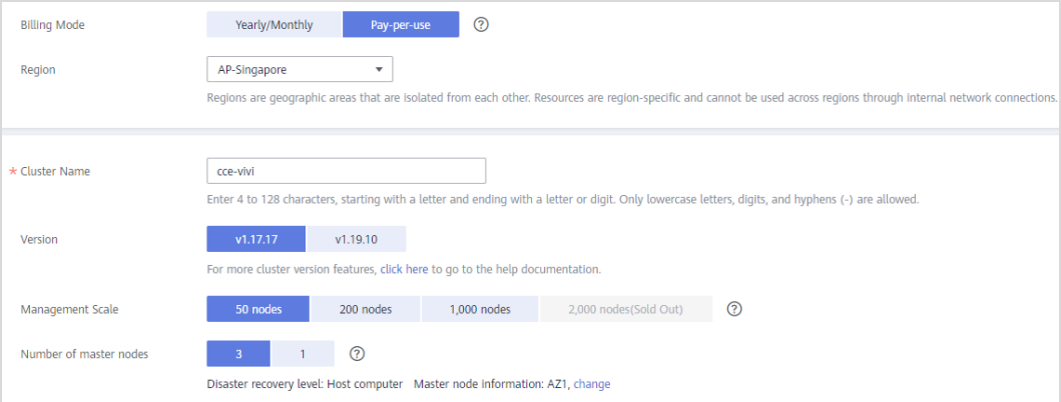


Figure 3-36 Creating a cluster

Configure CCE parameters such as billing mode, region, cluster version, management scale, and number of master nodes.

After the cluster is created, CCE can automatically scale the cluster (adding or releasing worker nodes) according to the scaling policies you configure. For example, when workloads cannot be scheduled into the cluster due to insufficient cluster resources, scale-out will be automatically triggered.

Upgrading a cluster is easy in CCE.

You can use the CCE console to upgrade the Kubernetes version of a cluster.

An upgrade flag will be displayed on the cluster card view if there is a new version for the cluster to upgrade.

Precautions:

- Upgraded clusters cannot be rolled back. Therefore, perform the upgrade during off-peak hours to minimize the impact on your services.

- Before upgrading a cluster, get familiar with the features and differences of each cluster version in the Kubernetes Release Notes. Exceptions may occur after the upgrade if applications are incompatible with the new cluster version.
- Do not shut down or restart nodes during cluster upgrade. Otherwise, the upgrade will fail.
- Before upgrading a cluster, disable auto scaling policies to prevent node scaling during the upgrade. Node scaling will cause the upgrade to fail.
- If you locally modify the configurations of a cluster node, the cluster upgrade may fail or the configuration may be lost after the upgrade. You are advised to modify the configurations on the CCE console (cluster or node pool list page) so that they will be automatically inherited during the upgrade.
- During the cluster upgrade, the running workload services will not be interrupted, but the API server will be temporarily inaccessible.

3.5.6 Application Scenarios

CCE is ideal for the following scenarios:

- Auto Scaling

  Traffic surges brought by promotions and flash sales on online shopping apps and websites Fluctuating service loads of live streaming Increase in the number of game players that go online in certain time periods

  CCE automatically adapts the amount of computing resources to fluctuating service loads according to auto scaling policies you configured. To scale computing resources at the cluster level, CCE adds or reduces cloud servers. To scale computing resources at the workload level, CCE adds or reduces containers.

- Traffic Governance

  Large enterprise systems are becoming more complex, beyond what traditional system architectures can handle. A popular solution is microservice. Complex applications are divided into smaller components called microservices. Microservices are independently developed, deployed, and scaled. The combined use of microservices and containers streamlines microservice delivery while improving application reliability and scalability.

  Microservices make distributed architectures possible. However, more microservices indicate more complexity in O&M, commissioning, and security management of these architectures. Developers are often troubled by writing additional code for microservice governance and integrating the code into their service systems. In this regard, CCE provides an efficient solution to free you from management workload.

  CCE is deeply integrated with Application Service Mesh (ASM), which allows you to complete grayscale release, observe your traffic, and control traffic flow without changing your code.

- DevOps

  Your applications and services may receive a lot of feedback and requirements. To release new features and improve user experience, you need fast continuous integration (CI). An efficient tool to support CI is container. By deploying containers,

you can streamline the process from development, testing, to release and realize continuous delivery (CD).

CCE works with SWR to support DevOps that will automatically complete code compilation, image build, grayscale release, and deployment based on source code. Traditional CI/CD systems can be connected to containerize legacy applications.

- Hybrid Cloud
  - **Multi-cloud deployment for DR and backup**: To achieve high service availability, you can deploy applications on container services from multiple cloud providers. When a cloud is down, application loads will be automatically distributed to other clouds.
  - **Load balancing and auto scaling**: Large enterprise systems often span cloud facilities in different regions. They also need to be automatically resizable — they can start small and then scale up as system load grows. This frees enterprises from the costs of planning, purchasing, and maintaining more cloud facilities than needed and transforms large fixed costs into much smaller variable costs.
  - **Migration to clouds and database hosting**: Finance, security, and other industries whose top concern is data confidentiality want to keep critical systems in local IDCs while moving other systems to the cloud. They want to manage these systems, no matter in local IDCs or in the cloud, in a unified manner.
  - **Decoupling development from deployment**: To ensure IP security, you can set up the production environment on a public cloud and the development environment in your local IDC.

Applications and data can be seamlessly migrated between your on-premises network and the cloud, facilitating resource scheduling and disaster recovery (DR). This is made possible through environment-independent containers, network connectivity between private and public clouds, and the ability to collectively manage containers on CCE and your private cloud.

3.6 Other Compute Services

- Cloud Phone

  Cloud Phone (CPH) is a cloud server that is virtualized based on the Huawei Cloud BMS. It runs the native Android OS and provides the virtual phone functions. Simply put, a cloud phone consists of a cloud server and Android OS. You can remotely control cloud phones to run cloud Android apps in real time, or leverage the basic computing power of cloud phones to efficiently build applications, such as cloud gaming, mobile office, and live entertainment.

- Dedicated Host

  Dedicated Host (DeH) provides dedicated physical hosts on which you can deploy ECS instances for your exclusive use, enhancing computing isolation, security, and performance. When migrating services to a DeH, you can continue using the existing server software license. The Bring Your Own License (BYOL) feature on the DeH improves the autonomous management of your ECS instances at a lower cost.

- Hyper Elastic Cloud Server

  Hyper Elastic Cloud Server (HECS) is a next-generation cloud server with an independent OS and network. It allows you to deploy and manage applications at ease and is cost-effective for low-load scenarios such as building websites and deploying development environments.

4 Network Cloud Services

Network resources are essential to the development of the ICT infrastructure. With network resources, devices and systems can communicate with each other.

This chapter describes the network services provided by Huawei Cloud.

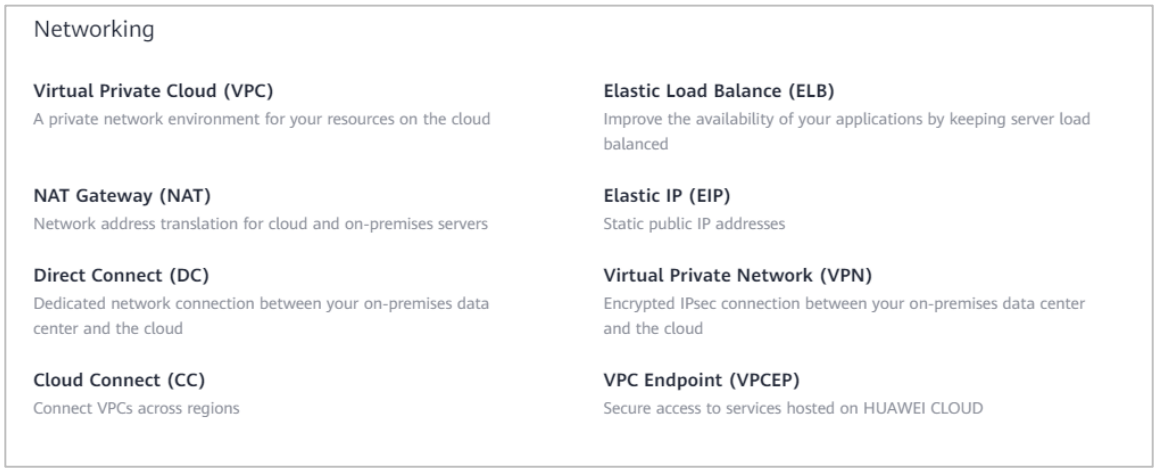


Figure 4-1 Network services

4.1 VPC

4.1.1 What Is VPC?

A Virtual Private Cloud (VPC) is a logically isolated, configurable, and manageable virtual network for cloud resources, such as cloud servers, containers, and databases. It improves resource security and simplifies network deployment on the cloud.

4.1.2 Key Concepts

After learning the definition of the VPC, let's learn some concepts about VPC.

- Subnet

  A subnet is a unique CIDR block with a range of IP addresses in your VPC. All resources in a VPC must be deployed on subnets. Subnets in a VPC cannot overlap with each other. Once a subnet has been created, its CIDR block cannot be modified. By default, ECSs in all subnets of the same VPC can communicate with one another, but ECSs in different VPCs cannot. You can create VPC peering connections to enable ECSs in different VPCs to communicate with one another.

- Elastic IP (EIP)

  The EIP service provides static public IP addresses and scalable bandwidths that enable your cloud resources to communicate with the Internet. You can easily bind an EIP to an ECS, BMS, virtual IP address, NAT gateway, or load balancer, enabling immediate Internet access. Various billing modes are provided to meet diverse business requirements. Each EIP can be used by only one cloud resource at a time.

- Security Group

  A security group provides access control for cloud servers that have the same security requirements within a given VPC. You can define inbound and outbound rules to control traffic to and from the cloud servers in a security group, making your VPC more secure. Your account automatically comes with a security group by default. The default security group allows all outbound traffic and denies all inbound traffic. Your ECSs in this security group can communicate with each other without the need to add rules.

- VPC Peering Connection

  A VPC peering connection is a network connection between two VPCs in one region that enables you to route traffic between them using private IP addresses. You can create a VPC peering connection between your own VPCs, or between your VPC and a VPC of another account within the same region. However, a VPC peering connection between VPCs in different regions will not take effect.

- Network ACL

  A network ACL is an optional layer of security for your subnets. After you associate one or more subnets with a network ACL, you can control traffic in and out of the subnets. Similar to security groups, network ACLs control access to subnets and add an additional layer of security for your subnets. Security groups operate at the ECS level, whereas network ACLs operate at the subnet level. You can use network ACLs together with security groups to implement access control that is both comprehensive and fine-grained.

- Virtual IP Address

  A virtual IP address can be shared among multiple ECSs. An ECS can have both private and virtual IP addresses, and you can access the ECS through either IP address.

- Elastic Network Interface

  An elastic network interface is a virtual network card. You can create and configure network interfaces and attach them to your instances (ECSs and BMSs) to obtain flexible and highly available network configurations.

- Layer 2 Connection Gateway (L2CG)

  An L2CG is a virtual tunnel gateway that works with a Direct Connect connection to establish network communication between cloud and on-premises networks. The gateway allows you to migrate workloads in data center or private cloud to the cloud without changing subnets and IP addresses.

- IP Address Group

  An IP address group is a collection of IP addresses that use the same security group rules. You can use an IP address group to manage IP addresses that have the same

security requirements or whose security requirements change frequently. An IP address group frees you from repeatedly modifying security group rules and simplifies security group rule management. If you set the source or destination to an IP address group when you configure a security group rule. The rule takes effect for all IP addresses in the IP address group.

4.1.3 Architecture

The VPC architecture consists of VPC components, security groups, and VPC connectivity services.

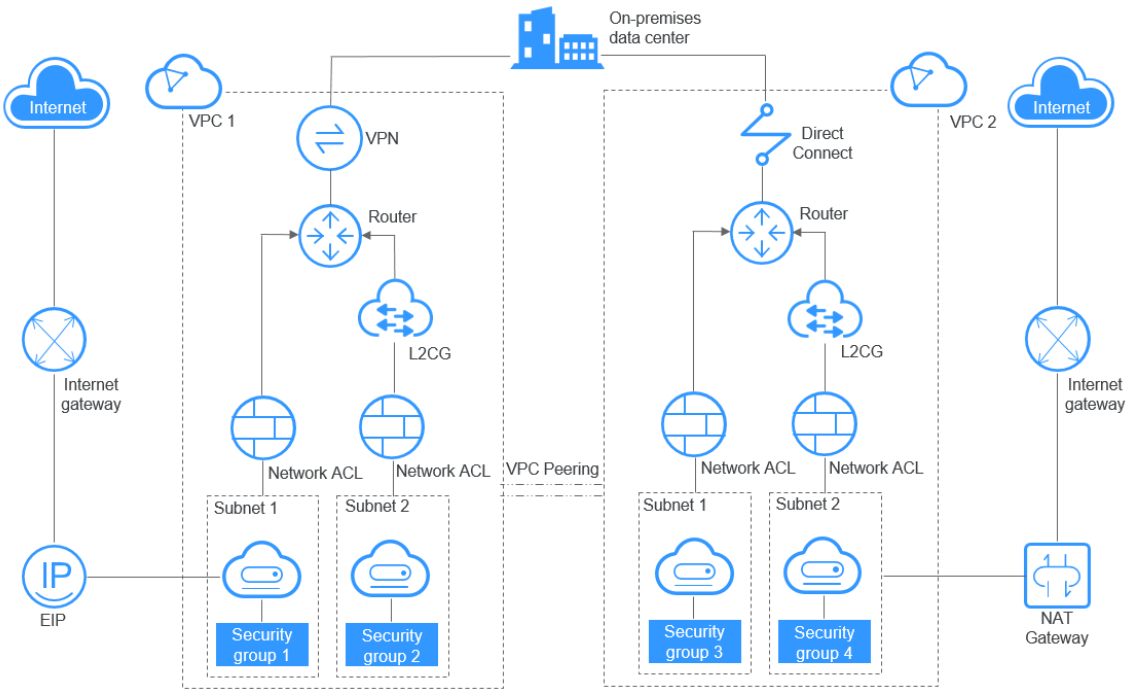


Figure 4-2 VPC architecture

- Each VPC consists of a private CIDR block, route tables, and at least one subnet.
  - CIDR block: When you create a VPC, you need to specify a private CIDR block for the VPC. The VPC service supports CIDR blocks 10.0.0.0/8-24, 172.16.0.0/12-24, and 192.168.0.0/16-24.
  - Subnet: Cloud resources, such as cloud servers and databases, must be deployed in subnets. After you create a VPC, you can divide the VPC into one or more subnets. Each subnet must be within the VPC.
  - Route table: When you create a VPC, the system automatically generates a default route table for the VPC. The route table ensures that all subnets in the VPC can communicate with each other. If the routes in the default route table cannot meet application requirements, you can create a custom route table.
- Security Group

  Security groups and network ACLs ensure the security of cloud resources deployed in a VPC. A security group acts as a virtual firewall to provide access rules for instances that have the same security requirements and are mutually trusted in a VPC. A

network ACL can be associated with subnets that have the same access control requirements. You can add inbound and outbound rules to precisely control inbound and outbound traffic at the subnet level.

- VPC Connectivity
  - Huawei Cloud provides multiple VPC connectivity options to meet diverse requirements.
  - A VPC peering connection allows two VPCs in the same region to communicate with each other using private IP addresses.
  - An EIP or a NAT gateway allows cloud servers in a VPC to communicate with the Internet.
  - VPN, Cloud Connect, Direct Connect, or L2CG can connect your on-premises data center to VPCs.

4.1.4 Advantages

Let's learn about the advantages of VPC.

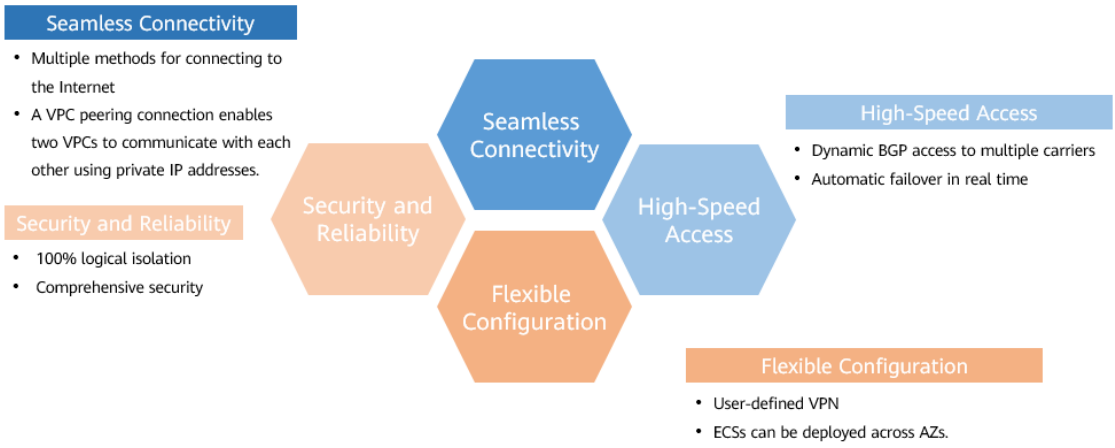


Figure 4-3 VPC advantages

As shown in the preceding figure, the VPC has the following advantages:

- Secure and reliable: VPCs are logically isolated from each other. By default, different VPCs cannot communicate with each other. Network ACLs protect subnets, whereas security groups protect ECSs.
- Flexible configuration: You can customize VPCs, divide subnets as required, and configure DHCP and route tables. ECSs can be deployed across AZs.
- High-speed access: Up to 21 dynamic BGP connections are established to multiple carriers. Dynamic BGP provides automatic failover in real time and chooses the optimal path when a network connection fails.
- Seamless Interconnection: By default, a VPC cannot communicate with the Internet. You can use EIP, ELB, NAT Gateway, VPN, and Direct Connect to enable communication with the Internet. By default, two VPCs cannot communicate with each other. You can create a VPC peering connection to enable the two VPCs in the same region to communicate with each other using private IP addresses.

4.1.5 How to Configure a VPC

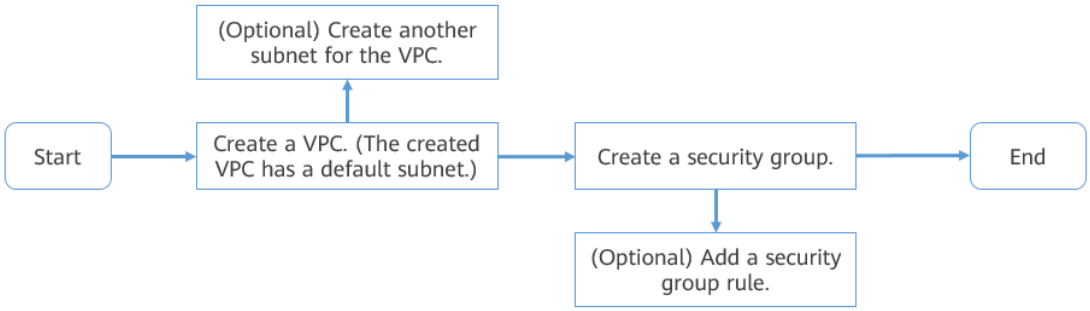


Figure 4-4 VPC configuration process

The preceding figure shows the process of creating a VPC and configuring a security group.

Each VPC comes with a default subnet. If the default subnet cannot meet your requirements, you can create one. A subnet is configured with DHCP by default. When an ECS in a subnet starts, the ECS automatically obtains an IP address using DHCP. Your account automatically comes with a default security group. You can add inbound and outbound rules to the default security group or create a security group.

- Inbound rules control incoming traffic to ECSs in the security group.
- Outbound rules control outgoing traffic from ECSs in the security group.

4.1.6 Application Scenarios

VPCs are commonly used in scenarios, such as dedicated cloud networks, web application or website hosting, and web application access control. Let's learn about these scenarios.

- Dedicated Networks on Cloud

  Each VPC represents a private network and is logically isolated from other VPCs. You can deploy your service system in a VPC to build a private network environment on the cloud. If you have multiple service systems, for example, a production system and a test system, you can deploy them in two different VPCs to isolate them. If you want to establish communication between these two VPCs, you can create a VPC peering connection between them.

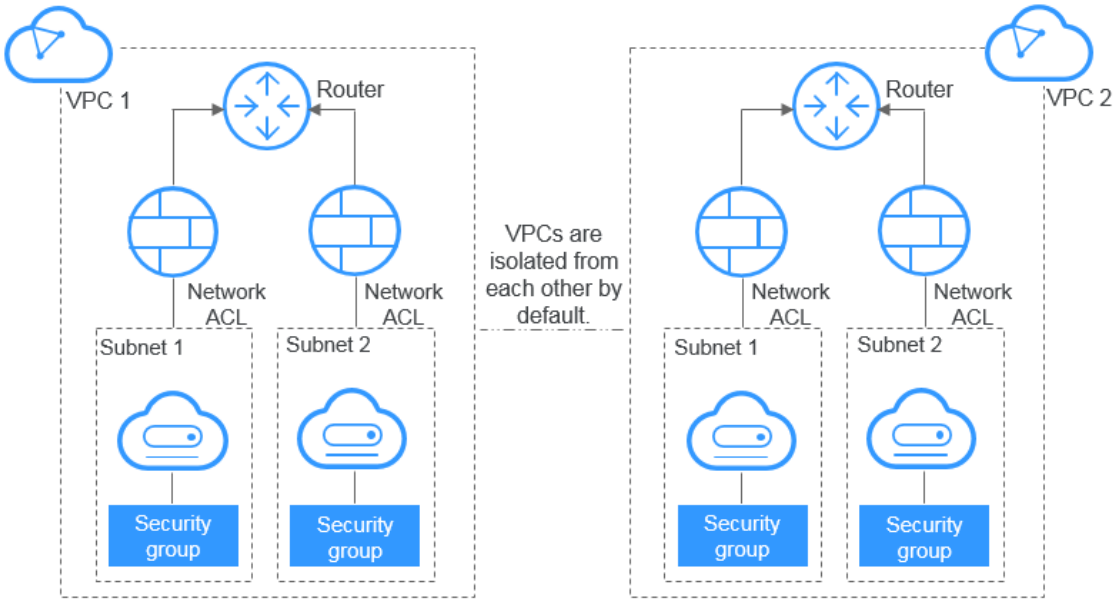


Figure 4-5 Dedicated networks on cloud

- Web Application or Website Hosting

  You can host web applications and websites in a VPC and use the VPC as a regular network. With EIPs or NAT gateways, you can connect ECSs running your web applications to the Internet. You can use load balancers provided by the ELB service to evenly distribute traffic across multiple ECSs.

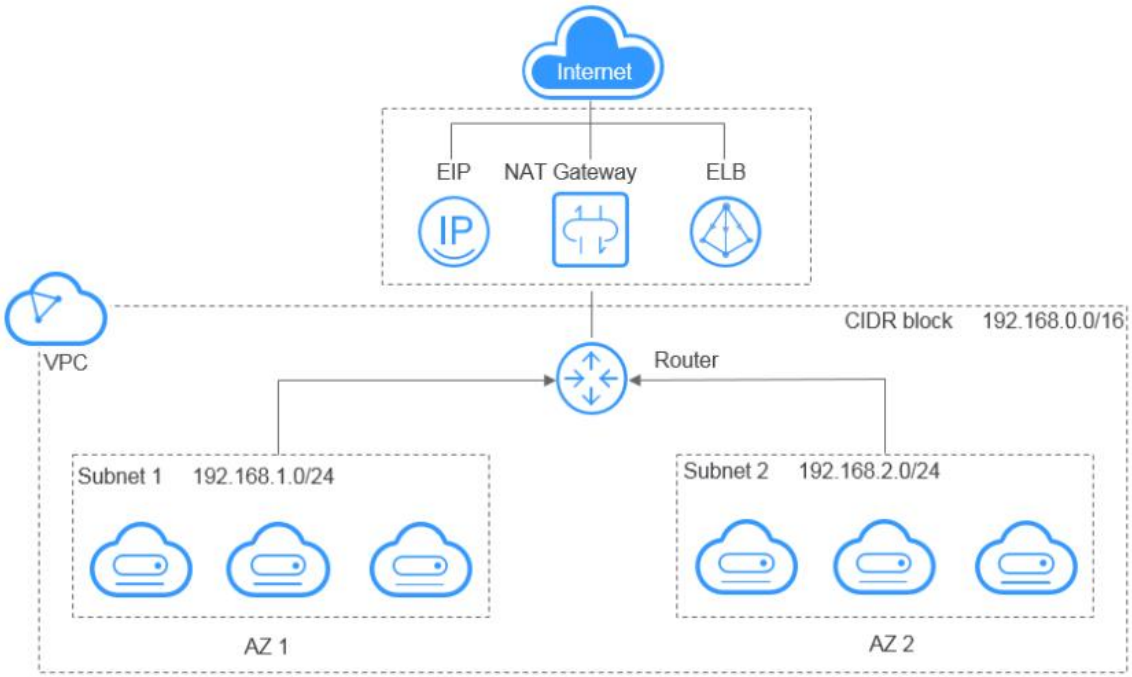


Figure 4-6 Web application or website hosting

- Web Application Access Control

You can place multi-tier web applications into different security groups, and configure rules for each security group as required. In a VPC, you can add the web servers and database servers to different security groups. You can launch web servers in a publicly accessible subnet, and also run database servers in subnets that are not publicly accessible. In this way, you can ensure high security.

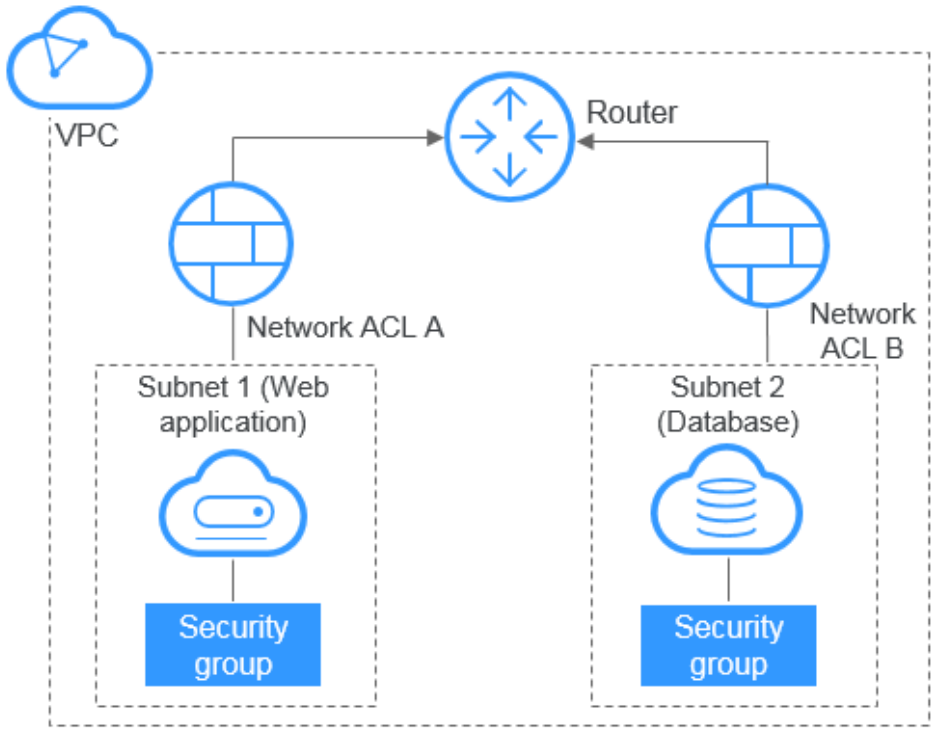


Figure 4-7 Web application access control

4.2 ELB

4.2.1 What Is ELB?

To understand a cloud service, we need to first define what it is. Elastic Load Balance (ELB) is a service that automatically distributes incoming traffic across multiple backend servers based on listening rules you configure. ELB improves the service capabilities of your applications and improves availability by eliminating single points of failure (SPOFs).

4.2.2 Architecture

Next, let's look at the ELB's architecture. ELB consists of three components: load balancers, listeners, and backend server groups.

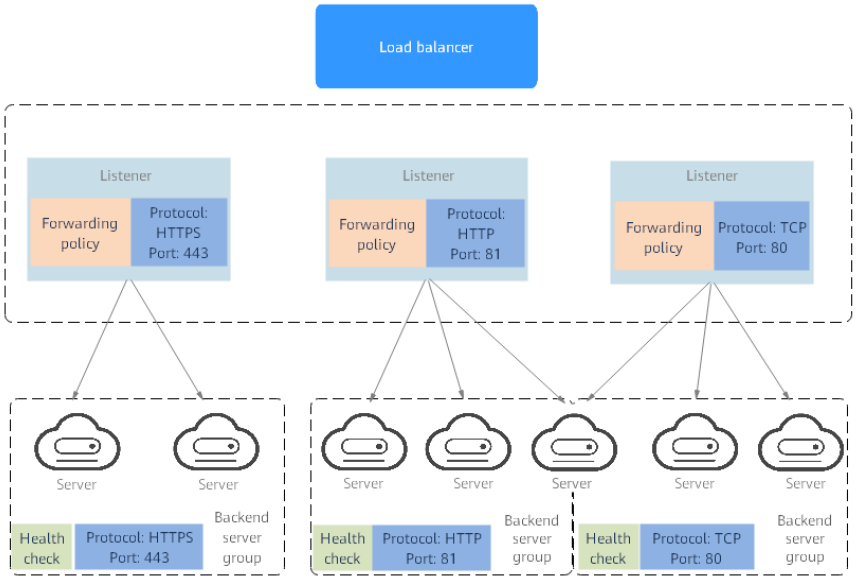


Figure 4-8 ELB architecture

In the ELB architecture:

- A load balancer is an instance that distributes incoming traffic across backend servers in one or more availability zones (AZs).
- A listener uses the protocol and port you specify to check for requests from clients and route the requests to associated backend servers based on the rules you define. You can add one or more listeners to a load balancer.
- A backend server group contains one or more backend servers, which use the protocol and port you specify to receive the requests routed to them by the listener. Each backend server group is associated with a listener. You can set a weight for each backend server and can also configure health checks to check the health of each backend server. If an exception occurs on a backend server, the load balancer automatically distributes new requests to other healthy backend servers. When the backend server recovers, traffic routing to this server will automatically resume.

