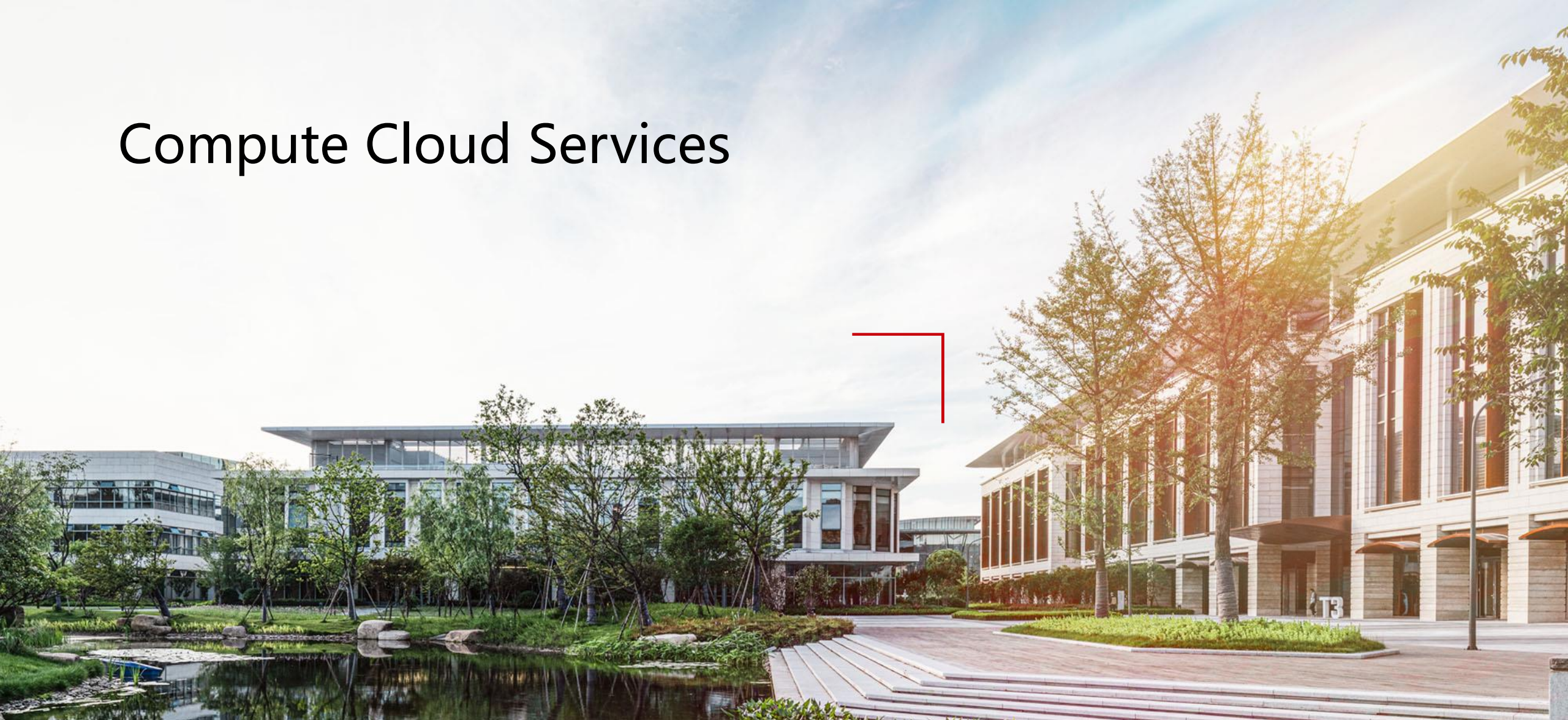


Compute Cloud Services



Foreword

- 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 on HUAWEI CLOUD.

Objectives

- Upon completion of this course, you will:
 - Understand common compute services available on HUAWEI CLOUD.
 - Understand the positioning, technical details, and usage of these compute services.

Compute Cloud Services



Elastic Cloud Server
(ECS)



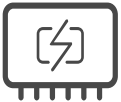
Bare Metal Server
(BMS)



Auto Scaling
(AS)



Cloud Container Engine
(CCE)



GPU Accelerated
Cloud Server
(GACS)



Dedicated Host
(DeH)



Image Management
Service
(IMS)



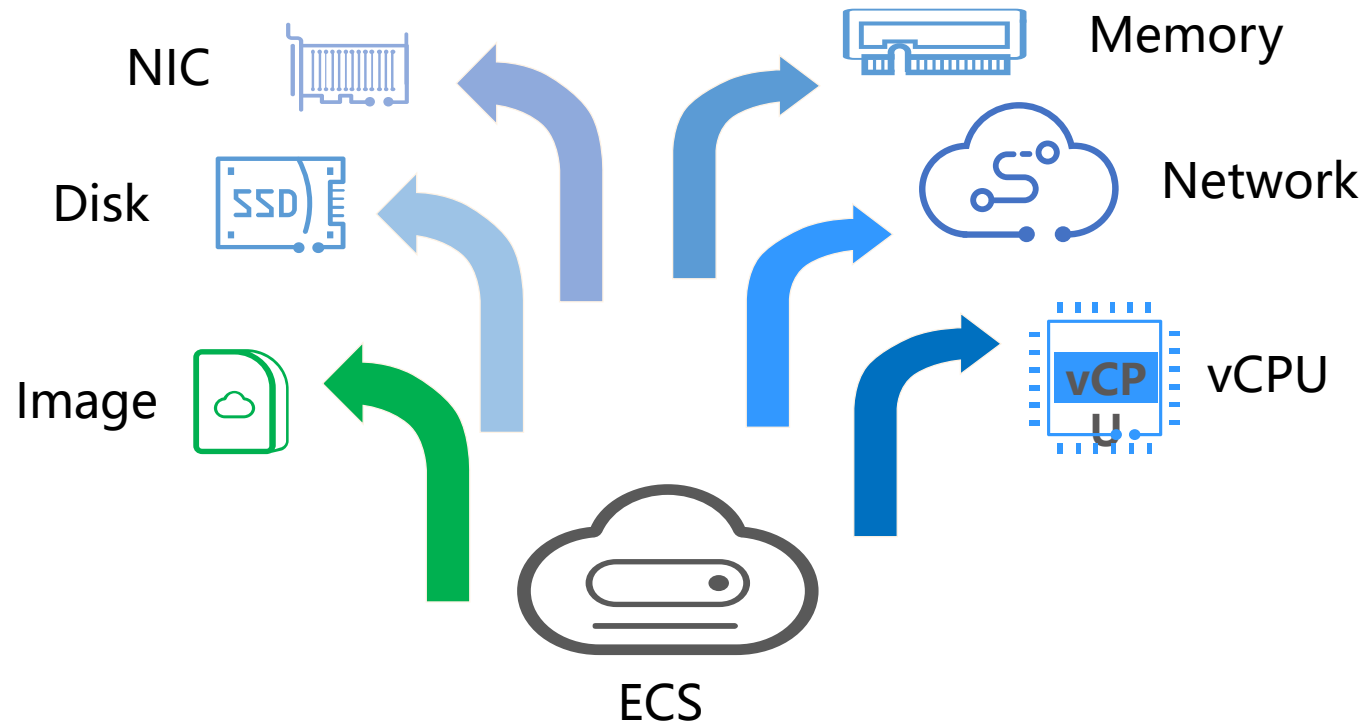
FunctionGraph

Contents

- 1. Elastic Cloud Server (ECS)**
2. Bare Metal Server (BMS)
3. Image Management Service (IMS)
4. Auto Scaling (AS)
5. Cloud Container Engine (CCE)
6. Other Compute Services

What Is ECS?

- An ECS is a basic computing unit that consists of vCPUs, memory, an OS, and Elastic Volume Service (EVS) disks. After an ECS is created, you can use it on the cloud similarly to how you would use your local computer or physical server.



Why ECS?

Auto Scaling

- Automatic adjustment of compute resources
- Flexible adjustment of ECS configurations
- Flexible billing modes

Stability and Reliability

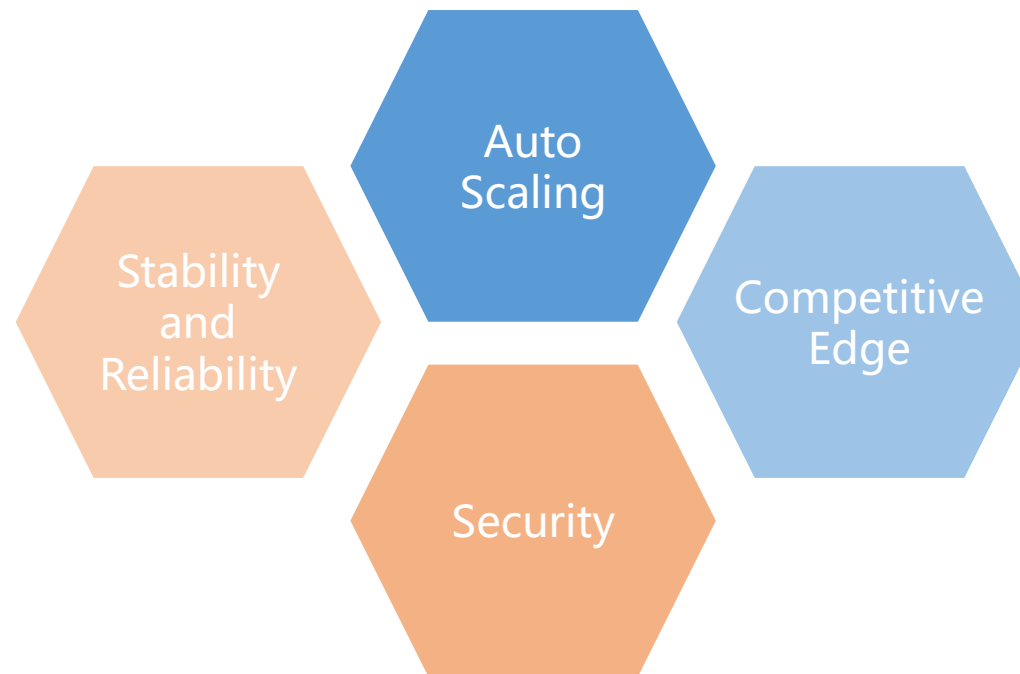
- A variety of EVS disk types
- Reliable data
- Backup and restoration of ECSs and EVS disks

Competitive Edge

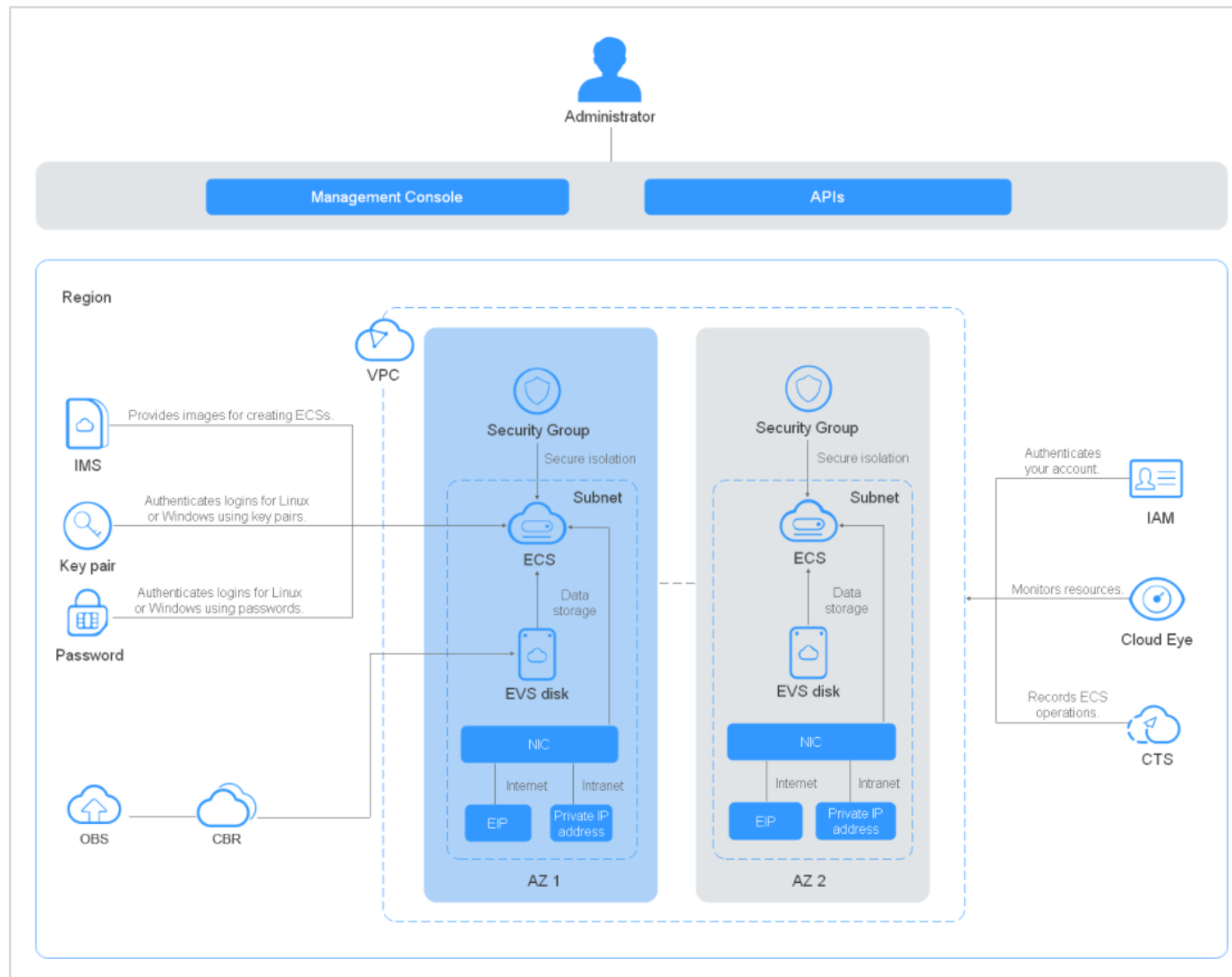
- Professional hardware devices
- Always available virtual resources

Security

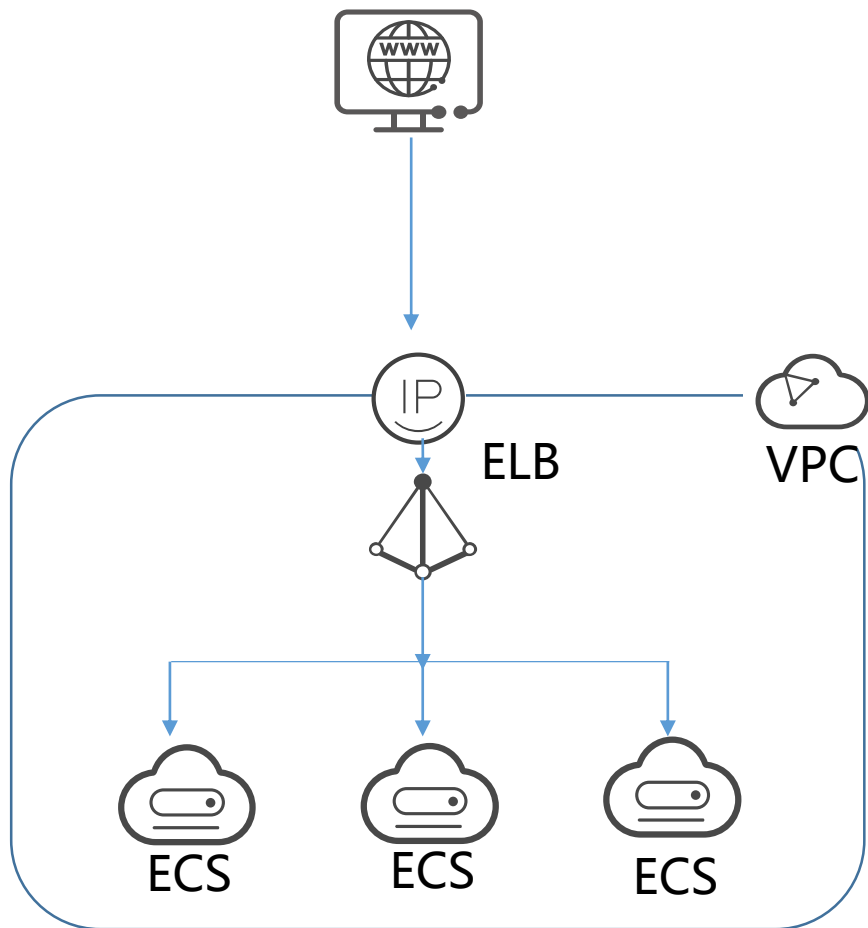
- A range of security services available for multi-dimensional protection
- Security evaluation
- Intelligent process management
- Vulnerability scans



ECS Architecture



Scenarios – Internet



Application Scenarios

Website R&D and testing, and small-scale databases

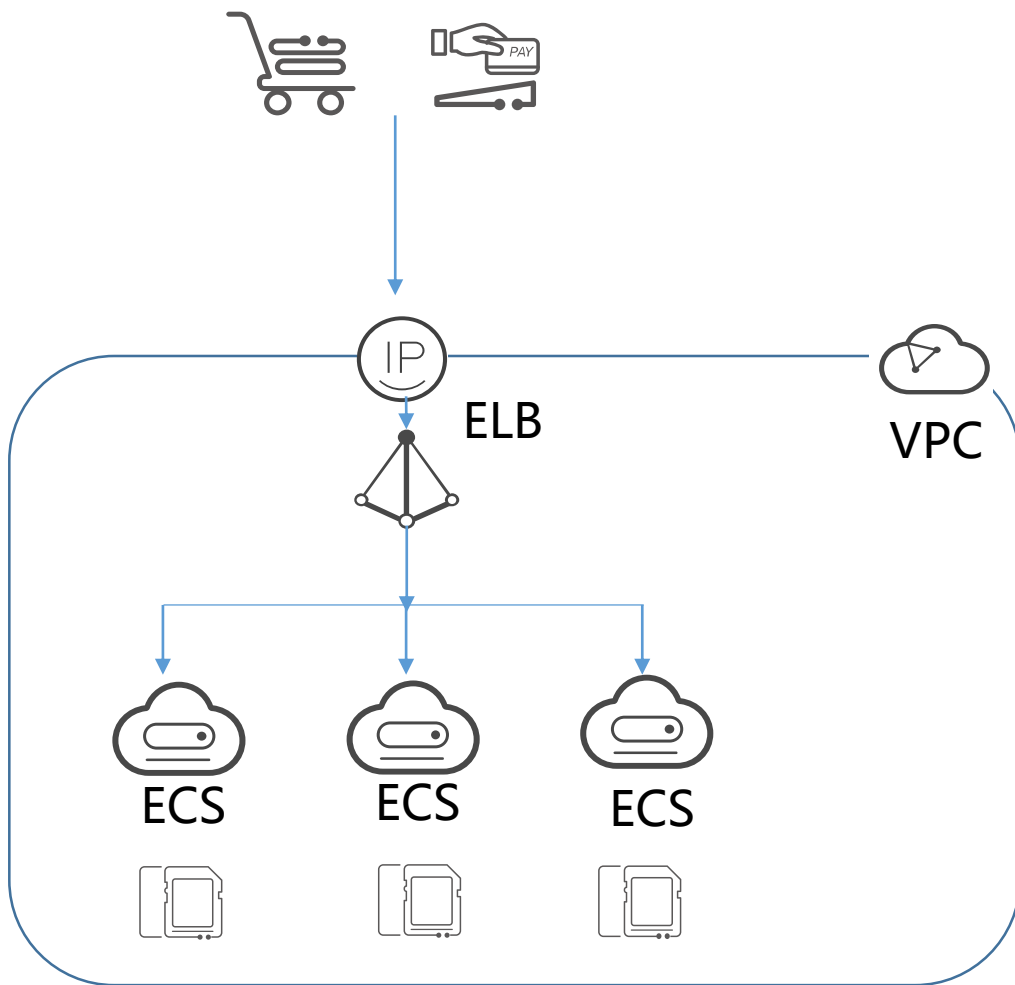
Recommended ECS

General-computing ECSs

Recommendation Reasons

- Requirements: To minimize upfront deployment and O&M costs, applications need to be deployed on only one or just a few servers, but there are no special requirements for CPU performance, memory, disk capacity, or bandwidth, strong security and reliability.
- Solution: **General-computing ECSs** provide a balance of compute, memory, and network resources. They are appropriate for medium-workload applications and meet the cloud service needs of both enterprises and individuals.

