

Enterprise Application Architectures and Cloud Technologies



Foreword

- Traditional application architectures are becoming insufficient for the growing businesses of enterprises. They need architectures that are more secure, efficient, and cost-effective.
- This course describes why enterprises are going cloud, explores the principles of architecture design, and goes through cases of migrating traditional e-commerce platforms to the cloud.

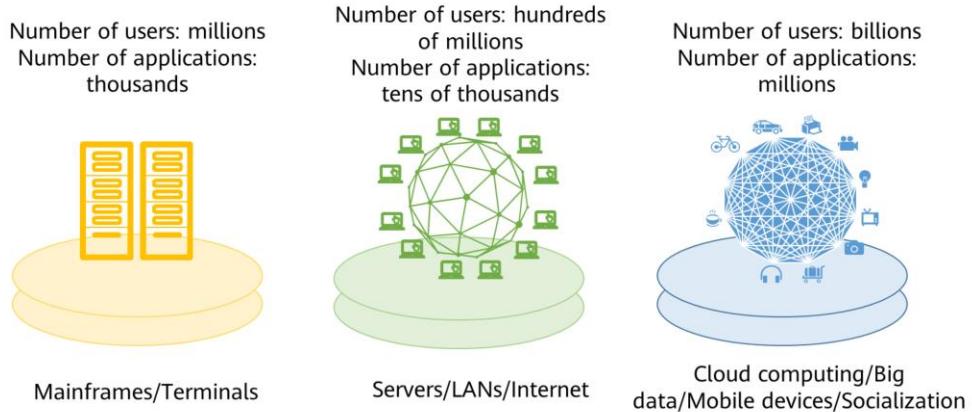
Objectives

- Upon completion of this course, you will understand:
 - IT architecture development trends and why enterprises need to go cloud
 - Basic principles of cloud-based architecture design

Contents

- 1. Background of Enterprise Cloud Migration**
2. Principles of Enterprise Application Architecture Design
3. Cloud-based Architecture Design Case Study

The Evolution of Computing Platforms

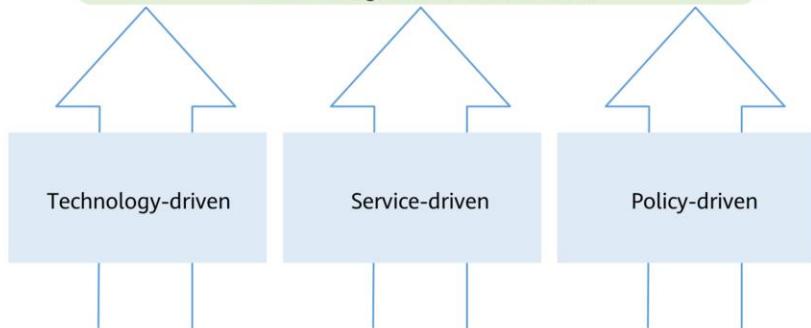


Source: International Data Corporation (IDC)

- The emergence of early computing technologies mainly consisted of mainframes and terminals.
- With the advent of personal computers (PCs) in the 1980s, the second platform emerged, which was characterized by the client/server system, Ethernet, RDBMS, and Web applications.
- Today we are using the third platform, which includes cloud computing, big data, mobile devices, and socialization technologies. At the core of these technologies is cloud computing. Customers use cloud providers' services to allocate IT resources. Big data turbocharges data analysis to achieve in-depth insights and for leaders' to make better-informed decisions. Mobile devices enable ubiquitous access to applications and information. Socialization technologies help connect people and ensure better collaboration and information exchanges.
- For more details, see https://en.wikipedia.org/wiki/Third_platform.

Enterprises Are Going Cloud

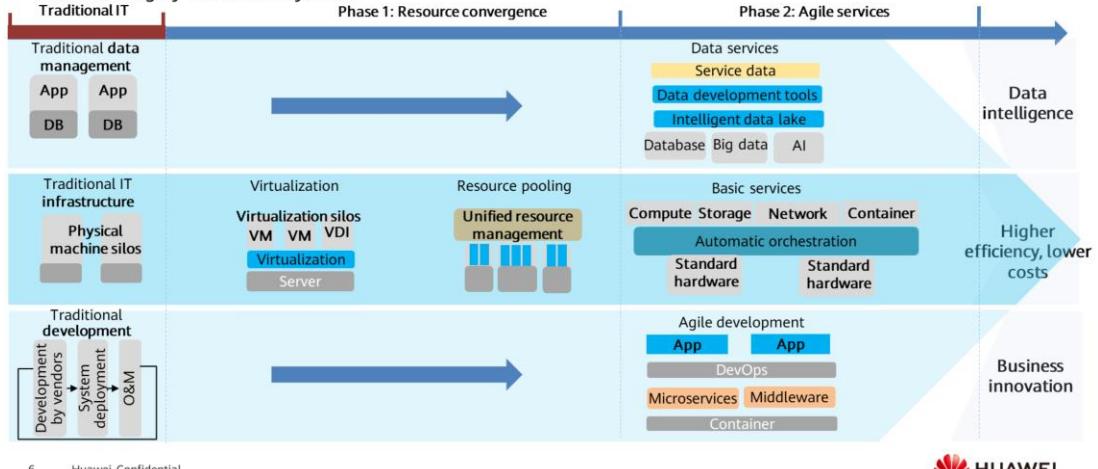
An open, flexible, easy-to-use, and secure cloud platform provides enterprises with a new choice in modernizing their IT architectures.



- The cloud platform provides enterprises with a new choice in adapting their IT architectures to receive growing volumes of data and business.

Technology-driven

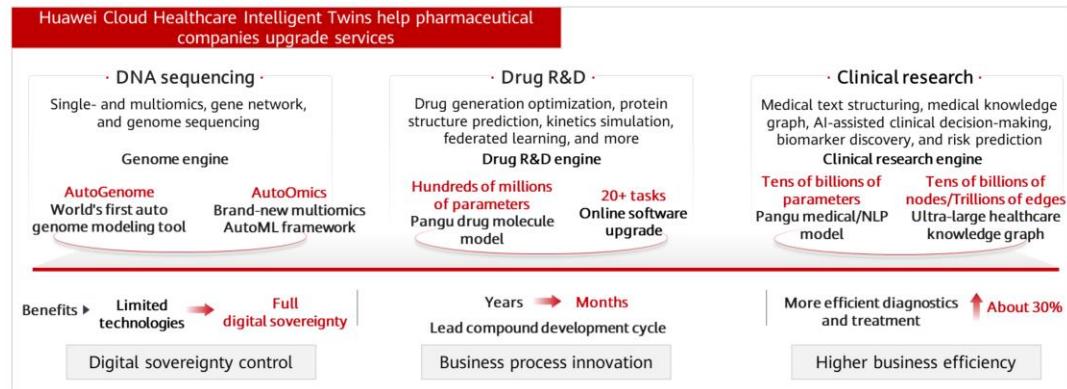
- Cloud migration is essential for enterprises to achieve agile service development, process massive amounts of data, and build highly resilient IT systems.



- Data management:
 - Traditional data management involves relational data and transactions but concurrent analysis throughput is low.
 - Modern data management adopts database and big data services to mine massive service data. Intelligent analysis contributes to better operations and better-informed decisions for new value.
- Infrastructure:
 - The resource utilization of physical machine-based deployment is low.
 - Virtualization deployment improves device utilization and simplifies O&M.
 - Resource pooling allows the management platform to integrate virtualization silos into resource pools for unified management and sharing.
 - Modern infrastructure management is automated and allows for self-service. Teams can collaborate and perform massive operations concurrently, and IT activities are fixed, standardized, and measurable. IT that used to support O&M is now oriented to operations.
- Development:
 - Traditional development methods use mature and reliable technologies. Services are stable and developed in advance, but the process is rigid.
 - A modern IT architecture is constructed in a distributed manner. That is, the architecture consists of microservices, and adopts DevOps and a development and test pipeline for quick roll-out and elastic scaling of new services.

Service-driven

- Cloud vendors provide a large number of PaaS and SaaS services and complete solutions to help enterprises explore new service requirements and stay relevant.



7

Huawei Confidential



- Huawei Cloud Healthcare Intelligent Twins breathe new life into the traditional healthcare industry and help improve efficiency.
- Agility and resource scheduling are embodied in the previous page. This page focuses on cloud service enablement. That is, the new capabilities that are offered in services.

Policy-driven

- Cloud computing is a key industry. Governments worldwide see the benefit of helping enterprises go cloud.

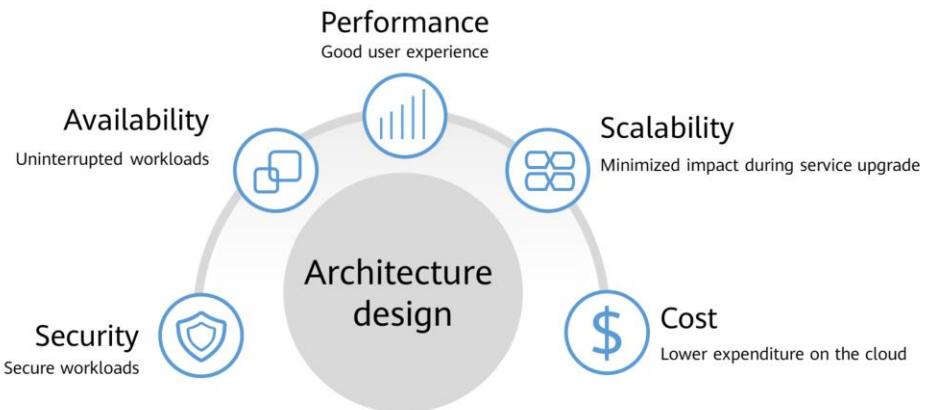


- The cloud computing software industry has been a national priority since the 12th Five-Year Plan.
- According to the 13th Five-Year Science and Technology Innovation Plan, cloud computing technologies and applications will be promoted to empower the next generation of ICT infrastructure.
- The 2019 Federal Cloud Computing Strategy — Cloud Smart — is a long-term, high-level strategy to drive cloud adoption in Federal agencies.
- Shaping Europe's Digital Future stresses the importance of cloud computing to digitization.
- In the Outline of the 14th Five-Year Plan for National Economic and Social Development and Long-Range Objectives Through the Year 2035, the development of StatChina was propelled to new heights and cloud computing has become key to that growth. Cloud computing software will embrace new opportunities.

Contents

1. Background of Enterprise Cloud Migration
- 2. Principles of Enterprise Application Architecture Design**
3. Cloud-based Architecture Design Case Study

Five Principles of Enterprise Application Architecture Design



- Building a cloud-based software system is very similar to building a house. If the foundation is not solid, structural problems may damage the integrity and functionality of the house. When designing a solution for migrating enterprise applications to the cloud, if you ignore security, reliability, scalability, performance, and cost optimization, it may be difficult to build a system that meets your expectations and requirements. Considering the following factors in the design will help you build a stable and efficient system:
- Security: System security is assessed to protect information, systems, and assets while unleashing business value.
- Availability: The system recovers from infrastructure or service faults and dynamically obtains resources to meet requirements and reduce service interruption. Single-AZ availability, cross-AZ DR, cross-AZ active-active, and remote DR deployment should be considered in the design.
- Performance: The system uses resources to meet performance requirements, including compute, network, storage, and data.
- Scalability: The system can be scaled out or scaled up according to the number of users or overall workload.
- Cost: Avoid or eliminate unnecessary costs or poor resources.

Significance of Security Design

- Protection against lingering security threats is essential for service continuity and data confidentiality. Compliance has become a basic aspect of enterprise business.



Lingering security threats

- In February 2019, GitHub was hit with a massive DDoS attack that peaked at 1.35 Tbps.
- In H1 2020, the number of DDoS attacks increased 2.5 times year-on-year.
- In June 2020, a CSP suffered the biggest attack of all time, peaking at 2.3 Tbps.
- Canalys estimated that the amount of data leaked in 2019 alone was greater than the amount in the previous 15 years combined. The number of ransomware attacks also increased by 60%.



Compliance: a basic business requirement

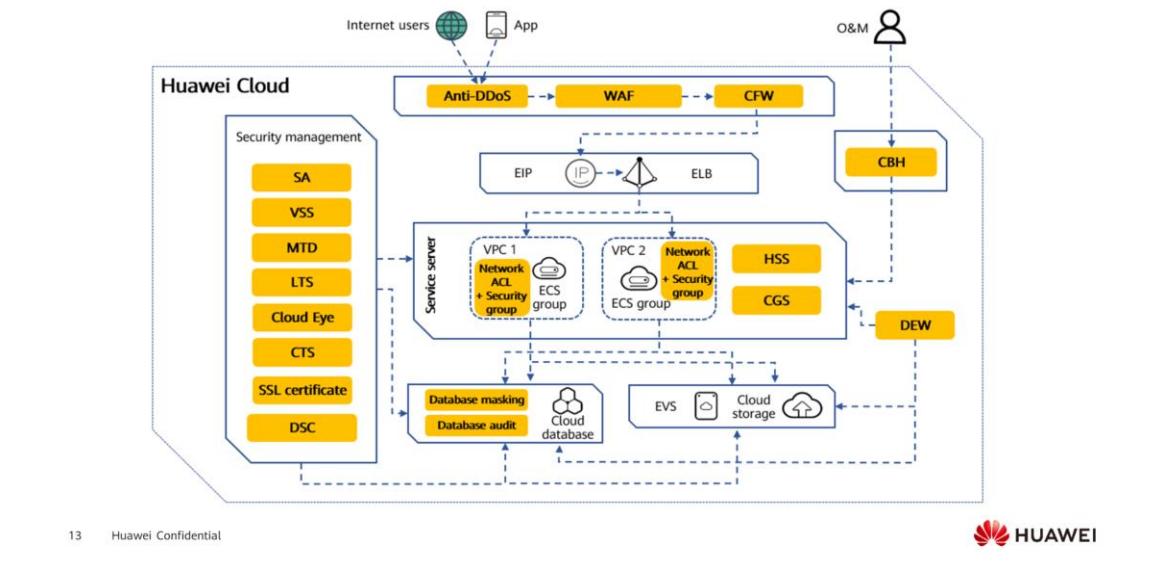
- EU: GDPR
- China: *Cybersecurity Law*, DJCP (MLPS) 2.0, and *Data Security Law*
- Countries are introducing laws to enhance data security and privacy protection.

Cloud-based Security Design (1)

- Use appropriate security services to enhance security at each layer.

Domain	Protective Measure	Service	
Workload security	Continuously monitor and eliminate threats to ensure cloud workload security.	Host Security Service (HSS) Cloud Bastion Host (CBH)	Container Guard Service (CGS)
Network security	Configure security services at the network layer to isolate cloud resources and protect network borders.	Cloud Firewall (CFW)	Advanced Anti-DDoS (AAD)
Application security	Configure security services at the application layer to block attacks.	Web Application Firewall (WAF)	Application Trust Center (ATC)
Data security	Manage data assets throughout their lifecycles to ensure that the entire data usage process is secure, visible, controllable, and traceable.	Data Security Center (DSC) Database Security Service (DBSS)	Data Encryption Workshop (DEW) Cloud Certificate Management (CCM)
Security management	Manage the cloud environment to minimize risks.	Identity and Access Management (IAM) Situation Awareness (SA)	Managed Threat Detection (MTD)

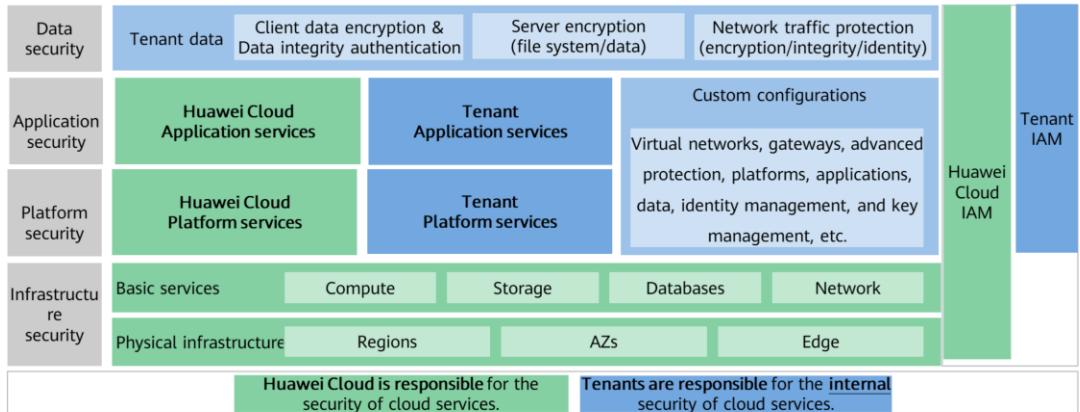
Cloud-based Security Design (2)



- Secure communication network
 - Anti-DDoS is used to defend against DDoS attacks.
 - Web Application Firewall (WAF) is used to defend against web attacks.
 - SSL certificates are used for communication encryption.
- Security zone border
 - The cloud firewall is deployed between Internet borders and VPCs.
- Secure compute environment
 - Host Security Service (HSS) and Container Guard Service (CGS) are deployed.
 - Network ACLs and security groups are used for access control within a VPC.
 - Data Security Center (DSC) manages data security throughout the data lifecycle.
 - Data encryption is enabled for storage by default.
 - Database Security Service (DBSS) is deployed for key databases.
- Security Management Center
 - The Situational Awareness (SA) service is used to ensure cloud resource security.
 - Cloud resources are periodically scanned to detect vulnerabilities.
 - Log Tank Service (LTS), Cloud Trace Service (CTS), and Cloud Eye are used to manage cloud resources.
 - Cloud Bastion Host (CBH) is used for O&M.

Cloud-based Security Design (3)

- According to the shared responsibility model, users need to effectively manage the internal security of cloud services and security of custom configurations.



- Tenants deploy and configure security service products, including security configurations and management tasks (such as updates and security patches) of cloud services, such as virtual networks, virtual hosts, and guest VMs in tenant space, as well as container security management. Tenants are also responsible for the internal security configurations of other cloud services they lease.
- Tenants are also responsible for the security management of any application software or utility they deploy on Huawei Cloud. Before deploying security workloads in the production environment, tenants should test these workloads to prevent adverse effects on their applications and services.
- Tenants own and control their data regardless of the Huawei Cloud service they use. Tenants take measures to guarantee data confidentiality, integrity, and availability, as well as the identity authentication and authorization for data access. For example, tenants using IAM and DEW need to configure rules to properly keep their own service login accounts, passwords, and keys.

Significance of Availability Design

Availability: The degree to which a product is available when it needs to and can execute a task at any time. This probability measure is called availability.

Reliability: The ability of a product to perform a specified function under specified conditions and within a specified period of time.

Maintainability: The ability of a product to be maintained and restored to a previous state under certain conditions and within a specified time period in accordance with specified procedures and methods.

In cloud computing, availability is committed through the service level agreement (SLA).

SLA = 1 - [Downtime/(Downtime + Uptime)]

The following table lists the downtime acceptable for different SLA commitments.

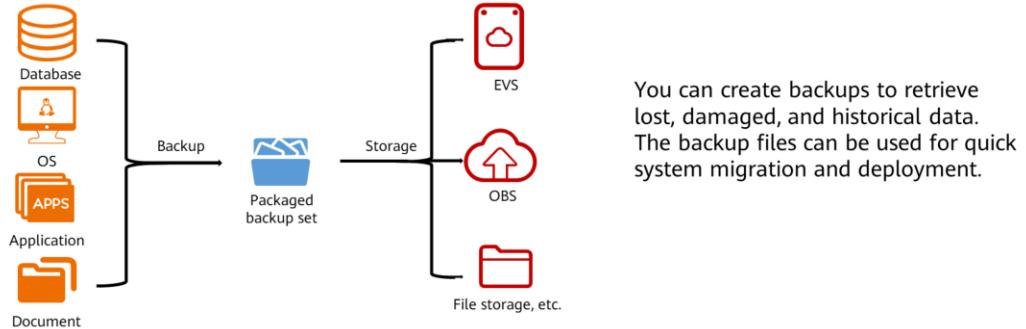
SLA	Weekly Downtime	Monthly Downtime	Annual Downtime
99%	1.68 hours	7.2 hours	3.65 days
99.90%	10.1 minutes	43.2 minutes	8.76 hours
99.95%	5 minutes	21.6 minutes	4.38 hours
99.99%	1.01 minutes	4.32 minutes	52.56 minutes
99.999%	6 seconds	25.9 seconds	5.26 minutes

Higher availability means a lower possibility of system breakdown and a better user experience. However, this is possible only when the system is improved, which will increase construction costs. Availability is closely related to service requirements. The availability requirement of a service depends on its importance.

- The longest annual downtime allowed for each SLA level is calculated as follows (365 days in a year):
 - 1 year = 365 days = 8760 hours
 - $99.9 = 8760 \times 0.1\% = 8760 \times 0.001 = 8.76 \text{ hours}$
 - $99.99 = 8760 \times 0.0001 = 0.876 \text{ hours} = 0.876 \times 60 = 52.6 \text{ minutes}$
 - $99.999 = 8760 \times 0.00001 = 0.0876 \text{ hours} = 0.0876 \times 60 = 5.26 \text{ minutes}$
- An annual downtime of 5.26 minutes means 99.999% SLA. A better SLA means higher requirements on the system. As a result, we need to consider whether the system is capable of meeting the increasing SLA requirements.

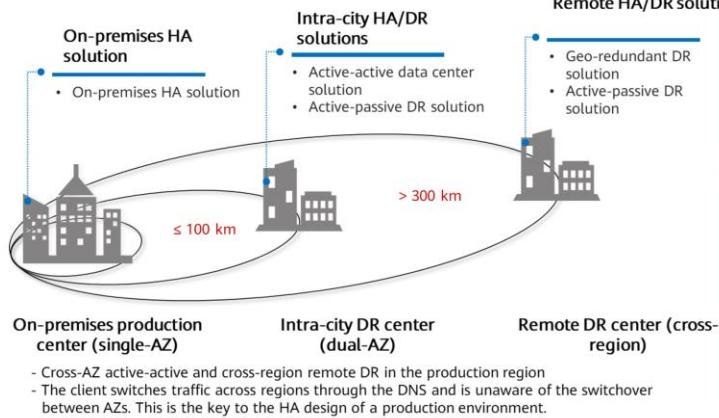
Cloud-based Availability Design (1)

- Backup design is the basis of data loss prevention.



Cloud-based Availability Design (2)

- Select single-AZ, multi-AZ, or cross-region deployment as required during HA/DR design.



HA refers to an on-premises or intra-city HA system.

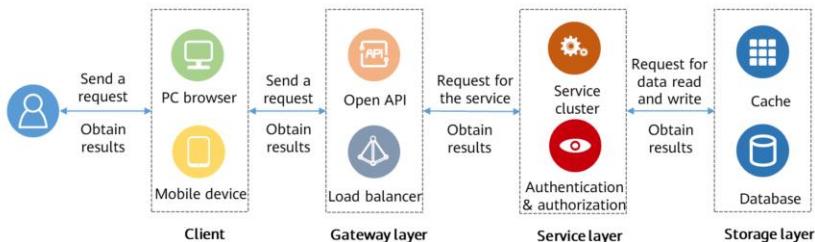
- When one or more applications are running on servers, ensure that the applications are not interrupted if any server is faulty and that the applications and system can be quickly switched to other servers, that is, on-premises system clusters and hot backup.

DR refers to the redundancy site built by users in addition to the production site. When the production site is damaged by a disaster, the redundancy site can take over services from the production site to ensure service continuity.

- Users can build more than one redundancy site for higher availability. Generally, a DR system is deployed in another city (more than 800 km away from the production site). When a disaster occurs, the DR system can recover data, applications, and services.

- The common cloud system HA design solutions are as follows:
 - The on-premises HA solution applies to on-premises production centers and single-AZ scenarios.
 - The intra-city HA/DR solutions, including an active-active data center solution and an active-passive DR solution, apply to the HA design of intra-city DR centers and dual-AZ scenarios.
 - The remote HA/DR solutions, including a geo-redundant DR solution and an active-passive DR solution, apply to remote DR centers and cross-region HA.

Significance of Performance Design



When a user interacts with the system, the user needs to access the system layer by layer. The information flow may suffer from latency during transmission. When many users access a website at the same time, user experience may be affected. User experience is one of the performance indicators. Performance improvements create the following benefits:

Prevention of performance bottlenecks

Better user experience

Appropriate resource allocation

- Prevention of performance bottlenecks

- Performance issues are detected and resolved in advance, such as high server CPU/MEM usage, program memory leakage, network congestion of application access links, insufficient database connection pools, application process suspension, and low cache hit ratio.

- Better user experience

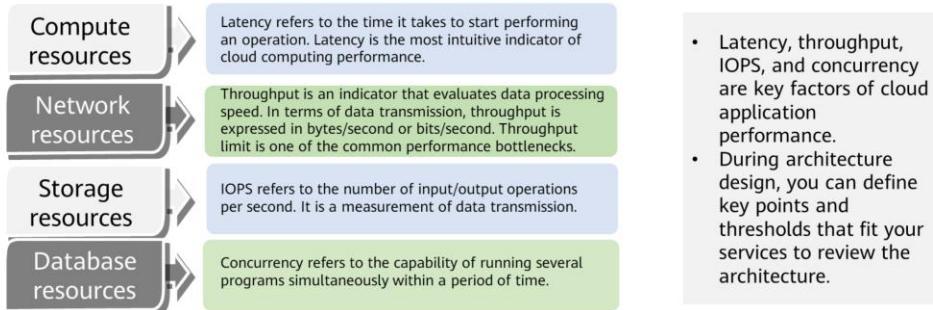
- User experience is improved by preventing the following problems: web page opening failures or slow response, video frame freezing, artifacts, delayed market data update, and disconnection and frame freezing during gaming.

- Appropriate resource allocation

- Cloud service specifications are appropriately allocated based on performance indicators. Nodes are added to or removed from service clusters.

Cloud-based Performance Design (1)

- Design performance measurement and review service status based on performance indicators to optimize cloud-based architecture.



- The performance of cloud applications is affected by many factors, including data transmission and software and hardware. These factors make performance evaluation complex.
- Cloud application performance can be affected by latency, throughput, IOPS, and concurrency, as well as compute, network, storage, and database resources.
- Compute resources: Large-scale infrastructure is shared, resulting in resource competition. Therefore, the appropriate distribution of limited resources is required to deal with load changes.
 - Compute resources affect the latency of applications.
- Network resources: The public cloud infrastructure is not located in the enterprise data center. As a result, the WAN must be used, which causes high bandwidth and latency. Multi-peer networks, encrypted offloading, and compression are factors that must be considered for architecture design.
 - Network resources affect the throughput of applications.
- Storage resources: read and write performance of storage products with different performance characteristics; unmeasurable disk I/O of elastic block storage
 - Storage resources affect the data transmission of applications.
- Database resources: If an application uses a database, the database resources affect application concurrency.
- The performance of cloud infrastructure can be unpredictable. Load changes may affect available CPU, network, and disk I/O resources. As a result, the performance of applications that work at the same time is unpredictable.

Cloud-based Performance Design (2)

- Select product types and specifications based on service scenarios for a flexible solution, and design a high-performance architecture to inexpensively meet performance requirements.

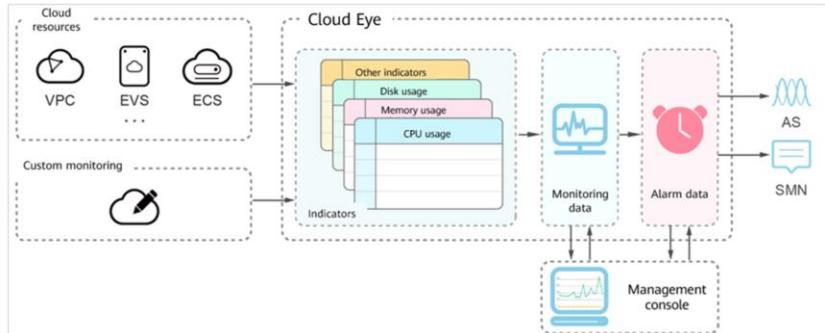
The following uses compute, storage, and network resources as an example:

Category	Sub-Category	Description	Cloud Service
Compute	Server	Virtual server instances have different series and sizes. They provide a variety of functions, including solid-state drives (SSDs) and graphics processing units (GPUs). When an ECS instance starts, the specified instance type determines the hardware of the instance host.	ECS, GPU acceleration, and FPGA acceleration
	Container	You can use Auto Scaling (AS) to define metric-based auto scaling for services so that more containers can be added as the service requirements grow.	CCE, AS, and ELB
Storage	Block	The low latency of data processing makes this storage mode ideal for high-performance computing.	EVS
	File	Low latency and high bandwidth are supported.	SFS
	Object	Latency is reduced and throughput is increased to ensure low-latency data access across geographic regions and smooth high-concurrency access.	OBS and CDN
Network	Enhanced networks	Enhanced networks provide higher I/O performance and lower CPU usage than legacy virtualized network interfaces. Enhanced networks also feature higher bandwidth and packets per second (PPS) and continuous reduction of inter-instance latency.	VPC
	Network functions	Network functions are used to reduce network distance or jitter.	DNS/CDN/ELB/VPN

- The overall design of the architecture system is also important. For example, to avoid remote data transmission, you can deploy resources near service sites, and adopt services such as CDN to reduce access latency.

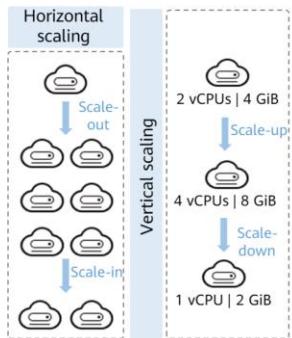
Cloud-based Performance Design (3)

- Enable monitoring in the cloud to remain informed so that you can adjust architecture loads in a timely manner.



Significance of Scalability Design

- Scalability is an attribute of a system or an application. Higher scalability means better processing of a massive number of requests. A system must be scalable to meet the increasing requirements of networks, task processing, database access, or file system resources.

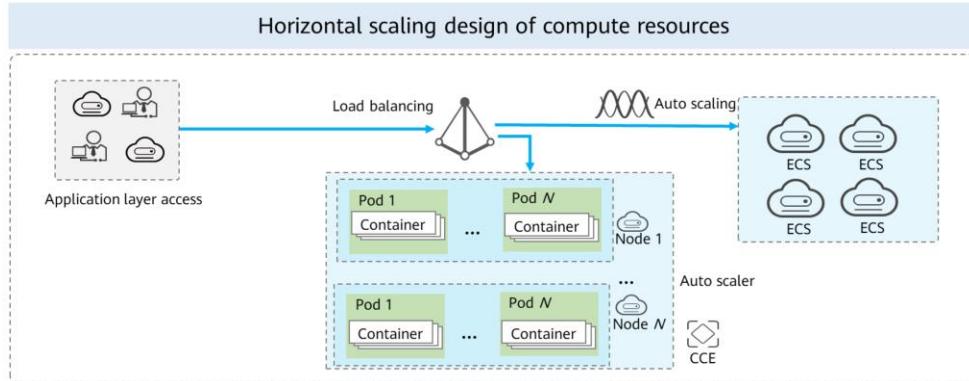


- The growing business of a company needs to be accompanied with system scaling. Continuous scalability improvement guarantees powerful software.
- When a system is highly scalable, its processing capability can be linearly increased with minimal changes or even just with the addition of hardware devices to achieve high throughput and performance at a low latency.

- Scalability is a design indicator that represents the computing and processing capabilities of a software system. High scalability indicates that the system can continue to run properly as the system expands and grows. The processing capabilities of the entire system can be linearly increased with only minimal modifications or hardware changes. In this way, high throughput and low latency can be achieved.
 - Horizontal scaling is a feature that allows the connection of multiple software and hardware products. In this way, multiple servers can be logically considered an entity. When a system is scaled out by adding new nodes with the same functions, the system can redistribute resources according to the loads of all nodes. The system is scaled out by adding more servers to the load balancing network so that incoming requests can be distributed among all of these networks.
 - Vertical scaling is to replace existing IT resources with new ones regardless of their capacity. That is, the CPU performance of the server is increased or reduced in place. You can add processors, main memory, storage devices, or network interfaces to nodes to handle the increasing requests of each system. The system is scaled up by adding more processors or main memory to host more virtual servers.
- Scalability of cloud computing allows users to use more resources as the load increases, and lets developers build scalable architectures. For example, microservices and containerized architectures encourage independent scaling.
- Latency and throughput are a pair of indicators for scalability. An ideal system architecture should deliver low latency and high throughput. Latency is the system response time that users can perceive. Shorter response time indicates lower latency. Throughput indicates the number of users who can perceive the low latency at the same time.

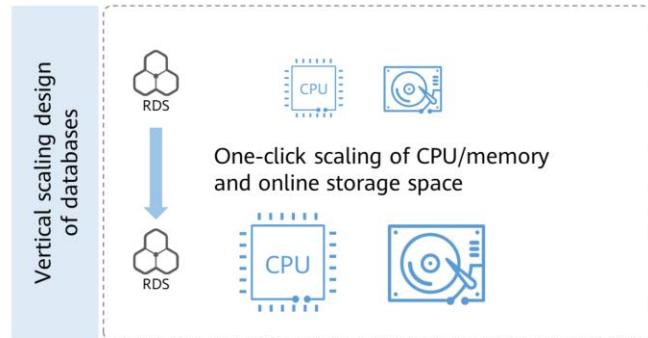
Cloud-based Scalability Design (1)

- Select scalable instances and use methods such as auto scaling to implement horizontal scaling.



Cloud-based Scalability Design (2)

- Select scalable instances to flexibly change instance specifications and adjust resources on demand.



Significance of Cost Design

- Budgets have a direct impact on the feasibility of cloud architecture, so enterprises need to control costs to ensure higher profits. A better cloud-based cost design contributes to a more flexible and controllable IT architecture.

Investment Item	Built by the Customer	Leased from Carriers	Cloud Service
Installation of servers, storage devices, network, and security implementation hardware	✓	✓	✗
Maintenance of servers, storage devices, network, and security implementation hardware	✓	✓	✗

No installation service or hardware maintenance fee is required after services are migrated to the cloud.

and

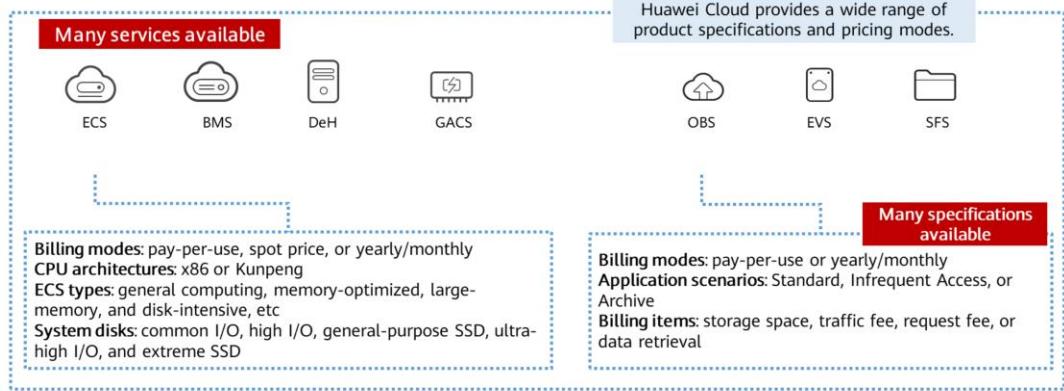


Cloud services can be purchased as needed.
Refined design helps control architecture costs.



Cloud-based Cost Design (1)

- Cloud resources are more flexible, so ensure that all cloud resources are in use. During planning, you can flexibly select product types and pricing modes to reduce infrastructure costs.

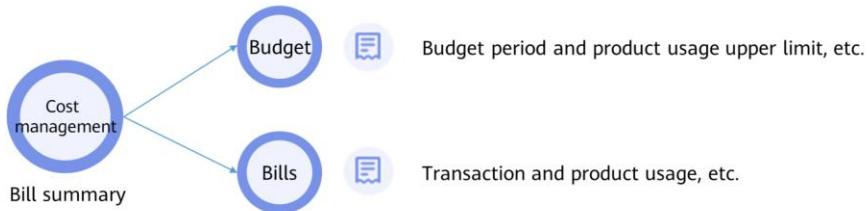


- Pay-per-use is preferred when service requirements fluctuate or flexible scale-out is needed. It is an ideal billing mode for development and test environments.
- Yearly/Monthly is a better option when resource requirements are stable and resources are used for a long period of time.

Cloud-based Cost Design (2)

- You must foster an awareness of expenditure and manage revenue and expenditure for cloud resource management.

The budget management and bill management functions provided by cloud service vendors allow for visualized fee management to optimize cloud costs.

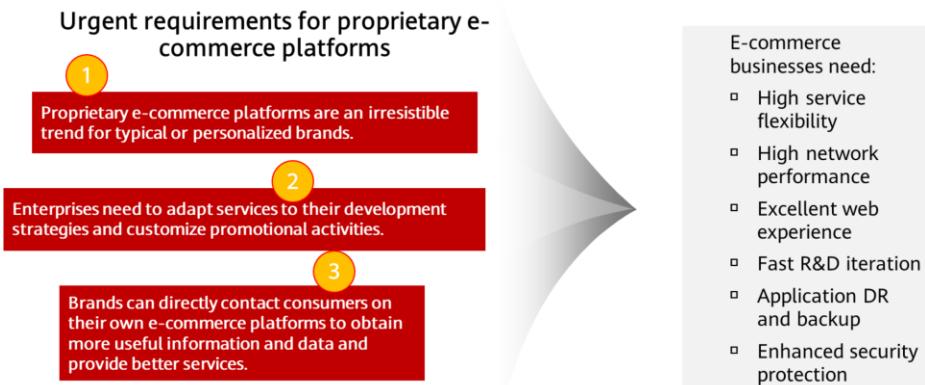


- Huawei Cloud provides budget and bill functions and visualized fee management to help customers optimize costs.
- A transaction bill includes the billing information of each order and of each billing cycle (a cloud service billing cycle can be hourly, daily, or monthly).

Contents

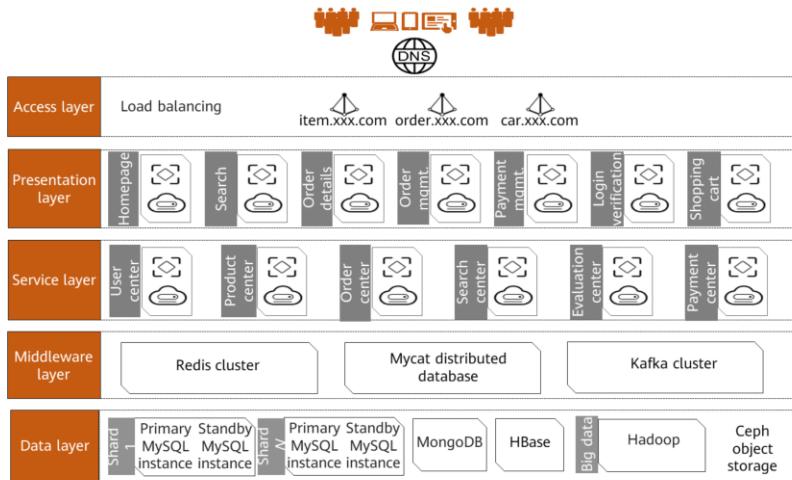
1. Background of Enterprise Cloud Migration
2. Principles of Enterprise Application Architecture Design
3. **Cloud-based Architecture Design Case Study**

Case Study: Migrating E-Commerce Platform Applications to the Cloud



- High service flexibility – scalability
- High network performance – performance
- Excellent web experience – performance
- Fast R&D iteration – scalability
- Application DR and backup – availability
- Enhanced security protection – security
- Cost is an important factor to consider when selecting a solution and plays a decisive role in the profitability of an enterprise.

Traditional Self-built E-Commerce Architecture



Common Challenges to the Traditional E-Commerce Architecture



Insufficient for service surges

E-commerce promotions such as flash sales can suffer from instant spikes in the number of requests (increasing tens or even hundreds of times), resulting in high server load, slow system response, and even system crash.



Risk-prone operations

The entire procedure, ranging from off-site traffic attraction, registration and login, browsing and comparison, and special offer claim to order placement, payment, delivery, and evaluation, are risk-prone, such as credential stuffing and brute force attacks, illegal resale, web page tampering, DDoS attacks, account leakage, and Trojan horse implantation.



Poor user experience

E-commerce services involve a large amount of static data. If static data is stored on servers in traditional mode, the loading is slow and expensive. When users on different networks access the same e-commerce website, problems such as slow web page response and network latency may occur.



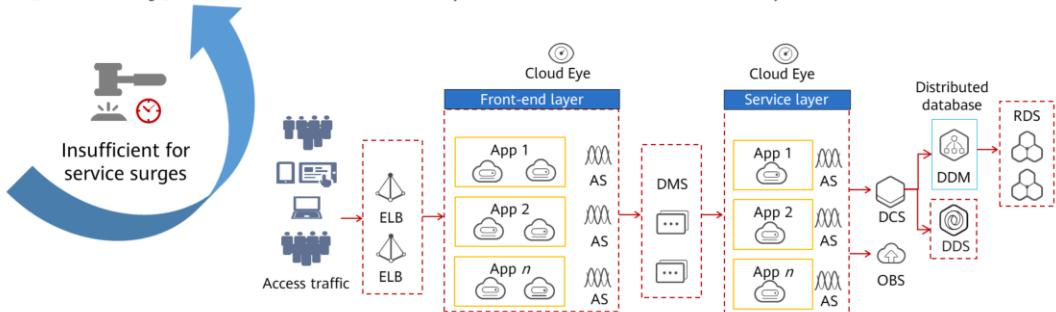
Expensive O&M

Traditional e-commerce architecture requires labor-intensive O&M of the infrastructure and cloud platform. Once completed, the architecture cannot be flexibly adjusted, which may cause high costs and resource waste.

- These are the key challenges that need to be addressed by e-commerce platforms built on on-premises infrastructures. A good cloud-based architecture design can solve these problems.

Cloud Solutions Help Handle Service Surges at Ease

- [Scalability] Scalable cloud resources, responsive to service requirements
- [Performance] Performance monitoring
- [Availability] An HA architecture that prevents server breakdown upon excessive loads



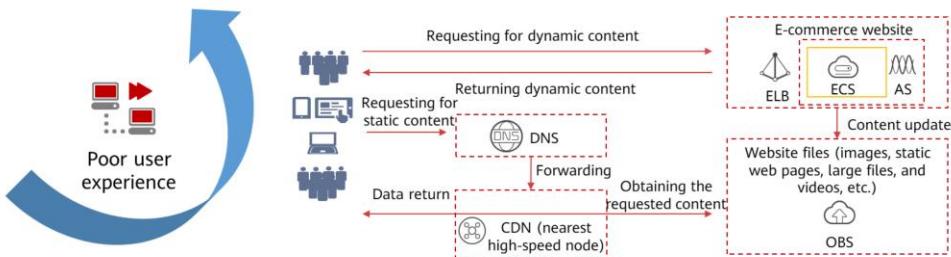
32 Huawei Confidential



- [Scalability] Auto Scaling (AS)
- [Performance] Cloud Eye
- [Availability] ELB provides multiple back-end ECS instances to prevent single points of failure (SPOFs).

Cloud Solutions Improve User Experience

- [Performance] Select high-performance cloud products (such as high network bandwidth and high-configuration ECSs). Flexibly select performance solutions and design static data cache to optimize user access.



- Dynamic content refers to the content obtained through asp, jsp, php, perl, and cgi requests, APIs, and dynamic interaction requests (such as post, put, and patch requests).
- Static content refers to the same content obtained through different access requests, such as images, videos, and file packages on websites. CDN can provide acceleration services for static content under acceleration domain names.
- CDN cannot cache dynamic content. As a result, dynamic content cannot be accelerated during the acceleration of websites, file download, and on-demand services. Static and dynamic content can be accelerated through whole site acceleration.

Cloud Solutions Ensure Secure Operations

- [Security] Follow Huawei Cloud security best practices to design the architecture, select cloud security protection services, maintain security monitoring, and give risk alarms in a timely manner.



Scenario	Security Protection Service	Protective Measure
DDoS attacks	AAD	Provides network traffic cleansing to defend against DDoS attacks at the network and application layers.
SQL injection, XSS cross-site scripting attack, web page Trojan horse upload, third-party application vulnerability attack, CC attack, and malicious crawler scanning, etc.	WAF	Filters out HTTP and HTTPS application attack traffic.
Host Trojan horses and mining intrusions	HSS	Manages e-commerce cloud host vulnerabilities and performs intrusion detection and baseline check.
Security analysis, resource change, compliance audit, and fault locating	CTS	Records operations on resources in a cloud account.
Malicious and unauthorized operations	MTD	Integrates detection models such as the AI engine, threat blacklist and whitelist, and rule baseline to identify potential threats in cloud service logs and provide analysis results.
More	More	More

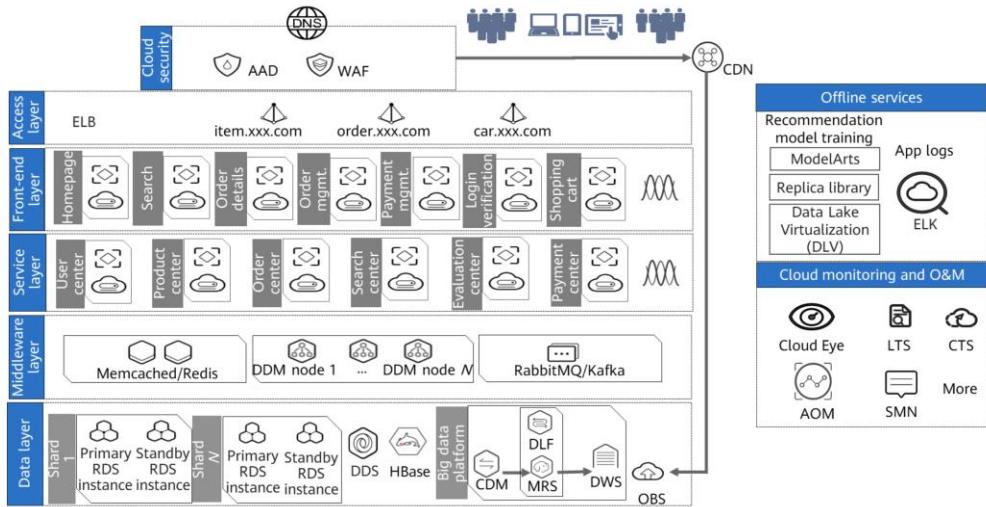
- High-security methods and tools are adopted.

Cloud Solutions for Less Expensive O&M

- [Cost] A cloud-based architecture allows for automated and less expensive O&M. In addition to cloud migration, you can flexibly configure products, standardize the usage of cloud resources, and clear non-standard or idle resources in a timely manner to reduce costs.



Cloud E-commerce Application Architecture



E-Commerce Application Architecture Comparison: Self-purchased Servers vs. Huawei Cloud Architecture

Architecture Design Consideration	Self-purchased Servers	Huawei Cloud Architecture
Security	Access control of multiple users on multiple servers is hardly supported.	Easy access control of multiple users on multiple servers
	Devices for traffic cleansing and black hole filtering are expensive.	Free anti-DDoS service, traffic cleansing, black hole filtering, Trojan horse scanning, and protection against brute force hacking attempts
	Vulnerabilities, Trojan horses and port intrusion are common.	Additional services such as scanning of vulnerabilities, Trojan horses, and port intrusion
	Security response is prone to delay due to a limited number of security experts.	A large number of cloud security experts for responsive security services
Availability	Insufficient hardware reliability results in frequent device problems.	Native active-active Huawei Cloud products for 99.95% of instance availability and at least 99.9999999% of data reliability
	Manual data backup and restoration are time- and labor-consuming.	Easy data recovery with automatic failover and snapshot backup
	Equipment rooms are mostly single-line and dual-line.	Border Gateway Protocol (BGP) multi-line equipment rooms that support nation-wide stable access
Performance	Resource configuration management is not real-time, which may cause insufficient or excessive performance.	Elastic resource scaling
	There are no online management tools for performance monitoring, making maintenance difficult.	Web-based online management and automatic monitoring of performance indicators, which is simple and convenient
Scalability	Fixed configurations cannot meet different service requirements.	Flexible service provisioning and online upgrade
	Modifying configurations requires lengthy hardware upgrade, and the duration of service interruption is uncontrollable.	Prevention of configuration data loss during upgrade, and controllable duration of service interruption
Cost	O&M is expensive and labor-consuming.	Low usage costs and no maintenance costs
	Pay-per-use billing is not supported and the quantity of resources to be purchased must be based on the largest service spike.	Shift from CAPEX to OPEX, flexible pay-per-use to deal with fluctuating service volumes

- OPEX indicates the operating expenses (OPEX) of an enterprise. The calculation is performed as follows: $OPEX = \text{Maintenance expense} + \text{Marketing expense} + \text{Labor cost} (+ \text{Depreciation})$. OPEX mainly refers to the cash cost of the current period.
- CAPEX indicates the capital expenditure, such as fund and fixed assets. For example, the once-off expenditure on network equipment, computers, and instruments is CAPEX, among which network equipment accounts for the largest proportion.

Quiz

1. (True or False) Migrating workloads from traditional architectures to the cloud means lower costs.

True

False

2. (Single-answer question) Which of the following is not one of the five principles of architecture design?

- A. Testability
- B. Security
- C. Performance
- D. Cost

- Answer: False. After migrating workloads to the cloud, organizations need to adopt cost design. If cloud resources are used without restrictions, the cost will far exceed that of the off-cloud architecture.
- Answer: A. The five principles of architecture design are security, performance, cost, availability, and scalability.

Summary

- This course has described the background of enterprise cloud migration and the principles of enterprise application architecture design, and uses a case study to explain the advantages of cloud-based architectures. You have now learned all about cloud-based architectures. In the courses that follow, we will dive into Huawei Cloud services.

Recommendations

- Huawei iLearning
 - <https://e.huawei.com/en/talent/portal/#/>
- Huawei Technical Support
 - <https://support.huaweicloud.com/intl/en-us/help-novicedocument.html>
- HUAWEI CLOUD Developer Institute
 - <https://edu.huaweicloud.com/intl/en-us/>

Acronyms and Abbreviations

- AAD: Advanced Anti-DDoS
- APIG: API Gateway
- AS: Auto Scaling
- ATC: Application Trust Center
- AZ: Availability Zone
- BMS: Bare Metal Server
- CBH: Cloud Bastion Host
- CCE: Cloud Container Engine
- CDN: Content Delivery Network
- CFW: Cloud Firewall
- CGS: Container Guard Service
- DDM: Distributed Database Middleware
- DeH: Dedicated Host
- DEW: Data Encryption Workshop
- DNS: Domain Name Service
- DSC: Data Security Center
- ECS: Elastic Cloud Server
- EVS: Elastic Volume Service

Acronyms and Abbreviations

- EIP: Elastic IP
- ELB: Elastic Load Balance
- FACS: FPGA Accelerated Cloud Server
- FPGA: Field Programmable Gate Array
- GACS: GPU Accelerated Cloud Server
- GPU: Graphics Processing Unit
- HSS: Host Security Service
- IAM: Identity and Access Management
- IOPS: Input/Output Operations per Second
- IT: Information Technology
- MDR: Managed Detection Response
- MTD: Managed Threat Detection

Acronyms and Abbreviations

- NAT: Network Address Translation
- OBS: Object Storage Service
- PaaS: Platform as a Service
- RDS: Relational Database Service
- SA: Situation Awareness
- SaaS: Software as a Service
- SFS: Scalable File Service
- SOC: Security Operations Center
- SMN: Simple Message Notification
- VPC: Virtual Private Cloud
- VPN: Virtual Private Network
- VSS: Vulnerability Scan Service
- WAF: Web Application Firewall

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©2022 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.



Compute Service Planning



Foreword

- Compute resources are critical for enterprise development. Services cannot run properly without compute resources. In the cloud computing era, computing services are the most important cloud services.
- In this course, we will learn about the compute services on Huawei Cloud.

Objectives

- Upon completion of this course, you will:
 - Understand common compute services on Huawei Cloud.
 - Understand how to select compute resources based on service scenarios and how compute services work with each other.

Contents

- 1. Compute Service Overview**
2. Compute Service Planning
3. IMS Planning
4. AS Planning

The Evolution of Computing

- With the innovation, convergence, spread, and upgrade of computing technologies, the way we live and work has undergone profound changes. Computing power has been playing an increasingly prominent role in driving economic and social development. There is a general consensus that more compute means more productivity.



- In the early stages, mainframes and midrange computers provided compute, storage, and network resources. We call this era the "exclusive computing" era. Under the leadership of well-known companies such as Intel, x86 chips emerged and were used commercially. A large number of data centers emerged as well, and the industry started shifting from exclusive computing to general computing, the age of computing 2.0. As the development of network and digital technologies accelerated, computing was no longer limited to data centers or x86 processors. Computing services and technologies started diversifying to meet full-stack, all-scenario service requirements. We call this era the "intelligent computing" era.
- Full-stack, full-scenario: a variety of development frameworks and languages.

The Evolution of Cloud Computing Models

- As science and technology have evolved, we have seen four main stages in the evolution of compute.



Huawei Cloud Compute Services



- Cloud computing has the following characteristics:
 - On-demand self-service: You can purchase software, servers, and other services by yourself through web portals.
 - Resource pooling: Thanks to the virtualization technologies, you can share systems and services in cloud data centers. Regions are physically isolated from each other.
 - Extreme elasticity: Compute resources can be flexibly scaled as service demand changes. For example, you can purchase more powerful servers to handle increased workloads without having to install new IT systems like you would with an on-premises infrastructure.
 - Pay-as-you-go: You only need to pay for what you use by the hour or even by the minute.
 - Widespread network access: Cloud computing resources are available over the network and can be accessed by diverse customer platforms. No additional tools are required.

Contents

1. Compute Service Overview

2. Compute Service Planning

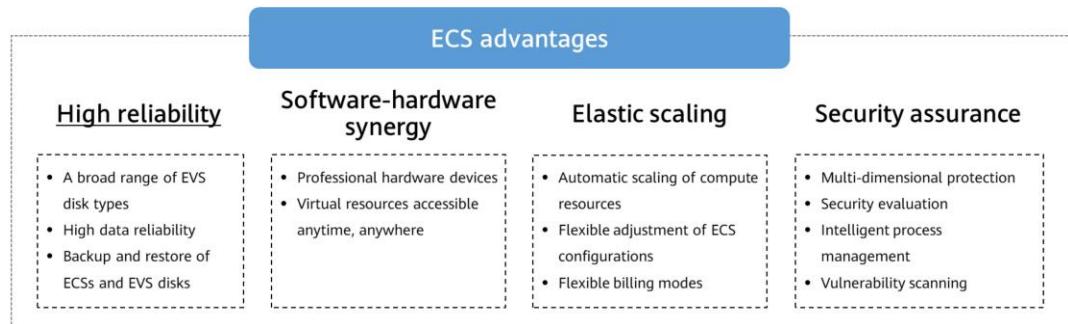
- ECS
 - DeH
 - BMS
 - Other compute services

3. IMS Planning

4. AS Planning

ECS

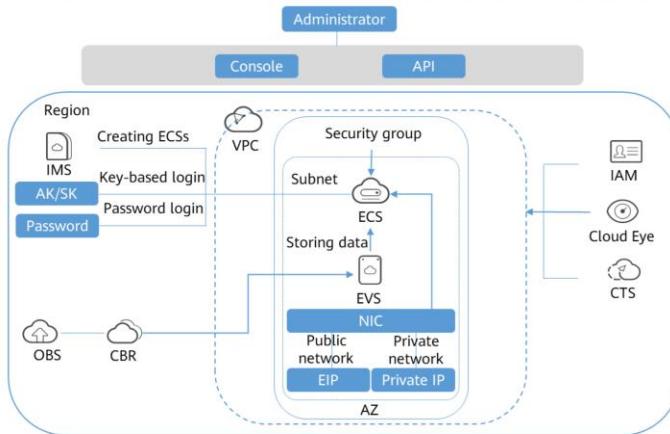
- ECS provides secure, scalable, on-demand compute resources, enabling you to flexibly deploy applications and workloads.



- **High reliability:**
 - A broad range of EVS disks: Common I/O, high I/O, general purpose SSD, ultra-high I/O, and extreme SSD disks are available for customer service requirements.
 - High data reliability: Scalable, reliable, and high-throughput virtual block storage is provided in a distributed architecture. This ensures that data can be quickly migrated and restored if any data replica is unavailable, preventing data from being lost because of a single hardware fault.
 - Backup and restoration of ECSs and EVS disks: You can configure automatic backup policies for in-service ECSs and EVS disks. You can also configure policies on the management console or use APIs to back up the data of ECSs and EVS disks at a specified time.
- **Security assurance:**
 - Multiple security services: Web Application Firewall (WAF), Vulnerability Scan Service (VSS), and other security services provide multi-dimensional protection.
 - Security evaluation: The security of cloud environments is evaluated to help you quickly identify security vulnerabilities and threats. Security configuration check and recommendations reduce or eliminate losses due to viruses or online attacks.
 - Intelligent process management: You can customize a whitelist to automatically prohibit the execution of unauthorized programs.
 - Vulnerability scanning: Comprehensive scanning services are provided, including general web vulnerability scans, third-party application vulnerability scans, port detection, and fingerprint identification.

ECS Architecture

- ECS works with other cloud services to provide compute, storage, and network resources.



ECS-related services:

Network services: There is a dedicated network environment that allows you to configure subnets and security groups for server access. You can also bind elastic IP addresses (EIPs) to ECSs for Internet access.

IMS: You can create an image from an existing ECS or use a private image to batch create ECSs for quick service deployment.

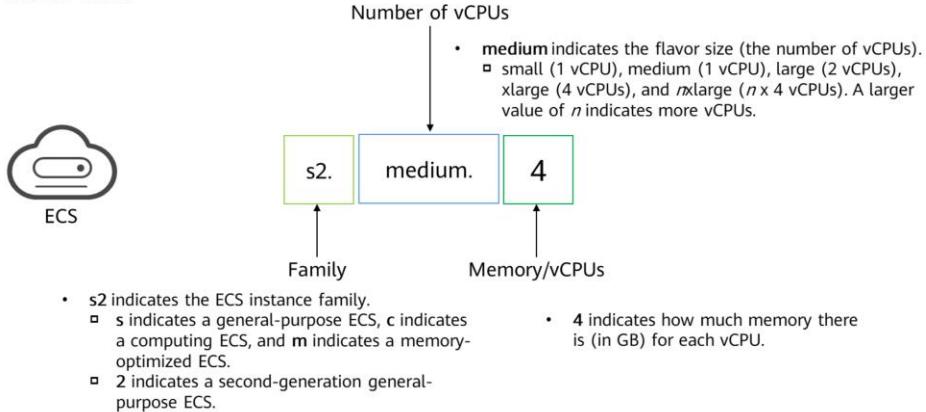
EVS: You can use EVS disks to store data. With Volume Backup Service (VBS), you can back up and restore data.

Cloud Backup and Recovery (CBR): You can back up data for EVS disks and ECSs, and use snapshot backups to restore the EVS disks and ECSs when necessary.

- ECS works with other cloud services to provide compute, storage, and network resources.
 - ECSSs are deployed in different AZs, so that if one AZ becomes faulty, other AZs in the same region will not be affected.
 - Cloud Eye lets you keep a close eye on the performance and resource utilization of ECSSs, ensuring their reliability and availability.

ECS Flavor Naming Rules

- Each ECS type provides various flavors with different vCPU and memory configurations for you to choose from.



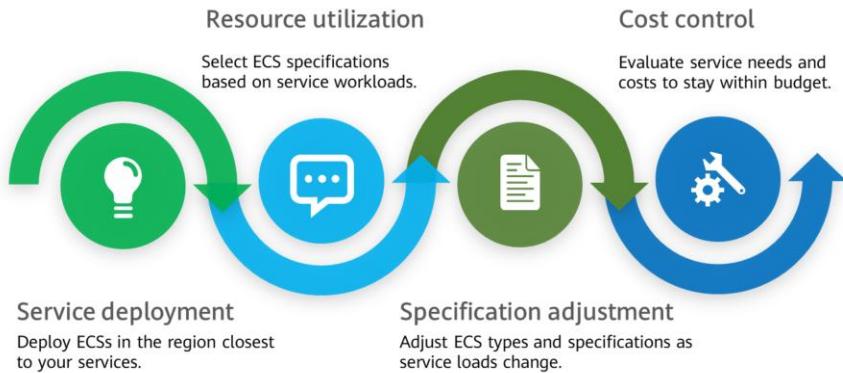
ECS Types

- Huawei Cloud provides a wide range of ECS types for you to choose from.

ECS Type	General Computing-basic	General Computing	General Computing-plus	Memory-optimized	Large-memory	Disk-intensive	High-Performance Computing	Ultra-high I/O	GPU-accelerated	AI-accelerated	FPGA-accelerated
x86	T6	S2	C3	M2	E3	D2	Hc2	Ir3	G6	Ai1	fp1c
		S3	C3ne	M3	E6	D3	H3	I3	G6v	Ai1s	fp1
		S6	C6	M6		D6			P2s		
		S7	C7	M7	E7	D7		I7/Ir7	Pi2		
Kunpeng			kC1	km1				kl1		kAi1s	

- General computing-basic
 - Suitable for scenarios that require moderate CPU performance generally but occasionally burstable high performance while keeping costs low
- General computing:
 - Suitable for websites and web applications, small-scale databases and cache servers, and light- and medium-workload enterprise applications with strict requirements on PPS
- General computing-plus
 - Suitable for heavy- and medium-load enterprise applications that have higher requirements on computing and network performance, such as web applications, e-commerce platforms, short video platforms, online games, insurance, and finance
- Memory-optimized
 - Suitable for massive parallel processing (MPP) data warehouses, MapReduce and Hadoop distributed computing, distributed file systems, network file systems, and log or data processing applications
- Disk-intensive
 - Suitable for distributed file systems, network file systems, and log or data processing applications
- High-performance computing
 - Computing and storage systems for genetic engineering, games, animations, and biopharmaceuticals
- The types displayed in the table were current as of the end of August 2022.

ECS Selection



- To select the right ECS type, consider the following factors:
 - Service deployment: Deploy ECSs in the region closest to your services to reduce network delay and improve the access speed.
 - Resource utilization: Make full use of purchased cloud resources. Do not buy more capacity than is needed.
 - Specification adjustment: In the subsequent content, we'll examine a hypothetical startup to look at how to select the right ECS types for different development stages (startup, growth, and maturity).
 - Cost control: Selecting the right ECS types and specifications help control costs. Evaluate your service scale and budget and scale up ECSs or change ECS types to meet service demands.

ECS Selection for Startups

- Generally, startups deploy services in the region where the companies are located. Their services are typically composed of website portals, small databases, and simple applications. As traffic is usually light in the early stages, they usually don't have high requirements on server performance.