4.2.3 Advantages

Figure 4-9 ELB Advantages

ELB has the following advantages:

- Robust performance: ELB is deployed in clusters, which can establish up to 100 million concurrent connections so that your applications can handle a massive volume of concurrent requests.
- High availability: ELB can distribute incoming traffic across AZs to ensure your services are never uninterrupted. If servers in one AZ are unhealthy, ELB automatically routes traffic to healthy servers in other AZs.
- Excellent scalability: ELB automatically scales in line with spikes in incoming traffic to ensure that your applications always stay online. It works with Auto Scaling (AS) to flexibly adjust the number of servers and intelligently distribute incoming traffic across them.
- Ease-of-use: ELB provides a diverse set of protocols and algorithms that you can use to configure traffic routing policies to meet your requirements while keeping the configuration simple.

4.2.4 How to Configure ELB

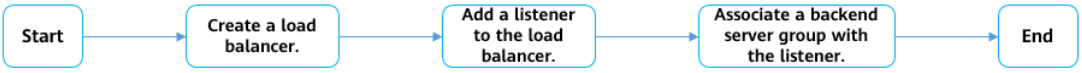


Figure 4-10 Configuring ELB

Configuring ELB includes three main steps.

Step 1 Create a load balancer.

You need to plan the region, type, protocol, and backend servers for the load balancer based on service requirements.

Step 2 Add at least one listener to the load balancer.

The listener listens for requests from clients and routes the requests to backend servers based on the settings that you configure when you add the listener.

Step 3 Associate a backend server group with each listener.

A backend server group is a collection of cloud servers that have the same features and receive the requests routed by the load balancer.

4.2.5 How to Use ELB

When you use ELB, note the following:

- You can delete a load balancer if you do not need it any longer. Deleted load balancers cannot be recovered.
- After a public network load balancer is deleted, its EIP will not be released. The EIP is made available for use by other resources.
- You can modify a listener as needed or delete a listener if you no longer need it. Deleted listeners cannot be recovered.

- After you disassociate a backend server from a load balancer, the backend server will no longer receive requests from the load balancer, but the server will not be deleted. You can associate it with the load balancer again when your business grows or reliability needs to be enhanced.

4.2.6 Application Scenarios

When would I need ELB? Well, let's look at some common application scenarios of this service.

- Heavy-Traffic Applications

  For an application with heavy traffic, such as a large web portal or mobile app store, ELB evenly distributes incoming traffic to multiple backend servers, balancing the load to ensure steady performance. Sticky sessions ensure that requests from a given client are forwarded to the same backend server, improving access efficiency.

- Applications with Predictable Peaks and Troughs in Traffic

  For an application that has predictable peaks and troughs in traffic volumes, ELB works with AS to add or remove backend servers to keep up with changing demands. An example is flash sales, during which application traffic spikes in a short period. Running only the required number of backend servers to handle the load of your application helps reduce costs.

- Zero SPOFs

  ELB routinely performs health checks on backend servers to monitor their health. If any unhealthy backend servers are identified, ELB will not route requests to these servers until they recover. This makes ELB a good choice for running services that require high reliability.

- Cross-AZ Load Balancing

  ELB can distribute traffic across AZs. When an AZ becomes faulty, ELB distributes traffic across healthy backend servers in other AZs. ELB is ideal for banking, policing, and large application systems that require high availability.

4.3 VPN

4.3.1 What Is VPN?

A Virtual Private Network (VPN) allows you to establish an encrypted, Internet-based communications tunnel between your on-premises network and a VPC, so you can access resources in the VPC remotely.

By default, ECSs in a VPC cannot communicate with your data center or private network. To enable communications between them, use a VPN.

4.3.2 Architecture

A VPN consists of two parts.

- VPN gateway

  A VPN gateway is an egress gateway for a VPC. With a VPN gateway, you can create a secure, reliable, and encrypted connection between a VPC and an on-premises data center or between two VPCs in different regions. Each data center needs a local gateway and remote gateway. The local gateway needs to be paired with a remote gateway, but each VPN local gateway can connect to multiple remote gateways, so you can set up point-to-point or hub-and-spoke VPN connections.

- VPN connection

  A VPN connection uses IPsec encryption to establish a secure and reliable communications tunnel between a VPN gateway and the gateway in an on-premises data center. Only IPsec VPN connections are supported. VPN connections use Internet Key Exchange (IKE) and IPsec protocols to encrypt and transmit data over the Internet. VPN is more cost-effective than other connection options like Direct Connect.

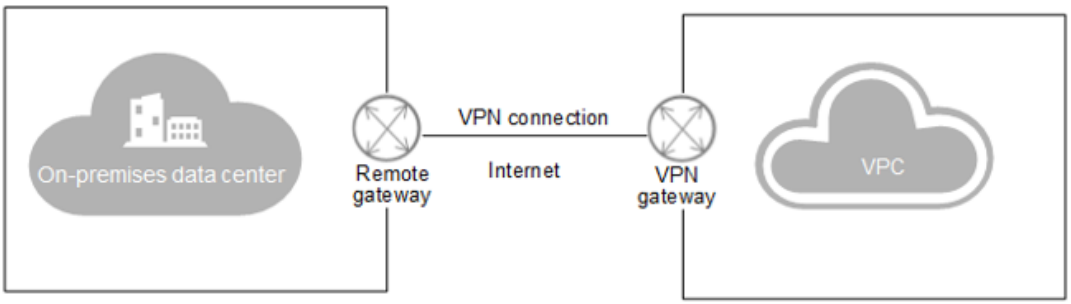


Figure 4-11 VPN architecture

4.3.3 Advantages

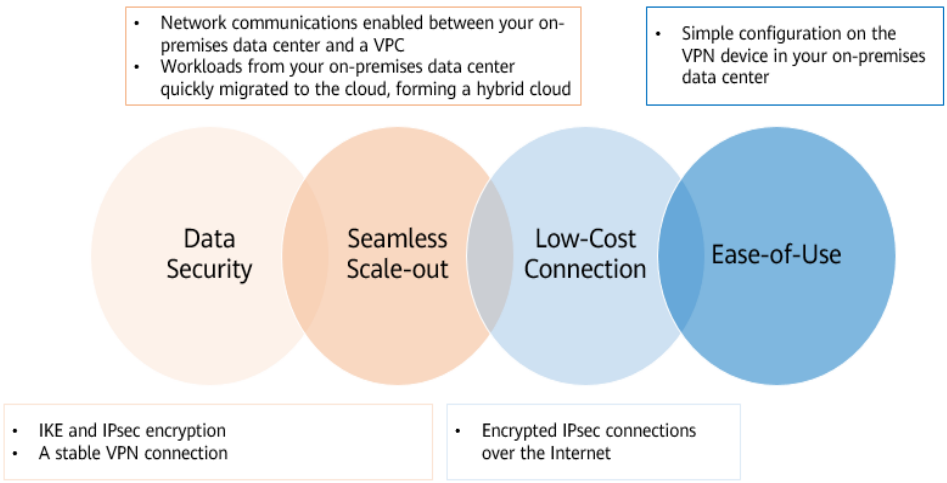


Figure 4-12 VPN advantages

VPN has the following advantages:

- Data security

  Huawei hardware uses IKE and IPsec to encrypt data to provide carrier-class reliability and ensure a stable VPN connection.

- Seamless scale-out

  With VPN, you can connect your on-premises data center to your VPC and quickly extend services from the data center to the cloud, forming a hybrid cloud.

- Low-cost connection

  Encrypted IPsec connections over the Internet provide a cost-effective alternative to Direct Connect connections.

- Ease-of-use

  Creating a VPN connection is easy. Just specifying a few parameters on the VPN console and configuring the VPN device in your on-premises data center.

4.3.4 Application Scenarios

With a VPN connecting a VPC to your on-premises data center, you can easily use cloud servers and storage resources deployed in the VPC. Your on-premises applications can be migrated to the cloud while your core data is retained in your on-premises data center. Your web servers can be scaled as required to meet ever-changing computing requirements while your IT O&M costs can be greatly reduced. The VPN service allows you to set up both site-to-site VPN connections and hub-and-spoke VPN connections.

- Site-to-site VPN connections

  You can use VPN to establish a hybrid cloud by connecting an on-premises data center to a VPC.

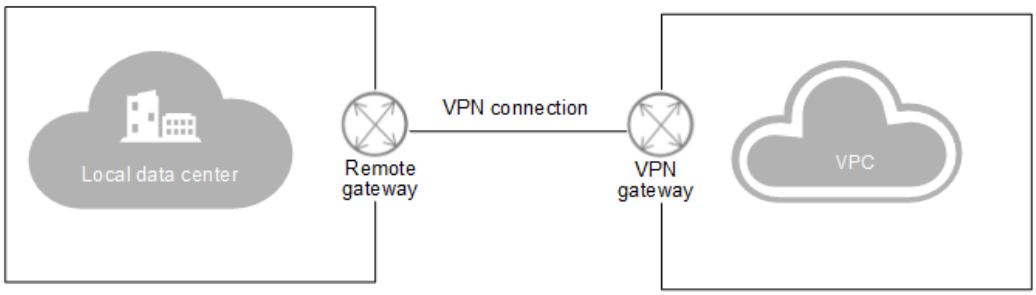


Figure 4-13 Site-to-site VPN connection

- Hub-and-spoke VPN connections

  You can also use VPN to establish a hybrid cloud by connecting multiple data centers to a VPC.

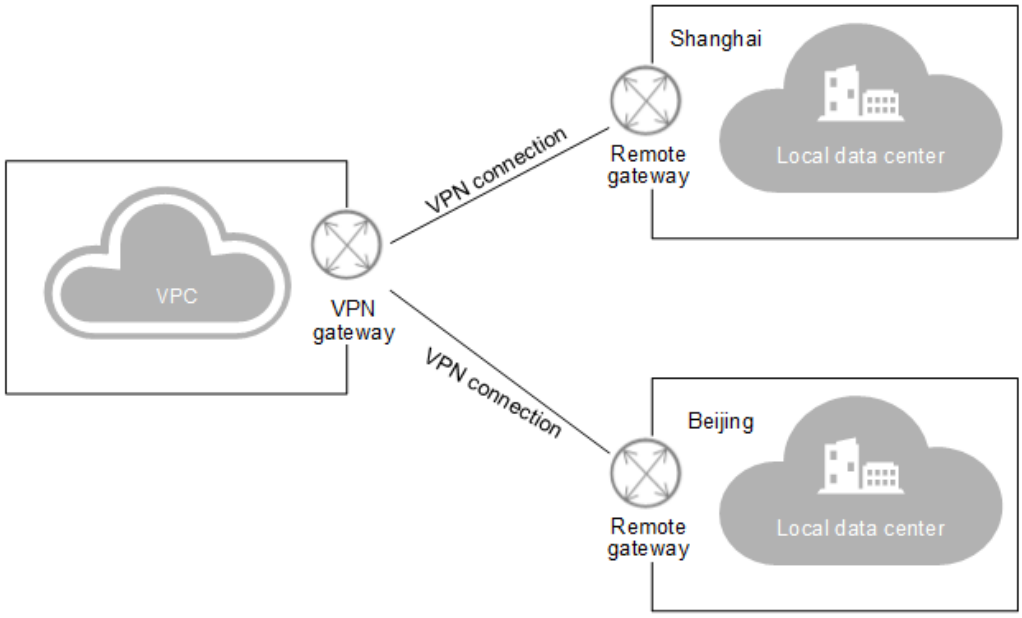


Figure 4-14 Hub-and-spoke VPN connections

4.3.5 How to Configure a VPN

The following figure shows the VPN configuration process. You can create a VPN gateway and a VPN connection on the management console.

Figure 4-15 VPN configuration process

To enable your ECSs in a VPC to access your on-premises network, you must first create a VPN gateway.

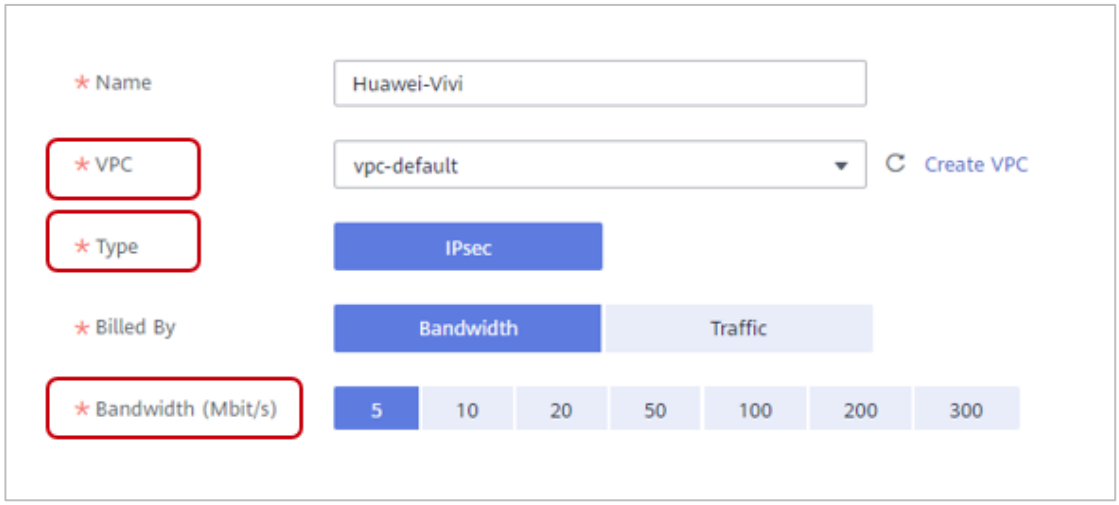


Figure 4-16 Buying a VPN gateway

The parameters in the preceding figure are as follows:

- **VPC**: the VPC that the on-premises network connects to via VPN
- **Type**: the VPN type. **IPsec** is selected by default.
- **Billed By**: There are two options available, bandwidth, and traffic.
  - **Bandwidth**: You specify a bandwidth and are charged based on the amount of time you use the bandwidth.
  - **Traffic**: You specify a bandwidth and pay for the total traffic you generate.
- **Bandwidth (Mbit/s)**: The bandwidth of the VPN gateway. The bandwidth is shared by all VPN connections created for the VPN gateway. The total bandwidth used by all VPN connections created for a VPN gateway cannot exceed the VPN gateway bandwidth.

If the network traffic exceeds the VPN gateway bandwidth, the network may get congested and VPN connections may be interrupted. Make sure you configure enough bandwidth. You can configure alarm rules on Cloud Eye to monitor the bandwidth.

To connect your ECSs in a VPC to your on-premises network, after the VPN gateway is created, you also need to create a VPN connection.

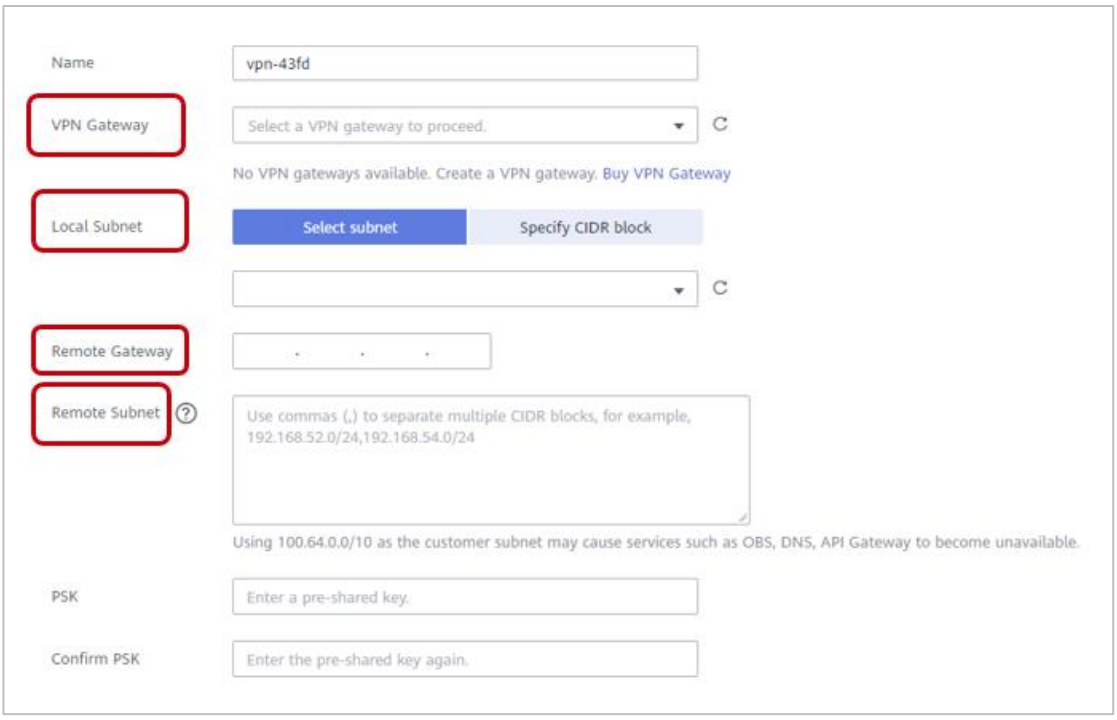


Figure 4-17 Creating a VPN connection

The parameters in the preceding figure are as follows:

- **VPN Gateway**: the name of the VPN gateway used by the VPN connection
- **Local Subnet**: the VPC subnets that will access your on-premises network through VPN. Possible values are **Select subnet** and **Specify CIDR block**.
- **Remote Gateway**: the public IP address of the VPN device translated by the VPN gateway in your on-premises network. This IP address is used for communications with your VPC.

- **Remote Subnet**: the subnets of your on-premises network that will access the VPC through a VPN. The remote subnet can include the CIDR block of the local subnet. The local subnet cannot include the CIDR block of the remote subnet.
- **PSK**: Enter 6 to 128 characters. The PSK at both ends of a VPN connection must be the same.

4.3.6 How to Use a VPN

When using the VPN, note that:

- A VPN connection is an encrypted communications channel established between the VPN gateway in your VPC and that in your on-premises data center. You can modify a VPN connection as required. You can delete a VPN connection to release network resources if it is no longer required. When you delete the last VPN connection for a pay-per-use VPN gateway, the associated gateway will be deleted along with it.
- You can modify the name and description of a VPN gateway if needed. If the bandwidth of a VPN gateway cannot meet your requirements, you can modify the bandwidth, too. If the number of VPN connections associated with a VPN gateway cannot meet your requirements, you can modify the VPN gateway specifications. You can change the billing mode of a VPN gateway billed by bandwidth from pay-per-use to yearly/monthly. If a VPN gateway is no longer required, you can delete it to release network resources as long as it has no VPN connections configured. A VPN gateway cannot be deleted if it is being used by a VPN connection. Delete the VPN connection before deleting the VPN gateway.

4.4 NAT Gateway

4.4.1 What Is NAT Gateway?

NAT Gateway provides network address translation (NAT). There are public NAT gateways and private NAT gateways.

A public NAT gateway provides both source NAT (SNAT) and destination NAT (DNAT) for cloud and on-premises servers and allows those servers to share EIPs to communicate with the Internet.

A private NAT gateway provides NAT for servers in a VPC, so that multiple servers can share a transit IP address to access or provide services accessible from an on-premises data center or other VPCs.

4.4.2 Advantages

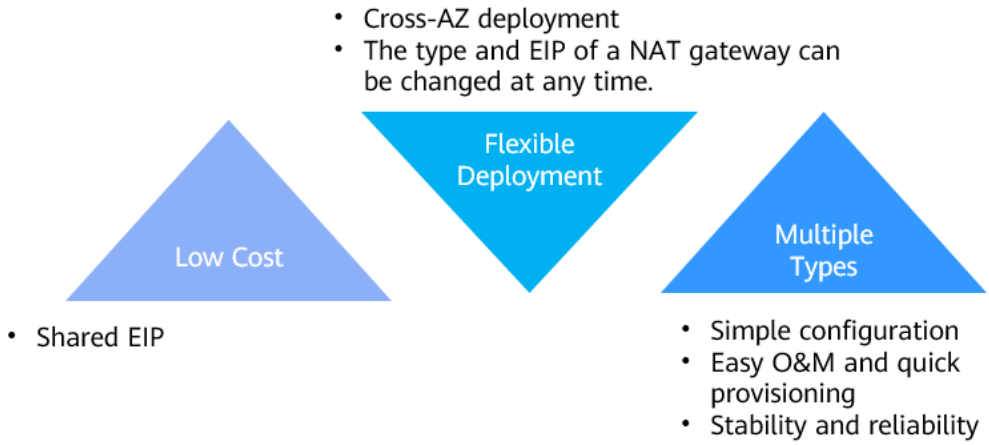


Figure 4-18 NAT Gateway advantages

NAT Gateway has the following advantages:

- Flexible deployment

  A NAT gateway can be shared across subnets and AZs. A fault in a single AZ does not affect the service continuity of a NAT gateway. The type and EIP of a NAT gateway can be changed at any time.

- Ease of use

  Multiple types of NAT gateways are available. NAT gateway configuration is simple, the O&M is easy, and they can be provisioned quickly. Once provisioned, they are stable and reliable.

- Cost-effectiveness

  When you use a NAT gateway to send data from a private IP address or your applications provide services accessible from the Internet, the NAT gateway translates the private IP address to a public IP address. NAT Gateway helps you save money on EIPs and bandwidth.

4.4.3 How to Configure a NAT Gateway

Figure 4-19 NAT gateway configuration process

The preceding figure shows how to configure a NAT gateway.

When you buy a public NAT gateway, you must specify its VPC, subnet, and type. Check whether the default route (0.0.0.0/0) of the VPC is in use by any other gateways. If yes, add another route for this gateway or add the default route to a new route table to be associated with this gateway after this gateway is created.

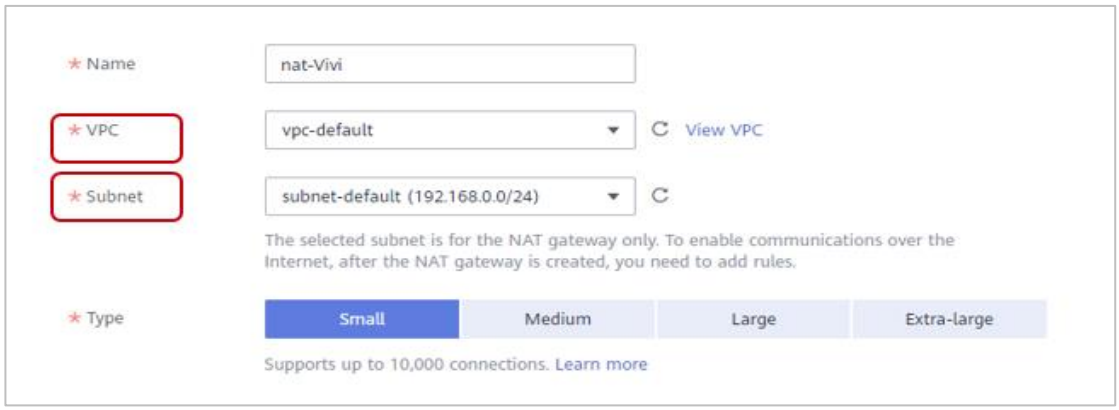


Figure 4-20 Buying a NAT gateway

The parameters in the preceding figure are as follows:

Subnet: the subnet where the public NAT gateway is deployed. The subnet must have at least one available IP address. The selected subnet cannot be changed after the public NAT gateway is created.

Type: The type can be **Small**, **Medium**, **Large**, and **Extra-large**. You can click **Learn more** on the page to view details about each type.

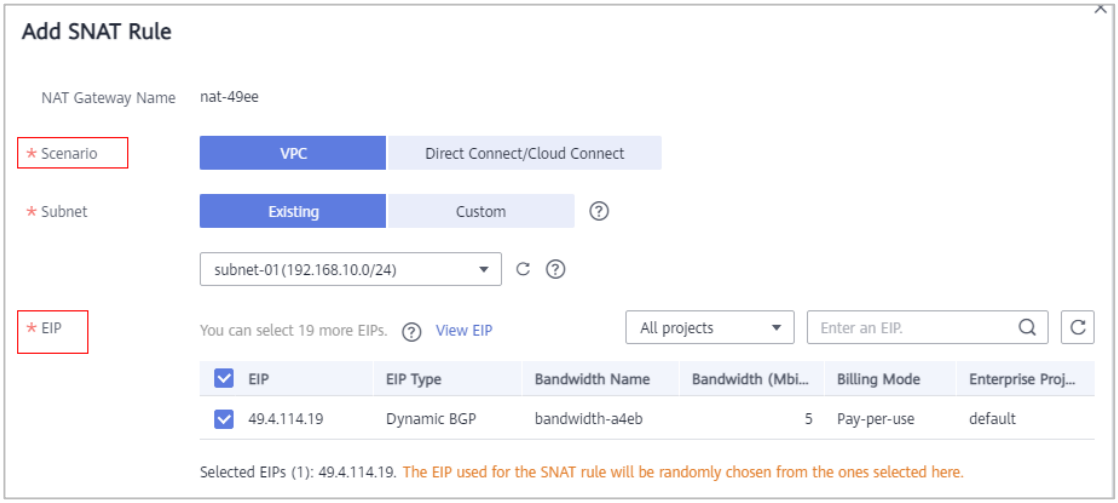


Figure 4-21 Adding an SNAT rule

After the public NAT gateway is created, add SNAT rules to enable your cloud or on-premises servers to share an EIP for access to the Internet.

Each SNAT rule is configured for one subnet. If there are multiple subnets in a VPC, you can create multiple SNAT rules.

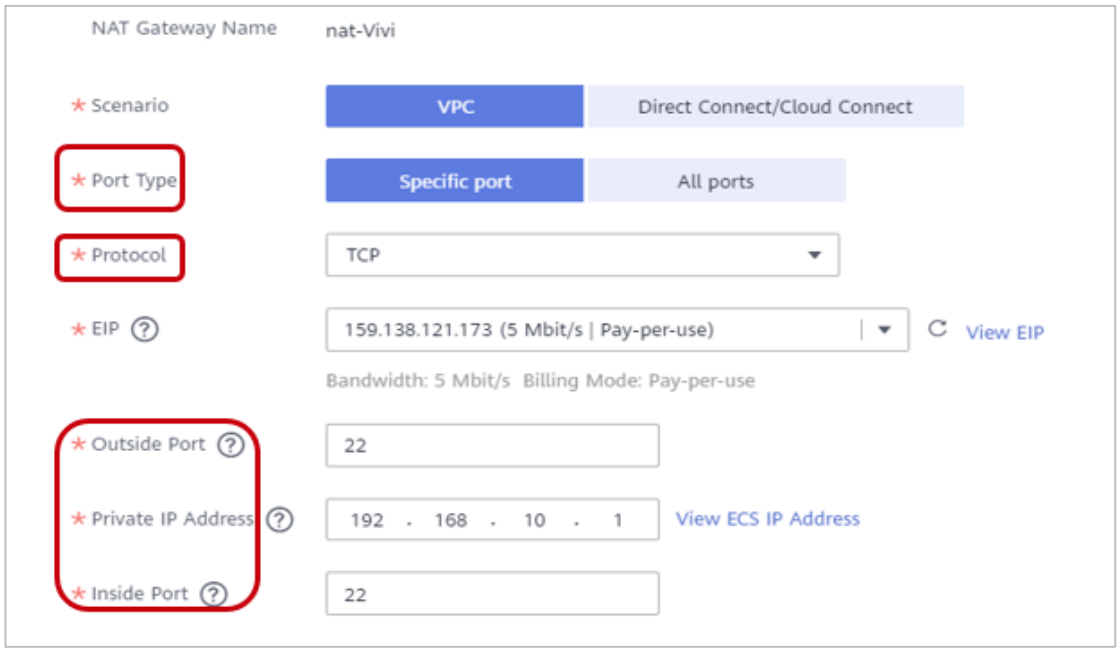


Figure 4-22 Adding a DNAT rule

After a public NAT gateway is created, you can add DNAT rules to allow servers in your VPC to provide services accessible from the Internet. You can configure a DNAT rule for each port on a server. If multiple servers need to provide services accessible from the Internet, create multiple DNAT rules.

4.4.4 Application Scenarios

Here are four application scenarios of NAT Gateway.