Scenarios – E-Commerce



Application Scenarios

Precision marketing, E-Commerce, and mobile apps

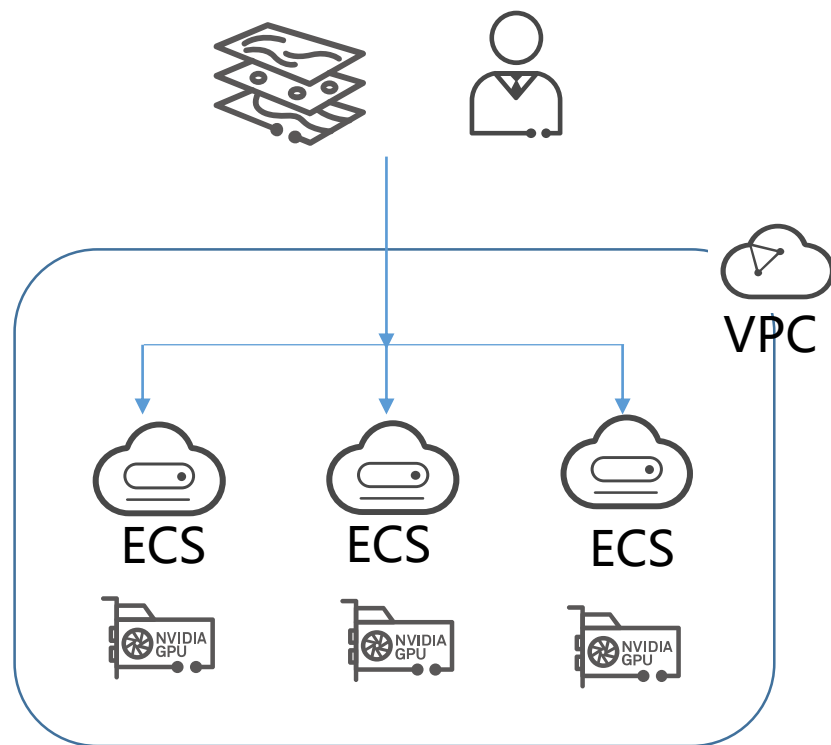
Recommended ECS

Memory-optimized ECSs

Recommendation Reasons

- Requirements: large amount of memory, rapid processing of large volumes of data, and fast network access
- Solution: **memory-optimized ECSs**, which feature a large amount of memory, ultra-high I/O EVS disks, and appropriate bandwidths

Scenarios – Graphics Rendering



Application Scenarios

Graphics rendering and engineering drawing

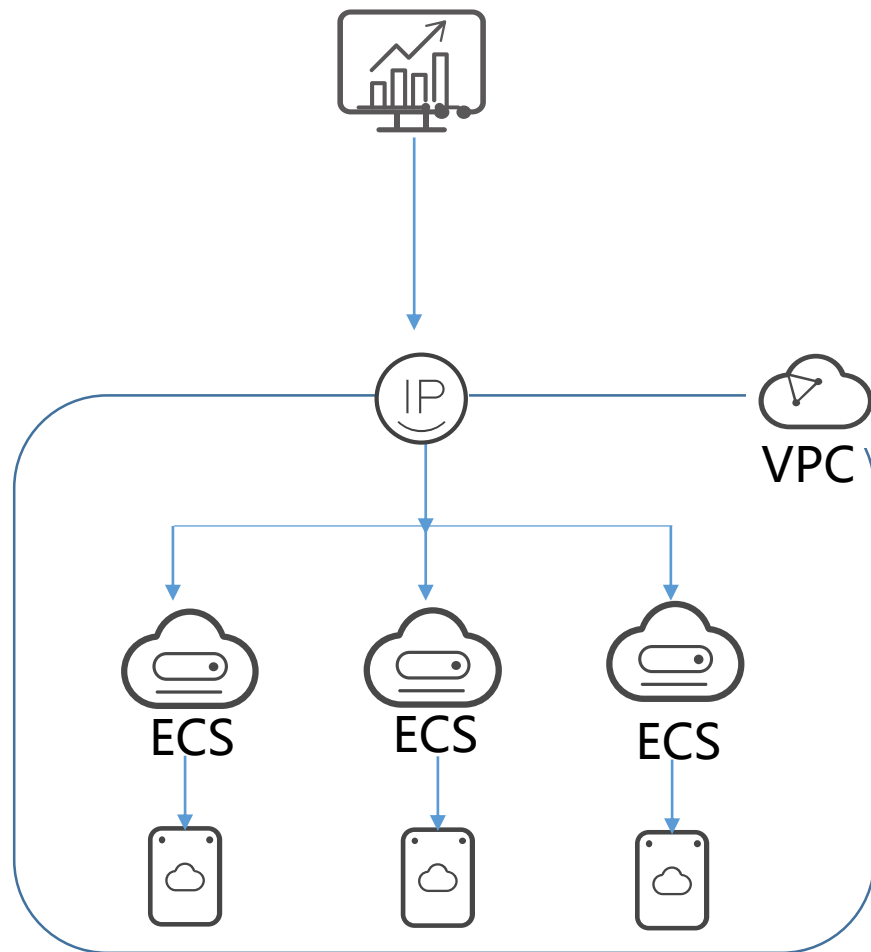
Recommended ECS

GPU-accelerated ECSs

Recommendation Reasons

- Requirements: high-quality graphics and video, lots of memory, processing of large volumes of data, high I/O concurrency, rapid data processing, and high GPU performance
- Solution: **GPU-accelerated ECSs**, which provide cost-effective graphics acceleration

Scenarios – Data Analysis



Application Scenarios

MapReduce and Hadoop

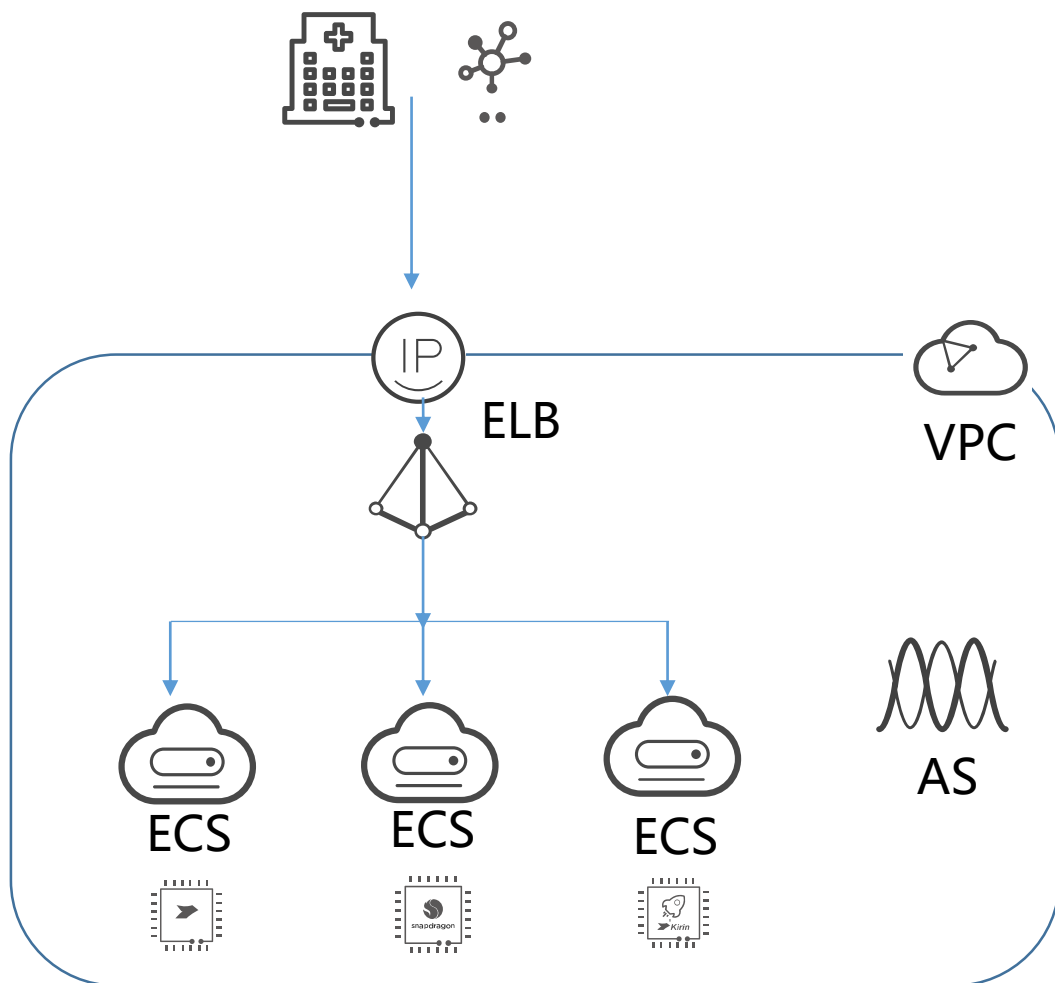
Recommended ECS

Disk-intensive ECSs

Recommendation Reasons

- Requirements: processing of large volumes of data; high I/O performance and rapid data switching and processing
- Solution: **disk-intensive ECSs**, which are suitable for applications requiring high-performance sequential read/write on ultra-large datasets in local storage

Scenarios – High-Performance Computing



Application Scenarios

Computing and storage systems for scientific computing, genetic engineering, games, animations, and biopharmaceuticals

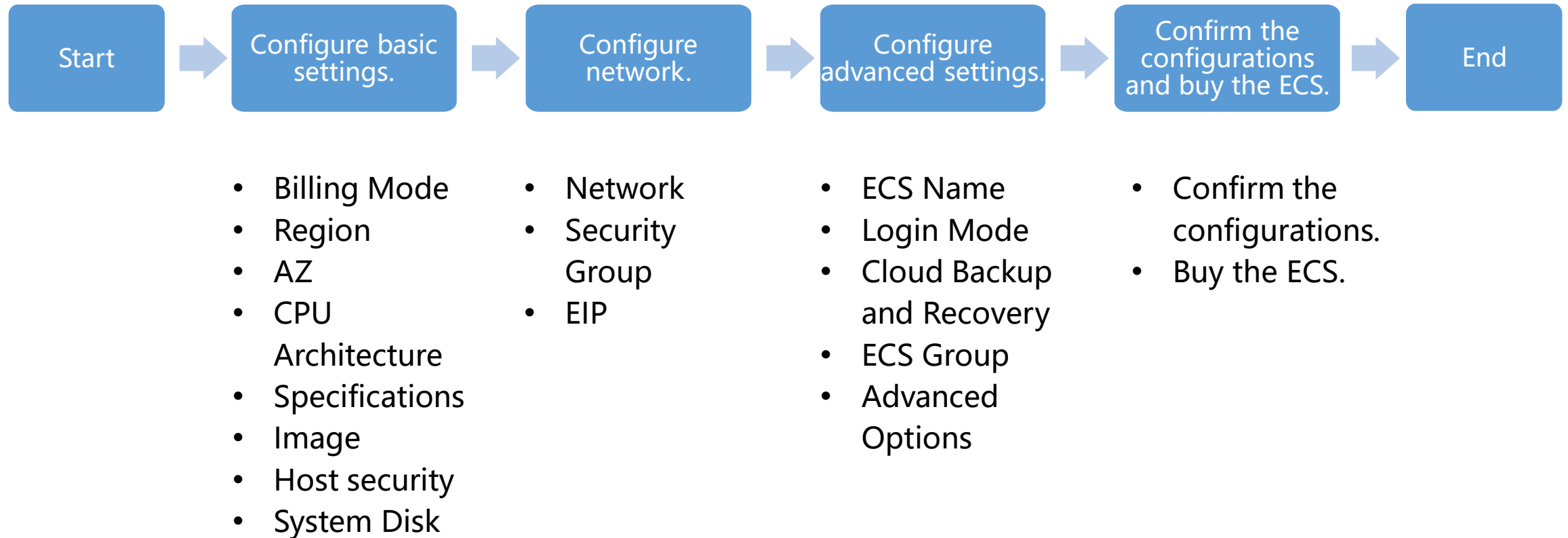
Recommended ECS

High-performance computing ECSs

Recommendation Reasons

Solution: **high-performance computing ECSs**, which meet the computing, storage, and rendering needs of high-performance infrastructure services and applications that require a large number of parallel computing resources.

Purchasing an ECS



Configuring Basic Settings

- Set Billing Mode, Region, AZ, CPU Architecture, and Specifications.

Billing Mode

Region

AZ

Yearly/Monthly

Pay-per-use

Spot price

AP-Singapore

Recommended

CN Southwest-... (0)

CN South-Guangz... (0)

CN North-Beijing4 (0)

CN East-Shanghai1 (0)

AP-Hong-Kong (0)

For low network latency and quick resource access, select the region nearest to your target users. [Learn how](#) to select a region.

Random

AZ1

AZ2

AZ3

CPU Architecture

Specifications

x86

Kunpeng

Latest generation

vCPUs

All

Memory


All


Flavor Name


Configuring Network


- Select a VPC, subnet, and security groups for the ECS.

Network


vpc-default(192.168.0.0/16) 

subnet-default(192.168.0.0/24) 


Automatically-assigned IP address 



Available private IP addresses: 250 

Create VPC


Extension NIC  Add NIC NICs you can still add: 1

Security Group

Sys-WebServer (6b832bf5-6804-4a8b-972e-31b2d3374c... 

 Create Security Group 

Similar to a firewall, a security group logically controls network access.

Security Group Rules 

Inbound Rules | Outbound Rules

Configuring Advanced Settings

- Set ECS Name, Login Mode, Cloud Backup and Recovery, ECS Group, and Advanced Options.

ECS Name

ecs-dc78

☐ Allow duplicate name

If multiple ECSs are created at the same time, the system automatically adds a hyphen followed by a four-digit number. For example, if the first ECS is named ecs-0001, the name of the second ECS will be ecs-0001-0001. If an ECS with the name ecs-0010 already exists, the name of the first new ECS will be ecs-0011.

Login Mode

Password

Key pair

Set password later

Username

root

Password

Keep the password secure. If you forget the password, you can log in to the ECS console and change it.

Enter a password.

Confirm Password

Enter the password again.

Cloud Backup and Recovery

To use CBR, you need to purchase a backup vault. A vault is a container that stores backups for servers.

Create new

Use existing

Not required

?

ECS Group (Optional)

Anti-affinity

?

--Select ECS group--

?

Create ECS Group

Advanced Options

☐ Configure now

Access Methods

- HUAWEI CLOUD provides a web-based management platform. You can access ECSs through the management console or HTTPS-based REST APIs.

API



Use an API if you need to integrate the ECSs into a third-party system for secondary development.

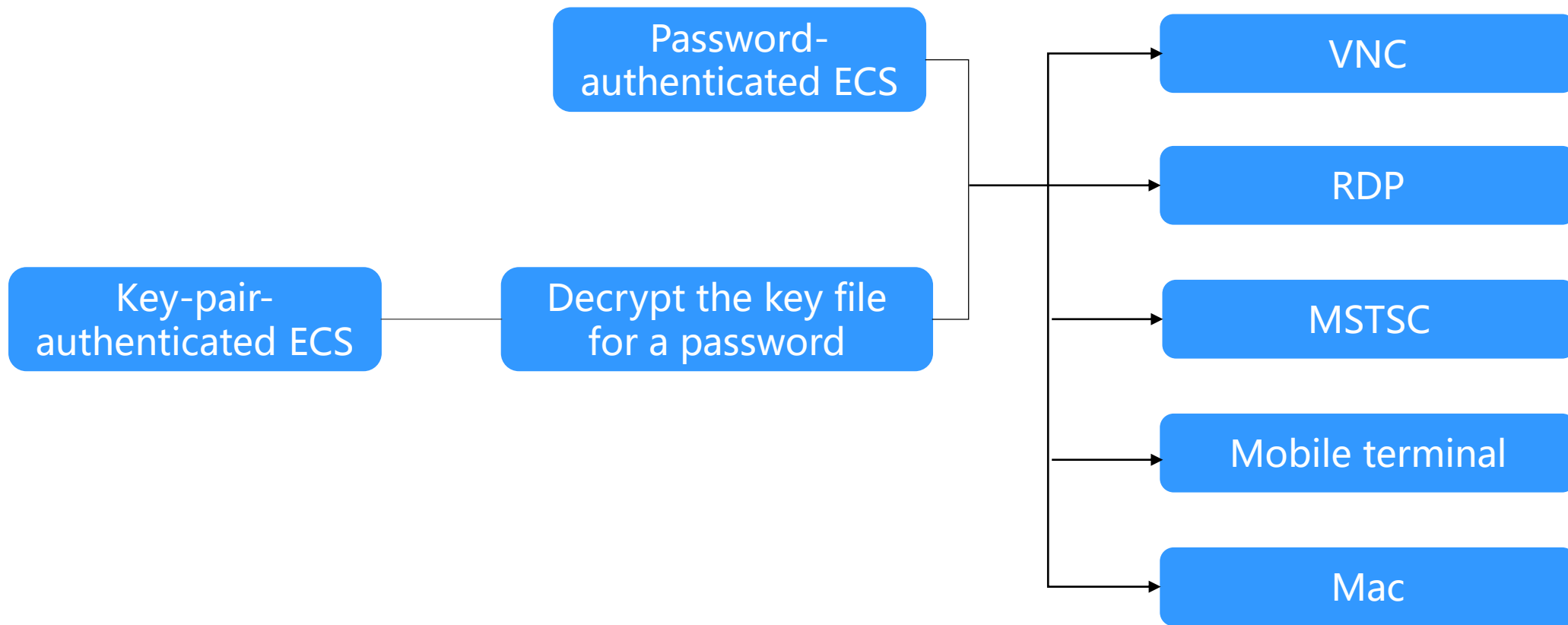
Management Console



After registering on HUAWEI CLOUD, log in to the management console and click Elastic Cloud Server under Compute on the homepage.

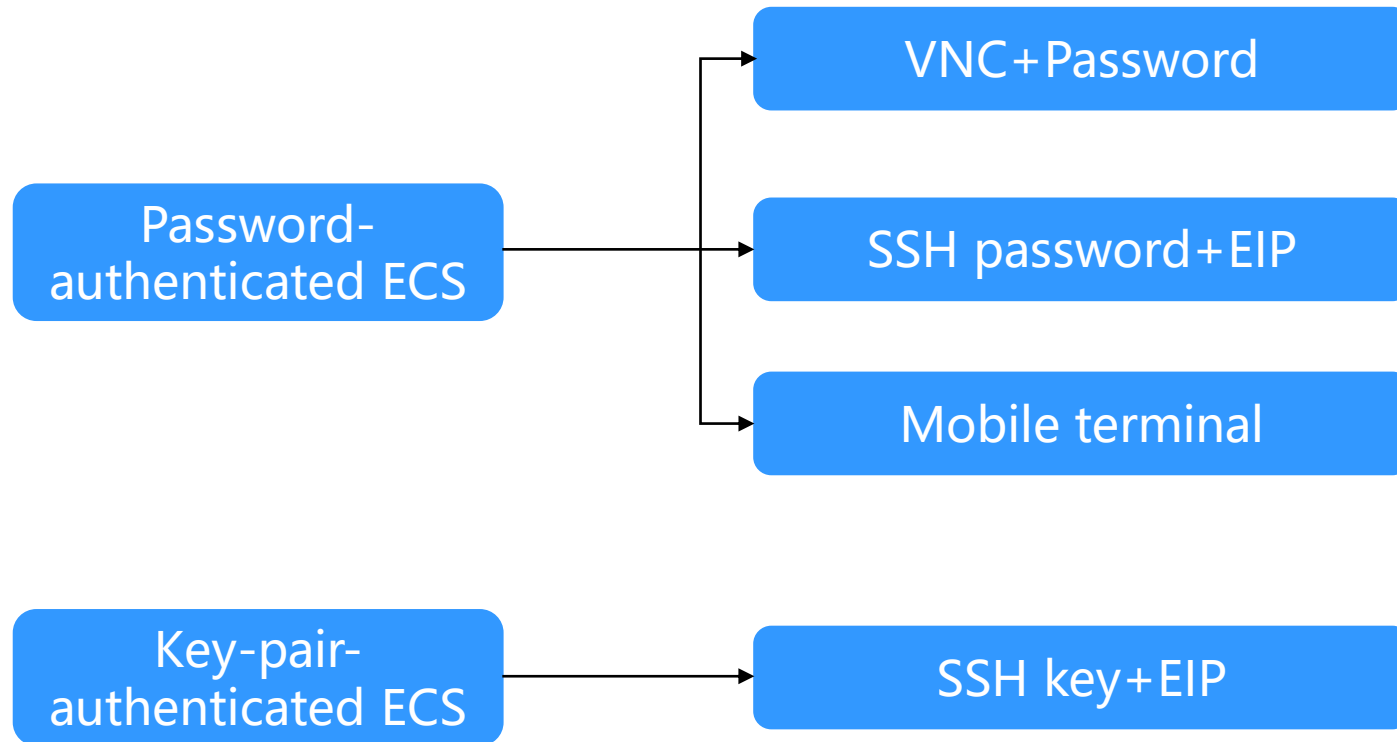
Logging In to a Windows ECS

- Select a login method and log in to the ECS.



Logging In to a Linux ECS

- The method of logging in to an ECS varies depending on the login authentication configured when you purchased the ECS.

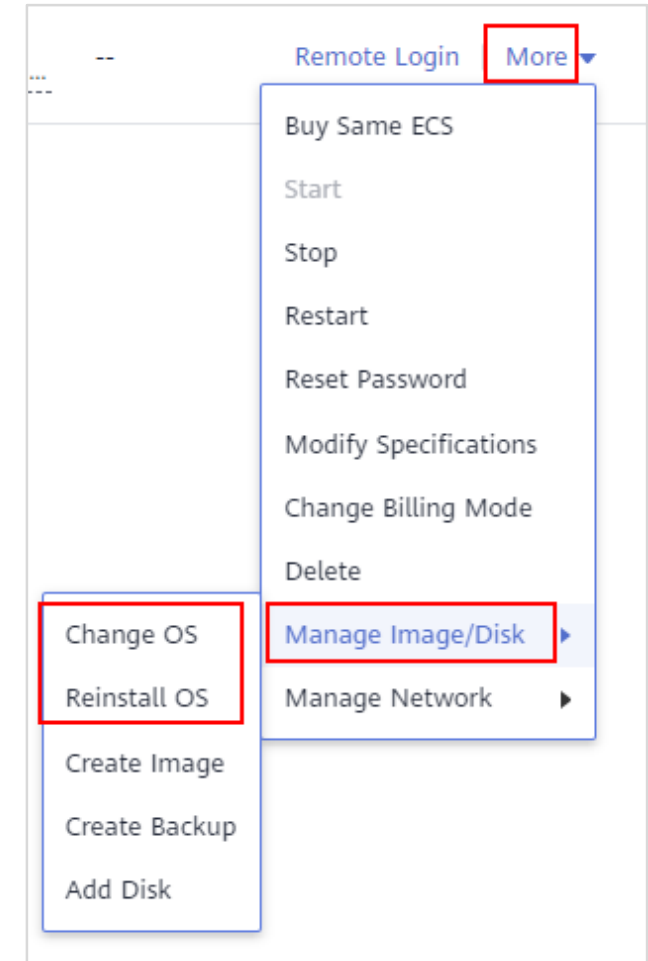


Reinstalling/Changing an ECS OS

- Scenarios: If the OS of an ECS fails to start, requires optimization, or cannot meet service requirements, reinstall or change the OS.