T6	S6	S7
General computing-basic 1-16 vCPUs Maximum PPS: 0.6 million Performance restricted by CPU credits	General-purpose computing 1-8 vCPUs Maximum PPS: 0.5 million 2nd Gen Intel® Xeon® Scalable processors	General-purpose computing 1-16 vCPUs Maximum PPS: 1 million 3rd Generation Intel® Xeon® Scalable processors
<ul style="list-style-type: none">✓ Development and staging environments✓ Small websites✓ Light-load applications	<ul style="list-style-type: none">✓ Websites and web applications✓ Lightweight database services✓ Light- and medium-load enterprise applications	<ul style="list-style-type: none">✓ Websites and web applications✓ Lightweight database services✓ Medium- or light-load enterprise applications

13 Huawei Confidential



- T6 family:
 - The performance of general-computing basic T6 ECSs is restricted by the benchmark performance and CPU credits.
 - Suitable for scenarios where the CPU usage is low but requires burstable CPU power, for example, microservices.
- S6 family:
 - S6 ECSs are equipped with second-generation Intel® Xeon® Scalable processors and Huawei 25GE high-speed intelligent NICs that cost-effectively provide high network bandwidth and PPS throughput.
 - Suitable for websites and web applications with high requirements for PPS
- S7 family:
 - S7 ECSs are equipped with third-generation Intel® Xeon® Scalable processors and Huawei 25GE high-speed intelligent NICs that cost-effectively provide high network bandwidth and PPS throughput.
- What is PPS?
 - PPS, short for packets per second, is the number of network data packets that can be processed by an ECS per second, including the number of sent and received packets, including both private and public traffic. The maximum PPS is the maximum number of data packets an ECS can process, both incoming and outgoing per second.

ECS Selection for Growing Companies

- As website traffic and app user volumes increase, companies enter a growth stage. As much as their budget permits, they will start purchasing more powerful servers to provide more powerful compute.

C3	C6s	C7
<p>General computing-plus</p> <p>2-60 vCPUs Maximum PPS: 5 million Intel® Xeon® Scalable processors</p> <ul style="list-style-type: none">✓ Websites and web applications✓ Small- and medium-sized database services✓ Enterprise-class applications of various types and scales	<p>General computing-plus</p> <p>2-64 vCPUs Maximum PPS: 8.5 million 2nd Gen Intel® Xeon® Scalable processors</p> <ul style="list-style-type: none">✓ Game or audio servers✓ General database services✓ Heavy- and medium-load enterprise applications	<p>General computing-plus</p> <p>2-128 vCPUs Maximum PPS: 12 million 3rd Generation Intel® Xeon® Scalable processors</p> <ul style="list-style-type: none">✓ High-performance web applications✓ General database services✓ Heavy-load enterprise applications
14 Huawei Confidential		 HUAWEI

- C3 family:
 - C3 ECSs use Intel® Xeon® Scalable processors to provide high and stable computing performance. Working in high-performance networks, the C3 ECSs deliver higher performance and stability, meeting enterprise-class application requirements.
 - Suitable for small- and medium-sized databases, cache clusters, and search clusters that have high requirements on stability.
- C6s family:
 - C6s ECSs use second-generation Intel® Xeon® Scalable processors to provide high performance, high stability, low latency, and cost-effectiveness.
 - Suitable for Internet, gaming, and rendering scenarios, especially those that require high computing and network stability.
- C7 family:
 - C7 ECSs use third-generation Intel® Xeon® Scalable processors to provide enhanced compute, security, and stability. A C7 ECS can be configured with up to 128 vCPUs and 3,200 MHz memory. C7 ECSs support secure reboot and provide secure, trusted cloud environment for applications to run in.
 - Suitable for heavy- and medium-load enterprise applications that demand more compute and network performance, such as web applications, e-commerce platforms, short video platforms, online games, insurance, and finance applications.

ECS Selection for Mature Companies

- As services continue to expand, companies focus more on service stability than cost effectiveness. Mature companies will select different ECS types to address specific service needs.

M7 Memory-optimized	D7 Disk-intensive	I7 Ultra-high I/O
2-128 vCPUs Maximum PPS: 12 million Maximum memory: 1 TB	4-72 vCPUs Maximum PPS: 9 million Maximum storage: 36 x 4 TB local storage	2-96 vCPUs Access to local disks responded within milliseconds Maximum IOPS per disk: 0.75 million
✓ High-performance databases ✓ In-memory databases ✓ Distributed computing	✓ Big data computing ✓ Data warehouse services ✓ Log search	✓ High-performance relational databases ✓ NoSQL databases ✓ ElasticSearch

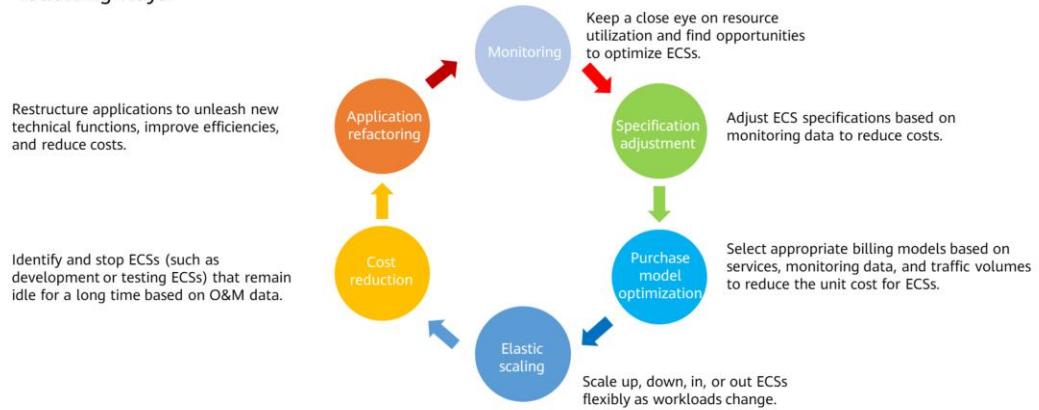
15 Huawei Confidential



- M7 family:
 - M7 ECSs use third-generation Intel® Xeon® Scalable processors to provide enhanced compute, security, and stability. An M7 ECS can be configured with up to 128 vCPUs and 3,200 MHz RAM frequency. M7 ECSs support secure reboot and provide secure, trusted cloud environment for applications.
 - Suitable for high-performance data warehouses, in-memory databases, MapReduce and Hadoop distributed computing, distributed file systems and network file systems, and log or data processing applications.
- D7 family:
 - D7 ECSs are mainly used for massively parallel processing (MPP) data warehouses, MapReduce and Hadoop distributed computing, and big data computing.
 - Suitable for distributed file systems, network file systems, and log or data processing applications.
- I7 family:
 - I7 ECSs use high-performance local NVMe SSDs to provide high IOPS and low read/write latency.
 - Suitable for high-performance relational databases, non-relational databases, and ElasticSearch search.

ECS Optimization

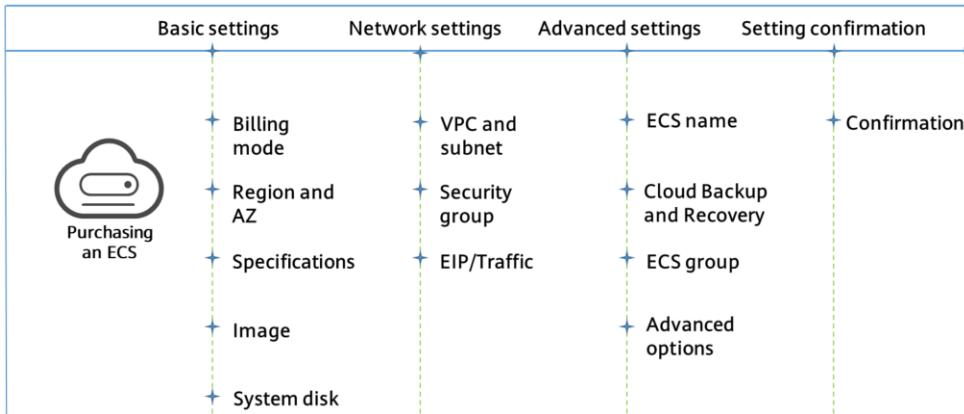
- To reduce costs and meet service needs, O&M personnel need to continuously optimize ECSs in the following ways.



- ECSs should be continuously optimized.

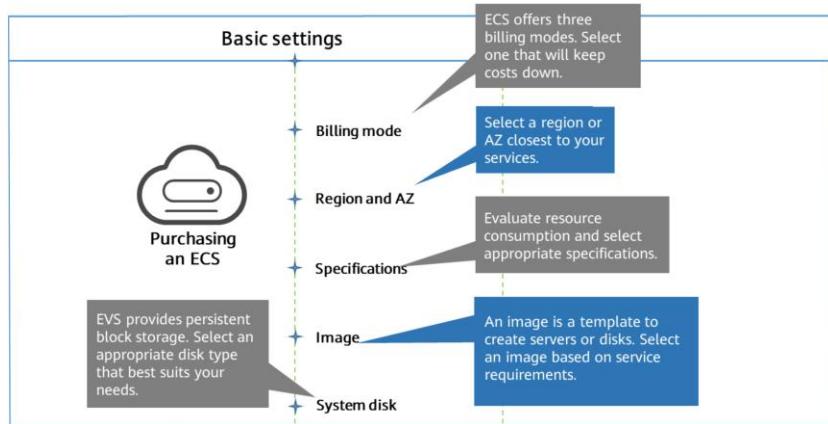
Purchasing an ECS – Overview

- When purchasing an ECS, you need to configure basic, network, and advanced parameters.



Purchasing an ECS – Basic Settings

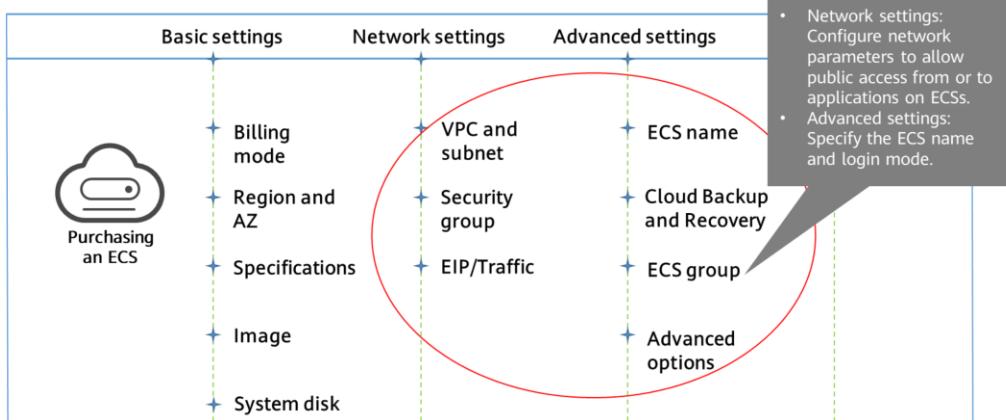
- Configure basic parameters.



- Billing modes:** yearly/monthly, pay-per-use, and spot price
 - Yearly/monthly:** You can purchase a yearly/monthly ECS subscription and enter your required duration. Yearly/monthly subscriptions are pre-paid with a single, lump sum payment.
 - Pay-per-use:** You do not need to set a required duration after setting ECS configurations. The system bills your account based on the service duration.
 - Spot price:** Huawei Cloud sells available spare compute resources at a discount. The price changes in real time depending on market supply and demand.
- Region and AZ:** ECSs in different regions cannot communicate with each other over a private network. Select a region closest to your services to ensure low network latency and quick access.
- Specifications:** A broad set of ECS types are available for you to choose from. You can choose from existing types and flavors in the list, or enter a flavor or specify vCPUs and memory size to search for the flavor suited to your needs.
- Image:** An image is a server or disk template that contains an OS or service data and necessary application software. IMS provides public, private, Marketplace, and shared images.
- System disk types:** high I/O, general-purpose SSD, ultra-high I/O, and extreme SSD. By default, you need to specify the type and size of the system disk.

ECS Creation – Other Configurations

- Configure network parameters and advanced settings.



- Network settings for an ECS:

- Subnet: A subnet is a range of IP addresses in your VPC and provides IP address management and DNS resolution functions for ECSs in it. The IP addresses of all ECSs in a subnet belong to the subnet.
- Security group: A security group is a collection of access control rules for ECSs that have the same security protection requirements and that are mutually trusted. It helps to enhance ECS security.
- Extension NIC: optional

- Advanced settings for an ECS:

- 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 the ECSs and EVS disks when necessary.
- ECS group (Optional): An ECS group allows ECSs within the group to be automatically allocated to different hosts.
- Advanced options: You can configure other advanced and optional settings.

ECS Use Cases

- Background: The convenience and product diversity of e-commerce platforms have had a tremendous impact on consumer shopping habits, even affecting offline shopping. To avoid being restricted by e-commerce platforms and to create a strong brand, a company planned to build an online store to acquire new customers as well as obtaining consumer information to improve retention.



Contents

1. Compute Service Overview
2. **Compute Service Planning**
 - ECS
 - DeH
 - BMS
 - Other compute services
3. IMS Planning
4. AS Planning

DeH

- A Dedicated Host (DeH) is a physical server fully dedicated for your own services, ensuring isolation, security, and performance for your ECSSs.

DeH advantages

Exclusive

- Fully dedicated, DeH hosts only your own ECSSs.

Flexible

- Scale up ECSSs deployed on DeH based on your service requirements to keep your services reliable.

Cost-effective

- Use existing server-bound software licenses. You can take advantage of these existing investments to keep costs down.

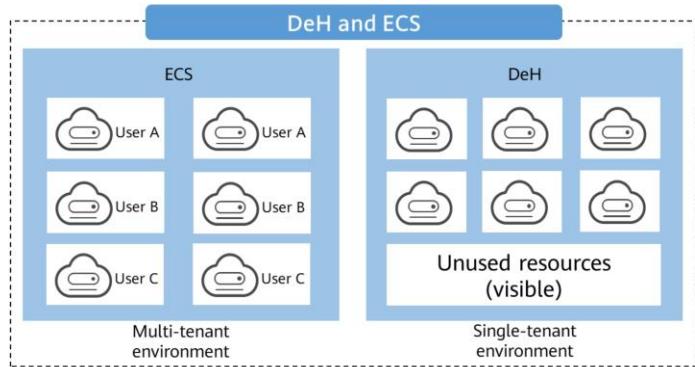
Secure

- Keep your services secure with host-level compute isolation, making it easier to meet compliance requirements.

- Cost-effectiveness: DeH allows you to bring your own license (BYOL), such as licenses for Microsoft Windows Server, Microsoft SQL Server, and Microsoft Office.
- Security: DeH isolates compute resources to prevent your workloads on DeHs from being affected by those of other tenants.
- Compliance: Physically isolated servers meet the compliance requirements of sensitive services.
- Flexibility: You can apply for your DeHs flexibly. Your DeHs will be allocated within several minutes.
- Reliability: DeH provides 99.95% availability.

Differences Between DeH and ECS

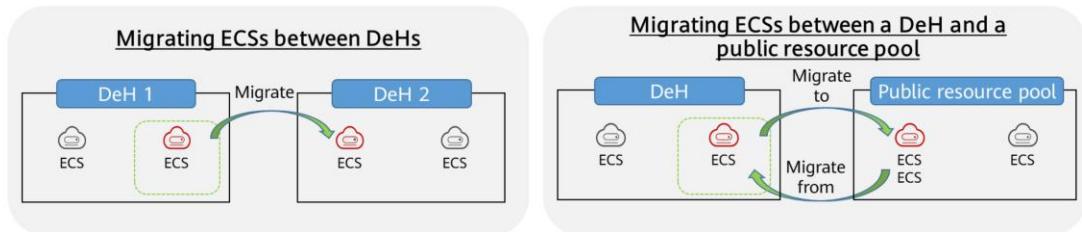
- The physical resources of the DeH are not shared with others. You can obtain the detailed information on the DeH, such as sockets, physical cores, CPU type, and memory size. So, you can create ECSs of specified flavors based on the DeH flavor.



- A DeH is fully dedicated for your own ECSs, ensuring the isolation, security, and performance. You can bring your own license (BYOL) to DeH to reduce the costs on software licenses and facilitate the independent management of ECSs.

Migrating ECSs

- For services that have high requirements for flexible resource usage, such as software R&D, resources need to be provisioned, deleted, and backed up on demand to facilitate deployment and destruction of development and testing environments for fault locating. DeH can meet aforesaid requirements.
- ECSs deployed on a DeH can be migrated to another DeH or a public pool to meet resource migration requirements during development.



- Notes:
 - Only stopped ECSs can be migrated.
- Application scenario: If you do not use the ECSs deployed on a DeH or want to delete them after a period of time, you can migrate the ECSs to a public resource pool.

DeH Application Scenarios



Commercial license

If you have a licensed OS or software (licensed based on the number of physical sockets or the number of physical cores), you can bring your own license and migrate your services to the cloud platform.

Resource isolation

DeH is ideal for service scenarios with higher requirements on server performance and stability such as finance, securities, and gaming applications. DeH guarantees the stability of CPUs and network I/O, ensuring smooth running of applications.

Independent resource planning

You can exclusively use a physically isolated host to meet your high compliance and security requirements.

Contents

1. Compute Service Overview
2. **Compute Service Planning**
 - Elastic Cloud Server
 - DeH
 - **BMS**
 - Other compute services
3. IMS Planning
4. AS Planning

BMS

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

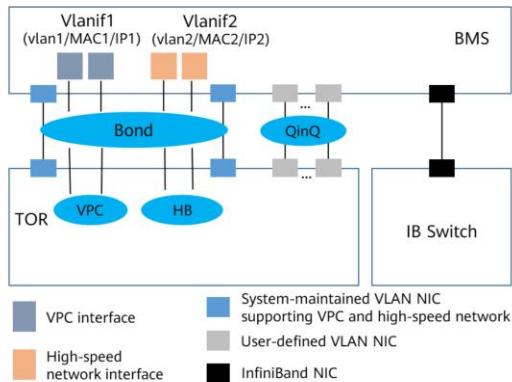
BMS advantages

Agile deployment	High performance	High security and reliability	Quick integration
<ul style="list-style-type: none">Quick provisioning (booted from EVS disks)Self-service lifecycle management and O&M	<ul style="list-style-type: none">Improved deployment density and performance that meet the requirements of mission-critical services such as enterprise databases, big data, containers, HPC, and AI	<ul style="list-style-type: none">Enterprise-grade data security, secure and reliable services compliant with regulations	<ul style="list-style-type: none">Easy cooperation with other cloud resources over VPC

- High security and reliability:
 - BMS provides you with dedicated computing resources. You can add servers to VPCs and security groups for network isolation and integrate related components for server security. BMSs run on a QingTian architecture and 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, so the compute resources are fully dedicated to running services. The QingTian they run on, an architecture from Huawei, 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 of when you submit your order. You can manage your BMSs throughout their lifecycle from the management console or using open APIs with SDKs.
- Quick integration:
 - BMSs can easily cooperate with the other cloud resources in a VPC, just like ECSs do, to run a variety of cloud solutions (such as databases, big data applications, containers, HPC, and AI solutions), accelerating cloud transformation.

BMS Network

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



28 Huawei Confidential



Virtual Private Cloud (VPC)

A VPC is a logically isolated, configurable, and manageable virtual network. It helps improve the security of cloud resources and simplifies network deployment.

Enhanced high-speed network

An enhanced high-speed network uses upgraded hardware and software and provides performance superior to a standard high-speed network.

User-defined VLAN

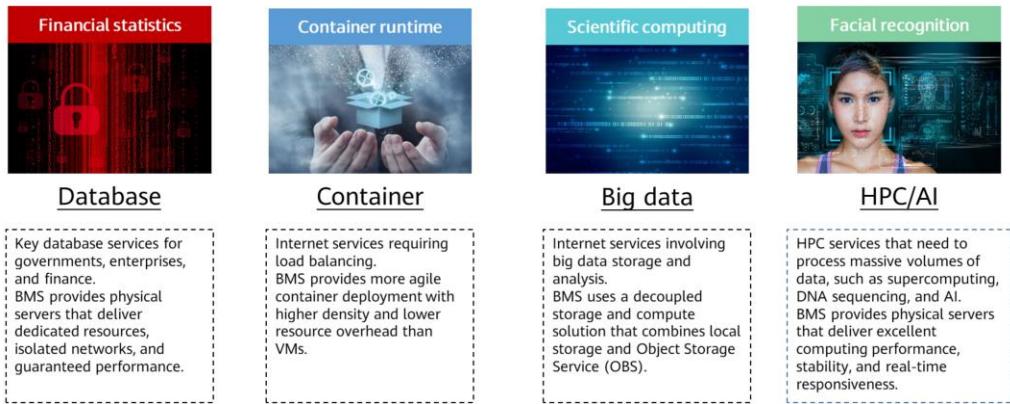
You can use Ethernet NICs that are not being used by the system to configure a user-defined VLAN. QinQ technology is used to isolate networks and provide additional physical planes and bandwidths. You can create VLANs to isolate network traffic.

InfiniBand network

An InfiniBand network features low latency and high bandwidth and is often used for High Performance Computing (HPC) projects. It uses the 100 Gbit/s Mellanox InfiniBand NIC, dedicated InfiniBand switch, and controller software UFM to ensure network communication and management.

- VPC:
 - 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.
- Enhanced 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:
 - User-defined VLAN NICs are deployed in pairs. You can configure NIC bonds to ensure high availability. User-defined VLANs in different AZs cannot communicate with each other.

BMS Application Scenarios



- Database:
 - Mission-critical database services of governments and financial institutions must be deployed on physical servers with dedicated resources, isolated networks, and guaranteed performance. BMS meets these requirements by providing high-performance servers dedicated to individual users.
- Big data:
 - Internet services involving big data storage and analysis. BMS uses a decoupled storage and compute solution that combines local storage and Object Storage Service (OBS).
- Container:
 - Internet services requiring load balancing. BMS provides more agile container deployment with higher density and lower resource overhead than VMs. Cloud native technologies reduce the cost of cloud transformation.
- HPC/AI:
 - High-performance computing applications, such as supercomputing and DNA sequencing, need to process massive volumes of data. BMS meets this requirement by providing excellent computing performance, stability, and real-time responsiveness.

Comparison Between a BMS and Other Servers

- A BMS is a physical server. Unlike ECSs whose underlying resources may be shared by different users, BMS resources are dedicated to individual users. This makes BMSs ideal for mission-critical applications or applications that require high performance and security.
- A Dedicated Host (DeH) is also a physical server, but the difference is that:
 - BMS uses a bare metal architecture and does not provide a virtualization platform.
 - DeH uses a virtualization architecture. After purchasing a DeH, you can use ECS public images to provision ECSs on it.

Item	BMS	DeH
Virtualization provided	No	Yes
Usage	Use each BMS as a whole server, or install virtualization software on it.	Provision ECSs on a DeH.
Specifications	BMS specifications	DeH specifications and ECS specifications
Image	BMS images	ECS images

Contents

1. Compute Service Overview
2. **Compute Service Planning**
 - Elastic Cloud Server
 - DeH
 - BMS
 - Other compute services
3. IMS Planning
4. AS Planning

CPH

- Cloud Phone (CPH) is a service that can migrate applications on mobile phones to virtual phones running native Android on Kunpeng cloud servers. You can remotely control the cloud phones in real time and run Android applications on the cloud. They are a convenient way to efficiently build applications using cloud computing.

CPH advantages

First in the public cloud industry

- Cloud Phones use the same Arm architecture as mobile phones. Huawei Kunpeng processors augment the performance of cloud phones by 80% compared with traditional simulators.

Flexible and elastic resources

- Massive BMSs allow for elastic capacity adjustment.

Powerful GPUs

- Professional GPU hardware acceleration for running large-scale games.

Flexible specification adjustment

- Adapt to fast-changing performance requirements of different applications.

Compatible with native apps

- Cloud phones are compatible with native commands and can run mainstream games and applications as mobile phones.

Data security

- Service data and confidential information is stored on the cloud, providing enterprise-grade security.

- Based on Huawei TaiShan Arm servers, Cloud Phone integrates multiple highly cost-effective GPUs to provide professional graphics processing capabilities.
- Cloud phones provide video, audio, and touch SDKs. You can develop applications based on terminals to obtain audios and videos of cloud phones. Alternatively, you can collect touch instructions, for example, touch, slide, or click instructions, and execute them on cloud phones.

CCE

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

CCE advantages			
Easy-to-use console	Excellent performance	Secure and reliable	Open and compatible
<ul style="list-style-type: none">Create Kubernetes clusters in just a few clicks.Scale clusters and workloads.Upgrade clusters.	<ul style="list-style-type: none">Supports high-concurrency and large-scale services.Improves computing performance with a high-performance architecture and high-speed InfiniBand network cards.	<ul style="list-style-type: none">Allows three master nodes in different AZs for the cluster control plane to ensure high availability.Gives users full control on clusters they create.	<ul style="list-style-type: none">Fully compatible with native Kubernetes and Docker.

- Easy-to-use:
 - Deployment and O&M of containerized applications can be automated and performed all in one place throughout the application lifecycle.
 - Helm charts are pre-integrated, delivering out-of-the-box usability.
- High performance:
 - CCE draws on years of field experience in compute, networking, 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:
 - 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.
 - Clusters are private and completely controlled by users with deeply integrated IAM and Kubernetes RBAC. You can set different RBAC permissions for IAM users on the console.
- 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.

FunctionGraph

- This service hosts and computes event-driven functions while maximizing scalability and efficiency. Simply write your code and set running conditions without provisioning or managing servers.

FunctionGraph advantages

Pay-per-Use

- Billing is by number of function requests and execution duration. No charge when code is not running.

Serverless

- Your code is automatically run without the need for provisioning or managing servers.

Dynamic Resource Allocation

- Resources are dynamically allocated to minimize usage and reduce costs.

High Scalability

- Resources are automatically scaled to service spikes.

Event-based Triggering

- FunctionGraph integrates with multiple cloud services (including SMN and OBS) for versatility and efficiency.

High Availability

- The system creates new instances in case of an instance error and reclaims affected resources.

- When using FunctionGraph, you do not need to apply for or pre-configure any compute, storage, or network services. Simply upload and run code in supported runtimes. FunctionGraph provides and manages underlying compute resources, including CPUs, memory, and networks. It also supports configuration and resource maintenance, code deployment, automatic scaling, load balancing, secure upgrade, and resource monitoring.

Contents

1. Compute Service Overview
2. Compute Service Planning
- 3. IMS Planning**
4. AS Planning

IMS

- Image Management Service (IMS) allows you to manage the entire lifecycle of your images. You can create Elastic Cloud Servers (ECSs) or Bare Metal Servers (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.

IMS advantages			
Convenient	Unified	Flexible	Secure
<ul style="list-style-type: none">• Batch creation of identical cloud servers simplifies service deployment• Multiple methods for creating private images• Cross-account, cross-region service migration	<ul style="list-style-type: none">• Centralized image management• Unified deployment and upgrade of applications, improving O&M efficiency• Cross-cloud migration enabled by public images that are compliant with industry standards	<ul style="list-style-type: none">• Image lifecycle management on the console or using APIs• A variety of image types• Useful in many different environments	<ul style="list-style-type: none">• Public images that have been thoroughly tested and include mainstream operating systems• Images are backed up using Object Storage Service (OBS) for high data reliability• Images encrypted using Key Management Service (KMS)

- Convenient: You can use a public, Marketplace, or private image to create ECSs in batches, simplifying service deployment. You can also share, replicate, or export images between different accounts, regions, or even cloud platforms.
- Secure: To ensure data reliability and durability, multiple copies of image files are stored using Object Storage Service (OBS). You can use the envelope encryption provided by Key Management Service (KMS) to encrypt private images.
- Flexible: You can manage the lifecycle of images using the management console or APIs as needed. IMS can meet your requirements no matter you want to migrate servers to the cloud, back up server environments, or migrate servers between different accounts or regions on the cloud.
- Unified: IMS provides a unified platform to simplify image maintenance. Images can be used to deploy and upgrade applications in a unified manner, improving application O&M efficiency and ensuring environment consistency.

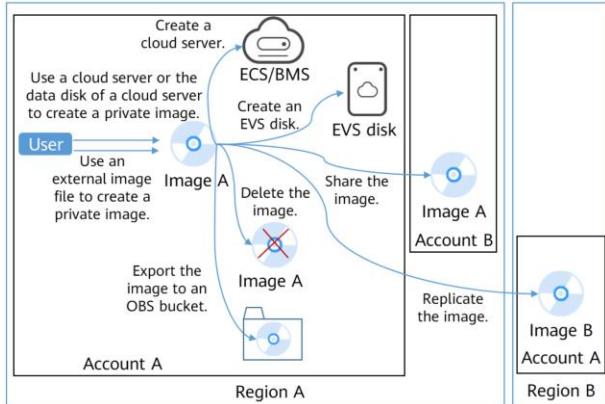
Image Types

- Images are classified as public, private, Marketplace, and shared. Public images are provided by the cloud platform, private images are those you create yourself, and shared images are private images that other tenants have shared with you.

Image Type	Description
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. Public images are very stable and their OS and any included software have been officially authorized for use. If a public image does not contain the application environments or software you need, you can use a public image to create an ECS and then deploy required software as needed. Public images include the following OSs to choose from: CentOS, Debian, openSUSE, Fedora, Ubuntu, EulerOS, and CoreOS.
Private image	A private image contains an OS or service data, preinstalled public applications, and a user's personal applications. Private images are only available to the users who created them. A private image can be a system disk image, data disk image, or full-ECS image.
Shared image	A shared image is a private image another user has shared with you.
Marketplace image	The Marketplace is an online store where you can purchase third-party images that have the OS, application environments, and software preinstalled. 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 in configuring and maintaining cloud servers. All the images are thoroughly tested and have been approved by Huawei Cloud before being published.

Private Image Lifecycle

- You can create an image by using an external image file, a cloud server, or the data disk of a cloud server.



Private image lifecycle management:

- After a private image is created and the image status is Normal, you can use it to create cloud servers or EVS disks.
- You can also share it with other accounts. These accounts can use the shared image to quickly create identical cloud servers or EVS disks.
- You can replicate the image to another region. Cross-region image replication is available only under the same account.
- If you want to save the image, you can export it to your OBS bucket and download it to your local PC.

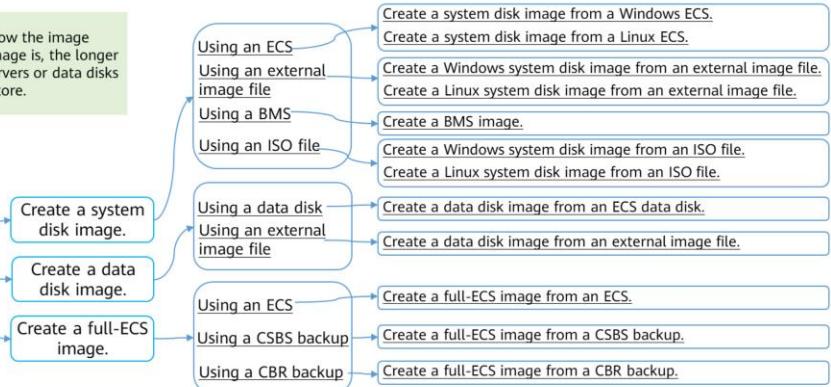
- A private image can be a system disk image, data disk image, or full-ECS image.
 - A system disk image contains an OS and pre-installed software for various services. You can use a system disk image to create cloud servers and migrate your services to the cloud.
 - A data disk image contains only service 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. A full-ECS image is created using differential backups and the creation takes less time than creating a system or data disk image of the same size.

Creating a Private Image

- A private image is an image available only to the user who created it. It contains an OS, preinstalled public applications, and a user's personal applications. A private image can be a system disk image, data disk image, or full-ECS image. You can create or import a private image.

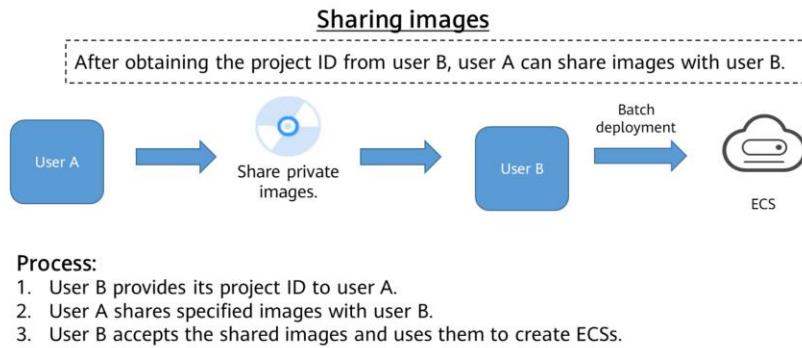
The image size depends on how the image was created. The larger an image is, the longer it will take to create cloud servers or data disks and the more it will cost to store.

Create a private image.



Sharing Images

- You can share your private images with other tenants. Then, the tenants can use the images to create identical ECSs in batches.

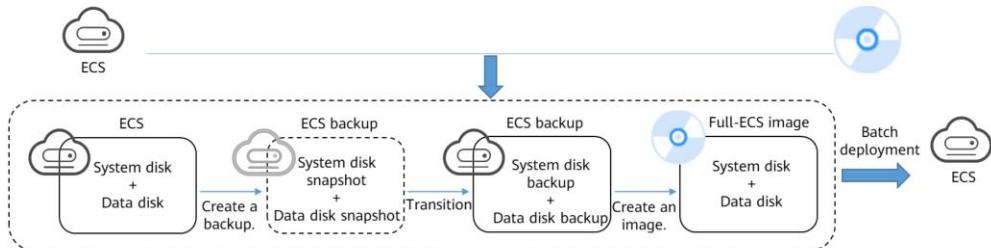


- Before you share an image, ensure that:
 - You have obtained the project ID of the target user.
 - Any sensitive data has been deleted from the image.
- Constraints:
 - You cannot share private images that have been published in Marketplace.
 - You can share images only within a given region. To share an image across regions, you need to replicate the image to the target region first.
 - A system disk image or data disk image can be shared with a maximum of 128 users, and a full-ECS image can be shared with a maximum of 10 users.
 - Encrypted images cannot be shared.
 - Only full-ECS images created from an ECS or a CBR backup can be shared.

Using an Image to Batch Deploy ECSs

Using an image to batch deploy ECSs

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



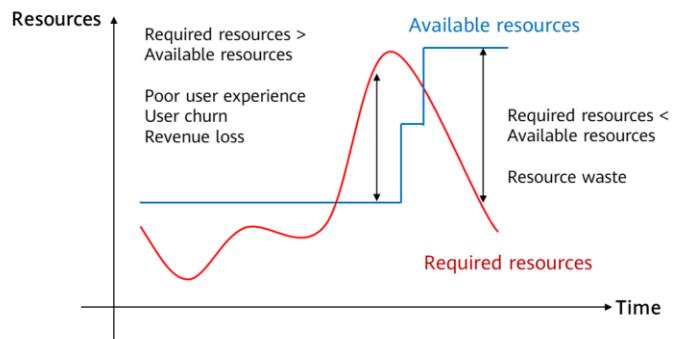
- When you submit a request for creating a full-ECS image from an ECS, the system will automatically create a backup for the ECS and then use the backup to create a full-ECS image.
- The time required for creating a full-ECS image depends on the disk size, network quality, and the number of concurrent tasks.
- The ECS used to create a full-ECS image must be in Running or Stopped state. To create a full-ECS image containing a database, use a stopped ECS.
- When a full-ECS image is being created, if you detach the system disk from the ECS or stop, start, or restart the ECS, the image creation will fail.
- If there are snapshots of the system disk and data disks but the ECS backup creation is not complete, the full-ECS image you create will only be available in the AZ where the source ECS is and can only be used to provision ECSs in this AZ. You cannot provision ECSs in other AZs in the region until the original ECS is fully backed up and the full-ECS image is in the Normal state.
- If you use a full-ECS image to change an ECS OS, only the system disk data can be written into the ECS. Therefore, if you want to restore or migrate the data disk data of an ECS by using a full-ECS image, you can only use the image to create a new ECS rather than use it to change the ECS OS.

Contents

1. Compute Service Overview
2. Compute Service Planning
3. IMS Planning
- 4. AS Planning**

Elastic Scaling

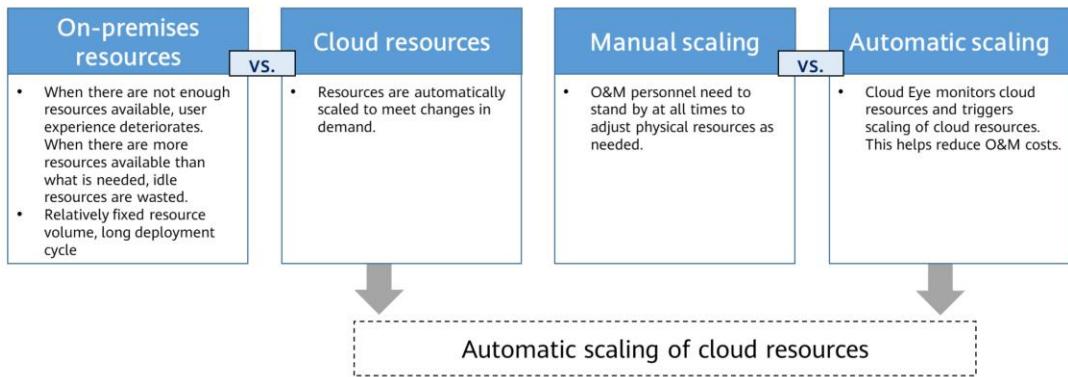
- Service requirements are always changing. Sometimes you need more resources but sometimes you need fewer. Elastic scaling lets you automate how groups of different resources respond to changes in demand. It allows you to automatically add and remove resources to respond to constantly changing service demand.



- When there are more resources available than what is needed, idle resources are wasted.
- When there are not enough resources available, user experience deteriorates. User churn increases and revenue is lost.

Automatic Scaling

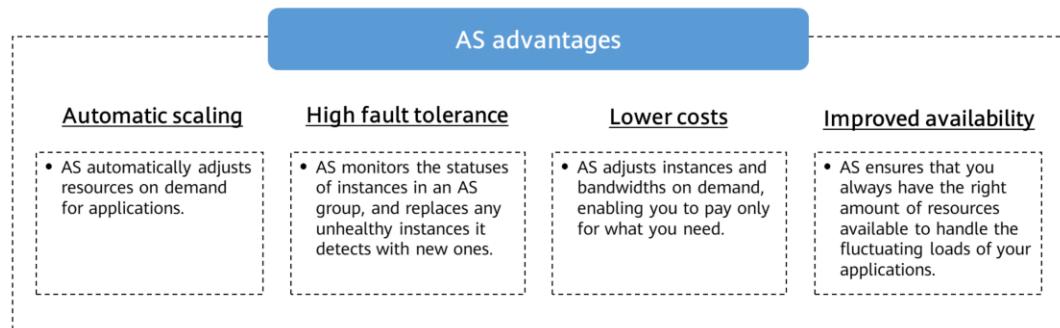
- Automatic scaling helps you make full use of IT resources and reduce labor costs.



- Automatic scaling helps you automatically meet customer requirements.

AS

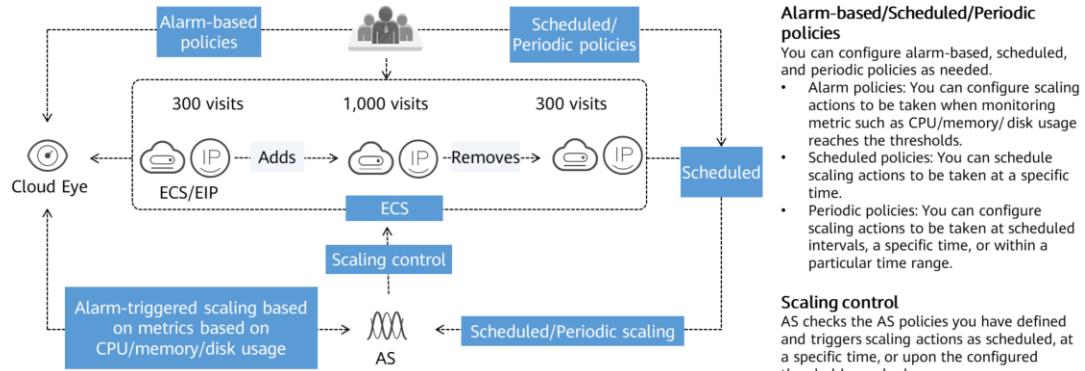
- Auto Scaling (AS) is a cloud service that can automatically adjust resources based on your service requirements and pre-configured policies. You can configure a scheduled, periodic, or alarm policy to adjust resources based on fluctuating service loads, avoiding overspending and ensuring stable services



- The process of using AS is as follows:
 - First, create an AS configuration. Then, create an AS group, and then configure an AS policy for the AS group you just created based on your service requirements.
- AS advantages:
 - Automatic scaling: When demand spikes, AS adds ECS instances and increases bandwidth to maintain service quality. When demand decreases, AS removes unneeded resources to avoid wasting resources.
 - Lower costs: AS can automatically adjust resources for applications. This enables you to allocate resources on demand, eliminate waste, and reduce costs.
 - Improved availability: With AS, your applications always have the right amount of resources at the right time. 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.
 - High fault tolerance: AS monitors the statuses of instances in an AS group, and replaces any unhealthy instances it detects with new ones.

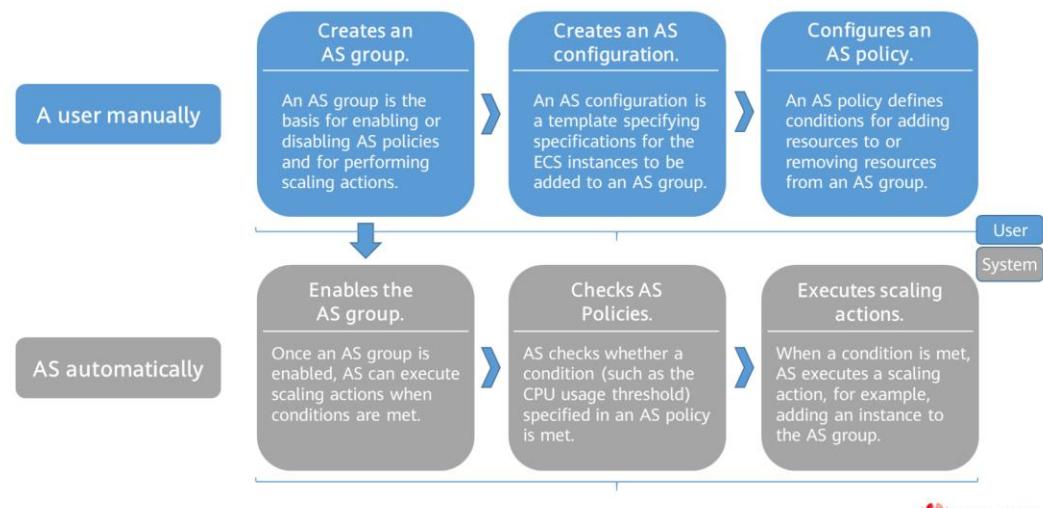
AS Architecture

- AS can scale ECS and bandwidth resources to keep up with changes in demand based on pre-configured AS policies.



- AS can work with Cloud Eye to make smarter scaling actions.
 - In the example shown here, when the number of access requests reaches 1,000, the existing resources cannot handle the demand. More resources are needed. When the peak hours pass, idle resources need to be removed to avoid waste and reduce costs.
 - AS can work together with Cloud Eye to do this automatically. When Cloud Eye detects resources reach a threshold you have specified in an AS policy, for example, CPU usage higher than 70%, memory usage higher than 80%, or access requests more than 500, AS automatically triggers scaling actions to add more resources.

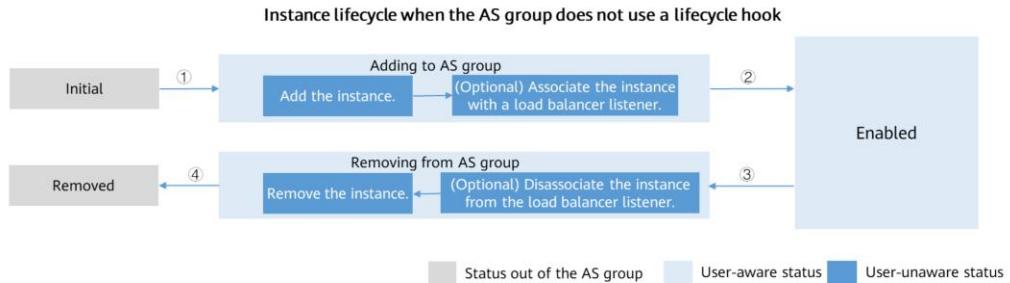
Process of Using AS



- When you use AS, you need to create an AS group, create an AS configuration, and then configure an AS policy for the AS group.
- Then AS checks whether the condition specified in the AS policy is met, and determines whether to execute a scaling action based on the results.
- An AS group consists of a collection of ECS instances and AS policies that have similar attributes and application scenarios. An AS group is the basis for enabling or disabling AS policies and performing scaling actions.
- An AS configuration defines the specifications of instances to be added to an AS group. The specifications include the ECS image and system disk size.
- An AS policy can trigger scaling actions to scale ECS and bandwidth resources for an AS group. An AS policy defines the conditions for triggering a scaling action and the operation that will be performed. When the condition is met, a scaling action is triggered automatically. AS supports alarm-based, scheduled, and periodic scaling policies.
- When creating an AS group, you need to configure parameters, such as Max. Instances, Min. Instances, Expected Instances, and Load Balancing.

Instance Lifecycle (1)

- An ECS instance in an AS group has a lifecycle that starts when it is created and ends when it is removed from the AS group.

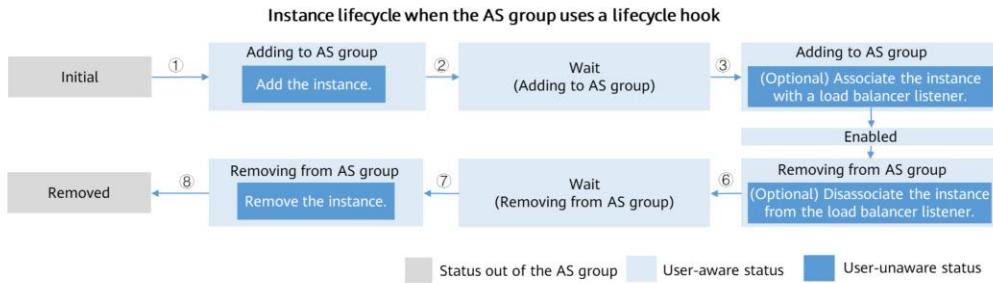


- When an ECS instance is added to an AS group, it goes through the **Adding to AS group**, **Enabled**, and **Removing from AS group** statuses. Then it is finally removed from the AS group.

- The instance status changes from Initial to Adding to AS group when either of following occurs:
 - You manually increase the expected number of instances for the AS group or AS automatically adds instances to the AS group.
 - You manually add instances to the AS group.
- The instance status changes from Enabled to Removing from AS group when any of the following occurs:
 - You manually decrease the expected number of instances for the AS group or the system automatically removes instances from the AS group.
 - AS removes unhealthy instances from the AS group.
 - You manually remove instances from the AS group.

Instance Lifecycle (2)

- An ECS instance in an AS group has a lifecycle that starts when it is created and ends when it is removed from the AS group.



- When a scaling action occurs in the AS group, the required instances are suspended by the lifecycle hook and remain in the wait status until the timeout period ends or you manually call back the instances.
- You can perform custom operations on the instances when they are in the wait status. For example, you can install or configure software on an instance before it is added to the AS group or download log files from an instance before it is removed.

- When a scale-out or scale-in event occurs in the AS group, the required instances are suspended by the lifecycle hook and remain in the wait status until the timeout period ends or you manually call back the instances. You can perform custom operations on the instances when they are in the wait status. For example, you can install or configure software on an instance before it is added to the AS group or download log files from an instance before it is removed.

AS Application Scenarios



Heavy-traffic websites

Service load on heavy-traffic websites keeps changing. AS dynamically scales in or out ECS instances based on monitored ECS metrics, such as CPU and memory usage.



E-commerce

During big promotions, E-commerce websites need more resources. AS automatically adds resources within minutes to ensure that promotions go smoothly.

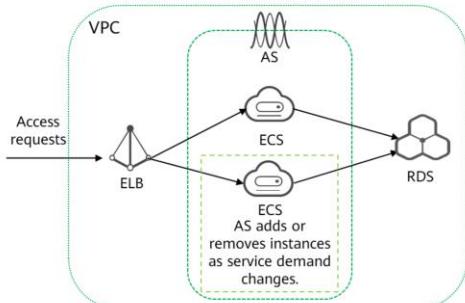


Live streaming

Live streaming websites broadcast popular programs during a fixed time of period every day. AS lets you automate resource scaling to respond to traffic changes, providing users with smooth video playback.

AS Use Cases

- Background: An e-commerce platform launched a large promotion, attracting a huge number of visits and resulting in traffic surges.



Solution
Automatic scaling + Traffic load balancing AS automatically adds or removes ECS instances to or from the backend ECS group of the ELB balancer. ELB then routes requests to backend servers based on the load balancing algorithm and health check results.
Benefits
<ul style="list-style-type: none">With the help of AS and ELB, they achieved high service stability and could ensure an excellent user experience even during peak hours.AS automatically removed idle resources, helping them reduce IT costs and avoid wasting resources.

51 Huawei Confidential



- You can use AS and ELB to confidently deal with changes in service demand. When the workload goes up or down, AS scales out or in instances to maintain steady performance at the lowest possible cost. ELB can manage incoming requests by optimizing traffic routing to prevent instance overload.
- After ELB is enabled for an AS group, AS automatically associates a load balancing listener with any instances newly added to the AS group. Then, ELB automatically distributes traffic to all healthy instances in the AS group through the listener to improve system availability. The instances in the AS group may be hosting multiple applications. You can bind different load balancing listeners to the AS group to listen to each of these applications. This further improves your service scalability.

Quiz

1. (Multiple-Answer Question) On Huawei Cloud, which of the following image types are supported by Image Management Service (IMS)?
 - A. Public image
 - B. Private image
 - C. Shared image
 - D. Marketplace image
2. (Multiple-Answer Question) Which of the following policies are supported by Auto Scaling (AS)?
 - A. Alarm policy
 - B. Scheduled policy
 - C. Periodic policy
 - D. Notification policy

- 1. Answer: ABCD
- 2. Answer: ABC

Quiz

1. (Discussion) What are the advantages of cloud servers over on-premises self-built servers?

2. (Discussion) What should be considered for using cloud servers in terms of security, cost, reliability, performance, and scalability?

- Discussion 1:
 - Cloud servers are managed by cloud service providers.
 - On-premises self-built servers are managed by users.

- Discussion 2:
 - Security: dynamic and static data security, network security, and access control
 - Cost: server selection and purchase model
 - Reliability: cluster deployment
 - Performance: meet service requirements and reserve redundancy
 - Scalability: use Auto Scaling (AS) to dynamically adjust compute resources

Summary

- In this course, you have learned about the compute cloud services on Huawei Cloud. In the technical evolution from physical hardware to cloud platforms and then to cloud services, new technologies and products have emerged, for example, Elastic Cloud Server (ECS) and Cloud Container Engine (CCE). Both of them can be used to deploy applications but their technical structures are totally different. Understanding technical principles of each cloud service will help you get more involved in the cloud migration journey.

Acronyms and Abbreviations

- API: Application Programming Interface
- AS: Auto Scaling
- BMS: Bare Metal Server
- CBR: Cloud Backup and Recovery
- CDN: Content Delivery Network
- DCS: Distributed Cache Service
- DRS: Data Replication Service
- DNS: Domain Name Service
- DDoS: Distributed Denial of Service
- DevOps: Development Operation
- DIS: Data Ingestion Service
- DLI: Data Lake Insight
- EIP: Elastic IP Address
- ECS: Elastic Cloud Server
- ELB: Elastic Load Balance
- EVS: Elastic Volume Service
- GSLB: Global Server Load Balance
- HA: High Availability
- HPC: High Performance Computing
- IMS: Image Management Service
- IDC: Internet Data Center

Acronyms and Abbreviations

- LTS: Log Tank Service
- NAT: Network Address Translation
- OLAP: Online Analytical Processing
- OLTP: Online Transaction Processing
- RDS: Relational Database Service
- SMN: Simple Message Notification
- SFS: Scalable File Service
- SDRS: Storage Disaster Recovery Service
- UFM: Unified Fabric Manager
- VM: Virtual Machine
- VPC: Virtual Private Cloud
- VPN: Virtual Private Network

Recommendations

- Huawei iLearning
 - <https://e.huawei.com/en/talent/portal/#/>
- Huawei Cloud Help Center
 - <https://support.huaweicloud.com/intl/en-us/index.html>
- HUAWEI CLOUD Developer Institute
 - <https://edu.huaweicloud.com/intl/en-us/>
- Huawei Talent Online
 - <https://e.huawei.com/en/talent/portal/#/>

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。

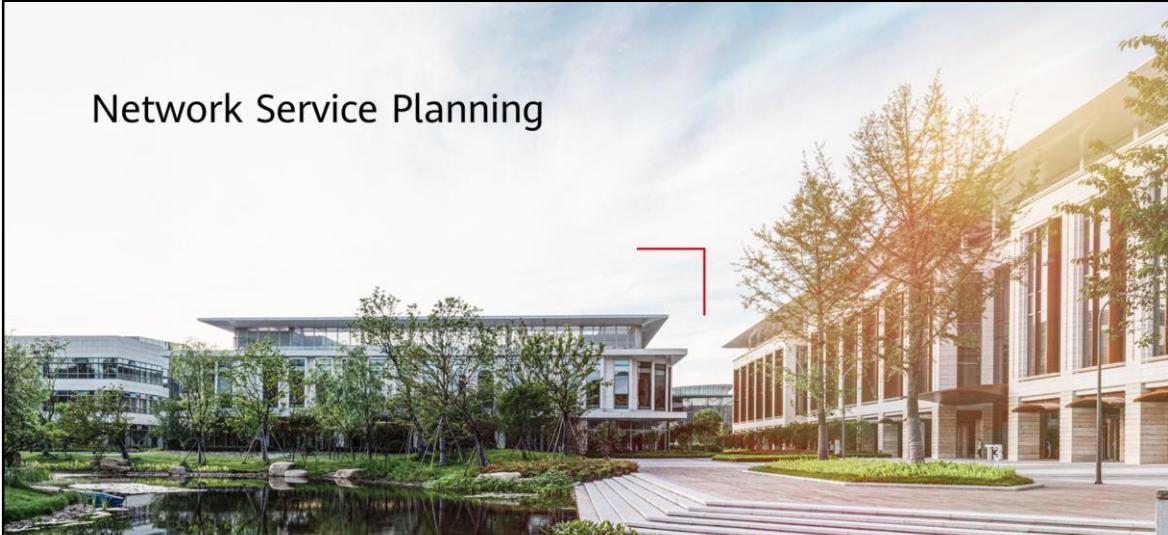
Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

Copyright©2022 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.



Network Service Planning



Foreword

- Network resources are essential to the development of the ICT infrastructure. They are necessary for devices and systems to communicate with each other.
- This course describes the network services provided by Huawei Cloud.

Objectives

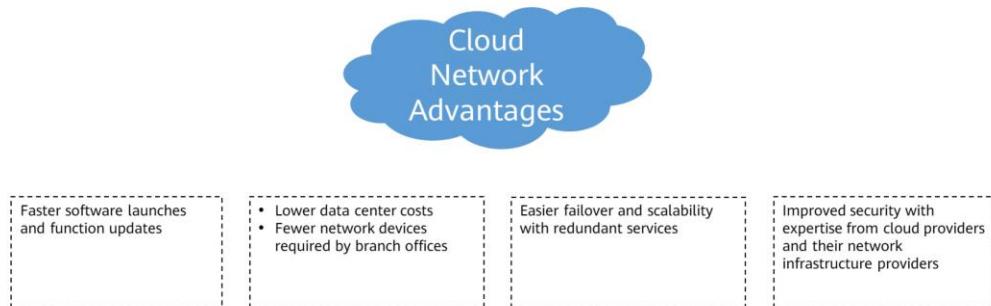
- Upon completion of this course, you will be able to:
 - Understand Huawei Cloud network services well enough to design a secure and reliable network architecture based on service scenarios and that minimizes costs.
 - Understand the principles and application scenarios of network services, and use network services together with other services.

Contents

1. Network Service Overview
2. Network Planning
3. Network Access

What Is Cloud Network?

- Cloud networks provide services for popular distributed architectures.



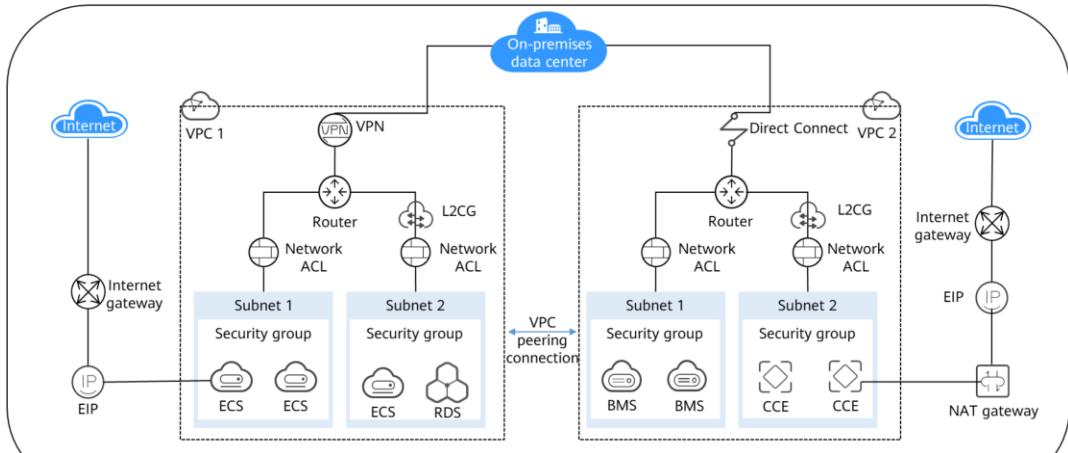
- Network functions and resources are hosted on public or private cloud platforms, managed by the platforms or by service providers, and provided on demand.
- Users and applications with high mobility require the flexibility and scale of cloud networks for assured performance, security, and easier management.
- Cloud networks also improve IT efficiency and save money for offices, schools, home office, healthcare, and public spaces.

Comparison Between Cloud Networks and On-premises Networks

- Cloud networks have great advantages over on-premises networks in terms of scale, IT personnel skills, security, and software versions.

Item	On-Premises Network	Cloud Network
Scale	Hardware procurement, rack space, cooling conditions, and power supply are required.	Only licenses for access points, switches, and gateways are required.
IT resources and skills	Hardware devices need to be maintained, and IT personnel need to be trained.	IT personnel can focus on company's business.
Security	External access permissions, firewall rules, administrator role access permissions, and software maintenance are required.	External access permissions are required, but cloud service providers and infrastructure providers have strict security methods and patch software as needed.
Software flexibility	Program updates and fixes must be downloaded and are limited by fixed version release cycles.	Software can be updated at any time as required, and functions can be added without affecting other services or release cycles.

Huawei Cloud Network Services



6 Huawei Confidential



- There are the following types of network services:
 - Cloud networks:
 - General networks and security policies: VPCs, security groups, and network ACLs
 - Communications within a given region on the cloud: VPC Endpoint and VPC Peering
 - Cross-region communications on the cloud: Direct Connect, Cloud Connect, and VPN
- Cloud network access: EIP, NAT Gateway, ELB, and DNS

Contents

1. Network Service Overview

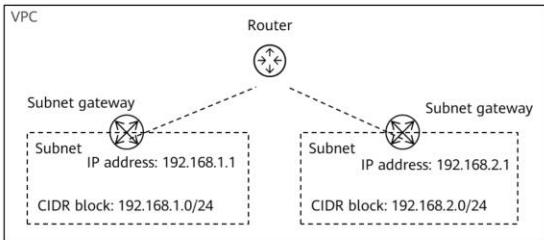
2. Network Planning

- VPCs
 - Network Security
 - Network Connectivity

3. Network Access

Virtual Private Cloud (VPC)

- The VPC service enables you to provision logically isolated private virtual networks for cloud resources, such as ECSs, containers, and databases. You can flexibly manage your cloud networks, including customizing subnets. You can connect your VPCs to an on-premises data center with Direct Connect or Virtual Private Network (VPN).



Easy Connectivity

Add ECSs from different availability zones to the same VPC, and control communication between VPCs in the same region with VPC peering and custom routes.

Secure and Reliable

Get an isolated network for your resources on the cloud. Control traffic between instances and subnets.

High-Speed Bandwidth

Choose dynamic BGP or static BGP bandwidth as needed.

Seamless Scaling

Extend your on-premises data center to Huawei Cloud.

Network Planning Principles

- When you deploy workloads on the cloud, you need to consider network isolation, scalability, and connectivity.

Isolation

Isolation is a basic requirement for network planning. By default, VPCs are isolated from each other and cannot communicate with each other over a private network even if they are in the same region.

Principle

- Isolate different services and departments from each other.
- Create different subnets in a VPC for workloads with different requirements and configure network ACLs for the subnets to control traffic between them.

Scalability

Network scalability needs to be considered as workloads constantly change over time.

Principle

- Reserve sufficient IP addresses for capacity expansion.
- Consider the connectivity between VPCs across regions.

Connectivity

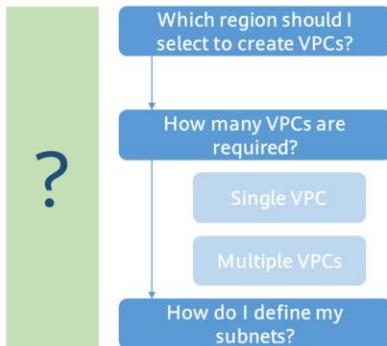
Network connectivity is closely related to network isolation and scalability.

Principle

- If internet access is required, use an EIP.
- If VPCs in the same region need to communicate, use VPC peering connections or VPC endpoints. For VPCs in different regions, use Cloud Connect or VPN.
- If VPCs need to communicate with on-premises data centers, use VPN or Direct Connect.

VPC Network Planning

- When you plan VPCs:



Which region should I select to create VPCs?

- Select the region closest to your target users. VPCs are region-specific. By default, VPCs cannot communicate with each other over a private network even if they are in the same region.

How many VPCs are required?

- If there is only one service, one VPC is enough. If services need to be isolated, create multiple VPCs.
- If an enterprise deploys services in different environments (development, test, and production), create multiple VPCs.
- If there are different permissions management requirements on resources, allocate the resources to different VPCs.

How do I define my subnets?

- Reserve sufficient IP addresses for workload expansion.
- Avoid IP address conflicts if you need to connect a VPC to an on-premises data center or connect two VPCs.

- IP address planning:

- Ensure that the VPC CIDR block does not overlap with the enterprise private network. If there are multiple VPCs in different regions, the VPC CIDR blocks cannot overlap.
- Select a VPC CIDR block based on expected service growth.
- Do not allocate all subnets and IP addresses at once. You should reserve space for future capacity expansion.

- Private CIDR blocks:

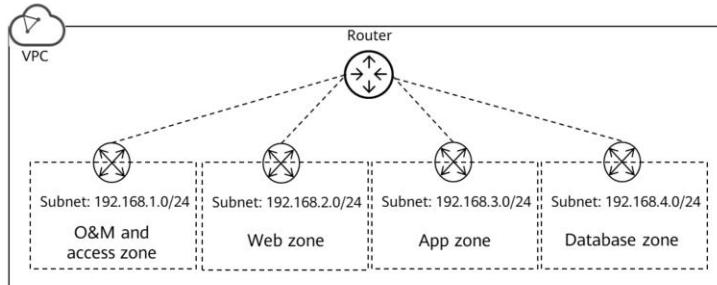
- Select private CIDR blocks for VPCs and subnets, which are used for private communications. If a public CIDR block is configured, conflicts may occur during internet access.
- 10.0.0.0-10.255.255.255 (10/8 prefix)
- 172.16.0.0-172.31.255.255 (172.16/12 prefix)
- 192.168.0.0-192.168.255.255 (192.168/16 prefix)

Case: Single VPC

- If your services do not require network isolation, a single VPC should be enough.

Example: A user wants to set up a website that does not need to communicate with other networks.