- Public NAT Gateway - Configuring SNAT Rules to Enable Servers to Access the Internet

  If you have servers in a VPC that requires Internet access, you can configure an SNAT rule to let the servers share one or more EIPs to access the Internet without exposing their IP addresses. You configure one SNAT rule for each subnet in a VPC and configure one or more EIPs for each SNAT rule. NAT gateways of different types support different numbers of connections. You can create multiple SNAT rules to meet different service requirements.

- Public NAT Gateway - Configuring DNAT Rules to Enable Servers to Provide Services Accessible from the Internet

  DNAT rules let servers in a VPC to provide services accessible from the Internet. You can configure an EIP for a DNAT rule. Then NAT gateways can forward requests from a specific port and over a specific protocol to the EIP first, and then to the specified port or ports of your server, or forwards all requests to your server through the EIP. NAT Gateway allows multiple servers to share an EIP, saving costs on bandwidth. A DNAT rule is configured for one server. If there are multiple servers, you can create multiple DNAT rules.

- Public NAT Gateway - Configuring SNAT or DNAT Rules to Communicate with the Internet at a High Speed

If on-premises servers that connect to a VPC through Direct Connect or VPN need secure, high-speed Internet access, SNAT and DNAT make it possible.

- Public NAT Gateway - Setting Up a Highly Available System by Configuring Multiple EIPs for an SNAT Rule

  EIPs bound to servers or other resources are vulnerable to online attacks. To improve system reliability, you can configure multiple EIPs for the same SNAT rule so that if one EIP is attacked, another one can take over to ensure service continuity. If an SNAT rule has multiple EIPs, the system randomly selects one EIP for servers using the SNAT rule to access the Internet. A maximum of 20 EIPs can be added to each SNAT rule. If an EIP added to an SNAT rule is blocked or unavailable, manually delete it from the EIP list.

4.4.5 Precautions

After learning the definition, advantages, configuration process, and application scenarios of NAT Gateway, we also need to pay attention to the following items when using NAT gateways:

- You can modify the name, type, or description of a public NAT gateway.
- Using a larger NAT gateway than you need does not affect services, but if you switch to a smaller gateway, make sure that the reduced capacity is still enough to meet your needs.
- You can delete NAT gateways that are no longer required to release resources and reduce costs.
- Before deleting a NAT gateway, ensure that all SNAT and DNAT rules of the NAT gateway have been deleted.

4.5 Other Network Services

In the cloud marketplace, there are many other network cloud services. In addition to VPC, ELB, VPN, and NAT Gateway mentioned previously, I'd like to briefly introduce some other networking services.

Domain Name Service (DNS) provides highly available and scalable authoritative DNS services that translate domain names (such as www.example.com) into IP addresses (such as 192.1.2.3) required for network connection. DNS is what allows users to visit your websites or web applications using domain names. The DNS service is free and is enabled by default.

The DNS service provides the following four functions:

- Public domain name resolution

  Maps domain names to public IP addresses so that your users can access your website or web applications over the Internet.

- Private domain name resolution

  Translates private domain names into private IP addresses to facilitate access to cloud resources within VPCs.

- Reverse resolution

  Obtains a domain name based on an IP address. Reverse resolution, or reverse DNS lookup, is typically used to affirm the credibility of email servers.

- Intelligent resolution

  Returns different resolution results for the same domain name based on the carrier networks or geographic locations of user IP addresses. This significantly reduces network latency for users from different carrier networks and geographic locations.

When you use the DNS service, ensure that the domain name format meets the following requirements:

- The different parts of a domain name are separated using periods (.).
- Each part of a domain name can contain only supported language-specific characters, letters, digits, and hyphens (-) and cannot start or end with a hyphen.
- Each part of a domain name can exceed 63 characters.
- The total length of a domain name, including the period at the end, cannot exceed 254 characters.

A domain name structure is divided into the following levels:

- Root domain: **.** (a period)
- Top-level domain: for example, .com, .net, .org, and .cn
- Second-level domain: subdomains of the top-level domain names, such as example.com, example.net, and example.org
- Third-level domain: subdomains of the second-level domain names, such as abc.example.com, abc.example.net, and abc.example.org
- The next-level domain names are similarly expanded by adding prefixes to the previous-level domain names, such as def.abc.example.com, def.abc.example.net, and def.abc.example.org.

5 Storage Cloud Services

Data is everywhere. We use USB flash drives and cloud disks to store data, and these devices are called storage devices. That is enough for most of us, but what do you use for enterprise storage? In today's age of cloud computing, what are the most common storage cloud services?

In this section, we will cover some common storage services on Huawei Cloud.

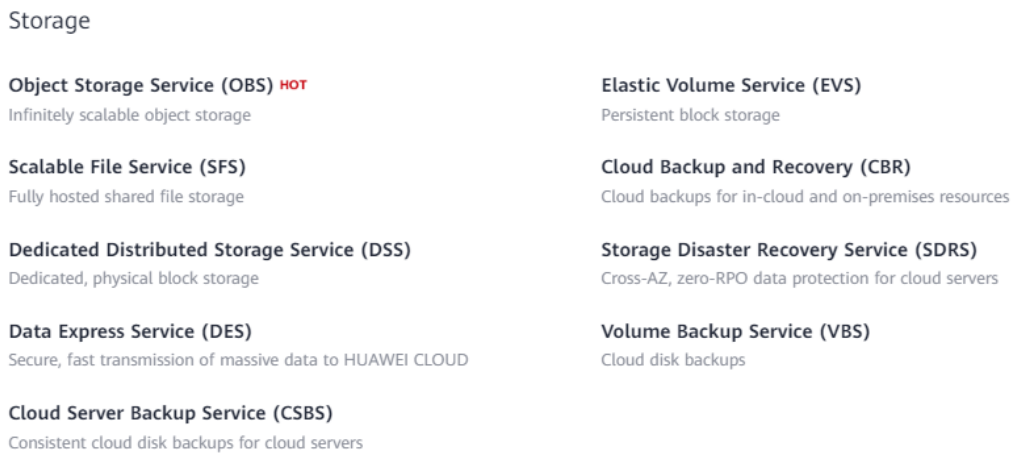


Figure 5-1 Storage service overview

5.1 EVS

5.1.1 What Is EVS?

Elastic Volume Service (EVS) offers scalable block storage for cloud servers such as Elastic Cloud Servers (ECSs) and Bare Metal Servers (BMSs). EVS disks offer high reliability and excellent performance. They can be used for distributed file systems, development and testing environments, data warehouse applications, and high-performance computing (HPC).

5.1.2 Architecture

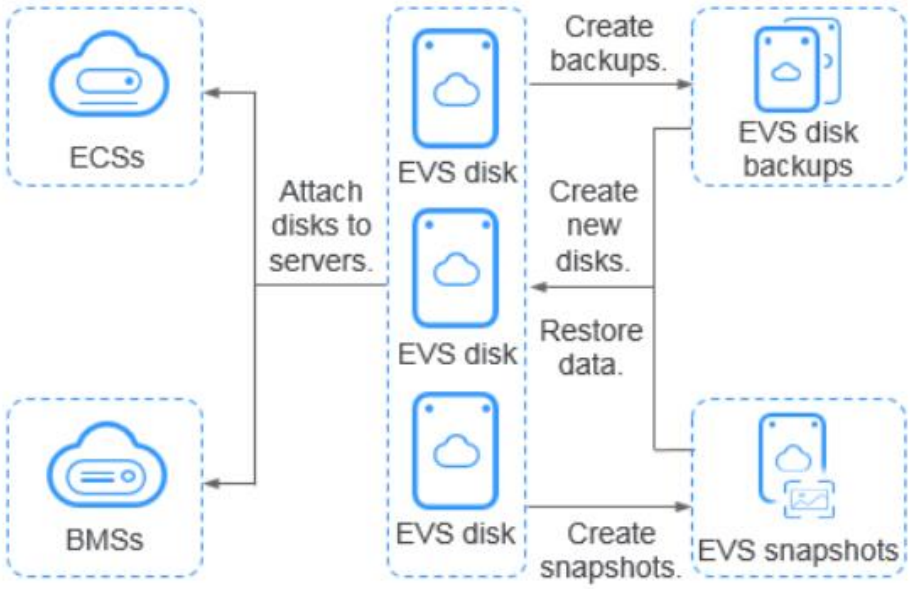


Figure 5-2 EVS architecture

EVS disks are like the hard disks on your local computer, except on the cloud. They need to be attached to cloud servers before you can use them. You can initialize EVS disks, create file systems, and then use them for persistent data storage. Alternatively, you can create backups and snapshots for your EVS disks to improve data reliability.

5.1.3 Advantages

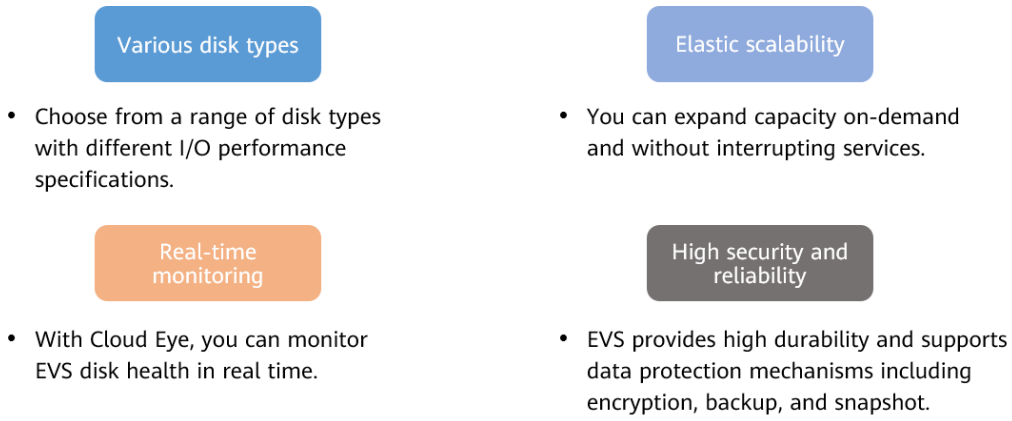


Figure 5-3 EVS advantages

EVS has the following advantages:

- Various disk types: EVS disks come in various specifications and can be attached to ECSs as data disks or system disks. You can select EVS disks based on your budget and service requirements.

- Elastic scalability: The size of a single EVS disk can be anything from 10 GB to 32 TB. As services migrate to the cloud, if a disk no longer meets you needs, you can expand disk capacity in increments as small as 1 GB without stopping services. Aside from the limit on individual disk capacity, space can also be limited by quotas.
- High security and reliability: Both system disks and data disks can be encrypted. Stored data can also be protected by backups and snapshots, which can help you restore disk data damaged or destroyed by a software exception or hacker attack.
- Real-time monitoring: On Cloud Eye, you can monitor the disk health and operating status at any time, and make changes accordingly.

5.1.4 Disk Types and Performance

On Huawei Cloud, EVS disks are classified as extreme SSD, ultra-high I/O, general purpose SSD, high I/O, or common I/O types, each type offering different performance characteristics. EVS disks differ in performance and price. Choose the disk type most appropriate for your applications. The brand new extreme SSD EVS disks use the congestion control algorithms for RDMA deployments, with the maximum throughput of a single disk reaching up to 1,000 MB/s, and extreme low single-channel latency.

EVS performance metrics include:

- IOPS: The number of read/write operations that can be performed by an EVS disk per second
- Throughput: The amount of data that can be read from and written into an EVS disk per second
- Read/write I/O latency: The minimum interval between two consecutive read/write operations on an EVS disk

EVS disk performance is closely related to data block size. An EVS disk gets its maximum performance by either reaching the maximum IOPS or maximum throughput, but it cannot reach both. EVS disks can get the maximum IOPS with 4 KB or 8 KB data blocks, and they can get the maximum throughput with data blocks of at least 16 KB.

5.1.5 Application Scenarios

Different types of EVS disks are recommended for different application scenarios. Figure 5-4 shows the details.

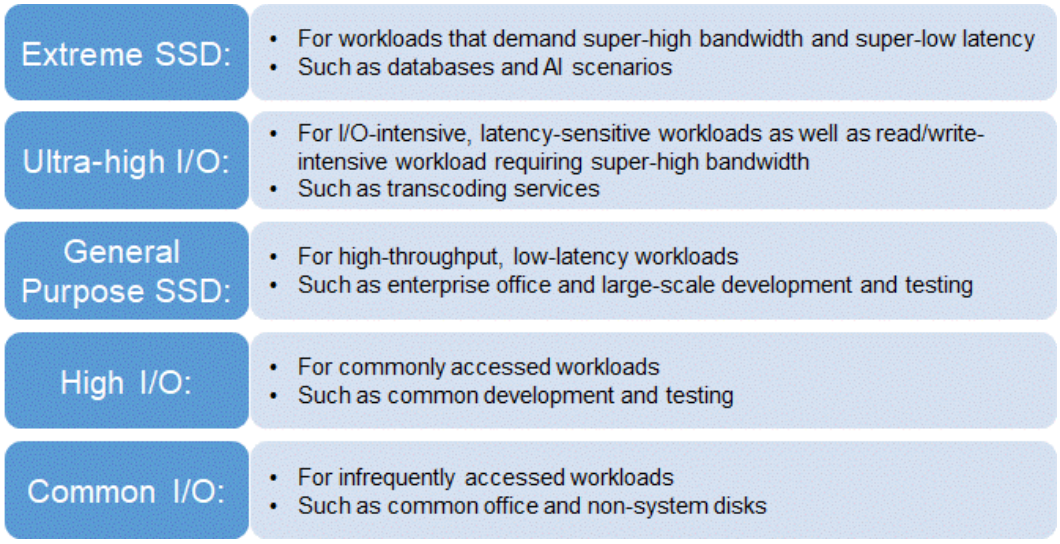


Figure 5-4 EVS application scenarios

Extreme SSD disks are suitable for workloads that demand super-high bandwidth and super-low latency.

- High-performance databases
  - Oracle
  - SQL Server
  - ClickHouse
- AI scenarios

  Ultra-high I/O disks are high-performance disks. They are excellent for enterprise mission-critical services as well as workloads demanding high throughput and low latency.
- Read/write-intensive applications that require ultra-large bandwidth
- Transcoding services
- I/O-intensive scenarios
  - NoSQL
  - Oracle
  - SQL Server
  - PostgreSQL
- Latency-sensitive scenarios
  - Redis
  - Memcache

  General purpose SSD disks are a cost-effective option. They are good for enterprise office applications, such as mainstream high-performance, low-latency interactive applications.
- Enterprise office applications
- Large-scale development and testing

- Transcoding services
- Web server logging
- Container system disks

  High I/O disks are suitable for commonly accessed workloads.
- Common development and testing

  Common I/O disks (previous-generation product) were used for infrequently accessed workloads. They are suitable for applications that required large capacity, medium read/write speed, but with fewer transactions.
- Common office applications
- Lightweight development and testing
- Non-system disks

5.1.6 Device Types

There are two EVS device types: Virtual Block Device (VBD) and Small Computer System Interface (SCSI). The following is a brief introduction to the two types:

- VBD is the default EVS device type. VBD EVS disks support only basic read/write SCSI commands.
- SCSI EVS disks support transparent commission of SCSI commands and allow the server OS to directly access the underlying storage media. In addition to basic read/write SCSI commands, SCSI EVS disks support advanced SCSI commands.

The device type selected during purchase cannot be changed later.

Usage Instructions

BMSs only support SCSI EVS disks.

Shared EVS disks must be used together with a distributed file system or clustered software. Because most clustered applications, such as Windows MSCS, Veritas VCS, and Veritas CFS, use SCSI reservations, you are advised to use SCSI for shared EVS disks. SCSI reservations take effect only when a shared SCSI EVS disk is attached to ECSs in the same ECS group with an enabled anti-affinity policy.

5.1.7 Major Features

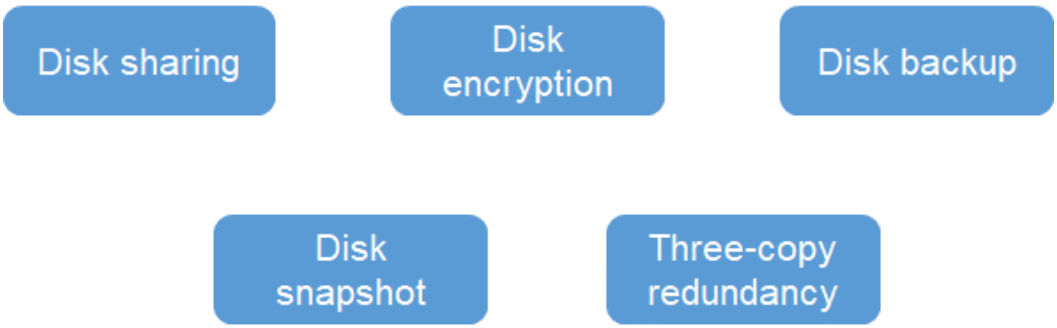


Figure 5-5 EVS feature overview

- Disk sharing

  A shared EVS disk can be attached to multiple ECSs or BMSs, and supports concurrent access. Shared EVS disks feature multiple attachments, high-concurrency, high-performance, and high-reliability. They are often used for mission-critical applications that require cluster deployment for high availability (HA).

  Precautions for using shared EVS disks:

  Before you use shared EVS disks, you must set up a shared file system or cluster management system. If you directly attach a shared EVS disk to multiple servers, data on the servers cannot be shared and may be overwritten. A shared EVS disk can be attached to a maximum of 16 servers.

- Disk encryption

  EVS disks can be encrypted in case your services require extra security. The security administrator can grant Key Management Service (KMS) permissions for EVS. These permissions allow you to use disk encryption. The system creates a default master key, which, if you have the permissions, you can use to encrypt EVS disks.

- Disk backup

  CBR cloud disk backup allows you to create backups of EVS disks to safeguard the data stored on them. You can back up EVS disks on the console without stopping the cloud servers. If anything happens to an EVS disk, you can restore the disk data to any point in the past when the backup was created. Disk backups help ensure the integrity and security of your data. After a backup policy is applied, the EVS disk data is automatically backed up based on the policy. You can use the backups as a baseline to create new EVS disks or to restore to original EVS disks.

- Disk snapshot

  An EVS snapshot is a complete copy or image of the disk data taken at a specific point in time for data disaster recovery. If data is lost, you can completely restore the disk data to the state from when the snapshot was taken. **The differences between backups and snapshots are as follows:**

  - Both snapshots and backups provide redundancy for the EVS disk data, but they are stored differently.
  - Snapshots are stored together with the EVS disk data, which facilitates fast backup and recovery. Backups are stored in OBS so that data can be restored even if the EVS disk itself is damaged.
  - Auto snapshots are currently not supported, but auto backups are. After you apply a backup policy, the system will automatically back up the disk data accordingly.

- Three-copy redundancy

  Three-copy redundancy means that there are three copies of each piece of data, with each copy stored on a different node in the storage system. The backend storage system of EVS uses three-copy redundancy to guarantee data reliability.

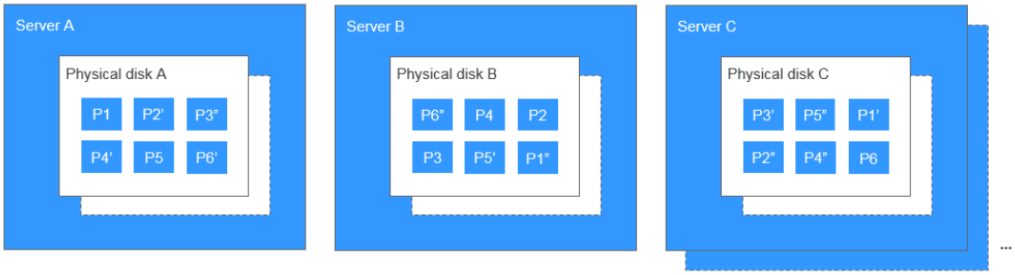


Figure 5-6 Three-copy redundancy

As shown in Figure 5-6, the storage system guarantees strong consistency among data copies using three-copy redundancy. For example, the storage system backs up data block P1:

- on physical disk A of server A as P1
- on physical disk B of server B as P1"
- on physical disk C of server C as P'

Data blocks P1, P1', and P1'' are all copies of the same data block. If physical disk A where P1 resides should fail for some reason, P1' and P1'' are still available to ensure service continuity.

- Data rebuild

  On Huawei Cloud, if a physical server or disk fault is detected, the storage system automatically rebuilds the data.

  Each physical disk in the storage system stores multiple data blocks, whose copies are distributed on cluster nodes according to certain rules. If a physical server or disk fault is detected, the storage system automatically rebuilds the data. Since data copies are distributed on different storage nodes, data can be rebuilt on different nodes at the same time and each node has only a small amount of data rebuilt. This prevents performance deterioration caused by restoration of a large amount of data on a single node, and minimizes impacts on upper-layer services.

5.1.8 How to Use EVS

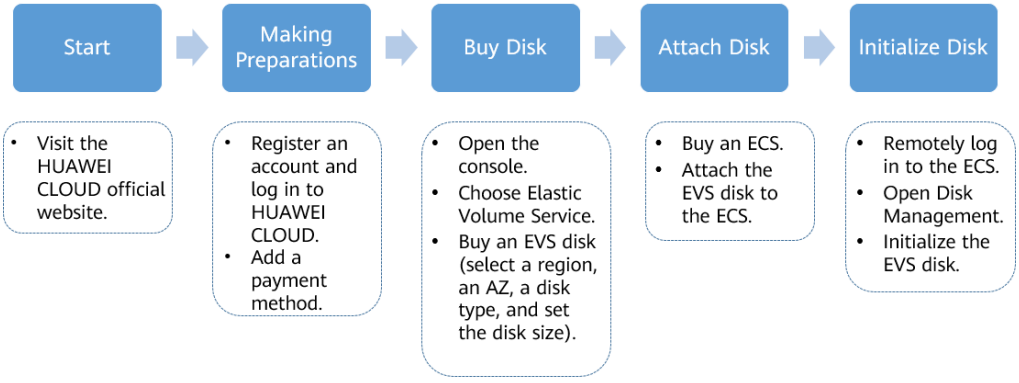


Figure 5-7 EVS configuration process

Figure 5-7 shows the EVS configuration process. You can find detailed usage instructions in the lab guides. Note that EVS disks cannot be used alone. They can be used only after being attached and initialized.

5.2 OBS

5.2.1 What Is OBS?

In addition to EVS, OBS is another commonly used storage service of Huawei Cloud. It offers secure, highly reliable, and inexpensive storage for massive amounts of data.

5.2.2 Architecture

Figure 5-8 shows the architecture of OBS. OBS basically consists of buckets and objects.

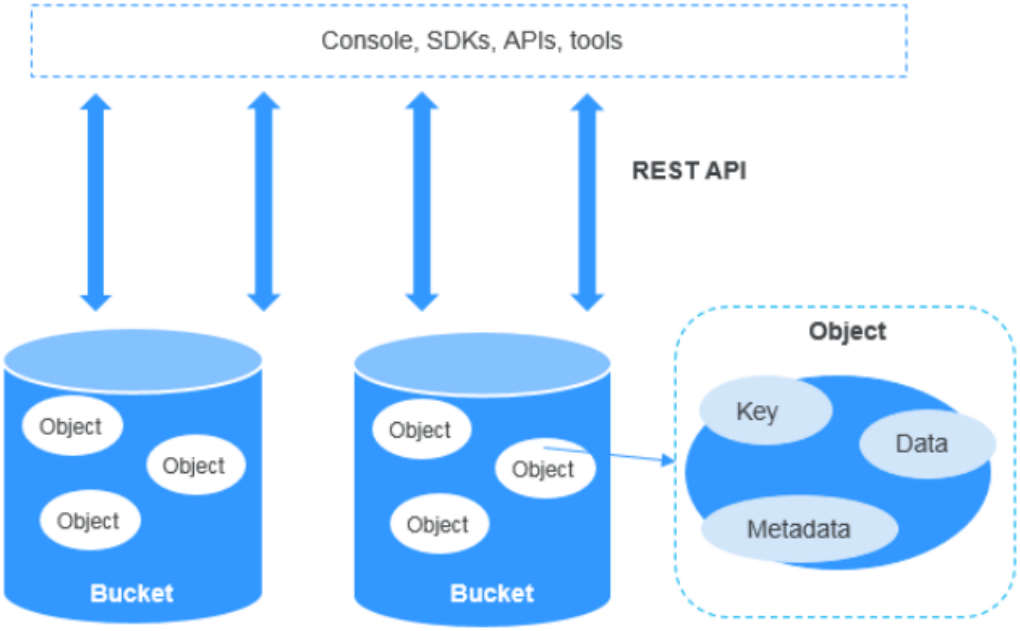


Figure 5-8 OBS architecture

Buckets are containers for storing objects in OBS. Each bucket has its own attributes, such as a storage class, a region, and access permissions. A bucket is accessible through its access domain name over the Internet.

The fundamental storage unit in OBS is the object. An object includes the file data and any metadata that describes it. An object is comprised of three parts: the data, a key, and the metadata.

- Data refers to the content of the object.
- A key specifies the name of an object. An object key is a UTF-8 string up to 1,024 characters long. Each object in a bucket has a unique key.
- Metadata describes an object, and can be system-defined or user-defined. The metadata is a set of key-value pairs that are assigned to the object stored in OBS.

- System-defined metadata is automatically assigned by OBS for processing objects. Such metadata includes Date, Content-Length, Last-Modified, Content-MD5, and more.
- User-defined metadata is specified when users upload objects and is used to describe objects.

Objects are generally managed as files, but OBS, as an object-based service, has no concept of files or folders. For easy data management, OBS provides a method to simulate folders. If you add a slash (/) to an object name, for example, **test/123.jpg**, OBS will present **test** to you as a folder and **123.jpg** as the name of a file stored inside of **test**, but the object key is still actually **test/123.jpg**.

When uploading an object, you can configure a storage class for it. If you do not specify a storage class, the object inherits the storage class of the bucket. The storage class can also be changed after the object is uploaded. On OBS Console or OBS Browser+, you can use folders the same way as you would use folders in a regular file system.

There are a number of different ways to access OBS. With secondary development based on OBS REST APIs, Huawei Cloud offers OBS Console, SDKs, and a variety of other tools for you to easily access your buckets and objects in different scenarios. You can use OBS SDKs and APIs to develop applications customized for your business needs. But no matter what method you use, OBS will always need to authenticate your identity.

OBS supports authentication using an AK/SK pair. It uses AK/SK-based encryption to authenticate requests. When you use OBS APIs for secondary development and use an AK/SK pair for authentication, the signature must be computed based on an algorithm defined by OBS and be added to requests.

OBS allows authentication using a permanent AK/SK pair, or using a temporary AK/SK pair and a security token.

- Permanent AK/SK Pair

  The access key ID (AK) is a unique identifier used in conjunction with a secret access key to sign requests cryptographically.

  The secret access key (SK) works with the AK to identify request senders and prevent their requests from being modified.

  You can create a permanent AK/SK pair on the **My Credentials** page.

- Temporary AK/SK Pair

  Temporary AK/SK pairs and security tokens issued by the system to users are valid for 15 minutes to 24 hours. They comply with the principle of least privilege and are used to access OBS temporarily. If a request does not have a security token, error code 403 will be returned.

  A temporary AK/SK pair works the same way as a permanent AK/SK pair, but is only valid for a limited time.

- Security token

  A security token must be used together with a temporary AK/SK pair to access all resources under a particular account.

5.2.3 Functions

OBS offers many advanced features for refined configurations and management. The following illustrates some commonly-used OBS features:

Server-Side Encryption

With this function enabled:

- OBS will encrypt your object before saving it on the server.
- When you download an encrypted object, OBS decrypts it to plaintext on the server first before returning the decrypted data to you.

URL Validation

Some website creators may steal links from other sites to enrich their content without any costs. This sort of activity places unnecessary strain on the servers where the content is held and hurts the interests of the content owners. URL validation is designed to address this issue. OBS also supports whitelist and blacklist settings.

- If **Whitelisted Referers** is left blank but **Blacklisted Referers** is not, all websites except those specified in the blacklist are allowed to access the target bucket.
- If **Whitelisted Referers** has websites configured, only the websites specified in the whitelist are allowed to access the target bucket, regardless of whether the **Blacklisted Referers** list is configured.

What is **Referer** and how does it work? In HTTP, the **Referer** field allows websites and web servers to identify where people are visiting them from. OBS URL validation takes advantage of this **Referer** field. The idea is that once you find that a request to your resource did not originate from an authorized source, you can have the request blocked or redirected to a specific web page. This way, OBS can prevent unauthorized access to data stored in your buckets.

Versioning

With versioning enabled, OBS can store multiple versions of an object. That way you can then quickly retrieve and restore different versions as needed, or restore data if an application destroys data by mistake. By default, versioning is disabled for new OBS buckets and new objects will overwrite any existing objects with the same names.

User-Defined Domain Name Binding

If you need to migrate files from a website to OBS, and you want to keep the website address unchanged, you can bind a user-defined domain name to a bucket.

For example, if the domain name of your website is **www.example.com** and your website file is **abc.html**, then the URL for accessing this file is **http://www.example.com/abc.html**.

To use **http://www.example.com/abc.html** to access the file stored in OBS:

1. Create a bucket on OBS, and upload **abc.html** to the bucket.
2. On OBS Console, bind the domain name **www.example.com** to the created bucket.
3. (Optional) If there are over 100,000 visits to a website, configure CDN acceleration. CDN can be used to accelerate page loads, file downloads, and VOD.
4. On the DNS server, add a CNAME rule and map **www.example.com** to the bucket's domain name.