Notes

- Only the original image of the ECS can be used to reinstall the OS.
- Changing the OS will change the system disk of the ECS. After the change, there will be a new system disk ID, and the original system disk will be gone.

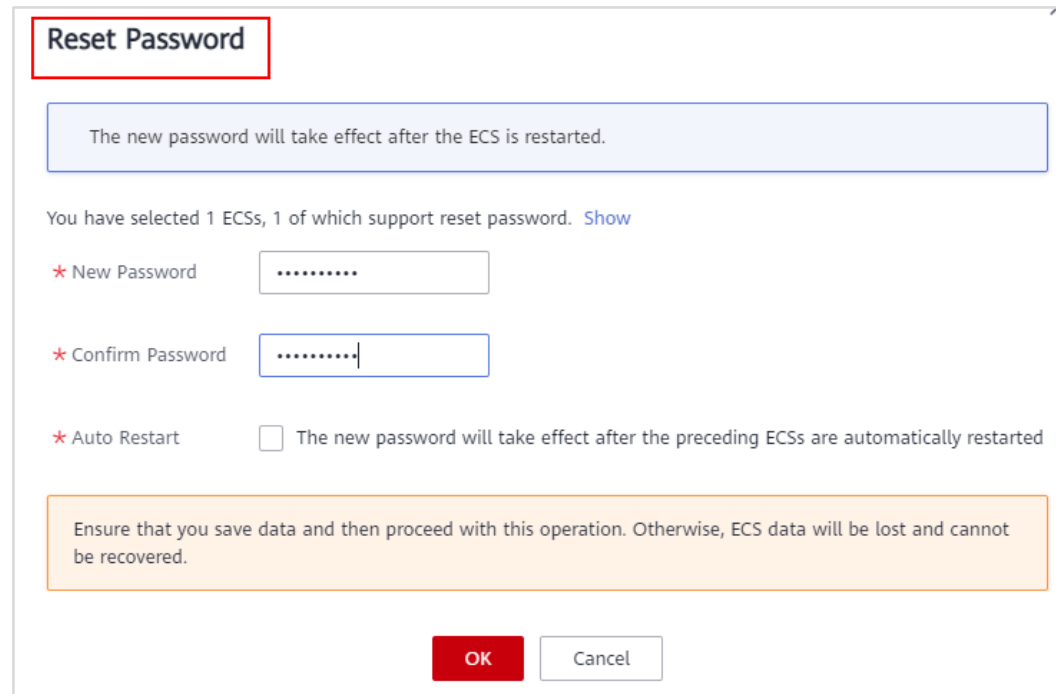


Modifying ECS Specifications

- If the specifications of an existing ECS cannot meet service requirements, modify the ECS specifications as needed, for example, by increasing the number of vCPUs or adding memory.
- Notes
 - To modify the specifications of a yearly/monthly ECS, select the target specification, pay the difference in price or claim the refund, and restart the ECS.
 - There is no need to make an additional up front payment and there are no refunds if you modify the specifications of a pay-per-use ECS.

Resetting the ECS Login Password

- Scenarios: The ECS password is lost or has expired.
- Prerequisites: One-click password reset plug-ins have been installed on the ECS.
- Notes: ECSs created using a public image have the one-click password reset plug-in installed by default.



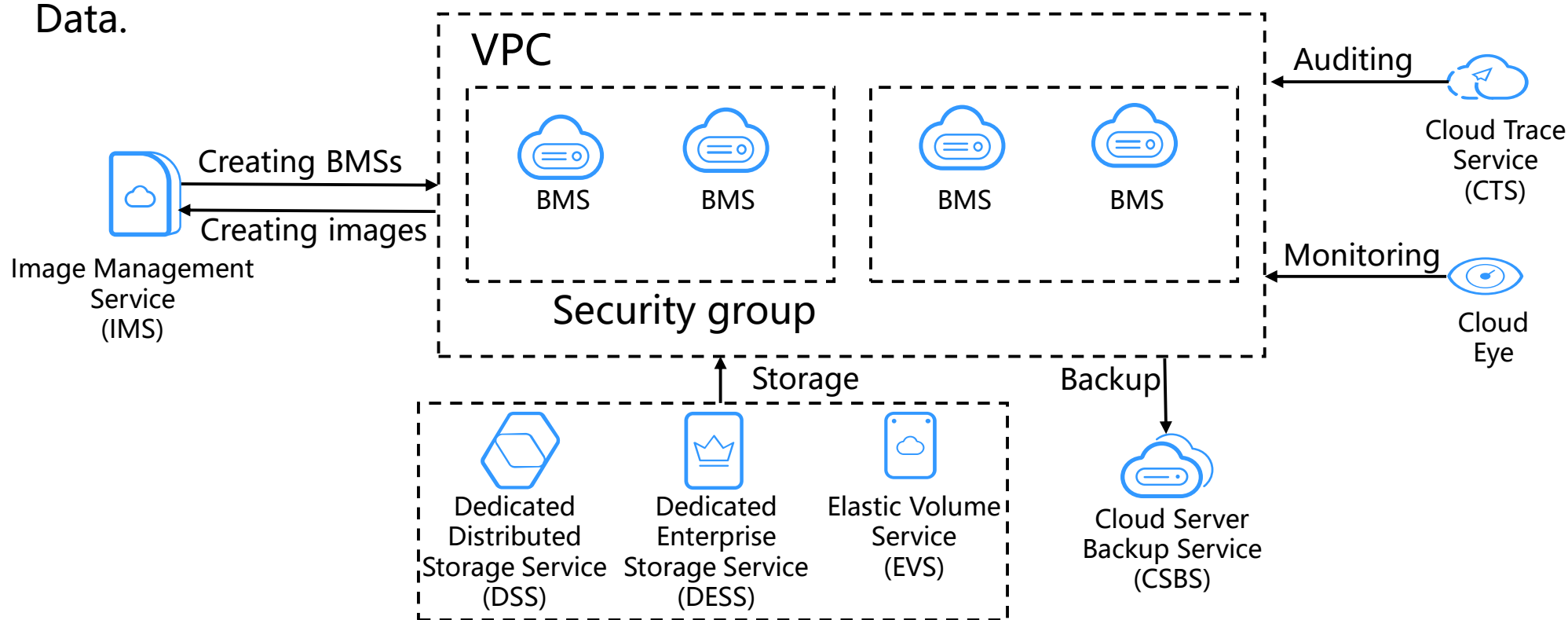
The screenshot shows a 'Reset Password' dialog box. At the top, the title 'Reset Password' is highlighted with a red box. Below the title, a light blue box contains the text: 'The new password will take effect after the ECS is restarted.' Underneath, a message states: 'You have selected 1 ECSs, 1 of which support reset password. [Show](#)'. The form includes three fields: 'New Password' with a red star icon and a masked input field; 'Confirm Password' with a red star icon and a masked input field; and 'Auto Restart' with a red star icon, an unchecked checkbox, and the text 'The new password will take effect after the preceding ECSs are automatically restarted'. At the bottom, an orange box contains a warning: 'Ensure that you save data and then proceed with this operation. Otherwise, ECS data will be lost and cannot be recovered.' Below the warning are two buttons: a red 'OK' button and a white 'Cancel' button.

Contents

1. Elastic Cloud Server (ECS)
- 2. Bare Metal Server (BMS)**
3. Image Management Service (IMS)
4. Auto Scaling (AS)
5. Cloud Container Engine (CCE)
6. Other Compute Services

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.



Why BMS?

High security

- Dedicated servers, VPC network, and security group
- Server security protection
- Disk backup and restoration
- Dedicated storage

High performance

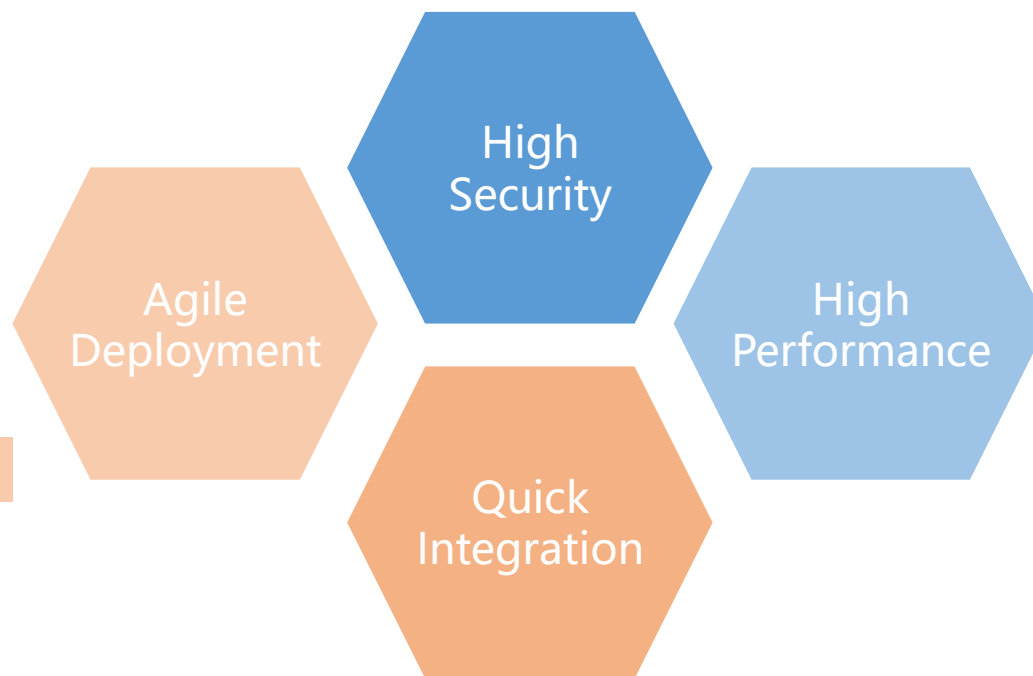
- No virtualization overhead or performance loss
- Cloud-based storage and network access
- Deployment density and performance for mission-critical services

Agile deployment

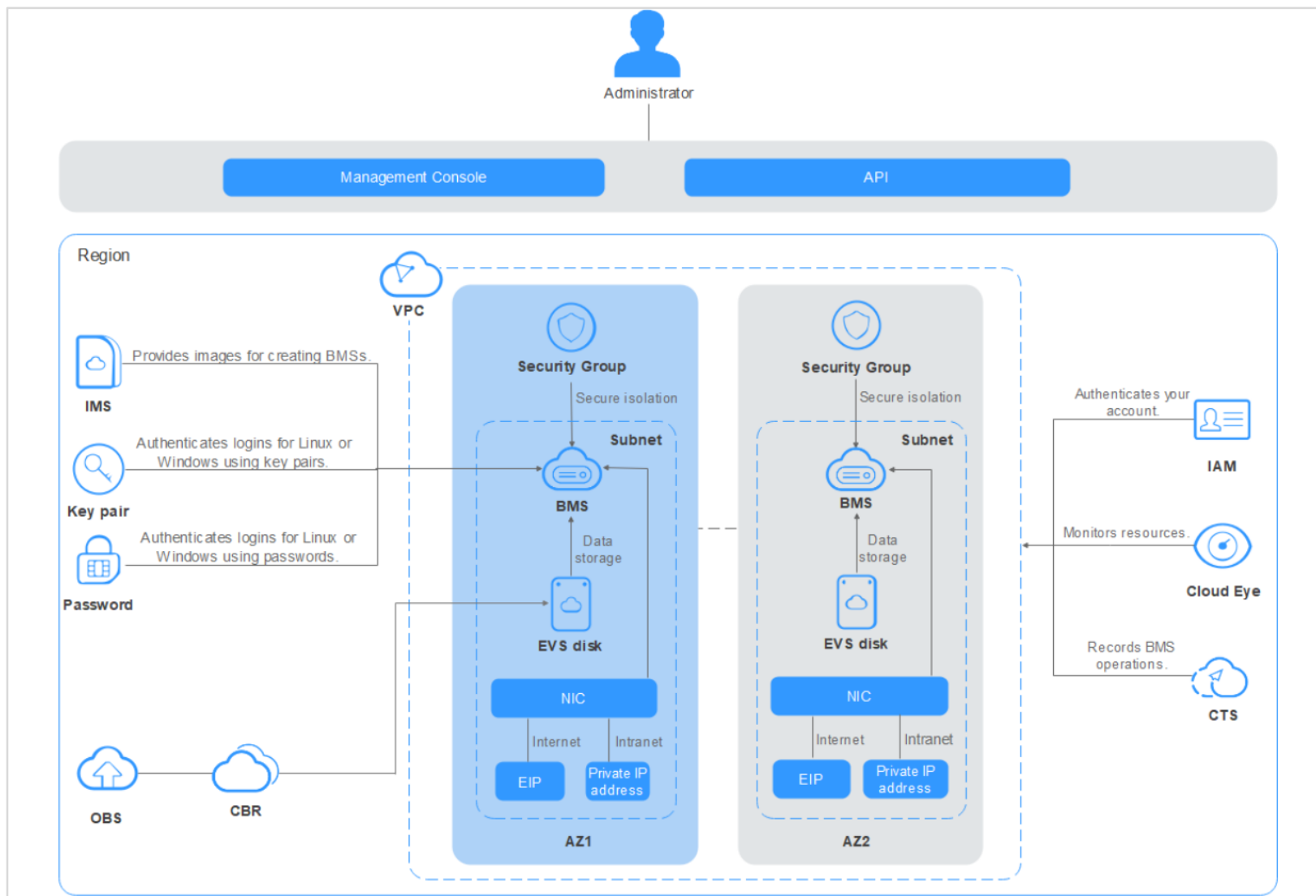
- Quick provisioning (can be booted from EVS disks)
- Self-service lifecycle management and O&M

Quick integration

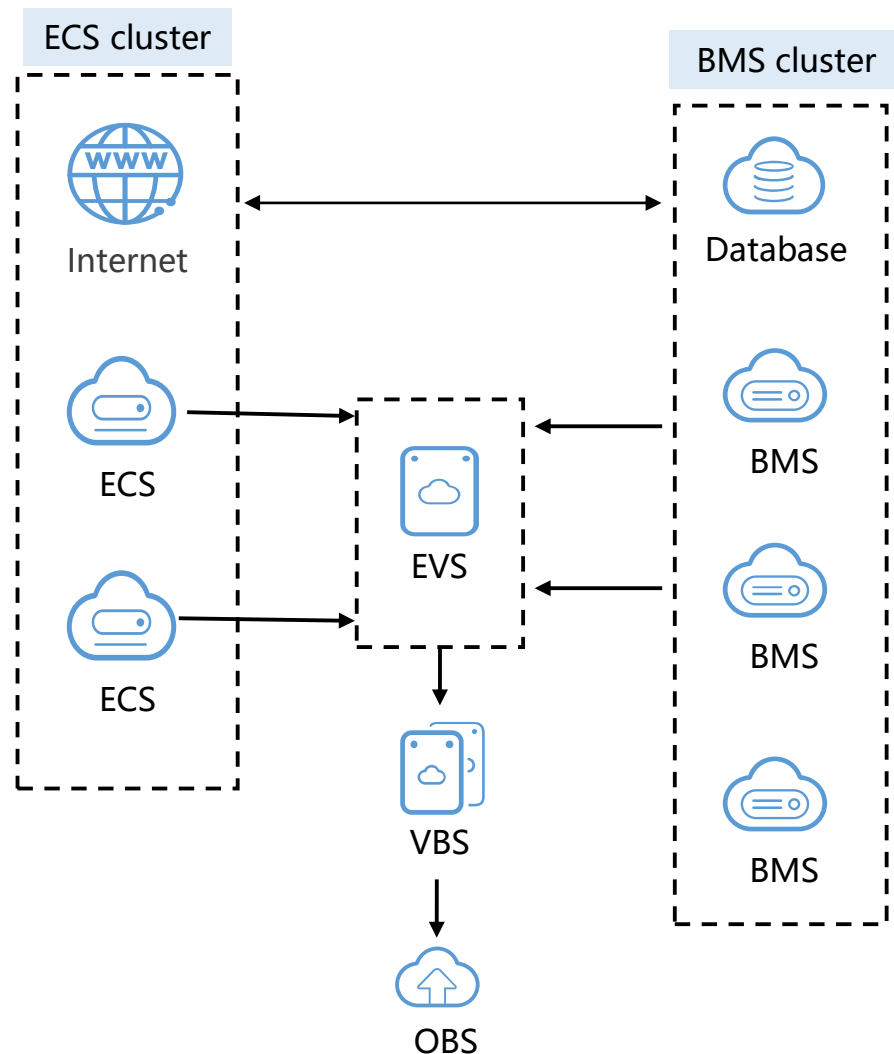
Quick integration with cloud services and cloud solutions for accelerated cloud transformation



BMS Architecture



Scenarios - Core Database



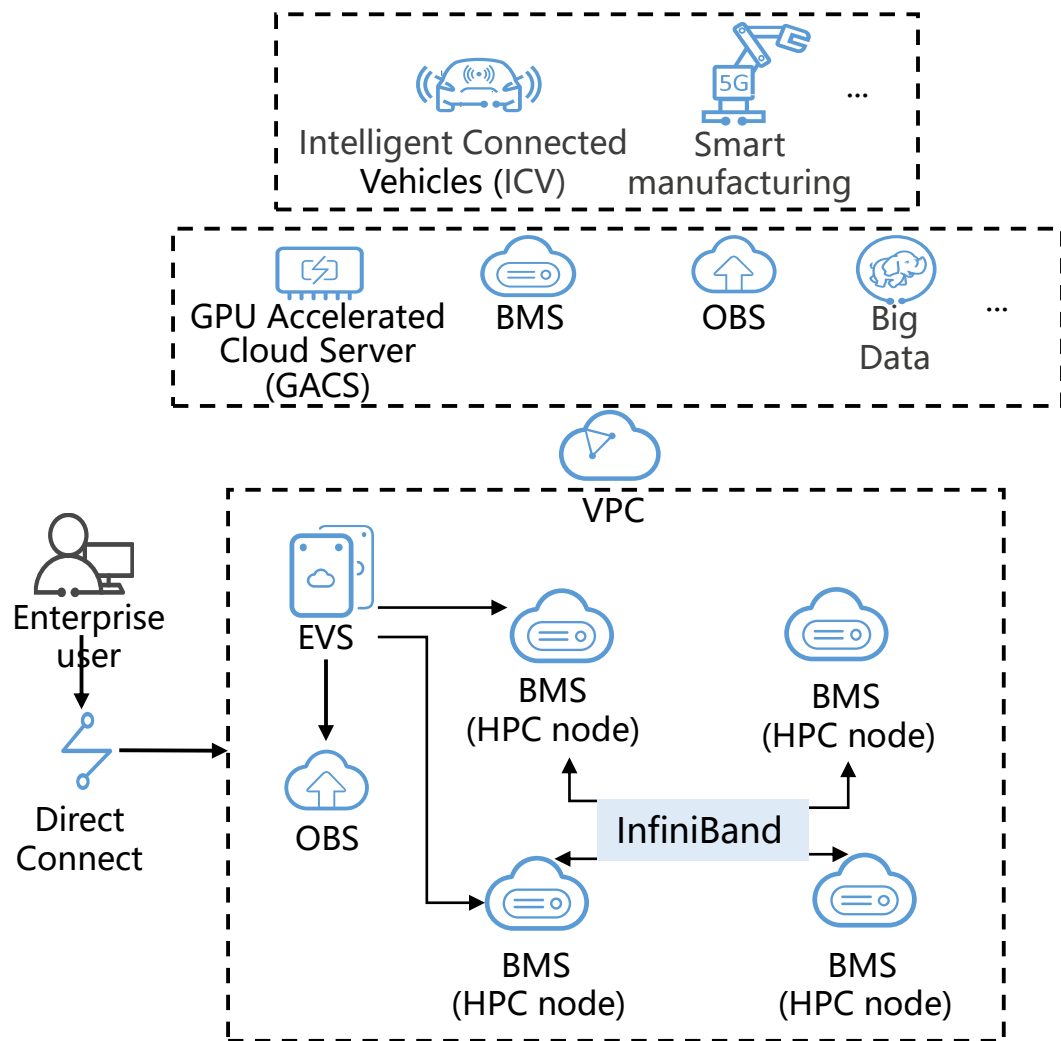
Application Scenarios

Core database. Multiple BMS flavors are available and shared EVS disks can be attached to BMSs, providing the performance and security required by core databases.

Recommendation Reasons

- Requirements: Some critical database services cannot be deployed on VMs and must be deployed on physical servers that have dedicated resources, isolated networks, and assured performance.
- Solution: The BMS service meets these database service requirements by providing high-performance servers dedicated to individual users.

Scenarios - High Performance Computing (HPC)



Application Scenarios

Supercomputing centers and DNA sequencing. For high performance and high throughput scenarios, BMSs with the latest CPUs, coupled with a 100 Gbit/s network, provide low latency and high performance services.