- Deploy all servers in one VPC.
- Create separate subnets for web, application, and database zones.
- Create a subnet for the O&M and access zone, which is used to deploy bastion hosts or management and authentication devices, facilitating remote access, service deployment, and O&M.
- Configure network ACLs to control traffic between subnets.

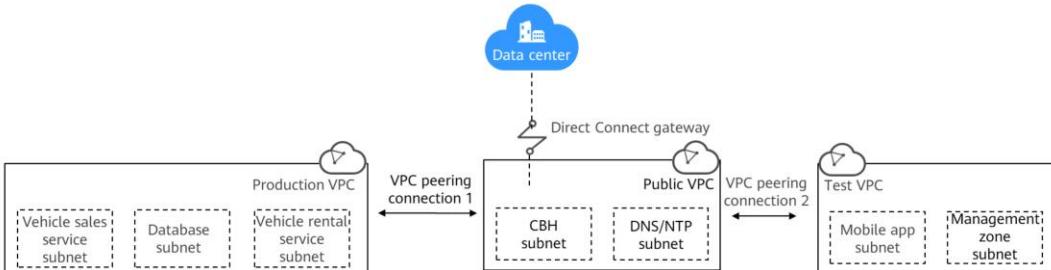


Case: Multiple VPCs

- If you have multiple service systems in a region and each service system requires an isolated network, you can create a separate VPC for each service system. You can use a VPC peering connection to enable communications between different VPCs.

Example: After services of a large automobile enterprise are migrated to the cloud, all services share the same Direct Connect connection to connect to the data center.

- Create separate VPCs for different services or departments.
- Use VPC peering connections to connect all VPCs to the public VPC.



Contents

1. Network Service Overview

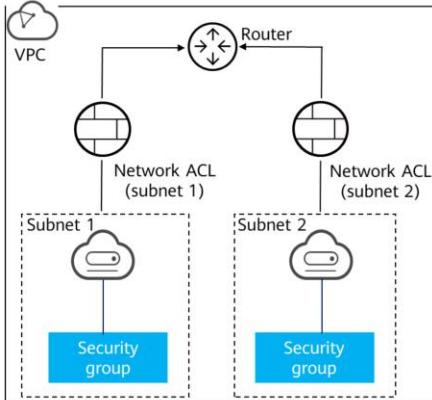
2. Network Planning

- VPCs
- **Network Security**
- Network Connectivity

3. Network Access

Security Groups and Network ACLs

- VPCs are logically isolated from each other using tunneling technology. By default, different VPCs cannot communicate with each other. Network ACLs protect subnets, and security groups protect ECSs. They work together to make your network more secure.



14 Huawei Confidential



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 within a VPC. You can define different access control rules for a security group, and these rules are then applied to all the ECSs added to this security group.

Network ACL

- A network ACL allows you to create rules to control traffic in and out of its associated subnets.

- Similar to security group rules, network ACL rules are used to determine whether data packets can enter or leave a subnet.

Differences Between Security Groups and Network ACLs

- You can configure security groups and network ACLs to protect ECSs in your VPCs. Security groups operate at the ECS level, and network ACLs operate at the subnet level.

Item	Security Group	Network ACL
Protected object	ECSs	Subnets
Action	Only supports Allow rules. (Deny rules are supported in certain regions.)	Supports both Allow and Deny rules.
Priority	If there are conflicting rules, the first security group associated will take precedence over those associated later, then the rule with the highest priority in that security group will be applied first.	If there are conflicting rules, only the rule with the highest priority takes effect.
Packet filtering	Only supports packet filtering based on 3-tuple (protocol, port, and destination IP address).	Only supports packet filtering based on 5-tuple (protocol, source port, destination port, source IP address, and destination IP address).
Application operation	By default, a security group must be selected during ECS creation and the security group will be automatically applied to the ECS.	You cannot select a network ACL when creating a subnet. You must create a network ACL, associate subnets with the network ACL, add inbound and outbound rules, and enable the network ACL. Then, the network ACL takes effect for the associated subnets and ECSs in the subnets.

Security Group and Network ACL Configuration Principles

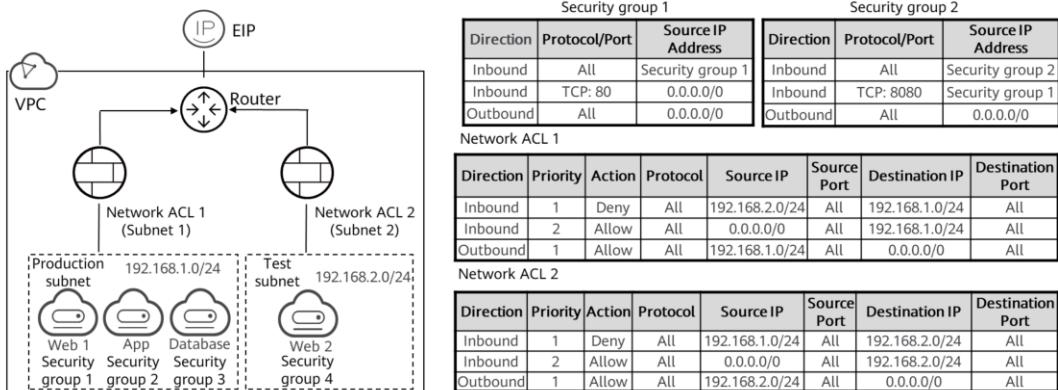
- When adding inbound and outbound rules for network security, you have to comply with the configuration principles of security groups and network ACLs.

	Security Group		Network ACL
Inbound	<ul style="list-style-type: none">Create security groups based on different resources, such as web servers and databases.Add a rule only if it is necessary and always following the principle of least privilege. Each rule allows traffic only on specific IP address range and ports. Do not allow traffic on any IP addresses and ports.Set the source to an IP address range or a specific IP address.Restrict traffic on certain ports.	Inbound	<ul style="list-style-type: none">Allow subnets in the same VPC to communicate with each other.Allow communications between VPCs only on specific IP address ranges.Control communications between subnets, such as load balancer subnet and data subnet.
Outbound	By default, all outbound traffic is allowed.	Outbound	By default, all outbound traffic is allowed.

Case: Configuring Security Groups and Network ACLs

An internet company deployed its services on the cloud and has the following requirements:

- The web server in the production subnet needs to provide services accessible from the internet through EIPs on port 80.
- The web server in production subnet needs to access the App server on port 8080.
- Production subnet and test subnet need to be isolated from each other.



17 Huawei Confidential



- Security group 1: The first rule allows Web 1 server to communicate with other web servers that may be added later for capacity expansion. The second rule allows internet access to Web 1 server. The third rule allows all outbound traffic.
- Security group 2: The first rule allows the App server to communicate with other App servers that may be added later for capacity expansion. The second rule allows Web 1 server to access the App server. The third rule allows all outbound traffic.
- Network ACL 1: The first rule denies the access from the test subnet. The second rule allows all inbound access, excepting the access from the test subnet denied by the first rule. The third rule allows all outbound traffic.
- Network ACL 2: The first rule denies the access from the production subnet. The second rule allows all inbound access, excepting the access from the production subnet denied by the first rule. The third rule allows all outbound traffic.

Contents

1. Network Service Overview

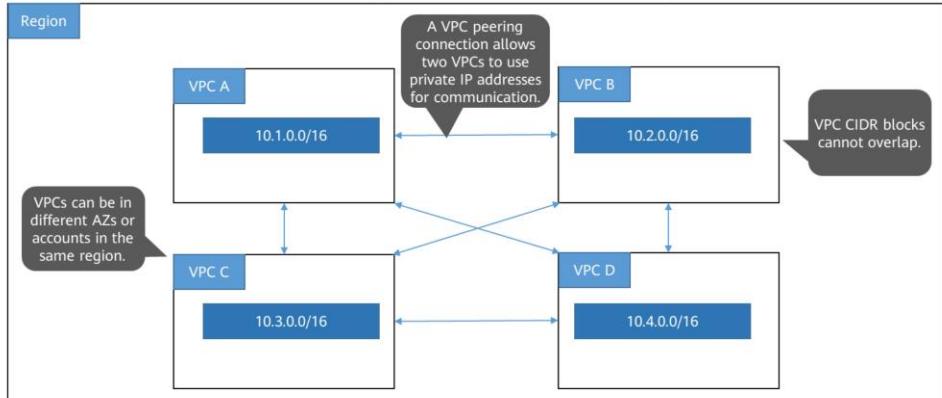
2. Network Planning

- VPCs
- Network Security
- **Network Connectivity**

3. Network Access

VPC Peering

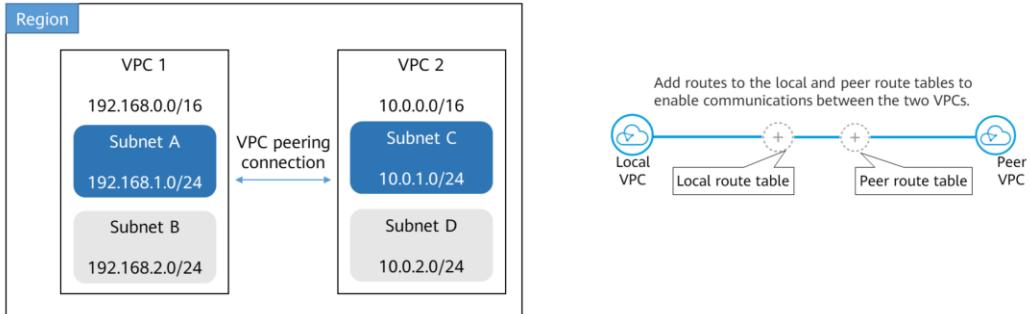
- A VPC peering connection is a network connection between two VPCs in the same region. It enables you to route traffic between them using private IP addresses as if they were in the same VPC.



- If two VPCs connected by a VPC peering connection overlap with each other, there will be route conflicts and the VPC peering connection may not be usable.
- If two VPCs connected by a VPC peering connection have overlapping CIDR blocks, the connection can only enable communications between non-overlapping subnets in the VPCs. If subnets in the two VPCs of a VPC peering connection overlap with each other, the connection may not take effect, so ensure that there are no overlapping subnets.
- If there are three VPCs, A, B, and C, and VPC A is peered with both VPC B and VPC C, but VPC B and VPC C overlap with each other, you cannot configure routes with the same destinations for VPC A.
- You cannot have more than one VPC peering connection between the same two VPCs at the same time.
- VPC peering does not support transitive peering relationships. For example, if VPC A is connected to both VPC B and VPC C, but VPC B and VPC C are not connected, VPC B and VPC C cannot communicate with each other through VPC A. You need to create a VPC peering connection between VPC B and VPC C.
- A VPC peering connection between VPCs in different regions will not be usable.

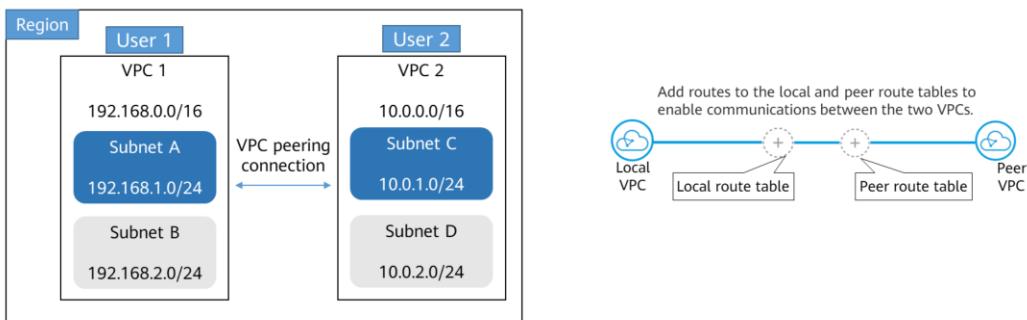
Creating a VPC Peering Connection Between VPCs of the Same Account

- You can request a VPC peering connection with another VPC of your account, but the two VPCs must be in the same region. The system automatically accepts the request.



Creating a VPC Peering Connection Between VPCs from Different Accounts

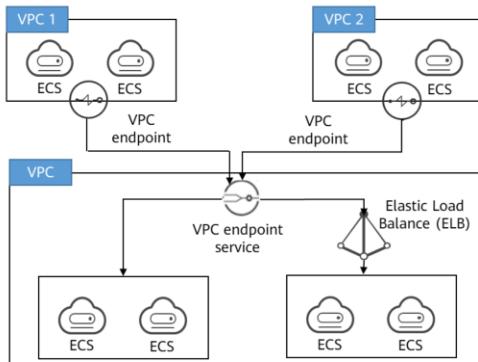
- You can request a VPC peering connection with a VPC of another account, but the two VPCs must be in the same region. Also, the connection is not valid until the peer account accepts the request.



- If you request a VPC peering connection with a VPC of another account, the connection takes effect only after the peer account accept the request. If you request a VPC peering connection with a VPC of your own, the system automatically accepts the request and activates the connection.

VPC Endpoint (VPCEP)

- VPCEP is a cloud service that provides secure and private channels to connect your VPCs to VPC endpoint services, including cloud services or your private services. It allows you to plan networks flexibly without having to use EIPs.
- VPCEP provides two types of resources: VPC endpoint services and VPC endpoints.



Excellent performance

Each gateway supports up to 1 million concurrent connections across a variety of use cases.

Instant availability

VPC endpoints take effect a few seconds after they are created.

Easy to use

You can use VPC endpoints to access resources over private networks, without having to use EIPs.

High security

VPC endpoints enable you to access VPC endpoint services without exposing server information, helping you minimize risks.

VPCEP provides two types of resources: VPC endpoint services and VPC endpoints.

- VPC endpoint services refer to cloud services or your private services that can be configured in VPCEP to provide services to users. For example, you can create an application in a VPC and configure it as a VPC endpoint service that VPCEP supports.
- VPC endpoints are channels for connecting VPCs to VPC endpoint services. You can create an application in your VPC and configure it as a VPC endpoint service. A VPC endpoint can be created in another VPC in the same region and then used as a channel to access the VPC endpoint service.

Differences Between VPC Peering Connections and VPC Endpoints

- A VPC peering connection differs greatly from a VPC endpoint in terms of security, communications, route configurations, and other aspects.

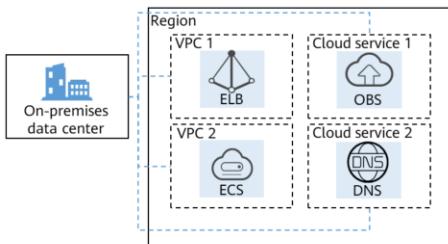
Item	VPC Peering Connection	VPC Endpoint
Security	All resources in a VPC, such as ECSs and load balancers, can be accessed.	Allows access to a specific service or application. Only the ECSs and load balancers in the VPC for which VPC endpoint services are created can be accessed.
CIDR block overlap	Not supported	Supported
Communications	VPCs connected through a peering connection can communicate with each other.	Requests can only be initiated from a VPC endpoint to a VPC endpoint service, but not the other way around.
Routing configurations	If a peering connection is established between two VPCs, add routes to the VPCs so that they can communicate with each other.	For two VPCs that are connected through a VPC endpoint, the route has been configured, and you do not need to configure it again.
Access using Virtual Private Network (VPN)/Direct Connect	Supported	Supported

- Function:
 - A VPC peering connection enables traffic between two VPCs so that instances in the subnets of the two VPCs can communicate with each other as if they were in the same network.
- Access scenario:
 - VPC peering connections, in most cases, are used to connect subnets of two VPCs belong to the same tenant.
 - VPCEP makes services available to tenants on a cloud platform.

Application Scenarios

Service Now

- After an enterprise migrates some of its workloads to the cloud through Direct Connect or VPN, its on-premises data center maintains a complex hybrid cloud architecture for a long time. Some production and testing workloads are running on on-premises data centers, and some production and testing workloads are running on Huawei Cloud or other cloud platforms. Therefore, the on-premises data center often needs an intranet to access cloud services. Unfortunately if only Direct Connect or VPN are used, many cloud resources and services will remain inaccessible.
- Enterprise requirements: The on-premises data center needs to access resources (such as ECSs) and other cloud services (such as OBS) in VPC 1 and VPC 2 (shown below) without using the public network.

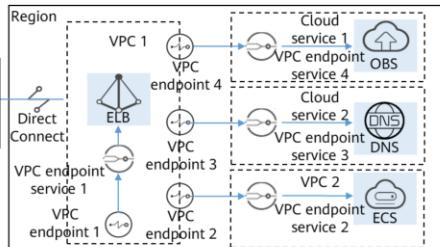


24 Huawei Confidential

Solution

To enable the user's on-premises data center to communicate with the cloud environment in a secure and high-speed manner, a combination of Direct Connect and VPCEP are used:

- Direct Connect establishes a high-speed, low-latency, stable, and secure dedicated network connection between the user's on-premises data center and a Huawei Cloud VPC.
- VPCEP uses the internal network of Huawei Cloud and does not require an EIP.

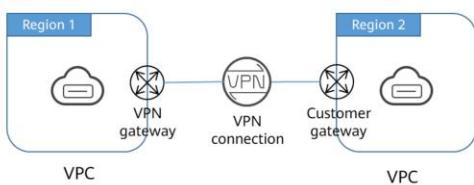


HUAWEI

- Direct Connect enables communication between the on-premises data center and VPC 1.
- With VPC endpoint 1, the user's on-premises data center can access ELB in VPC 1.
- With VPC endpoint 2, the user's on-premises data center can access Elastic Cloud Servers (ECSs) in VPC 2.
- With VPC endpoint 3, the user's on-premises data center can access Domain Name Service (DNS) over the intranet.
- With VPC endpoint 4, the user's on-premises data center can access Object Storage Service (OBS) over the intranet.

Virtual Private Network (VPN)

- VPN establishes a secure encrypted communications tunnel, in compliance with industry standards, between a local data center and a VPC or between two VPCs on the cloud.



High data security

Data is encrypted using Internet Key Exchange (IKE) and Internet Protocol Security (IPsec) technologies for secure and reliable transmission.

High availability

Active-active gateways can be deployed and dynamic routing is supported to achieve failover in seconds.

Cost-effectiveness

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

Easy to use

A VPN connection can be created in a few simple steps and is ready to use immediately after being created.

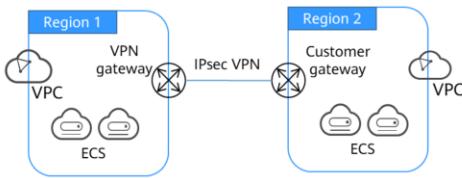


- High data security
 - Huawei hardware uses IKE and IPsec to encrypt data to provide carrier-class reliability and ensure a stable VPN connection.
- Seamless scale-out
 - With VPN, you can connect your local data center to your VPC and quickly extend services at the data center to the cloud, thereby forming a hybrid cloud architecture.

Application Scenarios

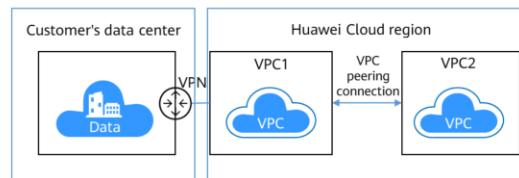
Interconnection between VPCs in different regions

- With VPN, you can connect VPCs in different regions on Huawei Cloud to enable the flow of user data and ensure always-on user services in these regions.



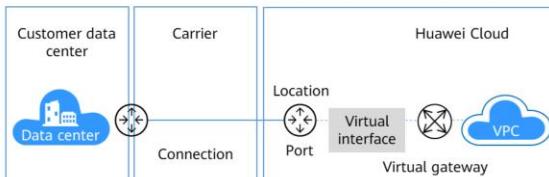
Communication between VPCs in the same region and a local data center

- Two VPCs are created in the same region on Huawei Cloud. The local data center is connected to one of the VPCs through VPN.
- A VPC peering connection is created between the two VPCs. The local data center can then communicate with both of the VPCs.



Direct Connect

- Direct Connect establishes a dedicated high-speed network connection with low latency, excellent stability, and robust security between your on-premises data center and the cloud. Direct Connect allows you to maximize legacy IT facilities and leverage cloud services to build a flexible, scalable hybrid cloud compute environment.



The key components of Direct Connect are a connection, virtual gateway, and virtual interface.

- The connection is a dedicated network connection between your premises and a Direct Connect location over a line you lease from a carrier.
- A virtual gateway is a logical gateway for accessing VPCs. A virtual gateway can be associated with the VPCs that you need to access.
- The virtual interface links a connection with one or more virtual gateways, each of which is associated with a VPC, so that your on-premises network can access all these VPCs.

Impenetrable security

A dedicated channel between your on-premises data center and one or more VPCs means airtight security.

Low latency

A dedicated network is used for data transmission, which ensures high network performance, low latency, and excellent user experience.

High bandwidth

A single connection supports up to 100 Gbit/s of bandwidth, which meets a diverse range of bandwidth requirements.

Seamless expansion

Connecting to the cloud allows for virtually unlimited cloud resources, meaning flexible and scalable hybrid deployment, so you can focus on what is important: growing your business.

- The connection is a dedicated network connection between your premises and a Direct Connect location over a line you lease from a carrier. You can create a standard connection by yourself or request a hosted connection from a partner. After you are certified as a partner, you can also create an operations connection.
 - A standard or operations connection has a dedicated port for your exclusive use and can be associated with multiple virtual interfaces.
 - A hosted connection allows you to share a port with others. Partners with operations connections can provision hosted connections and allocate VLANs and bandwidths for those connections. Only one virtual interface can be created for each hosted connection.
- The virtual gateway is a logical gateway for accessing VPCs. Each VPC can have only one virtual gateway associated, but multiple connections can use the same virtual gateway to access one VPC.
- The virtual interface links a connection with one or more virtual gateways, each of which is associated with a VPC, so that your on-premises network can communicate with all these VPCs.

Comparison Between VPN and Direct Connect

- Both VPN and Direct Connect are available for on-premises and cloud network communication. They are used in different scenarios.

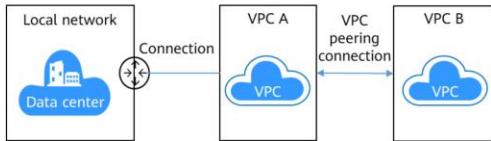
Item	VPN	Direct Connect
Accessing VPCs	Supported	Supported
Security	VPN < Direct Connect	
Access channel	Public network	Private network
Pricing	VPN < Direct Connect	
Bandwidth	Determined by the public network.	Determined by the connection capability of Direct Connect itself.
Latency	VPN > Direct Connect (common scenarios)	
Service provisioning	Be effective immediately	The duration depends on the construction speed of the carrier.

- VPN
 - IPsec VPN safeguards data transfer.
 - Ease of use and instant availability
- Direct Connect
 - Highly private with dedicated connections linking on-premises and cloud networks
 - Low and stable latency, low jitter level, and excellent performance

Application Scenarios

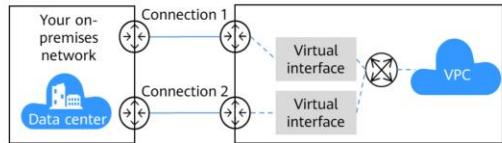
Communication between an on-premises data center and VPCs in the same region on the cloud

- You can use VPC Peering to connect the VPC your on-premises data center is accessing to other VPCs in the same region so that your on-premises data center can access all these VPCs.



Accessing a VPC over two connections that use BGP routing

- Connect your on-premises network to the cloud over two connections that are terminated at two locations in the same region and use BGP routing so that your on-premises network can access the VPC.

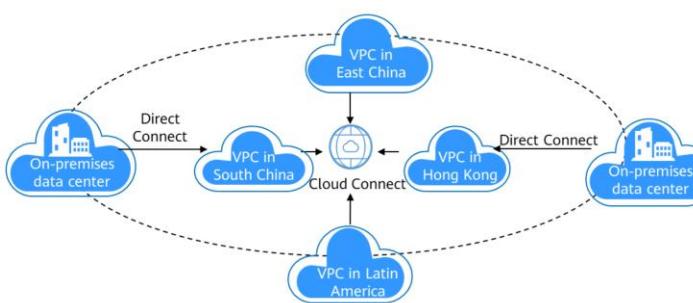


Prerequisites:

- Single-mode 1 GE, 10 GE, 40 GE, or 100GE optical modules must be used to connect to Huawei Cloud access devices.
- Auto-negotiation for the port has been disabled. The port speed and full-duplex mode have been manually configured.
- 802.1Q VLAN encapsulation is supported on your on-premises network.
- Your device supports Border Gateway Protocol (BGP) and does not use Autonomous System Number (ASN) 64512, which is used by Huawei Cloud.

Cloud Connect

- Cloud Connect allows you to connect VPCs in different regions to allow instances in these VPCs to communicate over a private network as if they were within the same network.



Full connectivity

You can connect VPCs in any region to build a multi-VPC network without using additional links. Over 10 regions are currently supported, with support for more regions coming soon.

Ease of use

In just four simple steps, you can build cross-region VPC connectivity to securely connect and use cloud resources in multiple VPCs.

High performance

Cloud Connect leverages Huawei's global network infrastructure to securely transmit data through the shortest network path possible for ultra-low latency. You can flexibly adjust bandwidth to meet your business requirements.

Globally compliant

Cloud Connect complies with laws and regulations worldwide, allowing you to focus on innovation and build business success.



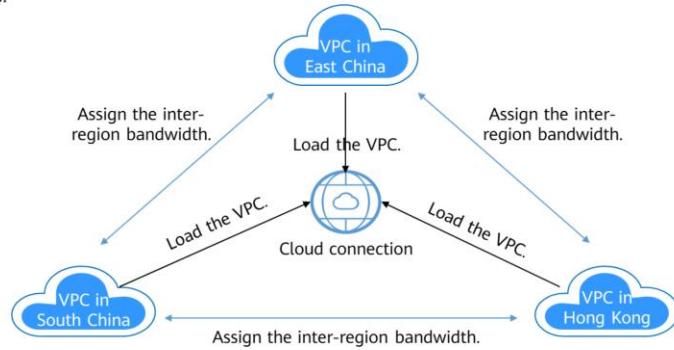
Constraints:

- A cloud connection cannot be created between VPCs that have overlapping CIDR blocks, or network communications will fail. In addition, IP addresses of network instances that will be loaded to a cloud connection cannot overlap.
- If you load a VPC to a cloud connection created using the same account, you cannot enter loopback addresses, multicast addresses, or broadcast addresses for the custom CIDR block.
- If a NAT gateway has been created for any VPC you have loaded to a cloud connection, a custom CIDR block needs to be added and set to 0.0.0.0/0.

Application Scenarios

Communications among VPCs in different regions

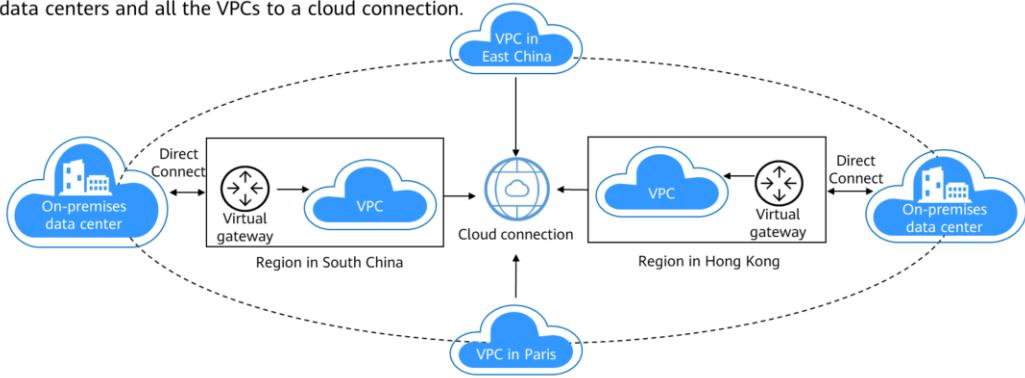
Cloud Connect helps you establish secure and reliable private network communications among VPCs in different regions.



Application Scenarios

Connecting on-premises data centers to VPCs in different regions

If you want to establish connectivity between multiple on-premises data centers and VPCs in different regions, you can use Direct Connect to connect the data centers to a VPC, and then load the virtual gateways configured for the data centers and all the VPCs to a cloud connection.

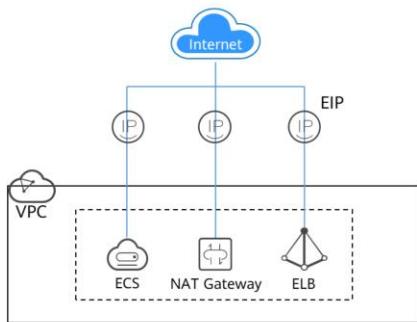


Contents

1. Network Service Overview
2. Network Planning
- 3. Network Access**
 - Elastic IP (EIP)
 - Elastic Load Balance (ELB)
 - NAT Gateway
 - Domain Name Service (DNS)

Elastic IP (EIP)

- The Elastic IP service provides static public IP addresses and scalable bandwidths that enable your cloud resources to communicate with the internet.



Ease of Use

You can easily bind an EIP to or unbind it from an ECS, BMS, virtual IP address, NAT gateway, or load balancer. You can also dynamically scale the bandwidth included in the EIP subscription to meet changing requirements.

Flexible Billing

EIPs are available on a pay-per-use (by bandwidth or traffic) and yearly/monthly basis.

Shared Bandwidth

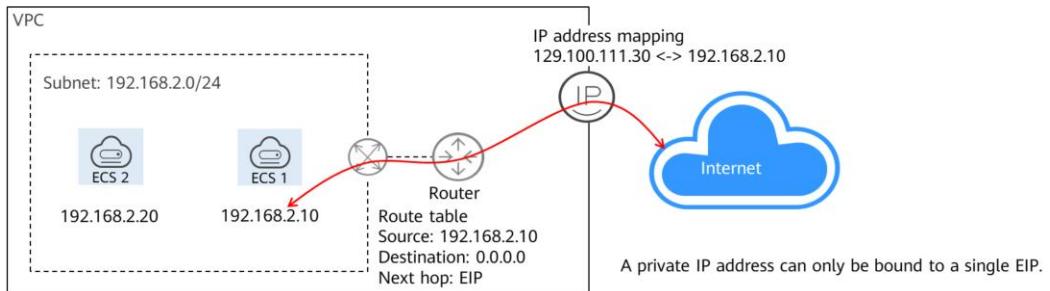
You can configure multiple EIPs to share the same bandwidth and thereby reduce costs.

Real-Time Adjustments

All EIP binding or unbinding or bandwidth adjustments you make are effective immediately.

How EIP Works

- To enable ECS 1 to access the internet, an EIP (129.100.111.30) is purchased and bound to the private IP address (192.168.2.10) configured for the ECS NIC. The system automatically delivers a route (source: 192.168.2.10, destination: 0.0.0.0, next hop: EIP). In addition, the EIP maintains a mapping with the private IP address.



Static BGP, Dynamic BGP, and Premium BGP

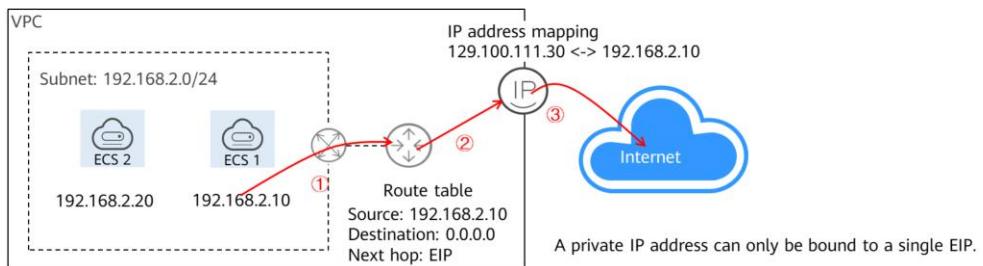
- There are different types of Border Gateway Protocols (BGP): static BGP, dynamic BGP, and premium BGP. The following table compares their differences.

Comparison Dimension	Static BGP	Dynamic BGP	Premium BGP
Definition	Static routes are configured manually and must be manually reconfigured anytime the network topology or link status changes.	Dynamic BGP provides automatic failover and the best path is chosen based on real-time network conditions and preset policies.	Premium BGP chooses the best path and ensures low-latency and high-quality networks. BGP is used to interconnect with lines of multiple mainstream carriers. Public network connections that feature low latency and high quality are directly established between CN-Hong Kong and Chinese mainland regions. Premium BGP is available only in CN-Hong Kong.
Assurance	When changes occur on a network that uses static BGP, the manual configuration takes some time and high availability cannot be guaranteed.	When a fault occurs on a carrier's link, dynamic BGP will quickly select another path to take over services, ensuring service availability.	Premium BGP has the same assurance capability as that of dynamic BGP. In addition, premium BGP ensures higher network quality and lower latency. Currently, mainstream carriers in Hong Kong (China) are supported.
Service availability	99%	99.95%	
Billing	Their price from least to most expensive: static BGP, dynamic BGP, and premium BGP.		

- Dynamic BGP:
 - Dynamic BGP provides automatic failover and chooses the best path based on real-time network conditions and preset policies.
- Static BGP:
 - Static routes are configured manually and must be manually reconfigured anytime the network topology or link status changes.
- Comparison in assurance:
- Dynamic BGP:
 - When a fault occurs on a carrier's link, dynamic BGP will quickly select another path to take over services, ensuring service availability.
 - Currently, carriers in China that support dynamic BGP routing include China Telecom, China Mobile, China Unicom, China Education and Research Network (CERNET), National Radio and Television Administration, and Dr. Peng Group.
- Static BGP:
 - When changes occur on a network that uses static BGP, the manual configuration takes some time and high availability cannot be guaranteed.

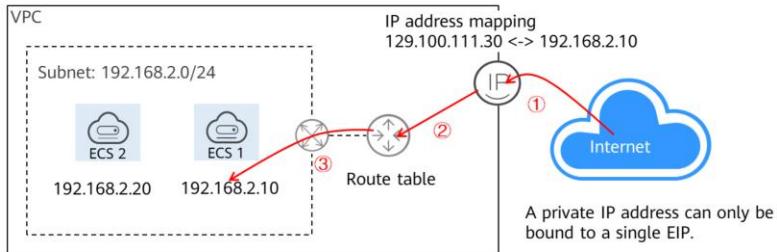
Accessing an ECS from the Internet

- When an ECS needs to access the internet:
 - If the destination IP address is not in the VPC of the ECS, the VPC agent searches the route table.
 - If there is no route that matches the destination IP address, the default route (0.0.0.0) is used. The packet is sent to the next hop (the EIP).
 - The EIP queries the IP address mapping and translates the source IP address from 192.168.2.10 to 129.100.111.30 and sends the packet to the internet.



Accessing from the Internet to an ECS

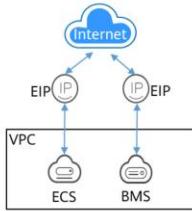
- To access an ECS from the public IP address 129.100.111.30:
 - After receiving a packet from the public network, the EIP queries the IP address mapping and translates the destination IP address of the packet from 129.100.111.30 to 192.168.2.10.
 - The EIP queries the route table of the VPC.
 - The route table forwards the packet to the corresponding ECS based on the destination IP address.



Application Scenarios

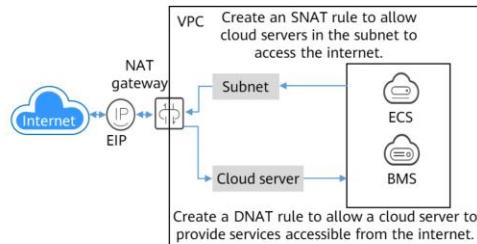
Binding EIPs to cloud servers

You can allow cloud servers to communicate with the internet.



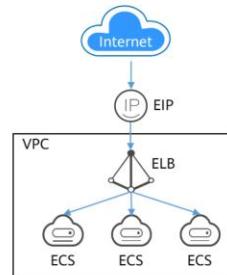
Binding an EIP to a NAT gateway

Multiple cloud servers (such as ECSS, BMSs, and desktops) can share an EIP to access the internet or provide services accessible from the internet.



Binding an EIP to a load balancer

You can route requests from clients to backend servers over the internet.



Contents

1. Network Service Overview

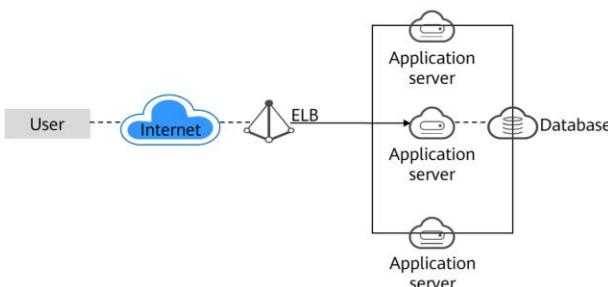
2. Network Planning

3. Network Access

- Elastic IP (EIP)
- **Elastic Load Balance (ELB)**
- NAT Gateway
- Domain Name Service (DNS)

Elastic Load Balance (ELB)

- ELB automatically distributes incoming traffic across multiple backend servers based on forwarding rules you configure, improving the service capabilities and fault tolerance of your applications.



High availability

Dedicated load balancers ensure service continuity. If servers in an AZ are unhealthy, load balancers automatically route traffic to healthy servers in other AZs.

Flexible scaling

ELB works with Auto Scaling to flexibly adjust the number of backend servers and intelligently distribute incoming traffic across them.

Robust performance

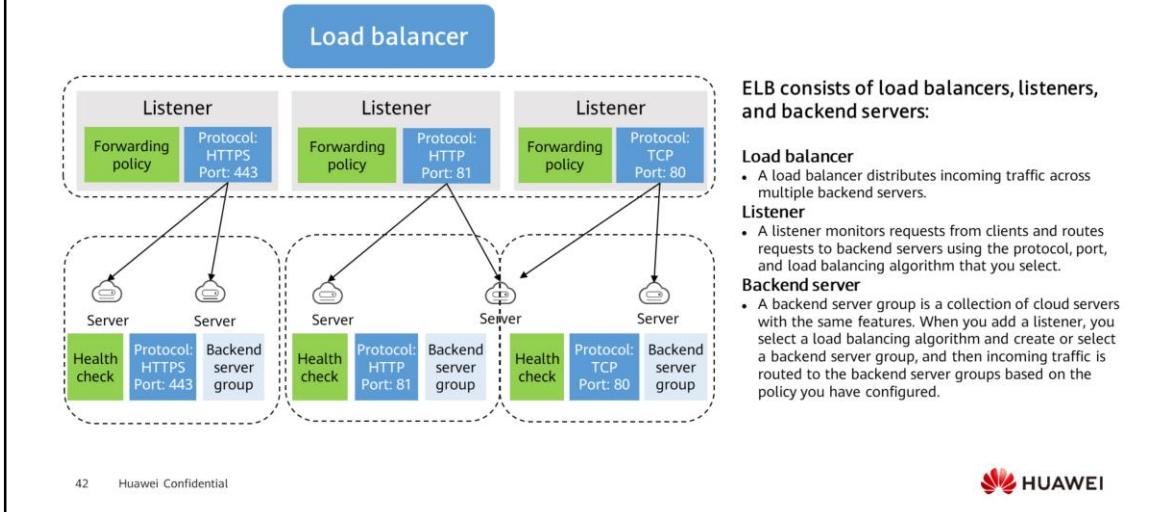
Each dedicated load balancer has exclusive use of resources and can handle up to tens of millions of concurrent connections.

Easy to use

ELB provides a diverse set of algorithms that allow you to configure custom traffic routing policies to meet your requirements while keeping the configuration simple.

- Dedicated load balancers give you exclusive access to their resources, so the performance of a dedicated load balancer is not affected by other load balancers. In addition, there are a wide range of specifications available for selection.
- Shared load balancers are deployed in clusters, where all the load balancers share resources. With a shared load balancer, the performance of one load balancer can be affected by other load balancers.

ELB Components



- ELB periodically sends heartbeat messages to associated backend servers to check their health to ensure that traffic is distributed only to healthy backend servers. This can improve the availability of applications.

Sticky Session

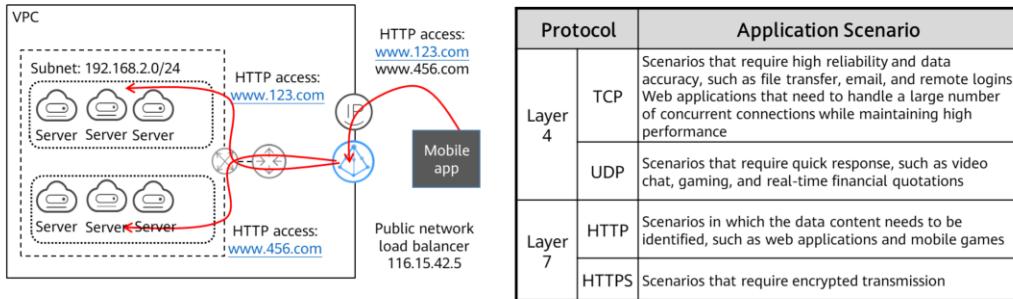
- Requests in a session may be stateful and depend on earlier requests. How can this problem be solved?
 - Sticky sessions ensure that all requests from the same client are routed to the same backend server for as long as the session lasts.
 - At Layer 4, the source IP address is used for maintaining sessions. At Layer 7, load balancer cookies and application cookies are used.

Sticky Session Type	Description
Source IP address	Requests from the same IP address are forwarded to the same backend server.
Load balancer cookie	The load balancer generates a cookie after receiving the first request from a client. All subsequent requests with the same cookie are distributed to the same backend server.
Application cookie	An application deployed on the backend server generates a cookie after receiving the first request from a client. All requests with the same cookie generated by backend applications are routed to the same backend server.

- The maximum stickiness duration at Layer 7 is 24 hours.
- The maximum stickiness duration at Layer 4 is one hour.

Protocol

- What protocols does ELB use to distribute traffic?
 - Protocols at Layer 4 forward packets to backend servers based on packet characteristics, like their IP addresses.
 - Protocols at Layer 7 forward packets to different backend server groups based on HTTP packet characteristics, for example, the URL.



Load Balancing Algorithms

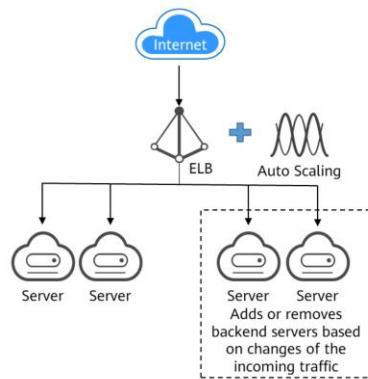
- What load balancing algorithms does ELB use to distribute traffic?
 - The round robin algorithm is suitable for short connections, and the least connections algorithm is good for persistent connections.
 - Weighted round robin and weighted least connections are often used in scenarios where the performance of servers in a backend server group varies.

Load Balancing Algorithm	Weight	Description
Round robin	The value ranges from 0 to 100.	Requests are distributed across backend servers in sequence based on their weights. Backend servers with higher weights receive proportionately more requests, whereas equally-weighted servers receive the same number of requests. This algorithm is typically used for short connections, such as HTTP connections.
Least connections	The value ranges from 0 to 100.	Requests are routed to the server with the lowest connections-to-weight ratio. Building on least connections, the weighted least connections algorithm assigns a weight to each server based on their processing performance. This algorithm is often used for persistent connections, such as connections to a database.
Source IP hash	Weights do not take effect even if they are not 0.	The source IP address of each request is calculated using a hash algorithm to obtain a unique hash key, and all backend servers are numbered. The generated key allocates the client to a particular server. This allows requests from different clients to be routed based on source IP addresses and ensures that requests from the same client are forwarded to the same server. This algorithm applies to TCP connections without cookies.

Application Scenarios

Applications with predictable peaks and troughs in traffic

For an application that has predictable peaks and troughs in traffic volumes, ELB works with Auto Scaling to add or remove backend servers to keep up with the changing demand. ELB routes requests to the required number of backend servers to handle the load of your application based on the load balancing algorithm and health check you set.

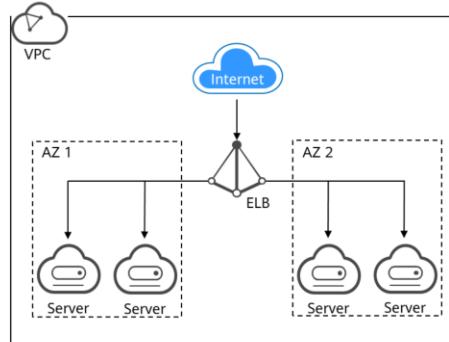


- Applications with predictable peaks and troughs in traffic
 - For an application that has predictable peaks and troughs in traffic volumes, ELB works with Auto Scaling to add or remove backend servers to keep up with the changing demand. ELB routes requests to the required number of backend servers to handle the load of your application based on the load balancing algorithm and health check you set. One example is flash sales, during which application traffic spikes in a short period. ELB can work with Auto Scaling to run only the required number of backend servers, helping to minimize IT costs.

Application Scenarios

Cross-AZ load balancing

For services that require high availability, ELB can distribute traffic across AZs. If an AZ becomes faulty, ELB distributes the traffic to backend servers in other AZs that are running properly.



- Cross-AZ load balancing:
 - For services that require high availability, ELB can distribute traffic across AZs. If an AZ becomes faulty, ELB distributes the traffic to backend servers in other AZs that are running properly.
 - ELB is ideal for banking, policing, and large application systems that require high availability.

Contents

1. Network Service Overview

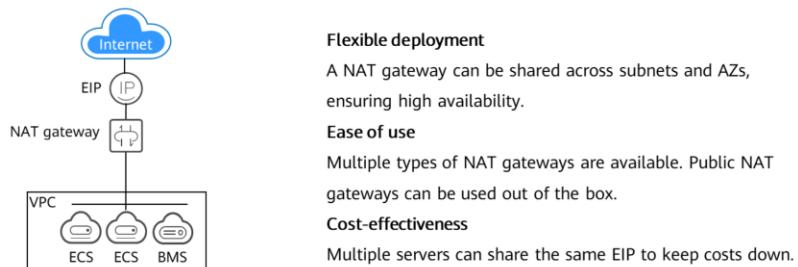
2. Network Planning

3. Network Access

- Elastic IP (EIP)
- Elastic Load Balance (ELB)
- **NAT Gateway**
- Domain Name Service (DNS)

Public NAT Gateways

- Huawei Cloud provides public NAT gateways and private NAT gateways.
- A public NAT gateway can translate private IP addresses into public IP addresses. After the translation, cloud resources can securely access the public network or provide services externally. In addition, private network information is protected from being exposed to the public network.
- A public NAT gateway enables instances in a private subnet to share EIPs to communicate with the Internet. The instances can be ECSs or BMSs in a VPC, or servers in on-premises data centers that connect to a VPC through Direct Connect or VPN. A public NAT gateway supports up to 20 Gbit/s of bandwidth.



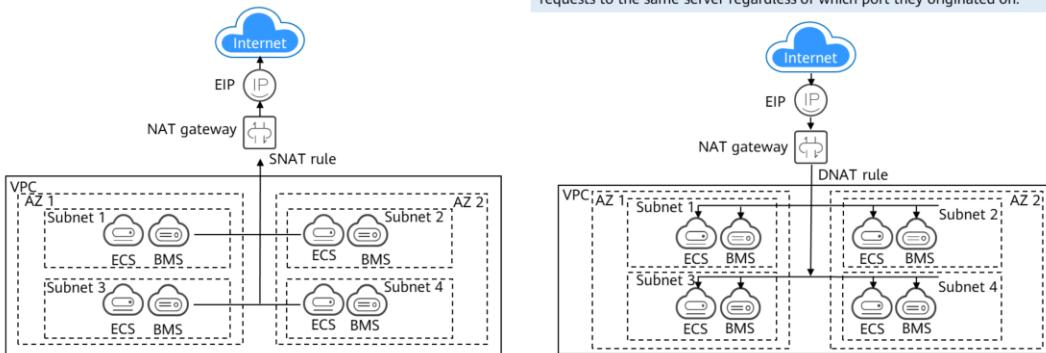
- **Flexible deployment**
 - A public NAT gateway can be shared across subnets and AZs, so that even if an AZ fails, the public NAT gateway can still run normally in another AZ. The type and EIP of a public NAT gateway can be changed at any time.
- **Ease of use**
 - Multiple types of public NAT gateways are available. Public 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**
 - With a public NAT gateway, when you send data through a private IP address or provide services accessible from the Internet, the public NAT gateway translates the private IP address to a public IP address. You no longer need to configure one EIP for each server, which saves money on EIPs and bandwidth.

Public NAT Gateways

- Public NAT gateways support source network address translation (SNAT) and destination network address translation (DNAT).

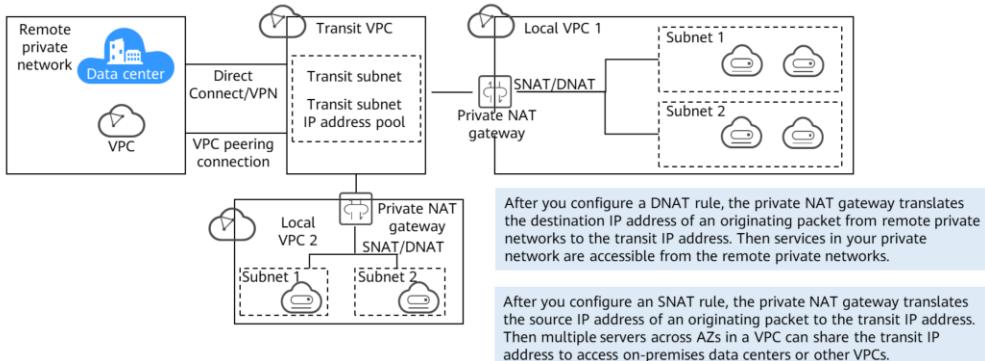
SNAT translates private IP addresses into EIPs, allowing traffic from a private network to go out to the Internet.

DNAT enables multiple servers within an AZ or across multiple AZs in a VPC to share EIPs to provide services accessible from the Internet. With an EIP, a NAT gateway forwards the Internet requests from only a specific port and over a specific protocol to a specific port of a server, or it can forward all requests to the same server regardless of which port they originated on.



Private NAT Gateways

- A private NAT gateway provides private network address translation for ECSs and BMSs in a VPC, allowing them to communicate with servers in other VPCs or on-premises data centers. You can configure SNAT and DNAT rules for the NAT gateway to translate the source and destination IP addresses into a transit IP addresses.



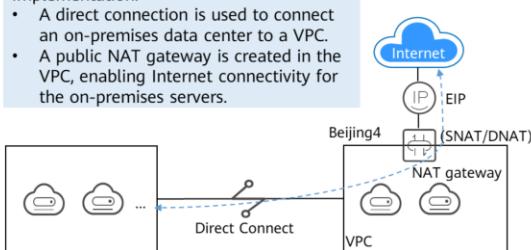
- Transit subnet: A transit subnet is where a transit IP address resides.
- Transit VPC: A transit VPC is where a transit subnet resides.
- Easier network planning
 - The private NAT gateway allows for communication between overlapping CIDR blocks. This frees customers from the time-consuming and stressful network replanning, so that customers can retain their original network while migrating workloads to the cloud.
- Strong security
 - Private NAT gateways help organizations meet industry regulatory requirements by mapping private IP addresses to specified IP addresses for access.
- Easy O&M
 - A private NAT gateway can map the CIDR block of each department to the same VPC CIDR block, which simplifies the management of complex networks.
- Zero conflicts
 - Thanks to IP address mapping, the private NAT gateways allow for communication between overlapping CIDR blocks.

Application Scenarios of Public NAT Gateways

Using a public NAT gateway and direct connection to accelerate Internet access: On-premises servers that connect to a VPC through VPN or Direct Connect need secure, reliable, and high-speed Internet access. This is useful in scenarios like Internet applications, online gaming, e-commerce, and finance.

Implementation:

- A direct connection is used to connect an on-premises data center to a VPC.
- A public NAT gateway is created in the VPC, enabling Internet connectivity for the on-premises servers.



Notes:

- The default route of the on-premises data center must be available for configuring Direct Connect.
- The CIDR block of the on-premises data center cannot overlap with the subnet CIDR block of the VPC. Otherwise, the communication between the on-premises data center and the VPC will fail.

52 Huawei Confidential

Advantages

- With Direct Connect, you can access a VPC on Huawei Cloud over secure, high-performance, low-latency networks. A single connection supports up to 100 Gbit/s of bandwidth, which meets a diverse range of bandwidth requirements.
- A public NAT gateway allows multiple servers to share an EIP, saving money on EIPs. The public NAT gateway types and EIPs bound to the gateway can be changed at any time. The configuration is simple and will take effect immediately.



- An SNAT connection consists of a source IP address, source port, destination IP address, destination port, and transport layer protocol. The source IP address refers to the EIP, and the source port refers to the EIP port. These five elements identify a connection as a unique session.
- Throughput specifies the total bandwidth of EIPs in a DNAT rule. For example, a public NAT gateway has two DNAT rules. If the EIP bandwidth in the first rule is 10 Mbit/s and that in the second rule is 5 Mbit/s, the throughput of the public NAT gateway is 15 Mbit/s.
- Each public NAT gateway supports up to 20 Gbit/s of bandwidth.
- Common scenarios and recommended NAT gateway types:
 - Small or medium: scenarios where there are a small number of destination addresses and connections, such as upload, download, and Internet access.
 - Large or extra large: scenarios where there are a large number of destination addresses or ports and connections, such as crawlers and client push.
- The maximum number of SNAT connections varies depending on the NAT gateway type. The details are as follows:
 - Small: 10,000
 - Medium: 50,000
 - Large: 200,000
 - Ultra-large: 1,000,000

Application Scenarios of Private NAT Gateways

Using private NAT gateway and Direct Connect to enable communication between a VPC and an on-premises data center: A private NAT gateway enables ECSs in a VPC to communicate with on-premises servers through a specific IP address over a direct connection.

Implementation:

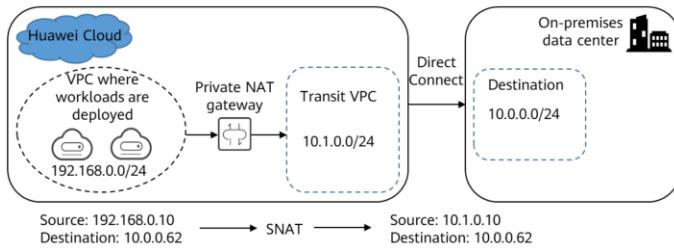
- A direct connection is used to connect the on-premises data center to the transit VPC.
- A private NAT gateway connects the service VPC to the transit VPC. The service VPC indicates the VPC where services are deployed.

Notes:

- The CIDR block of the on-premises data center cannot overlap with the subnet CIDR block of the transit VPC on the cloud, or communications between the on-premises data center and the transit VPC will fail.
- A CIDR block, in the transit VPC for performing NAT for resources in the service VPC, has to be determined.

Advantages

In this hybrid cloud scenario, a direct connection enables communications between a transit VPC and the on-premises network. Private NAT gateways can translate IP addresses of cloud servers in the service VPC to a specified private IP address (serving as a transit IP address) in the transit VPC, so that the cloud servers in the service VPC can share the transit IP address to communicate with the on-premises servers over the direct connection, meeting security compliance requirements.



Contents

1. Network Service Overview

2. Network Planning

3. Network Access

- Elastic IP (EIP)
- Elastic Load Balance (ELB)
- NAT Gateway
- Domain Name Service (DNS)

Domain Name Service (DNS)

- DNS is a highly available and scalable authoritative DNS service that translates domain names into IP addresses, redirecting visitors to the desired resources.



High performance

A single DNS node can handle millions of concurrent queries, allowing end users to access your website or application more quickly.

Security

DNS offers built-in DDoS mitigation and works with Anti-DDoS to ensure that requests from your legitimate end users are not affected.

Private domain name resolution

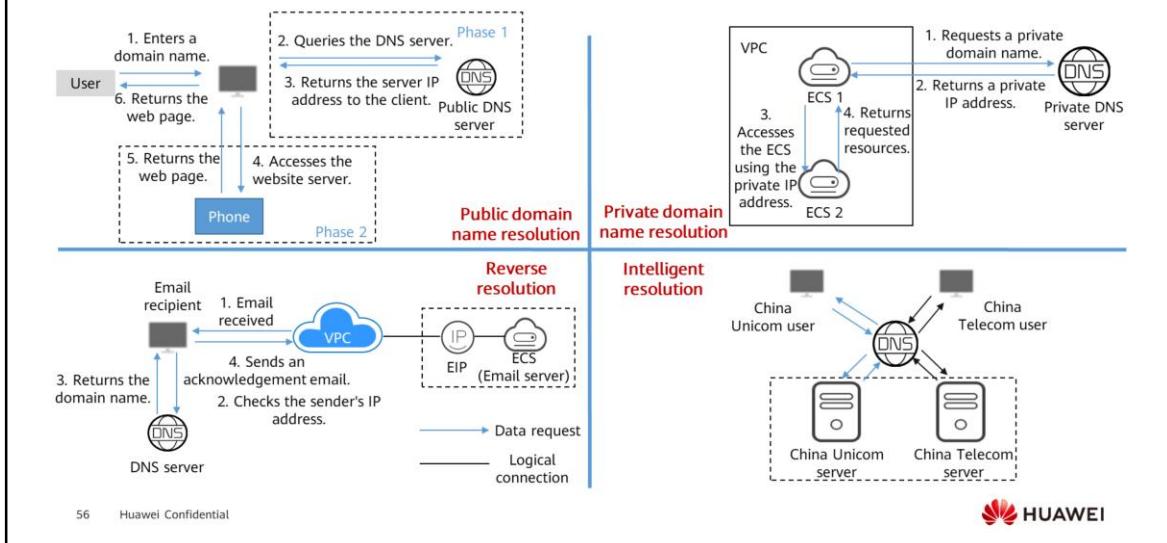
DNS provides you with secure private domain name resolution. You can have your own authoritative DNS servers in VPCs and avoid exposing your DNS records to the Internet. Private domain names improve resolution efficiencies, reduce latencies, and prevent DNS spoofing.

Reverse resolution

Pointer records (PTR) can be added to point IP addresses to domain names, reducing the number of junk mails.

- Smooth service migration
 - You can migrate an in-use website domain name to the DNS service. To ensure that your website services are not interrupted during the migration, we will create a public zone and add DNS record sets for your website in advance.

Types of Resolution Services Provided by DNS

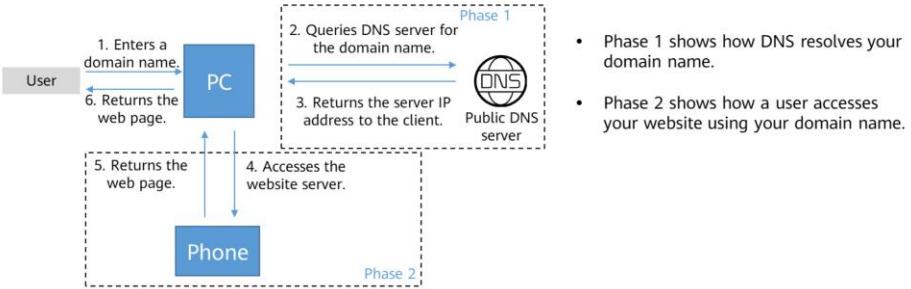


- **Public domain name resolution:** maps domain names to public IP addresses so that your users can access your website or web applications over the Internet. A public zone contains information about how a domain name and its subdomains are translated into IP addresses for routing traffic over the Internet.
- **Private domain name resolution:** Translates private domain names into private IP addresses to facilitate access to cloud resources within VPCs. A private zone contains information about how to map a domain name (such as `ecs.com`) and its subdomains used within one or more VPCs to private IP addresses (such as `192.168.1.1`). With private domain names, your ECSs can communicate with each other within the VPCs without having to connect to the Internet. These ECSs can also access cloud services, such as OBS and Simple Message Notification (SMN), over a private network.
- **Reverse resolution:** DNS 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. For example, if the visitor is a China Unicom user, the DNS server will return an IP address of China Unicom. With this function, you can improve DNS resolution efficiency and speed up cross-network access. You can also create more fine-grained resolution lines based on source IP addresses.

Public Domain Name Resolution

- A public zone contains information about how a domain name and its subdomains are translated into IP addresses for routing traffic over the Internet.

How DNS routes Internet traffic to a website

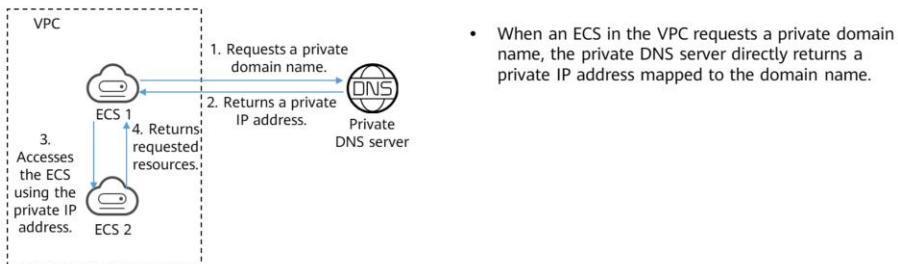


- Phase 1 shows how DNS resolves your domain name.
- Phase 2 shows how a user accesses your website using your domain name.

Private Domain Name Resolution

- A private zone contains information about how to map a domain name (such as ecs.com) and its subdomains used within one or more VPCs to private IP addresses (such as 192.168.1.1). With private domain names, your ECSs can communicate with each other within the VPCs without having to connect to the Internet. These ECSs can also access cloud services, such as OBS and SMN, over a private network.

Process for resolving a private domain name



Quiz

1. (Multiple-answer question) Which of the following are components of the Elastic Load Balance service?
 - A. Backend server group
 - B. Listener
 - C. Load balancer
 - D. NAT gateway
2. (Single-answer question) Two VPCs that are from different AZs but the same region need to communicate with each other, which of the following statements is true?
 - A. You can use a VPC peering connection to enable the communication.
 - B. They cannot communicate with each other.
 - C. By default, they can communicate with each other, but the communication can be disabled.
 - D. They can only use VPN to communicate.

- ABC
- A

Quiz

1. (Discussion) What are the advantages of cloud network services over traditional network services?

2. (Discussion) What should be considered for using cloud network services in terms of security, cost, reliability, performance, and scalability?

- Discussion 1:
 - Cloud network services are managed by cloud service providers.
 - Traditional network services are managed by users.

- Discussion 2:
 - To ensure network security, configure outbound and inbound rules and allow traffic only on specific ports.
 - To reduce costs, delete servers that are not working in a backend server group for load balancing immediately.
 - To ensure reliability, perform health check to ensure that backend servers are healthy and are of the same type.
 - To ensure performance, use the monitoring function to maximize the load capability of ECSs.
 - To ensure scalability, use the Layer 4 and Layer 7 forwarding capabilities of load balancers to prevent network access congestion.

Summary

- This course describes basic network knowledge and common cloud network services.
- After completing this course, you will understand the functions of networks as well as how network cloud services work and when you need to use these services. For example, a VPC is like the internal network used by an enterprise, and applications can provide internet-accessible services using EIPs. Mastering these concepts can help you better prepare for cloud migration.