After the request for **http://www.example.com/abc.html** reaches OBS, OBS finds the mapping between **www.example.com** and the bucket domain name, and redirects the request to the **abc.html** file stored in the bucket. OBS redirects the request to access **http://www.example.com/abc.html** to **http://***[bucket domain name]***/abc.html**.

Cross-Region Replication

OBS includes disaster recovery across regions, catering to your needs for remote backup. You can configure cross-region replication rules to automatically, asynchronously replicate data from a source bucket to a destination bucket in another region. However, this does require that the source and destination buckets be under the same account. Replicating data across accounts is currently not available.

You can configure a rule to replicate only objects with a specified prefix, or you can replicate all objects in a bucket. Replicated objects in the destination bucket are copies of those in the source bucket. Source objects and their copies have the same names, metadata, content, sizes, last modification time, creators, version IDs, user-defined metadata, and ACLs. By default, a source object and its copy have the same storage class, but you can also specify a different storage class for an object copy if you want.

Cross-region replication can be used to meet the following requirements:

- Compliance requirements: OBS stores data across AZs that are relatively far apart from each other, but regulatory compliance may require further distances. With cross-region replication, you can replicate data across distant OBS regions to meet regulatory compliance requirements.
- Minimized latency: There may be OBS resources that are frequently accessed from different locations. To minimize the access latency, you can use cross-region replication to create object copies in regions nearer to end users.
- Data migration: It can be used to migrate data from one region to another.
- Backup and disaster recovery: For data security and availability purposes, you can create explicit backups of the data written to OBS in one region to a data center in another region. That way you have a secure backup available even if your bucket is somehow irrevocably damaged.
- Easy maintenance: Users may have compute clusters used to analyze the same group of objects, in two different OBS regions and may need to maintain object copies in these two regions.

What bucket contents can be replicated?

With cross-region replication enabled, OBS will replicate the following objects to a destination bucket:

- Newly uploaded objects (excluding those in the Archive storage class)
- Updated objects, for example, the object content is updated or the copied ACL is updated
- Historical objects in a bucket if **Synchronizing Existing Objects** is enabled (excluding objects in the Archive storage class)

Lifecycle Management

Configure lifecycle rules to periodically delete objects or transition objects between storage classes.

You may configure lifecycle management rules to:

- Periodically delete files that are only meant to be retained for specified periods of time.
- Transition documents that are seldom accessed to the Infrequent Access or Archive storage class or delete them.
- Store various types of data in OBS for archive purposes, such as digital media, financial and medical records, original genome sequence data, long-term database backup, and data that must be retained for regulatory compliance.
- Schedule the deletion of all files from a bucket. Deleting objects manually is time-consuming, and only a limited number of objects can be deleted at a time.

Using lifecycle rules to transition objects that are infrequently accessed to the Infrequent Access or Archive storage class can save storage costs. Transition basically means that the object storage class is altered without copying the object. To transition the storage class of an object, you can configure a lifecycle rule or manually change the storage class on the **Objects** page.

Lifecycle rules can be configured for buckets with versioning enabled or disabled. By default, versioning is disabled, but you can enable it as needed. If versioning is enabled for a bucket, one current object version and zero or more non-current object versions will be maintained at the same time. You can manage object storage costs using versioning and lifecycle rules together. Predefined lifecycle management actions can facilitate management of the lifecycles of current and non-current object versions.

Permission Management

By default, OBS resources (buckets and objects) are private and can only be accessed by their owner. Other users cannot access your OBS resources without authorization. You can control OBS through granting access permissions to other accounts or IAM users by editing access policies. For example, you can authorize another IAM user to upload objects to your bucket or make your bucket public for anyone to access it over the Internet. OBS offers multiple methods to help you manage resource permissions for your buckets. Resource owners can flexibly customize permission control policies as needed to keep their data secure.

5.2.4 Advantages

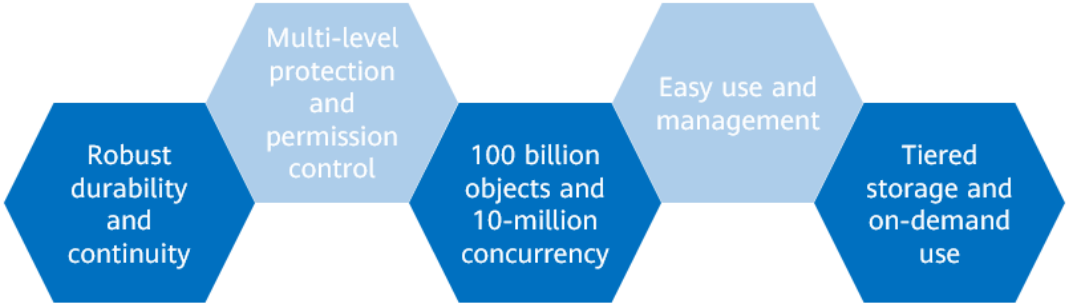


Figure 5-9 OBS advantages

OBS has the following five advantages:

- **Robust Durability and Reliability**

  Huawei mobile phones use OBS to provide stable and reliable backend storage with access supported for hundreds of millions of users. It delivers a data durability of up to 99.9999999999% and service continuity of up to 99.995% thanks to cross-region replication, cross-AZ disaster recovery, intra-AZ device and data redundancy, slow disk or bad sector detection, and other technologies.

- **Multi-Level Protection and Permission Control**

  Huawei Cloud OBS lets you store your data safely. It keeps your data secure and trusted by using versioning, server-side encryption, URL validation, VPC-based network isolation, log auditing, and fine-grained permissions control.

- **100-Billion Objects, 10-Million Concurrency**

  With intelligent scheduling and response, optimized data access paths, and technologies such as event notification, transfer acceleration, and big data vertical optimization, you can store hundreds of billions of objects in OBS, and still experience smooth concurrency with up to hundreds of billions of tasks, along with ultra-high bandwidth and low latency.

- **Easy Use and Management**

  RESTful APIs, SDKs in different languages, and data migration tools make it easier to migrate services to the cloud. There is no need to plan storage capacity beforehand or worry about capacity expansion or reduction, because storage resources are linearly scalable and effectively limitless.

- **Tiered Storage and On-Demand Use**

  OBS is billed on a pay-per-use or monthly/yearly basis. Data in Standard, Infrequent Access, and Archive storage classes is metered and billed separately to reduce costs.

5.2.5 How to Access OBS

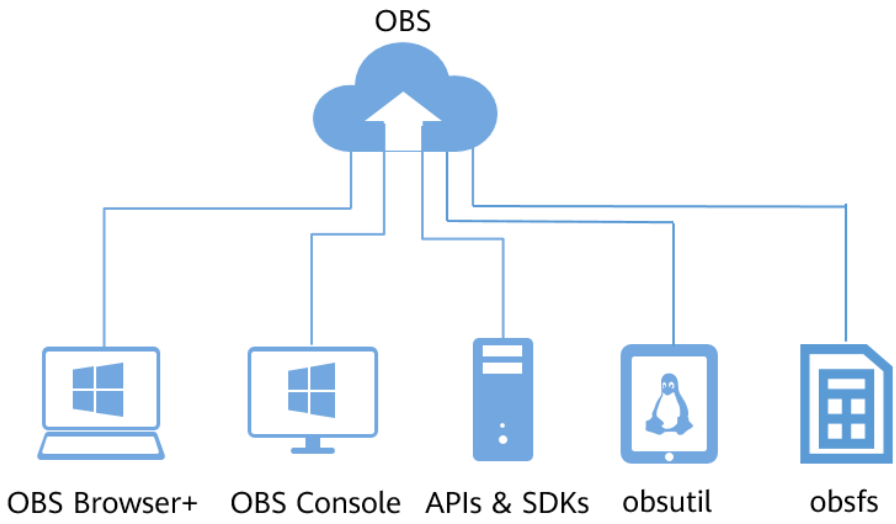


Figure 5-10 Ways for accessing OBS

You can access OBS in any of the following ways:

- OBS Console

  OBS Console is a web-based UI. You can manage all of your OBS resources through this console.

  To access OBS Console, you need an account or an IAM user account.

- SDKs & APIs

  OBS SDKs encapsulate APIs provided by OBS to simplify development. You can call API functions provided by OBS SDKs to access OBS capabilities. With OBS RESTful APIs, you can easily access OBS from web applications. Using APIs, you can upload and download OBS data through any Internet connected device anytime, anywhere.

- OBS Browser+

  OBS Browser+ is a Windows client that lets you to easily manage OBS resources from your desktop.

  When using OBS Browser+, you can use access keys (AK/SK) for identity authentication. To access OBS resources, you can use OBS Browser+ or configure a server address.

- obsutil

  obsutil is a command line tool for accessing OBS. You can use it to perform common configuration and management operations on OBS. If you are comfortable using the command line interface (CLI), obsutil is the recommended method for batch processing and automated tasks. You can download obsutil, configure the server address, and use access keys (AK and SK) for identity authentication.

- obsfs

  obsfs is an OBS tool based on Filesystem in Userspace (FUSE). It allows you to use OBS to mount parallel file systems to Linux operating systems. With obsfs, you can easily access the virtually unlimited storage space of OBS the same way as you would use a regular local file system.

5.2.6 Application Scenarios

OBS can be used in a wide range of application scenarios. The following are five common examples:

- **Big Data Analytics**

  OBS enables inexpensive big data solutions that feature high performance with zero service interruptions. It eliminates the need for capacity expansion. Such solutions are designed for scenarios involving storage and analysis of massive amounts of data, query of historical data details, analysis of a large number of behavior logs, analysis of public transactions, and statistics collection.

- **Static Website Hosting**

  OBS provides a website hosting solution that is cost-effective, highly available, and can scale automatically based on traffic volume. Combined with CDN and ECS, you can quickly build a website or an application system with static and dynamic content stored separately.

  Dynamic data on end user devices and apps interacts directly with the systems deployed on Huawei Cloud. Requests for dynamic data are sent to those systems for

processing and then returned to end users. Static data, in contrast, is stored on OBS and is processed over an intranet connection. End users can request and read the static data from OBS through nearby high-speed nodes.

- **Video on Demand**

  OBS provides high currency, low latency access to massive amounts of stored data and it does it reliably, inexpensively. Working with MPC, Content Moderation, and CDN services, OBS can help you quickly construct a fast, secure, and highly available video on demand (VOD) platform. OBS serves as the VOD origin server. After regular users or professional creators upload their video content to OBS, Content Moderation reviews the uploaded content and MPC transcodes it. Then CDN ensures the content is delivered swiftly to end terminals for playback.

- **Backup and Archiving**

  OBS offers a highly reliable, inexpensive storage system featuring high concurrency and low latency. It can hold massive amounts of data, meeting the archive requirements for unstructured data of apps and databases. You can use synchronization clients, mainstream backup software, Cloud Storage Gateway (CSG), or DES to back up your on-premises data to OBS. In addition, OBS provides lifecycle rules to automatically move objects between storage classes to save you money on storage. You can restore data from OBS to a DR host or test host on the cloud if necessary.

- **Enterprise Web Disks**

  OBS works with cloud services such as ECS, ELB, RDS, and VBS to provide enterprise web disks with a low latency, high concurrency storage system that is reliable, and inexpensive. The storage capacity automatically scales as the volume of stored data grows. Dynamic data on user devices such as mobile phones, PCs, and tablets interacts with the enterprise cloud disk business system on Huawei Cloud. Requests for dynamic data are sent to this system for processing and then returned to devices. Static data, in contrast, is stored on OBS and is processed by the business system over an intranet connection. End users can request and read static data directly from OBS. Additionally, OBS allows you to configure lifecycle rules to automatically transition objects between storage classes to reduce costs.

5.3 SFS

5.3.1 What Is SFS?

Scalable File Service (SFS) provides reliable, high-performance shared file storage hosted on Huawei Cloud. With SFS, you can enjoy shared file access spanning multiple ECSs, BMSs, and containers created on CCE and CCI.

5.3.2 Key Concepts

Before using SFS, there are a few concepts worth noting. NFS, CIFS, and file system are basic file storage concepts.

- NFS

Network File System (NFS) is a distributed file system protocol that allows different computers and operating systems to share data over a network.

- CIFS

  Common Internet File System (CIFS) is a protocol used for network file access. It is a public or open version of the Server Message Block (SMB) protocol, which was initiated by Microsoft. CIFS allows applications to access files on computers over the Internet and send requests for file services. Using the CIFS protocol, network files can be shared between hosts running Windows.

- File system

  A file system provides users with shared file storage service through NFS or CIFS. It is used for accessing network files remotely. After a user creates a file system on the management console, the file system can be mounted to multiple ECSs and is accessible through the standard POSIX.

5.3.3 Advantages

SFS has the following advantages over traditional file sharing storage:

- Elastic scalability: Storage can be scaled up or down on demand to dynamically adapt to service changes without interrupting applications. Resizing can be done with a few clicks.
- High performance and reliability: SFS enables file system performance to increase as capacity grows, and delivers a high data durability to support rapid service growth.
- Seamless integration: SFS supports NFS and CIFS protocols. With standard access protocols, a broad range of mainstream applications can read and write data in the file systems. In addition, SFS is compatible with SMB 2.0, SMB 2.1, and SMB 3.0, so Windows clients can access the shared space.
- Simple operation and low cost: You can create and manage file systems with ease in a GUI.

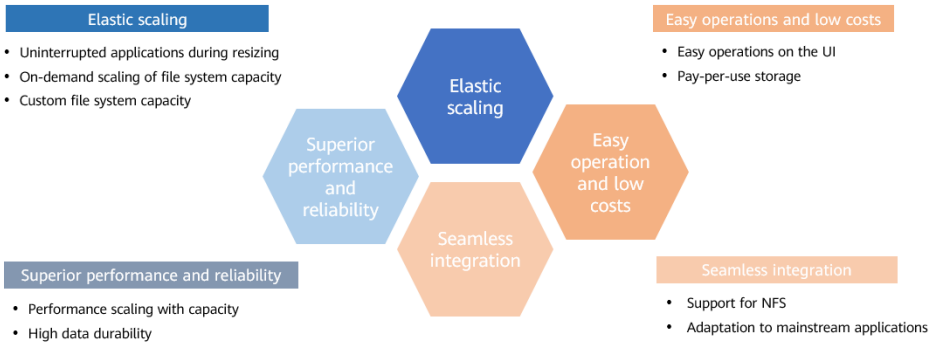


Figure 5-11 SFS advantages

5.3.4 Application Scenarios

5.3.4.1 SFS Capacity-Oriented

- HPC

  In industries that require a lot of HPC, for instance biopharmaceutical research, gene sequencing, image processing, scientific research, weather forecasts and other applications involving complex computer simulations. SFS can provide high bandwidth, large capacity shared storage.

- Media processing

  TV station and new media are more likely to be deployed on cloud platforms than before. Such services include streaming media, archiving, editing, transcoding, content distribution, and video on demand (VoD). In such scenarios, a large number of workstations are involved in the whole program production process. Different workstations using different operating systems may require file sharing based on the same file system. Recently HD/4K video has become a major trend in the broadcast industry. Take video editing as an example. Many high quality projects today involve HD editing projects with 30 to 40 layers. A single editing client may require a file system that can handle hundreds of MB per second. Usually, producing a single TV program needs several editing clients to process a lot of video materials concurrently. SFS provides customers with enough stable, bandwidth-intensive, and latency-sensitive performance to meet these requirements.

- Content management and web service

  SFS can be used in various content management systems to provide shared file storage for websites, home directories, online releasing, and archiving.

- Big data and analytic applications

  SFS delivers an aggregate bandwidth of up to 10 GB/s, enough to handle ultra-large data files such as satellite image. In addition, SFS has robust reliability to prevent service interruptions due to system failures.

5.3.4.2 SFS Turbo

Expandable to 320 TB, SFS Turbo provides a fully hosted shared file storage. It features high availability and durability to support massive small files and applications requiring low latency and high IOPS. SFS Turbo is suitable for high-performance websites, log storage, compression and decompression, DevOps, enterprise offices, and containerized applications.

- High-performance websites

  For I/O-intensive website services, SFS Turbo can provide shared website source code directories for multiple web servers, enabling low-latency and high-IOPS concurrent shared access.

- Log storage

  SFS Turbo can provide multiple service nodes for shared log output directories, facilitating log collection and management of distributed applications.

- DevOps

  The development directory can be shared to multiple VMs or containers, simplifying the configuration process and improving R&D experience.

- Enterprise office

  Enterprise office documents can be saved in an SFS Turbo file system for high-performance shared access.

5.3.5 How to Use SFS

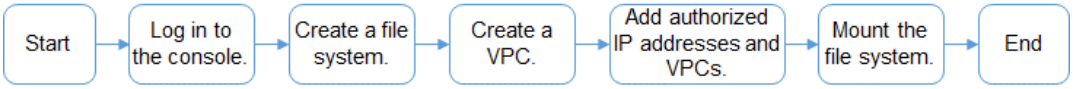


Figure 5-12 SFS configuration process

There are two steps involved. First, create a file system. You can create an SFS Capacity-Oriented or SFS Turbo file system. Then, once the files system has been created, you mount it to your ECSs so that they can share access. The following sections describe how to mount an NFS file system in Linux and how to mount a CIFS file system in Windows.

5.3.5.1 Mounting an NFS File System to a Linux ECS

After creating a file system, mount the file system to ECSs so that the ECSs can share the file system. The following example uses user **root** to log in to the Linux ECS and mount the file system:

1. Install the NFS client.
2. Run the following command to check whether the file system domain name can be resolved: (SFS Turbo file systems do not require domain name resolution. You can skip this step and directly mount the file system.)

   nslookup *File system domain name*
3. Run the following command to create a local path for mounting the file system:

   mkdir *Local path*

   Run the following command to mount the file system to the ECS: (Currently, the file system can be mounted to Linux servers using NFS v3 only.)

   mount -t nfs -o vers=3 timeo=600 *Mount address Local path*

   After the file system is mounted, check that you can access the file system on the server.

   In the Linux command line:

   - **nslookup** is used to resolve domain names.
   - **mkdir** creates a directory. We will use it to create a local path for the file system. *Local path* is the name of the folder to be created.

– **mount** mounts the filesystem. The **-t** argument specifies the type of the file system, in this example, **nfs**. The **-o** argument sets the protocol and configures a timeout interval, in this example, **v3** and **600s**.

5.3.5.2 Mounting a CIFS File System to a Windows ECS

To mount a CIFS file system to ECSs running Windows Server 2012, perform the following steps:

1. Log in to an ECS running Windows Server 2012.
2. Click **Start**, right-click **Computer**, and choose **Map network drive**.
3. In the dialog box that is displayed, for **Folder**, enter the file system mount address. The format is *File system domain name**Path*.

5.3.5.3 Unmounting a File System

The procedures for unmounting a file system in Linux and Windows are as follows.

- Linux
  (1) Log in to the ECS.
  (2) Run the following command:
  umount *Local path*
- Windows
  (1) Log in to the ECS.
  (2) Right-click the file system to be unmounted and choose Disconnect.
  (3) The file system has been unmounted when it disappears from the network locations.

5.3.5.4 Configuring VPCs

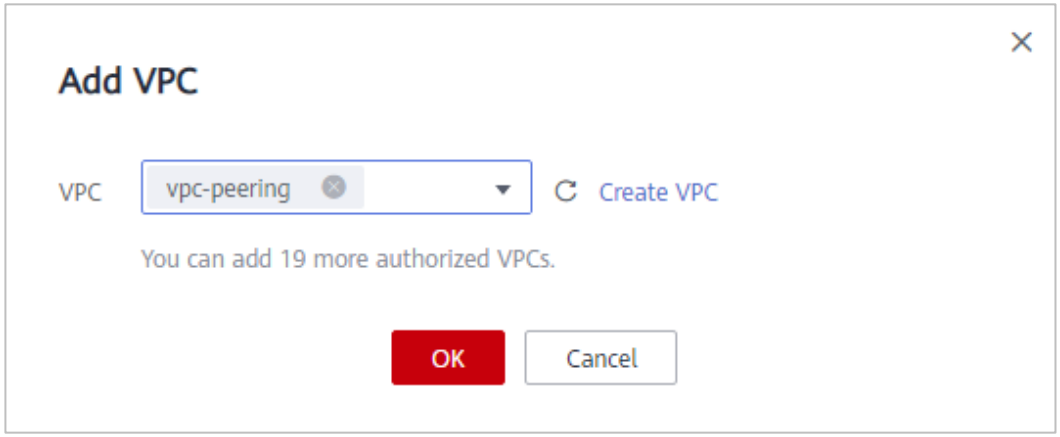


Figure 5-13 Configuring VPCs

Multiple VPCs can be configured for an SFS file system so that servers belonging to different VPCs can share the same file system.

Servers belonging to different VPCs can share the same file system as long as those VPCs are added to the VPC list of the file system or those servers are added to the authorized addresses of the VPCs.

6 More Cloud Services

In addition to compute, storage, and networking services, enterprise customers need services like database and security services, Content Delivery Network (CDN), and EI services for cloud transformation. These services are available on-demand and are easy to maintain. They help customers reduce CAPEX and facilitate O&M.

This chapter introduces database, security, CDN, and EI services.

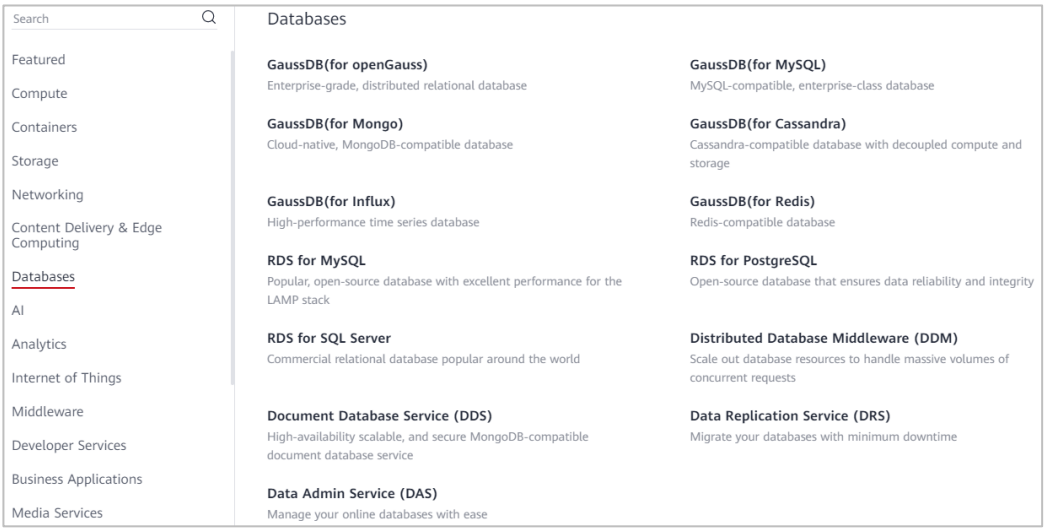


Figure 6-1 Overview of other cloud services

6.1 Database Services

6.1.1 Database Basics

As enterprise applications are increasingly migrated to the cloud, database migration is inevitable. A wide range of database services are available on Huawei Cloud. However, before introducing these services, we need to understand some basic database concepts.

First, we need to know what a database is and what an instance is.

A database is a collection of files that contain data organized using a given data model.

An instance contains a set of background processes and memory blocks. It is a data management software that connects the user and the operating system (OS).

We all know that data can be stored in multiple media, such as memory and disks. A database is also a medium for storing data.

All operations for the data of databases, such as defining data, querying data, maintaining data, and managing databases, are performed on database instances. Your applications can interact with the databases only through database instances.

Depending on their data model, databases are classified as either relational databases or non-relational databases.

6.1.1.1 Relational and Non-relational Databases

- Relational Databases

  (1) Definition

  A relational database organizes data using a relational model. The data is stored in rows and columns. A series of rows and columns in a relational database comprises a table, and a group of tables form a database. A user retrieves data from a database through a query, which is the execution code that qualifies certain areas of a database.

  A relational model can be understood as a two-dimensional table model, and a relational database is a data organization consisting of two-dimensional tables and their relationships.

  (2) Common Relational Databases

  Common relational databases include Oracle Database, SQL Server, IBM DB2, MySQL, and Microsoft Access. Each type of database has unique syntax. Enterprises have to pay a high price for using most relational databases, except open-source MySQL. MySQL is free, but has many restrictions on its performance.

  (3) ACID Properties

  ACID stands for Atomicity, Consistency, Isolation, and Durability. Relational databases are ACID-compliant to ensure data integrity for complex data queries. In addition, relational databases provide strong data consistency for transactions. They manage each transaction as an atomic unit. Any partial updates can be rolled back in the event of a failure during a transactional update.

- Non-relational Databases

  (1) Definition

  A non-relational database is also called a NoSQL database (Not Only SQL). According to Wikipedia, NoSQL first appeared in 1998. It was a lightweight, open-source, and SQL-incompatible relational database developed by Carlo Storzzi. The concept of NoSQL was introduced again at an event to discuss open-source, distributed databases in 2009. This time, NoSQL was referred to as a non-relational, distributed database designed on a non-ACID mode. At the NoSQL (east) seminar held in Atlanta of the same year, NoSQL was defined as "non-relational" databases featuring key-value stores and document stores, not simply a Relational Database Management System (RDBMS). Since then, NoSQL began to take more prominence on the world stage.

(2) Common Non-relational Databases

Common non-relational databases include Redis, Amazon DynamoDB, Memcached, Microsoft Azure Cosmos DB, and Hazelcast.

(3) ACID

Non-relational databases do not necessarily support the ACID properties for transaction processing. Most non-relational databases are distributed databases. Redis has become the most popular distributed database in recent year.

6.1.1.2 OLTP and OLAP

Databases can be classified as either OLTP (Online Transaction Processing) or OLAP (Online Analysis Processing) types. OLTP is a typical application of relational databases. It is used for processing basic and routine transactions, such as bank transactions. OLAP is a typical application of data warehouses. It supports complex analysis and provides clear query results, helping companies make informed decisions.

- OLTP

  OLTP is also known as a transaction-oriented processing system. It can immediately transmit customers' raw data to a computing center and provide results quickly. OLTP is also called a real time system for its lightning-fast data input and response. Response time is an important metric for an OLTP system. It is the time from a user enters data on a terminal to when the response is returned. OLTP databases enable transactional programs to write only required data so that a single transaction can be processed as quickly as possible.

- OLAP

  The concept of OLAP was first proposed by Dr. E. F. Codd, the Father of Relational Databases, in 1993. OLAP is a technology used to organize large business databases and support business intelligence (BI). An OLAP database consists of one or more cubes, and each cube is organized and designed by a cube administrator to fit the way you retrieve and analyze data. This allows you to easily create and use PivotTables and PivotCharts reports.

  OLAP is a software technology enabling multidimensional data sharing and fast online data access and analysis for specific problems. It allows decision makers to observe data in depth through quick, stable, consistent, and interactive access to multiple possible forms of observation. Data for decision-making is multidimensional and plays a fundamental role during making decision. OLAP is dedicated to complex analysis and provides support for decision makers and top management personnel. It can quickly and flexibly process complex queries on a large volume of data based on analysis personnel requirements, and it provides clear query results. In this way, they can precisely track enterprise processes and the needs of their customers. This information helps enterprises plan more appropriately.

  OLAP provides flexible analysis functions, intuitive user interface, and visualized analysis results, helping you easily analyze a large volume of complex data and make correct decisions. It provides analysis results in chart or table form to help

verify complex assumptions. OLAP does not show exceptions. It uses a knowledge verification method.

OLAP pre-establishes a multidimensional data model for users by imitating the thinking process of a human being. Here, dimensions refer to user perspectives, for example, sales volume, periods of time, product types, distribution channels, geographical distribution, and customer groups. Once a multidimensional data model is created, you can quickly obtain analyses based on a range of different dimensions, and can dynamically adjust the dimensions, or perform multi-dimensional comprehensive analysis as needed. This is the reason why OLAP has attracted so much attention in recent years. OLAP is essentially different from the traditional management information system in terms of design principle and implementation.

- OLTP vs. OLAP

  OLTP processes basic and routine transactions. For example, depositing or withdrawing money in a bank is a transaction. OLTP applications:

  - Support real-time analytical queries.
  - Handle a small amount of queried data.
  - Access deterministic data.
  - Handle concurrent requests while maintaining the integrity and security of transactions.

  OLAP is mainly used in data warehouses, typically used for complex dynamic report systems. OLAP applications:

  - Typically do not demand real-time execution.

    Few OLAP applications need to update their data more than once a day.
  - Need to analyze large amounts of data.

    OLAP supports dynamic query. Users need to collect statistics from large amounts of raw data to obtain desired information, such as time series analysis.
  - Provide support for decision-making.

    Users can submit query requests at any time. OLAP uses an important concept, "dimensions", to build a dynamic query platform which helps you figure out what you need to know.

6.1.2 Huawei Cloud Database Service Overview

Huawei cloud database services include both relational and non-relational databases. Huawei provides a wide range of database services to best suit customer requirements. The GaussDB series was developed by Huawei to provide reliable, high performance databases for government and enterprise. Relational Database Service (RDS) series is an open source, cost effective database service for small and medium enterprises.