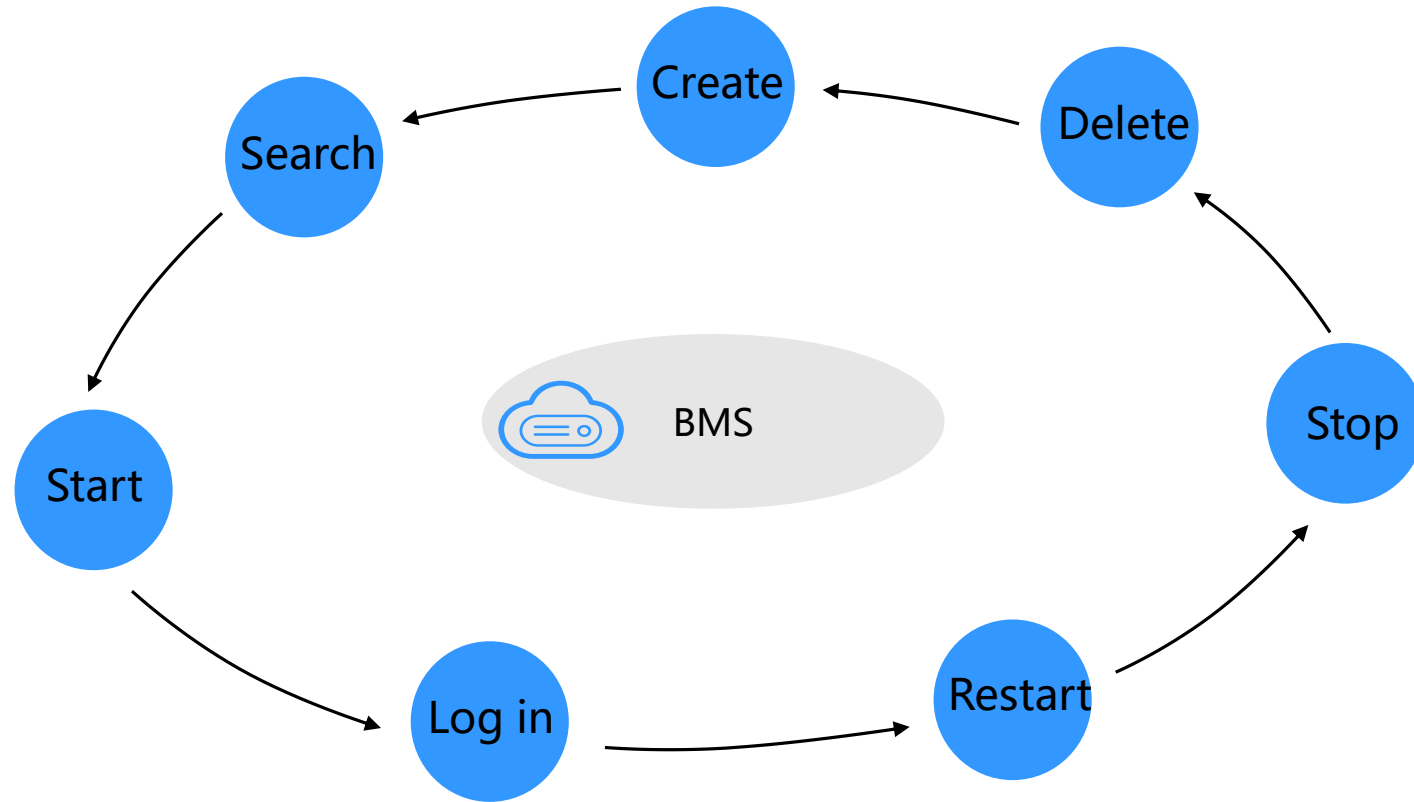
Recommendation Reasons

- Requirements: In HPC scenarios, such as supercomputer centers and DNA sequencing, massive volumes of data need to be processed and the computing performance, stability, and real-time responsiveness need to be stellar.
- Solution: HPC node (BMS)
 - Low latency: 100 Gbit/s, isolated, microsecond-level latency network
 - High performance: the latest Intel CPUs
 - Convenient scale-up: open APIs for easy ecosystem integration

Comparisons Between a BMS, ECS, and Physical Server

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

BMS Lifecycle Management



Self-service application, simple configuration, provisioning in minutes,
and full-lifecycle management

Creating a BMS - Basic Configuration

- Configure the region, AZ, flavor, and image.

Region

AP-Singapore

BMSs in different regions cannot communicate with each other over an intranet. For low network latency and quick resource access, select the nearest region.

AZ ?

AZ1AZ2AZ3

Flavor

Flavor name	CPU	Memory	Local Disk	Extended Configuration
<input checked="" type="radio"/> physical.d2.large	24 cores 2*12Core 5118...	12*16 GB DDR4	2*600G SAS System Disk RAID 1 + 12*1...	2x2*10GE

Image

Public imagePrivate imageShared image

CentOS

CentOS 7.4 64bit for BareMetal

Creating a BMS - Network Configuration

- Configure the VPC, NICs, enhanced high-speed NICs, security groups, and the EIP.

VPC ?

vpc-default

Create VPC

NIC

Primary NIC ?

subnet-default(192.168.0.0/24)

User-configured IP address

View In-Use IP Address

+ Add NIC You can add 1 more NICs.

Enhanced High-Speed NIC ?

Enhanced High-Speed NIC1

--Select--

+ Add Enhanced High-Speed NIC You can add 1 more NICs.

Security Group ?

default (Inbound:TCP/3389, 22 | Outbound: -)

Create Security Group

Inbound: TCP/3389, 22 | Outbound: -

[Learn how to configure a security group.](#)

Ensure that the selected security group allows access to port 22 (SSH-based logins for Linux), 3389 (logins for Windows) and ICMP (for ping operations).
[Configure Security Group Rules](#)

EIP ?

Automatically assign

Use existing

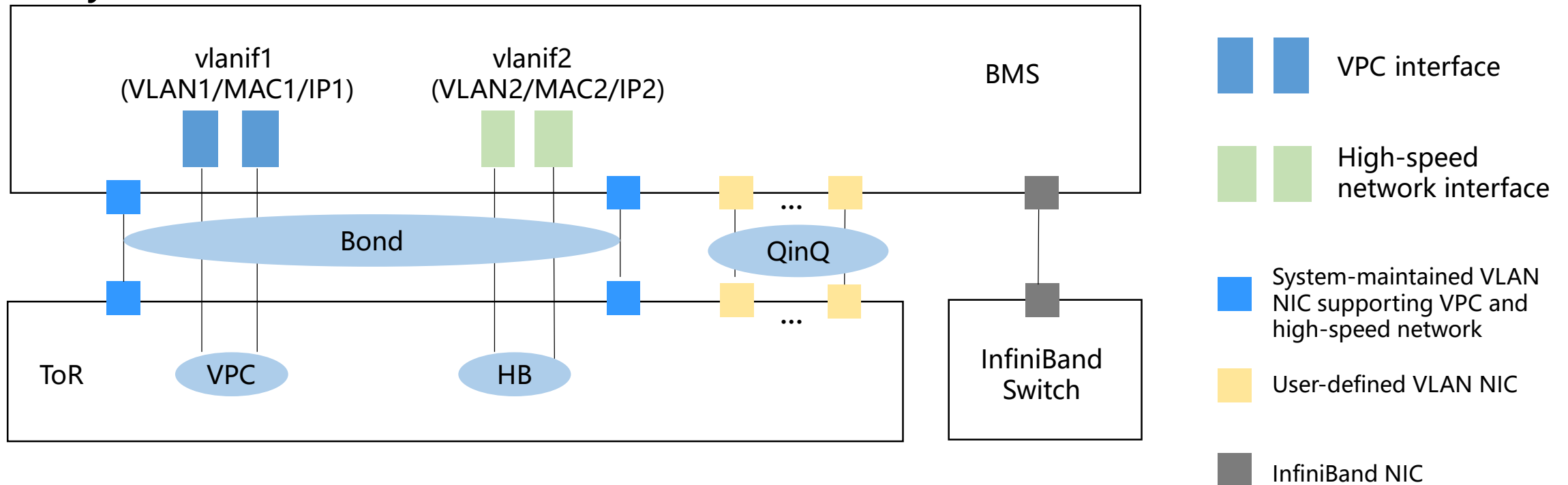
Not required

View EIPs

A BMS without an EIP cannot access the Internet. However, it can still be used as a service BMS deployed in a cluster or on a private network.

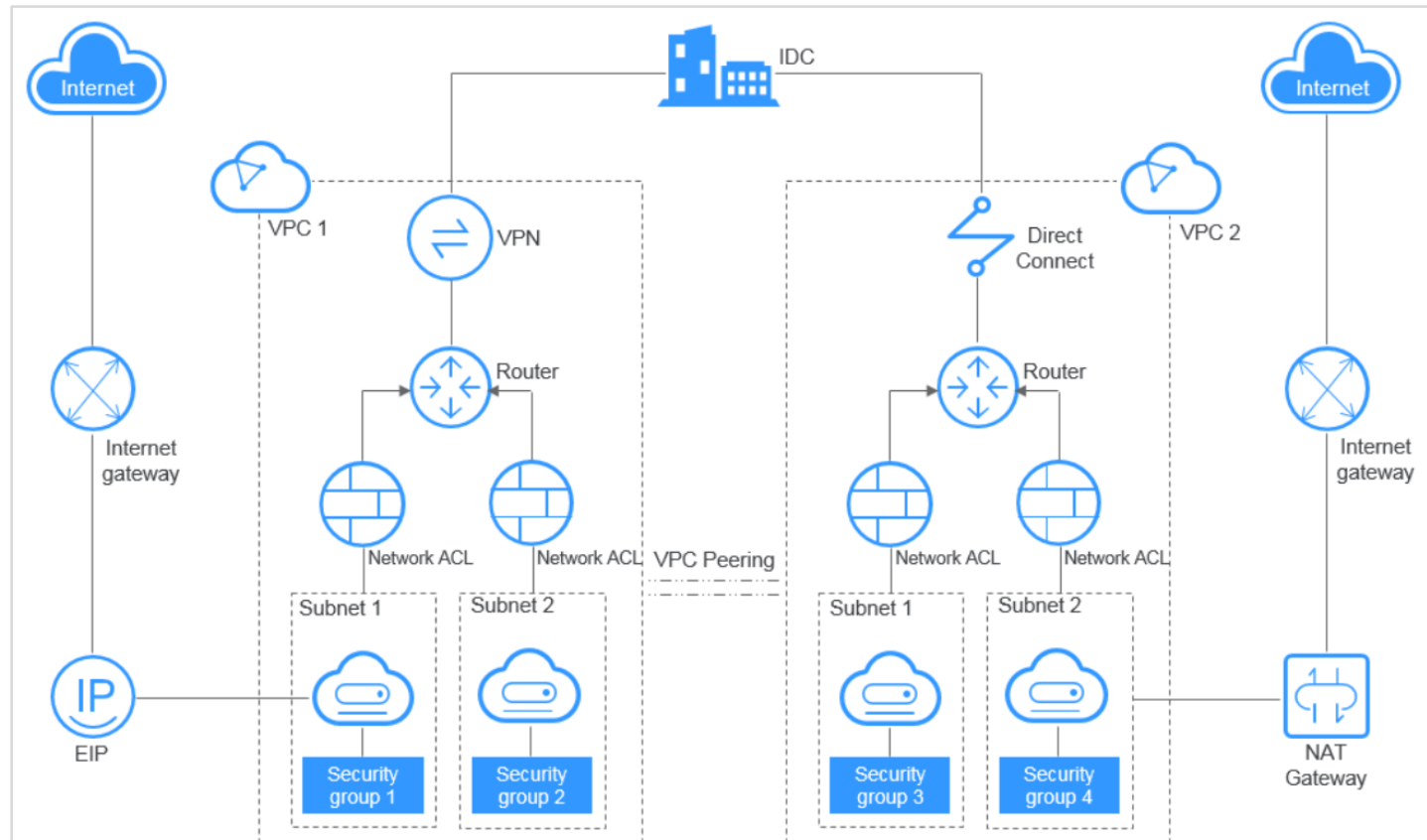
BMS Network

- 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.



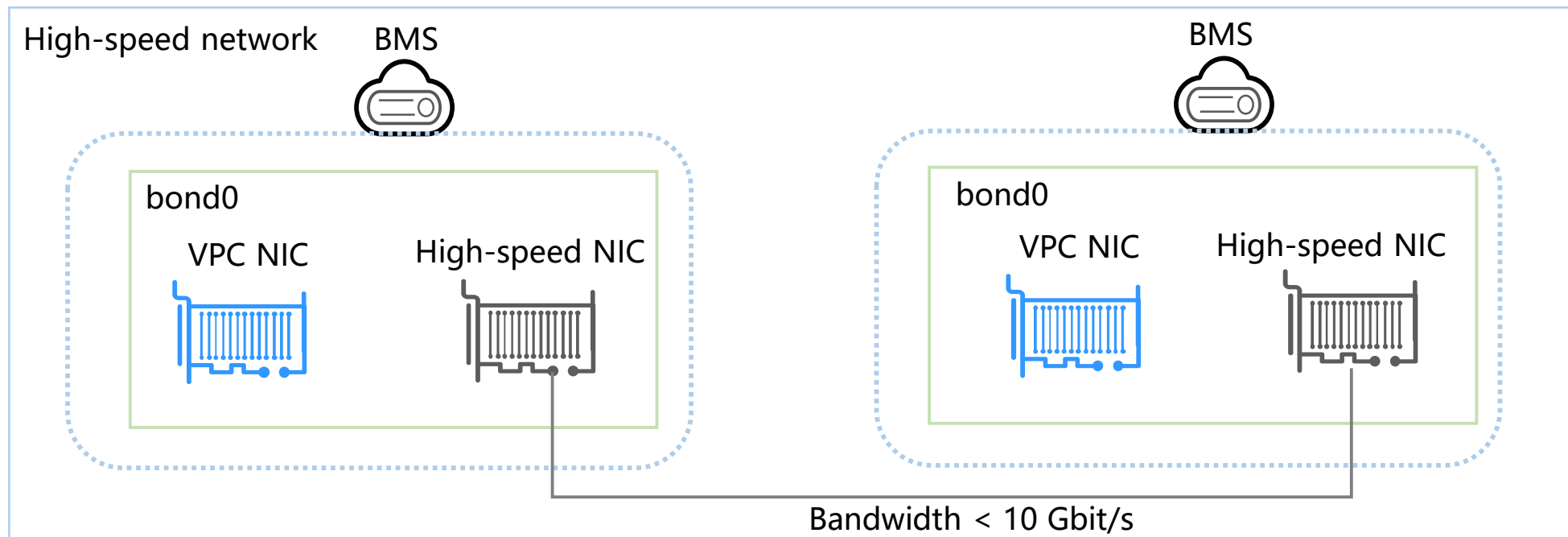
BMS Network - VPC

- 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.



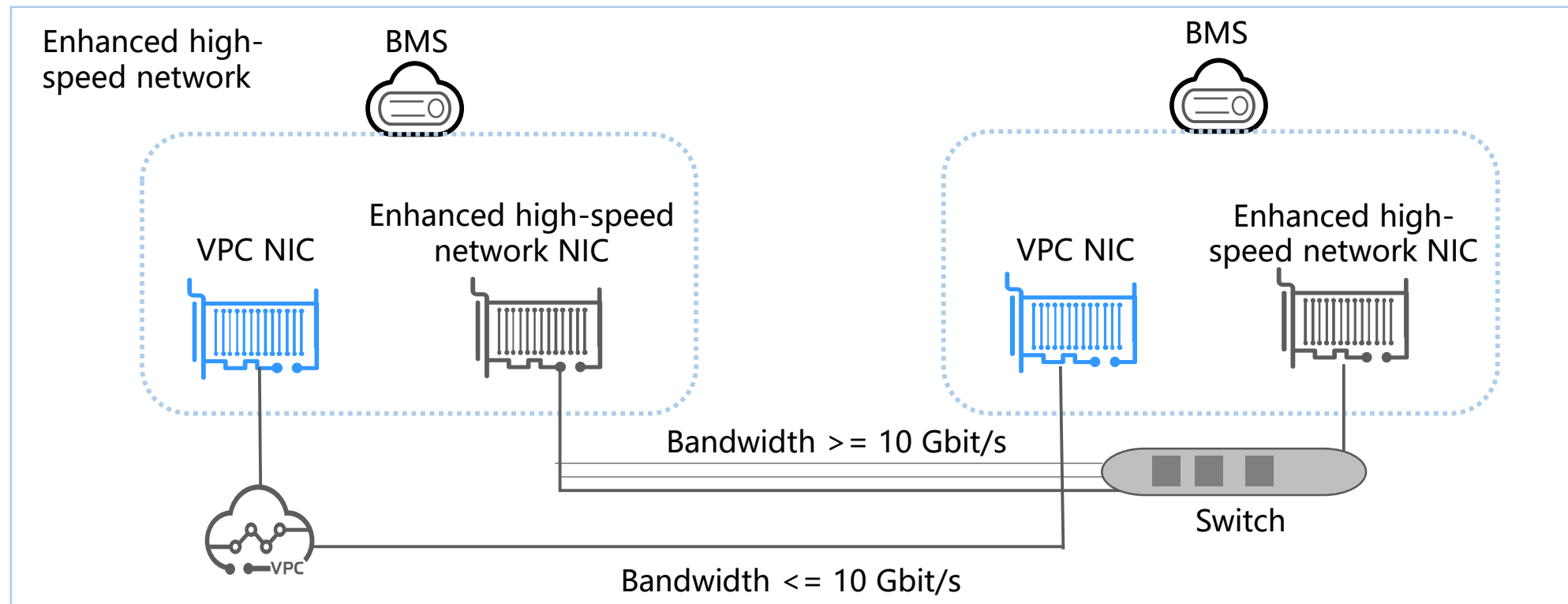
BMS Network - High-Speed Network

- A high-speed network is an internal network between 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.
- High-speed networks share the same physical plane with VPCs. A high-speed network carries only east-west traffic and supports only layer-2 communication because it does not support layer-3 routing.



BMS Network - Enhanced High-Speed Network

- An enhanced high-speed network is a high-quality, high-speed, low-latency internal network for BMSs to communicate with each other.



BMS Network - User-defined VLAN

- You can allocate VLAN subnets to isolate traffic in scenarios such as SAP HANA and virtualization. User-defined VLAN NICs are in pairs. You can configure NIC bonds to achieve high availability.

Create User-defined Network

The gateway IP address of a user-defined subnet cannot be used for any BMS.

Name

virtualnetwork-b439

VPC ?

vpc-default(192.168.0.0/16)

View VPC

* AZ ?

cn-south-1e

* User-defined Subnet

10 . 10 . 0 . 0 / 16

4000

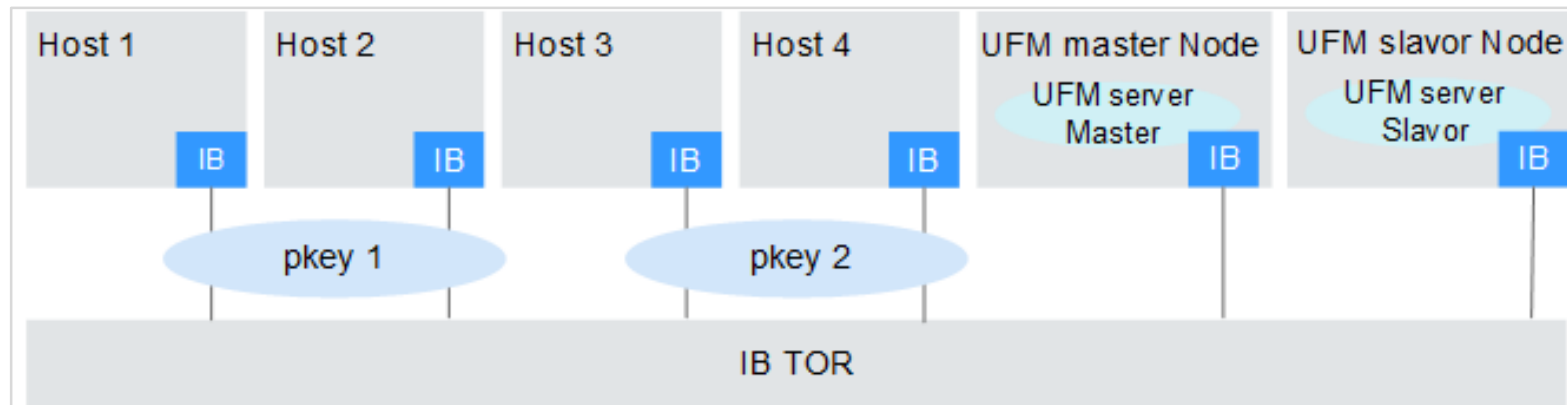
+ Add User-defined Subnet You can add 199 more subnets.

OK

Cancel

BMS Network - InfiniBand Network

- An InfiniBand network features low latency and high bandwidth, and is good for high performance computing (HPC) projects. An InfiniBand network supports two communication modes: RDMA and IPoIB.
- To create an InfiniBand network, you must select a flavor that supports InfiniBand NICs during BMS creation.



Creating a BMS - Advanced Configuration

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

Login Mode

Key pairPassword

The private key is required for you to log in to the BMS. It is important that this key not be lost.

Key Pair ? --Select-- Create Key Pair

To click Remote Login to log in to a Linux BMS in key pair login mode, you must set a login password after the BMS is created. [Learn how](#) to set the password.

Advanced Settings

Do not configureConfigure now

BMS Name

bms-837d

If you buy more than one BMS at a time, the system automatically adds a suffix to the name of each BMS, for example, bms-0001, bms-0002...