Acronyms and Abbreviations

- API: Application Programming Interface
- AS: Auto Scaling
- BMS: Bare Metal Server
- CBR: Cloud Backup and Recovery
- CDN: Content Delivery Network
- DCS: Distributed Cache Service
- DRS: Data Replication Service
- DNS: Domain Name Service
- DDoS: Distributed Denial of Service
- DevOps: Development and Operations
- DIS: Data Ingestion Service
- DLI: Data Lake Insight
- EIP: Elastic IP
- ECS: Elastic Cloud Server
- ELB: Elastic Load Balance
- EVS: Elastic Volume Service
- GSLB: Global Server Load Balance
- HA: High Availability
- IMS: Image Management Service
- IDC: Internet Data Center
- LTS: Log Tank Service

Acronyms and Abbreviations

- NAT: Network Address Translation
- OLAP: Online Analytical Processing
- OLTP: Online Transaction Processing
- RDS: Relational Database Service
- SMN: Simple Message Notification
- SFS: Scalable File Service
- SDRS: Storage Disaster Recovery Service
- VM: Virtual Machine
- VPC: Virtual Private Cloud
- VPN: Virtual Private Network

Recommendations

- Huawei iLearning
 - <https://e.huawei.com/en/talent/portal/#/>
- Huawei Cloud Help Center
 - <https://support.huaweicloud.com/intl/en-us/index.html>
- HUAWEI CLOUD Developer Institute
 - <https://edu.huaweicloud.com/intl/en-us/>
- Huawei Talent Online
 - <https://e.huawei.com/en/talent/portal/#/>

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©2022 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.



Storage Service Planning



Foreword

- 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 course, we will cover some common storage services provided by Huawei Cloud.

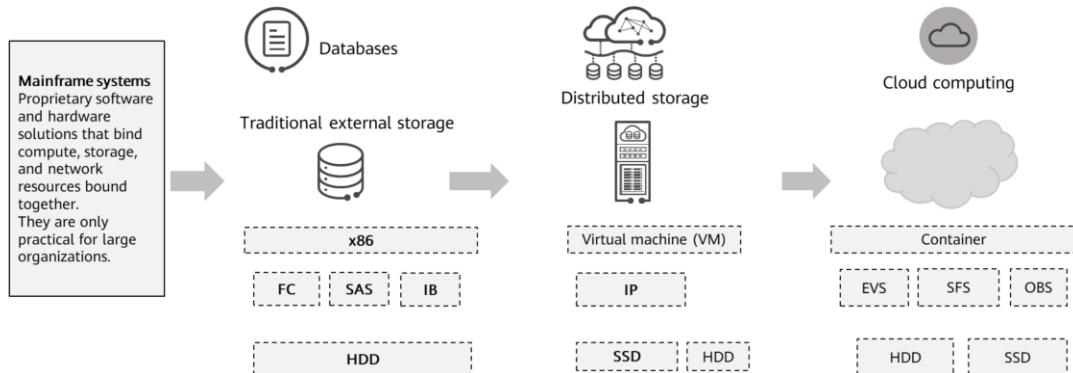
Objectives

- Upon completion of this course, you will:
 - Acquire a basic understanding of cloud storage.
 - Acquire the principles behind and uses of common storage services on Huawei Cloud.

Contents

- 1. Storage Service Overview**
2. Storage Service Planning
3. Content Delivery Network
4. Backup Solution Planning
5. DR Solution Planning

Data Storage Trends



4 Huawei Confidential



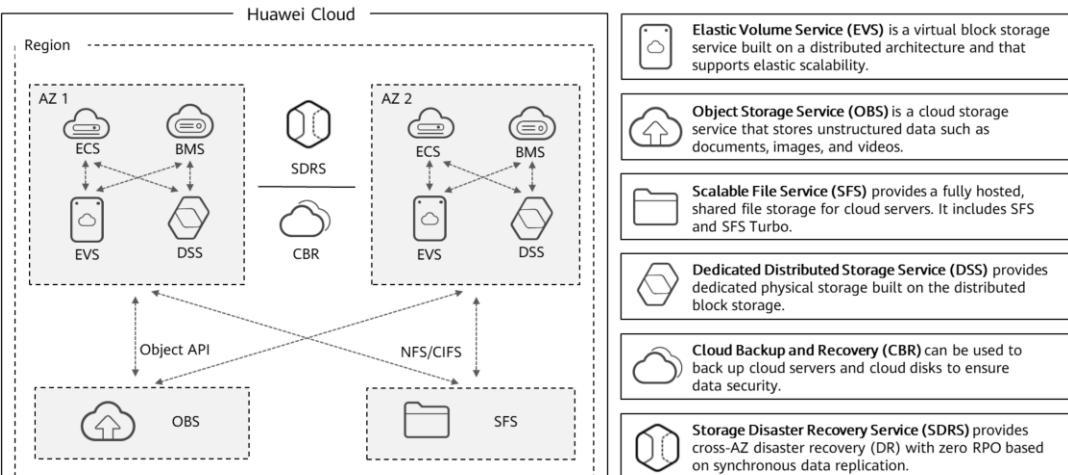
- In the IBM mainframe era, file, network, and storage capabilities are all encapsulated into one environment. In the x86 era, data in databases is stored in x86 servers. In the virtualization era, data is stored in VMs using the distributed technology. Services are migrating to the cloud, data is stored on the cloud, and all-IP network protocols become a major trend.

Traditional On-Premises Storage VS. Cloud Storage

Item	Traditional Storage	Cloud Storage
Storage Type	Direct Access Storage (DAS), Network Attached Storage (NAS), Storage Area Network (SAN)	Elastic Volume Service (EVS), Scalable File Service (SFS), Object Storage Service (OBS)
Typical architecture	Dedicated storage devices and storage systems	Software-defined storage
Flexibility	Inflexible capacity expansion, professional knowledge required	Fast, flexible, easy to expand, and ease of use
Cost	Procurement costs are high, the O&M depends on professional storage vendors, and after the service life of a device ends, the device needs to be replaced, all of which drives up TCO.	Pay-per-use billing, professional O&M provided by cloud vendors, and no more hardware O&M
Other	Resource utilization is low, data sharing difficult, technological grades difficult, and costs high	Resource utilization is high and service innovation continuous, with new cloud technologies constantly being incorporated

- Based on the server type, storage can be classified into closed storage and open storage. Open storage can be then classified as built-in storage and external storage. External storage can be further classified as Direct-Attached Storage (DAS), Network-Attached Storage (NAS), and Storage Area Network (SAN) based on the connection method and transmission protocol.
- DAS: Although DAS is old, it is still suitable for scenarios where the data volume is small and the requirement for access speed is not high.
- NAS: NAS is suitable for file servers to store unstructured data. Although their access speed is limited by the Ethernet, NAS can be flexibly deployed at low costs.
- SAN: SAN is suitable for large-scale applications or database systems. But SAN is costly and complex.
- Block storage: Block storage breaks up data into blocks and then stores those blocks as separate pieces, each with a unique identifier. Those blocks of data can be placed wherever it is most efficient. That means each block can be configured (or partitioned) to work with different operating systems.
- File storage: File storage is also referred to as file-level or file-based storage. File storage data is stored as single pieces of data in folders.
- Object storage: Object storage, which is also known as object-based storage, breaks data files up into pieces called objects. It then stores those objects in a single repository, which can be spread out across multiple networked systems.

Huawei Cloud Storage Services



6 Huawei Confidential



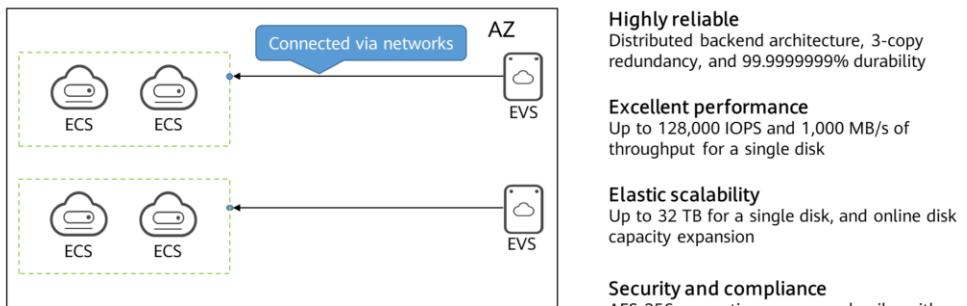
- Recovery Point Objective (RPO): the maximum tolerable amount of lost data
- Recovery Time Objective (RTO): the maximum tolerable service downtime, from the time when a disaster happened to the time when services were recovered

Contents

1. Storage Service Overview
2. **Storage Service Planning**
 - Block Storage – EVS
 - Object Storage – OBS
 - File Storage – SFS
 - Dedicated Distributed Storage – DSS
3. Content Delivery Network
4. Backup Solution Planning
5. DR Solution Planning

EVS

- Elastic Volume Service (EVS) offers scalable block storage for cloud servers. EVS disks offer high reliability, excellent performance, and come in a variety of specifications. They can be used for distributed file systems, development and test environments, data warehouses, and high-performance computing (HPC) applications.



- Precautions:
 - The maximum number of disks that can be attached to a cloud server varies with server specifications.
 - When attaching a disk, ensure that the server and disk reside in the same AZ. Or, the attachment will fail.
 - A backup of a disk will be created in the same AZ of the disk.

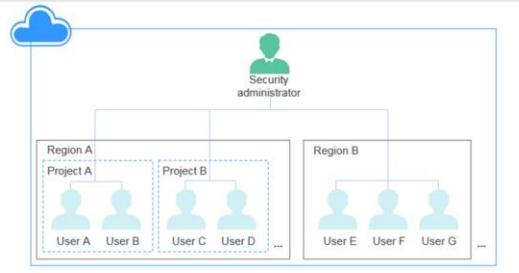
EVS Disk Types

- EVS disks include extreme SSD, ultra-high I/O, general purpose SSD, and high 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.

Disk Type	Max. IOPS	Max. Throughput	Latency	Data Durability	Max. Capacity	Service Scenario
High I/O	5,000	150 MB/s	1 to 3 ms	99.999999% (9 nines)	32 TB	Good for applications and development and test environments with common workloads
General Purpose SSD	20,000	250 MB/s	1 ms			Good for mainstream high-performance, low-latency interactive applications, such as enterprise OA, transcoding services, large-scale development and testing, web server logs, and performance-demanding system disks
Ultra-high I/O	50,000	350 MB/s	1 ms			Good for bandwidth-demanding workloads, including read/write-intensive applications, transcoding services, and I/O-intensive services, as well as latency-sensitive workloads
Extreme SSD	128,000	1 GB/s	Sub-millisecond			Good for databases and AI workloads

- Recommended use:
 - High I/O disks are recommended to be used as system disks.
 - SSD-based disks are recommended to be used as data disks.

Encryption

Encryption definition	User permissions of using encryption
<p>EVS uses the industry-standard XTS-AES-256 encryption algorithm and keys for EVS encryption. Keys used for encryption are provided by the Key Management Service (KMS) of Data Encryption Workshop (DEW), which is secure and convenient. You do not need to establish and maintain the key management infrastructure.</p> 	<p>Security administrator This account type can grant KMS access rights to EVS and to use the encryption function. Common user The first common user in a region or project to use encryption will need a security administrator to grant the necessary permissions first. Later users can use encryption directly.</p> <p>The following example describes the procedure when the security administrator or common user E is the first user in region B to use the encryption function.</p> <ul style="list-style-type: none">• The security administrator:<ol style="list-style-type: none">1. Grants KMS access rights to EVS.2. Selects a key.• Common user E:<ol style="list-style-type: none">1. Uses the encryption function, and the system responds a message showing that KMS access rights have not been granted to EVS.2. Contacts the security administrator to request KMS access rights to EVS.

10 Huawei Confidential



- If the security administrator is the first one to use the encryption function, the procedure is as follows:
 - Grants KMS access rights to EVS. After KMS access rights have been granted, the system automatically creates a Default Master Key (DMK) and names it evs/default. The DMK can be used for encryption.
 - Note: EVS encryption relies on KMS. When the encryption function is used for the first time ever, KMS access rights need to be granted to EVS. After KMS access rights have been granted, all the users in this region can use the encryption function, and KMS access rights do not need to be granted again.
 - Selects a key. Users can select one of the following keys: the DMK, evs/default.
 - CMKs, including existing CMKs or new CMKs.
 - After the security administrator has used the encryption function, all the users in region B can directly use the encryption function.
- If user E (common user) is the first one to use the encryption function, the procedure is as follows:
 - Uses the encryption function, and the system responds a message showing that KMS access rights have not been granted to EVS.
 - Contacts the security administrator to request KMS access rights to EVS.
- After KMS access rights have been granted to EVS, user E as well as all the users in region B can directly use the encryption function and do not need to contact the security administrator to request KMS access rights to EVS again.

Capacity Expansion

- If your EVS disk is running out of space, you can increase the disk size by expanding capacity.
- Disk capacity can be expanded, but cannot be reduced.



Customer pain points: service interruptions, long expansion process, and complex performance optimization after expansion

Advantages:

On-demand purchase: EVS disks are at least 10 GB and can be scaled up in increments of at least 1 GB.

Online capacity expansion: Storage space can be increased anytime **without interrupting services**.

No need for performance optimization: **Disk capacity increases linearly after expansion**.

Procedure:

1. Expand disk capacity on the management console.
2. Log in to the ECS and extend the disk partition and file system.

- Expanding capacity on the management console:
 - Choose an appropriate expansion method based on the disk status. View the disk status. If the disk status is In-use, the disk has been attached to a server. Check whether the disk can be expanded in the In-use state based on the constraints. If so, directly expand the disk capacity. If not, detach the disk and then expand the disk capacity. If the disk status is Available, the disk has not been attached to any server. You can directly expand the disk capacity.

Snapshot

- An EVS snapshot is a complete copy or image of the disk data taken at a specific point in time. If data is lost, you can use a snapshot to restore the disk data to the state when the snapshot was created.



Application scenarios: environment deployment, hacker attacks, and data backup before high-risk operations

Advantages:

By creating snapshots, you can quickly save EVS disk data at specified time points. You can also use snapshots to create new disks that already contain certain data.

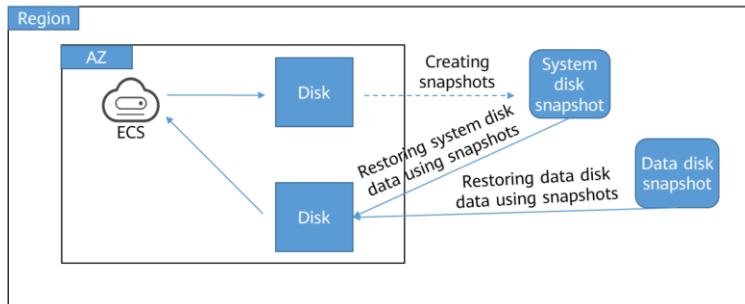
Application Scenarios:

- Routine data backup
- Rapid data restoration
- Multi-service quick deployment

- Routine data backup
 - You can create snapshots for disks regularly and use snapshots to recover your data in case that data is lost or inconsistent due to misoperations, viruses, or attacks.
- Rapid data restoration
 - You can create a snapshot or multiple snapshots before an application software upgrade or a service data migration. If an exception occurs during the upgrade or migration, service data can be rapidly restored to the state when the snapshot was created.
- Multi-service quick deployment
 - You can use a snapshot to create multiple disks containing the same initial data, and these disks can be used as data resources for various services.

Application Scenarios

- If data on an EVS disk is incorrect or damaged, you can restore the disk data using a snapshot.



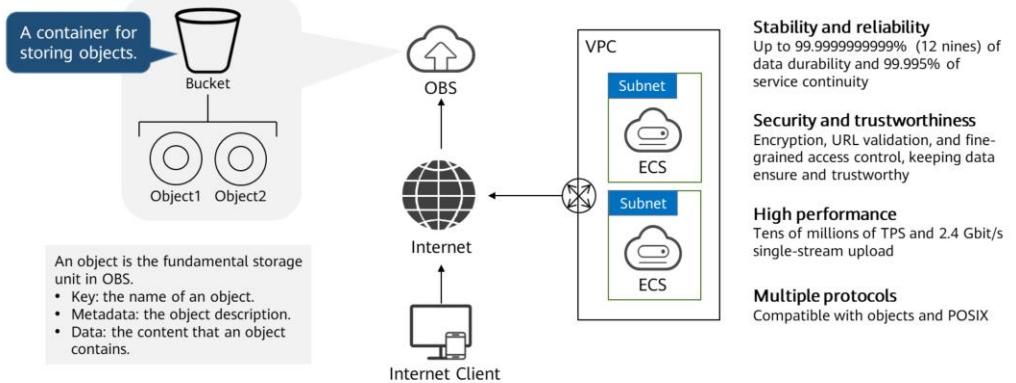
- A snapshot can be used to roll back data only to the source disk. Rollback to another disk is not possible.
- A snapshot roll backup can be performed only when the snapshot status is **Available** and the source disk status is **Available** (not attached to any server) or **Rollback failed**.

Contents

1. Storage Service Overview
2. **Storage Service Planning**
 - Block Storage – EVS
 - Object Storage – OBS
 - File Storage – SFS
 - Dedicated Distributed Storage – DSS
3. Content Delivery Network
4. Backup Solution Planning
5. DR Solution Planning

OBS

- Object Storage Service (OBS) provides massive, secure, and reliable storage for you to inexpensively store data of any format and of any size. You can use OBS for enterprise backup and archive, video on demand (VOD), video surveillance, and many other scenarios.



- A bucket is a container for storing objects in OBS. OBS offers a flat structure based on buckets and objects. This structure enables all objects to be stored at the same logical layer, rather than being stored hierarchically. Each bucket has its own properties, such as the storage class, access control, and region. You can create buckets with required storage classes and access control in different regions and further configure advanced settings, to meet storage requirements in a wide range of scenarios.
- OBS provides massive storage for files of any format, catering to the needs of common users, websites, enterprises, and developers. Neither the entire OBS system nor any single bucket has limitations on the storage capacity or the number of objects/files that can be stored. As a web service, OBS supports APIs over HTTP and HTTPS. You can easily access and manage data stored in OBS anytime, anywhere through OBS Console or OBS tools. In addition, OBS SDKs and APIs make it easy to manage data stored in OBS and to develop upper-layer applications.

OBS Storage Classes

- OBS offers the following three storage classes to meet a diverse range of needs for storage performance and cost.



Standard

With low latency and high throughput, it is good for storing frequently (multiple times per month) accessed files or small files (less than 1 MB).



Infrequent Access

Infrequent Access storage is for storing data that is accessed less than 12 times per year but when needed, the access has to be fast.



Archive

Archive storage is ideal for storing data that is rarely (once per year) accessed.

- Standard:
 - The Standard storage class is appropriate for a wide range of application scenarios, including big data analytics, mobile applications, hot videos, and social images.
- Infrequent Access:
 - The Infrequent Access storage class can be used for file synchronization and sharing, enterprise backups, and many other scenarios. It has the same durability, low latency, and high throughput as the Standard storage class, with a lower cost, but its availability is slightly lower than the Standard storage class.
- Archive:
 - The Archive storage class is ideal for scenarios such as data archive and long-term backups. It is secure and durable and delivers the lowest cost among the three storage classes. The OBS Archive storage class can be used to replace tape libraries. To save money, it may take hours to restore the archived data.

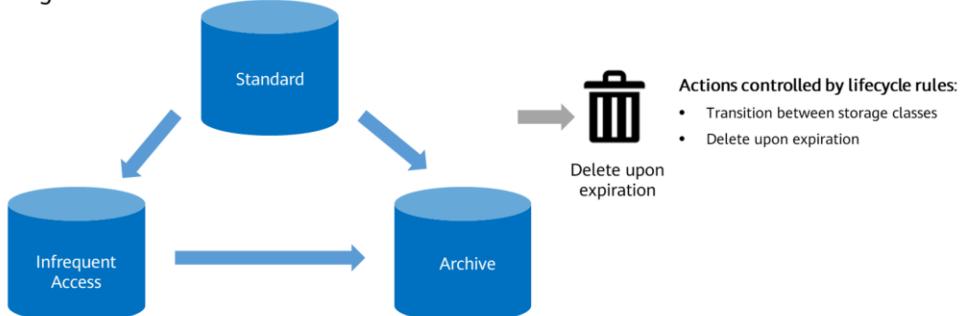
Comparison Between OBS Storage Classes

Item	Standard	Infrequent Access	Archive
Features	Top-notch performance, high reliability and availability	Reliable, inexpensive storage with real-time access	Long-term retention of archived data at a low cost
Application scenarios	Cloud applications, data sharing, content sharing, and hot data storage	Web disk applications, enterprise backups, active archive, and data monitoring	Storage of archives, medical imaging data, and videos, as well as replacement of tape libraries
Durability (single-AZ)	99.99999999%	99.99999999%	99.99999999%
Durability (multi-AZ)	99.999999999%	99.999999999%	Not supported
Availability (single-AZ)	99.99%	99%	99%
Availability (multi-AZ)	99.995%	99.5%	Not supported
Minimum storage duration	N/A	30 days	90 days
Data retrieval	N/A	Billed for each GB retrieved.	Data can be restored at a standard or an expedited speed. Billed for each GB restored.
Image processing	Supported	Supported	Not supported

- You can choose multi-AZ storage or single-AZ storage as your redundancy policy based on your business needs. The multi-AZ storage stores data in multiple AZs to deliver up to 99.999999999% of data durability and up to 99.995% of service continuity, far higher than those of a conventional architecture.
- The 12 nines of durability means that the average annual loss rate of objects is expected to be 0.0000000001%. For example, if you store 100 million objects in OBS, only one object may be lost every 10,000 years.
- The availability can be considered as service continuity. The 99.995% availability means that if you keep accessing OBS for 100,000 minutes (about 69 days), you can expect less than 5 minutes of unavailability.

Lifecycle Management

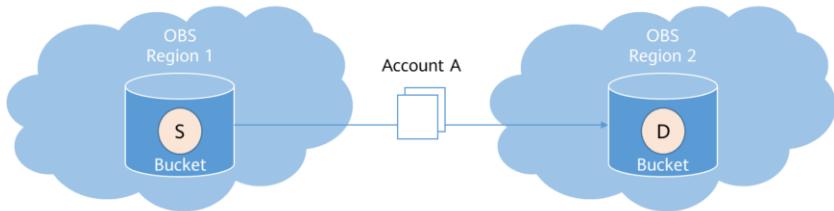
- You can configure lifecycle rules to periodically delete objects or transition objects between storage classes.



- Let's look at what these actions mean:
- Storage class transition: An object is transitioned from one storage class to another.
- Delete upon expiration: After an object has expired, it is deleted by OBS.
- The following rules are also important:
 - There is no limit on the number of lifecycle rules in a bucket, but the total size of XML descriptions about all lifecycle rules in a bucket cannot exceed 20 KB.
 - The minimum storage duration of Archive storage is 90 days. After an object is transitioned to the Archive storage class, if it stays in this storage class for less than 90 days, you still need to pay for a full 90 days.
- There are some restrictions on storage class transition using lifecycle rules:
 - Lifecycle rules can transition objects only from the Standard storage class to Infrequent Access storage class, or from the Standard or Infrequent Access storage class to Archive storage class.
 - If you want to change the storage class back from Infrequent Access to Standard, or from Archive to Standard or Infrequent Access, you must manually transition the storage class. In addition, to change the storage class of an archived object, you need to manually restore the object first.

Cross-Region Replication

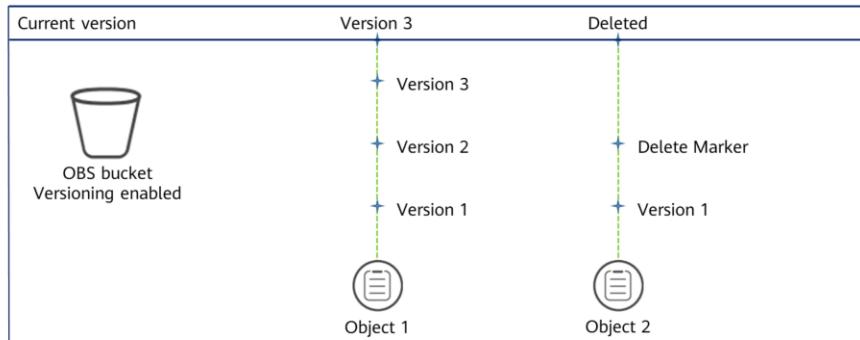
- Cross-region replication refers to the process of automatically, asynchronously replicating data from a source bucket in one region to a destination bucket in another region based on a configured replication rule. The source and destination buckets must be owned by the same account. Data currently cannot be replicated across accounts.



- You can configure a rule to replicate only objects with a specified prefix or to replicate all objects in a bucket. Replicated objects in the destination bucket are copies of those in the source bucket. Objects in both buckets 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.
- The content that is replicated includes:
 - 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 the function of synchronizing existing objects is enabled (excluding those in the Archive storage class).

Versioning

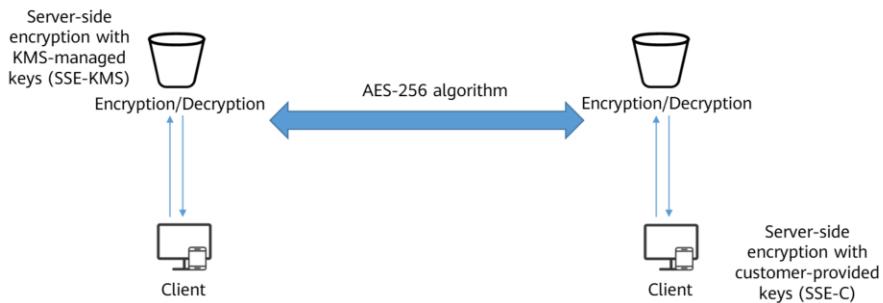
- With versioning enabled for a bucket, OBS can keep multiple versions of an object in the bucket. This allows you to recover an object if there is, for example, an accidental deletion or file overwrite, or an application failure. Versioning can also be used for data retention and archive.



- On this slide, the bucket stores two objects: object 1 and object 2. As versioning has been enabled for this bucket, the current version of object 1 is version 3. By querying the historical records, you can find that version 1 and version 2 are the noncurrent versions of object 1. In addition, the current version of object 2 has been deleted because there is a delete marker. By querying the historical records, you can find that version 1 is the noncurrent version of object 2.

Server-Side Encryption

- With server-side encryption enabled, objects uploaded to OBS will be encrypted before being stored on the server. When you download an encrypted object, the object will be decrypted first on the server and then returned to you in plaintext form. OBS server-side encryption uses SSE-KMS and SSE-C encryption, which both use AES-256 algorithms.



- Server-side encryption with KMS-managed keys (SSE-KMS)
 - With this method, you need to create a key using Key Management Service (KMS) or use the default key provided by KMS. The KMS key is then used for server-side encryption when you upload objects to OBS.
- Server-side encryption with customer-provided keys (SSE-C)
 - For this method, the customer-provided keys and their MD5 values are used for server-side encryption.

Event Notifications

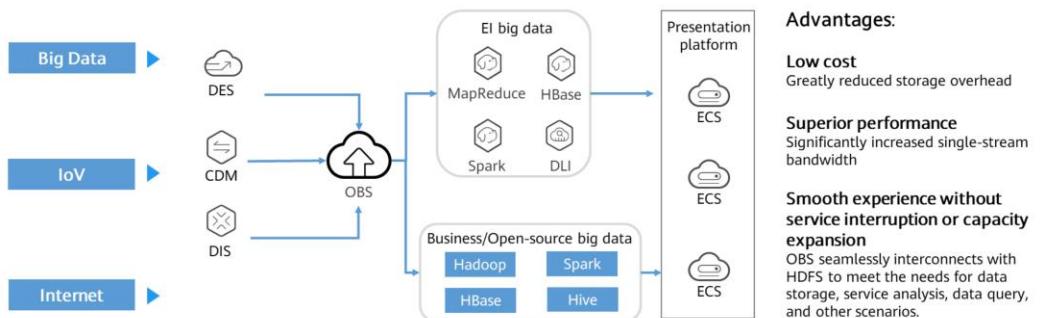
- OBS uses SMN to send notifications to specified subscribers when certain events (including uploads and deletions) happen in buckets. For example, after you configure a notification rule, whenever you upload objects to the specified bucket, SMN will send a notification to an email address you specify.



- Events supported by OBS are listed as follows:
- OBS provides APIs such as PUT, POST, and COPY for uploading objects. You can configure event types corresponding to these APIs. Then, when you use such an API to upload an object, you will receive a notification. You can also configure the ObjectCreated:* event type to obtain all object upload notifications.
 - ObjectCreated:*
 - ObjectCreated:Put (uploading an object)
 - ObjectCreated:Post (uploading an object with a browser)
 - ObjectCreated:Copy (copying an object)
 - ObjectCreated:CompleteMultipartUpload (merging parts)
- By configuring the ObjectRemoved event type, you can receive a notification when one or more objects are removed from a bucket.
- By configuring the ObjectRemoved:Delete event type, you can receive a notification when an object is deleted or an object version is permanently deleted. By configuring the ObjectRemoved:DeleteMarkerCreated event type, you can receive a notification when a delete marker is added to an object. You can also use ObjectRemoved:* to receive a notification each time an object is deleted.
 - ObjectRemoved:*
 - ObjectRemoved:Delete (deleting an object)
 - ObjectRemoved:DeleteMarkerCreated (adding a delete marker to an object)

Application Scenarios

- You can migrate massive amounts of data to OBS with a migration service, and then use Huawei Cloud big data services (like MapReduce) or open-source computing frameworks (such as Hadoop and Spark) to analyze data stored in OBS. Such analysis results will be presented in your programs or applications on Elastic Cloud Servers (ECSs).



23 Huawei Confidential



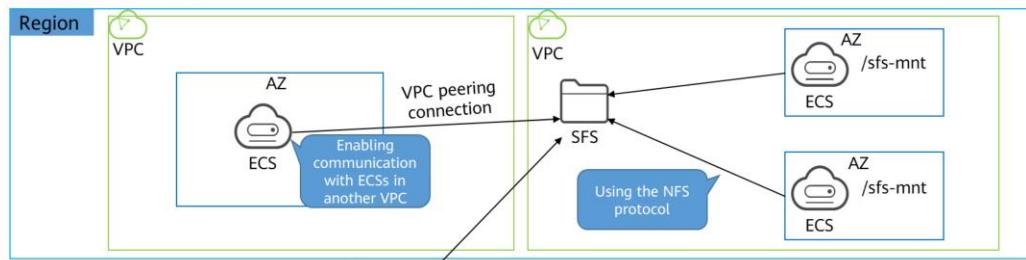
- The OBS big data solution is designed for a variety of scenarios, including storage and analysis of massive amounts of data, query of historical data details, analysis of a large number of behavior logs, and analysis and statistics of public transactions.
- Typical scenarios of storage and analysis of massive amounts of data include:
 - Storage for petabytes of data, batch data analysis, and response for data detail queries in seconds
- Typical scenarios of query of historical data details include:
 - Transaction audit, device energy consumption analysis, trail playback, driving behavior analysis, and fine-grained monitoring
- Typical scenarios of analysis of a large number of behavior logs include:
 - Analysis of learning habits and operation logs, as well as analysis and query of system operation logs
- Typical scenarios of analysis and statistics of public transactions include:
 - Crime tracking, associated case queries, traffic congestion analysis, and scenic spot popularity statistics

Contents

1. Storage Service Overview
2. **Storage Service Planning**
 - Block Storage – EVS
 - Object Storage – OBS
 - **File Storage – SFS**
 - Dedicated Distributed Storage – DSS
3. Content Delivery Network
4. Backup Solution Planning
5. DR Solution Planning

SFS

- Scalable File Service (SFS) provides high-performance file storage that is scalable on demand. With SFS, you can enjoy shared file access spanning multiple ECSs, BMSSs, and containers.

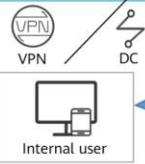


Easy to use

Fully-hosted file storage frees you from the complexities of hardware deployment and maintenance.

Secure

Huawei Cloud security comprehensively secures your data. VPC-based user authentication keeps your data isolated in your own cloud.



Using the NFS protocol
Allowing for connection with internal users

Durable

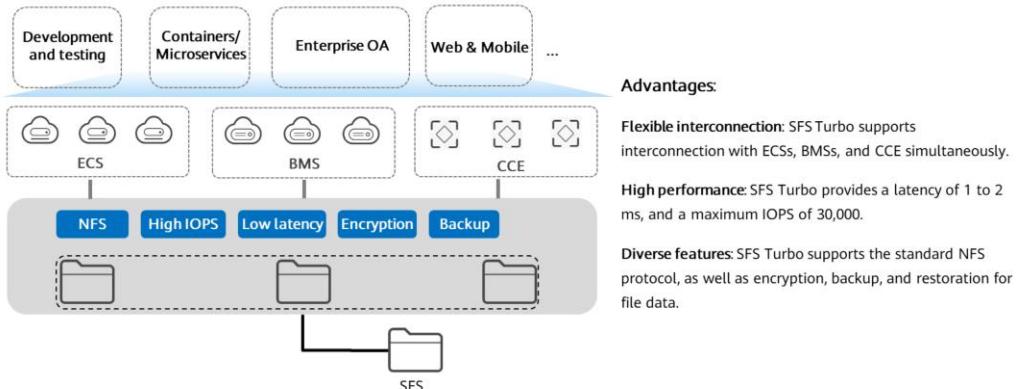
A multi-level reliability architecture ensures a data durability of 99.9999999% (10 nines) and service availability of 99.95%.

Efficient

High IOPS, low latency, and high bandwidth options are all offered so you can design a storage solution tailored to your specific performance needs.

SFS Turbo

- Expandable to 320 TB, SFS Turbo General provides fully hosted shared file storage, highly available and durable to support massive small file storage and support for applications requiring low latency and high IOPS.



Application Scenarios

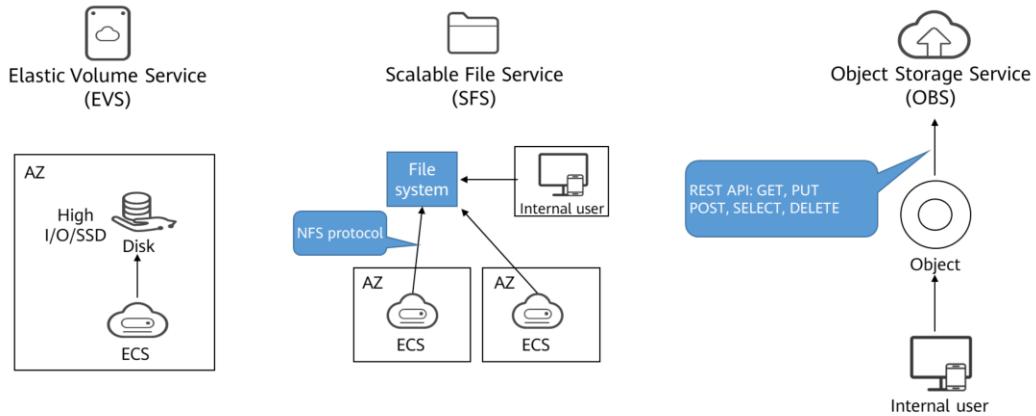
HPC	Media processing	File sharing	Web services
			
<p>High-bandwidth, large-capacity SFS is recommended for high-performance computing (HPC). Shared file storage facilitates industrial design (CAD/CAE), biomedicine, energy exploration, graphics rendering, and heterogeneous computing.</p>	<p>High-bandwidth, large-capacity SFS is recommended for media processing. Shared file storage facilitates multi-layer HD and 4K video editing, transcoding, composition, and video on demand (VoD).</p>	<p>High-IOPS, low-latency SFS Turbo is recommended for file sharing. For companies with a large number of departments and employees, documents and data can be shared and accessed company-wide.</p>	<p>High-IOPS, low-latency SFS Turbo is recommended for web services. The file systems secure data storage for content management systems and web applications, facilitating quick online publishing and archiving.</p>

Differences Among SFS, OBS, and EVS

- There are three data storage services available for you to choose from currently: EVS, SFS, and OBS. The differences are described in the following table.

Dimension	SFS	OBS	EVS
Definition	SFS provides on-demand high-performance file storage, which can be shared by multiple servers.	OBS provides massive, secure, reliable, and cost-effective data storage for data of any type and size.	EVS provides scalable block storage that features high reliability, high performance, and a variety of specifications to meet various service requirements.
Data storage logic	SFS stores files. Data is sorted and displays in files and folders.	OBS stores data as objects. Files can be saved directly to OBS and system metadata, which can also be customized, will be automatically generated.	EVS stores binary data. Files cannot be stored directly. To store files, you need to format the disk first.
Access method	SFS file systems can be accessed only after being mounted to ECSs or BMSS through NFS or CIFS. A network address must be specified or mapped to a local directory for access.	OBS buckets can be accessed through the Internet or Direct Connect. The bucket address must be specified for access, and both HTTP and HTTPS are used.	EVS disks can be used and accessed from applications only after being attached to ECSs or BMSS and initialized.
Application scenario	HPC, media processing, file sharing, content management, and web services	Big data analytics, static website hosting, online video on demand (VoD), gene sequencing, and intelligent video surveillance	HPC, enterprise clustered applications, enterprise application systems, and development and testing

Summary of EVS, SFS, and OBS



29 Huawei Confidential



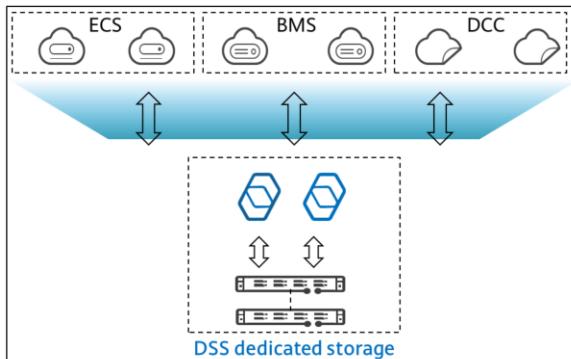
- EVS: Raw disk spaces are mapped entirely to hosts or VMs. You can format the disk with any file system and use it.
- SFS: Like a shared folder, for example, a remote shared directory in Windows, the file system already exists, and you can directly store data to the file system.
- OBS: Each piece of data corresponds to a unique ID. Object storage does not have the directory structure similar to file storage. Data is stored in a flat structure, and you can locate data by object ID.

Contents

1. Storage Service Overview
2. **Storage Service Planning**
 - Block Storage – EVS
 - Object Storage – OBS
 - File Storage – SFS
 - **Dedicated Distributed Storage – DSS**
3. Content Delivery Network
4. Backup Solution Planning
5. DR Solution Planning

DSS

- Dedicated Distributed Storage Service (DSS) provides you with dedicated, physical storage resources. By flexibly interconnecting with various compute services, such as ECS, BMS, and DCC, DSS offers excellent performance in a wide range of scenarios, including HPC, OLAP, or mixed workloads.



Dedicated storage

DSS provides dedicated storage resources to ensure high disk read/write speed and data security, which can help you obtain security certifications.

Abundant features

DSS supports disk sharing, encryption, backup, and snapshot, making it perfect for enterprises in a wide range of industries.

Wide-range of scenarios

By flexibly interconnecting with compute services, such as ECS, BMS, and DCC, DSS can easily accommodate HPC, OLAP, and mixed-workload scenarios.

High performance

With a distributed storage architecture and smooth expansion capabilities, DSS provides high-throughput and high-concurrency storage, all with improved performance.

- Various specifications:
 - High I/O storage is suitable for scenarios that require high performance, high read/write speed, and real-time data storage.
 - Ultra-high I/O storage is excellent for read/write-intensive scenarios that require extremely high performance and read/write speed, and low latency.
- Elastic scalability:
 - On-demand capacity expansion: Storage pools can be expanded based on service requirements.
 - Linear performance scaling: DSS disks can be expanded while services are running, and linear performance increase can be achieved.
- Security and reliability:
 - Three-copy redundancy ensures 99.9999999% data durability.
 - Both system disks and data disks can be encrypted for improved data security.
- Backup and restore:
 - CBR allows you to create backups for DSS disks and restore the disk data using backups. Backups can be created for a DSS disk, maximizing data security and integrity and ensuring service security.

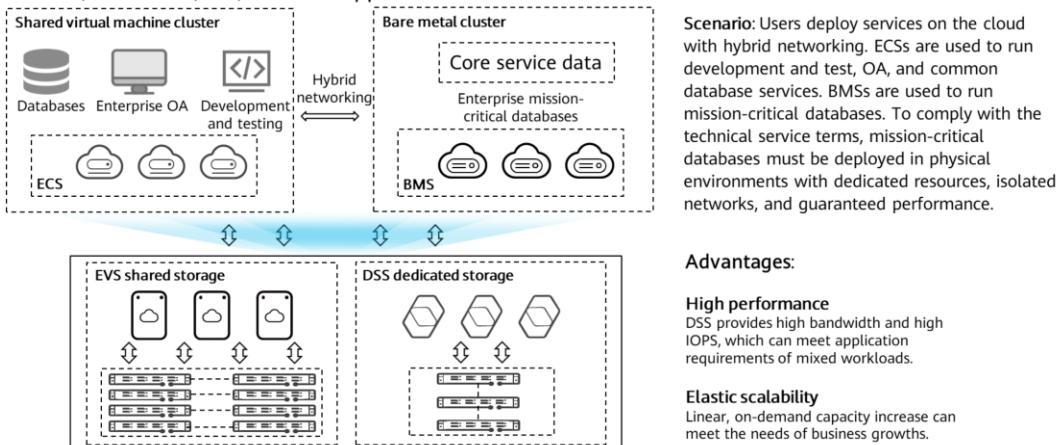
Differences Between DSS and EVS

- There are significant differences between DSS and EVS.

Service Name	Overview	Storage Type	Typical Application Scenarios	Performance and Specifications
DSS	DSS provides dedicated physical storage resources. DSS storage pools are physically isolated, and data durability reaches 99.9999999%. Multiple types of compute services, including ECS, BMS, and DCC, can be interconnected with DSS at the same time. DSS has robust features to guarantee data security and reliability.	Isolated storage pools and dedicated resources	Interconnection with compute services, such as ECS and BMS, in Dedicated Cloud Interconnection with compute services, such as ECS and BMS, in a non-Dedicated Cloud Mixed workloads. DSS supports hybrid deployment of HPC, databases, email, OA, and web applications. HPC OLAP applications	High I/O storage pool: The minimum capacity is 13.6 TB. It can be expanded to up to 435.2 TB in 13.6 TB increments. The maximum IOPS is 1,500 IOPS/TB. Ultra-high I/O storage pool: The minimum capacity is 7.225 TB. It can be expanded to up to 289 TB in 7.225 TB increments. The maximum IOPS is 8,000 IOPS/TB.
EVS	EVS provides scalable block storage that features high reliability, high performance, and is available in a variety of specifications.	Shared storage pools	Enterprise office applications Development and testing Enterprise applications, including SAP, Microsoft Exchange, and Microsoft SharePoint Distributed file systems Various databases, including MongoDB, Oracle, SQL Server, MySQL, and PostgreSQL	EVS disks start at 10 GB and can be expanded as required in 1 GB increments to up to 32 TB.

Application Scenarios

- DSS is designed for high concurrency, high throughput scenarios and supports hybrid deployment of HPC, databases, OA, and web applications.



Scenario: Users deploy services on the cloud with hybrid networking. ECSSs are used to run development and test, OA, and common database services. BMSs are used to run mission-critical databases. To comply with the technical service terms, mission-critical databases must be deployed in physical environments with dedicated resources, isolated networks, and guaranteed performance.

Advantages:

High performance
DSS provides high bandwidth and high IOPS, which can meet application requirements of mixed workloads.

Elastic scalability
Linear, on-demand capacity increase can meet the needs of business growths.



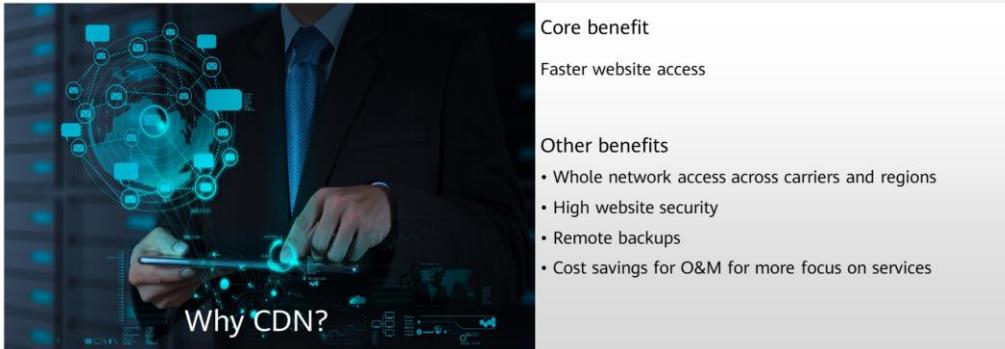
- Enterprise customers: IDC hosting customers, securities settlement companies, and more.
- Customers use EVS shared storage and DSS dedicated storage for their services. EVS provides storage for enterprise OA, development and testing, and databases. DSS provides storage for the mission-critical services running on BMSs.

Contents

1. Storage Service Overview
2. Storage Service Planning
- 3. Content Delivery Network**
4. Backup Solution Planning
5. DR Solution Planning

Why CDN?

- 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 edge nodes.



- CDN facilitates whole network access across carriers and regions. Websites cannot be accessed due to various factors, such as regional ISP limitation and egress bandwidth limitation. CDN can cover global lines. It cooperates with carriers to deploy Internet Data Center resources and edge nodes on networks of backbone node providers. CDN helps customers make the most of bandwidth resources and balance origin server traffic.
- Load balancing and distributed storage of CDN enhance website security and reliability to cope with most Internet attacks. The anti-attack system can also protect websites from malicious attacks.
- CDN supports remote backups. When a server is faulty, the system switches services to other adjacent healthy server nodes. The reliability is close to 100%, and websites never breaks down.
- With CDN, customers can delivery content to global users without worrying about server investments, subsequent hosting and O&M, image synchronization between servers, or O&M personnel. CDN helps customers save human, energy, and financial resources.
- CDN enables customers to stay focused on their core services. CDN vendors deliver one-stop services, including content delivery, cloud storage, big data, and video cloud services. In addition, CDN vendors provide 24/7 O&M and monitoring to ensure network connectivity at any time.

CDN

- Content Delivery Network (CDN) delivers content from origin servers to edge nodes nearer the users who want it. CDN prevents Internet congestion, speeds up website response, and ensures service availability.

Abundant nodes

- 2,000+ nodes (Chinese mainland) and 800+ nodes (elsewhere)
- ≥ 150 Tbit/s bandwidth network-wide
- Edge nodes are deployed on networks of smaller carriers and top carriers China Telecom, China Unicom, China Mobile, and China Education and Research Network (CERNET). CDN precisely schedules user requests to the most appropriate node for efficient and reliable acceleration.

Intelligent scheduling

- The IP address database allows CDN to schedule at 99.99% success rate.
- Net Turbo technology allows CDN to schedule user requests to the best quality nodes based on node load.

Security

- Huawei Cloud CDN provides secure and reliable content delivery services.
- Advanced network security functions include data transmission over HTTPS and URL validation throughout the entire network.

Easy operation

- Simple and quick domain access to Huawei Cloud CDN
- Customizable domain configuration items: URL validation, cache policies, and HTTPS certificates
- Statistics and logs for easy analysis

Diverse applications

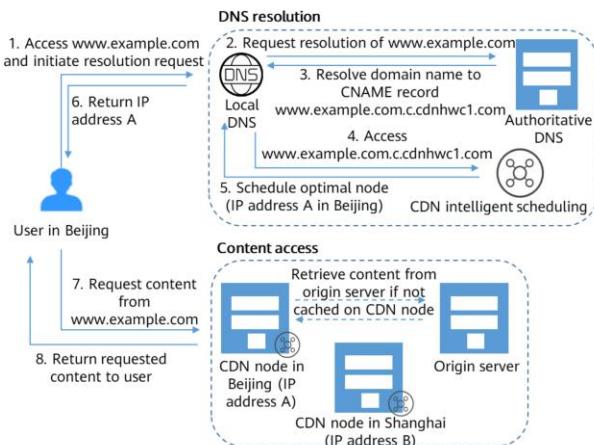
- Website, download, on-demand service, and whole site acceleration
- One-stop acceleration solutions for multiple scenarios improve the overall user experience



- Huawei Cloud CDN caches origin content on edge nodes across the globe. When a user accesses the content, the user does not need to retrieve it from the origin server. Based on a group of preset policies (including content types, geological locations, and network loads), CDN provides the user with the IP address of a CDN node that responds the fastest, enabling the user to obtain the requested content faster than would have otherwise been possible.
- Huawei Cloud CDN has over 2,000 edge nodes in the Chinese mainland and over 800 edge nodes outside the Chinese mainland. The network-wide bandwidth is at least 150 Tbit/s. Edge nodes are deployed on 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. Up to now, Huawei Cloud CDN covers more than 130 countries and regions, connecting to over 1,600 carrier networks. CDN precisely schedules user requests to the most appropriate node for efficient and reliable acceleration.

How Does CDN Work?

- CDN distributes content from origin servers to nodes close to users.



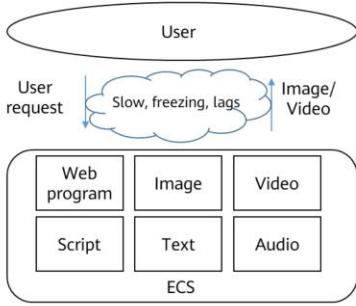
HTTP request process

- A user enters the domain name of a website to be accessed in the browser. A DNS request is sent to the local DNS server.
- 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.
- The authoritative DNS resolves the domain name and finds that the domain name points to the CNAME record of the domain name.
- The request is directed to the CDN service.
- CDN performs intelligent domain resolution and provides the user with the IP address of the Beijing CDN node, which responds the fastest.
- The user's browser obtains the IP address of the Beijing CDN node.
- 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.

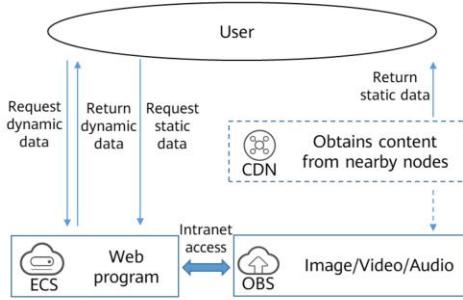


Static Content Acceleration

- In traditional architecture (dynamic and static data are not separate), website performance bottlenecks with increasing number of users, affecting user experience.



Traditional architecture



Website architecture that separates static and dynamic data

Anti-tampering

- Hypertext Transfer Protocol Secure (HTTPS) secures transmission through encryption and identity authentication.

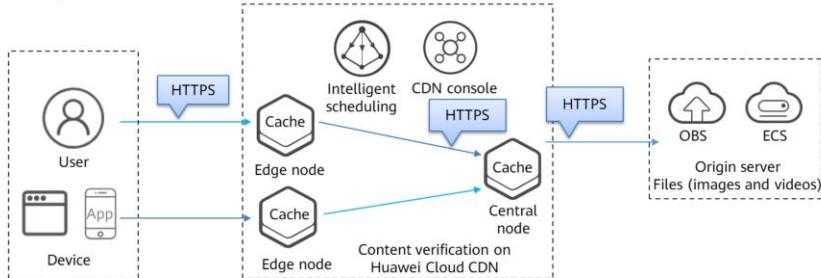
All links support HTTPS against content hijacking and tampering, and improve user experience.

- Customers purchase certificates on Huawei Cloud.

- Certificates stay secure by automatic push from the certificate management platform to CDN nodes.

Huawei CDN content verification

- The unique file ID is used to check whether the content is tampered with. If the content is tampered with, the system retries immediately.



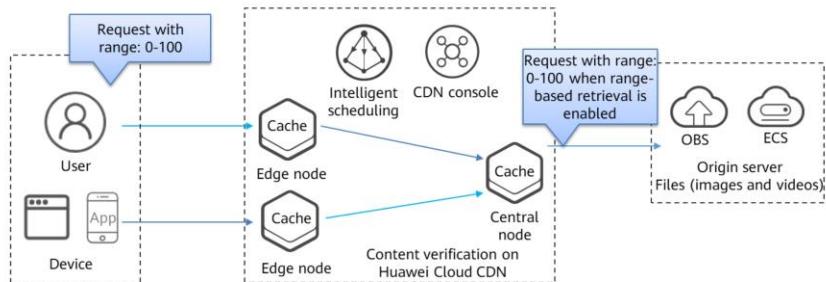
- CDN is widely used in security-sensitive communications on the World Wide Web, such as online payment.

Range-based Retrieval

- The origin server sends a specific range of data to a CDN node using the range information in the HTTP request header.

Lighter traffic and faster response

- The central node retrieves a range of content from the origin server and then sends the data to users, speeding up the response.
- Whole files do not need to be downloaded each time content retrieval is performed, reducing retrieval traffic.



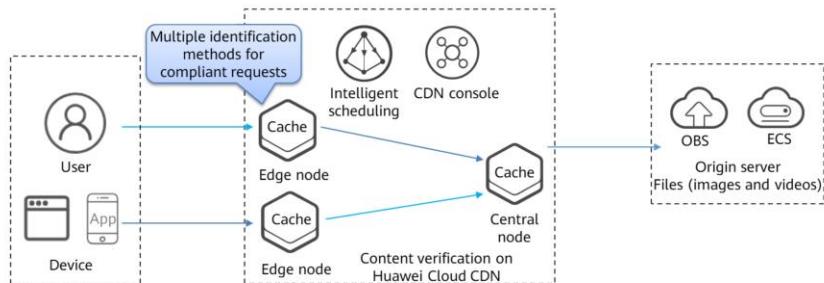
- Range information specifies the positions of the first and last bytes for the data to be returned. For example, Range: bytes=0-100 indicates that the first 101 bytes of the file are required.
- Range-based retrieval shortens the distribution time of large files, improves retrieval efficiency, and reduces content retrieval consumption.

URL Validation

- Customers configure a referer blacklist or whitelist to filter requests by referer value in request headers.

Multiple validation mechanisms prevent unauthorized requests while accelerating authorized content access.

- IP address and referer blacklist/whitelist
- Timestamp validation
- Custom validation for VIP customers



Use Case: Faster File Downloads from OBS Buckets

- Background: More and different enterprises use OBS buckets to store static resource files (images, videos, and software). OBS buckets are also the storage source for their services. OBS buckets are the solution to insufficient local storage. However, users accessing OBS buckets in other regions experience varying speeds, since files tend to be stored in only one region. Whenever frequent access is required, OBS bucket access consumes a large amount of traffic.



- If CDN is enabled for an OBS bucket, CDN nodes cache content in the OBS bucket and return content to users when requested. The cost of CDN is low. CDN acceleration reduces bandwidth costs by 50% to 57%.
- Huawei Cloud CDN has abundant acceleration resources and widely distributed nodes. It ensures that user requests are precisely scheduled to the optimal node, providing effective and stable acceleration.

Contents

1. Storage Service Overview
2. Storage Service Planning
3. Content Delivery Network
- 4. Backup Solution Planning**
5. DR Solution Planning

Why We Need to Back Up Data

- Cloud Backup and Recovery (CBR) provides redundancy and security for enterprises that want to maintain important data onsite or ensure that data is available in the event of a failure.

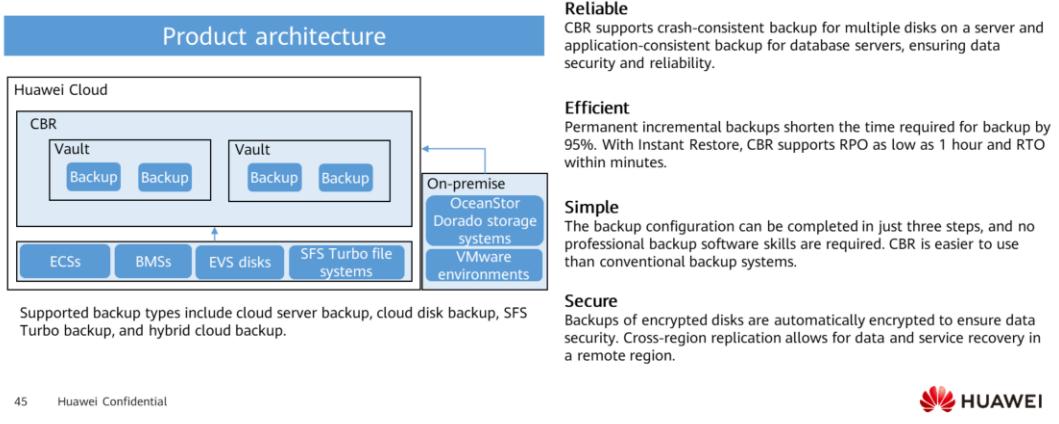
Key Threats to Data Security



- Data is the lifeline of enterprises.
- Data security threats are everywhere.
- Data protection is required by many laws and regulations.

CBR

- Cloud Backup and Recovery (CBR) enables you to easily back up ECSs, BMSs, EVS disks, and on-premises VMware virtual environments. If there is a virus attack, accidental deletion, or software or hardware failure, you can restore data to any point in the past when the data was backed up. CBR protects services by ensuring the security and consistency of data. CBR consists of backups, vaults, and policies.



- Backups:
 - A backup is a copy of a particular chunk of data and is usually stored elsewhere so that it can be used to restore the original data in the event of data loss.
- Vaults:
 - CBR uses vaults to store backups. Before creating a backup, you need to create at least one vault and associate the resource you want to back up with the vault. Then generated resource backups are stored in the associated vault.
 - Vaults can be either backup vaults or replication vaults. Backup vaults store resource backups, whereas replication vaults store replicas of backups.
 - The backups of different types of resources must be stored in different types of vaults.
- Policies: consist of backup policies and replication policies.
 - Backup policies: To perform automatic backups, configure a backup policy by setting the execution times of backup tasks, the backup frequency, and retention rules, and then apply the policy to a vault.
 - Replication policies: To automatically replicate backups or vaults, configure a replication policy by setting the execution times of replication tasks, the replication frequency, and retention rules, and then apply the policy to a vault. Backup replicas must be stored in replication vaults.

Differences Between Types of Backups

- CBR supports the following types of backups: cloud server backup, cloud disk backup, SFS Turbo backup, and hybrid cloud backup. The following table describes their differences.

Dimension	Cloud Server Backup	Cloud Disk Backup	SFS Turbo Backup	Hybrid Cloud Backup
Backup and restoration object	All disks (system and data disks) on a server	One or more specified disks (system or data disks)	SFS Turbo file systems	Backups synchronized from on-premises backup software and VMs
Recommended scenario	An entire cloud server needs to be protected.	Only data disks need to be backed up, because the system disk does not contain users' application data.	Data in the SFS Turbo file systems needs to be protected.	Backups for on-premises servers need to be managed and restored in the cloud.
Advantages	All disks on a server are backed up at the same time, ensuring data consistency.	Backup cost is reduced without compromising data security.	Backup data and file system data are stored separately. You can use the backup data to create new file systems.	On-premises data can be backed up to the cloud and used to re-build services in the cloud.

- The following are the types of CBR backups:
 - Cloud disk backup. This type of backup provides snapshot-based data protection for EVS disks.
 - Cloud server backup. This type of backup uses the consistency snapshot technology for disks to protect data of ECSs and BMSs. The backups of servers without deployed databases are common server backups, and those of servers with deployed databases are application-consistent backups.
 - SFS Turbo backup. This type of backup protects data of SFS Turbo file systems.
 - Hybrid cloud backup. This type of backup protects data of on-premises OceanStor Dorado storage systems and VMware VMs by storing their backups on the cloud. You can manage the backups on CBR Console.

CBR Backup Options

- CBR supports one-time backups and periodic backups. A one-time backup task is manually created by users and is executed only once. Periodic backup tasks are automatically executed based on a user-defined backup policy.

Item	One-Off Backup	Periodic Backup
Backup policy	Not required	Required
Number of backup tasks	One manual backup task	Periodic tasks driven by a backup policy
Backup name	User-defined backup name, which is manualbk_xxxx by default	System-assigned backup name, which is autobk_xxxx by default
Backup mode	An initial full backup and then incremental backups subsequently, by default	An initial full backup and then incremental backups subsequently, by default
Application scenario	Executed before an OS patch or upgrade or an application upgrade on a resource. A one-time backup can be used to restore the resource to the original state if the patch or upgrade fails.	Executed for routine maintenance of a resource. The latest backup can be used for restoration if an unexpected failure or data loss occurs.

- Two backup options can also be used together if needed. For example, users can associate all servers or file systems with a vault and then apply a backup policy to the vault for periodic backups, and manually perform one-time backups for the most important servers or file systems to further ensure data security.

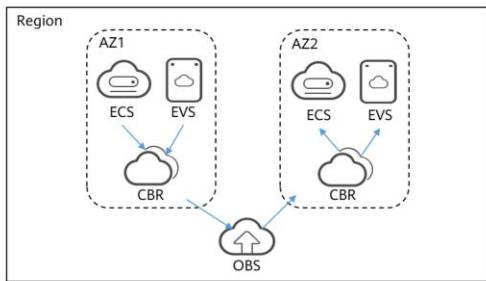
Differences Between Backups and Snapshots

- Both backups and snapshots provide data redundancy for improved reliability. The following table lists the differences between them.

Item	Storage Solution	Data Synchronization	DR Range	Service Recovery
Backup	Backups are stored in OBS, instead of on disks, and can be used to restore data even when the disk is damaged.	A backup is a copy of a disk at a given point in time. Cloud disk backup lets you configure automatic backups with backup policies. Deleting a disk will not delete its backups.	A backup is in the same AZ as its source disk. Cloud server backup supports cross-region replication.	Data can be recovered and services can be restored by restoring the backup data to source disks or creating new disks from backups, ensuring excellent data reliability.
Snapshot	Snapshot data is stored with disk data.	A snapshot records the state of a disk at a specific point in time. If a disk is deleted, all the snapshots created for this disk will also be deleted. If you reinstall or change the server OS, snapshots of the system disk are automatically deleted, but snapshots of the data disks are unaffected.	The snapshot is stored in the same AZ as its source disk.	Data can be recovered and services can be restored by rolling back the snapshot data to source disks or creating new disks from snapshots.

Application Scenarios

- CBR cloud server backup and cloud disk backup allow you to create backups for data restoration in the event of a software or hardware failure or accidental deletion. They maximize data security and integrity, and ensure service security.
- CBR provides backup and restoration services for cloud servers, cloud disks, and on-premises virtual environments, ensuring service security and reliability.



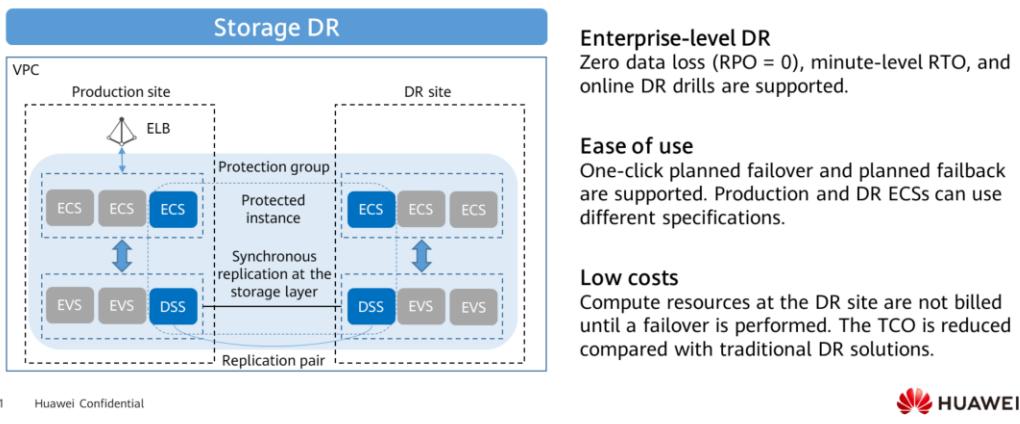
- Cloud server backup is recommended if the whole server, including server configurations and specifications as well as the data on its disks, needs to be protected, or if the users want to quickly replicate the service environment by creating servers from backups.
- Cloud disk backup is recommended if the system disks do not have user-defined data. This way, only the data disks are backed up to keep the costs down.