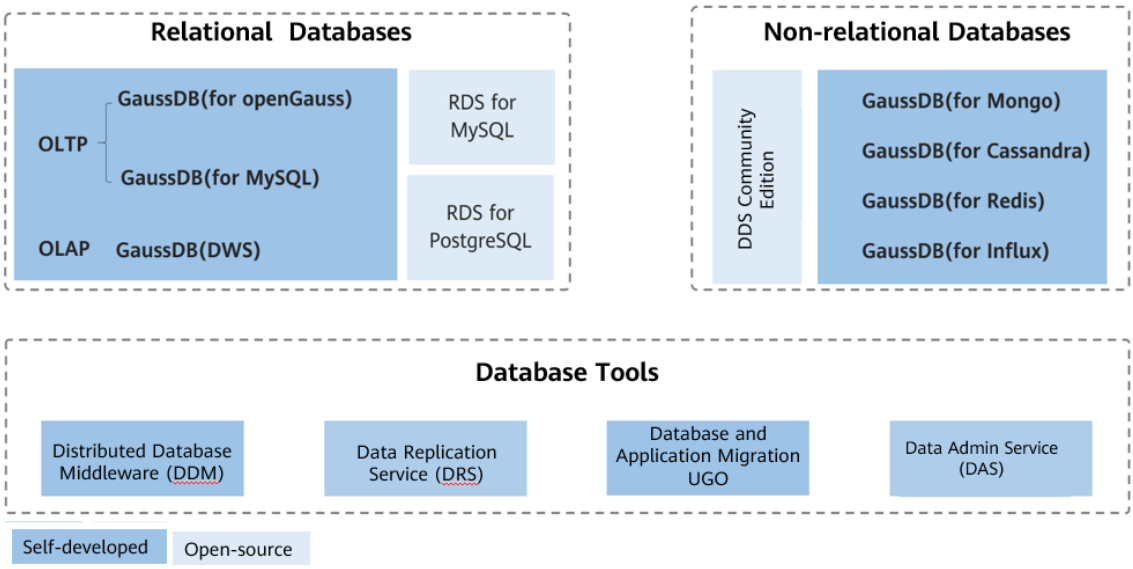


Figure 6-2 Database service overview

Huawei Cloud database services include GaussDB series and RDS series. Each series contains relational and non-relational database services.

- GaussDB series

  GaussDB is an enterprise-grade distributed relational database from Huawei. It features hybrid transactional/analytical processing (HTAP) workloads and intra-city cross-AZ deployment with zero data loss. With a distributed architecture, GaussDB(for openGauss) supports petabytes of storage and supports over 1,000 nodes per DB instance. It is highly available, secure, and scalable and provides capabilities including quick deployment, backup, restoration, monitoring, and alarm reporting for enterprises.

- RDS series

  RDS is an online cloud database service built on the cloud computing platform. It is stable, reliable, scalable, and easy to manage. RDS supports the following DB engines:

  - MySQL
  - PostgreSQL
  - SQL Server

  RDS provides a comprehensive performance monitoring system, multi-level security protection measures, and a professional database management platform, allowing users to easily set up and scale databases. On the RDS console, users can perform almost all necessary tasks and no programming is required. The console simplifies operations and reduces routine O&M workloads, so users can stay focused on application and service development.

  We will focus on the RDS databases in the following section.

6.1.3 RDS for MySQL

MySQL is one of the world's most popular open-source relational databases. It works with the Linux, Apache, and PHP (LAMP) stack to provide efficient web solutions. RDS for MySQL is reliable, secure, scalable, inexpensive, easy to manage, and immediately ready for use.

- It uses a stable architecture and supports various web applications. In addition, it is cost-effective and is preferred by small and medium enterprises.
- A web-based console is available for you to monitor the resources of your instances, making operations easy and visual.
- You can flexibly scale resources based on your service requirements and pay for only what you use.

6.1.3.1 Advantages

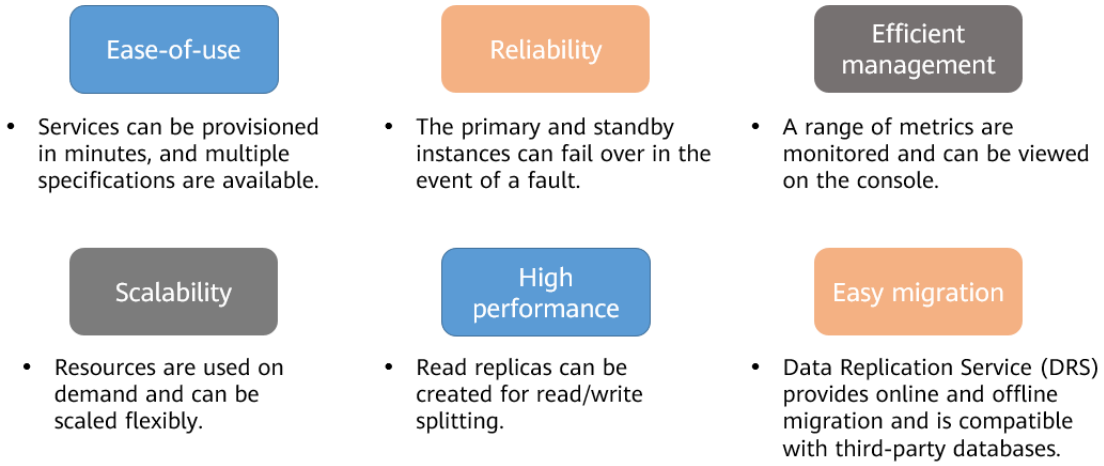


Figure 6-3 RDS for MySQL advantages

6.1.3.2 Architecture

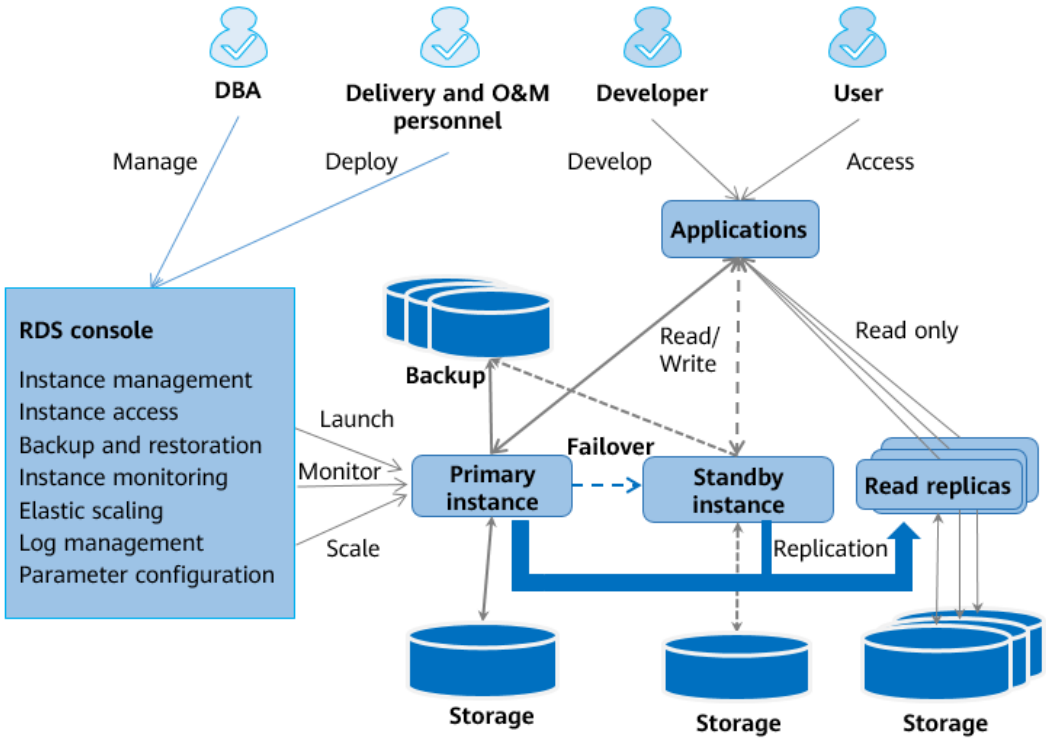


Figure 6-4 RDS for MySQL architecture

As shown in the preceding figure, the smallest management unit of RDS is a DB instance. A DB instance is an isolated database environment running in the cloud. RDS DB instances include single instances and primary/standby instances.

RDS for MySQL provides the following functions:

- Elastic scaling:
  - Horizontal scaling: Read replicas (up to 5 for each instance) can be created or deleted.
  - Vertical scaling: The instance vCPUs, memory, and storage (up to 4 TB) can be scaled.
- Backup and restoration:
  - Backup: Automated, manual, full, and incremental backups are supported. Backups can be added, deleted, queried, or replicated.
  - Restoration: Data can be restored to any point in time within the backup retention period, or to a new or original DB instance. The backup retention period is up to 732 days.
- Log management: Slow query logs and error logs can be queried and downloaded.
- Parameter configuration: Database administrators (DBAs) can adjust DB engine parameter configurations based on monitoring metrics and log information for database tuning. DB engine parameters can be added, deleted, modified, queried, reset, compared, and replicated through parameter template management.

6.1.3.3 Features

Three major features of RDS for MySQL are described as follows: cross-AZ HA, read/write splitting, and point-in-time recovery (PITR).

- Cross-AZ HA is an effective disaster recovery (DR) mechanism. If your workloads require high database reliability, you can use primary/standby DB instances and deploy your primary and standby instances across AZs to achieve AZ-level DR.
- Read/write splitting enables read and write requests to be automatically routed through a read/write splitting address. After read replicas are created, you can enable read/write splitting to automatically route write requests to the primary DB instance and read requests to read replicas by predefined weights. You can create up to five read replicas for each RDS for MySQL instance.
- You can use backups to restore data to any point in time. Binlog is a binary log used to record changes in MySQL database table structure and table data.

6.1.3.4 Application Scenarios

Figure 6-5 RDS for MySQL application scenarios

RDS for MySQL is mainly used in the following scenarios:

- Users of public cloud platforms other than HUAWEI CLOUD generally use RDS for MySQL.
- Start-ups choose RDS for MySQL in the early stages because they need ways to support fast growth on a limited budget.
- MySQL is used widely by Internet, e-commerce, and game enterprises. When migrating databases to the cloud, these types of enterprises choose RDS for MySQL.
- IoT applications tend to be very large scale and they need to be extremely reliable. RDS for MySQL is the first choice for IoT enterprises because it allows for a large number of concurrent connections and does not require customers to reconstruct their applications.

6.1.4 RDS for PostgreSQL

6.1.4.1 What Is RDS for PostgreSQL?

RDS for PostgreSQL is a typical open-source relational database that excels in data reliability and integrity. It is suitable for e-commerce, geographic location application systems, financial insurance systems, complex data object processing, and other application scenarios.

PostgreSQL is based on Postgres, which was developed at the University of California, Berkeley. After more than 30 years of development, PostgreSQL has become the most powerful open-source database in the world. It has earned a reputation for reliability, stability, and data consistency, and has become the preferred open-source relational database for many enterprises.

6.1.4.2 Architecture

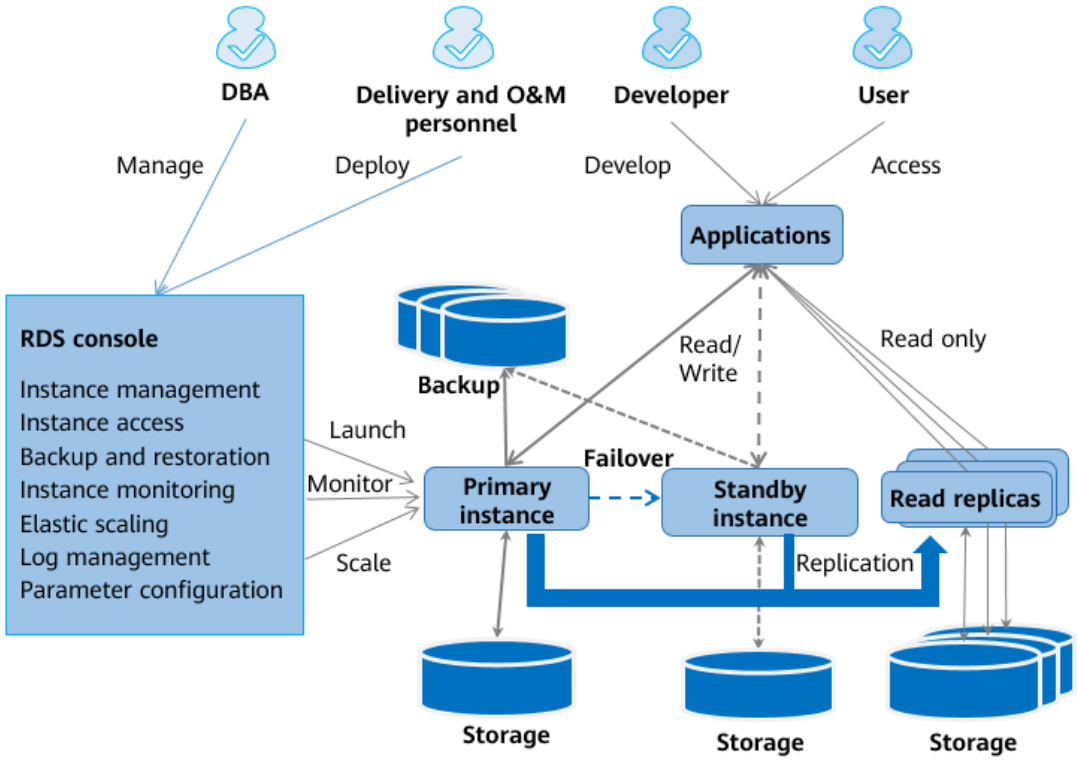


Figure 6-6 RDS for PostgreSQL architecture

As shown in the preceding figure, RDS for PostgreSQL has the following features:

- Database type: PostgreSQL 9.5, 9.6, 10.0, 11, 12, and Enhanced Edition are provided.
- Security: Multiple security measures such as VPCs, subnets, security groups, and SSL are provided to protect databases and user privacy.

- HA: Data is automatically synchronized from a primary DB instance to a standby DB instance. If the primary DB instance fails, services are quickly and automatically switched over to the standby DB instance.
- Monitoring: Key performance metrics of DB instances are monitored, including CPU usage, memory usage, storage space usage, I/O activities, database connections, QPS, TPS, buffer pool, and read/write activities.
- Elastic scaling
  - Horizontal scaling: Read replicas (up to 5 for each instance) can be created or deleted.
  - Vertical scaling: DB instance classes can be changed.
  - Instances can be scaled out in a few clicks with no downtime.
- Log management: Slow query logs and error logs can be queried.
- Parameter configuration: DBAs can adjust DB engine parameter configurations based on monitoring metrics and logs for database tuning.

6.1.4.3 Advantages

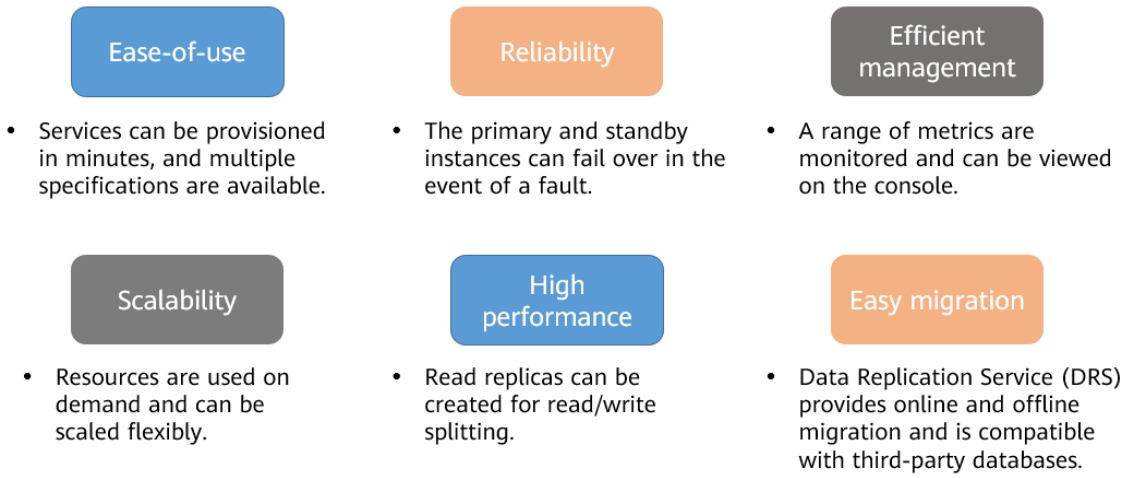


Figure 6-7 RDS for PostgreSQL advantages

Due to RDS for PostgreSQL's numerous advantages, customers have mainly used it to replace Oracle.

6.1.4.4 Application Scenarios

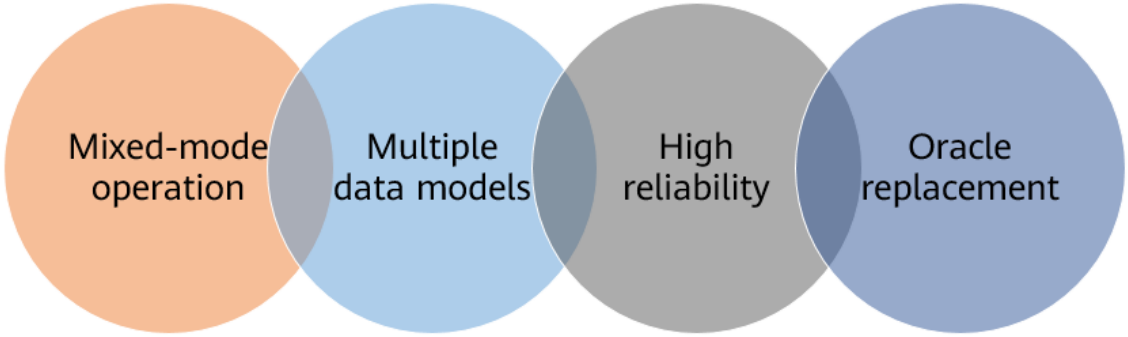


Figure 6-8 RDS for PostgreSQL application scenarios

Mixed-mode operations combining OLTP and OLAP are supported.

Multiple data models are applicable to spatiotemporal, geographic, heterogeneous, image, text retrieval, time series, stream computing, and multi-dimensional scenarios.

Huawei provides you with a reliable database service and keeps your data consistent.

To replace Oracle databases, there are two solutions available: RDS for PostgreSQL Enhanced Edition and RDS for PostgreSQL Community Edition plus Orafce.

6.1.4.5 Features

RDS for PostgreSQL has two main features: high availability and point-in-time recovery.

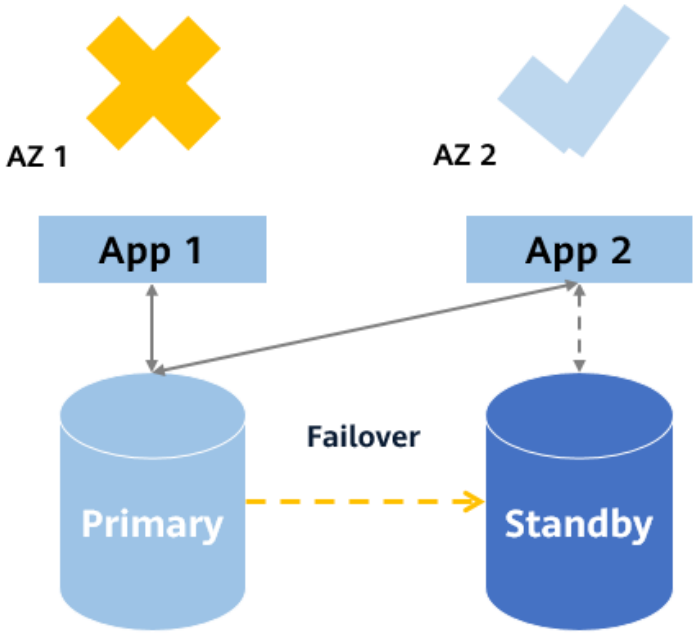


Figure 6-9 High availability

High Availability

- You can choose a failover policy to prioritize reliability or availability.
- DB instances can be deployed in one AZ or across AZs and can automatically fail over within a cluster.
- You can manually switch a primary instance to standby to simulate a fault.
- Read replicas can automatically associate themselves with a new primary instance.
- A failover can complete in seconds.
- The standby database does not handle traffic. It only ensures RTO.
- The Huawei-developed HA Monitor module is used.
- Virtual IP addresses can be switched completely invisibly to the applications.
- Multiple primary/standby switchovers can be performed.
- Automatic fault detection is provided.

Point-In-Time Recovery

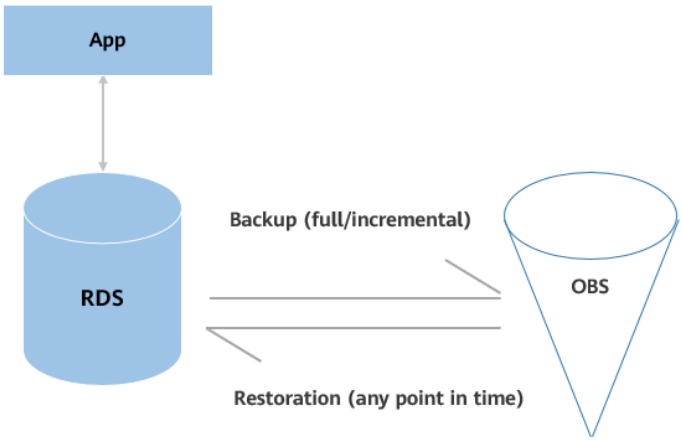


Figure 6-10 Point-in-time recovery

- Backup cycle: 7 to 732 days
- Pay-per-use: Free EVS storage space equal to the requested storage and virtually limitlessly expandable
- Up to 11 nines of data reliability
- Security encryption: KMS encryption and multiple protections
- Data archived in OBS can be restored to any point in time.

6.1.5 DDS

6.1.5.1 What Is DDS?

Document Database Service (DDS) is a high-performance, highly availability MongoDB-compatible database service that is scalable and secure. It provides one-click deployment, elastic capacity expansion, disaster recovery, backup, restoration, monitoring, and alarm

reporting. An instance is a basic management unit of DDS. A DDS instance consists of databases, collections, and documents.

Database: One or more databases can be created in a single DDS instance, and one or more collections can be created in each database.

Collection: A collection is a group of multiple MongoDB documents.

Document: A document is a group of key-value pairs (BSON). It is the basic unit of data in MongoDB.

6.1.5.2 Key Concepts

Each DDS cluster consists of a config node, and multiple mongos and shard nodes. The following diagram shows the node relationships.

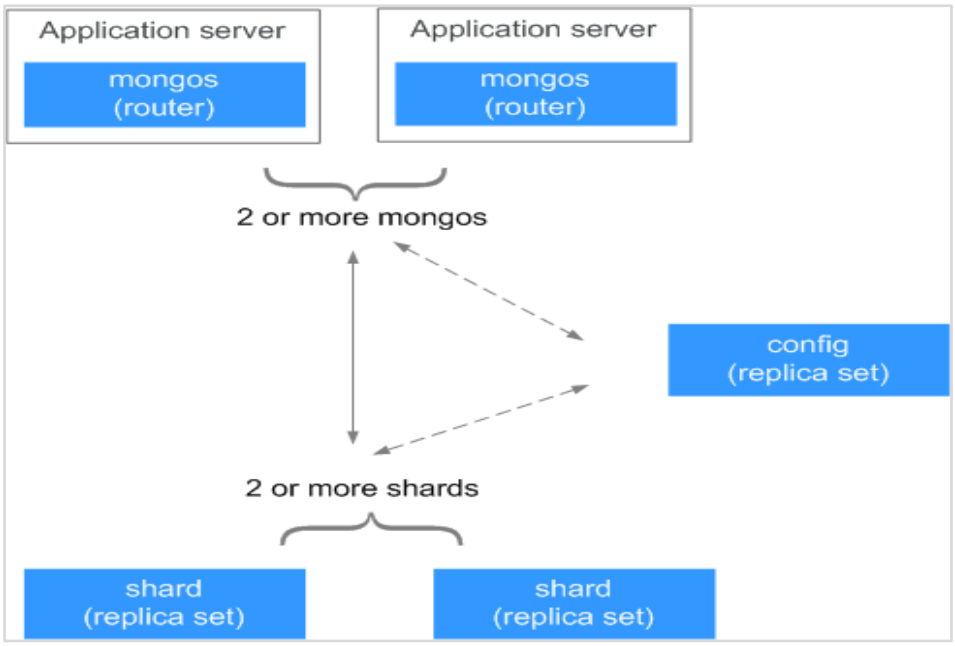


Figure 6-11 DDS cluster architecture

- mongos

  A mongos is a router for reading and writing data, providing a unified interface for accessing DB instances. Each DB instance has 2 to 32 mongos. You can specify the quantity. A mongos reads configuration settings from configs and allocates read and write requests to shards. You can connect to a mongos directly.

- config

  A config stores configuration settings for DB instances and consists of one replica set. You cannot connect to a config node directly.

- shard

  Shards are used to store user data. Each DB instance of Community Edition has 2 to 32 shards. You can specify the quantity. Each shard is deployed as a replica set to ensure data redundancy and high reliability. You cannot connect to a shard node directly.

6.1.5.3 Advantages

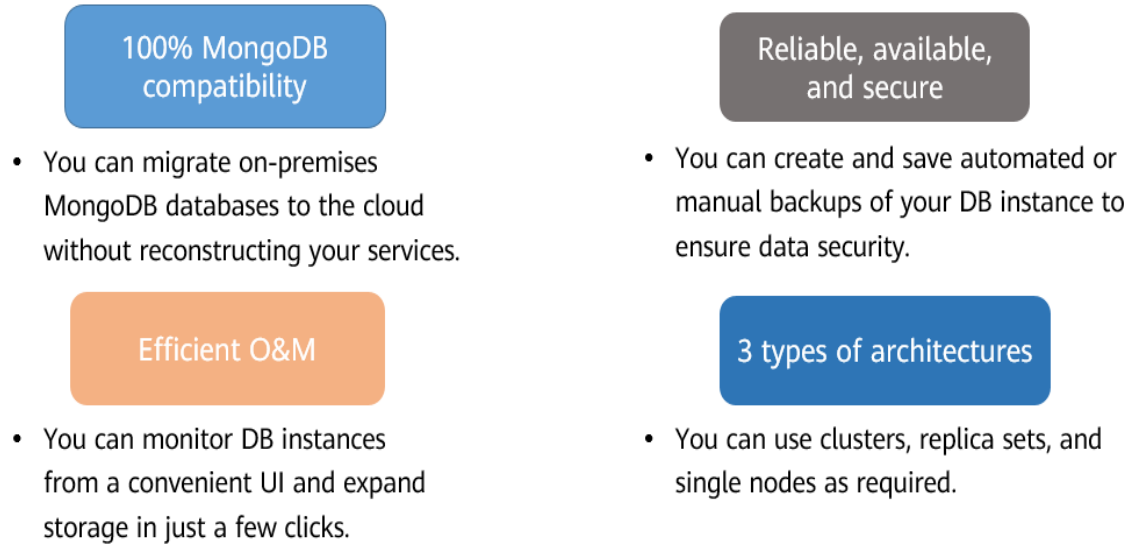


Figure 6-12 DDS advantages

DDS has the following four advantages:

- Fully compatible

  DDS is a document-oriented NoSQL database that is fully compatible with MongoDB.
- Efficient O&M

  DDS console is a visualized instance management platform. You can restart, back up, or restore an instance in just a few clicks.
- Backups and restorations

  DDS supports both automated and manual backup. The maximum retention period for an automated backup is 732 days. A manual backup can be retained until you delete it. You can restore a DB instance from a backup file. Replica sets support point-in-time recovery at the instance, database, and table-level.
- Three types of architectures

  DDS supports the following deployment architectures: cluster, replica set, and single node, meeting requirements of different service scenarios.

6.1.5.4 Architecture

DDS supports the following deployment modes.

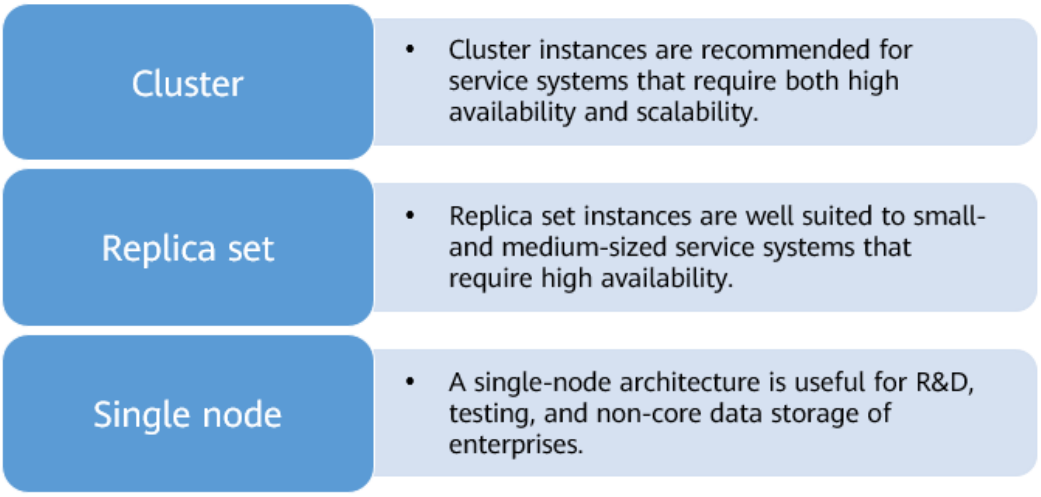


Figure 6-13 DDS architecture

- DDS provides sharded cluster instances comprised of a config node paired with multiple shards and mongos nodes.
- A replica set consists of three nodes: primary, secondary, and hidden. The three-node architecture is set up automatically, and the three nodes automatically synchronize data with each other to ensure data reliability.
- The single node architecture is a supplementary deployment mode that is useful for R&D, testing, and non-core data storage.