Using a BMS - Reinstalling the OS

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

Reinstall OS

Reinstalling the OS is to use the original image of a BMS to install the BMS again. This operation does not incur extra fees.

1. OS reinstallation will delete system disk data, including data on the system partition and other partitions. Back up your system disk data before performing this operation.
2. After the OS reinstallation on a BMS is successful, the BMS will be automatically started.
3. To reset the BMS password, you must install a plug-in on the BMS after it is created. [Click here for more details.](#)
4. After the OS is reinstalled, customized configurations (such as DNS and hostname) will be reset.

Image

CentOS 7.2 64bit for BareMetal

OS Bit64-bit

Login Mode

Key pair

Password

The private key is required for you to log in to the BMS. It is important that this key not be lost.

Key Pair

--Select--

Create Key Pair

OK

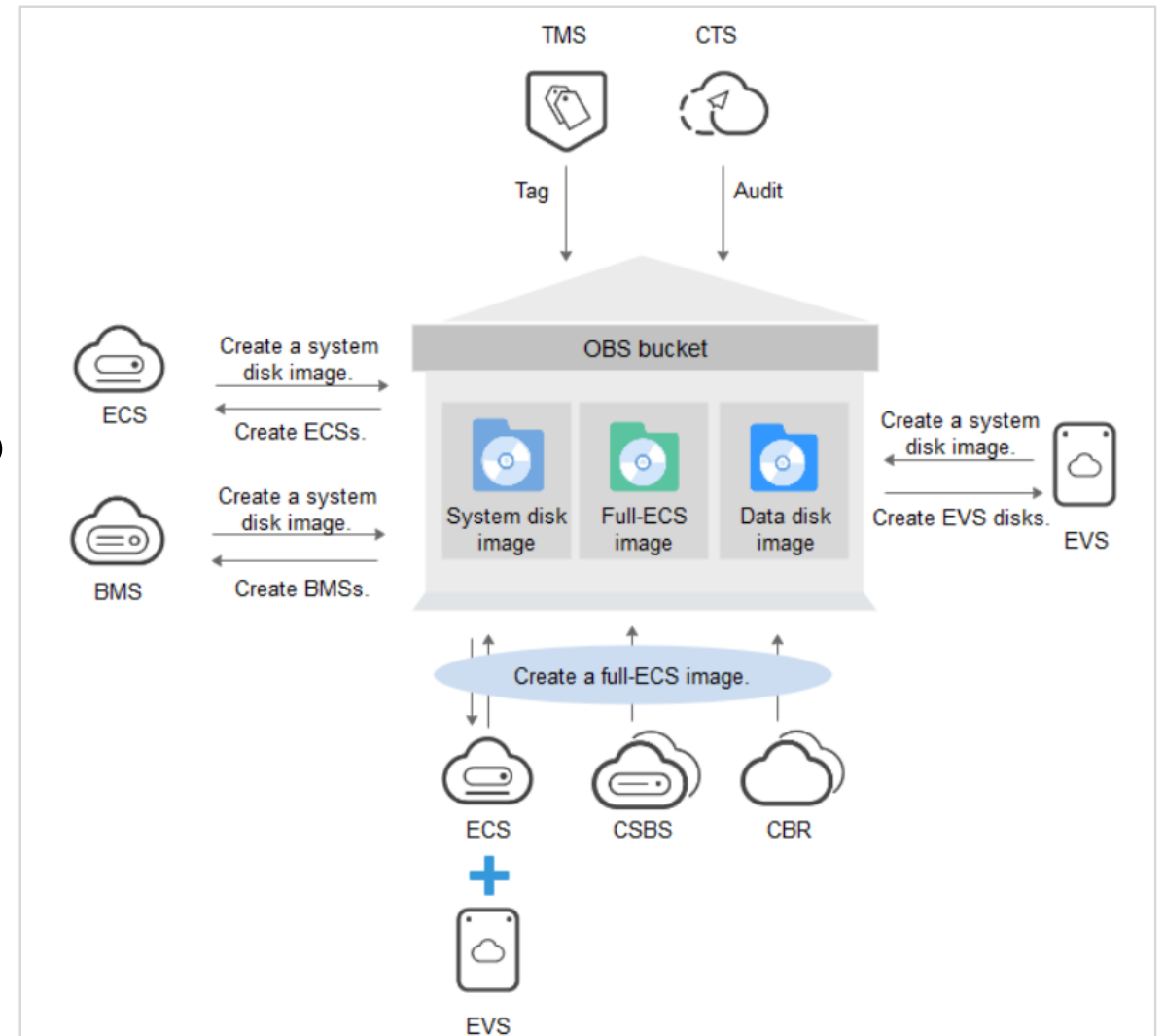
Cancel

Contents

1. Elastic Cloud Server (ECS)
2. Bare Metal Server (BMS)
- 3. Image Management Service (IMS)**
4. Auto Scaling (AS)
5. Cloud Container Engine (CCE)
6. Other Compute Services

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 on the cloud.



Why IMS?

Secure

- Thoroughly tested mainstream OSs for public images
- Multiple copies of image files stored on Object Storage Service (OBS)
- Private images encrypted by Key Management Service (KMS)

Convenient

- Different types of images for cloud server creation
- Multiple methods for private image creation
- Private images can be shared or replicated between accounts, regions, and cloud platforms

Centralized

- Self-service platform for image management
- Batch deployment and upgrade of application systems
- Compliance with industry standards for service migration between cloud platforms

Flexible

- Image management on the console or using APIs
- Public images, private images, and Marketplace images for flexible choices

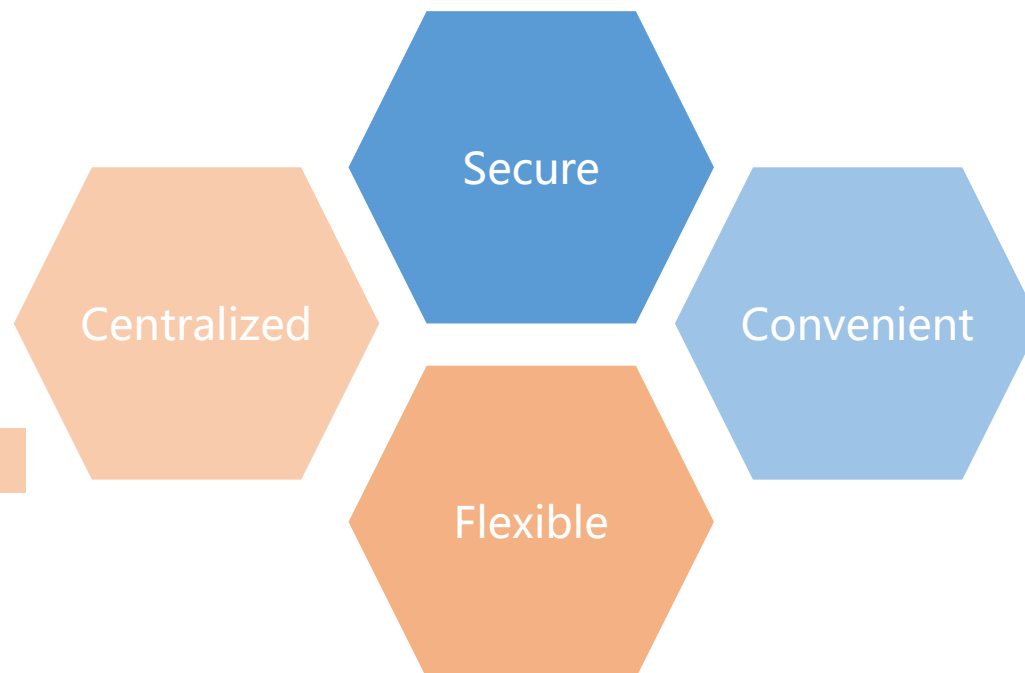


Image Types

- A public image is a standard image provided by the cloud platform. It contains an OS and various preinstalled applications, and is available to all users.
- A private image is created by users and is visible only to the user who created it.
- A shared image is a private image another user has shared with you.
- A Marketplace image is a third-party image published in the Marketplace. It has an OS, various applications, and custom software preinstalled.



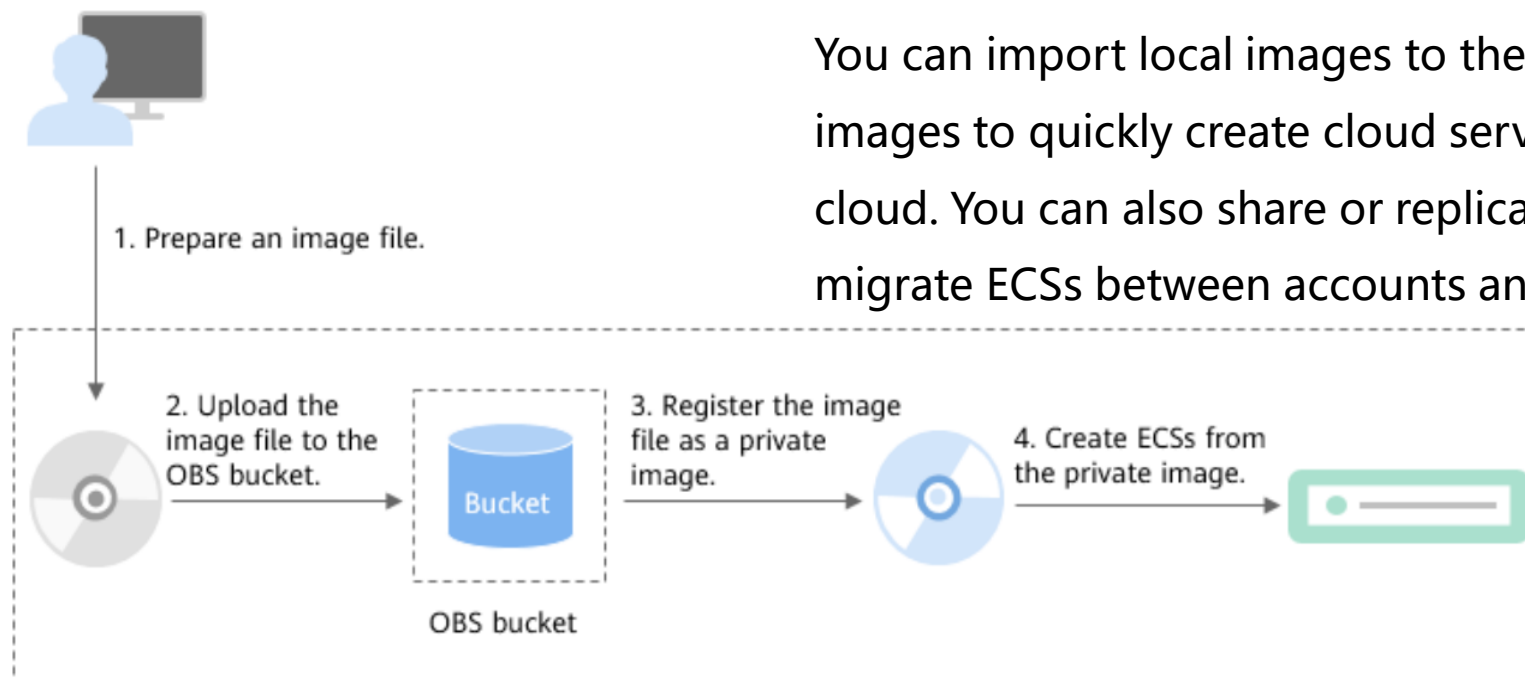
Scenarios - Migrating Servers to the Cloud or in the Cloud

Application Scenarios

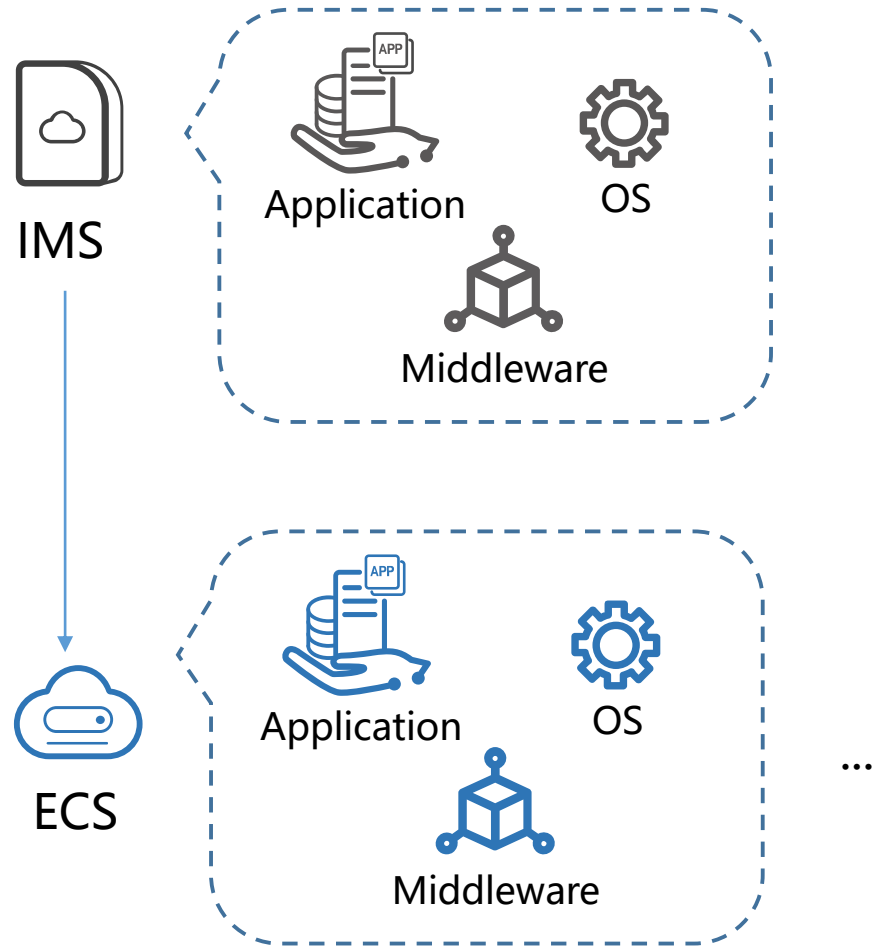
Migrating servers to the cloud or in the cloud

Recommendation Reasons

You can import local images to the cloud platform and use the images to quickly create cloud servers for service migration to the cloud. You can also share or replicate images across regions to migrate ECSs between accounts and regions.



Scenarios - Deploying a Specific Software Environment



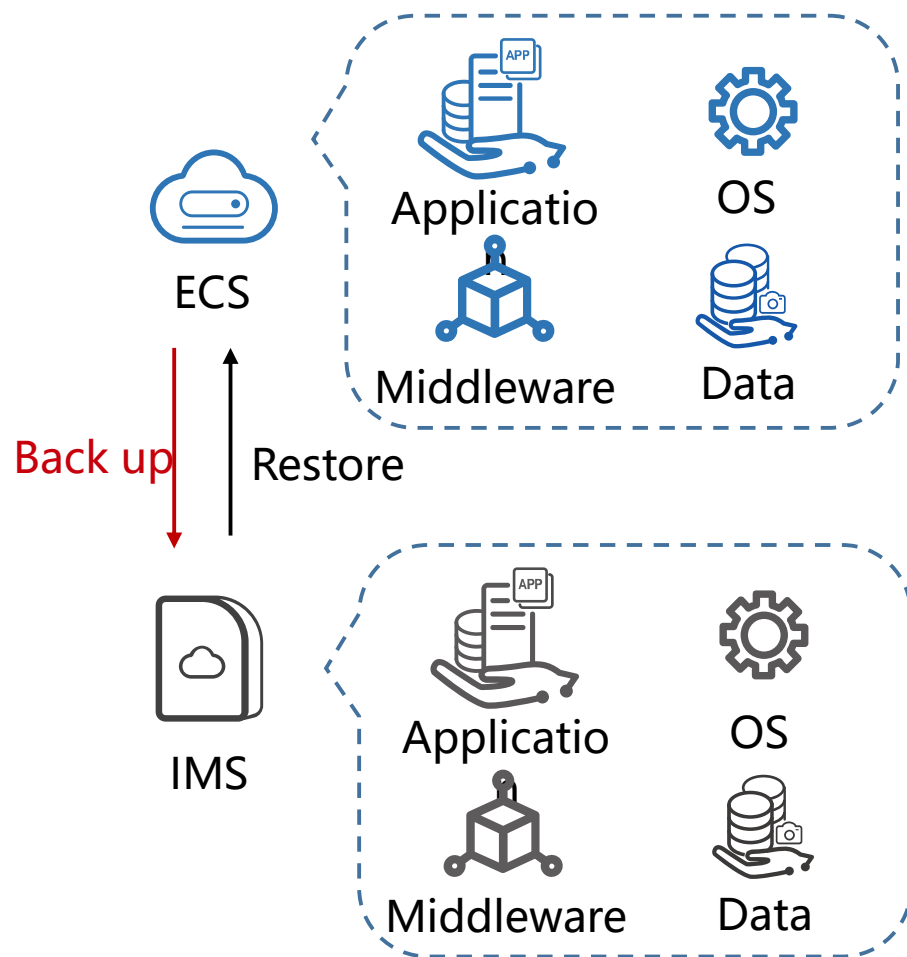
Application Scenarios

Deploying a specific software environment

Recommendation Reasons

You can 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.

Scenarios - Backing Up Server Environments



Application Scenarios

Backing up server environments

Recommendation Reasons

You can create an image from an ECS to back up the ECS. If the ECS breaks down for some reason, you can use the image to restore it.

Creating a Private Image

Creating a system disk image from a Windows ECS

Creating a system disk image from a Linux ECS

Creating a full-ECS image from an ECS

Creating a full-ECS image from a CSBS backup

Creating a data disk image from an ECS

Creating a Windows system disk image from an external image file

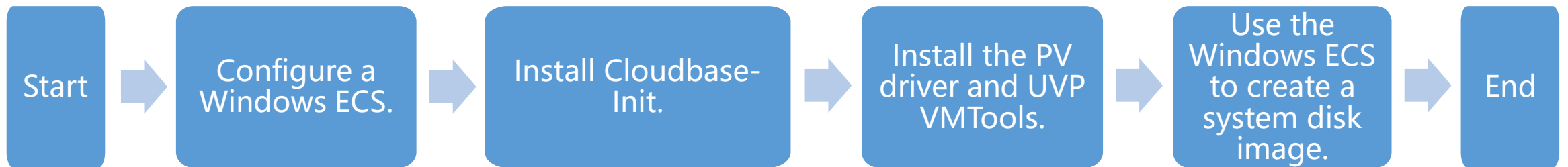
Creating a Linux system disk image from an external image file

Creating a data disk image from an external image file

...

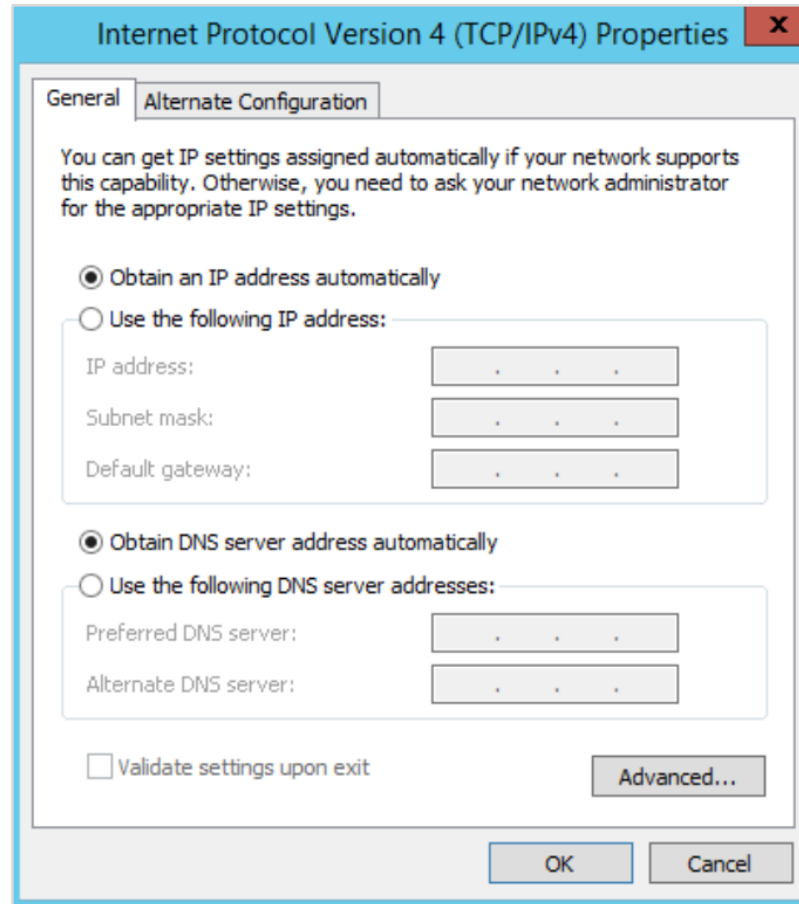
Creating a System Disk Image from a Windows ECS

This course will show how to create a system disk image from a Windows ECS as an example.