Contents

1. Storage Service Overview
2. Storage Service Planning
3. Content Delivery Network
4. Backup Solution Planning
5. DR Solution Planning

SDRS

- Storage Disaster Recovery Service (SDRS) provides disaster recovery (DR) for cloud services such as ECS, EVS, and DSS. By leveraging various techniques, including storage replication, data redundancy, and cache acceleration, SDRS can provide you with high data reliability and service continuity.



51 Huawei Confidential



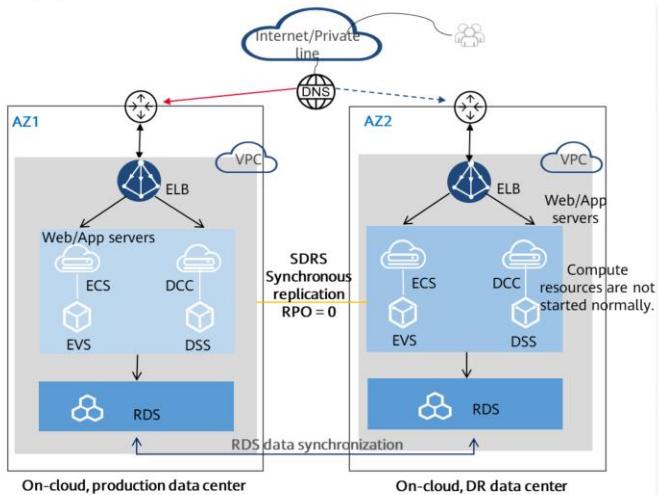
- RPO = 0: With the accumulation of 8+ years of Huawei-developed technology, the storage-layer, synchronous replication ensures zero data loss.
- Minute-level RTO: If a disaster occurs, a failover can be completed within minutes.
- Online DR drill: DR drills can be performed at any time when needed to verify the feasibility and effectiveness of the DR solution.
- Three-step DR: Cloud DR can be completed in only three steps: creating a protection group, creating protected instances, and enabling protection.
- One-click DR switchover: SDRS supports one-click DR switchover. After a switchover is complete, services can be quickly recovered by manually starting the ECSs.
- Workload-level protection: SDRS supports DR protection by workload, that is, adding ECSs running the same workload to the same protection group.
- No additional plug-ins: The deployment is simply. No additional plug-in is required on DR site ECSs.
- No fees for DR site ECSs: If everything is working normally, ECSs at the DR site are not started and are not charged.
- TCO reduced by 60%: SDRS saves the DR TCO by 60% compared with the traditional DR solutions, reducing the costs in hardware devices, power supply, O&M, and more.
- Automatic network migration: After a switchover is complete, the IP address, MAC address, and EIP of an ECS are automatically migrated, freeing you from reconfigure them again.

Differences Between SDRS and CBR

- A DR plan using SDRS+CBR is suitable for scenarios where ECSs and EVS disks are used to run services. But there are many differences between DR and backup in terms of data protection, such as the application scenarios and service recovery time. The following table describes their differences.

Item	DR	Backup
Application scenario	DR is used to protect services from natural disasters, such as a fire and an earthquake. A DR site must be located within a safe distance from the production site.	Backups provide protection from human error, viruses, and logic errors. They are used to restore the service system data. Usually, a system and its backup data are deployed in the same data center.
Service continuity	A DR system protects data but is more focused on guaranteeing service continuity.	A backup system only ensures that backups generated at different points in time can be restored. Usually, the system first performs an initial full backup, which takes a long period of time. Subsequent backups are incremental backups and can be done quicker.
Timeliness	The highest-level DR can achieve zero RPO.	A maximum of 24 executions in a day can be set in a backup policy, and the backups generated accordingly can be used to restore data to different points in time.
Service recovery	If a disaster occurs, such as an earthquake or a fire, a DR system only takes a few seconds to perform a failover.	A backup system takes several hours or even dozens of hours to restore data.

Application Scenario – Cross-AZ DR



Application Scenario

- Suitable for customers requiring on-cloud, intra-city DR and zero RPO

Solution Architecture

- User access traffic:** User traffic is controlled by the DNS service. If everything is working normally, 100% of the user traffic is directed to the production center. If there is a disaster, 100% of the user traffic is directed to the DR center.
- Web/App server data synchronization:** Web and application servers synchronize data using synchronous replication provided by SDRS. The RPO is 0, and the RTO is less than 30 minutes. Normally, compute resources at the DR site are not started.
- Database synchronization:** RDS databases are deployed in primary/standby mode across AZs to synchronize data.
- DR switchover:** If the production site fails, RDS database services automatically switch to the standby databases, application services can be taken over by the DR servers in just a few clicks, and all of the user traffic can be directed to the DR site using DNS.
- DR drill:** Users can use SDRS to perform DR drills with just a few clicks.

Solution Highlights

- Cross-AZ DR, RPO = 0, and RTO < 30 minutes
- Compute resources at the DR site are not started to reduce costs, except in the event of a disaster.
- One-click DR switchover and drills are available to simplify management.



Quiz

1. (Multiple-answer question) EVS disks differ in performance. Which of the following are EVS disk types?
 - A. Extreme SSD
 - B. Ultra-high I/O
 - C. General Purpose SSD
 - D. High I/O
2. (Multiple-answer question) Which of the following are basic components of OBS?
 - A. Buckets
 - B. Objects
 - C. Data
 - D. Metadata

- ABCD
- AB

Quiz

1. (Discussion) What are the advantages of cloud storage over on-premises, self-built storage?

2. (Discussion) What should be considered for using cloud storage in terms of security, cost, reliability, performance, and scalability?

- Discussion 1:
 - Cloud vendor offers management.
 - Self-built storage is managed by users.

- Discussion 2:
 - Security: Static data is protected using encryption, and dynamic data is protected using HTTPS access. If data is stored in OBS, data in the bucket can be protected by bucket policy.
 - Cost: Based on the data usage, select an appropriate type of storage to reduce storage costs.
 - Reliability: Determine whether storage redundancy is required based on service requirements. If redundancy is not used, cross-region replication can be used to back up data.
 - Performance: In large service volume scenarios, considering the high I/Os, SSD-based disk types are recommended.
 - Scalability: Determine appropriate storage classes based on the usage of data to be stored. Capacity expansion and reduction methods may vary with storage classes.

Summary

- Where there is data, there is a need for data storage. After studying the content presented here, we should have a new understanding of storage types and we should understand Huawei Cloud storage services a little better. As more and more enterprises migrate to the cloud, we are more able to better meet their storage requirements if we understand the positioning, principles, and usages of various storage services, for example, which storage services are suitable for video cloud and which are the best for databases.

Acronyms and Abbreviations

- API: Application Programming Interface
- AS: Auto Scaling
- BMS: Bare Metal Server
- CBR: Cloud Backup and Recovery
- CDN: Content Delivery Network
- DCS: Distributed Cache Service
- DRS: Data Replication Service
- DNS: Domain Name Service
- DDoS: Distributed Denial of Service
- DevOps: Development and Operations
- DIS: Data Ingestion Service
- DLI: Data Lake Insight
- EIP: Elastic IP
- ECS: Elastic Cloud Server
- ELB: Elastic Load Balance
- EVS: Elastic Volume Service
- GSLB: Global Server Load Balance
- HA: High Availability
- IMS: Image Management Service
- IDC: Internet Data Center
- LTS: Log Tank Service

Acronyms and Abbreviations

- NAT: Network Address Translation
- OLAP: Online Analytical Processing
- OLTP: Online Transaction Processing
- RDS: Relational Database Service
- SMN: Simple Message Notification
- SFS: Scalable File Service
- SDRS: Storage Disaster Recovery Service
- VM: Virtual Machine
- VPC: Virtual Private Cloud
- VPN: Virtual Private Network

Recommendations

- Huawei iLearning
 - <https://e.huawei.com/en/talent/portal/#/>
- Huawei Cloud Help Center
 - <https://support.huaweicloud.com/intl/en-us/index.html>
- HUAWEI CLOUD Developer Institute
 - <https://edu.huaweicloud.com/intl/en-us/>
- Huawei Talent Online
 - <https://e.huawei.com/en/talent/portal/#/>

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。

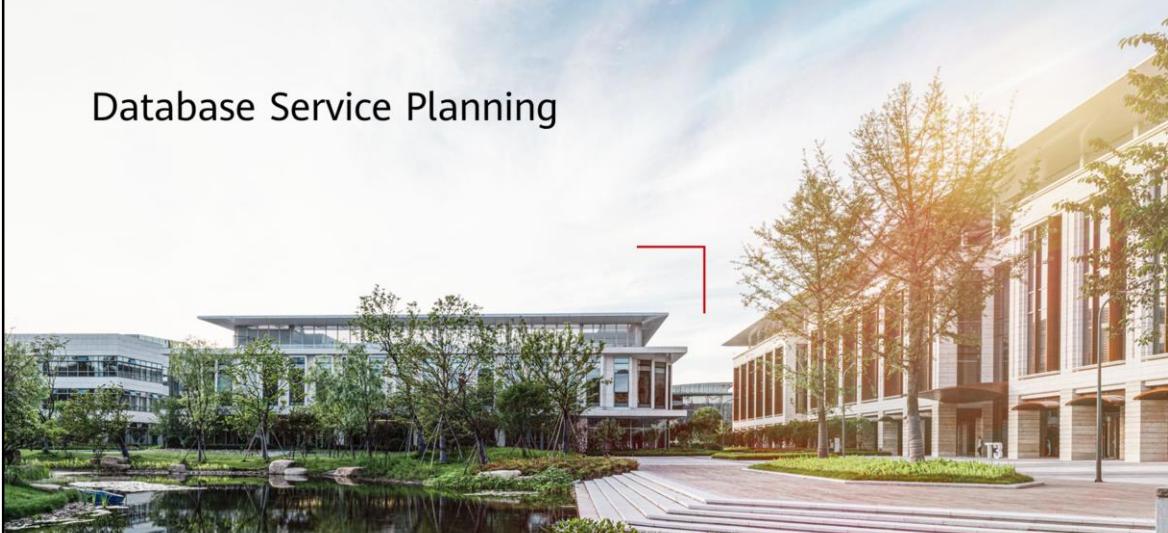
Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

Copyright©2022 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.



Database Service Planning



Foreword

- In addition to compute, networking, and storage services, enterprise customers need other services like database services and security services. These services can be billed on a pay-per-use basis and are easy to maintain, helping customers reduce investment and facilitate O&M.
- This course describes database services.

Objectives

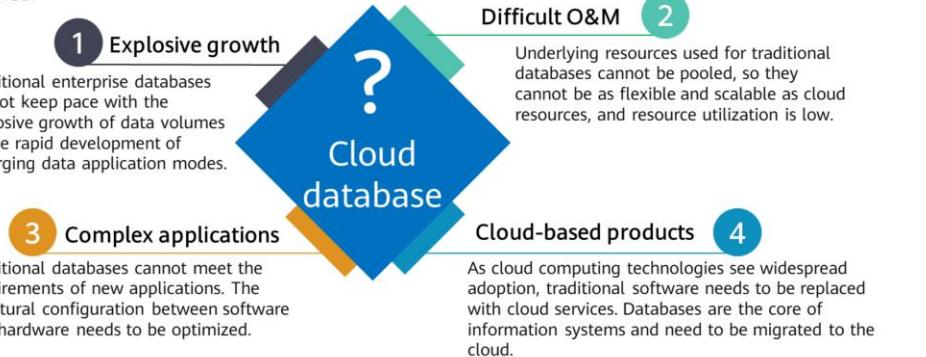
- Upon completion of this course, you will:
 - Understand common database services on Huawei Cloud.
 - Understand database service types, their application scenarios, and their related services.

Contents

- 1. Introduction to Database Services**
2. Cloud Database Services
3. Database Migration

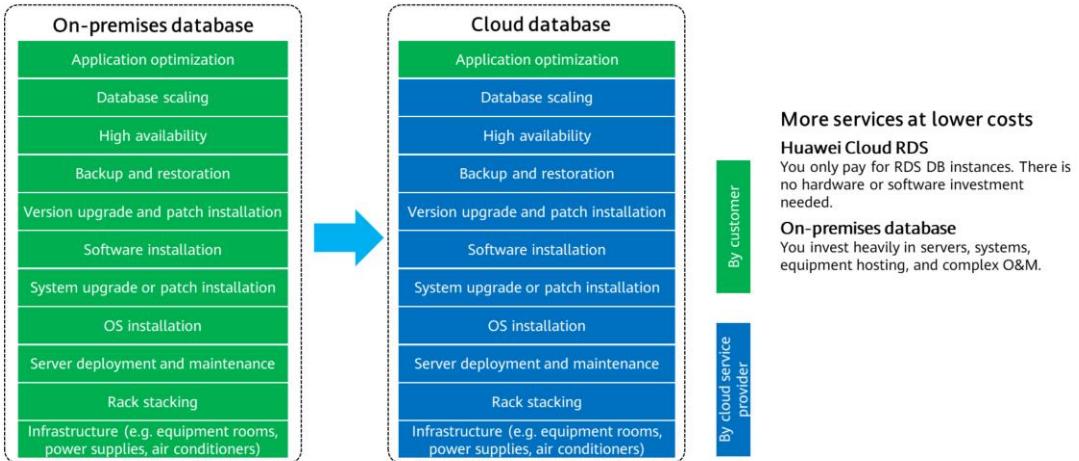
Database Development Trends

- Cloud computing technologies have driven the rapid development of database markets. Cloud databases have been favored by increasingly more users thanks to their pay-per-use billing models. Many enterprises have migrated applications to the cloud. They can choose cloud databases or cloud-native databases to keep up with fast business growth and the urgent need for intelligent maintenance.



- Enterprises are facing explosive data growth and ever more diverse types of data applications. Large-scale cloud transformation has been changing traditional business models.

Advantages of Cloud Databases



5 Huawei Confidential



- Compared with traditional databases, cloud databases have the following advantages:
 - Ease-of-use: Each cloud database is provided as a cloud service. You can easily create and run it on the cloud.
 - Scalable: Each cloud database service is an open-source database with decoupled storage and compute for more flexible scaling.
 - Cost-effective: Compared with a traditional database, using a cloud database service saves money on software and hardware, and pay-per-use billing helps you reduce the total cost of ownership (TCO).

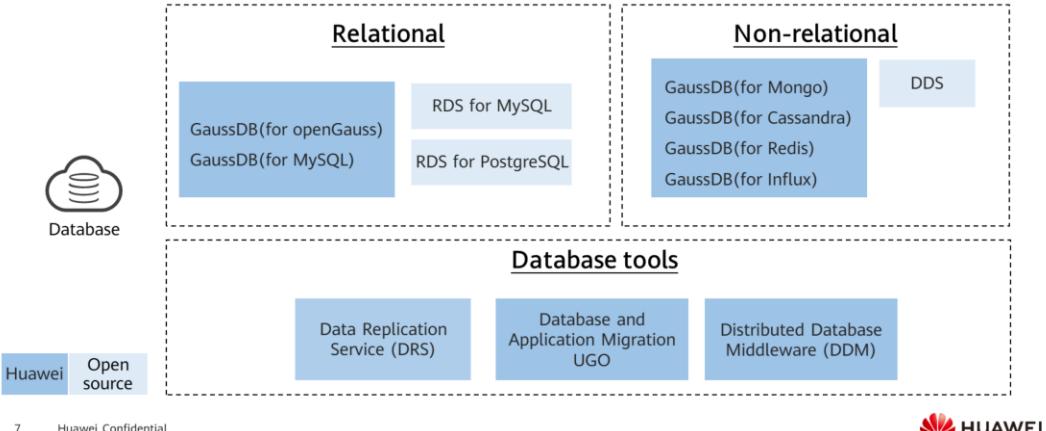
Database Categories: SQL and NoSQL

Comparison Item	Relational Database (SQL)	Non-Relational Database (NoSQL)
Database type	MySQL, SQL Server	MongoDB, Redis
Storage structure	Structured tables	Semi-structured dataset
Storage characteristics	Standardized logical tables avoid repetition and save space. Data operations involve multiple tables, making data management complex.	Data is stored in data sets. There may be redundant data stored, but data read and write operations are easier.
Scalability	Weak	Strong
Transaction support	Good	Not supported

- A relational database organizes data using a relational model. A relational model is a two-dimensional table model, and a relational database is a data organization consisting of two-dimensional tables and their relationships.
- A non-relational database is a non-relational, distributed data storage system that does not comply with ACID properties.
- Typical products:
 - Relational databases: SQL Server, MySQL, and PostgreSQL
 - Non-relational databases: Redis, Memcached, and MongoDB

Huawei Cloud Database Portfolio

- GaussDB is an open-source database designed for small and medium enterprises to achieve superior cost-effectiveness. GaussDB meets the reliability and performance requirements of government, enterprise, and Internet customers.



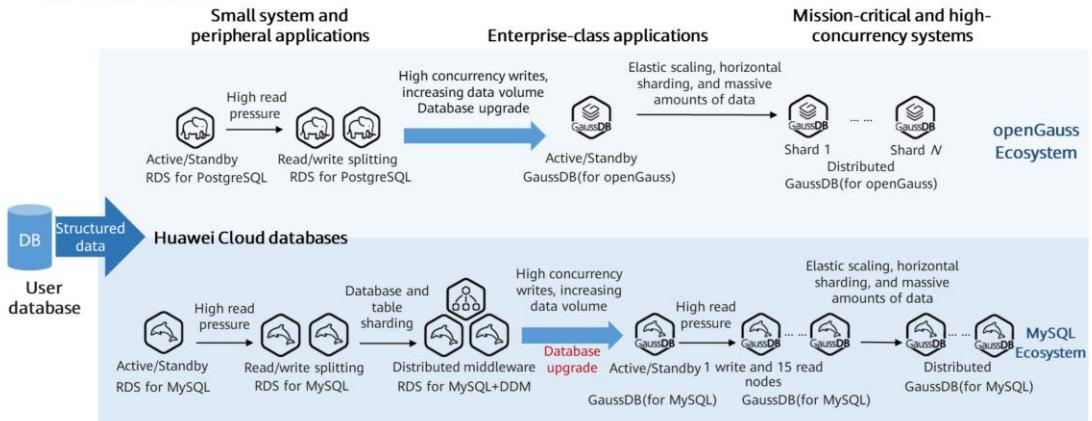
- Databases are classified as either relational databases or non-relational databases.
- Huawei Cloud relational database services include RDS for MySQL, RDS for PostgreSQL, RDS for SQL Server, GaussDB(for openGauss), and GaussDB(for MySQL).
- Huawei Cloud non-relational database services include GaussDB(for Mongo), GaussDB(for Cassandra), GaussDB(for Redis), GaussDB(for Influx), DDS, DCS.
- Database ecosystem services include DDM, DRS, and UGO.
- Distributed Database Middleware (DDM) breaks through the capacity and performance bottlenecks that plague traditional databases and addresses distributed scaling issues so you can handle highly concurrent access to massive volumes of data.
- DDM uses decoupled storage and compute. It provides functions such as database and table sharding, read/write splitting, elastic scaling, and sustainable O&M. Management of instance nodes has no impact on your workloads. You can perform O&M on your databases and read and write data from and to them on the DDM console, just like as operating a single-node MySQL database.
- Advantages: automatic database and table sharding, read/write splitting, and elastic scaling

Contents

1. Introduction to Database Services
2. **Cloud Database Services**
 - Relational Database Services
 - Non-Relational Database Services
3. Database Migration

Design Principles for SQL Databases

- If the source database is a relational database and your application scale is small, you can migrate your database to Huawei Cloud RDS. When the data amount increases significantly and your database struggles to handle high concurrency, you can migrate workloads to GaussDB.



9 Huawei Confidential

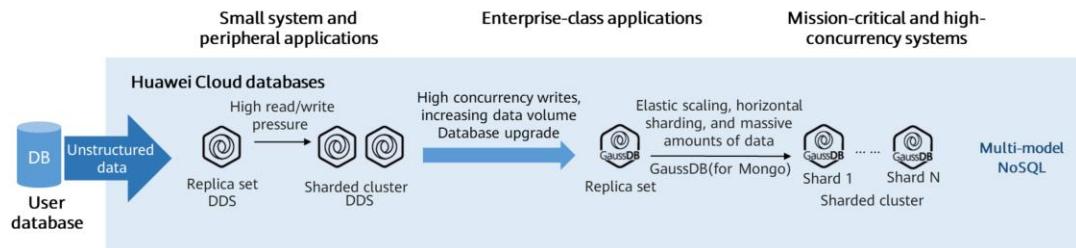


- Scenarios:

- Small systems and peripheral applications: 100,000 QPS, small-scale OLTP, and tens to hundreds of GB of data
- Enterprise-class applications: millions of queries per second, medium-scale OLTP, and terabytes, or even dozens of terabytes of data
- Mission-critical and high-concurrency systems: ultra-large OLTP, OLTP/OLAP, cloud-native distributed, dozens of terabytes of data

Design Principles for NoSQL Databases

- If the source database is a non-relational database and your application scale is small, you can migrate your database to Huawei Cloud DDS. If the amount of data increases significantly and your database struggles to handle high concurrency, you can migrate workloads to GaussDB.



RDS

- Relational Database Service (RDS) is a cloud-based web service that is reliable, scalable, easy to manage, and immediately ready for use. RDS provides a comprehensive performance monitoring system, multi-level security protection measures, and a professional database management platform, allowing users to set up and scale databases with ease.

Reliable	Secure	Excellent Performance	Cost-Effective
<ul style="list-style-type: none">RDS uses a hot standby architecture. If there is a fault, a failover just takes only a few seconds.RDS automatically backs up your data daily and can retain backups for up to 732 days. Using the RDS console, you can restore data from backup in just a few clicks.You can restore data from backups to any point in time.	<ul style="list-style-type: none">VPCs and security groups are used to isolate networks.RDS controls access with IAM users and security groups.TLS and SSL are used to ensure data security during transmission.RDS uses static encryption and tablespace encryption to encrypt the data to be stored.When an RDS DB instance is deleted, all data stored in the instance is also deleted.	<ul style="list-style-type: none">RDS offers high-performance database services backed by years of database R&D experience.RDS uses servers that have been proven time and again.SQL Tuning helps detect slow SQL queries and provides you suggestions on how to adjust your code.You can deploy RDS in the same region as your ECS to shorten the response time of your application.	<ul style="list-style-type: none">RDS can work with an ECS in the same region, communicating over an intranet connection. This minimizes costs by reducing Internet traffic.You can scale RDS DB instance specifications and storage if you identify a performance bottleneck detected by Cloud Eye.RDS is fully compatible with native database engines.You can easily manage your instance on the service console, performing actions such as rebooting the instance, resetting passwords, adjusting parameters, and viewing logs.

11 Huawei Confidential



- Security
 - Running a DB instance in a VPC improves security. You can configure subnets and security groups to control access to DB instances.
- Access control
 - When you create an RDS DB instance, an account is automatically created. To separate permissions, you can create IAM users and assign permissions to them as needed.
- Transmission encryption
 - You can download the Certificate Agency (CA) certificate from the console and upload it when connecting to a database for authentication.
- Storage encryption
 - RDS encrypts data before storing it. Encryption keys are managed by Key Management Service (KMS).
- Data deletion
 - Automated backup data and the data stored in the disks associated with your instance can be securely deleted. You can restore a deleted DB instance from a manual backup or rebuild the DB instance in the recycle bin during the retention period.

DB Engines supported by RDS

- RDS supports MySQL, PostgreSQL, and SQL Server engines.



RDS for MySQL



RDS for PostgreSQL



RDS for SQL Server

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

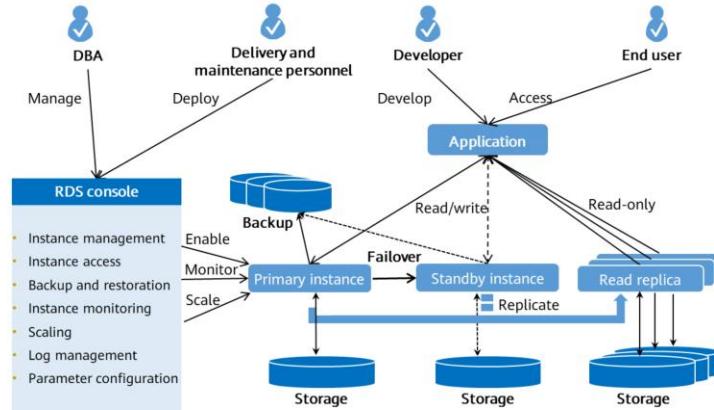
- PostgreSQL is designed for business-oriented online transaction processing (OLTP) scenarios and supports NoSQL (JSON, XML, or hstore) and geographic information system (GIS) data types. It has earned a solid, good reputation for reliability and data integrity.
- RDS for PostgreSQL is suitable for Internet websites, location application systems, complex data object processing and other application scenarios.

- SQL Server is a well-established commercial database with a mature enterprise-class architecture. It supports one-stop deployment and provides intelligent administration functions, helping reduce the costs of running and managing databases.
- RDS for SQL Server is widely used in government, finance, medical care, education, and the gaming industry.

- RDS for MySQL
 - It uses a stable architecture and supports a wide range of web applications. It is cost-effective and often preferred by small and medium enterprises.
 - A web-based console is available for you to monitor performance metrics so if there is an issue, you can identify it and take appropriate measures as soon as possible.
 - You can flexibly scale resources to meet business needs and pay for only what you use.
- RDS for PostgreSQL
 - RDS for PostgreSQL supports the postgis plugin and provides excellent spatial performance.
 - RDS for PostgreSQL is a cost-effective solution suitable for many business scenarios. You can flexibly scale resources based on your needs and pay for only what you use.
- RDS for SQL Server
 - RDS for SQL Server is reliable, scalable, inexpensive, and easy to manage. It supports high availability for your applications with automatic database failover that completes within several seconds. It also provides multiple options for backing up your data.

RDS for MySQL

- RDS for MySQL is reliable, scalable, inexpensive, easy to manage, and ready for use out-of the box, so you can stay focused on developing your services.



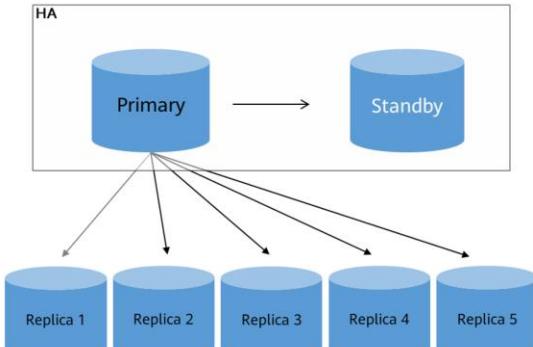
13 Huawei Confidential



- Database engine versions: MySQL 5.6, 5.7, and 8.0
- Data security: Multiple security policies protect databases and data privacy.
- Database reliability: Three-copy data storage ensures up to 9 nines of database data reliability and up to 11 nines of backup data reliability.
- High availability (intra-city disaster recovery): Primary/standby DB instances are deployed within an AZ or across AZs, ensuring service availability over 99.95%.
- Instance access: Multiple access methods are supported. You can use floating IP addresses, public IP addresses, or VPNs.
- Instance management: You can add, delete, modify, query, and reboot your DB instance on the console.
- Elastic scaling: Horizontal scaling: Read replicas can be created (up to five for each instance) or deleted. Vertical scaling: DB instance classes can be modified and storage space can be scaled up to 10 TB.
- Backup and restoration:
 - For backup, there are automated backup, manual backup, full backup, and incremental backup. Backups can be added, deleted, queried, or replicated.
 - For restoration, data can be restored to any point in time within the backup retention period, or to a new or an original DB instance. The backup retention period is up to 732 days.

Cross-AZ HA

- If a primary instance fails, workloads on it can be failed over to the standby instance within several seconds.



14 Huawei Confidential



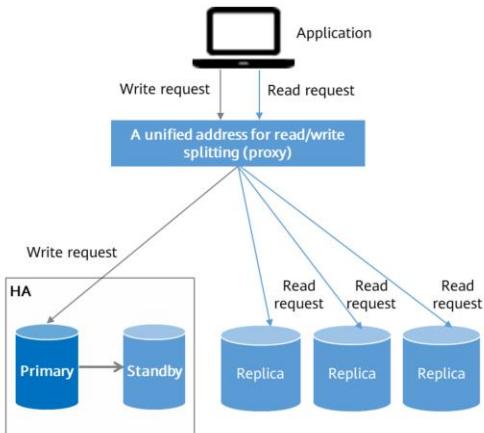
Functions and constraints

- Cross-AZ HA supports switchover in seconds.
- Up to 5 read replicas can be created to take over read traffic from the primary instance.
- The standby instance is invisible to the end user. Only its Virtual IP (VIP) is displayed.
- Read replicas cannot exist alone. They have to be paired with single or primary/standby DB instances.

- When creating a DB instance, you can select **Primary/Standby** as the instance type. If a primary instance fails, RDS automatically switches to the standby instance. If the standby instance also fails, a primary/standby instance in another AZ will automatically take over the workloads.
- Each RDS DB instance supports up to five read replicas and can scale out with Distributed Database Middleware (DDM) to further increase capacity. Write requests are routed to the primary instance and read requests are routed to read replicas.
- The primary and standby DB instances share the same virtual IP address (VIP) for communication with external systems. The DB instance associated with the VIP is the primary instance. If the primary instance is unavailable, RDS automatically associates the VIP with the standby instance and promotes it to be the new primary instance. Associating the VIP with the standby instance can be completed in seconds. There is no downtime. The switchover is imperceptible to users.
- Constraints: You can create read replicas only after purchasing a DB instance.

Read/Write Splitting

- Read/write splitting helps RDS distribute your requests to DB instances faster.



15 Huawei Confidential

Functions

- RDS provides a single read/write splitting address, which is transparent to applications.
- Read-only permissions can be configured for each node.
- If a health check detects that a read replica has failed or the I/O latency of a read replica has exceeded your preset threshold, read requests are no longer distributed to the read replica.

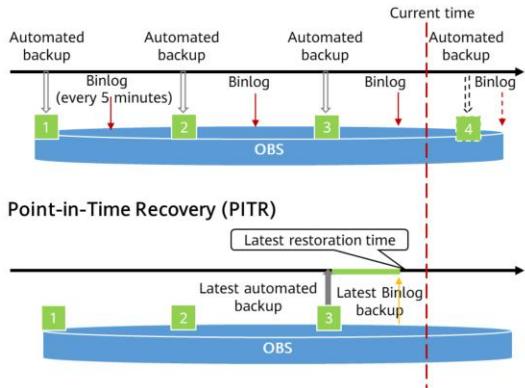


- After read replicas are created and read/write splitting is enabled for your DB instance, RDS will distinguish between read and write requests. Write requests are routed to the primary instance. Read requests are distributed to the read replicas.

Data Security

- You can restore data from backup to any point in time more than 5 minutes ago.

Full data backup + Binlog backup



Functions

- You can restore DB instances or database tables.
- You can configure the number of days that your automated backups can be retained.
- You can restore data to any point in time more than 5 minutes ago. You can restore data to a new DB instance or to your original DB instance.



- The automated backup retention period (1-732) is configurable.

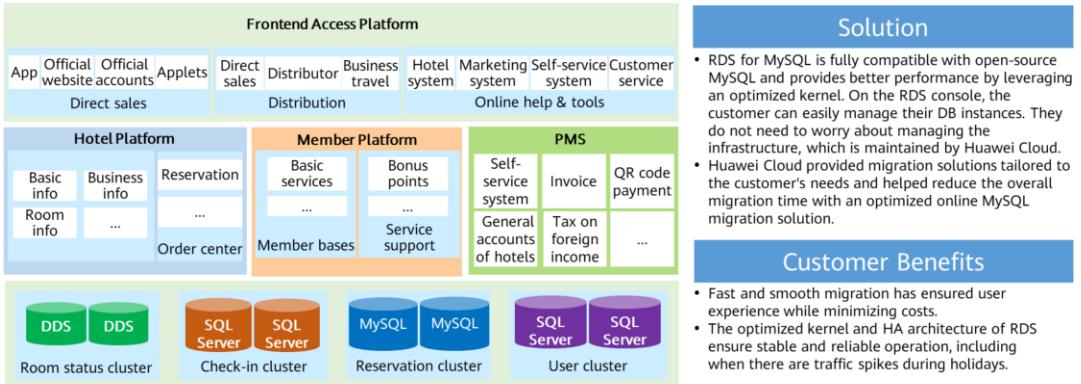
Kernel Optimization

- SQL statement concurrency control for RDS for MySQL instances restricts the execution of SQL statements during peak hours by controlling how many statements can be executed at the same time.

Feature	Description
No GTID restrictions	Enabling GTID for MySQL community edition databases comes with certain limitations. To work within these limitations, users have to change their applications when migrating the community edition databases to the cloud. Huawei RDS for MySQL addressed these issues by optimizing its kernel and is the first database in the industry to fully support GTID.
Thread pool	A thread pool function was introduced for the enterprise edition to support more concurrent connections and provide better performance.
Multi-threaded replication	Multi-threaded replication for transactions has been introduced in MySQL 5.6 and 5.7. It effectively reduces replication delays.
Converting MyISAM to InnoDB	To solve the problem of MyISAM not supporting transactions, RDS for MySQL automatically and transparently converts MyISAM tables to InnoDB.
Security control of system databases	Security controls are used to prevent users from deleting system databases by mistake or running commands that prevent instances from running normally.
MDL view	MDL information held or waited for by the thread can be obtained through metadata_lock_info.

Case: Smooth Migration of Hotel Companies to the Cloud

- A hotel using an open-source database system often encountered performance bottlenecks (slow queries and long wait times). The system had a large number of DB instances and O&M personnel were unable to handle the O&M workload. The hotel needed a high performance database system and wanted a professional migration solution to move their workloads to the cloud.



Solution

- RDS for MySQL is fully compatible with open-source MySQL and provides better performance by leveraging an optimized kernel. On the RDS console, the customer can easily manage their DB instances. They do not need to worry about managing the infrastructure, which is maintained by Huawei Cloud.
- Huawei Cloud provided migration solutions tailored to the customer's needs and helped reduce the overall migration time with an optimized online MySQL migration solution.

Customer Benefits

- Fast and smooth migration has ensured user experience while minimizing costs.
- The optimized kernel and HA architecture of RDS ensure stable and reliable operation, including when there are traffic spikes during holidays.



RDS for PostgreSQL

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

<u>Enhanced Features</u>	<u>Multilayer Network Security</u>	<u>Support for 70+ Plug-ins</u>	<u>Professional Database O&M Platform</u>
<ul style="list-style-type: none">• Up to five read replicas can be added to an existing DB instance to offload read-heavy database workloads.• Database proxies automatically distribute read and write requests through a read/write splitting address.• Tasks can be run automatically during off-peak hours based on a user-defined schedule.	<ul style="list-style-type: none">• Virtual Private Cloud (VPC) isolates tenant networks and security group rules are used to control traffic to and from specific IP addresses and ports, safeguarding your database.• TLS and SSL encrypt data during transmission, ensuring the security and integrity of your data.• Cloud Trace Service (CTS) records operations associated with RDS for PostgreSQL for later query, audit, and backtrack operations.	<ul style="list-style-type: none">• The PostGIS plugin is supported for 2D and 3D models, with space objects, indexes, operation functions, and operators.• The timescaledb time-series database plug-in, partition tables, and BRIN indexes are all supported.• A wide range of indexes are provided, including function- and condition-based indexes, for faster full-text search.	<ul style="list-style-type: none">• It is easy to manage your instances with flexible console-based capabilities.• You can view key operational metrics of your instances, including vCPU/storage utilization, I/O activity, and instance connections, and define alarm rules.• Point-in-time data restoration from backup is supported, and backups are saved for up to 732 days.• If an instance fails, your workloads do not, thanks to automated failovers.

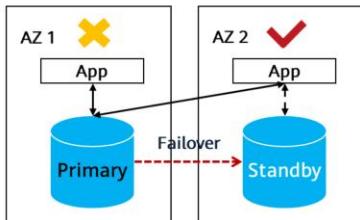
19 Huawei Confidential



- DB engine versions: 9.5, 9.6, 10.0, 11, and 12
- Security: Multiple security policies protect databases and data privacy.
- Data migration: There is online and offline migration to the cloud, to on-premises, and across clouds.
- HA: Data is automatically synchronized from a primary DB instance to a standby DB instance. If the primary DB instance fails, workloads are quickly and automatically switched over to the standby DB instance.
- Monitoring: Key performance metrics of RDS DB instances are monitored. These metrics include the CPU usage, memory usage, storage space usage, I/O activity, database connections, QPS, TPS, buffer pools, and read/write activities.
- Horizontal scaling: Read replicas (up to five for each instance) can be created or deleted. Vertical scaling: DB instance classes can be modified and storage space can be scaled without downtime.
- Backup and recovery: RDS supports automated and manual backups along with point-in-time recovery (PITR).

High Reliability and Availability

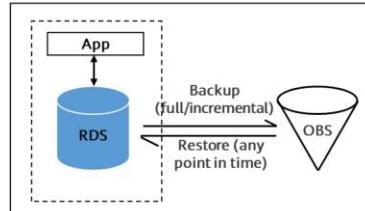
High availability



- You can choose to prioritize failovers for reliability or availability.
- Intra-AZ and multi-AZ HA and automatic failover are supported.
- You can manually switch primary to standby to simulate a fault.
- A read replica can automatically mount itself to a new primary node.
- A switchover can complete in seconds.
- The standby database does not bear traffic and ensures RTO.
- The Huawei-developed HA Monitor module is used.
- Virtual IP addresses can be switched, without any effect on applications.
- Multiple primary/standby switchovers can be performed.
- Automatic fault detection

20 Huawei Confidential

High reliability - recovery to any point in time



- Backup period: about two years
- Pay-per-use: free EVS storage space that equals to the requested storage and expansion without an upper limit
- High data reliability
- Security encryption: KMS encryption and multiple protection measures

OBS archive storage replaces tape libraries to restore data to any point in time.



- RDS for PostgreSQL supports cross-AZ HA. If the primary instance fails, the fault detection module attempts to start it three times. If the primary instance still cannot be started, a failover is automatically performed and completed within seconds. The standby instance is promoted to primary and read replicas are automatically associated with the new primary instance.
- RDS provides data backup and restoration. You can set an automated backup policy to back up your data daily. Automated backups can be retained for up to 732 days. An incremental backup is performed every 5 minutes for data consistency.
- If data is lost or deleted by mistake, you can restore the database to any point in time.
- Backup files are stored in OBS. OBS has no capacity upper limit and provides 99.99999999% data reliability.

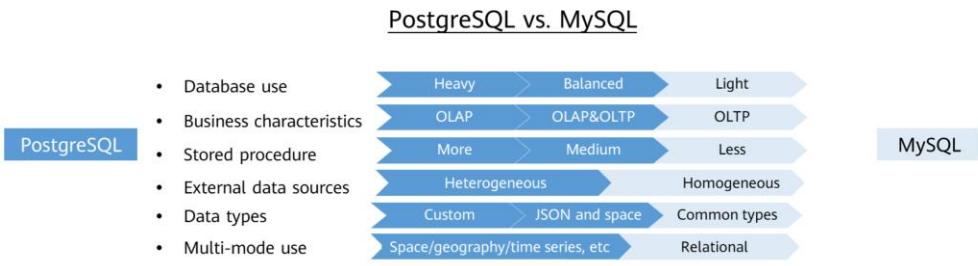
Database Selection

- PostgreSQL and MySQL have different engines and tools, but there is no single source for one of them is the best suited to all scenarios.
- Many organizations like PostgreSQL because it is an open-source relational database with a great reputation for reliability and data integrity. MySQL, however, can more flexibly meet many customers' needs.
- When selecting a database, organizations prioritize familiarity over features.

Don'ts	Dos
Don't just use what developers are familiar with.	No database is perfect for everything. Each type of database is designed to solve specific issues. You need to focus on whether the database is suitable for your business and may need to consider the dev-team's personnel preferences.
Don't just follow the leading Internet companies.	Choose a database for your application based on the industry and development phase your company is in.
Don't use too many different types of databases together, such as relational, document, buffer, time series, retrieval, and graph databases.	A multi-model database can be used instead, helping you reduce development and maintenance costs.
Don't select a database just for resolving current problems.	You also have to consider factors such as database code maturity, platform compatibility, open source license, technology reserves, the community ecosystem, learning curves, and the cost of maintenance.

Database Comparison

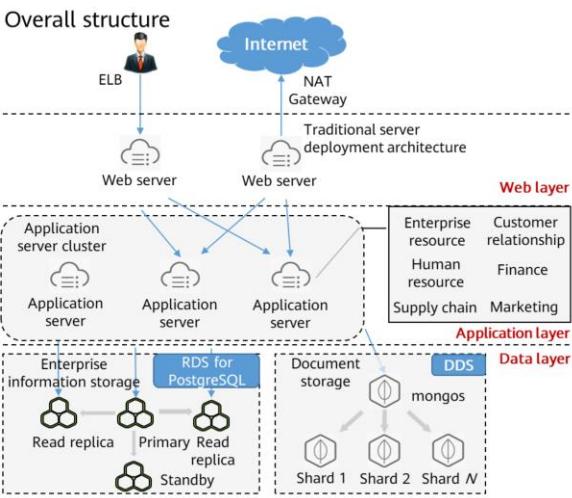
- The following figure compares PostgreSQL and MySQL. Both of them can meet business requirements in most cases.



- Database use:** If a database is only used as a data storage tool, both PostgreSQL and MySQL can be used. If many functions depend on database features, PostgreSQL is recommended.
- Business characteristics:** For transaction-heavy businesses, select either PostgreSQL or MySQL. For hybrid transaction processing, PostgreSQL is recommended.

- Note: Both of PostgreSQL and MySQL can be used in most scenarios.
 - When you are choosing a database, database use and design habits need to be considered. For example, some gaming and Internet companies just use databases to store data. Both PostgreSQL and MySQL are fine for this. But if many of your system's functions depend on more varied database features, PostgreSQL is recommended. It is a stable and reliable open-source database that is a good choice for many companies.
 - If your current DB system only processes transactions, choose a database using the same engine. If your database requires both transaction and analytic processing, PostgreSQL is recommended because it provides excellent analytical performance.
 - If many stored procedures are used, PostgreSQL is recommended. Use whatever your company is already used.
 - If your application has to access heterogeneous databases, PostgreSQL is recommended because it provides foreign data wrappers, which allows users to access heterogeneous data using SQL statements.
 - PostgreSQL is recommended for complex data types, such as complex arrays, spatial data, network data, JSON data, XML data, and certain custom types.
- PostgreSQL is recommended if your application requires geographic, spatial, image, time series, multi-dimensional data, access to heterogeneous DB, machine learning, text retrieval, or word segmentation and you do not want another dedicated database.

Case: Database Migration of ERP



23 Huawei Confidential

Customer Pain Points

Customer A wanted to migrate their ERP system to the cloud, but the commercial databases they used for the ERP blocked their digital transformation journey. The customer wanted to replace their commercial databases with cloud databases.

Solution

- RDS for PostgreSQL was selected as an alternative to using commercial databases.
- First, developers resolved the incompatibility between RDS for PostgreSQL and certain application functions by modifying application code.
 - Then, they optimized RDS for PostgreSQL so it could deliver better performance on par with commercial databases.

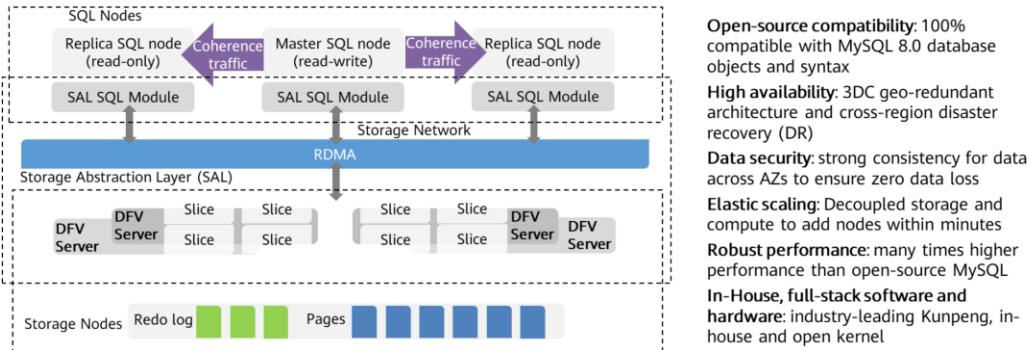
Customer Benefits

This solution helped the customer break free from the licensing constraints of commercial database, which reduced the cost of their ERP system. The customer smoothly transitioned to a SaaS model on Huawei Cloud.



GaussDB(for MySQL)

- Huawei Cloud GaussDB(for MySQL) is a MySQL-compatible, enterprise-grade distributed database. Data functions virtualization (DFV) is used to decouple storage from compute. With GaussDB(for MySQL), there is no need for sharding, and no need to worry about data loss. It provides the superior performance of commercial databases at the price of open-source databases.



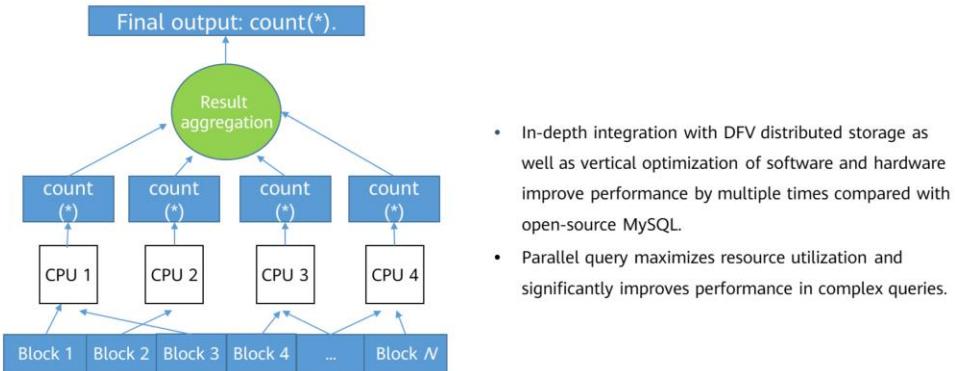
24 Huawei Confidential



- Shared DFV storage:
 - GaussDB(for MySQL) provides a shared storage pool. When adding a read node, you only need to add one compute node, and no additional storage is required. If there are more read-only nodes, more storage costs will be saved.
- Active-active architecture:
 - GaussDB(for MySQL) does not support standby instances. All read replicas are active, offloading read traffic from the primary node and improving resource utilization.
- A "logs as data" architecture:
 - GaussDB(for MySQL) does not use page flushing or double writes. All update operations are recorded in logs to save bandwidth.

Parallel Execution

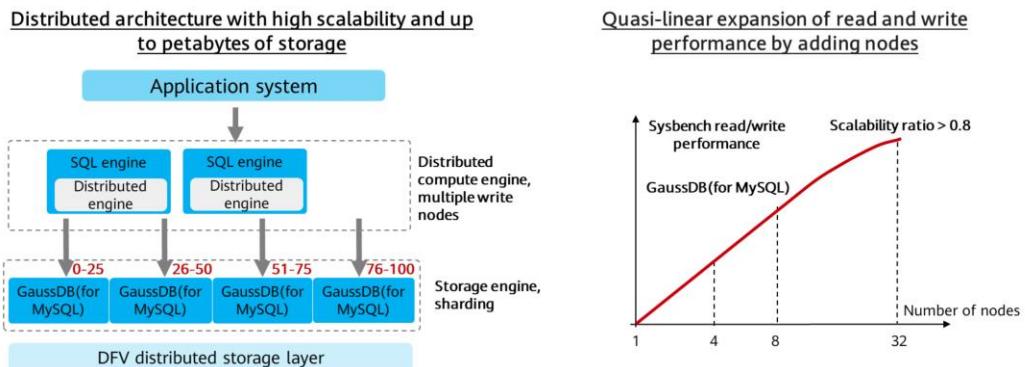
- In parallel execution, data tables that need to be processed are divided into independent data blocks, and then different threads process these divided data blocks in parallel.



- In TPC-H testing, if a DB instance (with 32 vCPUs and 256 GB of memory) handles 100 GB of data, its performance is improved by 8x when handling of 16-thread concurrency requests.

Horizontal Scalability

- GaussDB(for MySQL) supports distributed deployment and horizontal expansion, ensuring fast responses in the face of massive volumes of concurrent requests.



- Linear expansion of GaussDB(for MySQL) read and write performance:
- You do not need to re-divide storage for the new nodes because GaussDB(for MySQL) uses DFV distributed storage. The new nodes can share the same storage as the existing nodes.

Efficient Backup

- Thanks to log streams, terabytes of data can be backed up in seconds. Data can be restored to any point in time during the retention period.

A dedicated distributed storage system ensures fast backup and restoration.

- Snapshots in seconds**

Thanks to AppendOnly at the storage layer, data is stored in multiple copies at multiple points in time, and snapshots are generated in seconds.

- Parallel high-speed backup and restoration**

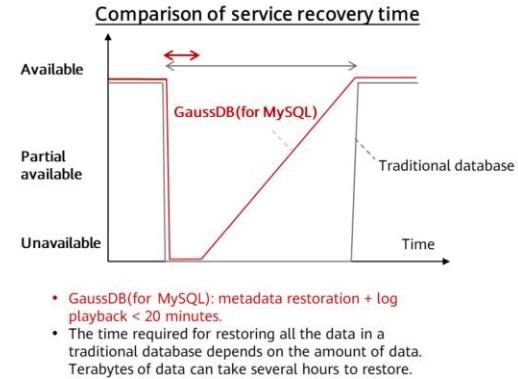
The backup and restoration logic is pushed down to each storage node. Data is accessed locally and directly interconnected with third-party storage systems, improving performance in high concurrency scenarios.

- Terabytes of data recovered in under 20 min**

After metadata is restored, the system can provide services for external systems. When users access the database, hot data responds first and cold data is asynchronously restored in the background.

- Quick rollback at any point in time**

Thanks to the multi-time point feature of the underlying storage system, logs can be rolled back to a specific point of the time without incremental log playback.



- When data is restored, GaussDB(for MySQL) can provide services before the restoration is complete. In contrast, traditional databases need to wait for all data to be fully restored before they can provide services again.

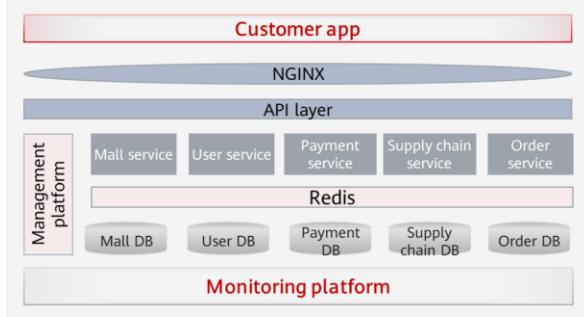
Case: A Customer Uses GaussDB(for MySQL) to Accelerate Business Upgrade

Challenges

Performance bottlenecks: To support rapid growth, the legacy architecture needs to be replaced with a microservice architecture.

Data silos: Cross-border services require streamlined processes.

Stability: The customer has demanding requirements for the performance, security, and stability of cloud platforms.



Solutions

Migration: To ensure a smooth migration, SoftWare Repository for Container (SWR) was provided to deploy containerized applications based on a Jenkins pipeline.

- Huawei Cloud Server Migration Tool (SMS) helped migrate VMs.
- Data from the MySQL and MongoDB databases was replicated to Huawei Cloud RDS.

Database: RDS and proxy were used to route read requests to read replicas and write requests to the primary node. This offloaded pressure from the primary node. GaussDB(for MySQL) can support up to 96 TB of storage and 15 read replicas and a primary node for a DB instance. DFV storage greatly improved the database performance.

Customer Benefits

- **High performance:** To eliminate performance and storage bottlenecks associated with single databases, Huawei databases use sharding to get rid of dependence on specific clouds. In this case, database performance was improved and strong consistency for data was ensured.

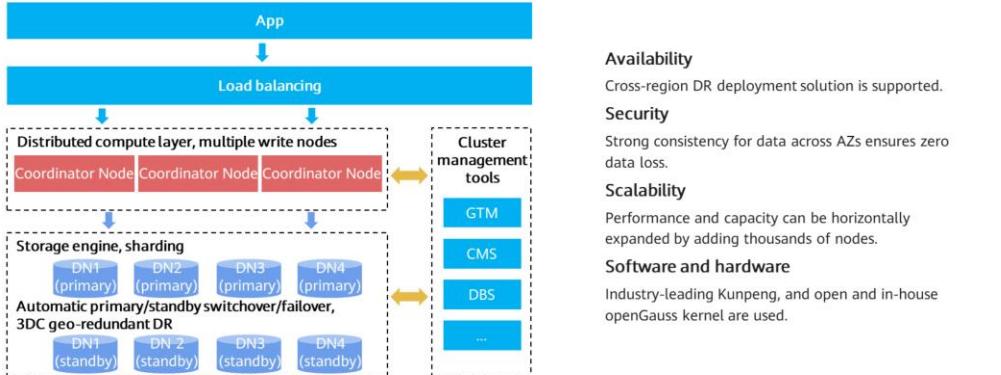
- **Flexible use:** Cloud-based development and test systems can be used on demand and support one-click capacity expansion, cloning, and copying for optimal flexibility.

- **Cost savings:** It took three months to migrate all of the data to the cloud. The migration was fast and the labor costs were low. There is no need to worry about O&M after the migration.



GaussDB(for openGauss)

- GaussDB(for openGauss) is an enterprise-grade relational database from Huawei. It features hybrid transactional/analytical processing (HTAP) workloads for high performance. With a distributed architecture, GaussDB(for openGauss) supports petabytes of storage.



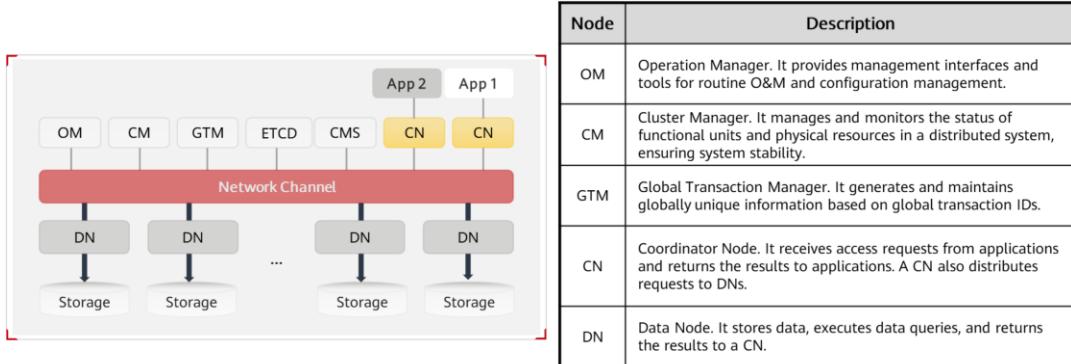
29 Huawei Confidential



- High security:
 - GaussDB(for openGauss) provides security equal to that of top commercial databases using dynamic data masking, transparent data encryption (TDS), row-level access control, and encrypted computing. This feature meets the core data security requirements of enterprises and finance institutions.
- Comprehensive tools
 - GaussDB(for openGauss) can be deployed in the Huawei Cloud and Huawei Cloud Stack for commercial use. It can also work with ecosystem tools such as DAS, UGO and DRS to make database development, O&M, tuning, monitoring, and migration easier.
- In-house, full-stack development
 - Developed based on Kunpeng ecosystem, GaussDB(for openGauss) performance is continuously optimized to meet ever-increasing demands across a wide range of scenarios.
- Open-source ecosystem
 - The primary/standby version is available for you to download from the openGauss community.

Key Nodes

- GaussDB(for openGauss) accelerates data queries by associating multiple modules and nodes.



- ETCD: Editable Text Configuration Daemon. It ensures data consistency.
- CMS: It manages cluster and controls primary/standby switchover to ensure high availability.

Performance: Distributed Parallel Execution Framework

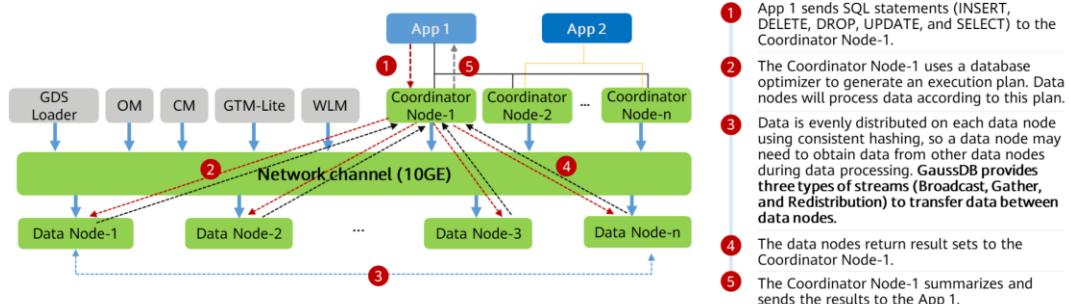
A distributed execution framework can generate an optimized execution plan based on SQL statements. Technologies such as operator pushdown and parallel execution are used to improve execution efficiency.

Operator pushdown

- Single-node execution: a CN delivers a SQL statement (such as INSERT, DELETE/DROP, UPDATE, or SELECT) directly to the DNs for execution.
- Cross-node distributed execution: When data between DNs is exchanged in an associated query, a CN delivers the execution plan to the DNs, and the DNs gather data using the streaming operator.

Parallel execution

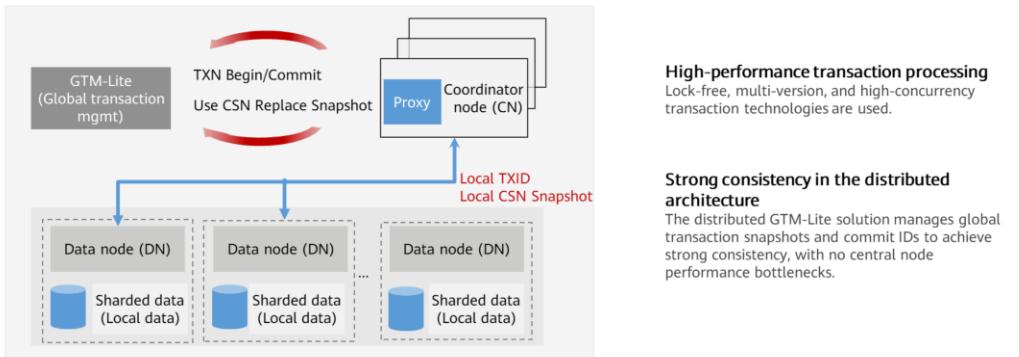
- SQL statements such as DDL and DML can be executed concurrently across nodes. Parallel query based on data pages is supported within a node.



Performance: GTM-Lite Distributed Transaction Processing

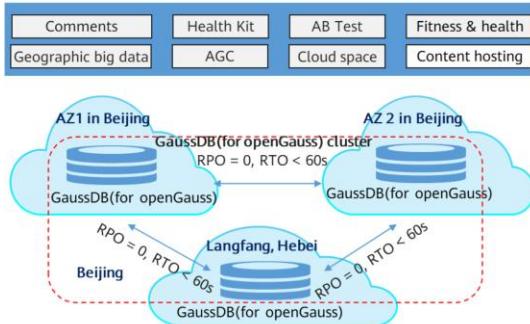
GTM-Lite enables high-performance transaction processing and ensures strong global consistency, eliminating the performance bottlenecks associated with single GTM.

- Instead of an active transaction list, a Commit Sequence Number (CSN) helps quickly determine which transactions are visible to a given transaction.
- GTM provides CSNs through lock-free atomic operations, ensuring there are no single-point bottlenecks.
- Only one CSN is required for transaction interaction between nodes, greatly reducing the network overhead needed to synchronize transaction statuses between nodes.



Case: GaussDB Facilitates Smart Operations

A customer uses cloud-based distributed databases for critical services



Challenges

A cloud-based big data platform of a health management customer used databases in a hybrid architecture to centrally store and manage data, including fitness & health data, cloud space data, and geographic big data. However:

- Fitness data was expanding rapidly, over 30% every year.
- The data analysis platform had to support real-time analysis to provide an intelligent user experience.
- Performance improvement had reached a bottleneck.

Solutions and Benefits

- **Database solutions:** On-demand sharding and elastic scaling can sustain rapid service expansion. Powerful hybrid transaction/analytical processing (HTAP) capabilities support real-time service analysis. Great improvements in linear performance can eliminate the existing architecture's performance bottlenecks.
- **Scalability:** On-demand scaling conserves resources without interrupting services.
- **Efficiency:** With the new data analysis model, analysis results can be obtained in real time. Both marketing precision and analysis efficiency are improved.
- **Costs:** The response time of typical visualized report query and analysis was reduced from minutes to seconds, and the time needed to produce reports was reduced from weeks to hours.

Contents

1. Introduction to Database Services
2. **Cloud Database Services**
 - Relational Database Services
 - Non-Relational Database Services
3. Database Migration

Document Database Service (DDS)

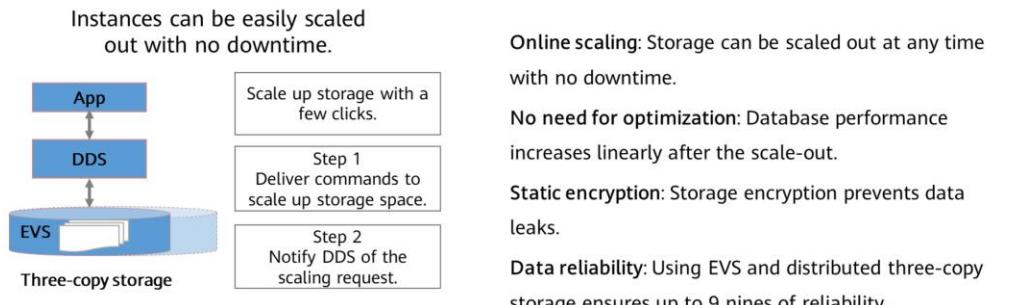
- Document Database Service (DDS) is a MongoDB-compatible database service that is secure, highly available, reliable, scalable, and easy to use. It provides DB instance creation, scaling, redundancy, backup, restoration, monitoring, and alarm reporting functions with just a few clicks on the DDS console. DDS offers three types of instance architecture to suit different scenarios: cluster, replica set, and single node.

Instance Architecture	Characteristics	Service Scenario
Single Node	Ultra-low cost: You only pay for a single node. 10 GB to 1,000 GB data storage is supported. High availability is not supported. If a node fails, workloads will become unavailable.	Requirements for non-core data storage, learning practices, and testing
Replica Set	Three-node architecture: If a primary node becomes faulty, a secondary node becomes the primary. If the secondary node is unavailable, a hidden node will take the role of the secondary to ensure high availability. After a replica set is created, you can add up to either 5 or 7 nodes. 10 GB to 2,000 GB of storage is supported.	Requirements for high availability and data storage less than 2 TB
Cluster	A cluster consists of a config node, and multiple mongos, and shard nodes. Each shard is a replica set that stores business data. You can create 2 to 32 shards with each 10 GB to 2000 GB data storage. The formula for calculating the cluster storage space is $(2 \text{ to } 32) * (10 \text{ GB to } 2000 \text{ GB})$. Online specification change and horizontal scaling are supported.	Requirements for high availability, massive data processing, and scale-out capabilities

- Database type and versions: compatible with MongoDB 4.0 and 4.2.
- Data security: Multiple security policies protect databases and data privacy.
- Data reliability: Three-copy data storage ensures up to 9 nines of database data reliability and up to 12 nines of backup data durability.
- High availability (intra-city disaster recovery): Cluster and replica set instances can be deployed within an AZ or across three AZs, ensuring service availability over 99.95%.
- DB instance monitoring: DDS monitors key performance metrics of DB instance OSs and DB engines, including CPU usage, memory usage, storage space usage, I/O activity, and database connections.
- Elastic scaling:
 - Horizontal scaling: Shards can be created (up to 32 for each instance) or deleted. You can also create 7-node replica sets and read replicas.
 - Vertical scaling: DB instance classes can be modified and storage space can be scaled up to $32 * 2 \text{ TB}$.
- Backup and restoration:
 - Backup: Multiple backup methods are available, such as automated backup, manual backup, full backup, and incremental backup. Backup files can be added, deleted, queried, or replicated.
 - Restoration: Data can be restored to any point in time within the backup retention period and can be backed up to the original DB instance or to a new one. The backup retention period is up to 732 days.