6.1.5.5 Application Scenarios

- Gaming

  DDS offers fast, reliable access to increasingly complex player profiles, including details such as character scores, items acquired and other details. For MMO games, the highly-available architecture of DDS clusters and replica sets can provide a smooth gaming experience even during peak hours.

  Schema-free and with MongoDB compatibility, DDS allows you to flexibly change table structures, so that you can continuously improve your game, always adding the coolest new features and staying competitive even in the fiercest markets. You can store structured data with a fixed schema in RDS, data with a flexible schema in DDS, and hot data in Distributed Cache Service (DCS) for Redis to efficiently access service data and reduce storage costs.

- IoT

  IoT applications feature high-concurrency writes, diverse data types, and sudden spikes in data volumes. With high performance and asynchronous data writes, DDS is able to process data as fast as in-memory databases when and where it is needed. In addition, the quantities and specifications of mongos and shard nodes in DDS cluster instances can be dynamically increased to meet growing demands, making DDS ideal for IoT applications. DDS also provides secondary indexes for dynamic queries and

uses a MongoDB-compatible map-reduce aggregation framework to perform the multidimensional data analysis needed by IoT applications.

- Internet

  Enterprise-class databases often need to process and store terabytes of data. Especially in big data scenarios, the databases need to handle writes in real time, support big data computing and analysis, and return analysis results. DDS replica sets use a three-node architecture to deliver reliability and enable disaster recovery. The three data nodes form an anti-affinity group and are deployed on different physical servers to automatically synchronize data. The primary and secondary nodes provide services. Each node has an independent private network address and works with the driver to distribute read load.

6.2 Security Services

6.2.1 Customer Requirements on Cloud Security

With more and more security risks, customers are increasingly concerned with security issues during cloud migration. They require:

- Assured service continuity

  Defense against network attackers and hackers as well as compliance with laws and regulations

- Controllable O&M

  Security policy configuration, risk detection and elimination, and auditable and traceable operations.

- Absolute data confidentiality

  Data breach prevention, making sure that data is accessible only to authorized staff.

6.2.2 Huawei Cloud Security Services

Addressing these requirements, Huawei Cloud provides customers with a series of high-quality security services.

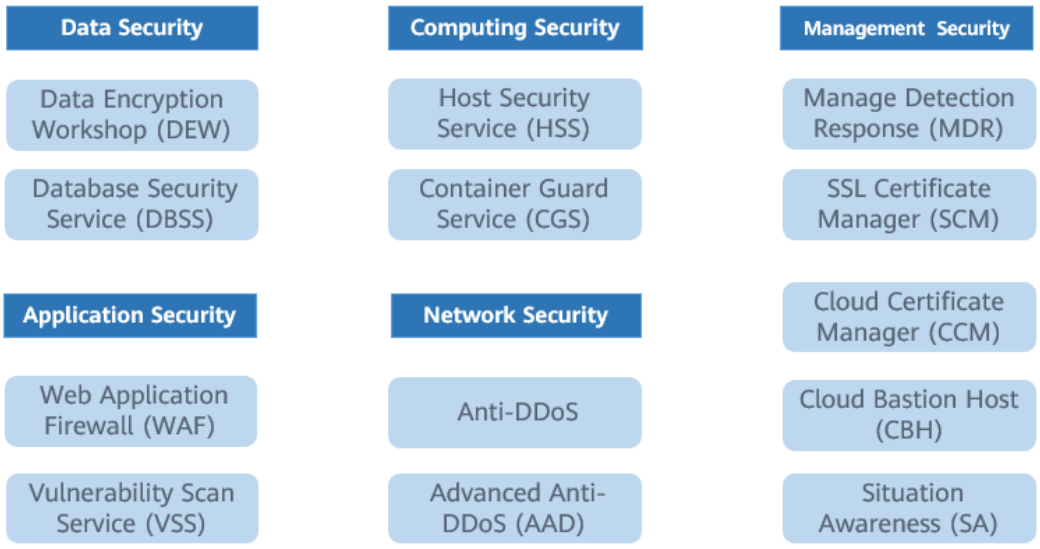


Figure 6-14 Huawei Cloud Security Services

This document describes five categories of security services: data, computing, management, application, and network security. As shown in Figure 6-14, each category has multiple cloud security services. Huawei Cloud has invested heavily in cloud security, assuring security for our customers.

6.2.3 HSS

6.2.3.1 What Is HSS?

The Host Security Service (HSS) improves host security. It provides functions such as account cracking prevention, weak password and malicious program detection, two-factor authentication, vulnerability management, and web tamper protection.

6.2.3.2 How It Works

After you install the HSS agent on your servers, you will be able to check the server security status and risks in a region on the HSS console.

Figure 6-15 shows how HSS works.

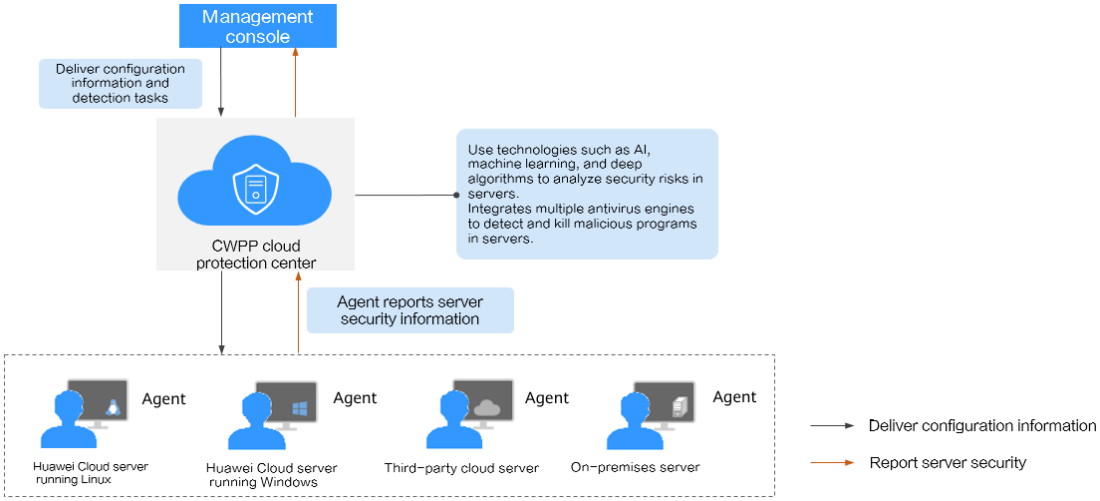


Figure 6-15 How HSS works

HSS components are described as follows.

- Console: a visualized management platform, where you can centrally apply configurations as well as view the defense status and scan results of servers in a region.
- HSS cloud protection center: This is the HSS server. It uses AI, machine learning, and in-depth algorithms to analyze security risks on hosts. The center integrates multiple antivirus engines to scan for and remove malicious programs on hosts; receives configuration information and detection tasks delivered by users on the console and forwards them to the agent installed on the server; receives host information reported by the agent, analyzes security risks and exceptions of hosts, and displays the analysis results in detection reports on the console.
- Agent: An agent is deployed on each host to communicate with the HSS cloud protection center through HTTPS and WSS, using port 443 by default. These scan all servers early every morning; monitor the security status of servers; and report the collected server information (including non-compliant configurations, insecure configurations, intrusion traces, software list, port list, and process list) to the cloud protection center.

Operating Process

1. Users deliver configuration information and detection tasks to the HSS cloud protection center through the management console.
2. The HSS cloud protection center receives these configurations and detection tasks, and then forwards them to the agent installed on the server.
3. Agents scan all servers, monitor their security status, and report the collected server information (including non-compliant configurations, insecure configurations, intrusion traces, software list, port list, and process list) to the cloud protection center.

4. The HSS cloud protection center receives server information reported by the agents, analyzes security risks and exceptions on servers, and displays the analysis results in detection reports on the console.

6.2.3.3 Advantages

HSS helps you manage and maintain the security of all your servers and reduce common risks.

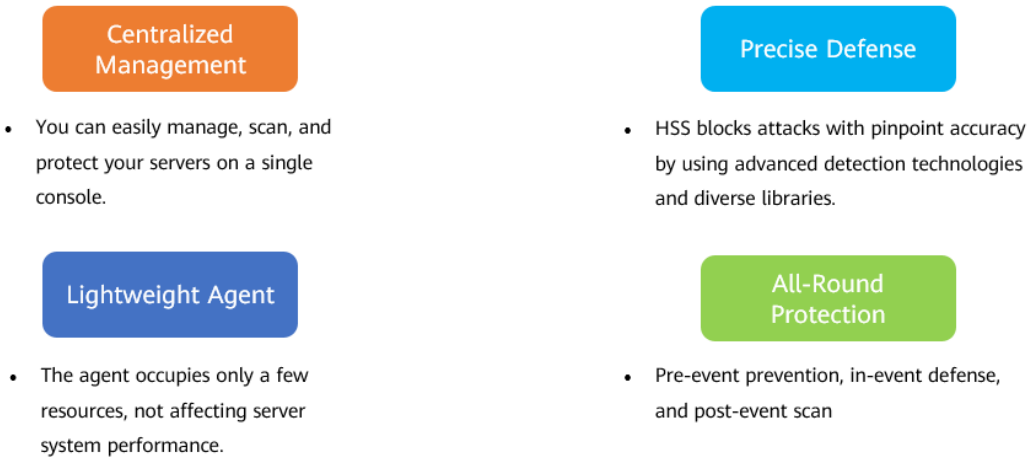


Figure 6-16 Advantages of HSS

HSS offers the following advantages:

- Centralized management

  It is easier to manage inspection and protection thanks to integrated management and control. You can install the agent on Huawei Cloud ECSs, BMSs, offline servers, and third-party cloud servers in the same region to manage all of them on a single console. On the security console, you can view the sources of server risks in a region, handle them according to displayed suggestions, and use filter, search, and batch processing functions to quickly analyze all regional risks.

- Accurate defense

  HSS blocks attacks with pinpoint accuracy by using advanced detection technologies and diverse libraries.

- Comprehensive protection

  HSS protects servers against intrusions through prevention, defense, and post-intrusion scans.

- Lightweight agent

  The agent occupies only a few resources, not affecting server system performance.

6.2.3.4 Application Scenarios

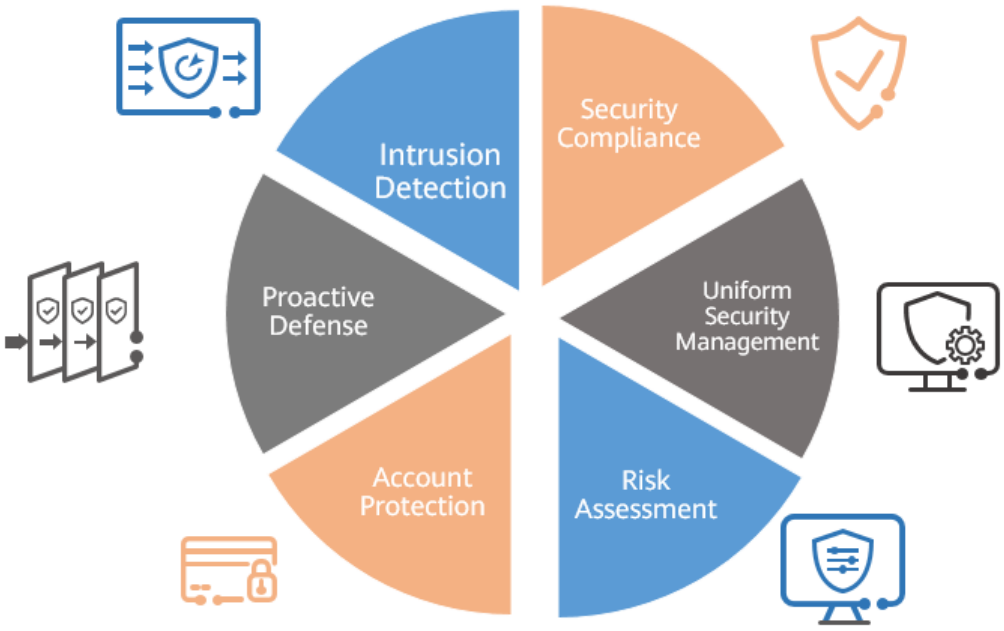


Figure 6-17 HSS applications

HSS is suitable for the following scenarios.

- Security compliance: HSS protects accounts and systems on cloud servers, helping enterprises meet compliance standards.
- Uniform security management: You can manage servers, security configurations, and security events all on the HSS console, reducing security risks and management costs.
- Risk assessment: HSS scans your servers for risks, including unsafe accounts, ports, software vulnerabilities, and weak passwords, and sends you timely prompts to eliminate security risks and harden the system.
- Account protection: Accounts are protected before, during, and after a security event. You can use 2FA to block brute-force attacks on accounts, enhancing the security of your cloud servers.
- Proactive defense: You can count and scan your server assets, check and fix vulnerabilities and unsafe settings, and proactively protect your network, applications, and files from attacks.
- Intrusion detection: You can scan all possible attack vectors to detect and fight APTs and other threats in real time, preventing system impact.

6.2.4 WAF

6.2.4.1 What Is WAF?

A Web Application Firewall (WAF) keeps web services stable and secure. It examines all HTTP and HTTPS requests to detect and block the following types of attacks: Structured Query Language (SQL) injection, cross-site scripting (XSS), web shells, command and code injections, file inclusion, sensitive file access, third-party vulnerability exploits, Challenge Collapsar (CC) attacks, malicious crawlers, and cross-site request forgery (CSRF).

6.2.4.2 Architecture

WAF offers several deployment modes for different service scenarios.

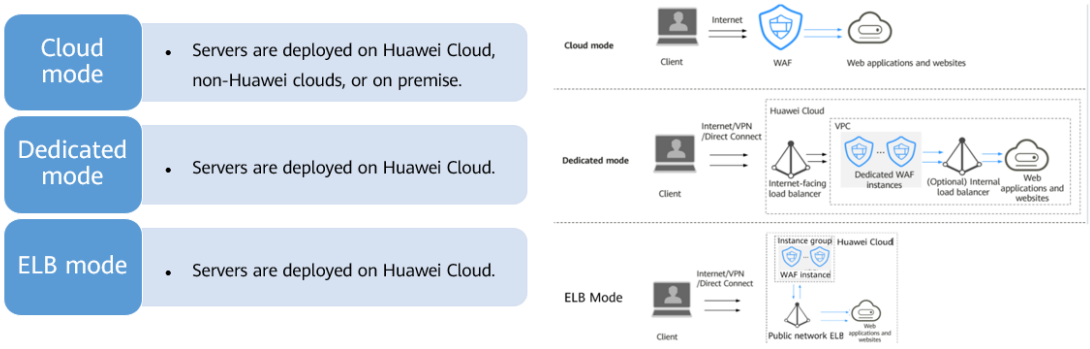


Figure 6-18 WAF Advantages

- Cloud
  - Protected objects: domains
  - Cloud WAF offers high scalability. A cloud WAF instance can protect web services deployed on Huawei Cloud, other cloud platforms, and on-premises data centers. It can also protect websites using IPv6 addresses.
- Dedicated
  - Protected objects: domains or IP addresses
  - Resources are exclusively used by users, meeting the protection requirements of large-scale traffic attacks with low latency.
- ELB
  - Protected objects: domains or IP addresses
  - When WAF instances are deployed out-of-line, there is zero impact on protected website services. If your WAF instance becomes faulty, the load balancer directly distributes your website traffic over the origin servers, preventing adverse impact on your normal business.

6.2.4.3 How It Works

After a website is connected to WAF, all website access requests are forwarded to WAF first. Then, WAF inspects the traffic, filters out malicious traffic, and routes only normal traffic to the origin server, keeping the origin server secure, stable, and available.

The process of forwarding traffic to the origin server through WAF is called back-to-source. WAF inspects traffic originating from the client and uses WAF back-to-source IP addresses to forward normal traffic to the origin server. The origin server interprets the source IP addresses of all requests as WAF back-to-source IP addresses. This hides the IP address of the origin server from the client.

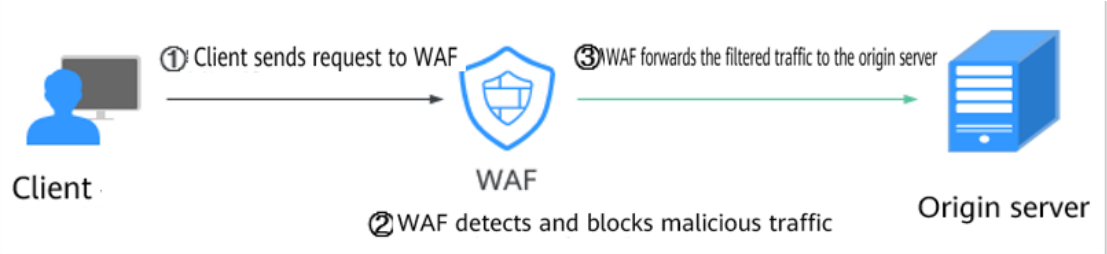


Figure 6-19 How WAF protects a website

6.2.4.4 Advantages

WAF examines web traffic from multiple dimensions to accurately identify malicious requests and filter attacks, reducing the risks of data being tampered with or stolen. The following figure describes the main advantages of WAF.

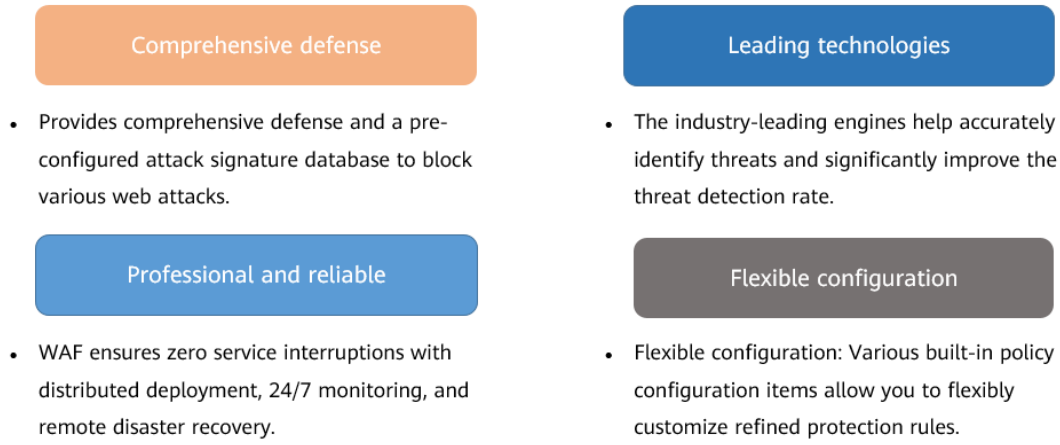


Figure 6-20 WAF advantages

6.2.4.5 Application Scenarios

WAF is widely used. Below, are some typical WAF scenarios.

- Basic protection: WAF helps you defend against common web attacks, such as command injection and sensitive file access.
- Promotions on e-commerce platforms: During online promotions, there may be many malicious requests sent to service interfaces. WAF enables customizable rate limiting rules to defend against CC attacks. This prevents services from breakdowns caused by too many concurrent requests while ensuring responses to legitimate ones.
- Defense against zero-day vulnerability: If website services fail to quickly recover from the impact of zero-day vulnerabilities in third-party web frameworks or plug-ins, WAF will update the preset protection rules immediately to ensure service security and stability. WAF functions as extra protection on a third-party network. Using WAF rules to block risks is quicker than directly fixing vulnerabilities on third-party architecture.
- Data leak prevention: WAF prevents malicious actors from using methods such as SQL injection and web shells to bypass application security and gain remote access to web databases and other sensitive information. You can customize WAF data masking rules for:
  - Precise identification

    WAF uses semantic analysis & regex to examine traffic from different dimensions, precisely detecting malicious traffic.
  - Distortion attack detection

    WAF detects a wide range of distortion attack patterns with 7 decoding methods to prevent bypass attempts.
- Web tamper prevention: WAF ensures that attackers cannot leave backdoors on protected web servers or tamper with web page content, preventing damage to customer credibility. You can configure web tamper protection rules on WAF to provide the following functions:
  - Website malicious code detection

    You can configure WAF to detect malicious code injected into web servers and ensure secure visits to web pages.
  - Web page tampering prevention

    Prevents attackers from tampering with or changing web page content, or publishing inappropriate information that can damage your reputation.

6.2.5 DEW

6.2.5.1 What Is DEW?

Data is an enterprise's core asset. Each enterprise has its core sensitive data, which needs to be encrypted and protected from breaches.

The Data Encryption Workshop (DEW) is a cloud data encryption service. It provides Key Management Service (KMS), Key Pair Service (KPS), and Dedicated Hardware Security Module (HSM). DEW uses HSMs to protect your keys, and can be integrated with other

Huawei Cloud services to address data security, key security, and key management. You can also develop your own encryption applications based on DEW.

6.2.5.2 DEW Services

DEW offers four main services: key management, credential management, key pair management, and Dedicated HSM.

- KMS

  KMS is a secure, reliable, and easy-to-use cloud service that helps users centrally create, manage, and protect keys. KMS uses HSMs to protect keys, helping you easily create and control customer master keys (CMKs). All CMKs are protected by root keys in HSMs to avoid key leakage.

- CSMS

  The Cloud Secret Management Service (CSMS) is a secure, reliable, and easy-to-use credential hosting service. Users or applications can use CSMS to centrally create, retrieve, update, and delete credentials throughout the credential lifecycle. CSMS can help you eliminate risks incurred by hardcoding, plaintext configuration, and permission abuse.

- KPS

  KPS is a secure, reliable, and easy-to-use cloud service designed to manage and protect your SSH key pairs (key pairs for short).

  KPS uses HSMs to generate random true numbers which are then used to produce key pairs. In addition, it adopts a complete and reliable key pair management solution to help users easily create, import, and manage key pairs. The public key of a generated key pair is stored in KPS while the private key can be downloaded and saved separately, which ensures the privacy and security of the key pair.

- Dedicated HSM

  Dedicated HSM is a cloud service used for encryption, decryption, signature, signature verification, key generation, and the secure storage of keys.

  Dedicated HSM provides encryption hardware certified by the China State Cryptography Administration (CSCA), guaranteeing data security and integrity on Elastic Cloud Servers (ECSs) and meeting compliance requirements. Dedicated HSM securely and reliably manages the keys generated by your instances, and uses multiple algorithms to encrypt and decrypt data.

6.2.5.3 Advantages

6.2.5.3.1 Advantages of KMS

- Extensive service integration

  KMS can be integrated with Object Storage Service (OBS), Elastic Volume Service (EVS), and Image Management Service (IMS), to manage keys of these services on the KMS console, and encrypt and decrypt your local data by making KMS API calls.

- Regulatory compliance

  Keys are generated by third-party validated HSMs. Access to keys is controlled and all operations involving keys are traceable by logs, and therefore compliant with Chinese and international laws and regulations.

6.2.5.3.2 Advantages of CSMS

- Secret encryption

  Secrets are encrypted by KMS before storage. Encryption keys are generated and protected by authenticated third-party Hardware Security Modules (HSMs). When you retrieve secrets, they are transferred to local servers via TLS.

- Secure Secret Retrieval

  CSMS calls secret APIs instead of hard-coded secrets in applications. Secrets can be dynamically retrieved and managed. CSMS centrally manages application secrets to reduce breach risks.

- Dual secret rotation

  Version management enables secret rotation at the application layer. Encryption keys are also automatically rotated to improve security.

- Centralized secret management and control

  IAM ensures that only authorized users can retrieve and modify credentials. CTS monitors access to credentials. These services prevent unauthorized access to and breach of sensitive information.

6.2.5.3.3 Advantages of KPS

- Reinforced login security

  Password-free log in to Linux ECSs prevents password interception and cracking, improving the security of Linux ECSs.

- Regulatory compliance

  Random numbers are generated by third-party validated HSMs. Access to key pairs is controlled and all operations involving key pairs are traceable by logs, and therefore compliant with Chinese and international laws and regulations.

6.2.5.3.4 Advantages of Dedicated HSM

- On-cloud protection: Dedicated HSM can transfer offline encryption capabilities to the cloud, reducing your O&M costs.
- Scalability: You can flexibly increase or decrease the number of HSM instances according to your service needs.
- Security management: Dedicated HSM separates device management from consent management (sensitive information). As a device user, you can control key generation, storage, and access. Dedicated HSM is only responsible for monitoring and managing devices and related network facilities. Even the O&M personnel of Dedicated HSM cannot obtain your keys.
- Permission authentication: Sensitive instructions are classified for hierarchical authorization, which prevents unauthorized access. Several authentication types are supported, such as username/password and digital certificate.

- Reliability: Dedicated HSM provides level-3 HSMs certified by CSCA and validated by FIPS 140-2 to protect your keys, guaranteeing high-performance encryption services to meet your stringent security requirements. Dedicated HSM chips are exclusively used by each instance. Even if some hardware chips are damaged, services will not be affected.
- Security compliance: Dedicated HSM provides HSM instances validated by CSCA, helping you protect your data on ECSs and meet compliance requirements.
- Wide application: Dedicated HSM offers finance, server, and signature server HSM instances for use in various service scenarios.

6.2.5.4 Application Scenarios

6.2.5.4.1 KMS Application Scenarios

- Small data encryption and decryption

  You can use the online tool on the KMS console or call the KMS APIs to directly encrypt or decrypt a small-size data, such as passwords, certificates, or phone numbers. Currently, a maximum of 4 KB of data can be encrypted or decrypted in this way.
- Encryption and decryption of large-size data

  If you want to encrypt or decrypt large volumes of data, such as pictures, videos, and database files, you can use the envelope encryption method, where the data does not need to be transferred over the network.

6.2.5.4.2 KPS Application Scenarios

When purchasing an ECS, you can either use the SSH key pair provided by KPS to authenticate a user to log in to the ECS, or use the provided key pair to obtain the password to log in to a Windows ECS.

- Logging in to a Linux ECS

  If you purchase a Linux ECS, you can select the key pair login option to log in. There are two types of key pairs:
  - Ones created on or imported to the ECS console
  - Ones created on or imported to the KPS console
- Obtaining the password for logging in to a Windows ECS

  If you purchase a Windows ECS, you need to use the private key of a key pair to obtain the login password. There are two types of key pairs:
  - Ones created on or imported to the ECS console
  - Ones created on or imported to the KPS console

6.2.5.4.3 Dedicated HSM Application Scenarios

If you purchase a Dedicated HSM instance, you can use the provided UKey to initialize and manage the instance. You can fully control key generation, storage, and access authentication.

You can use Dedicated HSM to encrypt your service systems (including encryption of sensitive data, payment, and electronic tickets). Dedicated HSM helps you encrypt sensitive enterprise data (such as contracts, transactions, and SNs) and sensitive user data (such as user IDs and mobile numbers). This prevents hackers from cracking the network and dragging the database, which may cause data leakage, as well as illegal access to or tampering with data by internal users.

- Sensitive data encryption

  It is used for government public services, Internet enterprises, and system applications that contain immense sensitive information

  Data is an enterprise's core asset. Each enterprise has its core sensitive data. Dedicated HSM provides integrity checks and encrypted storage for sensitive data, preventing data theft or tempering as well as unauthorized access.

- Financial payments

  It is used for system applications for payment and prepayment with transportation card, on e-commerce platforms, and through other means

  Dedicated HSM can ensure the integrity and confidentiality of payment data during transmission and storage, and ensure payer identity authentication and the non-repudiation of the payment process.

6.2.6 IAM

6.2.6.1 What Is IAM?

Identity and Access Management (IAM) enables you to easily manage users and control their access to Huawei Cloud services and resources. IAM is free of charge.

6.2.6.2 Advantages

Figure 6-21 IAM advantages

IAM has the following advantages:

- Fine-grained access control for Huawei Cloud resources

  An account is created after you successfully register with Huawei Cloud. Your account has full access permissions for your cloud services and resources and makes payments for the use of these resources.

  If you purchase multiple resources on Huawei Cloud, such as ECSs, EVS disks, and BMSs, for different teams or applications in your enterprise, you can create IAM users for the team members or applications and grant them permissions required to complete tasks. The IAM users use their own usernames and passwords to log in to Huawei Cloud and access resources in your account.

- Cross-account resource access delegation

  If you purchase multiple resources on Huawei Cloud, you can delegate another account to manage specific resources for efficient O&M.

  For example, you create an agency for a professional O&M company to manage specific resources with the company's own account. You can cancel or modify the delegated permissions at any time if the delegation changes.

- Federated access with existing enterprise accounts

  If your enterprise has an identity system, you can create an identity provider in IAM to provide single sign-on (SSO) access to Huawei Cloud for your employees. The identity provider establishes a trust relationship between your enterprise and Huawei Cloud, allowing the employees to access Huawei Cloud using their existing accounts.