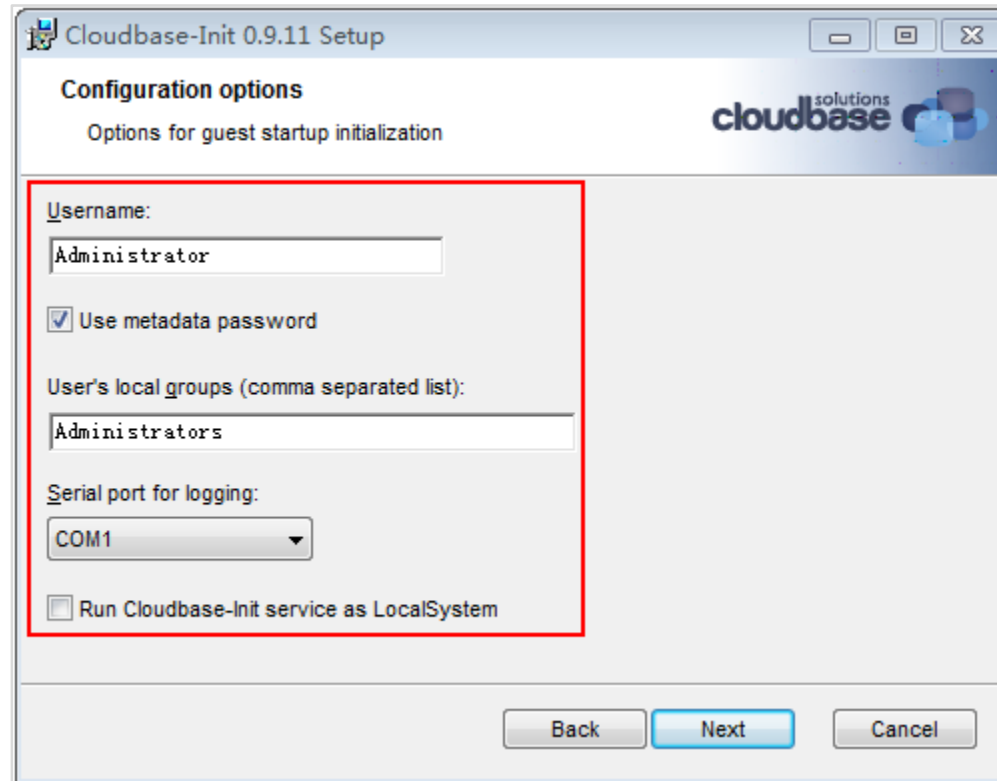
Configuring a Windows ECS

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



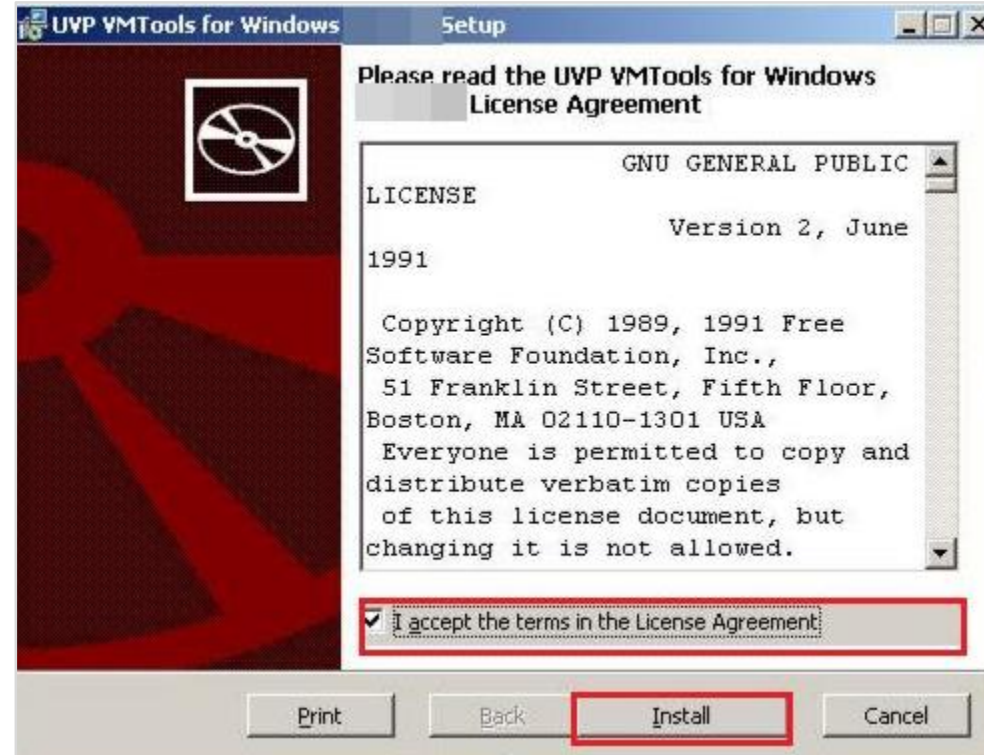
Installing Cloudbase-Init

- To ensure that ECSs created from a private image are configurable, you are advised to install Cloudbase-Init on the ECS before using it to create a private image.



Installing the PV Driver and UVP VMTools

- 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.



Using a Windows ECS to Create a 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 list.

Image Type and Source

★ Type

System disk image

Full-ECS image

Data disk image

ISO image

★ Source

ECS

Image File

- You can only use a running or stopped ECS to create a private image.
- You need to first customize and optimize the ECS to suit your needs. For example, you need to install Cloud-Init if the ECS runs Linux and install Cloudbase-Init if the ECS runs Windows. [Learn more](#)
- Do not perform any operation on the selected ECS or associated resources during image creation.

All statuses

Name

Q

C

Name	OS	Status	Private IP Address	Created
<div>▼</div> ecs-373896	Windows Server 2012 R2 Da...	Running	192.168.1.104	Aug 03, 2018 23:54:12 G...

Selected: ecs-373896|OS: Windows Server 2012 R2 Datacenter 64bit|System Disk: High I/O | 40 GB

Image Management - Modifying Image Information

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

Modify Image

★ Name

Description

0/1,024

Minimum Memory Ensure that the minimum memory size of an image is set to its original size before you reinstall OSs of the ECSs that were created using the image.

Unlimited	1 GB	2 GB	4 GB	8 GB
16 GB	32 GB	64 GB	128 GB	

Maximum Memory

Unlimited	4 GB	32 GB	64 GB
128 GB			

NIC Multi-Queue

Supported	Not supported
-----------	---------------

Boot Mode

BIOS	UEFI
------	------

The boot mode must be the same as that of the OS contained in the image file. Otherwise, ECSs created from this system disk image will fail to start.

OK Cancel

Image Management - Deleting an 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.

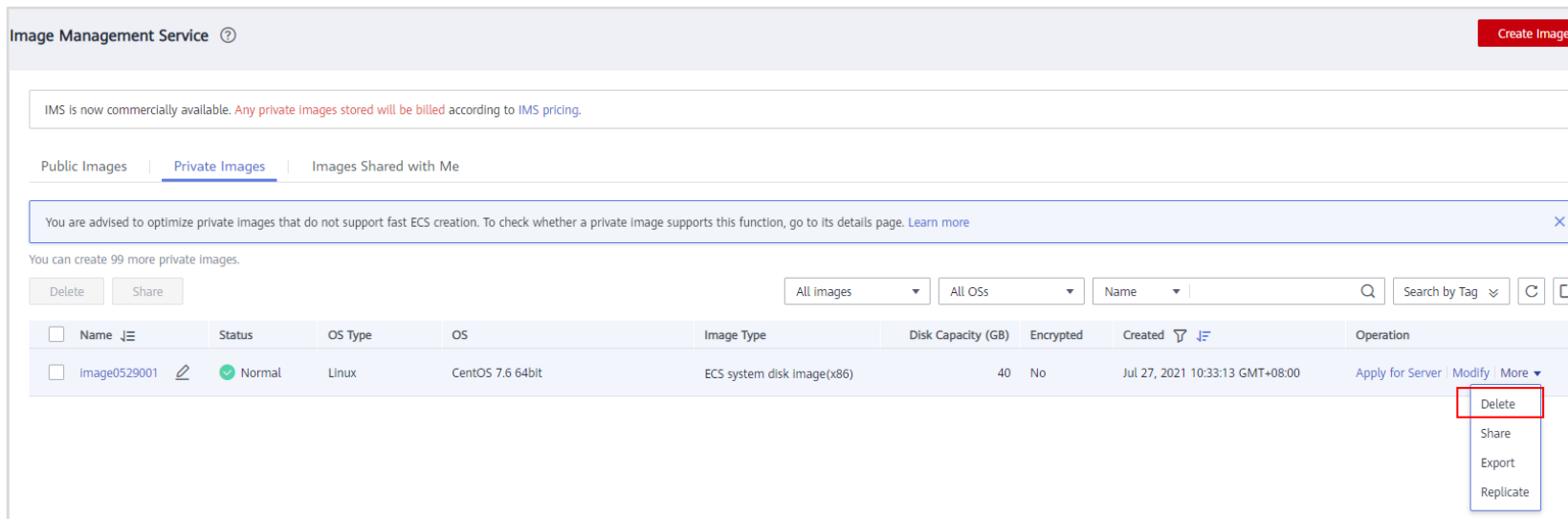


Image Management - Sharing an Image

- You can share your private images.

×

Share Image

Enter the account name of the image recipient. [Learn how](#) to obtain an account name.

✱

Enter an account name.

Add

Account Name	Project Name	Project ID	Operation
No data available.			

This image can be shared with a maximum of 256 tenants. You can share this image with 256 more tenants.

OK

Cancel

Image Management - Encrypting an Image

- You can create an encrypted image to securely store data.
- 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.

Image Information

☒ Enable automatic configuration [Learn more](#)

* Function ECS system disk image BMS system disk image

Architecture x86 ARM

If the system detects an architecture type different from that you set, the architecture type detected by the system prevails. If the system fails to detect the architecture type, the architecture type you set prevails.

Boot Mode BIOS UEFI

The boot mode must be the same as that of the OS contained in the image file. Otherwise, ECSs created from this system disk image will fail to start.

OS --Select OS-- --Select OS version--

If the OS you selected is different from the OS in the image file, IMS will use the OS in the image file if the OS can be detected. Otherwise, the OS you selected will be used for image creation. [View supported OSs](#)

* System Disk (GB) 40-1,024 The system disk size must be larger than the image file size.

[+ Add Data Disk](#) You can add 3 more data disks.

* Name

Encryption ☐ KMS encryption [?](#)

Image Management - Replicating an Image Within a Region

- You may need to replicate an image in the following scenarios:
- Creating an unencrypted version of an encrypted image
- Replicating an encrypted image
- Creating an encrypted version of an unencrypted image

×

Replicate Image

• The image size must be less than 128 GB.

Image Details

Name	discuz_centos6.5
Image Type	ECS system disk image
Image Size	1.04 GB
OS Type	Linux
OS	CentOS 6.5 64bit
Created	Jul 29, 2019 17:00:49 GMT+08:00

★ Name

copy_discuz_centos6.5

★ Enterprise Project ?

--Select an enterprise project--

↻

Description

0/1024

OK

Cancel

Image Management - Replicating an Image Across Regions

- You can replicate an image from one region to another and use the replicated image to create identical ECSs. This allows you to more quickly migrate services across regions.

Replicate Image

Image Size	9.7 GB
OS Type	Windows
OS	Windows-Server-2012-R2-Datacenter-64bit
Created	Dec 02, 2020 11:22:41 GMT+08:00

Replication Mode

Within Region

Across Regions

★ Name

copy_br-iaas-odin1_1202-new-2012-g5r

★ Destination Region

--Select--

★ Destination Project

★ IAM Agency

--Select--

[View Agency](#)

OK

Cancel

Image Management - Exporting an Image

- You can export an image if you want to:
 - Store the image on specified storage devices.
 - Use the image to create servers on other cloud platforms.

Export Image

- Images can be exported only to an OBS bucket with Standard storage, and the image size cannot exceed 1 TB.
- Images larger than 128 GB only support fast export, and the format of the exported images cannot be specified. See [Exporting Images](#).
- Images stored in the OBS will be billed according to standard [OBS Pricing](#).

Image Details

Image Name	image0529001
OS Type	Linux
OS	CentOS 7.6 64bit
Image Size	1.44 GB ?

Fast Export

☐ Enable

* Format

qcow2vmdkvhdzvhd?

* Name

.qcow2

* Storage Path

OK

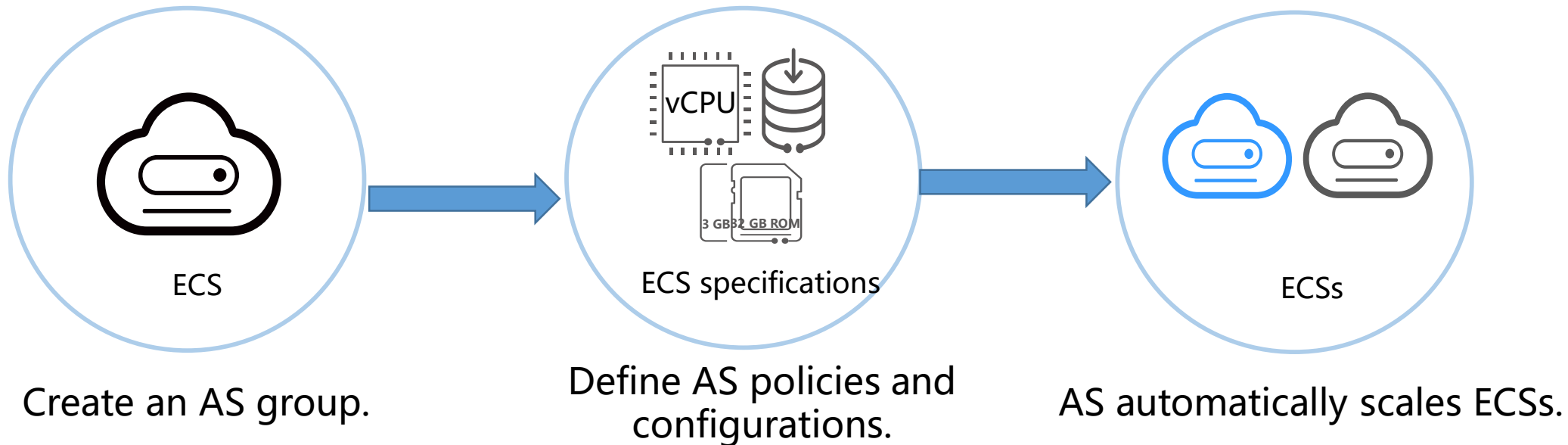
Cancel

Contents

1. Elastic Cloud Server (ECS)
2. Bare Metal Server (BMS)
3. Image Management Service (IMS)
- 4. Auto Scaling (AS)**
5. Cloud Container Engine (CCE)
6. Other Compute Services

What Is AS?

- Auto Scaling (AS) automatically adjusts resources to keep up with changes in demand based on pre-configured AS policies. You can specify AS configurations and policies based on service requirements. These configurations and policies free you from having to repeatedly adjust resources to keep up with service changes and spikes in demand, helping you reduce the resources and manpower required.



Why AS?

Automatic resource adjustment

AS automatically adjusts resources on demand for applications.

Enhanced cost management

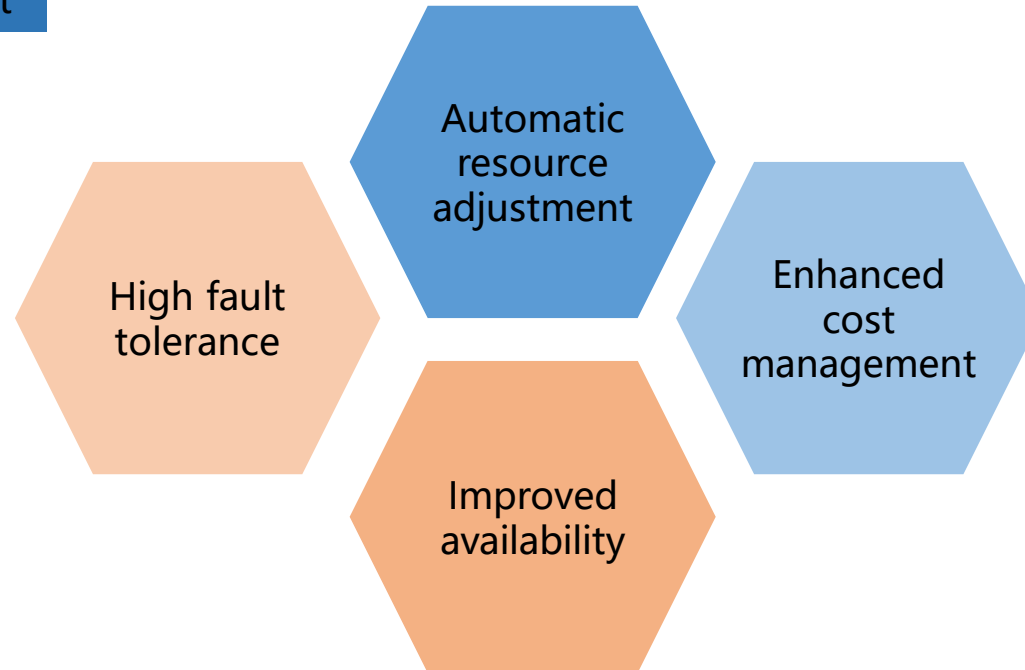
AS adjusts ECS instances and bandwidths on demand, enabling you to pay for what you need.

High fault tolerance

AS checks ECSs powering applications and replaces faulty instances with new ones.

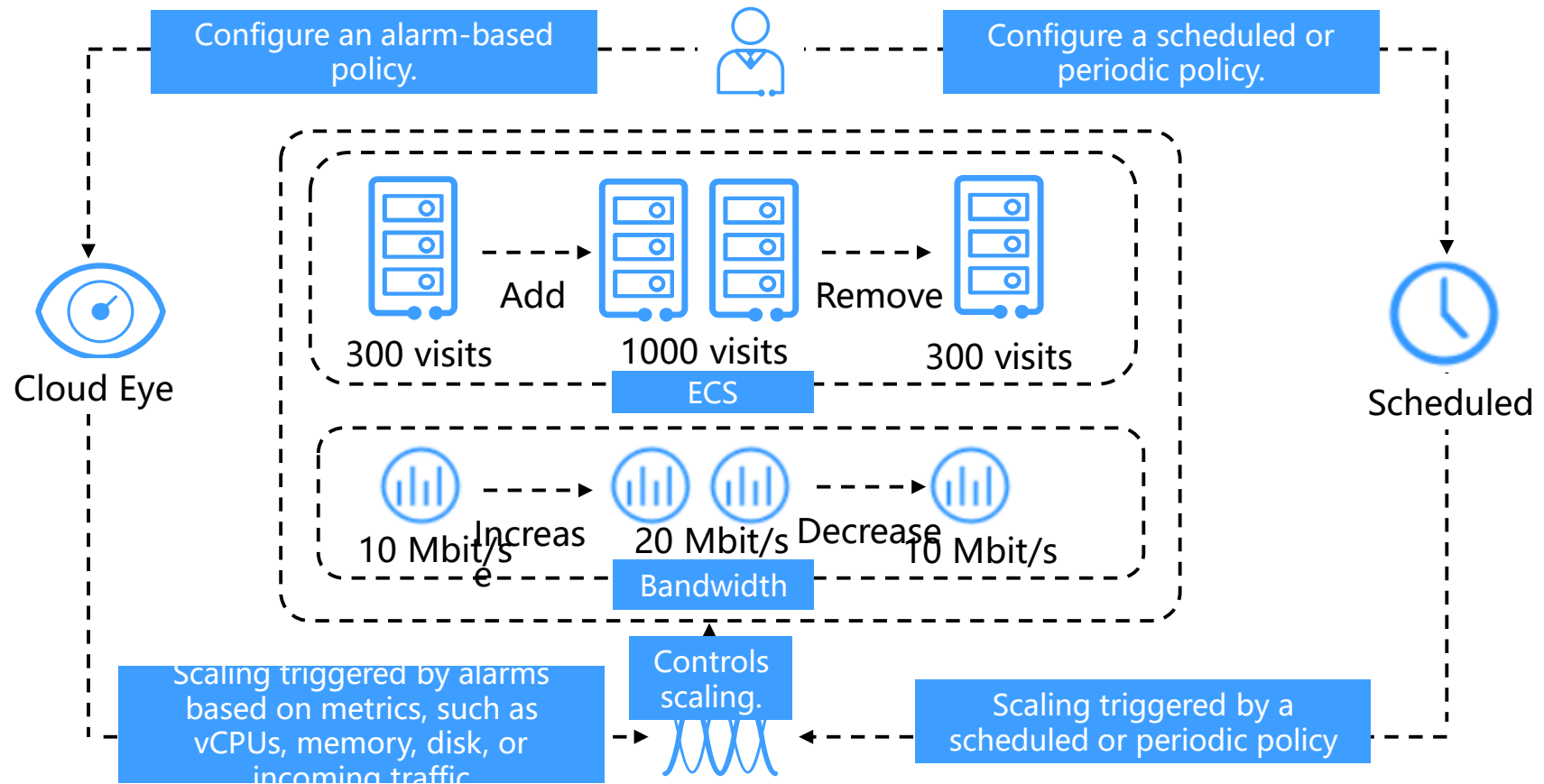
Improved availability

AS ensures proper resources deployed for applications.