High Reliability - Online Scaling With No Downtime

- DDS instances can be scaled out as needed.

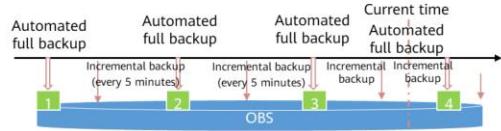


High Reliability - Data Archiving, Backup, and Restoration

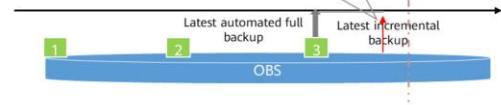
- DDS supports automated backup policies, real-time manual backup, and data restoration using backup files.

Full/Incremental backup and point-in-time restoration

Full backup + incremental backup



Point-in-Time Restoration (PITR)



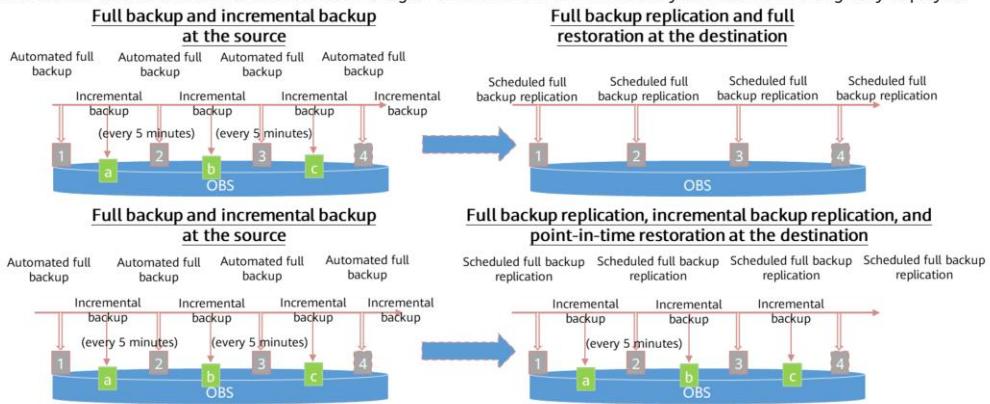
Basic Functions

- You can restore DDS DB instances.
- Backup data durability reaches 12 nines.
- The automated backup retention period is configurable (1-732 days).
- You can restore data to any point in time before the last 5 minutes and can restore data to a new DB instance or your original DB instance.

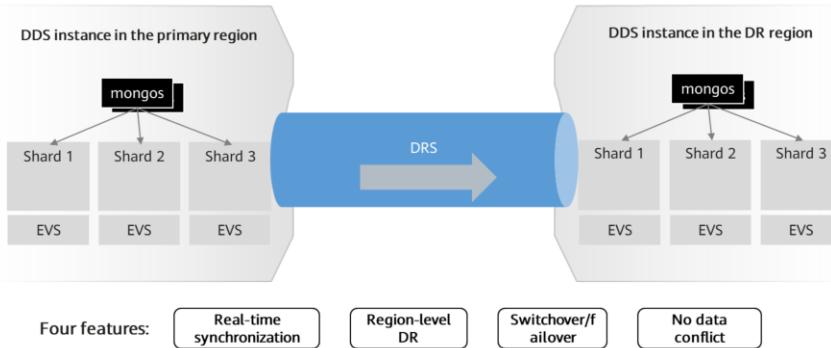


Cross-AZ Backup - Backup Data Replication and Restoration Across Regions

- DDS allows you to replicate backups to a different region. Then for disaster recovery, you can then use the backups in that new region to restore data to a new DDS DB instance in a region different from the one where your instance was originally deployed.

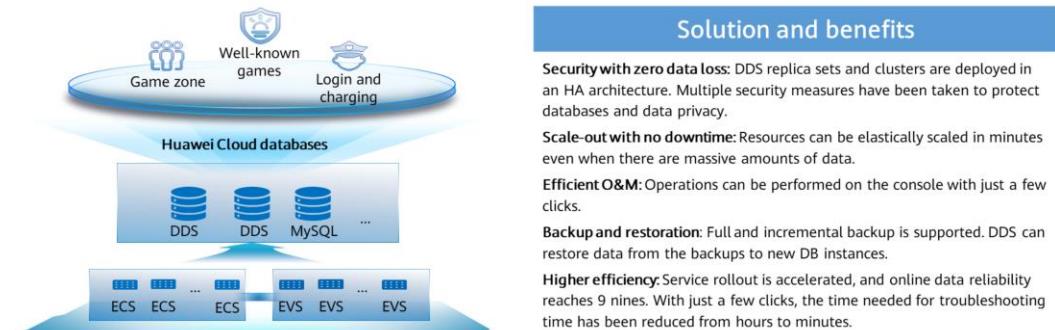


Cross-AZ DR - Real-Time DR and Data Synchronization Across Regions



Success Story in the Gaming Sector

- Background: A game company was having problems with its own databases, such as poor disaster recovery (DR), expensive remote DR, difficult system security assurance, and a lack of security-related O&M experts. Most resources required during peak hours were left idle later on, which was a huge waste. In addition, it was difficult to deploy and scale out on-premises databases, and also it took a long time to scale out database storage. To address all of these issues, DDS was a great choice.



40 Huawei Confidential

Solution and benefits

Security with zero data loss: DDS replica sets and clusters are deployed in an HA architecture. Multiple security measures have been taken to protect databases and data privacy.

Scale-out with no downtime: Resources can be elastically scaled in minutes even when there are massive amounts of data.

Efficient O&M: Operations can be performed on the console with just a few clicks.

Backup and restoration: Full and incremental backup is supported. DDS can restore data from the backups to new DB instances.

Higher efficiency: Service rollout is accelerated, and online data reliability reaches 9 nines. With just a few clicks, the time needed for troubleshooting time has been reduced from hours to minutes.



- Gaming:
- Player information, such as player items and bonus points, is stored in DDS databases. During peak hours, DDS cluster instances can handle large amounts of concurrent requests. DDS clusters and replica sets provide high availability to ensure games are stable in high-concurrency scenarios.
- In addition, DDS is compatible with MongoDB and provides a no-schema mode, which means you do not have to change the table structure when the game play mode changes. DDS can easily meet many flexible gaming requirements. You can store structured data with fixed patterns in RDS, data with flexible patterns in DDS, and hot data in GaussDB(for Redis) to speed up data access and reduce data storage costs.

GaussDB NoSQL

GaussDB NoSQL is a distributed, multi-model NoSQL database service with decoupled storage and compute. It is highly available, secure, and scalable and can provide robust performance and service capabilities like quick deployment, backup, restoration, monitoring, and alarm reporting.

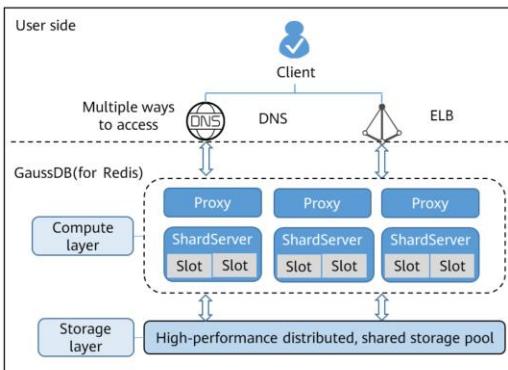
Highlights

Compatible with multiple types of NoSQL APIs	Elastic scaling	Easy O&M	High data security
Cassandra, MongoDB, Redis, and InfluxDB APIs	Decoupled storage and compute allows you to scale compute resources in minutes and storage resources in seconds, without interrupting your workloads.	You can create or delete instances in a visual way on a web-based console. Backup and restoration, configuring alarms, and adding or deleting nodes is just as easy.	A multi-layer security system, including VPCs, subnets, security groups, and Anti-DDoS, protects your databases against a wide range of network attacks.

- Cassandra APIs for:
 - Wide-column data model
 - Ultra-high write performance, making GaussDB NoSQL a huge fit for IoT and financial fraud detection scenarios
- MongoDB APIs for:
 - Document-oriented data model
 - Outstanding read/write performance, low latency, and high reliability
- Redis APIs for:
 - Redis databases with decoupled storage and compute
 - High reliability, scalability, and cost-effectiveness
- InfluxDB APIs for:
 - Efficiently handling time series data
 - High write performance and compression ratio

GaussDB(for Redis)

GaussDB(for Redis) is a NoSQL time-series database with decoupled storage and compute. It is compatible with Redis and breaks Redis' memory limits. GaussDB(for Redis) keeps hot data in memory of compute nodes to ensure low latency and stores cold data in a distributed, shared storage pool for persistence, minimizing storage costs.



Benefits:

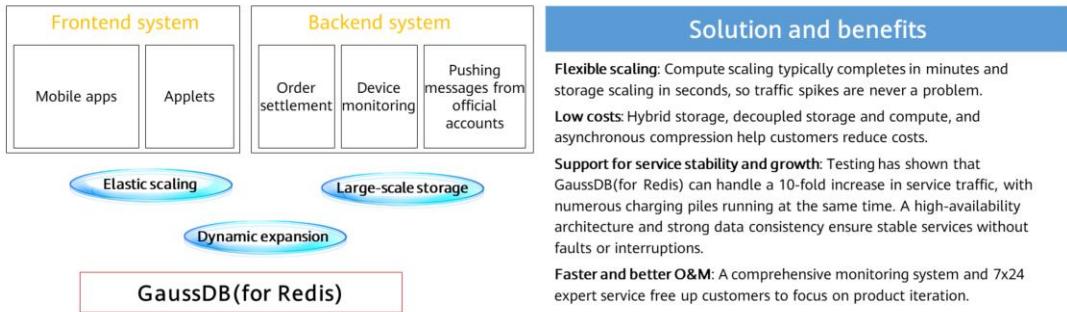
- All data is flushed to disks. Three copies of data are stored in the distributed shared storage pool, ensuring zero data loss.
- All compute nodes support reads and writes.
- Three copies of data keep data strongly consistent, and there are no dirty reads even when multiple nodes handle service requests concurrently.
- If a node fails, GaussDB(for Redis) automates failover to another node and can balance loads among all nodes in a cluster dynamically.
- Storage and compute resources are decoupled, so they can be scaled separately without interrupting workloads.

GaussDB(for Redis) has the following features:

- High cost-effectiveness
 - Thanks to shared storage, GaussDB(for Redis) is able to inexpensively process massive amounts of data.
 - All data is stored in disks with cold and hot data separated. Hot data can be read from the cache directly, making programs run fast.
- Hitless scaling
 - RocksDB is customized to allow storage to be scaled up in seconds.
 - Scaling is fast and smooth because no data needs to be migrated.
 - Proxies ensure that upper-layer applications are not affected by data sharding in the storage layer.
- Cold and hot data separation
 - Hot data is loaded to the memory and cold data is stored persistently, so there is no need to use an extra MySQL database.
 - Cold and hot data is automatically exchanged, making coding easier than before.

Success Story in the Energy Sector

Background: An energy company encountered many problems with its own databases. For example, it had to keep adding database storage to meet rapid service growth, which was driving up costs. Each expansion took a long period of time, and services were unavailable during the scale-out, which would impact user experience. What's worse, there were no dedicated personnel with the skills needed to monitor their databases, so problems could not be quickly diagnosed and resolved when they occurred. To address all of these issues, GaussDB(for Redis) was a great choice.



GaussDB(for Redis) is Redis-compatible and can store a large amount of data inexpensively and reliably, so it is a great fit for persistent storage scenarios.

GaussDB(for Mongo)

GaussDB(for Mongo), with MongoDB compatibility, is a flexible, scalable, and reliable NoSQL database designed for enterprise-grade performance.

Highlights

High performance

- GaussDB(for Mongo) provides 3x read and write performance of open-source MongoDB, allowing you to write data 24/7.

High elasticity

- Decoupled storage and compute allows you to scale compute resources in minutes and storage resources in seconds, without interrupting your workloads.

High reliability

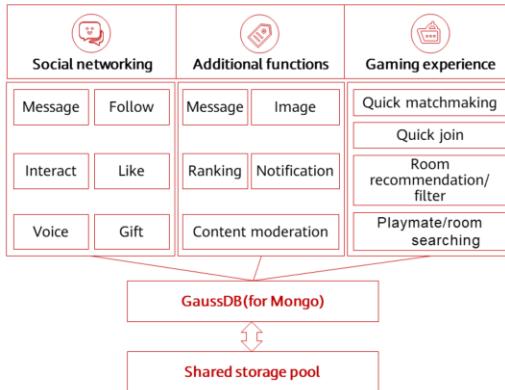
- GaussDB(for Mongo) runs in VPCs, so you can isolate your database in your own virtual network. You can also configure security group rules and enable SSL to secure database access.
- You can deploy nodes across three AZs and quickly back up or restore data once a fault occurs.
- The distributed architecture provides superlative fault tolerance ($N-1$ reliability).

Simple O&M

- A friendly UI makes it easy to manage instances, configure alarms, and add or delete nodes visually.

Success Story in the Gaming Sector

Background: A gaming company used a community edition database to support real-time communication between players. When the number of players grew, the company was unable to handle traffic bursts by adding shards because data sharding is slow. More and more requests increased the latency between primary and secondary nodes of a replica set. In addition, the community edition database does not support high availability. Services became unavailable when the primary node failed, and user experience suffered significantly.



45 Huawei Confidential

Solution and benefits

Fast, flexible sharding: New shard nodes can be added in minutes to handle traffic peaks. Since shard nodes are stateless, data ownerships can be changed from one chunk to another without the need to copy data.

High scalability: Read replicas can be linearly added to a replica set, without increasing the pressure on the primary node or the primary/standby replication latency. GaussDB(for Mongo) supports larger storage space and more nodes than MongoDB, so engineers do not have to worry about complex logics for database sharding.

High availability: Each cluster can contain many shard nodes and supports faulty nodes up to the number of cluster nodes minus 1. Strong data consistency is ensured with no data rollback risks.



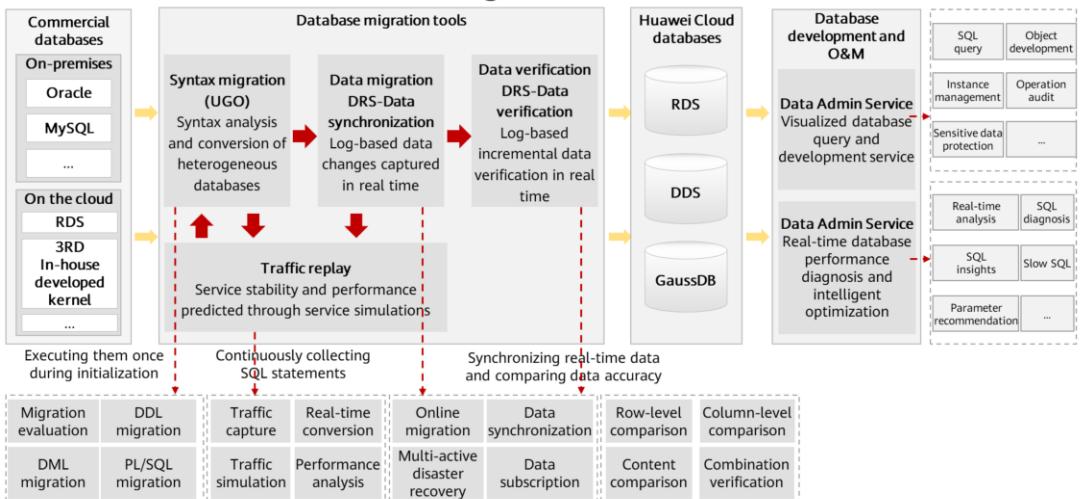
Gaming:

GaussDB(for Mongo) is compatible with MongoDB and allows you to keep track of gaming data like equipment or points earned. Adding compute nodes is so easy, making GaussDB(for Mongo) an excellent choice for high-concurrency scenarios often involved in online gaming.

Contents

1. Introduction to Database Services
2. Cloud Database Services
- 3. Database Migration**
 - Data Replication Service
 - Database and Application Migration UGO

Huawei Cloud Database Migration Solution



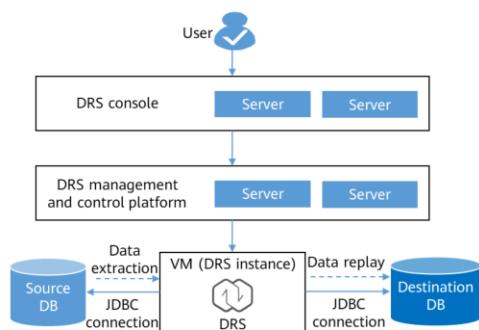
47 Huawei Confidential



- Database Migration Method:
- In most cases, you can migrate databases using both UGO and DRS. When migrating databases from on-premises or other clouds to Huawei Cloud, you can use UGO to analyze the source databases and migrate the databases based on the actual scenario and the suggestions provided by UGO. You can also use the full + incremental migration provided by DRS to migrate data from one database to another.

Data Replication Service (DRS)

- Data Replication Service (DRS) enables you to migrate databases to a cloud with no downtime. It supports migration between homogeneous databases, heterogeneous databases, distributed databases, and sharded databases. DRS enables data integration and transmission from databases to databases, data warehouses, and big data within seconds, laying a solid foundation for enterprise data integration and digital transformation.



Minimum permission design

- Java Database Connectivity (JDBC) is used to connect to the source and destination databases, so you do not have to deploy programs on the databases.
- A task runs on an independent and exclusively-used VM. Data is isolated between tenants.
- The number of IP addresses is limited. Only the DRS instance IP address is allowed to access the source and destination databases.

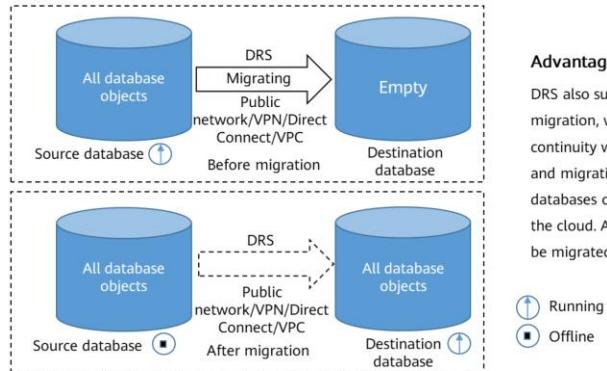
Reliability design

- Automatic reconnection: If the connection between DRS and your database breaks down due to a network disconnection or database switchover, DRS automatically retries the connection until the task is restored.
- Resumable data transfer: If the connection between the source and the destination fails, DRS automatically marks the log file position for replay. After the fault is rectified, you can resume data transfer from the log file position to ensure no data is lost.
- If the VM where the DRS replication instance is located fails, workloads are automatically switched to a new VM with the IP address unchanged to ensure that the migration task is not interrupted.

- Easy to use
 - Traditional migration requires professional technical personnel and migration procedures are complex.
- Fast setup
 - Traditional migration takes several days, weeks, or even months to set up.
- Low costs
 - Traditional migration is expensive and there is no pay-per-use pricing.
- Secure
 - Traditional migration involves downtime and if there is a migration failure, data may be lost.

Real-Time Migration

- For a real-time migration, DRS needs to be connected to both the source DB and destination DB. In addition, the source DB, destination DB, and migration objects must be configured, and then DRS can perform the migration automatically. By comparing multiple items and data between source and destination, you can determine the best time for migration to minimize downtime.



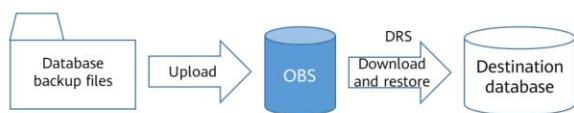
Advantages

DRS also supports incremental migration, which ensures your service continuity while minimizing downtime and migration impacts. In this way, databases can be smoothly migrated to the cloud. Also, all database objects can be migrated.

Backup Migration

- For security reasons, it is often necessary to hide the real IP address of your database. Migrating data through dedicated connections is an option, but it is expensive. DRS allows you to export data from your source database for backup and upload the backups to OBS. You can then restore the backups to a destination database to complete the migration. Using this method, data migration can be completed without exposing your source databases to the Internet.

Typical scenario: migration from on-premises databases to the cloud

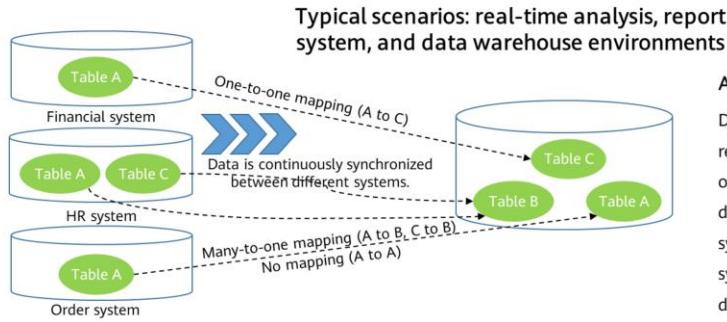


Advantages

DRS can help you complete data migration without connecting to your source databases.

Real-Time Synchronization

- Data synchronization refers to the real-time flow of key service data from one source to the other while the consistency of the data is ensured. Real-time synchronization is different from data migration. Migration means moving your whole database from one platform to another. Synchronization refers to the continuous flow of data between different services.



Advantages

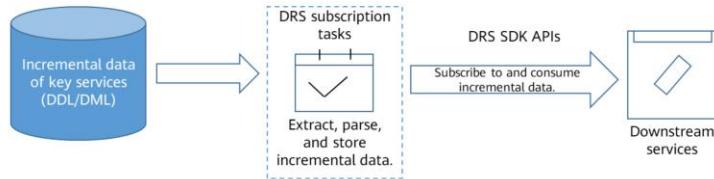
DRS can meet various synchronization requirements, such as many-to-one, one-to-many synchronization, dynamic addition and deletion of synchronization tables, and synchronization between tables with different names.



Data Subscription

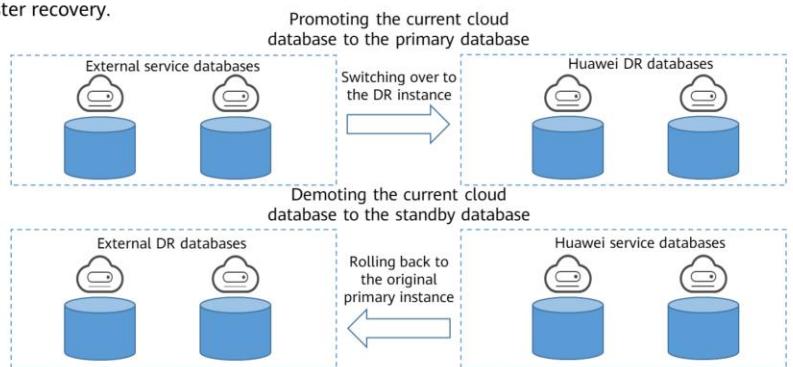
- Data subscription is how you learn of data changes made to key services. This type of information is often required by downstream services. Data subscription helps cache incremental data and provides a unified SDK interface for downstream services to subscribe to and consume the incremental data.

Typical scenario: Kafka subscribes to RDS MySQL incremental data.



Real-Time Disaster Recovery

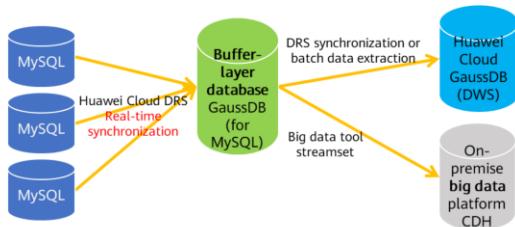
- To prevent service interruptions caused by regional faults, DRS provides disaster recovery. You can easily perform disaster recovery between on-premises databases and cloud databases, or between databases across cloud platforms.
- Geo-redundancy (two-site three-DC or two-site four-DC) disaster recovery architectures are supported. A primary/standby switchover can be implemented by promoting a standby node or demoting a primary node during disaster recovery.



Success Story in the Automotive Sector

- Background: An automobile company used an on-premises database to aggregate data in their report system. There were nearly 20,000 extraction tasks. The entire dataset was inspected automatically every night to detect changes. The timeliness and data governance could not meet the requirements of service reports. In the original solution, a synchronization link was created for each source table, and a total of 200 schemas were required. Each schema had 80 tables to be synchronized, and this much synchronization was driving up costs. As their business grew, so did the volume of data they needed to process. The report system had to store over 10 TB of data, and the database they were using was underperforming.

Multiple MySQL databases are integrated into the buffer-layer database GaussDB(for MySQL), and then into the data warehouse GaussDB(DWS) and on-premises big data platform CDH.



Solution and benefits

Data synchronization: A database synchronization tool was used to aggregate different data sources to the same database and synchronize data from the cloud database to the on-premises platform. The source database system has multiple DB instances, and each DB instance has multiple databases with the same table structure.

Performance and scalability: GaussDB(for MySQL) provides ultra-large capacity and supports millions of QPS, ensuring fast responses even when there is a massive volume of concurrent requests.

Real-time reports: The availability time of sales report data is reduced from days to minutes.

Low O&M cost: DRS synchronizes data by instance. The customer has less than 100 DB instances, so the number of O&M connections is reduced from tens of thousands to no more than 100, and tens of thousands of data tables are aggregated to hundreds of tables.

Data security and reliability: The data synchronization delay is short. DRS eliminates the data inconsistencies that may occur during synchronization.

Large capacity, high performance, and scale-out in seconds: DRS can quickly handle large amounts of concurrent data and scale out in seconds to easily handle traffic spikes.

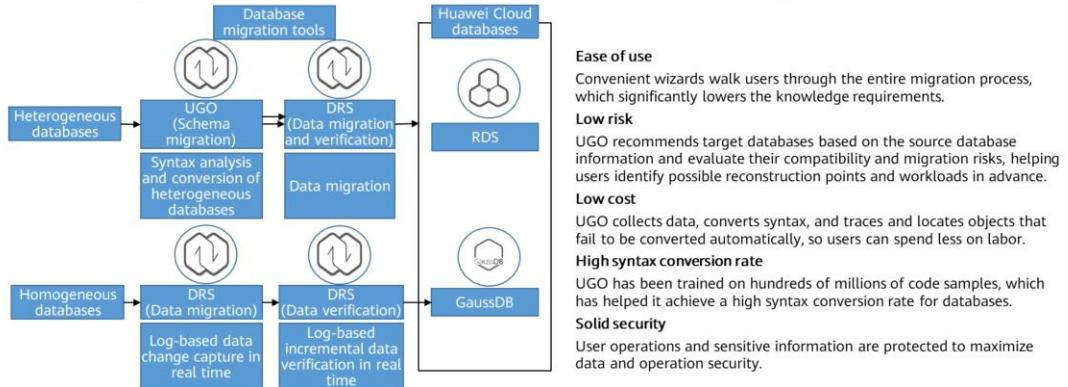


Contents

1. Introduction to Database Services
2. Cloud Database Services
- 3. Database Migration**
 - Data Replication Service
 - Database and Application Migration UGO

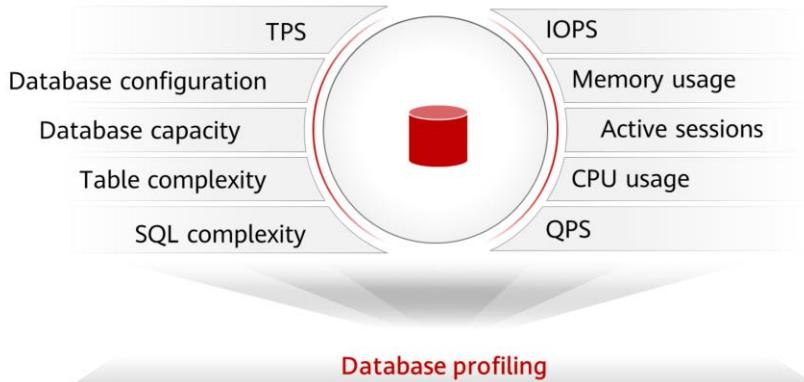
Database and Application Migration UGO

- Database and Application Migration UGO (UGO) is for heterogeneous database schema migration and application SQL conversion. With the functions like database evaluation, object migration, and automatic syntax conversion, UGO makes it simple and cost-effective to migrate databases.



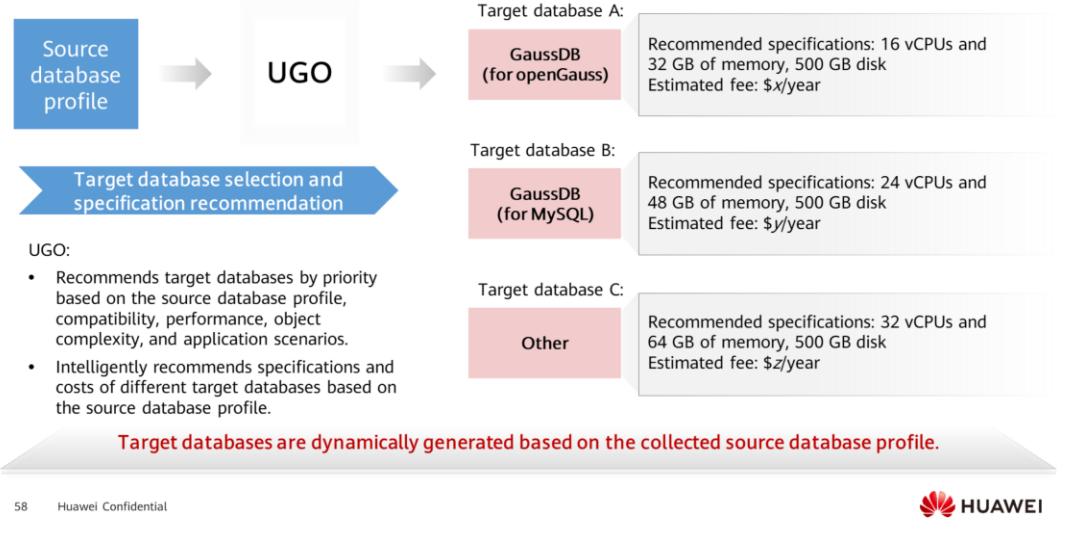
- UGO has been deployed commercially only in CN South-Guangzhou and AP-Singapore.

Source Database Profiling



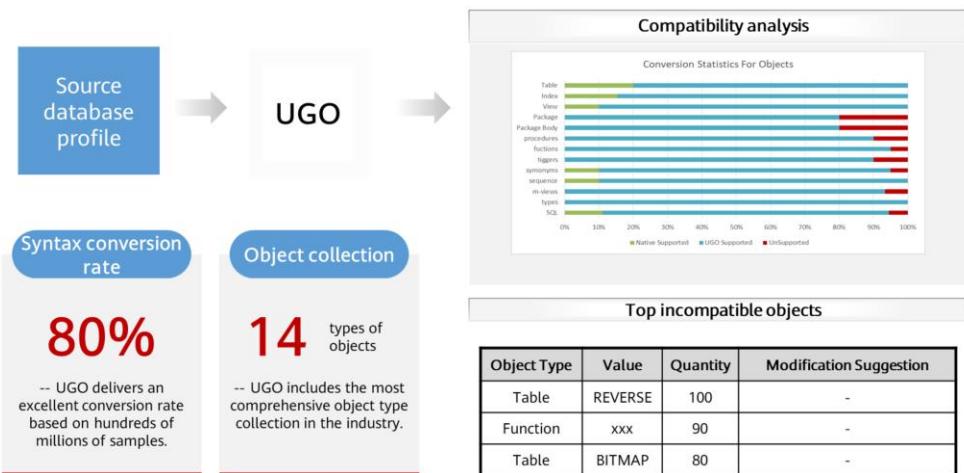
- Core Features 1: Source Database Profiling
 - Source database profiling uses a massive collection of actual service scenarios as samples and key database metrics as features for training to present a picture of the source database. Source database profiling provides a basis for a fast follow-up precision analysis of source database's application scenarios and user habits.

Recommended Specifications for Target Databases



- Core Features 2: Recommended Specifications for Target Databases
 - UGO recommends different types of target databases by priority based on source database profiling, compatibility, performance, object complexity, and application scenarios. UGO also intelligently recommends specifications and estimates costs of the target databases.

Compatibility Analysis



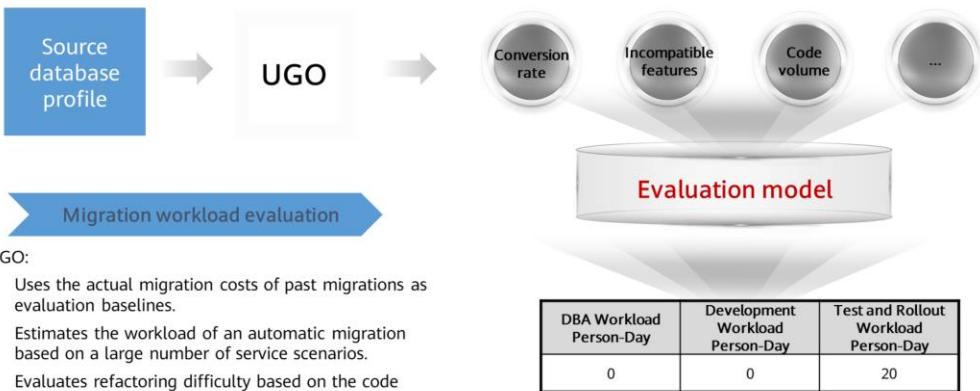
59 Huawei Confidential



- Core Features 3: Target Database Compatibility Analysis

- UGO analyzes the compatibility of up to 14 core object types between source and target databases based on source database profile and on a high syntax conversion rate. UGO delivers an excellent conversion rate based on hundreds of millions of samples.

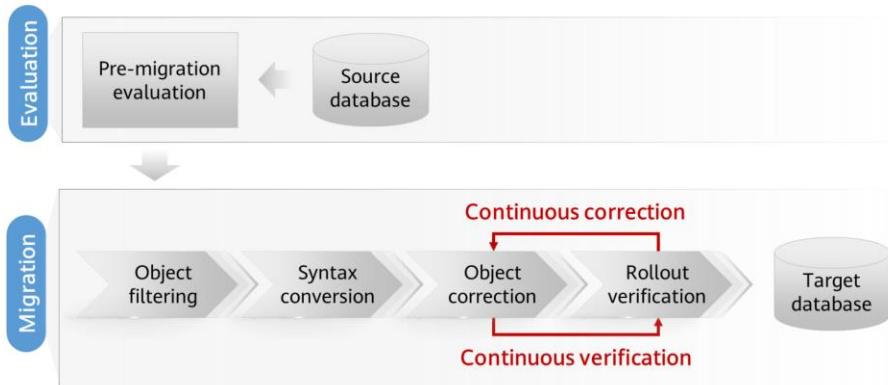
Workload Evaluation



- Core Features 4: Workload evaluation

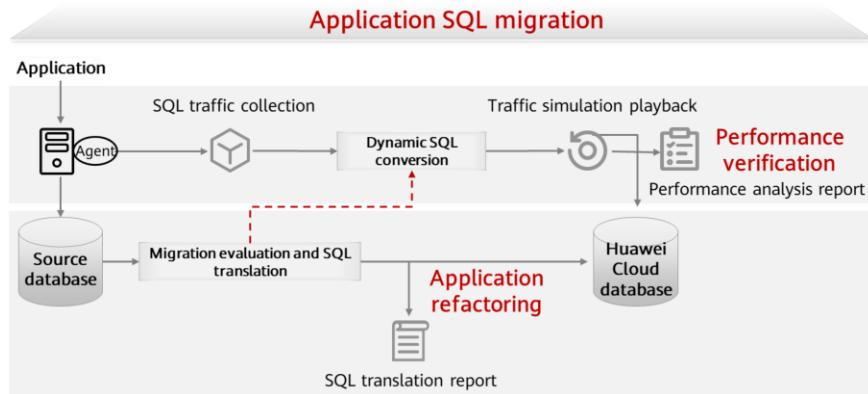
- The cost of labor for a typical database migration is used as a baseline, and then the workloads involved in automatic database migration are added in. Additionally, UGO evaluates the migration workloads based on the amount of code involved, the conversion rate, and how hard it will be to modify incompatible objects.

Database Schema Migration



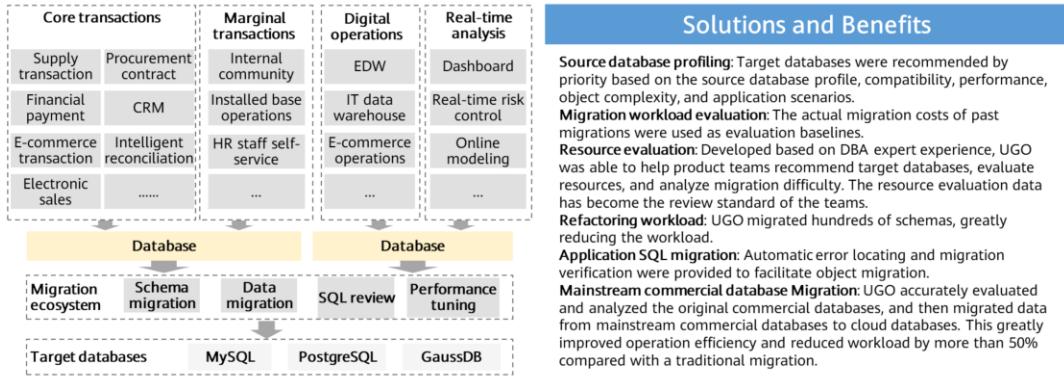
- Core Features 5: Database schema migration
 - After evaluating the source database, UGO allows users to filter the objects to be migrated, and then verifies and migrates the objects. Failed objects are modified and the process is repeated until all objects pass.

Application SQL Migration



Case: UGO Facilitates the Communications Industry

- Background: To meet service continuity requirements, a communications company wanted to migrate their data from various commercial databases to Huawei Cloud databases. However, they ran into some problems. For example, the workload involved in migrating and refactoring data objects was huge. There were more than 100,000 tables and about 100,000 stored procedures, so it would have taken about 200,000 person days to migrate database objects. Additionally, it was difficult and complex to select optimal target databases for various types of services such as ERP, transactions, and decision-making.



Quiz

1. (Multiple-Answer Question) Which of the following DB engines are supported by RDS?
 - A. MySQL
 - B. PostgreSQL
 - C. SQL Server
 - D. Redis
2. (Multiple-Answer Question) Which of the following products are included in Huawei Cloud GaussDB NoSQL?
 - A. GaussDB(for Cassandra)
 - B. GaussDB(for Mongo)
 - C. GaussDB(for Influx)
 - D. GaussDB(for Redis)

- ABC
- ABCD

Quiz

1. (Discussion) What are the advantages of cloud databases over on-premises databases?
2. (Discussion) What should be considered for using cloud database services in terms of security, cost, reliability, performance, and scalability?

- Discussion 1:
 - Cloud databases are managed by cloud vendors.
 - On-premises databases are managed by users.
- Discussion 2:
 - Security: Do not open external network access to databases. Open only internal ports. Use cloud security services to prevent malicious attacks and perform data DR drills.
 - Costs: Evaluate the costs and performance of different engines. When the number of access requests is small, reduce the cluster scale.
 - Reliability: Preferentially select primary/standby instances.
 - Performance: Select a proper DB engine based on data types. Design the association between tables. Use caches to improve the access speed.
 - Scalability: Select cloud-native databases to facilitate future data expansion. Consider intra-region and cross-region data backup.

Summary

- This course covered relational databases and non-relational databases. In this course, we studied the application scenarios and key features of different database services, and learned about the importance of database services. A good grasp of database services and their associated services is very important for developing enterprise services.

Acronyms and Abbreviations

- API: Application Programming Interface
- AS: Auto Scaling
- BMS: Bare Metal Server
- CBR: Cloud Backup and Recovery
- CDN: Content Delivery Network
- DCS: Distributed Cache Service
- DRS: Data Replication Service
- DNS: Domain Name Service
- DDoS: Distributed Denial of Service
- DevOps: Development and Operations
- DIS: Data Ingestion Service
- DLI: Data Lake Insight
- EIP: Elastic IP
- ECS: Elastic Cloud Server
- ELB: Elastic Load Balancer
- EVS: Elastic Volume Service
- GSLB: Global Server Load Balance
- HA: High Availability
- IMS: Image Management Service
- IDC: Internet Data Center
- LTS: Log Tank Service

Acronyms and Abbreviations

- NAT: Network Address Translation
- OLAP: Online Analytical Processing
- OLTP: Online Transaction Processing
- RDS: Relational Database Service
- SMN: Simple Message Notification
- SFS: Scalable File Service
- SDRS: Storage Disaster Recovery Service
- VM: Virtual Machine
- VPC: Virtual Private Cloud
- VPN: Virtual Private Network

Recommendations

- Huawei iLearning
 - <https://e.huawei.com/en/talent/portal/#/>
- Huawei Cloud Help Center
 - <https://support.huaweicloud.com/intl/en-us/index.html>
- HUAWEI CLOUD Developer Institute
 - <https://edu.huaweicloud.com/intl/en-us/>
- Huawei Talent Online
 - <https://e.huawei.com/en/talent/portal/#/>

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©2022 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.



Cloud Security Service Planning



Foreword

- An increasing number of companies are migrating their workloads to the cloud. Cloud security threats like data breaches are drawing more attention.
- This course gives an overview of corporate cloud security, and introduces Huawei Cloud security services and their usage.

Objectives

- Upon completion of this course, you will be able to:
 - Understand the importance of cloud security and learn basic concepts.
 - Get to know the Huawei Cloud security service system and be able to select required security services based on corporate security requirements.
 - Learn how to use cloud security services, design the security architecture, and select required products.

Contents

1. Cloud Security Design and Huawei Cloud Security System
2. Workload Security
3. Network Security
4. Application Security
5. Data Security
6. Security Management

Why Cloud Security Is Important?



Security threats on the cloud

Top 11 cloud security risks

(Updated by CSA in September 2020)

1. Data Breaches
2. Misconfiguration and Inadequate Change Control
3. Lack of Cloud Security Architecture and Strategy
4. Insufficient Identity, Credential, Access, and Key Management
5. Account Hijacking
6. Insider Threat
7. Insecure interfaces and APIs
8. Weak Control Plane
9. Metastructure and Applisstructure Failures
10. Limited Cloud Usage Visibility
11. Abuse and Nefarious Use of Cloud Services



Compliance requirements

EU: *General Data Protection Regulation*

China: *Cybersecurity Law*, *DICP (MLPS) 2.0*, and *Data Security Law*

Countries/Regions all over the world are enacting **cybersecurity laws** to enhance data and privacy protection.

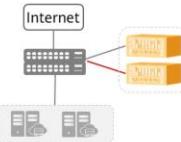
The country implements key protection of public communication and information services, power, traffic, water resources, finance, public service, e-government, and other critical information infrastructure that if destroyed, loses function, or experiences leakage of data might seriously endanger national security, national welfare and the people's livelihood, or the public interest, on the basis of the cybersecurity multi-level protection system.

Cybersecurity Law of the People's Republic of China

- CSA: Cloud Security Alliance

Corporate Security Requirements on the Cloud

Key security requirements for enterprise cloudification



Service continuity

- Defense against DDoS attacks
- Intrusion prevention
- High availability of security products

Convenient control and O&M

- Security policy configuration
- Risk identification & handling
- Auditable and traceable operations

Data confidentiality

- External breach prevention
- Invisible without authorization (blocking malicious insiders)
- Invisible to cloud service providers

Compliance

- DJCP
- *Data Security Law*
- *Personal Information Protection Law*

Meeting Cloud Security Requirements in Five Aspects

Workload security

Continuously monitor and eliminate threats to ensure cloud workload security.

Network security

Configure security services at the network layer to isolate cloud resources and protect network borders.

Application security

Configure security services at the application layer to block attacks.

Data security

Manage data assets throughout their lifecycles to ensure that the entire data usage process is secure, visible, controllable, and traceable.

Security management

Manage the cloud environment to minimize risks.

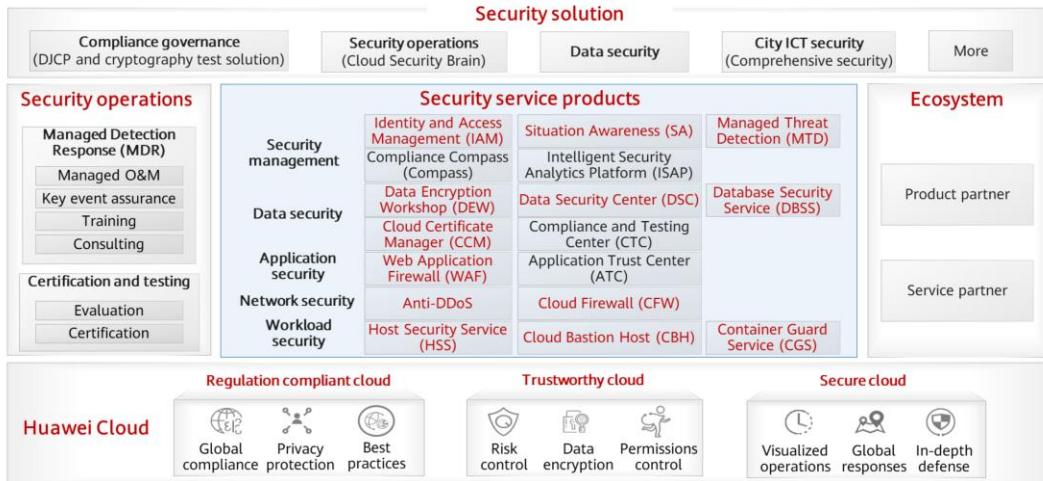
Service continuity

Convenient control and O&M

Data confidentiality

Compliance

Huawei Cloud Security Services

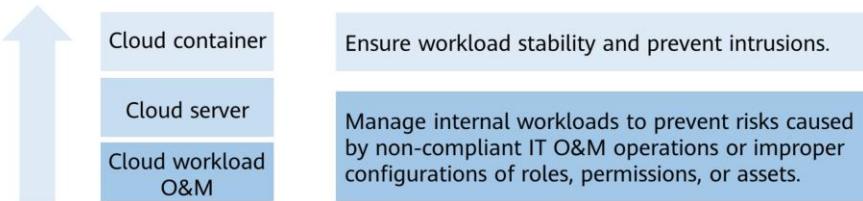


Contents

1. Cloud Security Design and Huawei Cloud Security System
- 2. Workload Security**
3. Network Security
4. Application Security
5. Data Security
6. Security Management

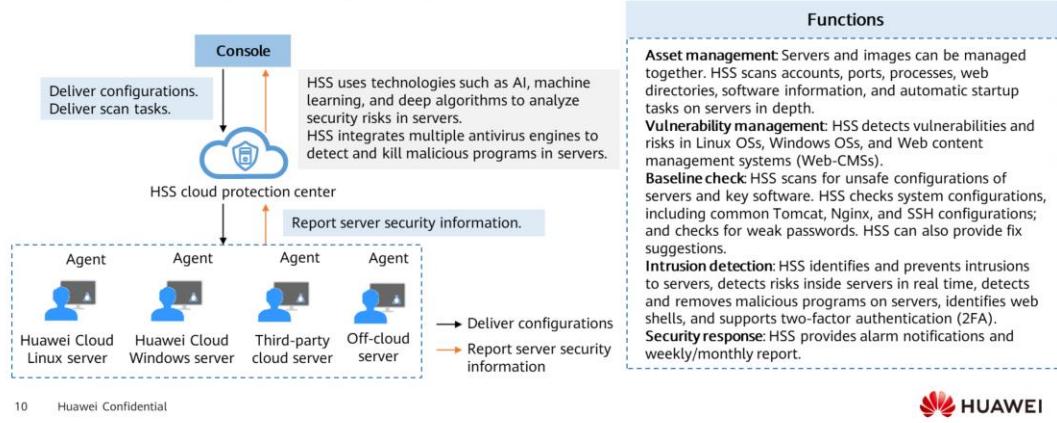
Workload Security

- The workloads of cloud resources like VMs and containers are the core of enterprise business on the cloud.



Host Security Service (HSS)

- HSS checks your assets and protects them from harm you may or may not have noticed, including intrusions, vulnerabilities, and unsafe settings. HSS can identify and manage data assets on your servers, scan for risks in real time, and defend against intrusions to your servers. With HSS, you can easily build a security system to protect your servers.



10 Huawei Confidential



- Console: a visualized management platform, where you can apply configurations in a centralized manner and view the defense status and scan results of servers in a region.
- The HSS cloud protection center receives these configuration information and detection tasks and then forwards them to the Agent installed on the server. Agents block attacks based on security policies and scan all servers every early 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.
- The cloud protection center presents analysis results as reports on the console.
- Other functions:
 - Web Tamper Protection (WTP): WTP can detect and prevent tampering of files in specified directories in real time, including web pages, documents, and images, and quickly restore them using valid backup files.
 - Advanced defense: application recognition service (ARS), file integrity monitoring, and ransomware protection.
 - Unified multi-cloud management: HSS can manage hundreds of thousands of servers running mainstream OSs, such as Linux and Windows, no matter what cloud they are deployed and which architectures (x86 or Arm) they are using.

HSS Application Scenarios

Risk prevention	Real-time protection	Secure operations	Regulation compliance
<ul style="list-style-type: none">Comprehensive preventive measures, 10+ types of baseline checks, management of 6 types of assets, and 170,000+ known vulnerabilities, making it possible for HSS to reduce your attack surface by 90%	<ul style="list-style-type: none">An advanced intrusion detection system that leverages brute-force cracking prevention and 2FA, preventing 100% brute-force attacks against accountsWeb tamper protection for dynamic and static web pages, preventing website information of important systems from being tampered with	<ul style="list-style-type: none">You can enable alarm notifications and HSS will notify you of server security issues in real time.Periodic security reports, making it easier for you to learn of major events on and risks to your servers	<ul style="list-style-type: none">The intrusion detection and vulnerability management functions of HSS comply with the server intrusion prevention clauses in DJCP (MLPS).HSS can detect malicious programs of HSS. The vulnerability management function of HSS meets the malicious code prevention clause in DJCP.

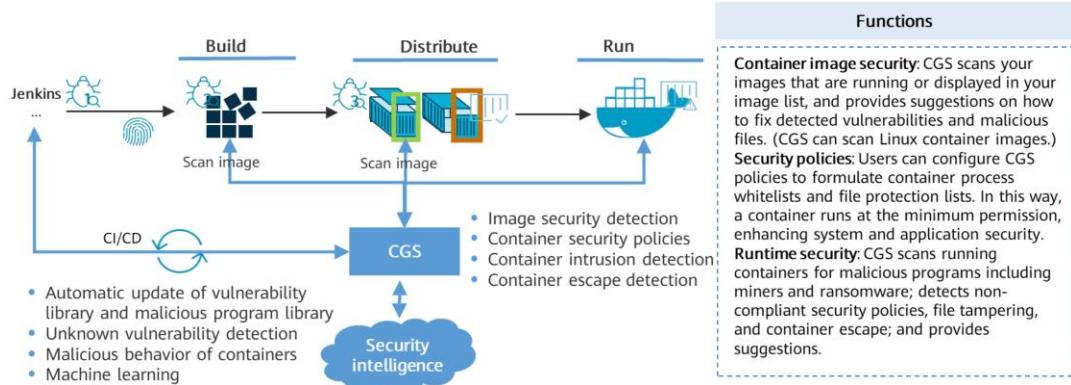
11 Huawei Confidential



- HSS provides comprehensive and effective security solutions for 230,000 companies and individual users in government, Internet, and education industries in and outside China. HSS provides comprehensive risk prevention and real-time protection capabilities, periodically generating security reports to meet DJCP requirements. HSS implements comprehensive protection by providing prevention, detection, and operations functions.

Cloud Container Security (CGS)

- Huawei Cloud CGS addresses container security risks. The core functions of CGS meet the DJCP requirements for intrusion prevention and malicious code prevention. CGS can scan vulnerabilities in container images, prevent escapes, and allow users to configure security policies.



12 Huawei Confidential



- With CGS, you can detect and eliminate risks in your containers and images throughout their lifecycles, including building, distributing, and running.

CGS Application Scenarios



Container image security

External images are susceptible to vulnerabilities. CGS can check image repositories, official Docker images, and running images for vulnerabilities and malicious files; perform baseline checks; and provides fixing suggestions.

Container runtime security

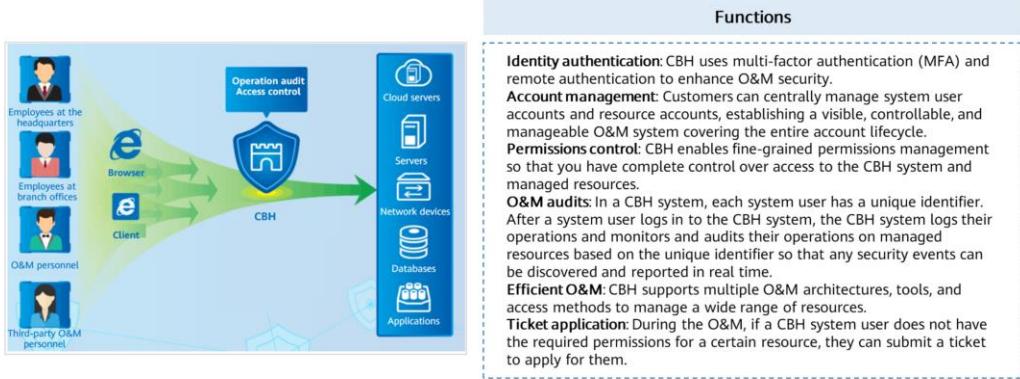
Container behaviors are immutable. CGS helps enterprises develop a whitelist of container behaviors to ensure that containers run with the minimum permissions required and secure containers against potential threats.

DJCP compliance

The core functions of CGS comply with the intrusion prevention and malicious code prevention clauses in DJCP.

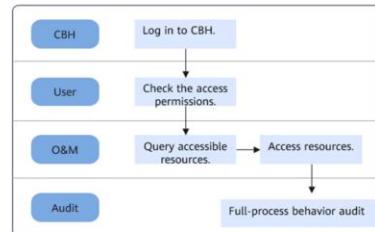
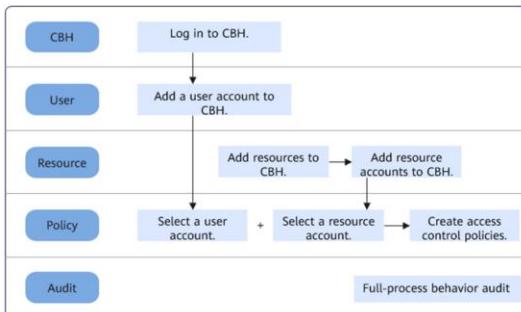
Cloud Bastion Host (CBH)

- CBH can monitor the usage of the CBH system, monitor O&M activities of each managed resource, and identify suspicious O&M actions in real time. CBH protects resources and data from being accessed or damaged by external or internal users. CBH reports alarms to customers, who can then more easily handle or audit O&M issues in a timely, centralized manner.



- CBH also enables collaborative O&M tasks and batch management of servers and databases.

How to Use CBH



- How an administrator creates access policies:
 - Adding resources to CBH: The administrator adds resources to be managed. They can add a wide range of resources, such as servers, network devices, security devices, applications, and databases. CBH allows users to edit resource details, including the system type, department name, resource name, resource address, protocol type, and applications.
 - Creating user accounts: The administrator creates user accounts. A user account is a unique account that is used by a specific O&M engineer to log in to the CBH system. Each user account maps to a real O&M individual.
 - Adding resource accounts to CBH: The administrator adds resource accounts to the CBH system. A resource account is used to log in to a specific resource managed in CBH. Each resource account has its own username and a password. Resource accounts can be used for automatic, manual, or semi-automatic logins. A regular resource account can be escalated to a privileged account, or even given sudo privileges. Beyond that, passwords of resource accounts can be updated by CBH periodically.
 - Creating access control policies: The administrator creates access policies based on combinations of time ranges, primary accounts, resource accounts, and permissions.
 - Full-process behavior audit: CBH automatically logs the administrator's behavior, including how they manage resources, system users, and policies, for monitoring and audits.

CBH Application Scenarios

Strict compliance audit

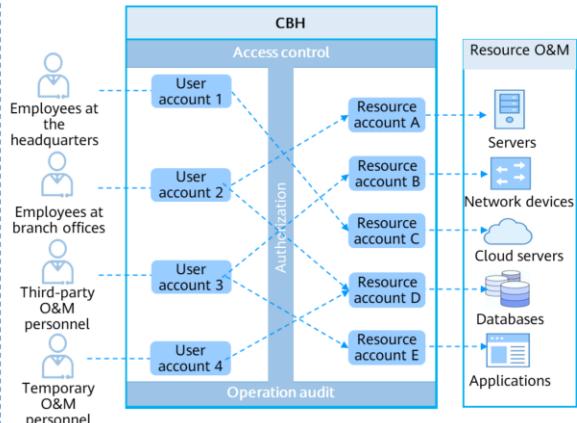
In insurance and finance, there is a lot of personal data involved, frequent financial transactions, and many third parties. So, there are big legal risks, such as regulatory risks and risks of abuse of power. CBH can help users meet regulatory requirements by centrally managing accounts and resources while following the principle of segregation of duties.

Stable Efficient O&M

Some enterprises, such as fast-growing Internet enterprises, have a lot of sensitive data, such as operations data, transmitted over the Internet, which makes them more vulnerable to data breaches. CBH can hide the real addresses of assets during remote O&M, protecting asset information from leakage. The administrator can monitor O&M activities by monitoring operation logs.

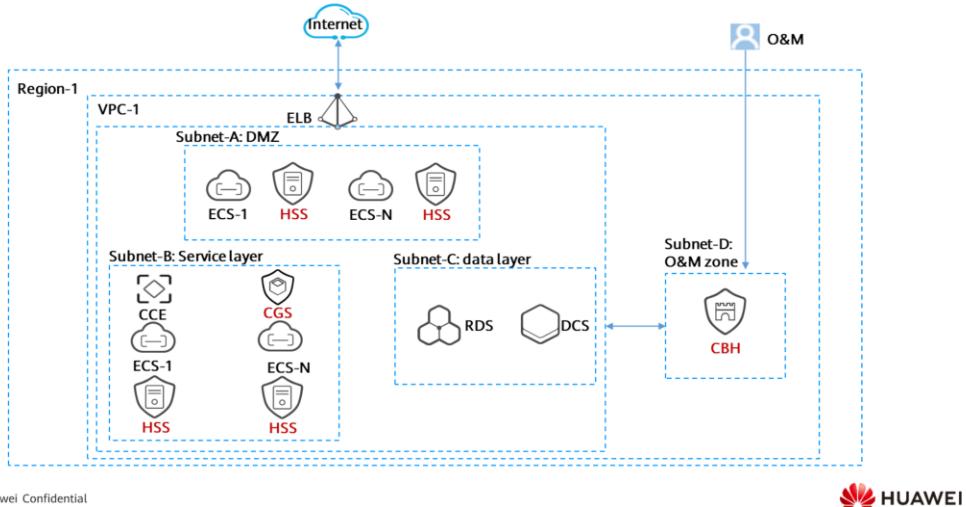
Management of a large number of assets and personnel

Cloud-based O&M is a big challenge for traditional enterprises as they move workloads to the cloud. To save money on human resources, many companies outsource O&M to system suppliers or third-party O&M providers. However, with more suppliers and staff involved, if there is insufficient monitoring, then there is also more risk. CBH can be used to manage O&M personnel and resources at scale, audit all users, discover risks in a timely manner, and effectively locate who should be responsible for a certain O&M activity.



- Customers can use IAM accounts to control who can access a CBH system. The administrator of a CBH system can create users in the CBH system and assign role-based permissions. This figure shows account permissions assigned to different O&M personnel. Only the administrator has the permissions needed to manage roles in the CBH system.
- Strict compliance audit:
 - CBH gives the customers the ability to establish a sound O&M audit system, making it easier for them to comply with regulatory requirements no matter what industry they are in and no matter how strict the requirements are. CBH provides a single point of entry for cloud resource management that enables customers to centrally manage accounts and resources, grant permissions by department, configure multi-level review for operations on mission-critical assets, and require double approvals for sensitive operations.
- Efficient and stable O&M
 - During remote O&M, CBH hides the actual IP addresses so the details of remotely managed assets can be kept secure. CBH provides comprehensive O&M logs that let customers can effectively monitor and audit the operations of O&M personnel both inside and outside of their organizations, reducing network security incidents and keeping service systems stable.
- Management of a large number of assets and personnel:
 - CBH provides a system to securely manage a large number of O&M accounts and a wide range of resources. It also allows O&M personnel to access resources using single sign-on (SSO) tools, improving the O&M efficiency. CBH uses fine-grained permissions control so that all operations on a managed resource are recorded and operations of all O&M staff are auditable. Any O&M incidents are traceable, making it easier to locate the operators.

Workload Security Products in a Cloud Architecture



17 Huawei Confidential

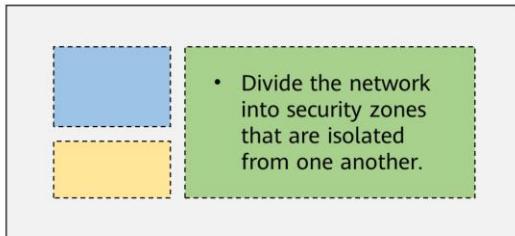
- The solid line indicates the access traffic.
- Demilitarized Zone (DMZ) is a special network area different from the external network or internal network. Generally, the DMZ houses public servers that do not contain confidential information, such as web servers, email servers, or FTP servers. Users from the external network can only access the services in the DMZ, but cannot access the information on the internal network. So, the information on the internal network cannot be impacted even if the servers in the DMZ were attacked.

Contents

1. Cloud Security Design and Huawei Cloud Security System
2. Workload Security
- 3. Network Security**
4. Application Security
5. Data Security
6. Security Management

Network Security

- The network is the basis for communication between resources on the cloud. As such, network security is an important part of cloud security.

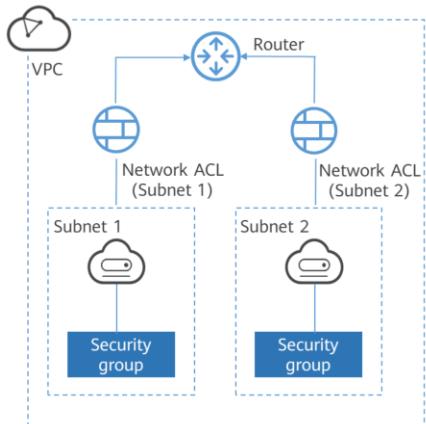


- Divide the network into security zones that are isolated from one another.

- Meet the redundancy requirements for network connections and devices.

- Implement protection and monitoring at network borders.

Security Group and ACL



- A security group provides security at the instance level. It is a collection of access control rules for cloud resources, such as cloud servers, containers, and databases, that have the same security protection requirements and have mutual trust within a VPC.
- A network ACL is an optional layer of security at the subnet level. You can control traffic in and out of a subnet by associating ACLs with it.