6.3 CDN

6.3.1 What Is CDN?

Content Delivery Network (CDN) is an intelligent virtual network running over the Internet. It distributes content from origin servers to nodes around the world so that users can access desired content faster by accessing nodes nearby. CDN speeds up site response and improves site availability, breaking through the bottlenecks caused by low bandwidth, heavy user access traffic, and uneven distribution of nodes.

6.3.2 Advantages

Huawei CDN has the following advantages:

- Global presence

  Huawei Cloud CDN has over 2,000 edge nodes in the Chinese mainland and over 500 edge nodes outside the Chinese mainland. The network-wide bandwidth is at least 150 Tbit/s. The edge nodes are connected to the networks of top carriers in China such as China Telecom, China Unicom, China Mobile, and China Education and Research Network (CERNET), as well as many small- and medium-sized carriers. CDN precisely schedules user requests to the most appropriate edge nodes, providing efficient and reliable acceleration.

- Intelligent scheduling

  The IP address database allows CDN to deliver up to 99.99% of a scheduling success rate. Net Turbo technology schedules user requests to the best quality nodes based on node load.

- Security

  Huawei Cloud CDN provides secure and reliable content delivery services. It supports advanced network security functions, such as data transmission over HTTPS and hotlink protection throughout the entire network.

- Easy operations

  You can simply and quickly access domain names to Huawei Cloud CDN. You can customize configuration items including hotlink protection, cache policy, and HTTPS certificates for domain names, and easily analyze statistics and manage logs.

- Diverse applications

  Huawei Cloud CDN can speed up the delivery of content like a web page, an entire website, a live stream, or a large file for download. It provides one-stop acceleration solutions for a wide range of scenarios, improving the overall user experience.

6.3.3 How It Works

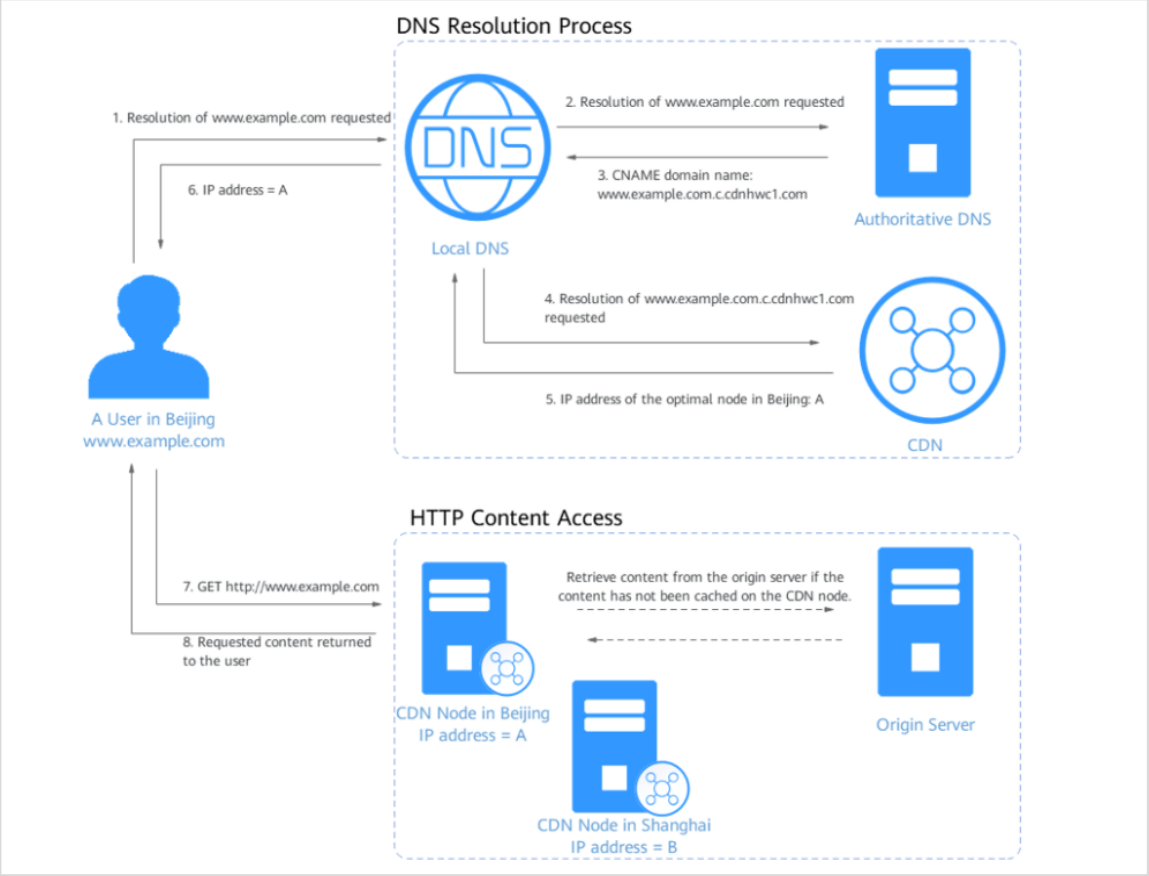


Figure 6-22 Request process

HTTP requests are processed as follows:

1. A user enters the domain name of a website to be accessed (for example, www.example.com) in the browser. A DNS request is sent to the local DNS server.
2. The local DNS checks whether its cache includes the IP address of www.example.com. If yes, the local DNS directly returns the cached information to the user. If no, the local DNS sends a resolution request to the authoritative DNS.
3. The authoritative DNS resolves the domain name and finds that the domain name points to www.example.com.c.cdnhwc1.com (CNAME record of the domain name).
4. The request is directed to the CDN service.
5. CDN performs intelligent domain resolution and provides the user with the IP address of the Beijing CDN node, which responds the fastest.
6. The user's browser obtains the IP address of the Beijing CDN node.
7. The user's browser sends the access request to this CDN node.
   - If this CDN node has cached the content, it sends the desired resource directly to the user and ends the request.
   - If this CDN node has not cached the content, it retrieves the content from the origin server. The retrieved content is cached on this CDN node based on custom cache policies. Then, the node sends the desired content to the user and ends the request.

6.3.4 Application Scenarios

CDN is used in the following acceleration scenarios:

- Website acceleration

  CDN is perfect for web portals, e-commerce platforms, news apps, and user generated content (UGC)-focused apps. It provides excellent acceleration for static content associated with an acceleration domain name. In addition, it supports custom cache policies. You can set the maximum cache age as needed. The files that can be cached include but are not limited to **.zip**, **.exe**, **.wmv**, **.gif**, **.png**, **.bmp**, **.wma, .rar**, **.jpeg**, and **.jpg**.

- File download

  CDN is useful for download clients, game clients, app stores, and websites that provide download services based on HTTP or HTTPS. An increasing number of new services, such as apps and mobile games, require software updates in real time. Conventional download services need to provide even more and larger downloads. If origin servers have to handle all of these requests directly, it places tremendous strain on these servers and results in a significant bottleneck. With CDN download acceleration, content to be downloaded is distributed to edge nodes, easing the pressure on origin servers and ensuring high-speed downloads.

- On-demand service acceleration

  If you provide on-demand audiovisual services, CDN is a must. On-demand services include online education, video sharing, music or video on demand, and other audiovisual content. Conventional on-demand audiovisual content puts significant load on servers and consumes an enormous amount of bandwidth, affecting user

experience. CDN ensures fast, reliable, secure acceleration for such services by delivering content to all CDN nodes. Users are then able to obtain that content from nearby nodes anywhere, anytime.

- Whole site acceleration

  CDN is perfect for websites that consist of both dynamic and static content and for sites that involve a large number of ASP, JSP, or PHP requests. CDN's whole site acceleration accelerates both dynamic and static content. Static content can be accessed from nearby nodes, while dynamic content is retrieved from origin servers through the optimal route. As such, dynamic pages can be loaded more quickly by bypassing congested routes.

6.4 EI Services

6.4.1 AI and Big Data

Huawei Cloud provides AI and big data cloud services to facilitate the intelligent upgrades of enterprises and build ubiquitous and pervasive AI.

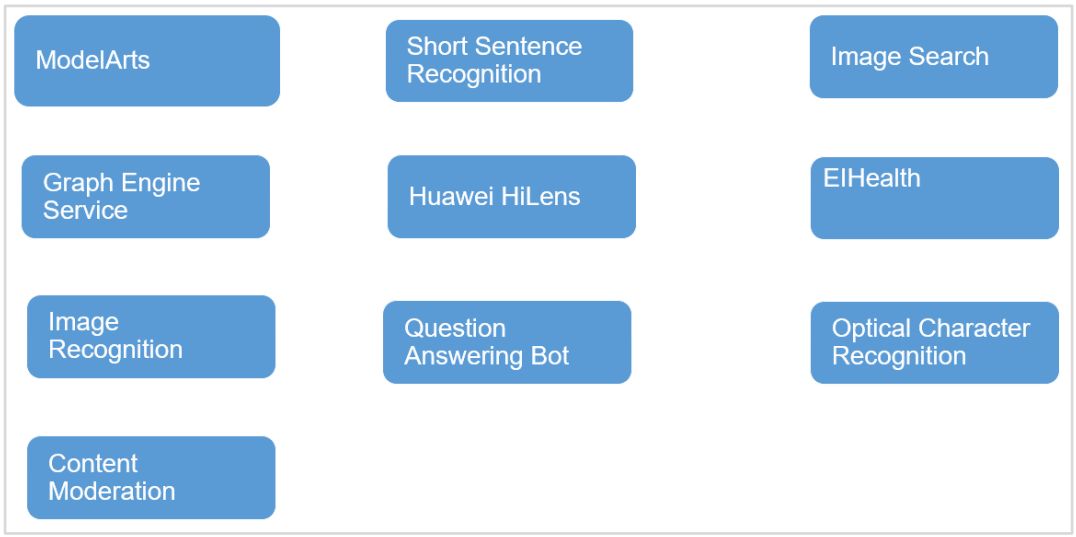


Figure 6-23 AI services

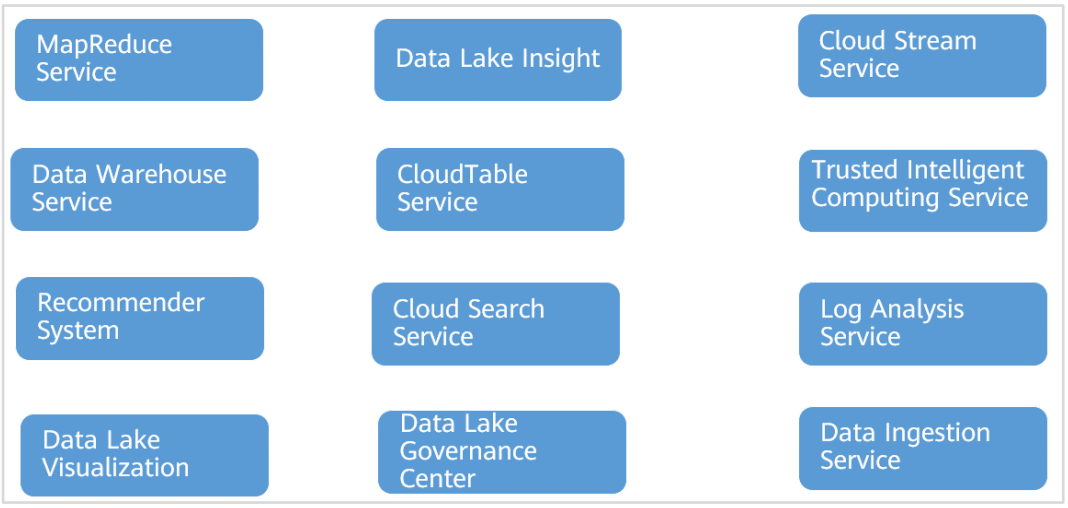


Figure 6-24 Big data services

6.4.2 ModelArts

ModelArts provides data preprocessing, semi-automated data labeling, distributed training, automated model building, and model deployment services on the device, edge, and cloud, helping AI developers build models quickly and manage lifecycles.

Figure 6-25 One-stop development platform ModelArts

The one-stop ModelArts platform covers all stages of AI development, including data processing, algorithm development, and model training and deployment. The underlying technologies support various heterogeneous computing resources, allowing developers to flexibly select and use resources. ModelArts supports mainstream open-source AI development frameworks such as TensorFlow and MXNet. Developers can use self-developed algorithm frameworks tailored to their usage habits.

ModelArts aims to simplify AI development, and is suitable for developers of any skill level. For example, service developers can use ExeML to quickly build applications without coding, AI beginners can use built-in algorithms to create applications, and AI engineers can use multiple development environments to quickly compile code for modeling and application development.

6.4.3 FusionInsight Intelligent Data Lake

As the data foundation of Huawei Cloud data enablement solutions, Huawei FusionInsight provides a cloud-native big data solution with converged lakes and warehouses for enterprises.

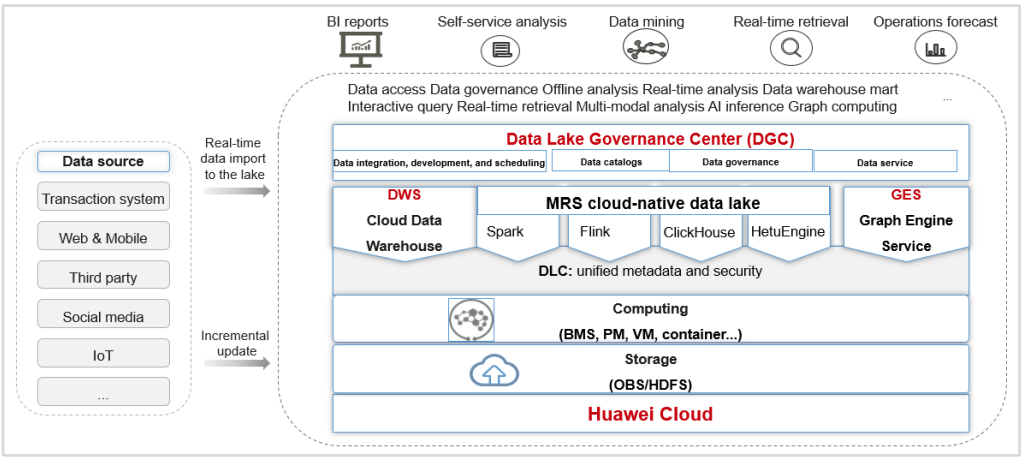


Figure 6-26 FusionInsight intelligent data lake

Highlights:

FusionInsight offers full-stack cost-effectiveness. Huawei Kunpeng chips have multiple cores and are naturally advantageous in providing stronger computing power, higher integration, and larger network bandwidth. Many typical algorithms such as compression, encryption, and decryption are preset in Kunpeng chips. Huawei Cloud's next-generation intelligent data lake service optimizes the full-stack architecture based on Kunpeng chips and accelerates data scanning through vectorized processing. In addition, lock hierarchy reduces the random competition for critical resources accessed by multiple threads, and uses Kunpeng multi-core chips to fully utilize the data lake service in computing- or I/O-intensive scenarios.

7 Huawei Cloud O&M Basics

Huawei Cloud not only provides resource services to meet enterprise needs to migrate their service systems to the cloud, but also ensures the normal running of the service systems on the cloud to meet the enterprise governance requirements.

This section will help you understand Huawei Cloud O&M.

7.1 O&M Key Concepts and Principles

7.1.1 O&M Key Concepts

This section describes cloud O&M. Before learning how to operate and maintain services on the cloud, we need to understand some basic concepts of O&M.

O&M refers to operations and maintenance. In the ICT industry, those who perform O&M operations are typically referred to as O&M engineers. O&M personnel are responsible for planning information, networks, and services based on service requirements and ensuring the long-term stability and availability of services by using various means, including but not limited to the following:

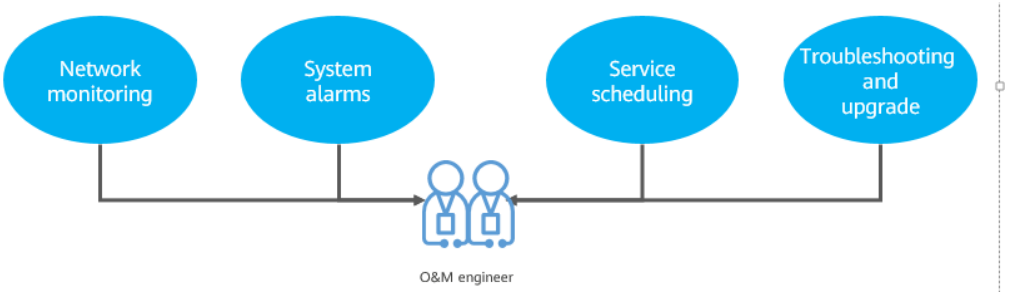


Figure 7-1 Responsibilities of O&M personnel

The following figure shows the classification of O&M position.

Figure 7-2 O&M position classification

As the number and complexity of devices, operating systems, and applications deployed in ICT data centers increase, enterprises need more robust O&M capabilities. They need more specialized O&M. Common categories include:

- Hardware O&M:
  - Equipment room planning (including equipment room location selection, network deployment, and server deployment planning)
  - Network system maintenance (including network adjustment, capacity expansion, bandwidth monitoring, and network QoS)
  - Server management (procurement, receipt, deployment in cabinets, system installation, delivery, and maintenance)
- System O&M:

  O&M based on OS usage includes system optimization and performance monitoring
- Database O&M:

  Software installation, configuration optimization, backup policy selection and implementation, data restoration, data migration, troubleshooting, preventive inspection, and other services for user databases
- Application O&M: mainly refers to the O&M of user services to ensure the stability of services during continuous iteration.
  - Change management, ensuring system stability in the process of continuous iteration
  - Fault management, including application monitoring, fault locating, fault rectification, and application optimization
  - Resource management ensuring the application system runs properly with optimal resource allocation, and evaluation of whether capacity expansion is required for future service needs.
  - In most cases, system O&M and application O&M are combined, because the stability of applications depends on the stability of the system.
- Network O&M:

  A series of management activities to ensure normal, secure, and effective operation of enterprise networks and services. This process is often called O&M management, but is also referred to as operation, administration, and maintenance (OAM). Network O&M is about maintaining and ensuring high availability for the entire service system, and continuously optimizing the system architecture to improve deployment efficiency.

7.1.2 O&M Principles

We have learned what the various O&M positions are, but how should ICT O&M personnel go about their jobs? O&M seems simple at first, but to better serve enterprise business systems, we need to first understand the main principles governing ICT O&M.

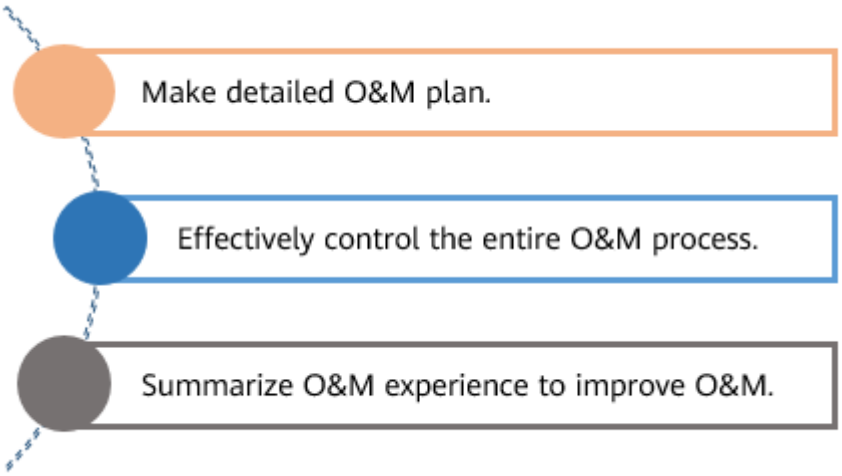


Figure 7-3 O&M principles

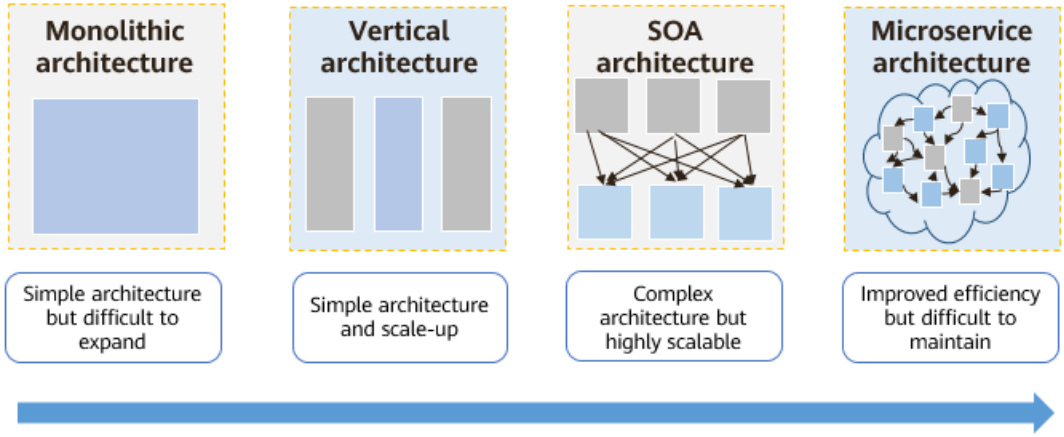


Figure 7-4 Enterprise architecture development trends

As IT architecture evolves, it tends to become more and more complex. In an enterprise, development and O&M are usually two independent departments with different work objectives and technical directions. When a project needs to be completed by the two departments, their communication is often not smooth. Communication issues can slow down the application development progress, and greatly reduce enterprise efficiency. The entire system architecture needs to evolve continuously, moving from traditional O&M to automated O&M, so it can break down the barriers between O&M engineers, development engineers, and quality assurance engineers, and produce a more efficient system.

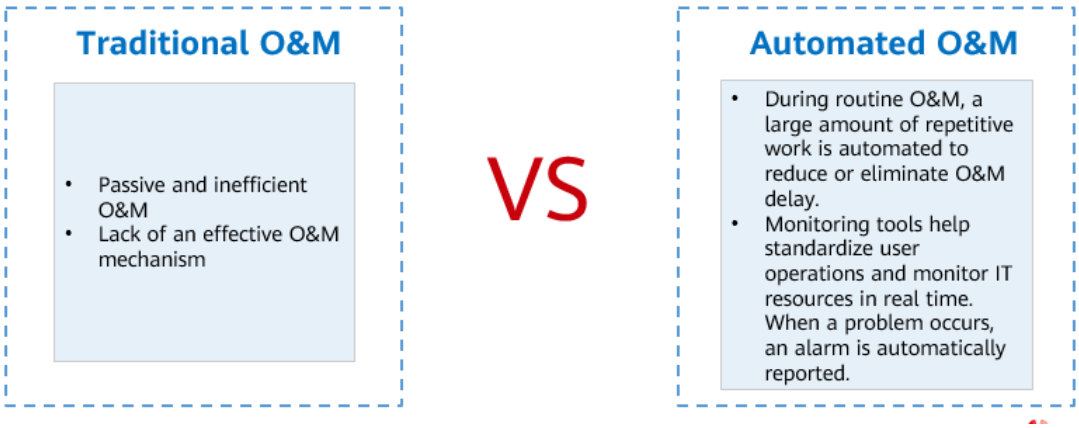


Figure 7-5 Automated O&M

After more than a decade of development, IT operations is now facing a new direction: automation. Automation is the inevitable result of IT development. Nowadays, the complexity of IT systems calls for O&M that is digital and automated. Automated O&M refers to the automation of repetitive daily IT tasks. IT operations are evolving from manual work to automation. IT operations automation is not only a maintenance process, but also a management improvement process. It is the highest level of IT operations and also the development trend in the future.

DevOps is a group of processes, methods, and systems used to promote communication, collaboration, and integration between development, technical operation (O&M), and quality assurance (QA) departments. DevOps greatly reduces the gap between O&M and development, which means much faster delivery.

The following are three major O&M services on Huawei Cloud.

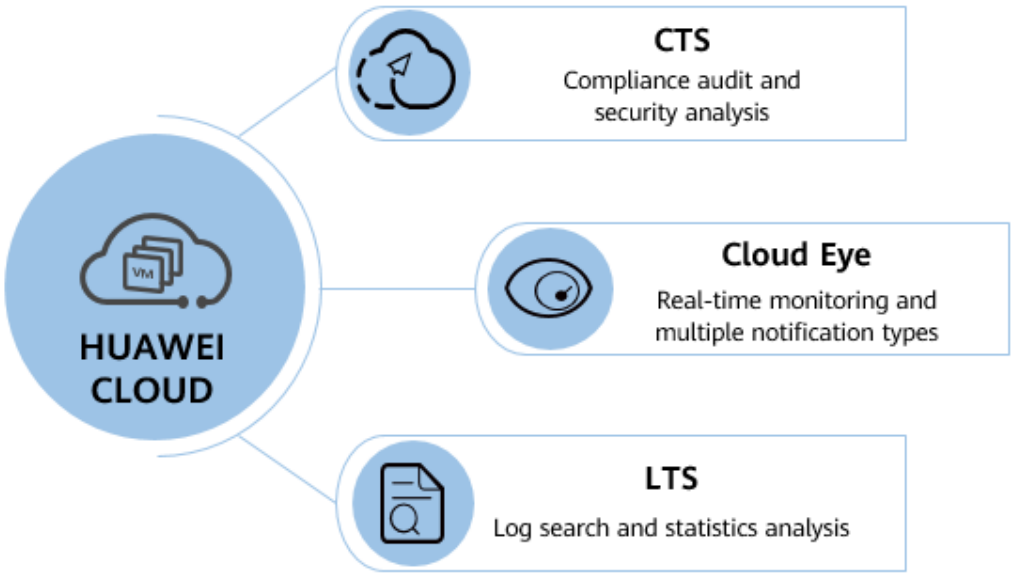


Figure 7-6 Overview of cloud O&M services

7.2 CTS

7.2.1 What Is CTS?

Before getting started with Cloud Trace Service (CTS), let's first look at auditing.

Auditing is the process of gathering and analyzing evidence to evaluate an enterprise's financial statements, drawing conclusions and producing reports on their compliance with generally accepted standards, and communicating the results to stakeholders. An audit in the information and communications technology (ICT) industry is mainly an examination of the entire lifecycle of information systems.

Auditing enterprises will usually compare enterprise financial statements with actual operations. It checks whether the information presented in an enterprise's financial statements is fair and accurate, helping the enterprise operate properly. In the ICT industry, audits usually aim to examine whether information systems are running as they should.

Log audits are the core of information security audits. They are essential for the security risk control of information systems in both private and public sectors. As information systems migrate to cloud, several information and data security management departments around the world, including the Standardization Administration of the People's Republic of China/Technical Committee (SAC/TC), have released multiple standards. These include ISO/IEC 27000, GB/T 20945-2013, COSO, COBIT, ITIL, and NIST SP 800.

CTS is Huawei Cloud's log audit service, which keeps track of user activities and resource changes on your cloud resources. It helps you collect, store, and query operational records for security analysis, audit and compliance, and fault location.

7.2.2 Advantages

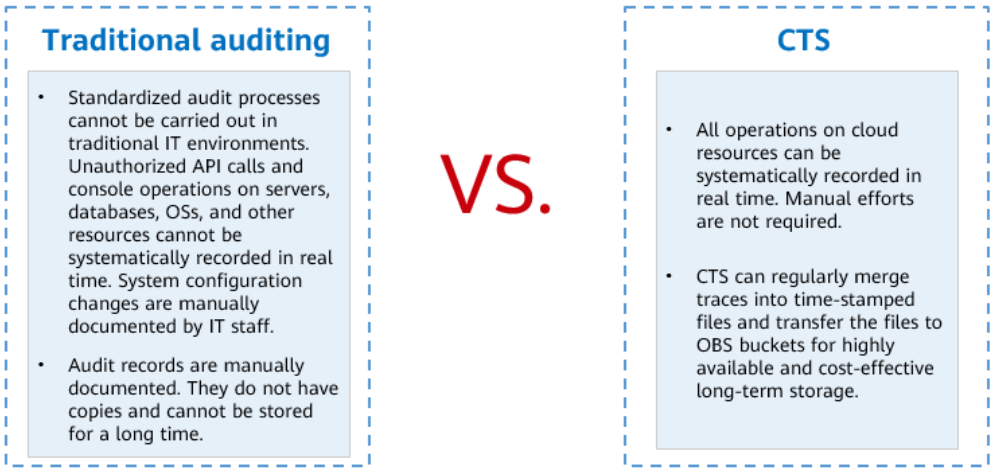


Figure 7-7 CTS Advantages

7.2.3 Architecture

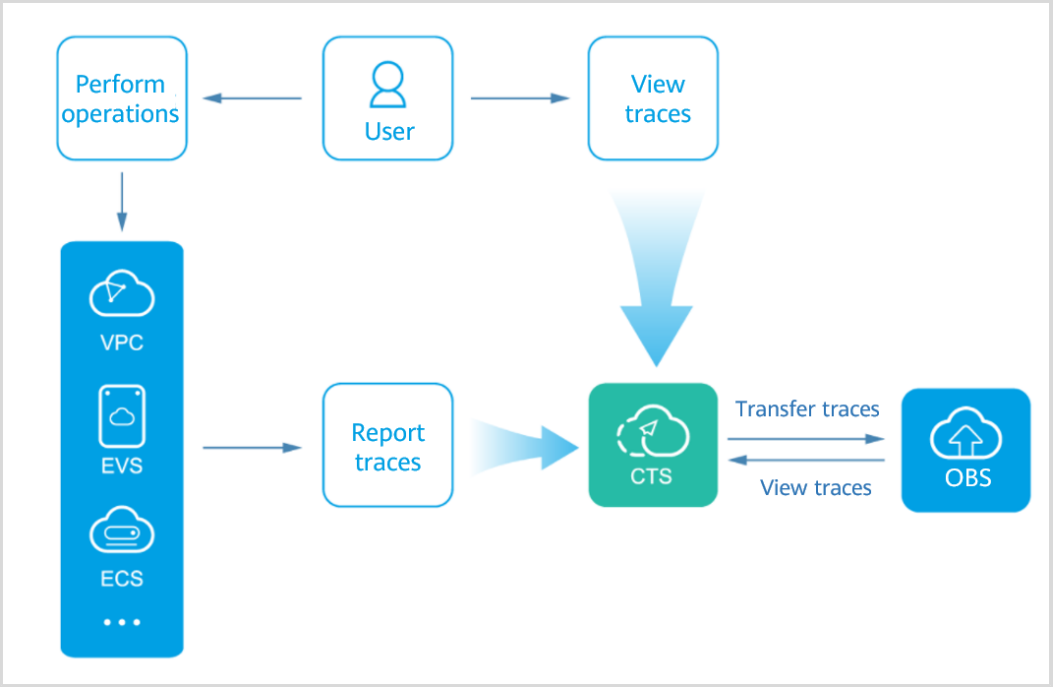


Figure 7-8 CTS architecture

CTS provides the functions detailed below.