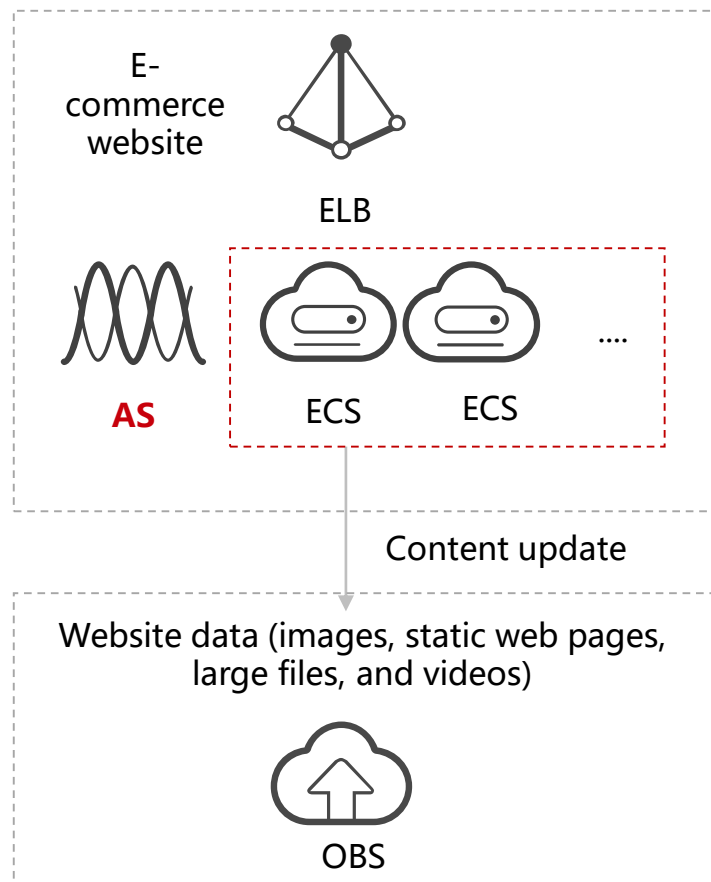


AS Architecture

- AS automatically adjusts compute resources based on service demands and configured AS policies. The number of ECS instances changes to match service demands, ensuring service availability.



Scenarios – Web Applications



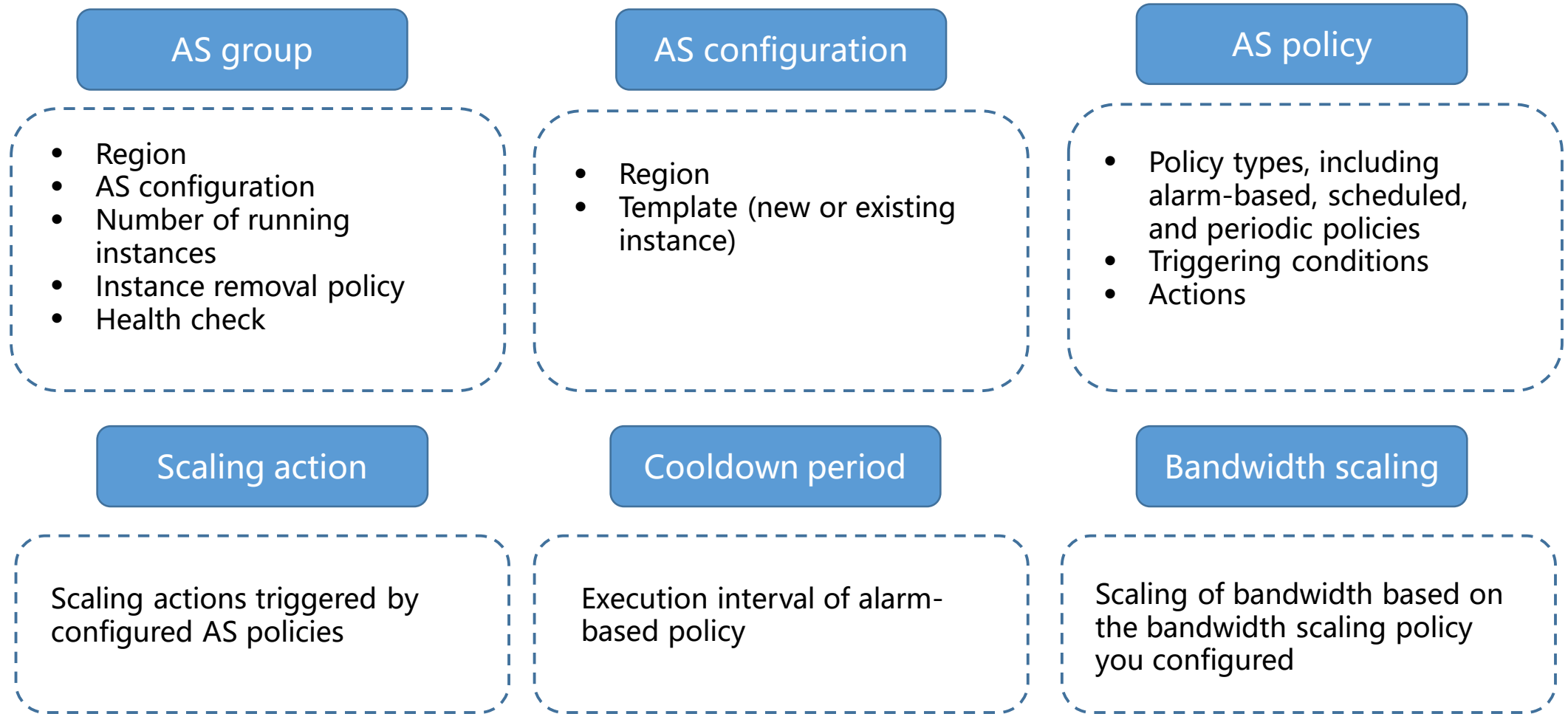
Application Scenarios

- E-commerce websites
- Heavy-traffic web portals

Recommendation Reasons

- E-commerce: During big promotions, E-commerce websites need more resources. AS automatically scales out ECS instances and bandwidth within minutes to ensure that promotions go smoothly.
- Heavy-traffic portals: Service load changes are difficult to predict for heavy-traffic web portals. AS dynamically scales in or out ECS instances based on monitored ECS metrics, such as vCPU usage and memory usage.

AS Basic Concepts



Getting Started with AS



Creating an AS Configuration

- Configuration Template options

Create a specifications template

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

Use specifications of an existing ECS

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

Creating an AS Group

- An AS group consists of a collection of instances and AS policies that have similar attributes and apply to the same scenario. It is the basis for enabling or disabling AS policies and performing scaling actions.
- AS automatically scales in or out instances or maintains a fixed number of instances in an AS group through scaling actions triggered by configured AS policies.
- When creating an AS group, you need to configure parameters, such as Max. Instances, Min. Instances, Expected Instances, and Load Balancing.

1

★ Multi-AZ Extension Policy ? ☒ Load-balanced ☐ Sequenced

★ Name

★ Max. Instances

★ Expected Instance ?

★ Min. Instances

2

★ Instance Removal Policy

EIP Do not release

If you select Release, EIPs bound to ECSs are released when the ECSs are removed from the AS group. Otherwise, EIPs will only be unbound from the ECSs.

Data Disk Do not release

If you select Release, data disks attached to ECSs are deleted when the ECSs are removed from the AS group. Otherwise, data disks will only be detached from the ECSs.

★ Health Check Method ?

When a protected instance is detected to be abnormal in a health check, AS removes the instance from the AS group and creates a new one.

★ Health Check Interval ?

Creating an AS Policy

- Main parameters: Policy Type and Cooldown Period

Add AS Policy

Policy Name

as-policy-8da5

Policy Type

Alarm

Scheduled

Periodic

Alarm Rule

Create

Use existing

Rule Name

as-alarm-8db0

Monitoring Type

System monitoring

Custom monitoring

Trigger Condition

CPU Usage

▼

Max.

▼

>

▼

%

To determine if an OS supports metrics Memory Usage, Inband Outgoing Rate, and Inband Incoming Rate, see [Elastic Cloud Server User Guide](#). Before using Agent to monitor metrics, make sure that the Agent plug-in has been installed on all instances in the AS group. [Learn how](#) to install the Agent plug-in.

Monitoring Interval

5 minutes

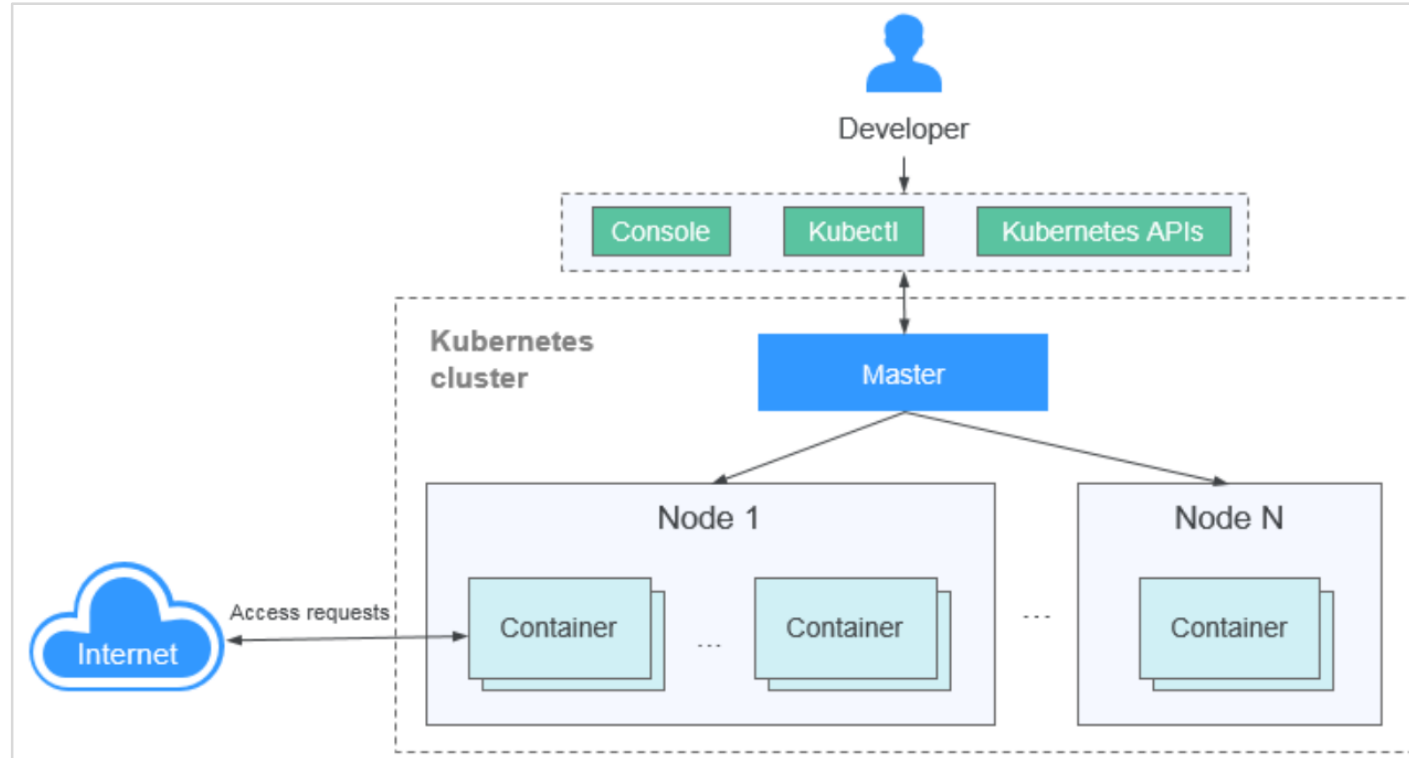
▼

Contents

1. Elastic Cloud Server (ECS)
2. Bare Metal Server (BMS)
3. Image Management Service
4. Auto Scaling (AS)
- 5. Cloud Container Engine (CCE)**
6. Other Compute Services

What Is CCE?

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



Why CCE?

User-friendly

- Create Kubernetes clusters in a few clicks on the console.
- Scale clusters and workloads on the console.
- Upgrade Kubernetes clusters on the console.
- Experience out-of-the box usability.
- Enjoy auto deployment and O&M of containerized applications.

High-performance

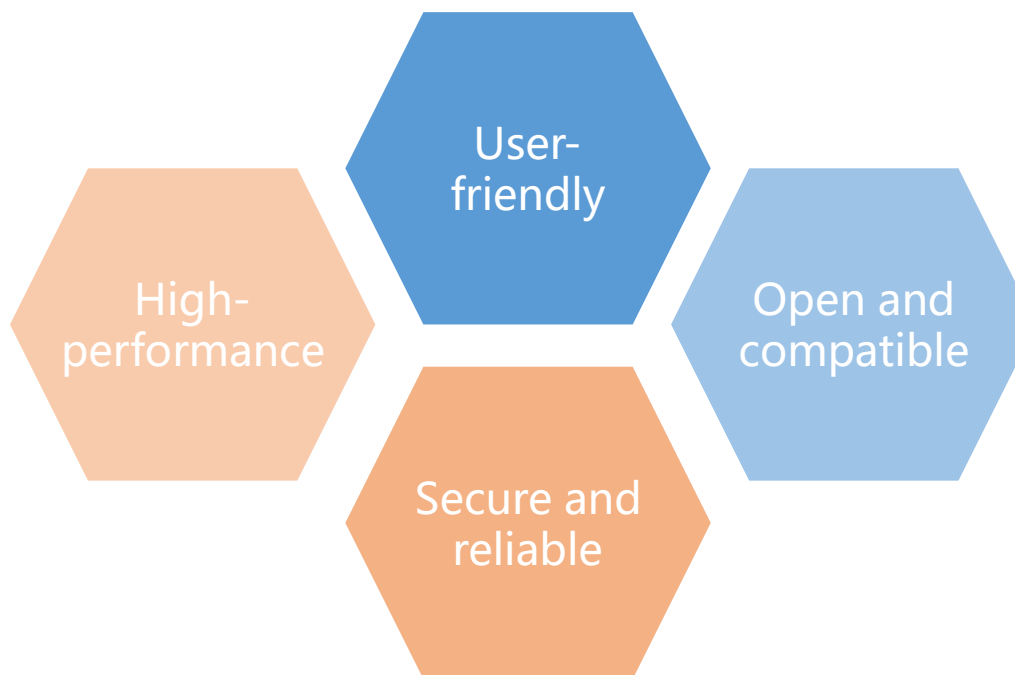
- Bare-metal servers with NUMA and high-speed InfiniBand NICs
- Industry-leading container engine

Open and compatible

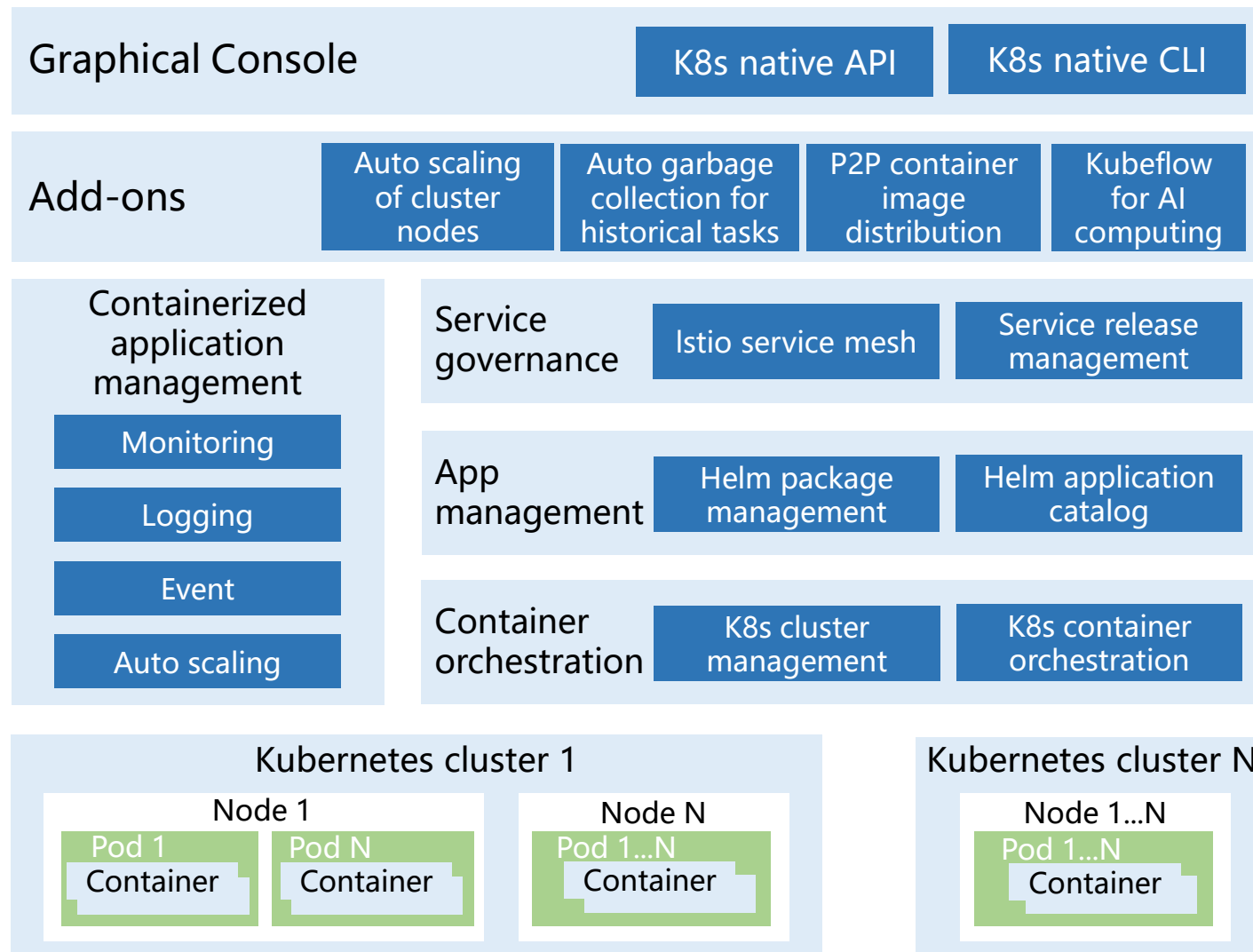
- Fully compatible with Kubernetes APIs and kubectl
- Easy management of large-scale container clusters

Secure and reliable

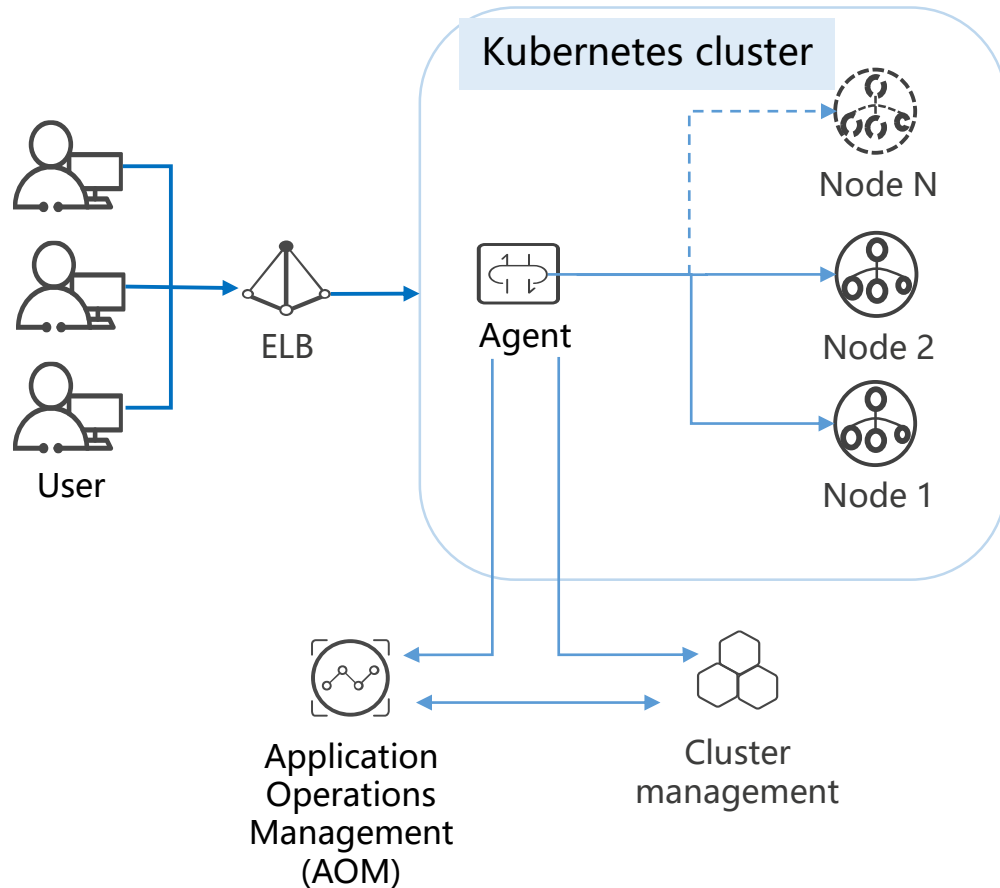
- You can deploy 3 master nodes on the cluster control plane for high availability (HA).
- Users have complete control of clusters they create.