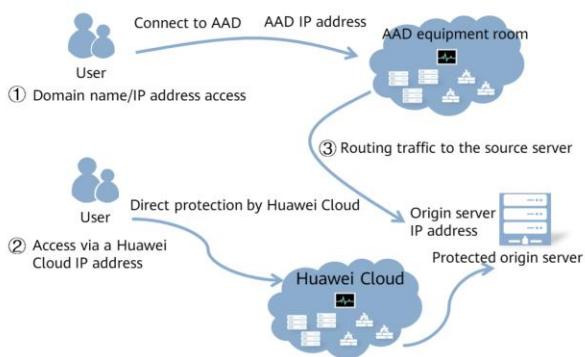
Anti-DDoS Service (ADS)

- DDoS (Denial of Service) attacks are also called flood attacks. They are intended to exhaust network or system resources on the target computer, which then becomes unresponsive. When an attacker compromises multiple computers to launch attacks on the targeted server, this is called Distributed Denial of Service Attack.
- Huawei Cloud Anti-DDoS Service (ADS) can defend against large-scale DDoS attacks. It includes Cloud Native Anti-DDoS Basic (Anti-DDoS traffic scrubbing), Cloud Native Anti-DDoS Advanced (CNAD Advanced), and Advanced Anti-DDoS. These products apply to different scenarios.

Huawei Cloud Anti-DDoS Products	Application Scenario
Cloud Native Anti-DDoS Basic	Defends public IP addresses on Huawei Cloud against DDoS attacks. Provides 2 Gbit/s DDoS attack defense for free. Its maximum defense capacity is 5 Gbit/s.
Cloud Native Anti-DDoS Advanced	Defends workloads deployed on Huawei Cloud and accessible via public IP addresses from large-scale DDoS attacks and ensures high network quality.
Advanced Anti-DDoS	Defends workloads deployed on Huawei Cloud, other clouds, and IDCs against DDoS attacks, ensuring the continuity of mission-critical services.

Advanced Anti-DDoS (AAD)

- Advanced Anti-DDoS (AAD) works as a proxy. It directs traffic to AAD IP addresses for scrubbing, ensuring the origin server is not exposed. It can be deployed on hosts used in Huawei Cloud, other clouds, and IDCs.



Product Features

Regulatory compliance: AAD complies with the anti-DDoS service requirements in DJCP 2.0 and other cybersecurity laws.

Massive bandwidth: AAD defends against network- and application-layer DDoS attacks. Its overall defense capacity is over 5 Tbit/s, including 600 Gbit/s for a single IP address.

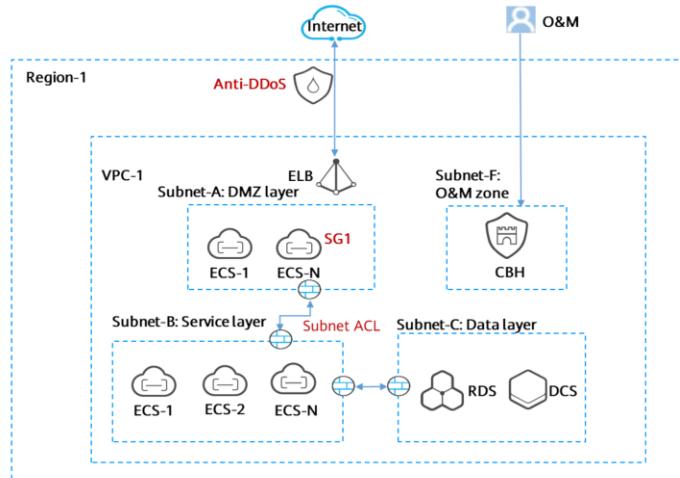
High availability: AAD automatically detects attacks and matches them with defense policies in real-time. Service traffic is distributed in clusters, yielding high performance, low latency, and good stability.

Elastic defense: You can purchase a basic bandwidth plus an elastic one. You can adjust the anti-DDoS defense baseline and upgrade the protection level at any time.

Professional operation team: A professional operation team responds to your questions and requests 24/7, safeguarding your applications and workloads.



Network Security Protection Products in the Cloud Architecture

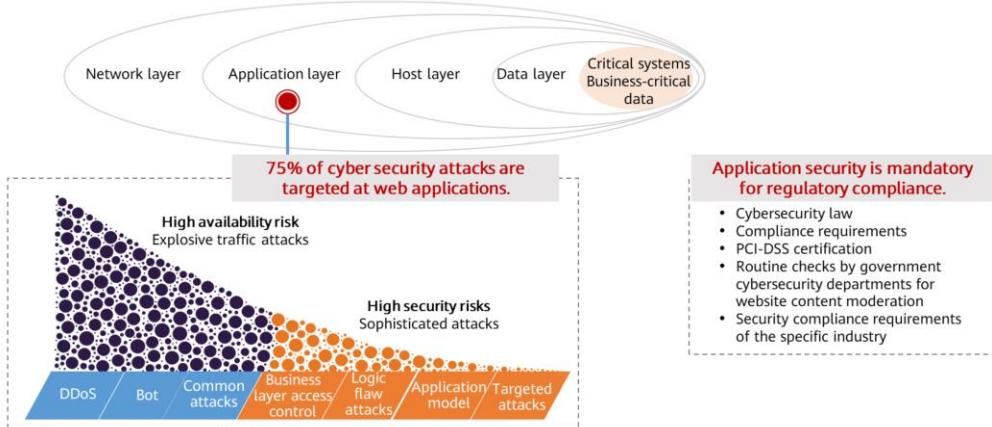


Contents

1. Cloud Security Design and Huawei Cloud Security System
2. Workload Security
3. Network Security
- 4. Application Security**
5. Data Security
6. Security Management

Application Security

- For applications on the cloud, application-level security is critical, as web application attacks are the most common type of online attack.

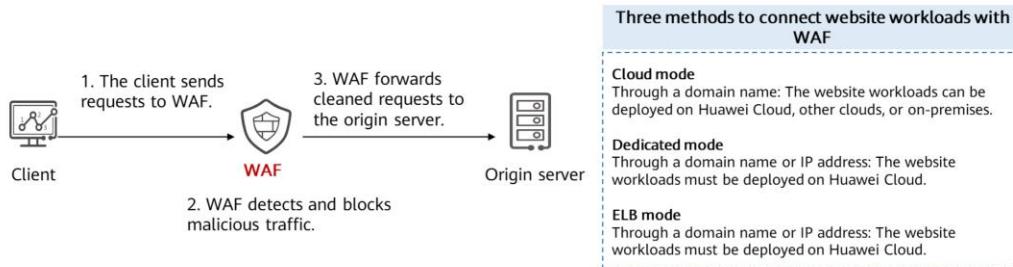


Application security is mandatory for regulatory compliance.

- Cybersecurity law
- Compliance requirements
- PCI-DSS certification
- Routine checks by government cybersecurity departments for website content moderation
- Security compliance requirements of the specific industry

Web Application Firewall (WAF)

- WAF is a security service that keeps your web applications secure and stable. It can block web attacks, such as SQL injection, cross-site scripting (XSS), command injection, Trojans, Challenge Collapsar (CC) attacks, and malicious web crawlers.



- In an SQL injection attack, an attacker tricks the database server into executing unauthorized queries. Attackers use exploits or logic flaws in application code to bypass security controls. They manipulate the database server behind a web application, tricking the system into doing what they want by executing specially constructed SQL statements.
- Cross-Site Scripting (XSS) is a common type of web security vulnerability. Attackers can exploit XSS vulnerabilities to inject malicious scripts into web pages that are provided for other users. In most types of attacks, there are only two parties involved: the attacker and the site they attack, but in an XSS attack, web clients, and web applications are also involved, so website visitors also suffer. XSS attacks are designed to steal cookies stored on a client or sensitive information used by other websites to identify a client.
- In command injection attacks, attackers construct and submit special command strings to embedded or web applications as these applications typically do not check data submitted by users very strictly. After receiving the constructed commands, applications are tricked into executing external programs or launching OS attacks so that attackers can steal data or network resources.
- In a Trojan attack, attackers upload a Trojan to a legitimate website. When a user visits the website, the Trojan is downloaded and executed automatically. The user's computer is attacked and even manipulated by the attacker.
- Challenge Collapsar (CC) attacks are web attacks against web servers or applications. In CC attacks, attackers send a large amount of standard GET/POST requests to target system to exhaust web servers or applications. For example, attackers can send requests to URLs of databases or other resources to make the servers unable to respond to normal requests.

WAF Application Scenarios

Threats to web applications: ✓

Other types of threats: x

Basic protection

WAF helps customers defend against common web attacks, such as command injection and sensitive file access.

Protection for promotions on e-commerce platforms

A large number of malicious requests may be submitted to service interfaces during online promotions. 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 requests.

Protection against zero-day vulnerabilities

If website services fail to recover quickly from impacts 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.

Data leak prevention

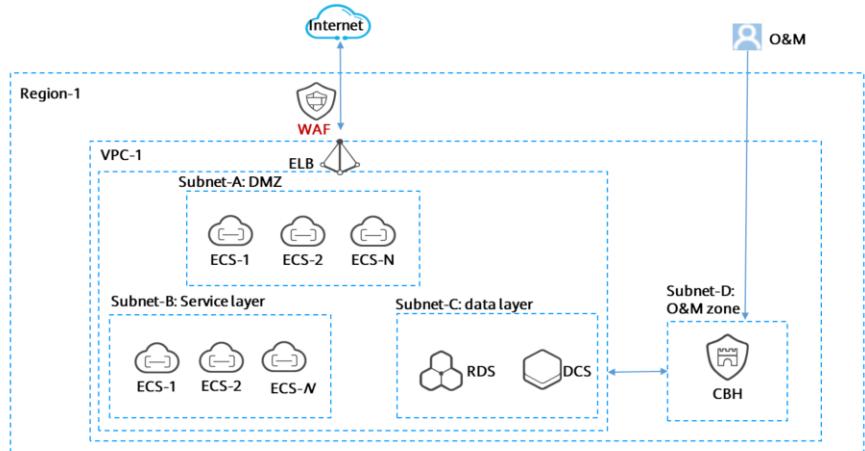
WAF prevents malicious actors from using methods such as SQL injection and web shells to bypass application security and gain remote access to web databases and other sensitive information. Users can configure information leakage prevention rules on WAF. WAF uses semantic analysis and regular expression engines to accurately detect attack traffic.

Web tamper prevention

WAF protects customer credibility by ensuring attackers cannot leave backdoors on web servers or tamper with web page content. Users can configure web tamper protection rules in WAF to scan for malicious code and protect website visitors from phishing attacks.

- A zero-day vulnerability is a vulnerability in a system or device that has been disclosed but has not been patched yet. No one except the one who discovered the vulnerability is aware of it. This person may exploit the vulnerability to launch attacks, and such attacks are often unpredictable and destructive.

Application Security Products in a Cloud Architecture



28 Huawei Confidential

HUAWEI

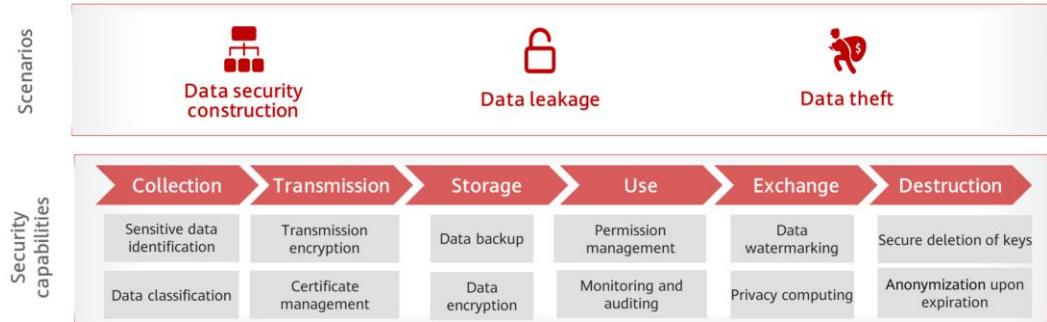
- The solid line indicates the access traffic.
- Demilitarized Zone (DMZ) is a special network area different from the external network or internal network. Generally, the DMZ houses the public servers that do not contain confidential information, for instance, web servers, email servers, or FTP servers. Users from the external network can only access the services in the DMZ, but cannot access the information on the internal network. So, the information on the internal network cannot be impacted even if the servers in the DMZ were attacked.

Contents

1. Cloud Security Design and Huawei Cloud Security System
2. Workload Security
3. Network Security
4. Application Security
- 5. Data Security**
6. Security Management

Data Assets Security

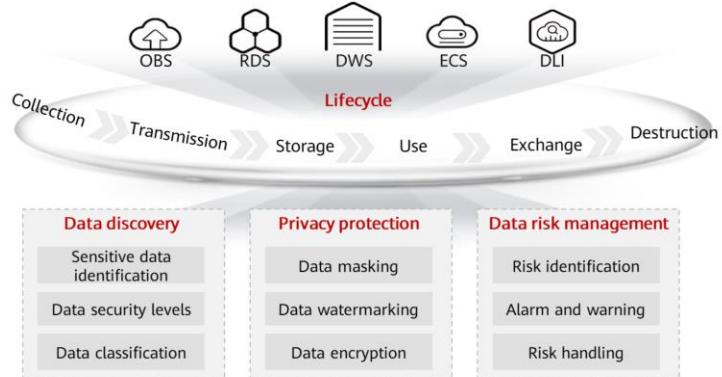
- Data security is critical for cloud computing and is evolving from border defense to security management covering all phases of data lifecycle.



The global average total cost of data breaches increased by 10% from 2020 to 2021.

Data Security Center (DSC)

- DSC is a new cloud data security platform.
- It protects data security in all phases of data lifecycle. This includes sensitive data identification, privacy protection, and risk management.

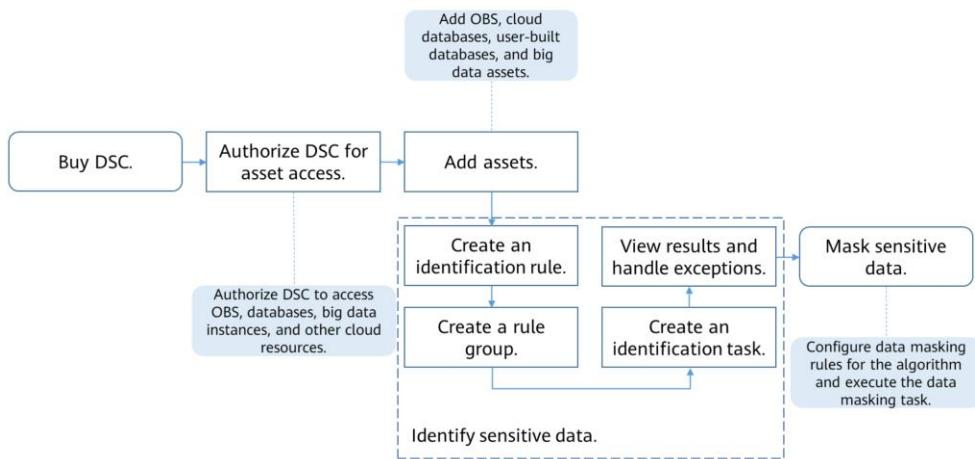


31 Huawei Confidential



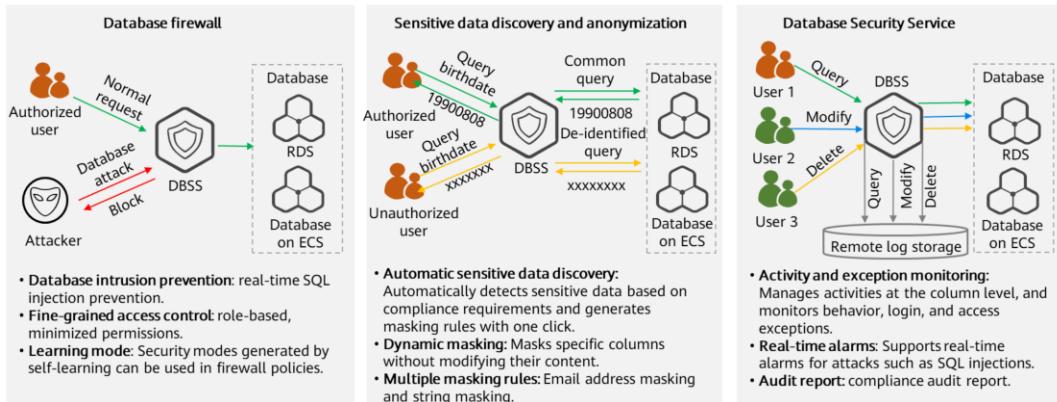
- Application scenarios:
 - Data identification: Scanning massive amounts of data, DSC can automatically identify sensitive data and analyze how it is being used. It also scans structured data in RDS and unstructured in OBS, and classifies the data by risk level for further security handling.
 - Behavior analysis: DSC analyzes user behavior, using deep learning to establish a user behavior library. Any behavior uncovered in the library is deemed abnormal and an alarm will be reported in real time. You can then trace user behaviors and correlate the events with the users to identify who performed the risky operations. DSC also detects data breaches and generates alarms so that you can take immediate protective actions.
 - Data masking: The DSC data masking engine leverages a wide range of preset and user-defined masking algorithms. It then masks structured and unstructured data for storage.
 - Compliance: DSC provides dozens of templates that can be used to check for compliance with regulations and standards such as GDPR, PCI DSS, and HIPAA. DSC checks your data protection measures against multiple rules stored in templates, and generates reports to propose corrective measures.

DSC Service Process



Database Security Service (DBSS)

- DBSS intelligently enhances data security during database running. Based on the machine learning mechanism and big data analytics technologies, the service can audit your databases, detect SQL injection attacks, and identify high-risk operations.



Digital Certificate Used for Identity Assurance During Data Transmission

- A digital certificate is an electronic authentication provided for secure communication between two parties. It implements identity verification and electronic information encryption the Internet and intranet.

How to ensure data security when transferring it to a website?



Network O&M personnel	User
How can I prove my identity to my visitors?	Is the website legitimate?
How can I build a secure transmission channel with my customers?	The URL shows it is the website of xx company. But can I trust it?
What technical measures can be taken to align with laws and regulations?	Will my usernames and passwords be stolen?

- A certificate is issued by a certificate authority (CA) to authenticate a user's public key.
- A digital certificate contains the identification information about the owner of a key pair (a public key and a private key). The identity of the certificate owner is authenticated by verifying the authenticity of the identification information.

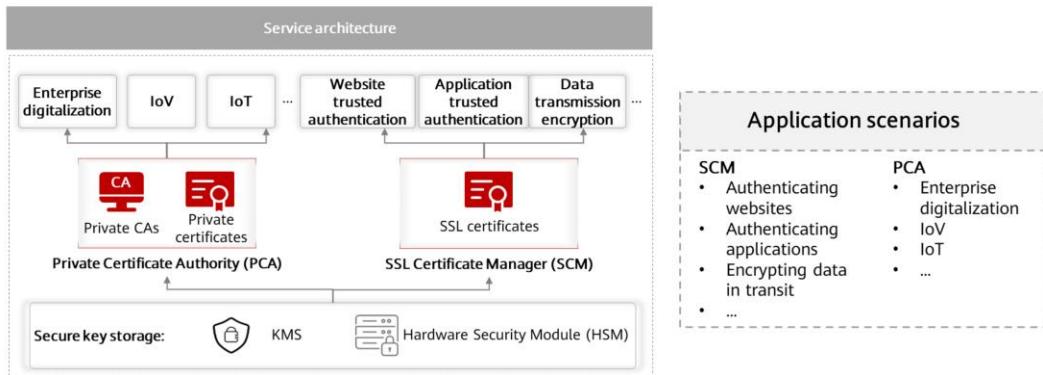
Digital certificates are classified into the following types:

- Public certificate**, which can be used by a web server to identify websites that use Secure Sockets Layer (SSL)/Transport Layer Security (TLS), and to establish encrypted connections to these websites.
- Private certificate**, which identifies and protects resources such as applications, services, devices, and users in an organization.

- Public certificate can be used by a web server to identify websites that use Secure Sockets Layer (SSL)/Transport Layer Security (TLS), and to establish encrypted connections to these websites.
 - A public certificate is issued by a public CA to authenticate resources on the Internet.
 - A public certificate is trusted by applications and browsers by default, because the corresponding CA root certificates have been stored in the trusted area of the browser and OS.
 - A public certificate complies with security standards specified by browser and operating system vendors and provides operation visibility.
- Private certificates identify and protect resources such as applications, services, devices, and users in an organization.
 - A private certificate is issued by a private CA and is used for authenticating internal resources of an organization.
 - Servers, websites, clients, devices, and VPN users
 - Resources in a private network
 - Untrusted by default: You need to install private certificates in the trusted zone on the client.
 - Advantages:
 - It can be used to identify any resource.
 - Users can define issuance rules for verification and naming.
 - It is not restricted by public CA certificate/agency rules.

Cloud Certificate Management (CCM)

- CCM is a cloud service that lets users easily manage millions of SSL and private certificates in one platform.



35 Huawei Confidential



- Currently, the SSL certificates issued by international certificate authorities are valid for one year. In CCM, users can configure a rotation schedule for a private certificate based on its expiration date.
- SSL certificate management:
 - Building a trusted website. SSL certificates can authenticate websites to effectively prevent the websites from being forged.
 - SSL certificates can also authenticate cloud and mobile applications. For example, a wide range of cloud applications, such as customer relationship management (CRM), office automation (OA), and enterprise resource planning (ERP) applications, can be authenticated to prevent unauthorized access.
 - SSL certificates can encrypt transmission between websites, applications, and clients. This effectively ensures data integrity and prevents data in transit from being stolen or tampered with.
- Private Certificate Authority (PCA)
 - Enterprises can use PCA to establish a unified certificate management system and manage certificates throughout the entire lifecycle. The system integrates continuous monitoring and automation to reduce the risk of improper certificate management.
 - Telematics Service providers (TSPs) use PCA to issue certificates to vehicular terminal, thus providing security capabilities such as authentication and encryption during vehicle-vehicle, vehicle-cloud, and vehicle-road interactions.
 - Internet of Things (IoT) platforms can use PCA to issue certificates to IoT devices for identity authentication, ensuring that only authenticated devices are connected.

Data Encryption Workshop (DEW)

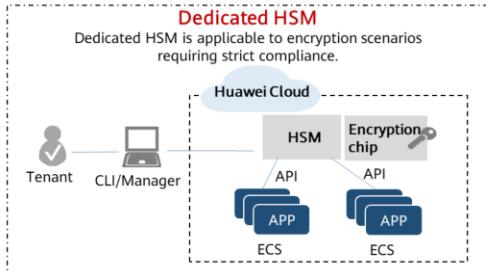
- DEW is a comprehensive cloud data encryption service. It uses keys to encrypt data.
- DEW provides Key Management Service (KMS), Key Pair Service (KPS), Dedicated Hardware Security Module (Dedicated HSM), and Cloud Secret Management Service (CSMS).

Scenario	DEW Service Module
Use HSMs to encrypt data on the cloud.	Dedicated HSM
Use keys for security management of cloud resources.	KMS
Manage SSH key pairs.	KPS
Host all the secrets in a system.	CSMS

- Dedicated HSM: A customer can use dedicated HSMs to meet strict compliance requirements (large-scale high concurrency services, such as payment services).
- KMS is used for cloud service encryption (integrated in cloud services), data disk encryption, and small-size data encryption.
- KPS is used for server login.
- CSMS is used for password and token storage.

DEW Service Module - Dedicated HSM

- Dedicated Hardware Security Module (Dedicated HSM) is a service provided by Huawei Cloud for encryption, decryption, signature, signature verification, key generation, and the secure storage of keys.



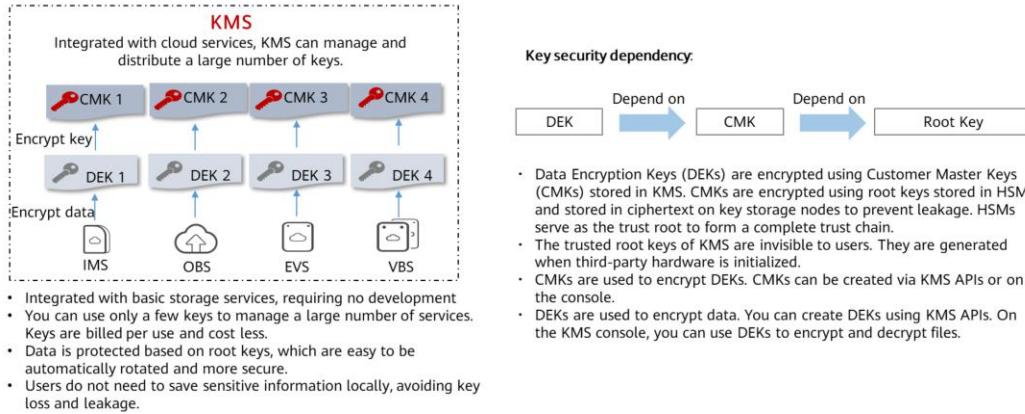
- Secondary development is required for using API requests to connect with applications.
- Dedicated HSMs ensure encryption performance.
- International algorithms are supported, meeting compliance requirements.

Item	Local HSM	Dedicated HSM
Connection costs	High	N/A
Subsequent O&M costs (OPEX)	High. Physical equipment requires regular on-site support from the vendor.	Low. Only the cloud service provider and HSM provider need to perform online maintenance, saving offline O&M manpower.
Key usage	Local network only	Cloud resources
Encryption resources	Not scalable	Scalable
User controls access keys	Satisfied	Compliant

- If you have purchased an instance, you can use Dedicated HSM to initialize and manage the instance. You can fully control the generation and storage of keys, as well as access authentication for keys.

DEW Service Modules - KMS

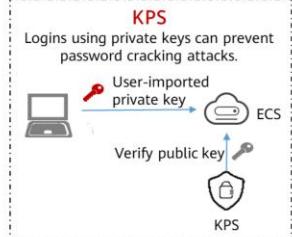
- Key Management Service (KMS) is a secure, easy-to-use service that uses HSMs to protect your keys. It seamlessly interworks with other services to protect service data and can be used to develop encryption applications.



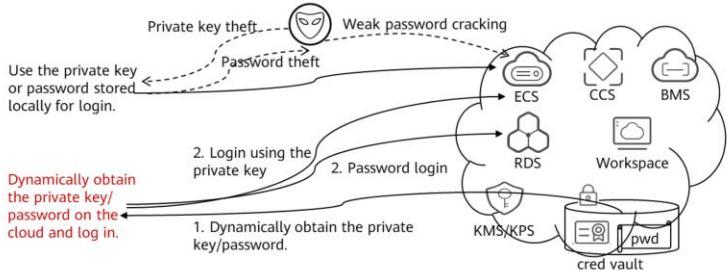
- KMS is integrated with a range of Huawei Cloud services. Customers can create keys on the KMS console or import external keys to encrypt data stored in more than 45 cloud services, such as RDS, ECS, OBS, SFS, DDS and EVS.

DEW Service Module - KPS

- Key Pair Service (KPS) is a secure, reliable, and easy-to-use service that helps users centrally manage and protect SSH key pairs.



- Enables you to log in to ECSs with key pairs, secure and convenient.
- Local private keys can be hosted, avoiding any loss.
- Key pairs can be reset or replaced, providing higher security.



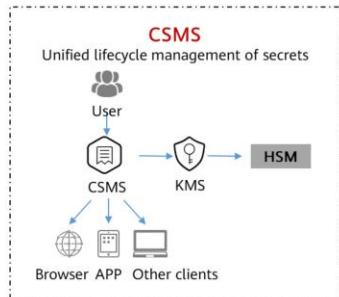
- KPS uses true random numbers generated by HSMs to generate key pairs, and provides a comprehensive and reliable key pair management solution to prevent brute force cracking of weak passwords, mitigate the threats of private key theft, and prevent credential stuffing attacks.
- Users can easily create, import, and manage SSH key pairs in KPS. The public key of a generated key pair is stored on Huawei Cloud while the private key can be downloaded and saved separately, which ensures the privacy and security of the key pair.

- A pair of public and private keys are used in the encryption method commonly known as the asymmetric encryption method. The key pair, consisting of a public key and a private key, is generated based on an algorithm. The public key is open while the private key is not. A public key can be used to encrypt a session key, verify a digital signature, or encrypt data that can be decrypted using a private key. The public and private key pair is unique across the whole world. If one key is used to encrypt a piece of data, the other key must be used to decrypt the data. If you use either key to encrypt a piece of data, the encrypted data can only be decrypted using the other key or the decryption fails.
- RDS/WKS password management is enhanced. Users do not need to record their passwords. Strong passwords are randomly generated, blocking credential stuffing attacks. A key pair can be dynamically bound to an ECS. Users can switch to the key pair login mode in one click and avoid using weak passwords.
- Private keys and passwords are not statically stored on the user side, reducing the risk of private key and password leakage. KMS and KPS manage private keys and regularly rotate keys in a unified manner, reducing the attack time window. Private keys and passwords are encrypted by KMS/KPS on the cloud and then securely stored. They are dynamically obtained after IAM/MFA authentication. They are easy to use and can be accessed anytime, anywhere. Users can use IAM credentials and MFA to obtain private keys and passwords anywhere to access resources.

DEW Service Module - CSMS

- CSMS is a secure, reliable, and easy-to-use secret hosting service.

Reducing key exposure is important to key security. CSMS implements full lifecycle management of secrets. The differences between CSMS and traditional key management are as follows.



- Secrets are encrypted for storage.
- Secrets are queried via APIs to reduce exposure.
- Secrets and keys are rotated to improve security.

Scenario	Traditional Method	CSMS
Unified secret management	Scattered sensitive information and inefficient management	The storage, retrieval, and usage of secrets can be managed in a unified manner and are more reliable.
Secure secret retrieval	Identity information used for verification is often written in plaintext in configuration files, susceptible to data breaches.	Dynamic secrets can be queried via API. Sensitive information is not contained in query results and is highly secure.
Secret and key rotation	It is difficult to update secrets for all applications, which may interrupt services.	Through version management, APIs/SDKs can be called to quickly rotate secrets at the application layer.

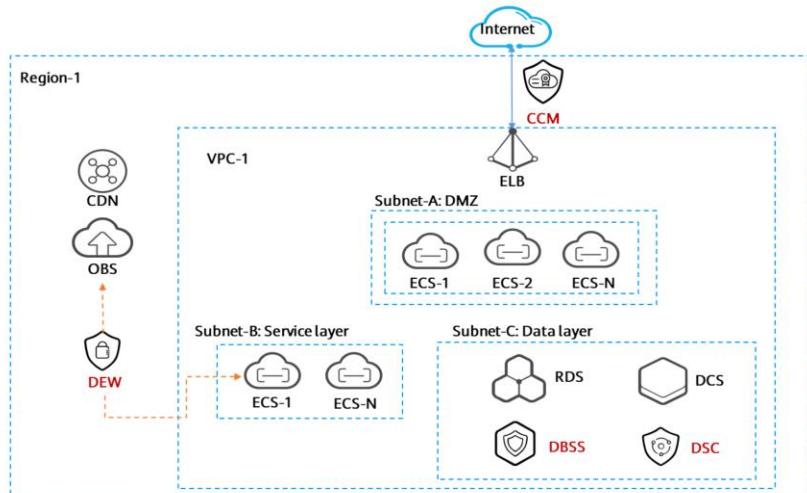
- Users or applications can use CSMS to create, retrieve, update, and delete credentials in a unified manner throughout the credential lifecycle. CSMS can help you eliminate risks that stem from insecure practices such as hardcoded, plaintext configuration, and inadequate permission control.

DEW Application Scenarios

- DEW service modules can be used to meet different data encryption requirements.

Scenario	Dedicated HSM	KMS	KPS	CSMS
Storage service encryption DEW is integrated with OBS, EVS, VBS, and IMS to create encrypted storage.		Recommended		
Application encryption Users can call APIs to encrypt applications.	Recommended	Recommended		
High-performance encryption and decryption Meet cryptographic computing requirements in a range of industry applications, such as identity authentication, data protection, and SSL offloading.	Recommended			
Financial cryptography Cryptographic calculation in financial systems, such as card issuing systems and point of sale (POS) systems	Recommended			
Signature verification service Signature usage in Certificate Authority (CA) systems, encrypted transmission of a large amount of data, and identity authentication	Recommended			
Defense against brute-force attacks A user can log in to a server through private key authentication instead of password authentication.			Recommended	
Sensitive data management Sensitive data storage, including passwords, access keys, OAuth keys, tokens, and API keys				Recommended

Data Security Products in the Cloud Architecture



42 Huawei Confidential

HUAWEI

- In this figure, DEW modules include KPS and KMS.

Contents

1. Cloud Security Design and Huawei Cloud Security System
2. Workload Security
3. Network Security
4. Application Security
5. Data Security
- 6. Security Management**

Security Management (1)

- Identity management is a top priority for security management. Unauthorized access and abuse of user credentials are common security issues.



Verizon investigated 41,686 security incidents, of which

- **15%** involved abuse by authorized users
- **29%** involved use of stolen credentials
- **56%** took months or longer to detect
- **36%** involved internal personnel or partners

Source: *2019 Data Breach Investigations Report* from Verizon



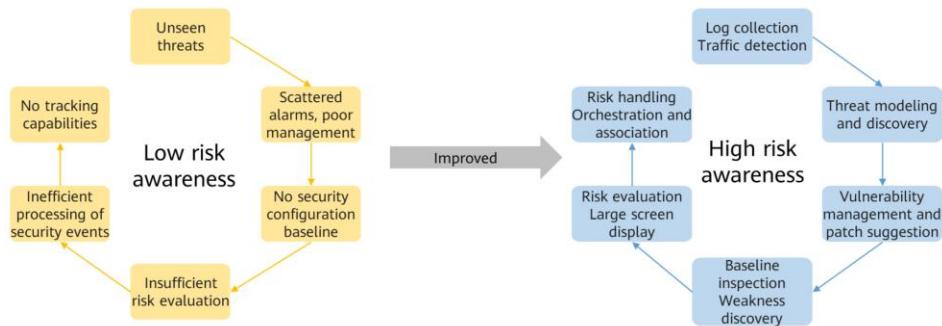
Continuous authentication

- Least privilege access
- Zero trust, a security framework requiring all users, devices, and applications, whether in or outside the organization's network, to be authenticated, authorized, and continuously validated for security configuration

- Verizon is the largest wireless carrier in the United States, with over 140 million subscribers.

Security Management (2)

- Proactively identifying and handling security risks can minimize losses.
- Focusing on situation awareness improves O&M efficiency and enhance system security.



Security Management (3)

- Security management helps ensure security compliance. Security compliance design covers security design and continuous improvement based on compliance requirements, effective monitoring of security risks and security events, and taking countermeasures in a timely manner to continuously reduce security risks and losses caused by security incidents. Security design also requires routine drills to ensure security system availability.

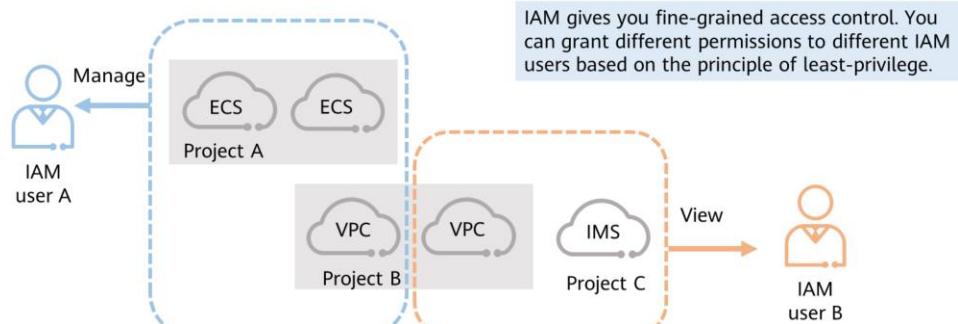
Laws and regulations	National standards	Industry standards
<ul style="list-style-type: none">• Cybersecurity Law of the People's Republic of China• The Cryptography Law of the People's Republic of China• <i>Cybersecurity Review Measures</i>• <i>Data Security Law of the People's Republic of China</i>• Personal Information Protection Law of the People's Republic of China...	<ul style="list-style-type: none">• GB/T 37973-2019 Information Security Technology—Big Data Security Management Guide• Information Security Technology - Personal Information Security Specification• Information Security Technology — Baseline for Classified Protection of Cybersecurity...	<ul style="list-style-type: none">• Implementation Guide for Classified Protection of Information System of Financial Industry• Several Provisions on the Management of Automotive Data Security (for Trial Implementation)• National Energy Administration Guidelines on Strengthening Cyber Security in the Electric Power Industry...

- Companies can use professional services offered by cloud service providers to improve confidence in security management.

- In digital transformation, companies face stringent security compliance requirements. Complying with security requirements is a huge responsibility, and non-compliance may result in severe penalties. Security compliance is the first and most important thing that enterprises are concerned with in cloud migration. Compliance standards determine the security level companies need to be able to comply with on the cloud.

Identity and Access Management

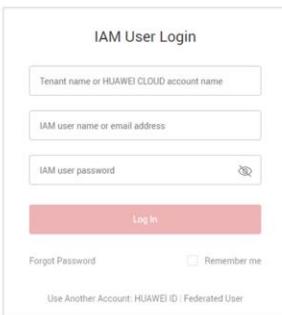
- Identity and Access Management (IAM) enables you to easily manage users and control their access to Huawei Cloud services and resources.



- A project can contain different resources. You can attach policies to different user groups to grant permissions for accessing specific resources. In the figure, user A is granted access to all resources in project A and to specific resources in project B. User B is granted access to specific resources in project B and all resources in project C.

IAM User Credentials

1. Password



IAM User Login

Tenant name or HUAWEI CLOUD account name

IAM user name or email address

IAM user password

Log In

Forgot Password

Remember me

Use Another Account: HUAWEI ID · Federated User

Password as the login credential

2. Access Key ID/Secret Access Key (AK/SK)



Access Keys

Access keys can be downloaded only once after being generated. Keep them secure, change them periodically, and do not share them with anyone.

Create Access Key

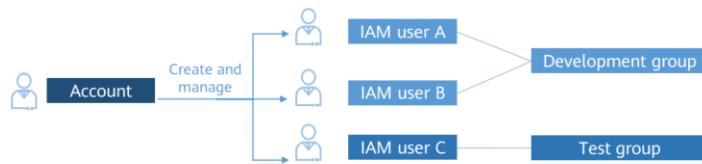
Access keys available for creation: 2

- Each IAM user can create two pairs of access keys.
- Each access key file can be downloaded only once.
- Access keys can be used for API access.
 - An AK contains 20 characters.
 - An SK contains 40 characters.

- AK: An access key ID is a unique ID associated with an SK. An AK is used together with an SK to cryptographically sign requests.
- SK: A secret access key is used in conjunction with an AK to sign requests cryptographically. It identifies a request sender and prevents the request from being modified.

IAM Basic Concepts

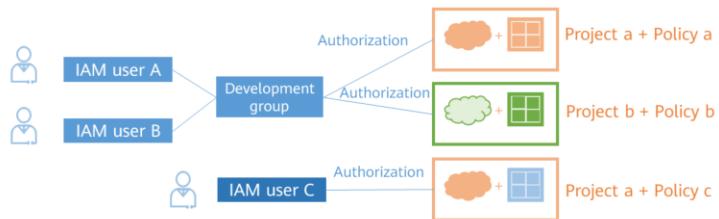
- **Account:** An account is created after you successfully register with Huawei Cloud. An account owns resources and pays for the usage of resources. An account has full access to the resources it owns. As a best practice, do not use the account to perform operations.
- **IAM user:** You can use your account to create IAM users and assign permissions for specific resources. Each IAM user has their own identity credentials (password and access keys) and uses cloud resources based on assigned permissions.
- **IAM user group:** An IAM user group is a collection of IAM users, which facilitates batch permissions control. An IAM user can be added to different IAM user groups.



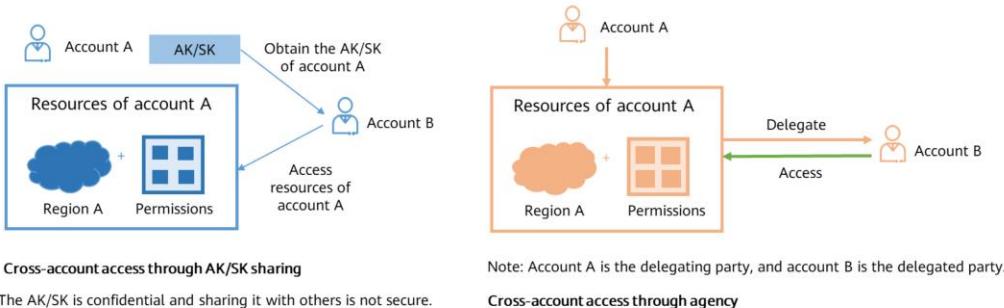
- IAM users do not have their own resources and cannot pay for the resources they use. The account assigns permissions to IAM users and pays for the use of the resources.
- You can assign permissions to IAM users through user groups. By default, new IAM users do not have any permissions assigned. To assign permissions to new users, add them to one or more groups, and assign permissions to these groups. The users then inherit permissions from the groups they belong to, and they can perform operations on cloud services based on the assigned permissions.

Fine-grained Access Control for Huawei Cloud Resources

- **Authorization:** The process of granting required permissions for a user to perform a task. After you attach a system-defined or custom policy to a user group, users in the group can inherit the permissions defined by the policy to access resources.
- **Policy:** A policy defines who can access what resources under which conditions based on the principle of least privilege. Policies enable you to achieve flexible fine-grained permissions management.
- **Project:** A region corresponds to a project by default. You can create subprojects in a project, purchase resources in the subprojects, and grant permissions by subproject. Projects physically isolate resources across regions. IAM users can only access resources in authorized projects.



Cross-Account Resource Access

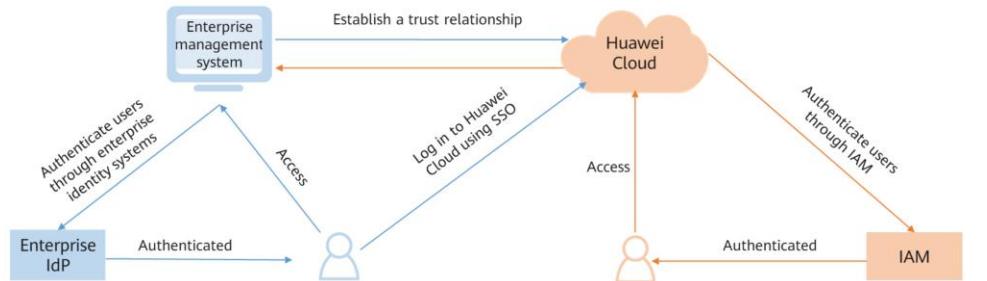


- **Agency:** You can delegate your account permissions to other Huawei Cloud accounts or services for more efficient O&M.

- **Account delegation:** You can delegate permissions to other Huawei Cloud accounts only. You cannot delegate permissions to federated accounts or IAM users.
- **Cloud service delegation:** Huawei Cloud services interwork with each other. Some cloud services depend on other services. You can create an agency to delegate a cloud service to call other services on your behalf. For example, if Container Guard Service (CGS) needs to scan container images, you need to delegate SoftWare Repository for Container (SWR) permissions to CGS.

Cloud Resources Access Using Your Enterprise Account

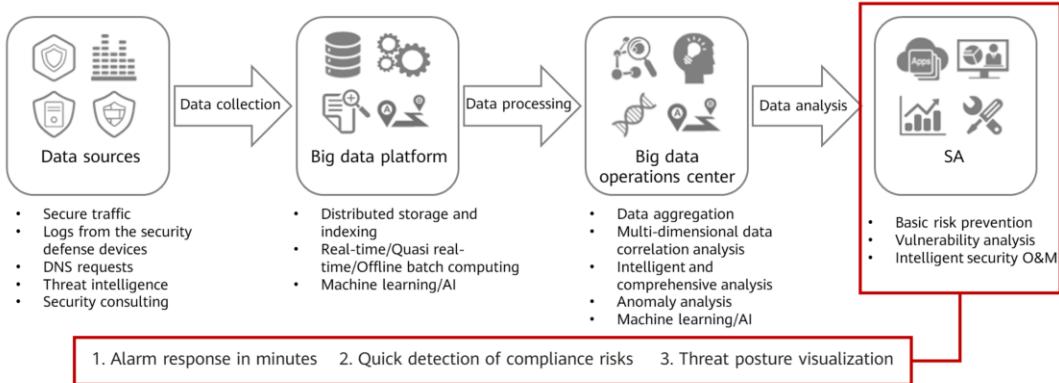
- **Identity federation:** IAM supports IdPs that are compatible with Security Assertion Markup Language 2.0 (SAML 2.0) and OpenID Connect (OIDC). You can use an identity provider (IdP) to enable the identities in your enterprise system to log in to Huawei Cloud using single sign-on (SSO).



- OpenID Connect (OIDC): a standard identity authentication protocol that runs on top of the OAuth 2.0 protocol.
- Security Assertion Markup Language (SAML): Security Proposition Markup Language. It is an XML-based open-standard for transferring identity data between two parties: an identity provider (IdP) and a service provider (SP).
- Identity provider (IdP): collects and stores user identity information, such as usernames and passwords, and authenticates users during login. For identity federation between an enterprise and Huawei Cloud, the IdP refers to the identity authentication system of the enterprise.
- Identity federation process:
 - Create an IdP and establish a trust relationship.
 - OIDC-based IdP: Create OAuth 2.0 credentials in the enterprise IdP and create an IdP in Huawei Cloud to establish a trust relationship between the enterprise and Huawei Cloud.
 - SAML-based IdP: Exchange the metadata files (SAML 2.0-compliant interface files that contain interface addresses and certificate information) of the enterprise IdP and Huawei Cloud. Then, create an IdP in Huawei Cloud to establish a trust relationship between the enterprise and Huawei Cloud.
 - Configure identity conversion rules: Map the users, user groups, and their permissions in the enterprise IdP to Huawei Cloud.
 - Configure a login link: Configure a login link in the enterprise management system to allow users to access Huawei Cloud using SSO.

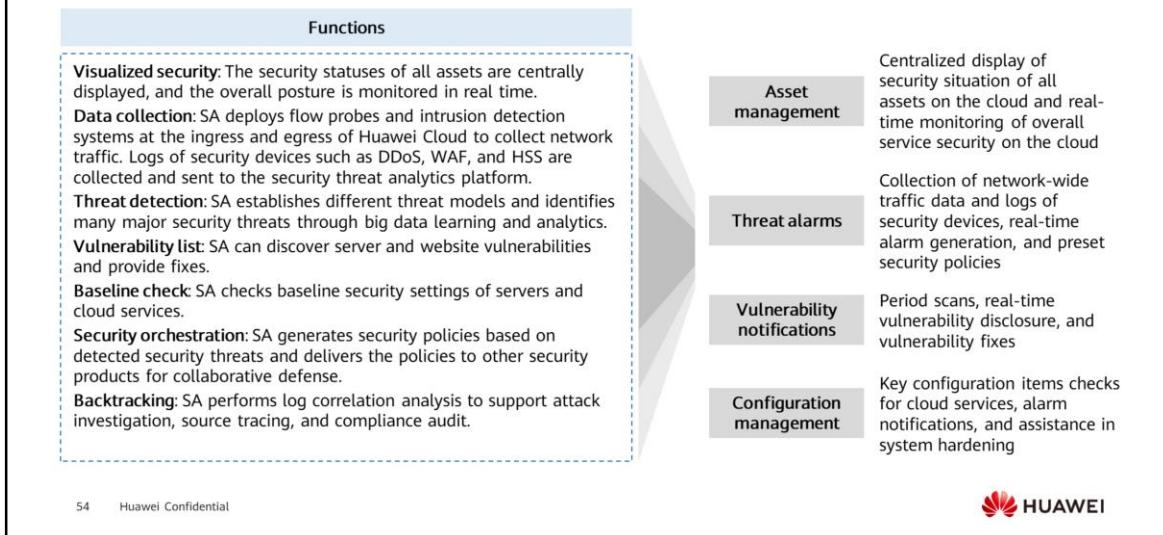
Situation Awareness (SA)

- SA is a security management and situation analysis platform. It can detect more than 20 types of cloud security risks and leverage big data to give users a comprehensive overview of their security posture.



- After data is collected, it is batch processed by the big data platform and then analyzed by the big data operations center. Analysis results are reported to SA so that SA can take appropriate protective actions such as event analysis and alarm reporting.

SA Application Scenarios

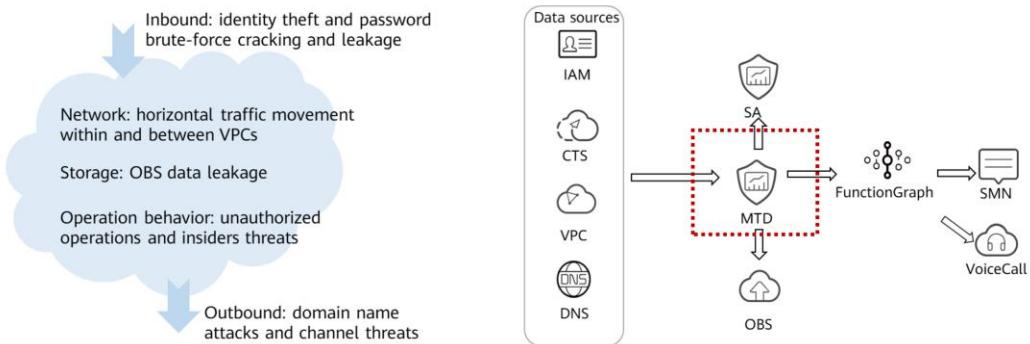


- Asset management: As enterprises migrate more workloads to the cloud, more cloud assets are used, and there are frequent changes made to those assets. This means more security risks on the cloud.
 - SA gives customers a comprehensive view of the security status of assets on the cloud. SA monitors the security status of all assets in the cloud in real time and visualizes vulnerabilities, threats, and attacks on servers, making it easier for customers to handle risks.
- Threat event alarms: Security threats to clouds never stop, and a variety of new threats are emerging every day.
 - By collecting network-wide traffic data and security device logs, SA can detect and monitor security risks on the cloud in real time, display statistics on security events in real time, and aggregate event data from other security services. SA uses preset security policies to effectively defend against common brute-force attacks, web attacks, Trojans, and zombie bots, greatly improving defense and O&M efficiency.
- Vulnerability notifications: Service security is of top priority during cloud migrations. To prevent vulnerabilities from being exploited, we need to find and fix as many vulnerabilities as possible.
 - Apart from reporting latest vulnerabilities based on emergency security notices issued on Huawei Cloud, SA periodically scans OSs, software, and websites for vulnerabilities by working with linked security services, making it easier for customers to centrally manage server and website vulnerabilities. SA also provides mitigation suggestions. With centralized vulnerability management on the cloud, SA helps customers quickly identify key risks and vulnerable assets and harden their service system.
- SA can scan for unsafe settings of cloud services, report scan results by category, generate alarms for unsafe settings, and provide hardening suggestions and

guidelines.

Managed Threat Detection (MTD)

- MTD continuously monitors cloud service logs in real time and generates alarms for malicious activities and unauthorized behavior.



55 Huawei Confidential



- MTD collects logs from IAM, DNS, CTS, OBS, and VPC and uses an AI engine, threat intelligence, and detection policies to continuously detect potential threats, malicious activities, and unauthorized behavior, such as brute-force cracking, penetration attacks, and mining attacks.
- Inbound bandwidth is the bandwidth consumed when data is transferred from the Internet to Huawei Cloud. For example, when resources are downloaded from the Internet to ECSs, that consumes inbound bandwidth.
- Outbound bandwidth is the bandwidth consumed when data is transferred from Huawei Cloud to the Internet. For example, when ECSs provide services accessible from the Internet and external users download resources from the ECSs, that consumes outbound bandwidth.

MTD Application Scenarios

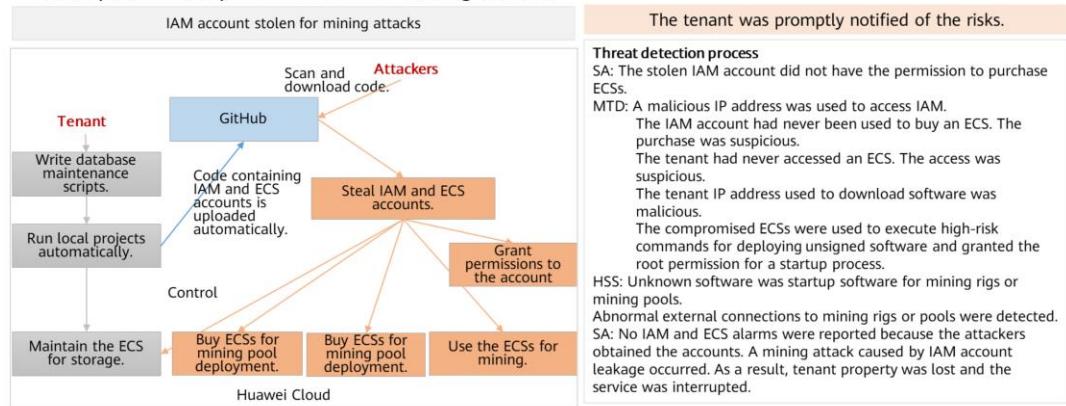
DJCP Multi-level protection scheme (MLPS) compliance	Network protection and key event assurance	Routine operations
DJCP 2.0 general requirements and special security requirements for cloud computing: border defense, access control, intrusion prevention, malicious code, security audit, centralized management and control, network architecture, and image and snapshot security	Annual network protection and key assurance for governments, companies, state-owned enterprises, finance institutions, and education organizations. Continuous detection of malicious activities and unauthorized behavior, alarm reporting, and handling during the attack-defense exercises for large IT companies	Identification of potential threats, including VPC attacks and malicious behavior in a wide range of cloud service networks, storage devices, and operation logs. Threat analysis and service security assurance

Differences Between MTD and SA

Feature	MTD	SA
Supported product/service	<ul style="list-style-type: none">Identity and Access Management (IAM)Domain Name Service (DNS)Cloud Trace Service (CTS)Virtual Private Cloud (VPC)Object Storage Service (OBS)	<ul style="list-style-type: none">Host Security Service (HSS)Anti-DDoSWeb Application Firewall (WAF)Cloud Bastion Host (CBH)Container Guard Service (CGS)Vulnerability Scan Service (VSS)
Data source detection/analysis	<ul style="list-style-type: none">IAM logsDNS logsCTS logsVPC logs of global servicesOBS logs	<ul style="list-style-type: none">Mobile trafficLogs from security defense devicesDNS requestsThreat intelligenceSecurity information
Threat detection	<ul style="list-style-type: none">MTD reports over 40 types of alarms for threats detected using the AI engine, threat intelligence, and detection policies.	<ul style="list-style-type: none">SA detects and displays eight types of alarm events and provides over 200 second-level alarm types. You can set alarm notifications to quickly take proactive actions.SA allows you to implement preset security orchestration policies to strengthen asset security.

SA Collaborates with MTD to Detect Identity Risks

- Incidence background: An IAM account for database O&M was accidentally uploaded to open source communities (such as GitHub) in plaintext. Attackers easily obtained the account and logged in to the cloud platform to purchase ECSs for mining attacks.



58 Huawei Confidential



- MTD uses advanced detection technologies, such as threat intelligence, AI detection engine, and correlation models, to scan IAM, CTS, VPC, and DNS service logs for cracking attacks. Additionally, MTD tracks and audits network behaviors, and identifies traffic changes of network devices and servers for an abnormal number of connections. MTD reports alarms to SA and collaborates with other security services for overall situation monitoring. System security issues can be detected in a timely manner.
- MTD identifies threats to IAM accounts and vulnerabilities to DNS and looks for intrusions by checking CTS logs. These security risks cannot or can barely be detected by other security services. When risks increase, multi-factor verification or biometric recognition is required by MTD for using an IAM account.

Quiz

1. (True or False) Compliance is a secondary demand for cloud security. If the budget is limited, compliance requirements can be ignored.

True

False

2. (Single-answer question) Which of the following products is most suitable for directly encrypting storage services such as OBS on the cloud?

- A. KMS
- B. HSS
- C. KPS
- D. DHSM

- 1. Answer: False.
 - Security compliance is the foundation of businesses on the cloud. If companies do not meet compliance requirements, they may be punished financially.
- 2. Answer: A.
 - KMS is integrated into cloud services.

Quiz

1. (Discussion) A company that recently migrated their services to the cloud needs to quickly build a security protection system. What factors should be considered when selecting security products and which products are suitable for this company?

2. (Discussion) In addition to the security products, how can this company improve cloud security?

- Discussion 1:
 - The five security dimensions should be considered.
 - Application scenarios and features of Huawei Cloud security products should be considered.

- Discussion 2:
 - Security organizations and personnel: Internal security and personnel security awareness of key positions, core services, and confidential services of the company
 - Infrastructure security: Isolated deployment of service planes (secure traffic distribution) and water-proof, electricity, and equipment room security of on-premises resources
 - Engineering security: Secure coding and review and approval processes of third-party software
 - O&M security: Business continuity planning and testing (periodical testing of DR and infrastructure HA)

Summary

- This course describes cloud security, five dimensions of security services, and security products and application scenarios in each dimension.
- You now know about cloud security design. We will learn about some Huawei Cloud services in the next course.

Acronyms and Abbreviations

- AAD: Advanced Anti-DDoS
- ACL: Access Control List
- ADS: Anti-DDoS Service
- AK: access key ID
- API: application programming interface
- ATC: Application Trust Center
- CA: certificate authority
- CBH: Cloud Bastion Host
- CCE: Cloud Container Engine
- CDN: Content Delivery Network
- CFW: Cloud Firewall
- CGS: Container Guard Service
- CSA: Cloud Security Alliance
- CSMS: Cloud Secret Management Service
- DC: data center
- DEW: Data Encryption Workshop
- DHSM: Dedicated Hardware Security Module
- DMZ: Demilitarized zone
- DNS: Domain Name Service
- DSC: Data Security Center
- ECS: Elastic Cloud Server
- ELB: Elastic Load Balance
- EVS: Elastic Volume Service

Acronyms and Abbreviations

- HSS: Host Security Service
- IAM: Identity and Access Management
- IMS: Image Management Service
- KMS: Key Management Service
- KPS: Key Pair Service
- MDR: Managed Detection Response
- MTD: Managed Threat Detection
- OBS: Object Storage Service
- OIDC: OpenID Connect
- RDS: Relational Database Service
- SA: Situation Awareness
- SAML: Security assertion markup language
- SFS: Scalable File Service
- SK: secret access key
- SOC: Security Operations Center
- SSH: Secure Shell Protocol
- SSL: Secure Sockets Layer
- VBS: Volume Backup Service
- VPC: Virtual Private Cloud
- VPN: Virtual Private Network
- VSS: Vulnerability Scan Service
- WAF: Web Application Firewall

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©2022 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.



Cloud Native Technologies



Foreword

- Compute technologies (containers and Kubernetes) are in demand, so migrating enterprise services to cloud is a major trend and necessary skill.
- This course introduces open source technologies (Docker and Kubernetes), Huawei Cloud container solutions, and the Huawei serverless solution.

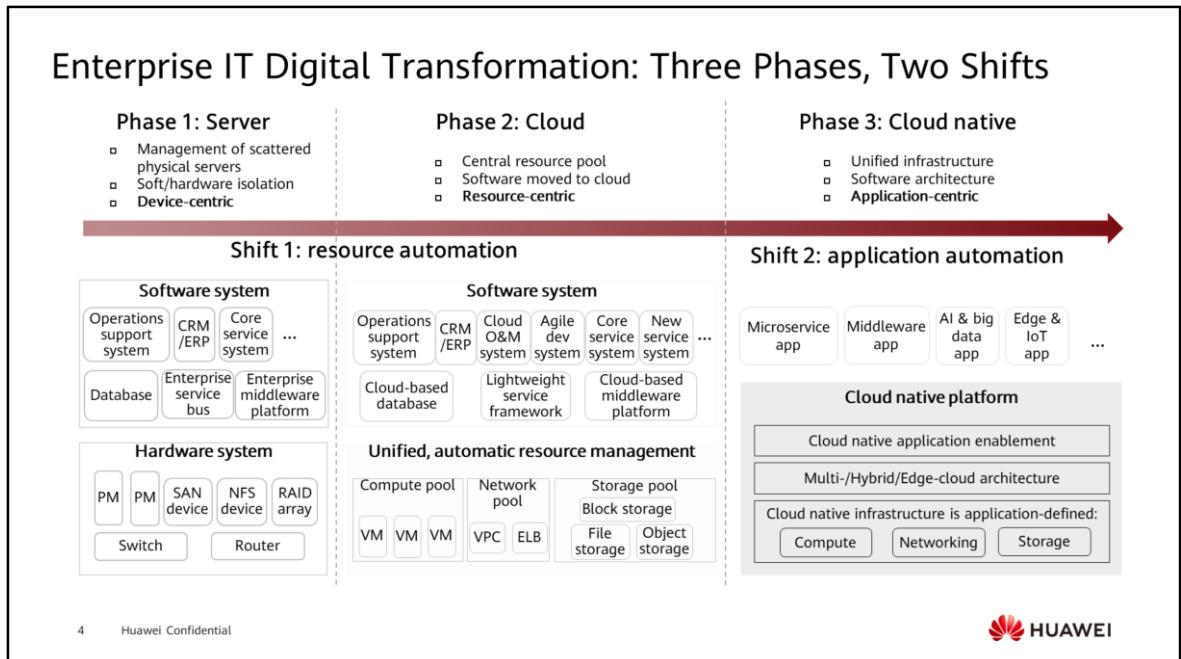
Objectives

- Upon completion of this course, you will understand:
 - Open source technologies
 - Huawei Cloud Container Engine (CCE) and Cloud Container Instance (CCI)
 - Huawei serverless solution (FunctionGraph)

Contents

- 1. Cloud Native Concepts and Background**
2. Open Source Container Technologies
3. Huawei Cloud Container Services
4. Serverless Overview

Enterprise IT Digital Transformation: Three Phases, Two Shifts



- Server-based phase: With hardware devices as the center, service applications are customized based on devices, OSs, and virtualization software. Device installation and commissioning, and application deployment and O&M are performed manually, so the automation level is low and unified device and application management capabilities are unavailable. With the emergence of virtualization software, resource utilization and container scaling flexibility are improved. However, the infrastructure is still separate from software and O&M is still complex.
- Cloud-based phase: Devices that are separately distributed in traditional mode are unified to form resource pools. A unified virtualization software platform automatically manages resources of upper-layer service software to enhance application universality. However, vendors strengthen virtualization software platforms with different commercial capabilities which cannot be shared among vendors, so applications cannot be built in a fully standardized mode, and application deployment is still resource-centric.
- Cloud native phase: Enterprise digital transformation is now shifting to cloud native. Agile application delivery, rapid scaling, smooth migration, and hitless DR are under the spotlight. Therefore, enterprises start to consider how to integrate the infrastructure with their service platforms to run service applications in a unified manner by taking advantages of standard app running, monitoring, and governance capabilities to implement application automation.

What Is Cloud Native?

- Cloud Native Computing Foundation (CNCF) definition v1.0:
- Cloud native technologies empower organizations to build and run **scalable applications** in modern, dynamic environments such **as public, private, and hybrid clouds. Containers, service meshes, microservices, immutable infrastructure, and declarative APIs** exemplify this approach.
- Loosely coupled systems (resilient, manageable, and observable) and robust automation allow engineers to make high-impact changes frequently and predictably.

CNCF was founded by Google, Huawei, and other enterprises on July 21, 2015. Huawei Cloud is the only founding member of CNCF in Asia and the only platinum member in China.