- Trace collection: CTS records operations performed on the console, API calls, and system-triggered actions.
- Trace query: On the CTS console, you can query the traces of the last seven days using multiple filters, such as trace type, trace source, resource type, user, and trace status.
- Trace transfer: Traces can be periodically transferred to OBS buckets for long-term storage. During transfer, traces are merged into trace files corresponding to specific services.
- Trace file encryption: Trace files can be encrypted using keys provided by the Data Encryption Workshop (DEW) during transfer.

7.2.4 Key Concepts

To understand the architecture, advantages, and process of CTS, you need to understand the following concepts:

- Tracker: You need to enable CTS before using it. A tracker is automatically created when CTS is enabled. The tracker automatically identifies all cloud services you are using and records all operations performed on the services.
- Trace: Traces are operation records captured and stored by CTS. They help you identify when a particular operation was performed by a specific user on a given resource. There are two types of traces.
  - Management traces: traces reported by cloud services.

- Data traces: traces of read and write operations reported by OBS.

7.2.5 Application Scenarios

CTS can be used in the following four scenarios.

- Compliance auditing

  CTS helps you obtain certifications for auditing in industry standards, such as classified protection of cybersecurity, PCI DSS, and ISO 27001, for your service systems. If you want to migrate your services to the cloud, you will need to ensure the compliance of your own service systems and the ones used by the cloud vendor you choose. CTS plays an important role in Huawei Cloud's own compliance. The service records operations of almost all Huawei Cloud services and resources, and carries out security measures such as encryption, disaster recovery, and anti-tampering to ensure the integrity of traces during their transmission and storage. In addition, you can use CTS to design and implement solutions that help you obtain compliance certifications for your service systems.

- Key event notifications

  CTS works with FunctionGraph to send notifications to natural persons or service APIs when any key operation is performed.

  You can configure HTTP or HTTPS notifications targeted at your independent systems and synchronize traces received by CTS to your own audit systems for auditing.

  You can also select a certain type of log (such as file upload) as a trigger for a preset workflow (for example, file format conversion) in FunctionGraph, simplifying service deployment and O&M as well as avoiding risks.

- Data mining

  CTS mines data in traces to facilitate service health analysis, risk analysis, resource tracking, and cost analysis. You can also obtain the data from CTS and explore its value yourself. A trace contains 19 fields, providing information such as when an operation was performed by a specific user on a specific resource and from which IP address.

  By configuring HTTP or HTTPS notifications, you can synchronize traces to your own system for analysis. In addition, CTS is connected to Cloud Eye and Log Tank Service (LTS) to help you monitor high-risk operations, detect unauthorized operations, and analyze resource usage, service health, and cost.

- Fault locating and analysis

  If a fault occurs, CTS allows you to use filters to quickly search for unusual operations. This accelerates troubleshooting and reduces staffing requirements. You can search by filters such as trace source, resource type, operator, and trace status. Each trace contains the request and response of an operation. Querying traces is one of the most efficient ways to locate a fault. If an issue occurs when you use cloud services, you can set filters to search for suspicious operations in a specified time period, and send the related traces to customer service or O&M engineers to handle the issue.

7.2.6 How to Use CTS

7.2.6.1 Security Analysis

Each trace records details about an operation. You can identify when an operation was performed by a specific user and from which IP address. Based on traces, you can also perform security and user behavior pattern analysis as well as configuring notifications for key operations.

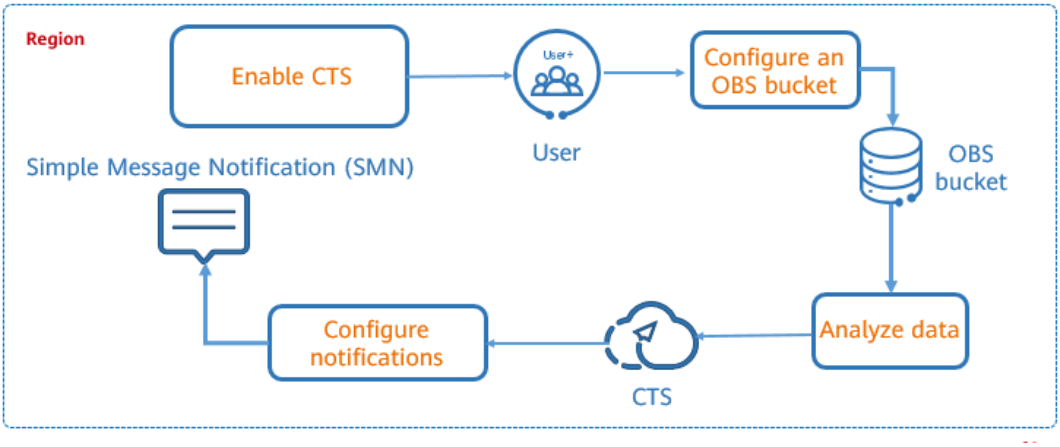


Figure 7-9 Security analysis

The security analysis process follows these steps:

1. CTS records all operations under your account after you enable CTS.
2. The traces are stored in an OBS bucket.
3. The data analytics component can download traces from buckets for analysis.
4. Analysis results can be set as triggers for sending notifications using SMN.

7.2.6.2 Resource Change Tracking

CTS records resource changes and the change results, allowing you to track and analyze resource usage statistics.

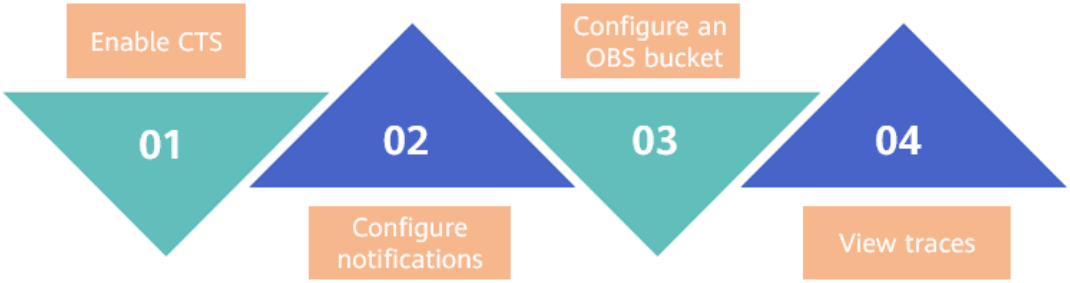


Figure 7-10 Resource change tracking

The resource change tracking process follows these steps:

1. All changes on cloud services are recorded by CTS.
2. You can configure notifications for key operations.
3. Change records are permanently saved.
4. You can query traces for details about resource changes.

7.2.6.3 Fault Locating

If a fault occurs, you can view CTS traces to figure out the cause and quickly rectify the fault. For example, you can quickly determine that the deletion of a system volume during configuration led to a failure in ECS capacity expansion.

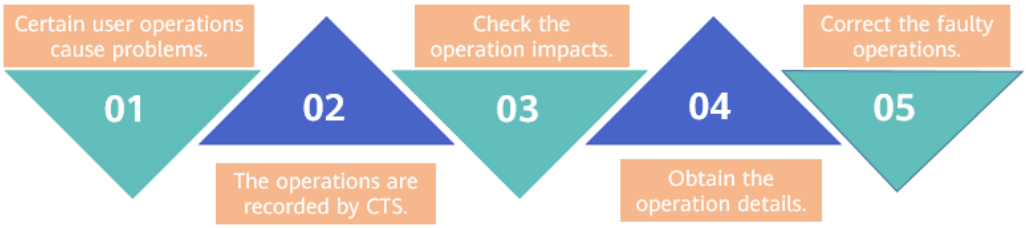


Figure 7-11 Fault locating

The fault locating process follows these steps:

1. A user performs operations that cause problems.
2. All operations are recorded by CTS.
3. You can search the related traces by resource name and check the operation impacts.
4. You can obtain the operation details, including the time and the user who performed the operations.
5. You can correct the faulty operations based on the obtained information.

7.2.6.4 Compliance Auditing

CTS records operations and allows you to query the records, making it easy to comply with internal policies and regulatory standards. This helps you meet the requirements of IT compliance certifications (for example, certifications for financial cloud and trusted cloud).

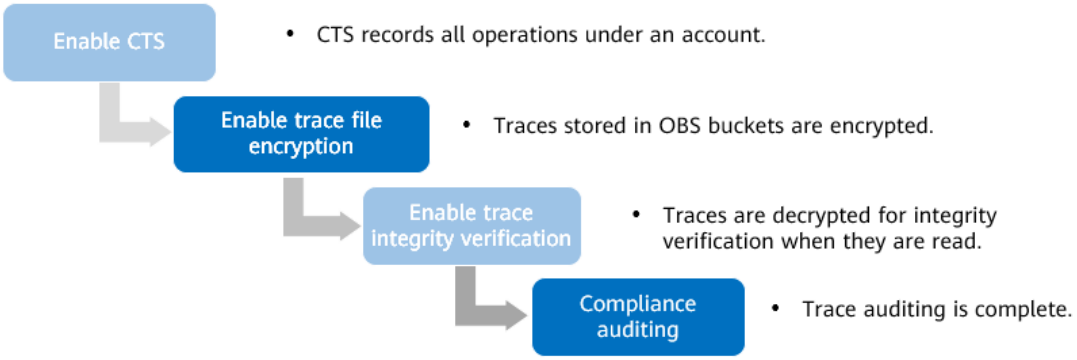


Figure 7-12 Compliance auditing

7.2.6.5 Key Event Notifications

Traces can be set as triggers for notifications sent to emails, mobile phones, and system interfaces or as triggers to invoke FunctionGraph functions.

Main functions:

- You can be alerted of changes to core system components, networking, and security configurations so that risks can be detected and mitigated as soon as possible.
- Traces collected by CTS can be synchronized to your own audit systems through HTTP/HTTPS notifications for independent auditing.
- FunctionGraph can be triggered by traces to execute specific functions.

7.3 Cloud Eye

7.3.1 What Is Cloud Eye?

Monitoring helps identify risks. Through monitoring, we can learn the status of an enterprise network. If a security risk is detected, O&M personnel can be informed of the risk in a timely manner, so that they have time to mitigate the risk. This prevents service systems from being affected, resolving issues at the earliest possible opportunity.

Cloud Eye is a multi-dimensional monitoring service. You can use Cloud Eye to monitor resources, set alarm rules, identify resource exceptions, and quickly respond to resource changes.

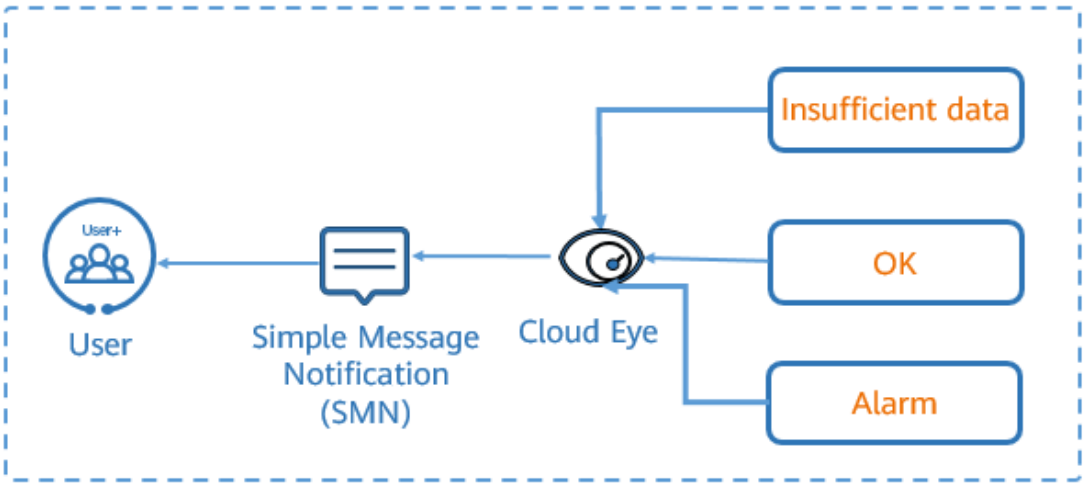


Figure 7-13 Cloud Eye

Cloud Eye provides the following functions:

- Automatic monitoring: Monitoring starts automatically after cloud resources such as ECSs or AS groups are created. After you deploy a cloud service, you can view its running status and configure alarm rules on the Cloud Eye console.
- Real-time notification: Users can enable **Alarm Notification** when creating alarm rules. If the status of a cloud service changes and metrics reach the thresholds specified in the alarm rules, Cloud Eye notifies you by text messages or emails, or by sending HTTP or HTTPS messages to servers. In this way, you can monitor the cloud resource status and changes in real time.
- Panels: Panels enable you to view cross-service and cross-dimension monitoring data. Panels display key metrics centrally, providing an overview of the service operating status and allowing you to check monitoring details when troubleshooting.
- Resource groups: A resource group allows you to add and monitor resources, such as ECSs, EVS disks, EIPs, bandwidths, and databases related to a certain service. They provide a way to track the collective health of all the resources related to a service. Resources of different types, alarm rules, and alarm history are managed based on the service, facilitating O&M.
- OBS dump: Raw data for each metric is kept for only two days on Cloud Eye. If you need to retain data for longer, use OBS buckets, and raw data will be automatically synchronized and saved to OBS.

7.3.2 Advantages

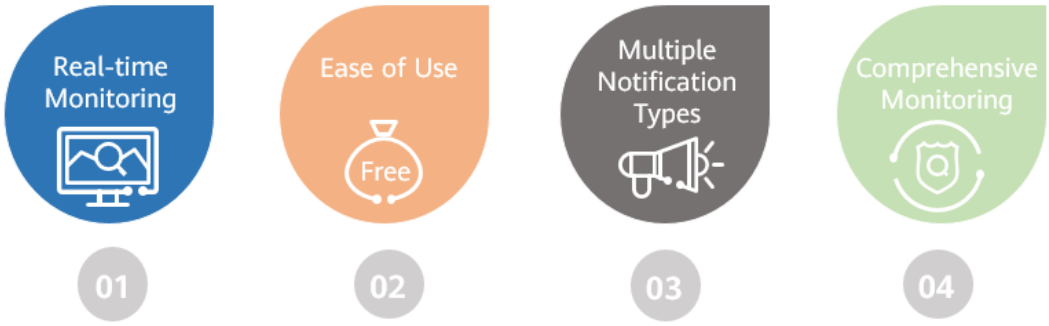


Figure 7-14 Advantages of Cloud Eye

Cloud Eye has the following advantages:

- Automatic provisioning

  Cloud Eye is automatically provisioned for all users. You can use the Cloud Eye console or APIs to view cloud service statuses and configure alarm rules.

- Reliable real-time monitoring

  Raw data is reported to Cloud Eye in real time for monitoring of cloud services. Alarms are generated and notifications are sent to you in real time.

- Visualized monitoring

  You can create monitoring panels and graphs to compare multiple metrics. The graphs automatically refresh to display the latest data.

- Multiple notification types

  You can enable **Alarm Notification** when creating alarm rules. When metrics reach the thresholds specified in the alarm rules, Cloud Eye notifies you by email or text, allowing you to keep track of the status of cloud services. Cloud Eye can also send HTTP/HTTPS requests to an IP address of your choice, enabling you to build smart alarm handling programs.

- Batch creation of alarm rules

  Alarm templates allow you to create alarm rules in batches for multiple cloud services.

7.3.3 Architecture

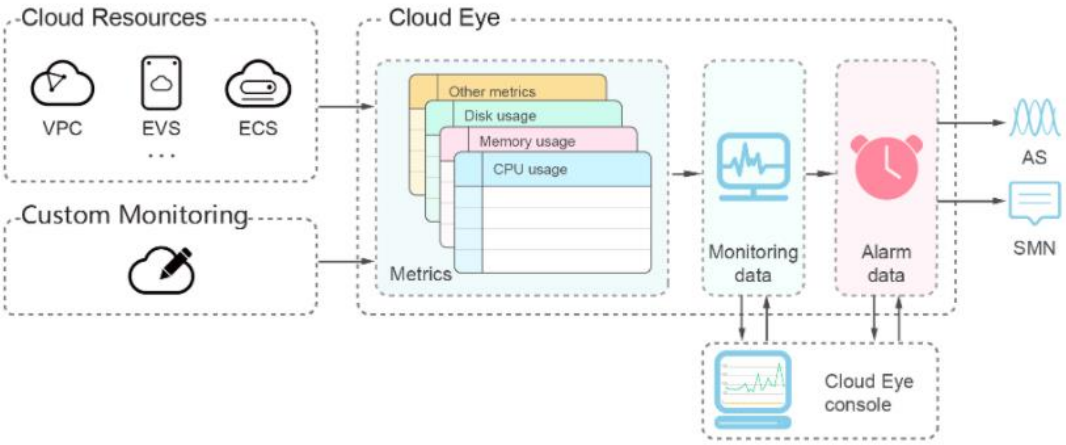


Figure 7-15 Architecture of Cloud Eye

7.3.4 Application Scenarios

Cloud Eye can be used for scenarios like cloud service monitoring, server monitoring, and troubleshooting.

- Cloud service monitoring

  After enabling a cloud service supported by Cloud Eye, you can view the service status and metrics, and create alarm rules based on those metrics on the Cloud Eye console.

- Server monitoring

  By monitoring the ECS and BMS metrics, such as CPU usage, memory usage, and disk usage, you can ensure that your ECSs or BMSs are running normally, and you can avoid service interruptions caused by overused resources.

- Troubleshooting

  When an alarm rule's conditions are met, Cloud Eye generates an alarm and invokes an SMN API to send notifications, allowing you to identify root causes of performance issues.

- Capacity expansion

  After you create alarm rules for metrics such as CPU, memory, and disk usage, you can track the statuses of your services. If service volume increases, Cloud Eye sends you an alarm notification, so you can manually expand capacity, or the alarm notification triggers the AS policies you configured for automatic capacity expansion if needed.

7.3.5 How to Use Cloud Eye

7.3.5.1 Panels

You can use panels to view core metrics and compare performance data of different services from different dimensions. You can create 20 panels, add 24 monitoring graphs to each panel, and add 20 items to each graph.

7.3.5.2 Metrics

This is the core concept of Cloud Eye. A metric is a quantitative value of a resource dimension on the cloud platform, such as the ECS or BMS CPU or memory usage. A metric is a time-dependent variable that generates a certain amount of monitoring data over time. It helps you understand the changes over a specific period of time.

7.3.5.3 Server Monitoring

Server monitoring is comprised of basic monitoring, OS monitoring, and process monitoring for servers.

Basic monitoring provides Agent-free monitoring for basic ECS and BMS metrics.

OS monitoring provides proactive and fine-grained OS monitoring for servers, and it requires the Agent (a plug-in) to be installed on all servers that will be monitored.

Process monitoring is used to monitor active processes on hosts. By default, Cloud Eye collects the CPU usage, memory usage, and number of opened files of active processes.

Functions of server monitoring:

Server monitoring provides more than 40 metrics, such as metrics for CPU, memory, disk, and network, to meet the basic monitoring and O&M requirements for servers.

After the Agent is installed, data of Agent-related metrics is reported once a minute.

CPU usage, memory usage, and the number of opened files used by active processes give you a better understanding of the ECS or BMS resource usages.

7.3.5.4 Website Monitoring

Website monitoring is free and is available in the CN North-Beijing1 region. If you want to use this function in other regions, ensure you have the CES FullAccess permissions configured in project **cn-north-1 [CN North-Beijing1]**.

Advantages:

You can create, modify, disable, enable, or delete monitors.

The configuration is simple and quick, allowing you to improve efficiency and save resources that you would otherwise use to configure complex open-source products.

You receive notifications of website exceptions in real time.

7.3.5.5 Custom Monitoring

The **Custom Monitoring** page displays all custom metrics reported by you and other users. You can use simple API requests to report collected monitoring data of those metrics to Cloud Eye for processing and display.

7.3.5.6 Event Monitoring

In event monitoring, you can query system events and custom events reported to Cloud Eye through the API. You can create alarm rules for both system events and custom events. When specific events occur, Cloud Eye generates alarms for them.

Events are key operations on cloud service resources that are stored and monitored by Cloud Eye. You can view events to see operations performed by specific users on specific resources, such as deleting or rebooting an ECS. Event monitoring is enabled by default. Event monitoring provides an API for reporting custom events, which helps you collect and report abnormal events or important change events generated by services to Cloud Eye.

You can see from the following description that event monitoring and custom monitoring are similar but not the same.

The differences between custom event monitoring and custom monitoring are as follows:

- Monitoring of custom events is used to report and query monitoring data for non-consecutive events, and generate alarms in these scenarios.
- Custom monitoring is used to report and query periodically and continuously collected monitoring data, and generate alarms in these scenarios.

7.4 LTS

7.4.1 What Is LTS?

Logs are files generated by system processes and record important system information. They provide useful details for fault location and program commissioning.

Log Tank Service (LTS) collects logs from hosts and cloud services for centralized management, and processes large volumes of logs efficiently, securely, and in real-time. LTS provides you with the insights needed to optimize the availability and performance of cloud services and applications. It helps you make faster data-driven decisions, perform device O&M, and analyze service trends.

LTS provides the following basic functions:

- Real-time log collection

LTS collects logs from hosts and cloud services in real time and displays them on the LTS console in an intuitive and orderly manner. You can query logs or transfer them for long-term storage.

You can define log structuring rules so LTS will extract logs that are in a fixed format or share a similar pattern based on the rules. Then you can use SQL syntax to query the structured logs.

- Log query and real-time analysis

  Collected logs can be quickly queried by keyword or fuzzy match. You can analyze logs in real time to perform security diagnosis and analysis, and obtain operations statistics, such as cloud service visits and clicks.

- Log monitoring and alarms

  LTS works with Application Operations Management (AOM) to count the frequency of specified keywords in logs retained in LTS. For example, if the keyword ERROR occurs frequently, it can indicate that services are not running normally.

- Log transfer

  Logs reported from hosts and cloud services are retained in LTS for seven days by default. You can set the retention period between one and thirty days. Logs older than the retention period will be automatically deleted. For long-term storage, you can transfer logs to the Object Storage Service (OBS), Data Ingestion Service (DIS), or Distributed Message Service (DMS).

7.4.2 Advantages

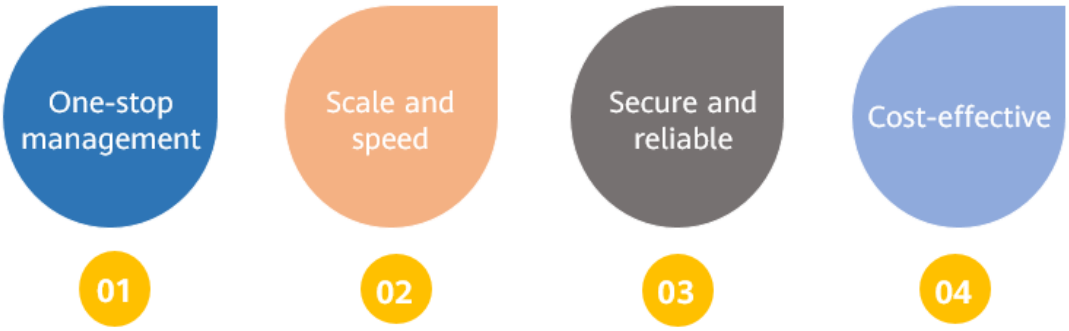


Figure 7-16 LTS advantages

LTS offers the following advantages:

- One-stop management: You can easily manage your logs with this all-in-one platform. You can ingest, store, transfer, and search for logs in LTS. Advanced functions are also available for you to structure or visualize logs, or run SQL queries.
- Scale and speed: LTS allows you to manage petabytes of logs with ease. It can ingest logs at up to 200 TB/day. You can obtain results in seconds even when searching gigabytes of logs, or when running SQL aggregate queries at hundreds of megabytes.
- Secure and reliable: LTS keeps your data secure with HTTPS encryption and rights- and domain-based control. It has an availability of 99.95%.

- Cost-effective: LTS helps you save on maintenance, and there are no upfront commitments. You can scale up resources at any time to meet spikes in log volume, but you only pay for what you use.

7.4.3 Application Scenarios

LTS applies to the following scenarios:

- Log collection and analysis

  Without proper management, there are too many logs about hosts and cloud services, which are difficult to query and are cleared periodically. Using LTS, collected logs are displayed on the console in a clear and orderly manner for fast query, and can be stored for a long time if necessary. Collected logs can be quickly queried by keyword or fuzzy match. You can analyze logs in real time to perform security diagnosis and analysis, and obtain operations statistics, such as cloud service visits and clicks.

- Service performance optimization

  The performance and quality of website services play an important role in customer satisfaction. By analyzing the network congestion logs, you can identify your websites' performance bottlenecks, and take measures such as improving website caching policies or network transmission policies to optimize performance.

- Quick network fault locating

  Network quality is the cornerstone of service stability. LTS centralizes logs from different sources, helping you quickly detect and locate faults as well as enabling backtracking.

  For example, you can quickly locate an ECS that causes an error, such as one with excessive bandwidth usage. In addition, you can judge whether there are ongoing attacks, leeching, and malicious requests by analyzing access logs, and locate and rectify faults as soon as possible.

7.4.4 How to Use LTS

7.4.4.1 Key Concepts

When learning about LTS, it is helpful to first understand some concepts including log groups, log read and write, and ICAgent.

Log groups can be created in two ways. They are either automatically created when other Huawei Cloud services are connected to LTS, or you can create one manually on the LTS console.

Data is written to and read from a log stream. You can configure different types of logs, such as operation logs and access logs, to be written into different log streams. ICAgent will package and send the collected log data to LTS on a log-stream basis. To view logs, you can go to the corresponding log stream and query them. In short, the use of log streams significantly reduces the number of log reads and writes, thereby improving efficiency.

ICAgent is the log collection tool of LTS. If you want to use LTS to collect logs from a host, you need to install ICAgent on the host. Batch ICAgent installation is also available if you want to collect logs from multiple hosts. After ICAgent installation, you can check the ICAgent status on the LTS console.

7.4.4.2 Viewing Real-Time Logs

You can view logs in real time on the **Real-Time Logs** tab, where the logs are updated every three seconds.

Logs are reported to LTS once every minute. You may wait for up to 1 minute before the logs are displayed on the **Real-Time Logs** tab. In addition, you can control log display by clicking **Clear** or **Pause** in the upper right corner.

If you click **Clear**, displayed logs will be cleared from the real-time view.

If you click **Pause**, loading of new logs to the real-time view will be paused. After you click **Pause**, the button changes to **Continue**. You can click **Continue** to resume real-time loading.

7.4.4.3 Structuring Logs

After you add extraction rules, LTS uses these rules to convert raw logs to a structured format, facilitating the execution of SQL queries.

7.4.4.4 Visualizing Logs

You can visualize SQL query results in tables, trend charts, bar charts, or pie charts.

LTS supports SQL queries on structured logs. You can set rules for LTS to structure raw logs and then run SQL queries in one or two minutes.

8 Conclusion

This document focuses on the virtualization technology and some essential cloud services of Huawei. Other mature technologies, such as containers and OpenStack, will be detailed in the latest HCIP courses. We will update some useful documents on our official website to keep you abreast of the latest cloud developments.

Any suggestions, comments, and technical questions are welcome on the official HCIA - Cloud Service forum:

https://forum.huawei.com/enterprise/en/index.html

Learn more at:

1. Huawei Certification: https://e.huawei.com/en/talent/#/cert
2. Huawei Talent Online: https://ilearningx.huawei.com/portal/subportal/EBG/51
3. Huawei ICT Academy: https://e.huawei.com/en/talent/#/ict-academy/home?t=1561101910908