CCE Architecture



Scenario - Auto Cluster Scaling



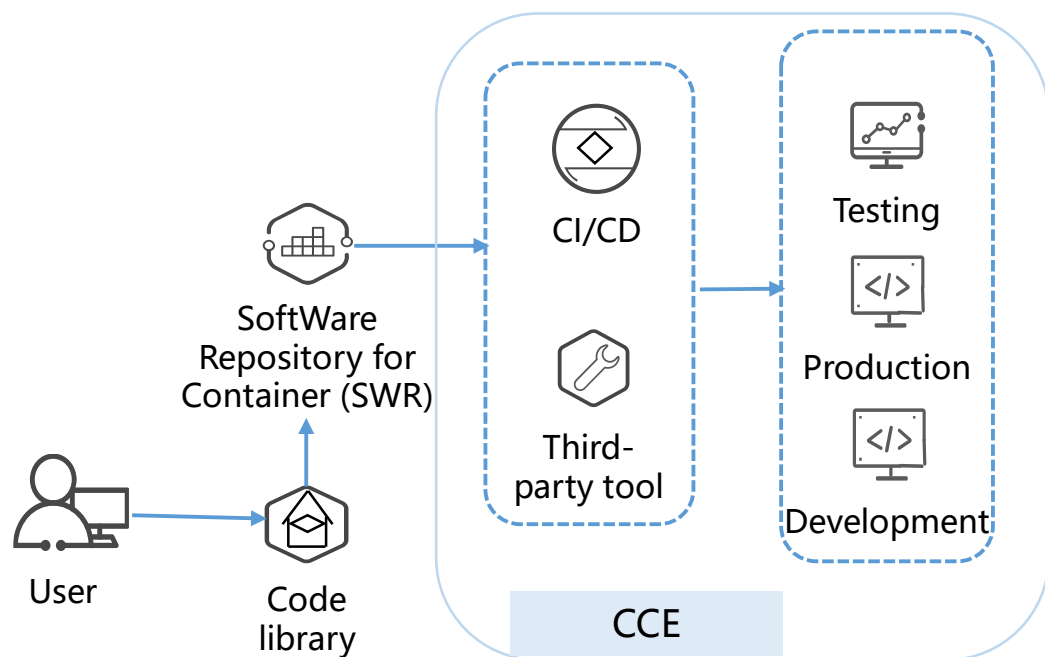
Function Description

CCE adjusts compute resources based on auto scaling policies to handle fluctuating service loads. Specifically, CCE automatically adds or reduces cloud servers for your cluster or containers for your workload.

Benefits

- Flexible: Multiple scaling policies are supported and containers can be provisioned within seconds when specific conditions are met.
- Highly available: Pods are automatically monitored and unhealthy pods will be replaced with new ones to ensure high service availability.
- Low cost: You are billed only for the cloud servers you use.

Scenario - DevOps



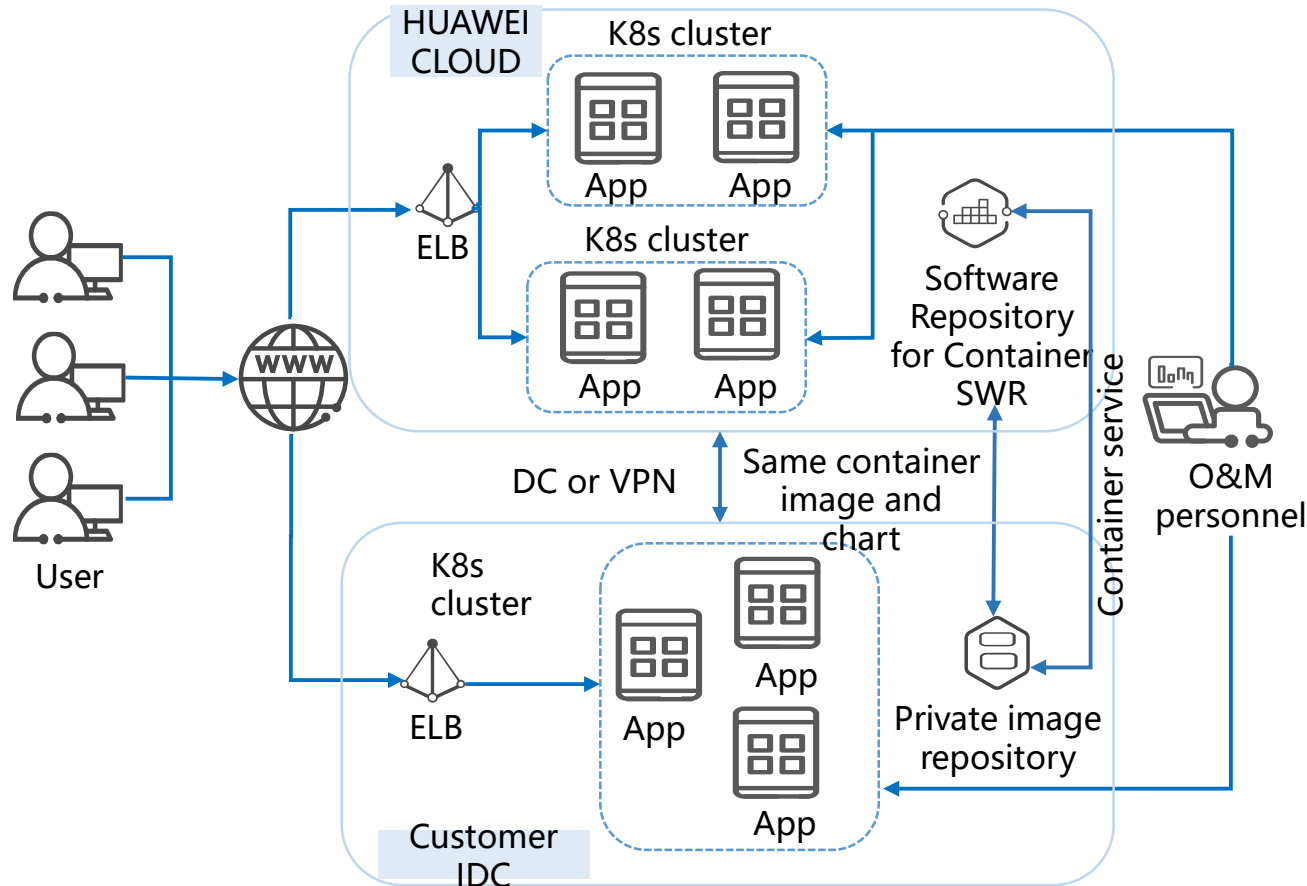
Function Description

CCE automatically completes code compilation, image build, grayscale release, and container-based deployment based on code sources. CCE can interconnect with your CI/CD systems. You can containerize traditional applications and deploy them in the cloud.

Benefits

- **Efficient CI/CD management:** Reduces scripting workload by more than 80% through streamlined process interaction.
- **Flexible integration:** Provides various APIs to integrate with existing CI/CD systems, facilitating customization.
- **High performance:** Allows for flexible task scheduling with a fully containerized architecture.

Scenario - Hybrid Cloud



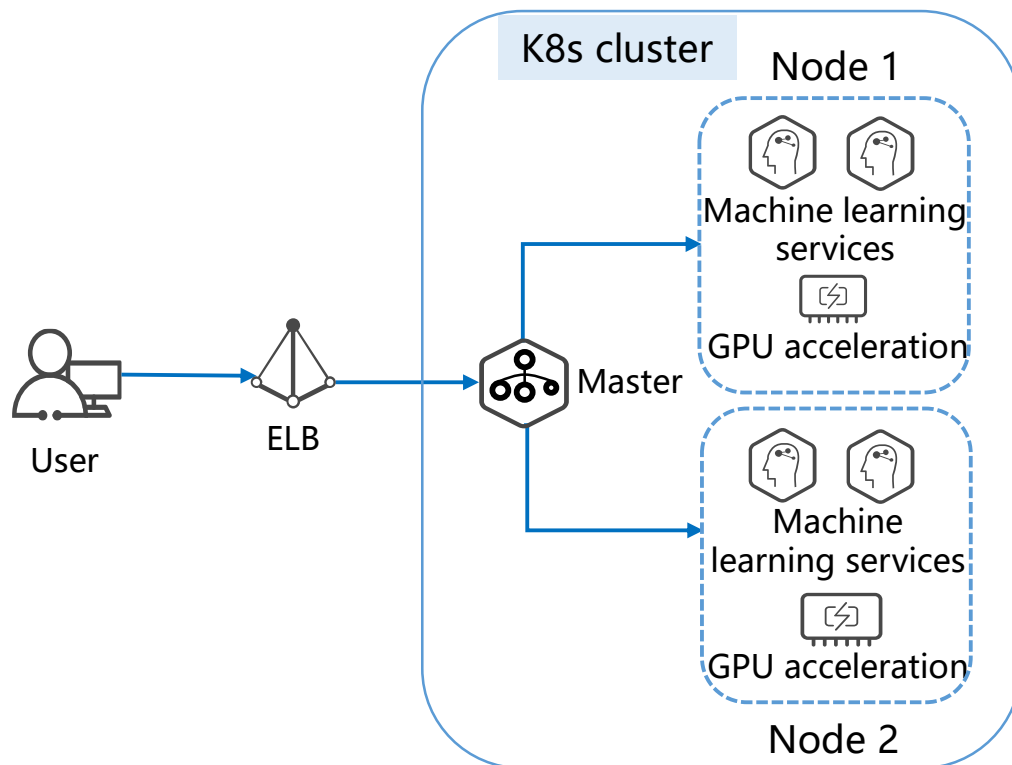
Function Description

Environment-independent containers allow you to seamlessly migrate applications and data between private and public clouds. You can achieve efficient resource usage and realize disaster recovery (DR).

Benefits

- **Lower costs:** Resource pools on HUAWEI CLOUD support rapid service scaling during peak hours, for only a fraction of the cost involved in building private clouds from scratch.
- **On-cloud DR:** Your services can be deployed both on-premises and in the cloud. The on-premises system provides services while the cloud system ensures DR.
- **Shared base:** The on-premises system shares the technical base with the cloud system. HUAWEI CLOUD resources are available on the on-premises system whenever required.

Scenario - AI Computing



Use Case

AI computing

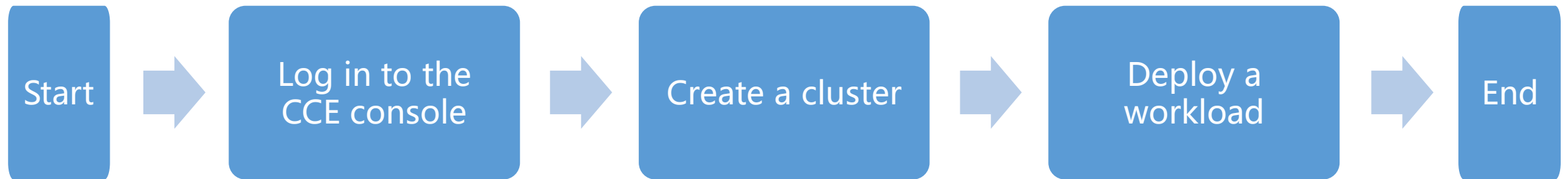
Benefits

- **Outstanding performance:** The bare-metal NUMA architecture and high-speed InfiniBand NICs drive a three- to five-fold improvement in AI computing performance.
- **Efficient computing:** GPUs are shared and scheduled among multiple containers, greatly reducing computing costs.
- **Proven success:** AI containers are compatible with all mainstream GPU models and have been used at scale in HUAWEI CLOUD's Enterprise Intelligence (EI) products.

CCE Concepts

Cluster	A cluster is a collection of cloud resources required for running containers, such as cloud servers and load balancers.
Pod	A pod consists of one or more related containers that share the same storage and network space.
Node	A node is a server (a VM or PM) on which containerized applications run.
Service	A Service is an abstraction which defines a logical set of pods and a policy by which to access them (sometimes this pattern is called a microservice).
Container	A container is a running instance of a Docker image. Multiple containers can run on the same node.
Image	An image is a binary that includes all of the requirements for running a container.

CCE Configuration Process



Creating a Cluster

- When creating a CCE cluster, set the billing mode, region, cluster version, management scale, and number of master nodes.

Billing Mode

Yearly/Monthly

Pay-per-use

?

Region

AP-Singapore

Regions are geographic areas that are isolated from each other. Resources are region-specific and cannot be used across regions through internal network connections.

* Cluster Name

cce-vivi

Enter 4 to 128 characters, starting with a letter and ending with a letter or digit. Only lowercase letters, digits, and hyphens (-) are allowed.

Version

v1.17.17

v1.19.10

For more cluster version features, [click here](#) to go to the help documentation.

Management Scale

50 nodes

200 nodes

1,000 nodes

2,000 nodes(Sold Out)

?

Number of master nodes

3

1

?

Disaster recovery level: Host computer Master node information: AZ1, [change](#)

Scaling a Cluster

- CCE automatically scales a 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.

Events | **Auto Scaling** | Kubectl | Resource Tags | Istioctl

Scale-out Settings | Scale-out Policies

Edit

Maximum Nodes: 10 | Cooldown Period (s): 900

Node Configuration

Billing Mode	Pay-per-use	AZ	Specifications
Node Quantity		System Disk	Data Disk
EIP		Subnet: subnet-e79e	Network Model: Tunnel network
Cloud Server Fee	¥0.00 /hour		

Currently automatic scale-in is not supported. Manual scale-in can be performed according to resource usage.

Upgrading a Cluster

- Currently, you can upgrade only CCE clusters containing VM nodes. CCE clusters consisting of BMS nodes or nodes created from private images, CCE Turbo clusters, and Kunpeng clusters cannot be upgraded.

The screenshot displays two CCE cluster details panels. The left panel shows a cluster with 2/2 nodes, 6 vCPUs, and 12 GB of memory. The right panel shows a cluster with 3/3 nodes, 6 vCPUs, and 12 GB of memory. In the right panel, the 'Version' field is highlighted with a red box, and a context menu is open over it, showing options like 'Auto Scaling', 'Add to Cluster', 'Upgrade', 'Delete', 'Hibernate', 'Wake', 'Download X.509 Certificate', and 'Enable Istio'. The 'Upgrade' option is highlighted in blue. A red 'Upgrade' banner is visible in the top right corner of the right cluster's panel.

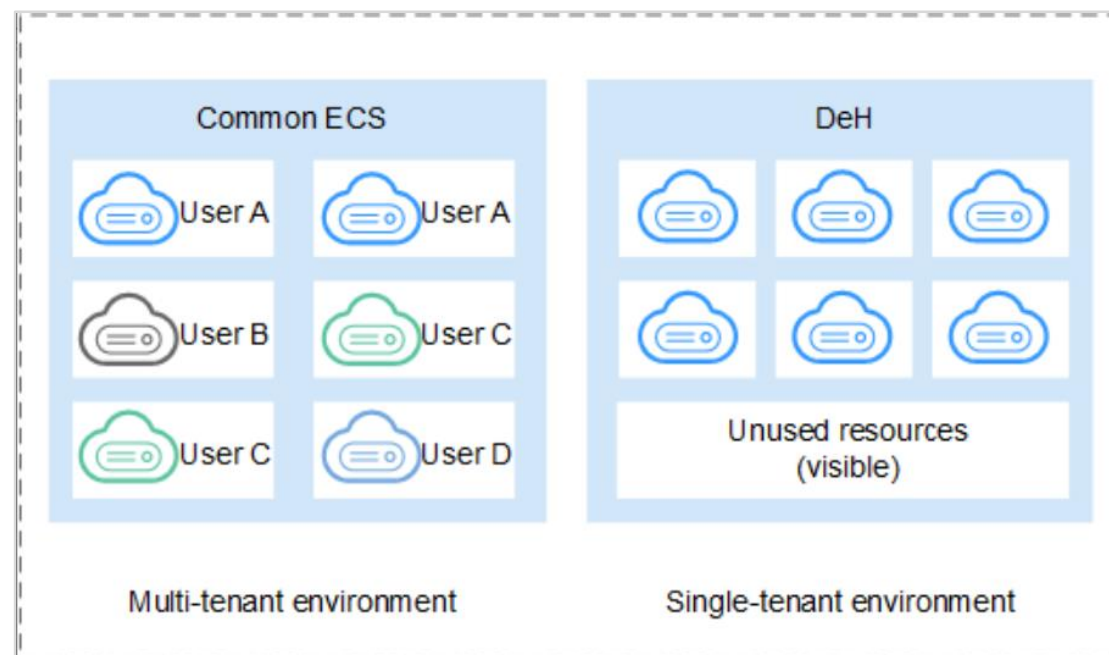
Cluster Name	Status	Version	Nodes (Available/Total)	Total vCPUs	Total Memory
W...	Available	v1.17.9-r0	2 / 2	6	12 GB
W...	Available	v1.11.7-r2	3 / 3	6	12 GB

Contents

1. Elastic Cloud Server (ECS)
2. Bare Metal Server (BMS)
3. Image Management Service
4. Auto Scaling (AS)
5. Cloud Container Engine (CCE)
- 6. Other Compute Services**

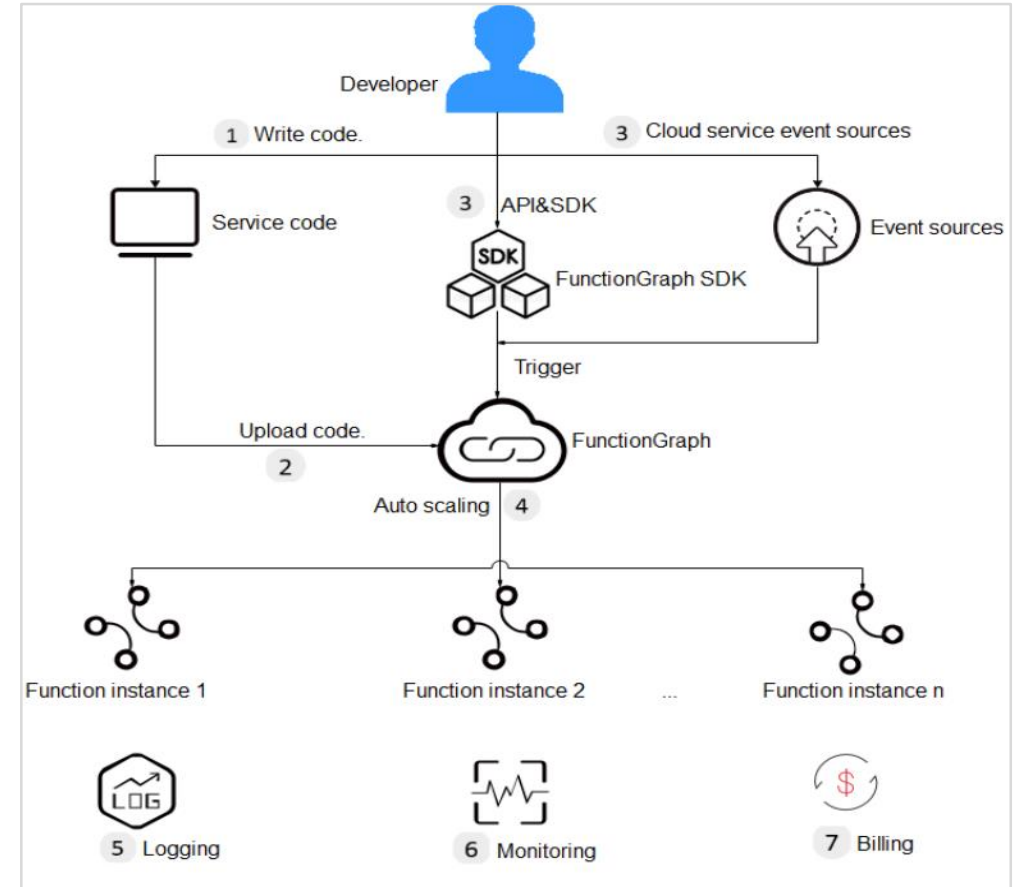
What Is DeH?

- A Dedicated Host (DeH) is a physical server fully dedicated for your own services. DeH allows you to ensure performance by keeping compute resources isolated. DeH also allows you to use your existing software licenses, so you can leverage existing investments to save money.



What Is FunctionGraph?

- FunctionGraph allows you to run your code without provisioning or managing servers, while ensuring high availability and scalability. All you need to do is upload your code and set execution conditions, and FunctionGraph will take care of the rest. You pay only for what you use and you are not charged when your code is not running.



Quiz

1. (True or False) There is a hypervisor layer in containerization, just like the traditional virtualization featuring VMs.
 - A. True
 - B. False
2. (True or False) The functions of an IMS image are the same as those of an ISO image.
 - A. True
 - B. False

Summary

- This chapter described compute cloud services. After completing this course, you will be able to understand each phase of technical transformation, from hardware, virtualization, cloud platform, and to cloud services. In this process, many new products, such as Elastic Cloud Server (ECS) and Cloud Container Engine (CCE) will be used. Both of these products can be used to deploy application systems, but the technical architectures are different. Therefore, to better help enterprises migrate their service systems to the cloud, you need to clearly understand the technical details of each cloud service.

Recommendations

- Huawei iLearning
 - <https://e.huawei.com/en/talent/>
- HUAWEI CLOUD Help Center
 - <https://support.huaweicloud.com/intl/en-us/help-novice.html>
- HUAWEI CLOUD Academy
 - <https://edu.huaweicloud.com/intl/en-us/>

Acronyms and Abbreviations

- AI: Artificial intelligence
- API: Application Programming Interface
- AS: Auto Scaling
- BMS: Bare Metal Server
- CCE: Cloud Container Engine
- CI/CD: Continuous Integration/Continuous Delivery
- CISC: Complex Instruction Set Computer
- CPH: Cloud Phone
- CPU: Central Processing Unit
- DeH: Dedicated Host

Acronyms and Abbreviations

- DevOps: Development and Operations
- DHCP: Dynamic Host Configuration Protocol
- ECS: Elastic Cloud Server
- EI: Enterprise Intelligence
- GPU: Graphics Processing Unit
- HPC: High Performance Computing
- HTTPS: Hypertext Transfer Protocol over Secure Sockets Layer
- IB: InfiniBand
- IMS: Image Management Service
- K8s: Kubernetes

Acronyms and Abbreviations

- IPoIB: Internet Protocol over Infiniband
- NUMA: Non-Uniform Memory Access
- RDMA: Remote Direct Memory Access
- RISC: Reduced Instruction Set Computer
- SR-IOV: Single Root Input/Output Virtualization
- VLAN: Virtual Local Area Network
- VPC: Virtual Private Cloud

Thank you.

把数字世界带入每个人、每个家庭、
每个组织,构建万物互联的智能世界.

Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

**Copyright©2021 Huawei Technologies Co., Ltd.
All Rights Reserved.**

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