- In July 21, 2015, Cloud Native Computing Foundation (CNCF) was founded by Google, Huawei, and other enterprises, marking the shift of cloud native from a technical concept to an open source implementation. Huawei Cloud is the only CNCF founding member from Asia and the only platinum member from China.
- CNCF is committed to fostering and maintaining a vendor-neutral open source ecosystem. We democratize state-of-the-art patterns to make these innovations accessible for everyone.
- CNCF is committed to fostering and maintaining a vendor-neutral open source ecosystem and aims to make cloud native technologies available to the public. Providing a clearer, more understandable definition on cloud native, CNCF lays a foundation for the wide adoption of cloud native in a variety of industries. As surveyed by CNCF, more than 80% of users have used or plan to use the microservice architecture for service development and deployment. Users' awareness and use of cloud native technologies are at a new height, and the technology ecosystem is experiencing rapid changes.

The Development of Cloud Native



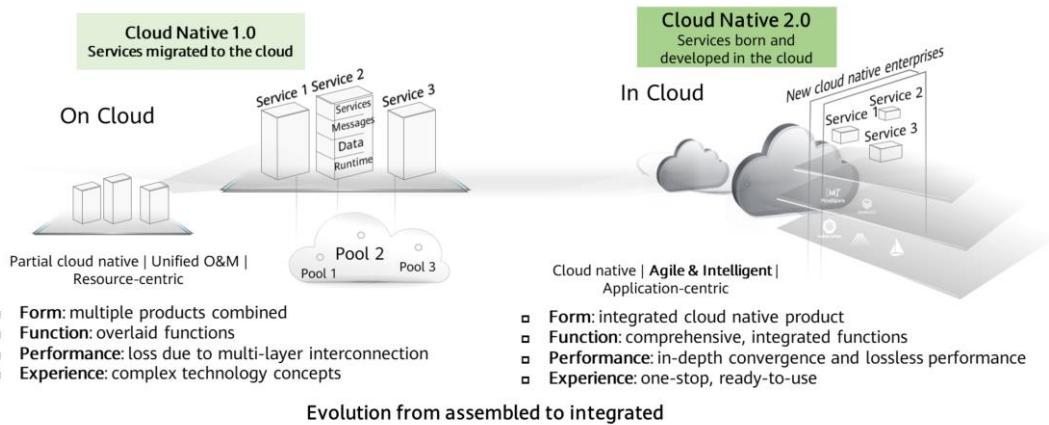
6 Huawei Confidential



- Starting from the basic container engine, the cloud native open source project continuously expands the application field and improves the adaptation capability to various scenarios. From Docker (an early open source container engine), to Kubernetes, Swarm, and Mesos (for efficient container orchestration), to Istio (for microservice governance using service meshes), KubeEdge (for edge scenarios), K3s (a lightweight Kubernetes distribution), and Volcano (for high-performance heterogeneous computing), these projects have accelerated the convergence of cloud native and industries and promoted innovation in various industries.
- In 2020, CAICT compiled the Cloud Native Development White Paper 2020 after in-depth survey and analysis on cloud native technologies and the industry in China. In 2020, Huawei Cloud first proposed the concept of Cloud Native 2.0, aiming to help every enterprise become a cloud native enterprise.

Cloud Native 2.0

- New enterprise applications are built on cloud native technologies. The cloud fully manages applications, data, and AI. Existing and new applications are organically coordinated.



- At the early stage of enterprise digital transformation, services were migrated from on premises to the cloud and deployed and run on the cloud. This is called "On Cloud". In this mode, the cloud-based resource pool simplifies service deployment, O&M, and capacity expansion in the IDC era. However, monolithic applications with their siloed architectures may lead to many application-level problems. The benefits of the cloud are still mostly limited to resource provisioning.
- In Cloud Native 1.0, technologies focus on the infrastructure layer and the monolithic architecture is resource-centric. The application ecosystem is simple. Cloud native technologies are mainly used in Internet companies.
- As digital transformation thrives, enterprises need to build and develop services in the cloud and integrate legacy capabilities with the new ones. In Cloud Native 2.0, "born in cloud" means using cloud native technologies, architectures, and services to build applications. "Grow in cloud" means these new apps run and expand fully on the cloud to build digital, intelligent services.
- From "On Cloud" to "In Cloud": New enterprise applications are built on cloud native technologies. Applications, data, and AI are managed in the cloud throughout their lifecycle. Existing applications are organically coordinated with new ones.
- New Cloud Native Enterprises: Cloud Native 2.0 is a new phase for intelligent upgrade of enterprises. Legacy capabilities co-exist and work well with new ones to achieve efficient resource utilization, agile applications, service intelligence, and secure, trustworthy services.

Advantages of Cloud Native 2.0

Efficient resources

Diverse computing power for different scenarios on an efficient, reliable distributed ubiquitous platform. Multi-cloud management and edge-cloud synergy on an efficient, application-centric resource scheduling and management platform with one-click deployment, intelligent scheduling, and comprehensive monitoring and O&M.

Agile applications

Agile application development and iteration for enterprise responsiveness, and E2E security with the latest DevSecOps.

Intelligent services

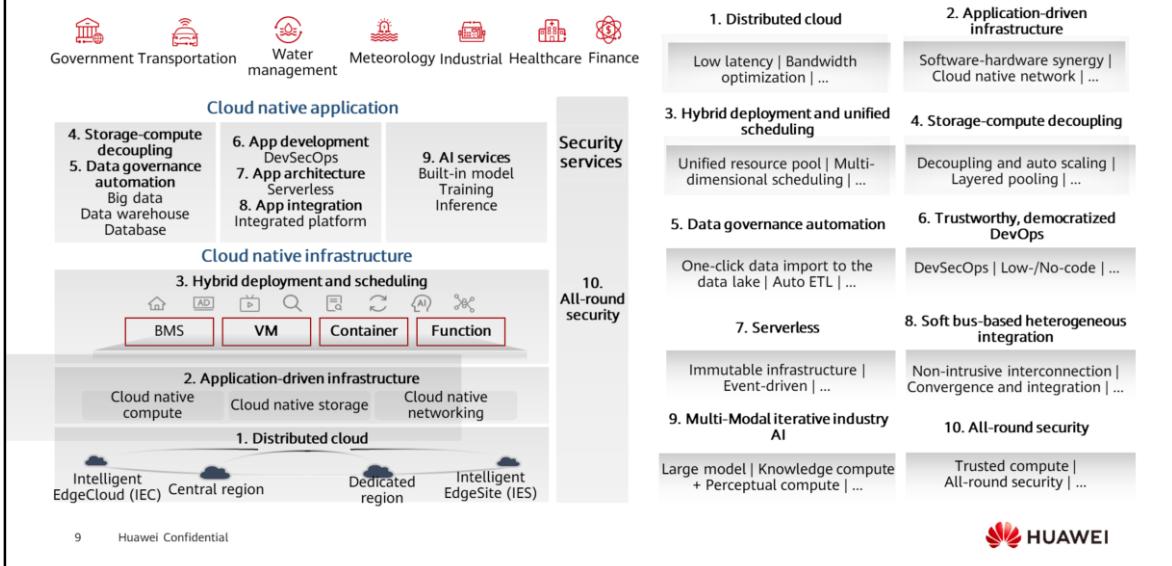
Manages data and quickly builds operations capabilities to turn data into enterprise assets for mined value and fuel service upgrade with data and AI capabilities.

Security and trustworthiness

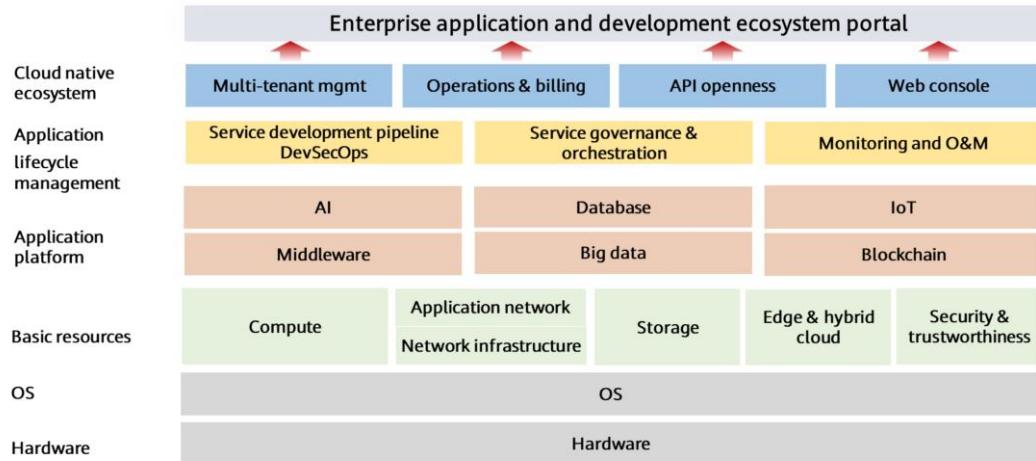
Enterprise-class security and compliance help enterprises build and run applications securely in the cloud.

- In Cloud Native 2.0,
 - cloud native technologies shift from resource-centric to application-centric. Cloud native infrastructure can be aware of application features, and applications can use cloud native infrastructure more intelligently and efficiently.
 - The multi-cloud architecture allows cloud-native applications to be distributed. Clouds can collaborate with devices, edges, and clouds themselves in multiple scenarios.
 - Cloud Native 2.0 is an open system that allows organic collaboration and co-existence between new and legacy applications.
 - Cloud Native 2.0 features full stack, where cloud native is extended to fields such as application, big data, database, and AI.

The Ten Architectures of Cloud Native 2.0



Huawei Cloud Native 2.0 at a Glance



10 Huawei Confidential



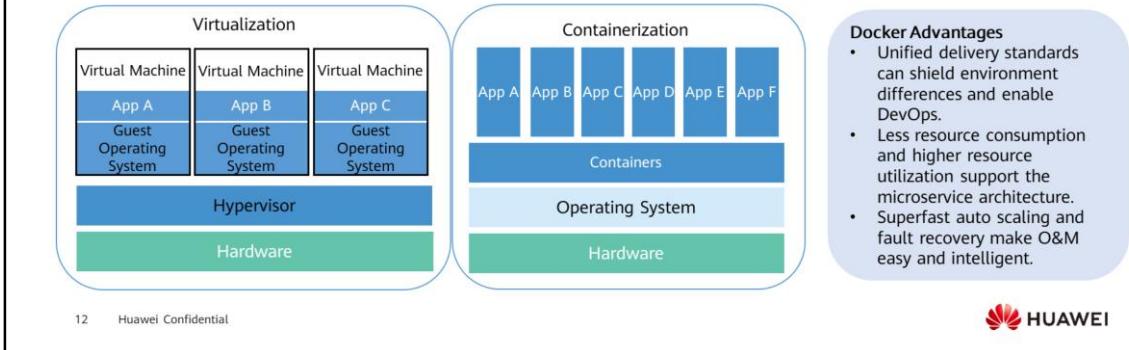
- **Hardware layer:** Introducing the cloud-infrastructure-aware hardware PCI card (SDI/Qingtian offloading card), self-developed universal CPU (Kunpeng), and heterogeneous NPU (Ascend), through a series of hardware offloading and in-depth software-hardware synergy oriented to homogeneous and heterogeneous compute, Huawei builds the most cost-effective computing power platform that works with containers and VMs.
- **OS layer:** In addition to standard OS functions, this layer distributes resources. Physical server resources are divided into multiple VMs and containers. The EulerOS supports upper-layer intelligent resource scheduling and flexible computing. Hardware passthrough minimizes the overheads of storage and network virtualization.
- **Elastic resource layer:** This layer integrates resources. For example, for cloud native compute, especially Kubernetes container clusters, their extended tasks, and Alkaid intelligent scheduling system, streamline cloud-edge and regionless scheduling, as well as cloud native capabilities such as network virtualization, distributed storage, disaster recovery, and high reliability.
- **Application and data enablement layer:** This layer covers blockchain, cloud security enablement, AI ModelArts (inclusive AI development platform), and cloud native distributed middleware, edge, database, big data, video, and IoT.
- **Application lifecycle management:** includes DevSecOps (service development pipeline), cloud native service governance and orchestration, CMDB (for tenants to deploy services), and monitoring and O&M services.
- **Multi-tenant framework:** provides cloud services with multi-tenant authentication and permission management (identity authentication for cloud service and resource access, and access permission management for cloud service objects), cloud native operations and billing, API openness, and cloud native console.

Contents

1. Cloud Native Concepts and Background
- 2. Open Source Container Technologies**
3. Huawei Cloud Container Service
4. Serverless Overview

Container Technologies

- Originated from Linux, containers are lightweight kernel virtualization technologies used to isolate processes and resources. Containers become popular since the emergence of Docker.
- Docker was launched in 2013. The idea is to standardize software delivery as shipping containers. Software in each container runs independently and does not affect each other. Docker has redefined the container industry as a unified standard for cloud computing PaaS technologies. In 2015, the Docker-led Open Container Initiative (OCI) was established, which sets up the industry-recognized container engine technology standards.



12 Huawei Confidential



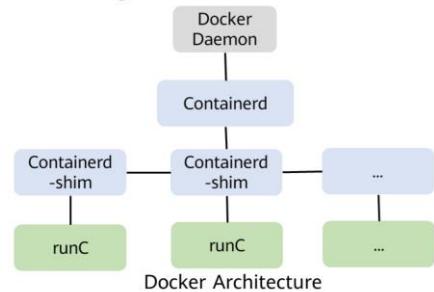
- Docker is the first system that allows containers to be portable in different machines. It simplifies the packaging of both the application and the application libraries and dependencies. Even the OS file system can be packaged into a simple portable package, which can be used on any other machine that runs Docker. Docker proposed OCI to set up container engine technology standards followed by many companies.
- Containers have the following advantages over VMs:
 - Higher system resource utilization: With no overhead for virtualizing hardware and running a complete OS, containers outperform VMs in application execution speed and memory loss.
 - Faster startup: Traditional VMs usually take minutes to start an application. However, Docker containerized applications run directly on the host kernel with no need to boot the OS, so they can start within seconds.
 - Consistent running environments: A common problem in development is the consistency of application running environments. Due to inconsistent development, testing, and production environments, some bugs cannot be discovered prior to rollout. A Docker container image provides a complete runtime to ensure consistency in application running environments.
 - Easier migration: Docker ensures the consistency in execution environment, so migrating applications becomes much easier. Docker can run on many platforms, and no matter on physical machines or virtual ones, its running results remains the same.
 - Easier maintenance and extension: Tiered storage and images in Docker facilitate the reuse of applications and simplify application maintenance and update. In addition, Docker collaborates with open source project teams to maintain a large number of high-quality official images. You can directly use them in the production environment or form new images based on them, greatly reducing the image production cost of applications.

Key Technologies

- Key technologies of Docker
 - Namespace: encapsulates the kernel resources so that each namespace has its own resources. In this way, resources for processes with different namespaces are isolated.
 - Cgroup: limits the resource usage (CPU, memory, block I/O, etc.) of a collection of process, isolating resources to prevent resource preemption and conflicts between containers.
 - Union Filesystem: a hierarchical, lightweight, and high-performance file system. It supports the overlay of file system modifications as one submission, which is the basis of container images.

Concepts of Docker containers

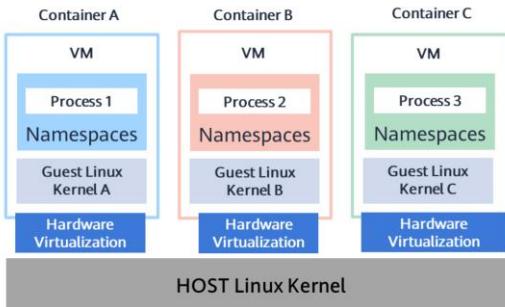
- Image: A Docker image packages an application and its dependent environments.
- Image repository: An image repository stores Docker images.
- Container: A container is a running instance created from an image and can be started, stopped, and deleted. A container image can be used to create multiple containers.



- Container technology was first invented by Linux developers. Docker has popularized containers by making the technology accessible through an open source tool and reusable images.
- Namespaces isolate the running environments, that is, each container is an independent process.
- Cgroups isolate running resources and make them exclusive for each container. You can specify the amount of resources for each container.
- The union filesystem is a filesystem service that Docker uses to layer images. Container images allow for standard container running, but container images are not containers. A container image is a series of layered read-only files managed by the storage driver. When a container image runs as a container, a writable layer, that is, a container layer, is added to the top of the image. All modifications to a running container are actually modifications to the container read/write layer. Such modifications, such as writing a new file and modifying an existing file are only applied to the container layer.
- Containers share the host kernel and do not need to boot an OS or virtualize resources. Therefore, they are more lightweight and have low resource overheads. In addition, a container image packages an application and its runtime environment. These highly portable and standardized packages allow for large-scale application scaling and management.
- Docker Daemon is a background system process in the Docker architecture.
- Containerd is an intermediate communication component between dockerd and runc. Docker manages and operates containers through containerd.
- Containerd-shim is a carrier for running containers. Each time a container is started, a new containerd-shim process is created.
- RunC is a command-line tool used to run the OCI applications.

Kata Containers

- Kata Containers builds a secure container runtime with lightweight VMs that feel and perform like containers, but provides stronger workload isolation using hardware virtualization as a second layer of defense.
- Kata Containers (kata-runtime) with QEMU/KVM makes VMs lightweight and is OCI-compliant. It supports Kubernetes Container Runtime Interface (CRI) and can replace CRI shim runtime (runC) to create pods or containers through Kubernetes.



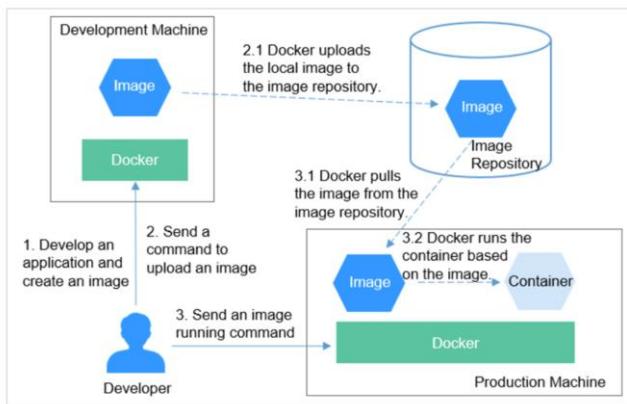
Features

- Security: Kata containers run in a dedicated kernel, providing isolation of network, I/O and memory.
- Compatibility: Kata containers support OCI and Kubernetes CRI.
- Performance: Kata containers deliver the security advantages of VMs and are as lightweight as containers.
- Simplicity: Kata Containers use standard interfaces.



- Kata Containers is an open source container project initiated by Intel, Huawei, and Red Hat. It runs container management tools on bare metal servers and provides strong, secure workload isolation. Kata containers are as lightweight and fast as containers and as secure as VMs.

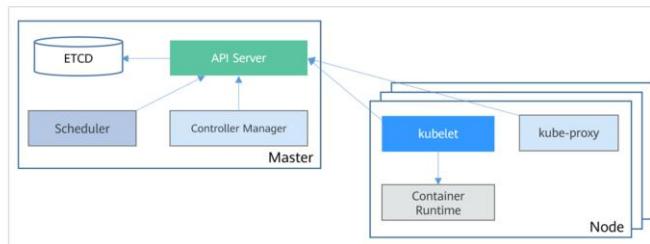
Process of Using Docker Containers



1. A developer develops an application and creates an image in the development machine. Docker runs the commands to create an image and store it on the machine.
2. The developer sends a command to push the image. After receiving the command, Docker pushes the local image to the image repository.
3. The developer sends an image running command to the machine. After the command is received, Docker pulls the image from the image repository to the machine, and then runs a container based on the image.

Kubernetes

- Kubernetes is an open container orchestration management technology built on Google's 15-year experience in large-scale cluster management and Docker technologies. Red Hat, Huawei, and other top companies are members of its open source ecosystem. Together, they set up the de facto standards of container orchestration technologies.
- In 2015, these top players established Cloud Native Computing Foundation (CNCF), which becomes the top open source organization in the cloud computing field. As the most important open source project of the CNCF community, Kubernetes has become the industry standards of the container technologies.



Master node

- A master node is the machine where the control plane components (API server, Scheduler, Controller manager, and etcd) run. In the production environment, multiple master nodes are deployed to ensure cluster high availability. For example, you can deploy three master nodes for your CCE cluster.

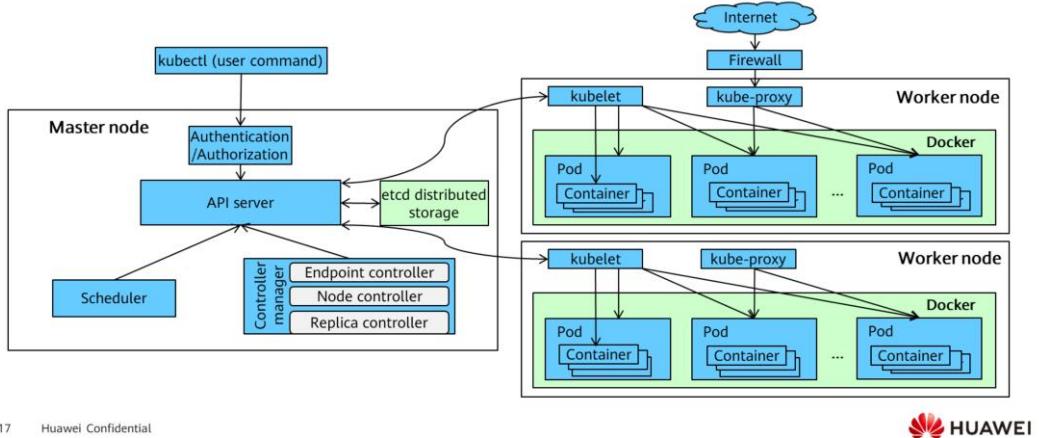
Worker node

- A worker node is a compute node in a cluster, that is, a node running containerized applications. A worker node has kubelet, kube-proxy, and Container Runtime.

- The name Kubernetes originates from Greek, meaning helmsman or pilot. K8s as an abbreviation results from counting the eight letters between the "K" and the "s". Kubernetes is open-sourced by Google from its internal cluster management system Borg after Google's own service attributes are removed. Kubernetes is a recognized de facto standard in the container orchestration field. Almost all container technologies of public cloud vendors are built on Kubernetes.
- In the standard architecture of Kubernetes, a cluster is a complete set of Kubernetes products. Most enterprises encapsulate the management plane on clusters for cluster-level management.
- For application developers, Kubernetes can be regarded as a cluster operating system. Kubernetes provides service discovery, scaling, load balancing, self-healing, and even leader election, freeing developers from infrastructure-related configurations.
- With Kubernetes, applications can be automatically deployed, restarted, migrated, and scaled based on the application status. Kubernetes can be compatible with different infrastructures (public/private cloud) using plug-ins. Kubernetes also provides flexible resource isolation for different teams to set up running environments quickly.
- A master node in the cluster manages the entire container cluster. In HA scenarios with etcd used, there are at least three master nodes in a cluster. There are many worker nodes in a cluster, which are used to run containerized applications. The master node installs kubelet on each worker node as the agent for managing the node.

Kubernetes Cluster Architecture

- A Kubernetes cluster consists of master nodes (masters) and worker nodes (nodes). Applications are deployed on worker nodes, and you can specify the nodes for deployment. A Kubernetes cluster has one or multiple master nodes and multiple worker nodes. A node is a server (VM or PM).



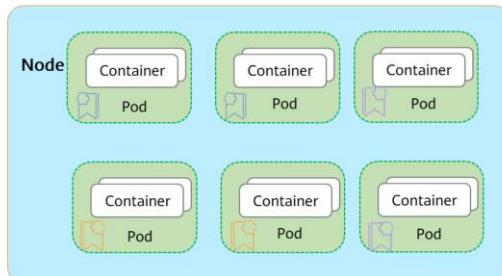
17 Huawei Confidential

HUAWEI

- Master node:
 - API server: functions as a transit station for component communication, receives external requests, and writes information to etcd.
 - Controller manager: performs cluster-level component replication, node tracing, and node fault fixing.
 - Scheduler: schedules containers to nodes by conditions (such as available resources and node affinity).
 - etcd: serves as a distributed data storage component that stores cluster configurations and object status.
- Worker node:
 - kubelet: communicates with the container runtime, interacts with the API server, and manages containers on the node. The cAdvisor monitors resources and containers on nodes in real time and collects performance data.
 - kube-proxy: serves as an access proxy between application components.
 - Container runtime: runs container software, such as Docker, containerd, CRI-O, and Kubernetes CRI.
- kubectl is a command line tool for Kubernetes clusters. You can install kubectl on any machine and run kubectl commands to operate your Kubernetes cluster.
- When using Kubernetes, users call the API server on the master node to use required resource objects such as applications and Services in the declarative APIs. The master node controller and scheduler create resources on the node based on the user definition and monitor the status at any time, ensuring that the resources meet requirements. Unified access to containerized applications on nodes can be achieved through kube-proxy.

Resource Management – Pod

- In Kubernetes, pods are the minimum deployment units that can be created, scheduled, and managed. A pod is a set of containers.
- Containers in the same pod share the same namespace, IP address, port range, and volumes.
- Pods are short-lived applications. Pods remain on the nodes to which they are scheduled until being deleted.

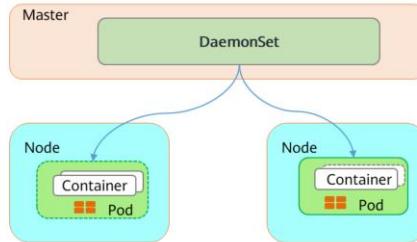


- Kubernetes uses labels to classify resources. Almost all resources in Kubernetes can be managed by labels.
- A label is a key-value pair. It can be set either during or after resource creation. You can easily modify it when needed at any time.
- Container Runtime Interface (CRI) provides computing resources when a container is running. It shields differences between container engines and interacts with each container engine through a unified interface.

- The minimum unit of Kubernetes orchestration is pod. The idea comes from pea pod. A pod can contain many containers, just as a pea pod can contain many peas.
- In most cases, containers are used to carry microservices (small and single services). During microservice design, it is recommended that one process be borne by one application. If the bearer is a container, one process is borne by one container. However, to manage microservices, you need to install service monitoring software or data reading software. That is, multiple software, or processes, need to be installed in a container. This undermines the principle of one container for one process. To comply with the microservice design principles, Kubernetes designed pods. Generally, a pod contains multiple containers, including one application container (used to provide services) and multiple sidecar containers (used to monitor the application container or manage data). For example, a pod contains three containers: web container, monitoring container, and log reading container. The web container only runs web software, and port 80 is exposed externally. The monitoring software of the web container, running in the monitoring container, monitors the web service through 127.0.0.1:80, because containers in the pod share the IP address. The log reading container reads files in the corresponding path and report the files to the log management platform, because containers in the pod share the data storage volumes.
- Container Runtime Interface (CRI) defines the interfaces of container and image services. The lifecycle of a container is separated from that of an image, two services need to be defined. CRI is responsible for the communication between kubelet and containers.

Resource Detection

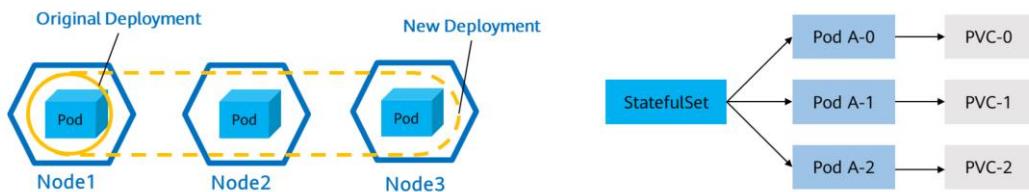
- Liveness, readiness, and startup probes check the running status of pods in Kubernetes.
- DaemonSet ensures that a pod replica runs on all or some nodes. When a node is added to a cluster, a pod is added for the node. When a node is removed from the cluster, the pod is also reclaimed. If a node becomes faulty, the DaemonSet will not create the same pod on other nodes.



- Probe types:
 - Liveness probe: checks whether the container is running. If the check fails, kubelet kills the container and restarts the container based on the restart policy.
 - Readiness probe: checks whether the container is ready to handle requests. If the check fails, the endpoint controller deletes the IP address of the pod from the endpoint of all services that match the pod.
 - Startup probe: checks whether the containerized application is started. Other probes are disabled until the startup probe is successfully detected. If the check fails, kubelet kills the container and restarts the container based on the restart policy.
- Application scenarios of DaemonSets
 - DaemonSet for clusters on the node.
 - DaemonSet for log collection on the node.
 - DaemonSet for node monitoring.
- Kubernetes supports node-level and pod-level affinity and anti-affinity. You can configure custom rules to achieve affinity and anti-affinity scheduling. For example, you can deploy frontend pods and backend pods together, deploy the same type of applications on a specific node, or deploy different applications on different nodes.

Resource Scheduling

- Controllers create and manage pods for Kubernetes and provide replica management, rolling upgrade, and self-healing. The most commonly used controller is Deployment. Pods under a Deployment have the same characteristics except for the name and IP address. Deployments can create pods using pod templates and delete pods.
- However, when each pod requires its own status or storage, a persistent identifier is needed. In this case, StatefulSets can maintain sticky identity for each pod. In addition, StatefulSets use PVCs for persistent storage to ensure that the same data can be accessed after pods are deleted.



20 Huawei Confidential



- Deployment: Controllers create and manage pods for Kubernetes and provide replica management, rolling upgrade, and self-healing.
- Deployment: the most commonly used controller. When a Deployment is created, a ReplicaSet is automatically created. A Deployment can manage multiple ReplicaSets and use them to manage pods.
- All pods under a Deployment have the same characteristics except for the name and IP address. If required, a Deployment can use the pod template to create a new pod. If not required, the Deployment can delete any one of the pods. Generally, a pod contains one container or several containers that are closely related. A ReplicaSet contains multiple identical pods. A Deployment contains one or more different ReplicaSets.
- StatefulSets provide a fixed identifier for each pod. A fixed suffix ranging from 0 to N is added to the pod name. After pods are rescheduled, the pod name and host name remain unchanged. StatefulSets provide a fixed access domain name for each pod through the headless Service (described in following sections). StatefulSets create PersistentVolumeClaims (PVCs) with fixed identifiers to ensure that pods can access the same persistent data after being rescheduled.
- Jobs and cron jobs allow you to run short lived, one-off tasks in batch. They ensure the task pods run to completion.
 - Job: a resource object used by Kubernetes to control batch tasks. Jobs are different from long-term servo workloads (such as Deployments and StatefulSets). The former is started and terminated at specific times, while the latter runs unceasingly unless being terminated. The pods managed by a job automatically exit after successfully completing the job based on user configurations.
 - CronJob: runs a job periodically on a specified schedule. A cron job object is similar to a line of a crontab file in Linux.

Resource Configurations

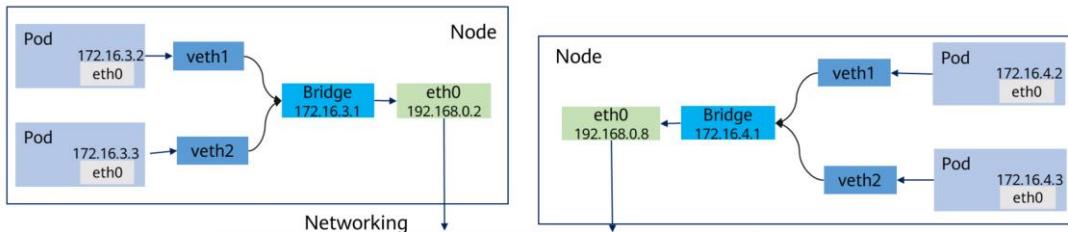
- A ConfigMap stores the configurations (configuration data or files) required by applications in key-value pairs. ConfigMap decouples image and application configuration files, command line parameters, and environment variables so you can use different configurations in different environments.
- A secret lets you store and manage sensitive information (password, authentication information, certificates, and private keys). Storing confidential information in a secret is safer and more flexible than putting it verbatim in a pod definition or in a container image.
- Similar to a ConfigMap, a secret stores data in key-value pairs. The difference is that a secret is encrypted. ConfigMap stores configuration files, while Secret stores sensitive data (passwords, tokens, and keys).



- Similar to a ConfigMap, a secret stores data in key-value pairs. The difference is that the value must be encoded using Base64.
- secret
 - A secret provides better security in the process of creating, viewing, and editing pods.
 - The system takes extra precautions for secret objects, for example, preventing it from being written to a location on the disk.
 - Only the secret requested by the pod is visible in its container. One pod cannot access the secret of another pod.

Kubernetes Networking

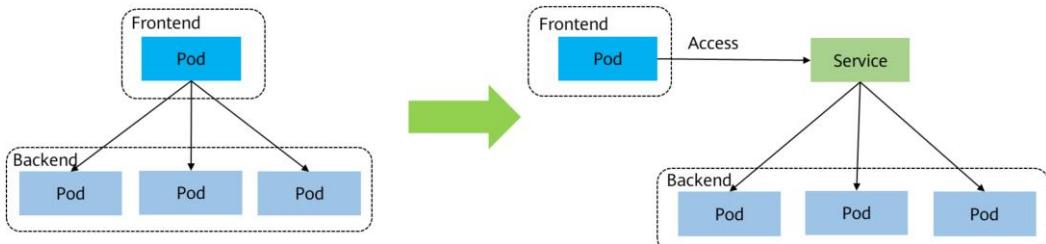
- Kubernetes provides Container Network Interface (CNI) for networking among pods, clusters, and nodes. There are many open source CNI plug-ins, such as Flannel and Calico. Huawei Cloud CCE also provides customized CNI plug-ins for you to use Huawei Cloud VPC networks when running Kubernetes.
- A pod is connected to external systems through a virtual Ethernet interface pair (veth pair). Pods on the same node communicate with each other using a Linux bridge. When bridges on different nodes are connected, the cluster requires the pod address to be unique. Therefore, cross-node bridges use different CIDR blocks to prevent duplicate pod IP addresses.



- Bridges between different nodes can be implemented in multiple modes. However, in a cluster, the pod IP address must be unique. Therefore, cross-node bridges use different CIDR blocks to prevent duplicate pod IP addresses.
- Communication between containers: Containers in a pod share the same network namespace, which is provided by IaaS. Each pod has its own IP address and has no conflicts with each other.
- Pods and nodes in the cluster can directly communicate with each other using the IP address. This communication does not require any network address translation, tunneling, or proxy. The same IP address is used internally and externally in a pod, which also means that the standard naming and discovery mechanisms, such as DNS, can be directly used. This type of communications also requires Kubernetes network plug-ins (for example, flannel) to configure a layer network fabric, routed network, and more.
- Communication between pods: Pods can communicate with each other through IP addresses only when pods know the IP addresses of each other. In a cluster, pods may be frequently deleted and created. That is, the IP addresses of pods are not fixed. To solve this problem, a Service provides an abstraction layer for accessing pods. No matter how the backend pod changes, the Service functions as a stable frontend to enable external access. In addition, a Service supports HA and load balancing, forwarding requests to the correct pod.
- Flannel is a network planning service designed by the CoreOS team for Kubernetes. It enables containers created on different nodes in a cluster to have a unique virtual IP address in the cluster.
- Calico is famous for its performance and flexibility compared to Flannel's simplicity. Calico provides more comprehensive functions, not only network connections between hosts and pods, but also network security and management.

Kubernetes Networking – Service

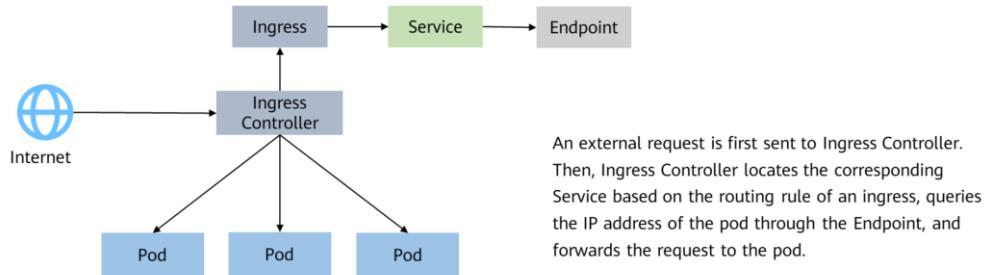
- Services in Kubernetes enable pod access. You can add a Service (with a fixed IP address) for the frontend pods to access the backend pods. The Service forwards the traffic to the backend pods. In this way, the frontend pod does not need to be aware of the changes on the backend pods. The Service can also perform load balancing for these pods.



- After a pod is created, the following problems may occur when you directly access a pod:
 - The pod can be deleted and recreated at any time by a controller such as a Deployment, and the result of accessing the pod becomes unpredictable.
 - The IP address of the pod is allocated only after the pod is started. Before the pod is started, the IP address of the pod is unknown.
 - An application is usually composed of multiple pods that run the same image. Accessing pods one by one is not efficient.
- ReplicationControllers, ReplicaSets, and Deployments only ensure the number of microservice pods that support services, but do not solve the problem of how to access these services. A pod is only an instance that runs services. It may be stopped on a node at any time and recreated on another node using a new IP address. Therefore, services cannot be provided using a fixed IP address and port number. To provide services stably, service discovery and load balancing are required. Service discovery finds the target backend service requested by the client. In a Kubernetes cluster, the service that the client needs to access is the Service object. Each Service corresponds to a valid virtual IP address in the cluster. The cluster uses the virtual IP address to access a Service.
- In Kubernetes, a Service is an abstraction which defines a logical set of pods and a policy by which to access them, usually this pattern is called a microservice. The set of pods targeted by a Service is usually determined by a selector.
- The implementation types of Service are as follows:
 - ClusterIP: provides an internal virtual IP address for pods to access (default mode).
 - NodePort: enables a port on the node for external access.
 - LoadBalancer: allows access through an external load balancer.

Kubernetes Networking – Ingress

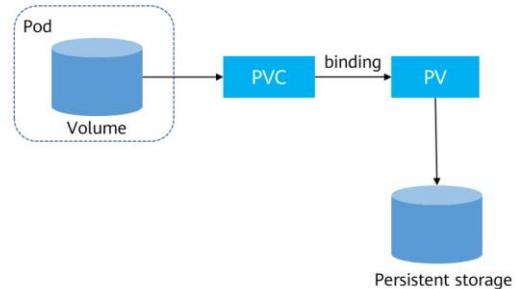
- Ingress is a set of rules that allow access from outside the cluster to Services within the cluster. A typical access mode is HTTP.
- Services forward requests based on Layer 4 TCP and UDP protocols. Ingresses can forward requests based on Layer 7 HTTPS and HTTP protocols and make forwarding more targeted by domain names and paths.



- Ingress provides load balancing, SSL termination, and name-based virtual hosting. Ingress exposes HTTP and HTTPS routes from outside the cluster to Services within the cluster. Traffic routing is controlled by rules defined on the ingress resource.
- To use an Ingress, you must install Ingress Controller on your Kubernetes cluster. Ingress Controller can be implemented in multiple modes. The most common one is NGINX Ingress Controller maintained by Kubernetes. In Huawei Cloud, Cloud Container Engine (CCE) works with Elastic Load Balance (ELB) to implement layer-7 load balancing (via ingresses).

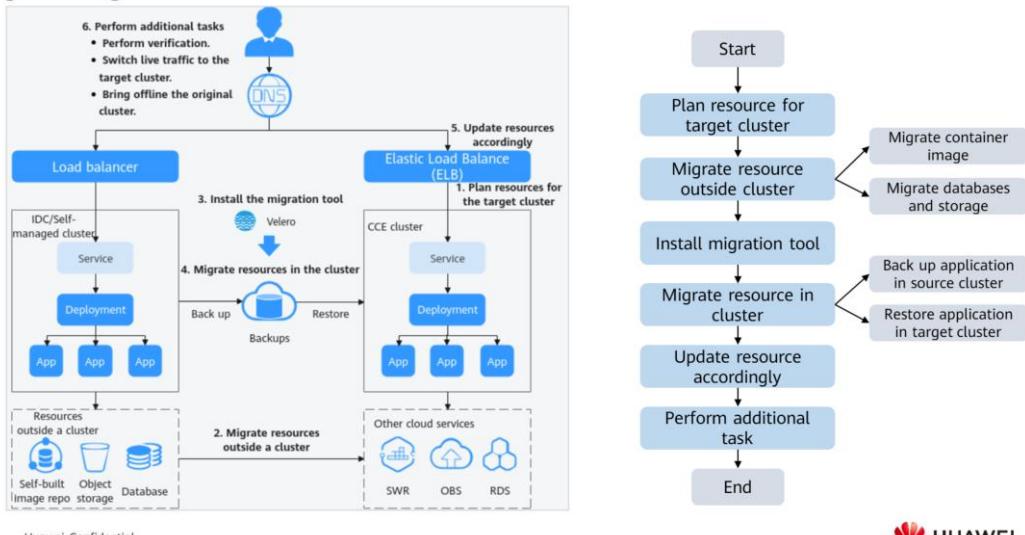
Persistent Storage

- On-disk files in a container are ephemeral. When a container crashes and is then restarted, the files in the container will be lost. When multiple containers run in a pod, files often need to be shared between these containers. Kubernetes volume abstraction solves these two problems. Volumes, as part of a pod, cannot be created independently and can only be defined in pods.
- If you want to read the previously written data after a pod is rebuilt and scheduled again, you need to use the network storage. There are multiple network storage types (for example, block storage, file storage, and object storage). Kubernetes provides PersistentVolumes (PVs) and PersistentVolumeClaims (PVCs) to abstract this problem. You can request specific size of storage when needed, just like pods can request specific levels of resources (CPU and memory).
- Kubernetes Container Storage Interface connects your containers to various types of storage resources.



- A volume will no longer exist if the pod to which it is mounted does not exist. However, files in the volume may outlive the volume, depending on the volume type. All containers in a pod can access its volumes, but the volumes must have been mounted. Volumes can be mounted to any directory in a container.
 - PV: defines a directory for persistent storage on a host machine, for example, a mount directory of a file system.
 - PVC: describes the attributes of the PV that a pod wants to use, such as the volume capacity and read/write permissions.
- Although PVs and PVCs allow you to consume abstract storage resources, you may need to configure multiple files to create PVs and PVCs. Therefore, they are generally managed by the cluster administrator. To resolve this issue, Kubernetes supports dynamic PV provisioning to create PVs automatically. The cluster administrator can deploy a PV provisioner and define the corresponding StorageClass. In this way, developers can select the storage class to be created when creating a PVC. The PVC transfers the StorageClass to the PV provisioner, and the provisioner automatically creates a PV.
- StorageClass describes the storage class used in the cluster. You need to specify StorageClass when creating a PVC or PV.
- To allow a pod to use PVs, a Kubernetes cluster administrator needs to set the network storage class and provides the corresponding PV descriptors to Kubernetes. You only need to create a PVC and bind the PVC with the volumes in the pod so that you can store data.

Migrating a Kubernetes Cluster to the Cloud



26 Huawei Confidential



- The cluster migration process is as follows:
 - Plan resources for the target cluster. For details about the differences between CCE clusters and on-premise clusters, see "Key Performance Parameter" in "Planning Resources for the Target Cluster". Plan resources as required and ensure the performance configuration of the target cluster is the same as that of the source cluster.
 - Migrate resources outside a cluster. Huawei Cloud provides migration solutions to migrate resources outside the cluster. These solutions involve the migration of container images, databases, and storage.
 - Install the migration tool. After resources outside the cluster are migrated, you can use the migration tool to back up and restore application configurations in the source and target clusters.
 - Migrate resources in the cluster. You can use open source DR software, such as Velero, to back up resources in the source cluster to the object storage and restore the resources in the target cluster.
 - No need to configure, update, or manage servers. Managing servers, VMs, and containers involves personnel, tools, training, and time.
 - FaaS and BaaS products can be scaled flexibly and precisely to process each request. For developers, a serverless platform does not need capacity planning or auto scaling triggers or rules.
 - Update resources accordingly. After the migration, cluster resources may fail to be deployed. You need to update the faulty resources. The possible adaptation problems lie in images, Services and ingresses, StorageClasses, and databases.
 - Perform additional tasks. After cluster resources are properly deployed,

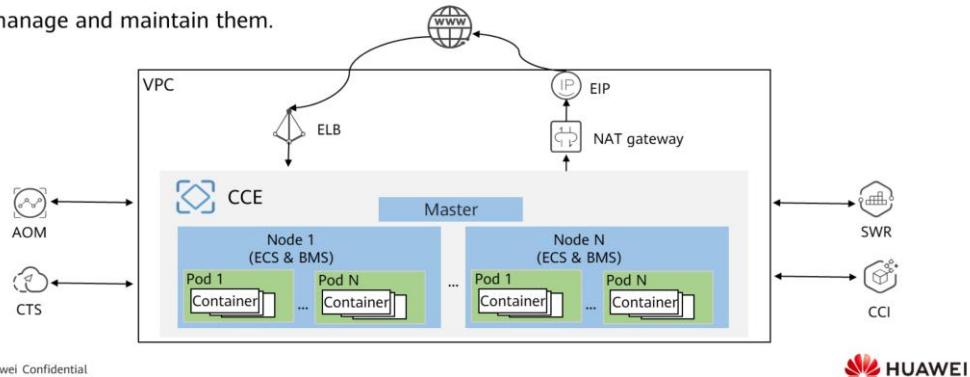
verify application functions after the migration and switch service traffic to the target cluster. After confirming that all services are running properly, bring the source cluster offline.

Contents

1. Cloud Native Concepts and Background
2. Open Source Container Technologies
- 3. Huawei Cloud Container Service**
4. Serverless Overview

Cloud Container Engine (CCE)

- CCE is a highly scalable, high-performance, enterprise-class Kubernetes service for you to run containers. With CCE, you can easily deploy, manage, and scale containerized applications in the cloud.
- CCE is a hosted Kubernetes service that simplifies the deployment and management of containerized applications. With CCE, you can easily create Kubernetes clusters, deploy containerized applications, and manage and maintain them.

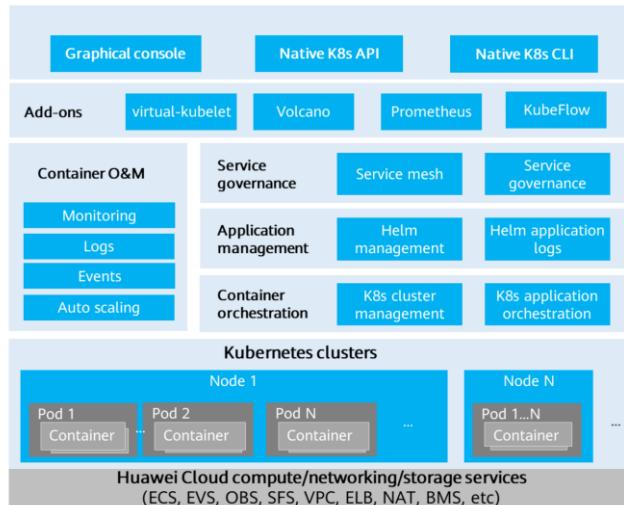


28 Huawei Confidential

HUAWEI

- CCE is deeply integrated with high-performance HUAWEI CLOUD computing (ECS/BMS), network (VPC/EIP/ELB), and storage (EVS/OBS/SFS) services, and supports heterogeneous computing architectures such as GPU and Arm. You can build high-availability Kubernetes clusters secured by multi-AZ, cross-region disaster recovery (DR) and auto scaling.

CCE Cluster Architecture and Features



Heterogeneous compute

- Supports various Huawei Cloud compute instances and manages existing instances.
- Supports secure Kunpeng instances.
- Supports GPUs and Huawei Ascend to run AI and big data services.
- Supports hybrid deployment of VMs and BMs.

High-performance cloud native network

- VPC-Router: 30% higher performance than open source Flannel
- Supports network policies and container-scoped network isolation.
- Cloud Native Network 2.0, passthrough networking, zero resource loss

Comprehensive cloud native security

- Shields open source vulnerabilities with a security-hardened container runtime.
- Scans container images using multiple policies to identify risks.
- Monitors container runtimes in real time to detect exceptions.

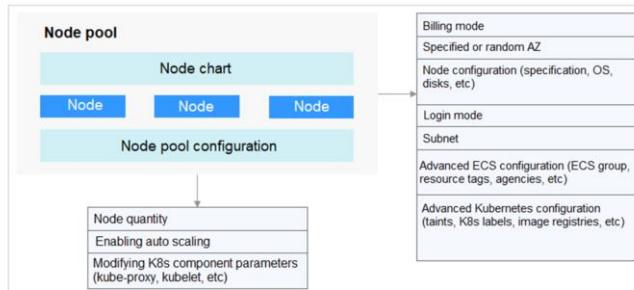
Unified, large-scale cloud native scheduling

- A single cluster supports a maximum of 10,000 nodes, meeting enterprises' requirements on large-scale service production.
- Improves scheduling efficiency by 30% with Volcano.

- Huawei is amongst the first developers of the Kubernetes community in China. Huawei is a major contributor to the open source community and a leader in the container ecosystem. Huawei Cloud CCE is the earliest commercial Kubernetes service in China, and is also one of the first products that passed the CNCF Certified Kubernetes Conformance Program. CCE features benefits such as access to open ecosystems, enhanced commercial features, and adaptation to heterogeneous infrastructure.
- Volcano: Native Kubernetes has weak support for batch computing services. Volcano provides two enhanced batch computing capabilities. One is advanced job management, such as task queuing, priority setting, eviction, backfilling, and starvation prevention. The other is intelligent scheduling, such as topology-aware affinity-based scheduling and dynamic driver-executor ratio adjustment. In addition, scheduling and distributed frameworks such as gang scheduling and PS-Worker are supported.
- You can use CCE via the CCE console, kubectl, or Kubernetes APIs.

CCE Nodes

- A Kubernetes cluster consists of master nodes and worker nodes. The nodes described in this section refer to worker nodes, the computing nodes that run containerized applications. They are the basic elements of a container cluster.
- A node can be a virtual machine (VM) or a physical machine (PM). The components on a node include kubelet, container runtime, and kube-proxy.
- A node pool contains one node or a group of nodes with identical configuration in a cluster.

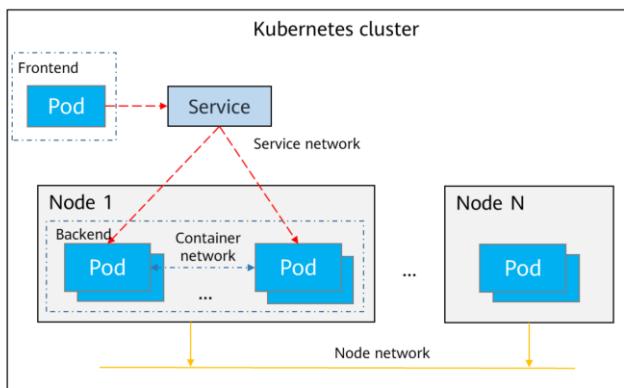


- You can create a node pool for a CCE cluster to quickly create, manage, and destroy nodes without affecting the entire cluster.
- All nodes in a custom node pool have identical parameters and node type. You cannot configure a single node in a node pool; any configuration changes affect all nodes in the node pool.

- A node is a basic element of a container cluster. CCE uses high-performance Elastic Cloud Servers (ECSs) or Bare Metal Servers (BMSs) as nodes to build highly available Kubernetes clusters.
- Kata containers are distinguished from common containers in a few aspects. The most important difference is that each Kata container (pod) runs on an independent micro-VM, has an independent OS kernel, and is securely isolated at the virtualization layer. CCE provides container isolation that is more secure than independent private Kubernetes clusters. With Kata containers, kernels, computing resources, and networks are isolated between different containers to protect pod resources and data from being preempted and stolen by other pods.
- 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.
- CCE supports Kubernetes-native deployment and lifecycle management of container workloads, including creation, configuration, monitoring, auto scaling, upgrade, uninstall, service discovery, and load balancing.

Cluster Networks

- A cluster involves node network, container network, and Service network.

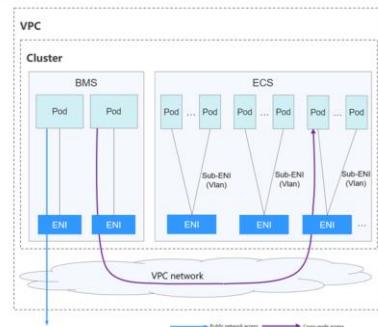


- Node network: assigns IP addresses to hosts (nodes) in the cluster. You need to select a subnet in the VPC for the node network. The number of available IP addresses in the subnet determines the number of nodes that can be created in a cluster.
- Container network: assigns IP addresses to containers in the cluster. CCE inherits the container network model of Kubernetes. Each pod has an independent IP address on each network plane.
- Service network: Each Service has a fixed IP address. When creating a cluster on CCE, you can specify the Service CIDR block, which cannot overlap with the node or container CIDR block. The Service CIDR block is cluster-scoped.

- Recommendations on CIDR block planning:
 - CIDR blocks cannot overlap. Otherwise, a conflict occurs. All subnets (including those created from the secondary CIDR block) in the VPC where the cluster resides cannot conflict with the container and Service CIDR blocks.
 - Ensure that each CIDR block has sufficient IP addresses. The IP addresses in the node CIDR block must match the cluster scale. Otherwise, nodes cannot be created due to insufficient IP addresses. The IP addresses in the container CIDR block must match the service scale. Otherwise, pods cannot be created due to insufficient IP addresses.
- In the Cloud Native Network 2.0 model, the container CIDR block and node CIDR block share the IP addresses in the same VPC. Therefore, you are advised not to set the container subnet and node subnet to the same. Otherwise, containers or nodes may fail to be created due to insufficient IP resources.
- CCE supports the following container network models: container tunnel network, VPC network, and Cloud Native Network 2.0.
- The container tunnel network is constructed on but independent of the node network through tunnel encapsulation. This network model uses VXLAN to encapsulate Ethernet packets into UDP packets and transmits them in tunnels. Open vSwitch serves as the backend virtual switch. Though at some costs of performance, packet encapsulation and tunnel transmission enable higher interoperability and compatibility for most scenarios that do not require high performance.

Cloud Native Network 2.0

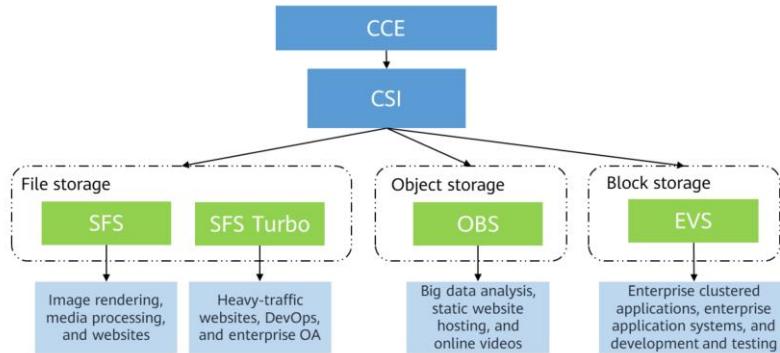
- Developed by CCE, Cloud Native Network 2.0 deeply integrates Elastic Network Interfaces (ENIs) and sub-ENIs of Virtual Private Cloud (VPC). Container IP addresses are allocated from the VPC CIDR block. ELB passthrough networking is supported to direct access requests to containers. Security groups and elastic IPs (EIPs) can be bound to deliver high performance. This model is ideal for scenarios that require high performance and large-scale networking.
 - Pods on BMS nodes use ENIs, whereas pods on ECS nodes use Sub-ENIs. Sub-ENIs are attached to ENIs through VLAN sub-interfaces.
 - In-node/Cross-node access: Packets are forwarded through the ENI or sub-ENI.



- Advantages: The container network directly uses the VPC, making it easy to locate network problems and improve the networking performance. Requests from external networks in a VPC can be directly routed to a container IP address. Load balancing, security groups, and EIPs provided by the VPC can be directly used.
- Disadvantages: The container network consumes the IP addresses in the VPC. You need to plan the container CIDR block before creating a cluster.
- This network model is available only to CCE Turbo clusters.

Container Storage

- Container storage provides storage for container workloads. It supports multiple storage classes. A pod can use any amount of storage.
- Container services allow you to use local disks, EVS/SFS/SFS Turbo/OBS volumes, snapshots, and backups.



33 Huawei Confidential



- In CCE, container storage is backed both by Kubernetes-native objects, such as emptyDir, hostPath, secret, and ConfigMap, and by cloud storage services. These cloud storage services can be accessed via Container Storage Interface (CSI).
- CSI enables Kubernetes to support various classes of storage. For example, CCE can easily interconnect with Huawei Cloud block storage (EVS), file storage (SFS), and object storage (OBS).
- CCE provides an add-on named everest to serve as CSI. Everest is a cloud native container storage system. Based on CSI, clusters can interconnect with Huawei Cloud storage services such as EVS, OBS, SFS, and SFS Turbo. everest is a system resource add-on. It is installed by default when a cluster of Kubernetes v1.15 or later is created.

Comparison Between CCE and CCE Turbo

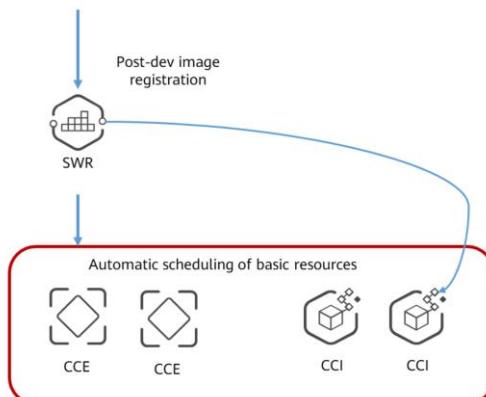
Dimension	Sub-dimension	CCE Turbo Cluster	CCE Cluster
Cluster	Positioning	Next-generation container cluster for Cloud Native 2.0 with accelerated computing, networking, and scheduling	Standard cluster for common commercial use
	Node specifications	Hybrid deployment of VMs and bare metal servers	Hybrid deployment of VMs and bare metal servers
	Supported models	Models based on the QingTian architecture with software-hardware synergy	General-purpose models
Networking	Model	Large-scale and high-performance scenarios	Scenarios that do not require high performance or involve large-scale deployment. Models: tunnel network and VPC network
	Performance	The VPC network and container network are flattened into one, achieving zero performance loss.	The VPC network is overlaid with the container network, causing certain performance loss.
	Container network isolation	Pods can be directly associated with security groups to configure isolation policies for resources inside and outside a cluster.	Tunnel network model: Network isolation policies are supported for intra-cluster communication (by configuring network policies). VPC network model: Isolation is not supported.
Security	Isolation	Bare metal server: You can select Kata containers for VM-level isolation. VM: Common containers are deployed.	Common containers are deployed and isolated by Cgroups.
Applications		Deploy, schedule, and autoscale containerized applications in Kubernetes clusters, manage microservice traffic, etc.	Run containerized applications that have higher requirements on computing, networking, scheduling, and security.

Comparison Between Self-managed Kubernetes Clusters and CCE

Item	Self-managed Cluster	CCE
Usability	Cluster management is complex. You have to handle all the complexity in installing, operating, scaling, configuring, and monitoring Kubernetes cluster infrastructure. Each cluster upgrade requires tremendous manual adjustment, imposing a heavy burden on O&M personnel.	You can create and upgrade Kubernetes clusters in just a few clicks. CCE supports automated deployment and one-stop O&M of containerized applications. CCE is deeply integrated with service mesh and Helm charts to offer out-of-the-box usability.
Scalability	You have to manually evaluate service load and cluster health before deciding to resize a cluster.	CCE can automatically resize clusters and workloads as resource usage changes. Combined use of auto scaling policies can flexibly scale clusters and workloads to meet fluctuating demands.
Reliability	Only one master node is available in a cluster. Once the master node is down, the entire cluster as well as all the applications in the cluster will become out of service.	If the high availability option is enabled when you create a cluster, three master nodes will be created in the cluster, avoiding single points of failure on the cluster control plane.
Efficiency	You have to either build image repositories or revert to third-party image repositories. Images are pulled from repositories in serial.	CCE works with SoftWare Repository for Container (SWR) to support DevOps pipelines and eliminate the need to manually write Dockerfiles or Kubernetes manifests. With ContainerOps pipeline templates, you can define how to build container images, push them to repositories, and deploy them. Images are pulled from repositories in parallel.
Cost	Heavy upfront investment is required in installing, managing, and scaling the cluster infrastructure.	You only need to pay for the cluster master nodes and the infrastructure resources (such as ECSSs, EVS disks, elastic IPs/bandwidth, and load balancers) used to run applications and store data.

SoftWare Repository for Container (SWR)

- Easily manage the full lifecycle of your container images, and secure deployment of images for your applications.



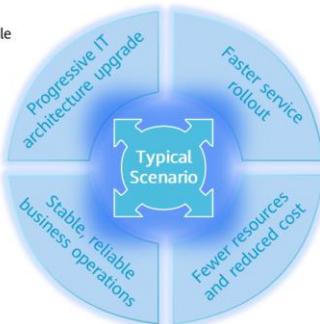
Features:

- Container image hosting
- Complete lifecycle management of images with multiple tags
- Free storage and traffic
- Compatible with Docker Hub and Docker commands
- Integration with CCE
- Event triggers

- Ease of use:
 - You can directly push and pull container images without platform build or O&M.
 - SWR provides an easy-to-use management console for full lifecycle management over container images.
- Security and reliability:
 - SWR supports HTTPS to ensure secure image transmission, and provides multiple security isolation mechanisms between and inside accounts.
 - SWR leverages professional storage services of Huawei to ensure reliable image storage.
- Faster image pull and build:
 - P2P acceleration technology developed by Huawei brings faster image pull for CCE clusters during high concurrency.
 - Intelligent node scheduling around the globe ensures that your image build tasks can be automatically assigned to the idle nodes nearest to the image repository.

CCE Application Scenarios

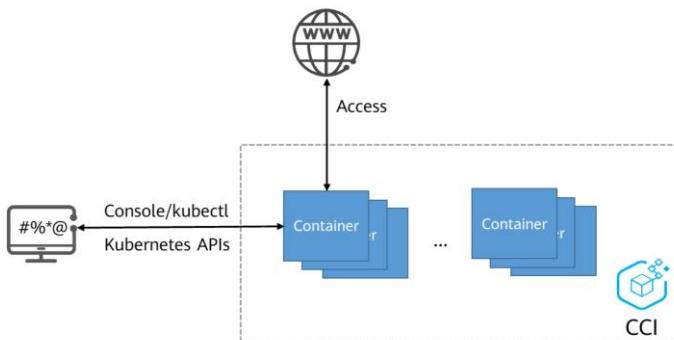
- A complex application is **decoupled** into multiple lightweight modules. Each module can be **flexibly** upgraded and scaled to respond to market changes.
- Fast **auto scaling** of containers ensures service performance even in the case of traffic bursts.
- The same container image can be used through each phase from R&D to O&M to ensure **consistency** of service running environments. Services can be used **out of the box** and rolled out faster.
- With containers, host resources can be divided at a **finer granularity** to improve resource utilization.



- From the practices of customers and partners, there are four typical scenarios of using CCE:
 - First, progressive IT architecture upgrade. With CCE, complex applications in traditional architectures are decoupled into multiple lightweight modules. Each module is run as a Kubernetes workload. For example, stateless applications run as Deployments and stateful applications run as StatefulSets. In this way, modules can be flexibly upgraded and scaled to meet changing market demands.
 - Second, faster service rollout. The same container image can be used through each phase from R&D to O&M to ensure the consistency of service running environments. Services can be used out of the box and rolled out faster.
 - Third, auto scaling upon service traffic fluctuation. Containers can be quickly scaled within seconds to ensure service performance.
 - Fourth, fewer resources and reduced cost. With containers, host resources can be divided at a finer granularity to improve resource utilization.

Cloud Container Instance (CCI)

- This serverless container engine allows you to run containers without creating or managing server clusters.
- Now you can directly create and use container workloads through the console, kubectl, and Kubernetes APIs. Your servers are invisible from the user perspective.

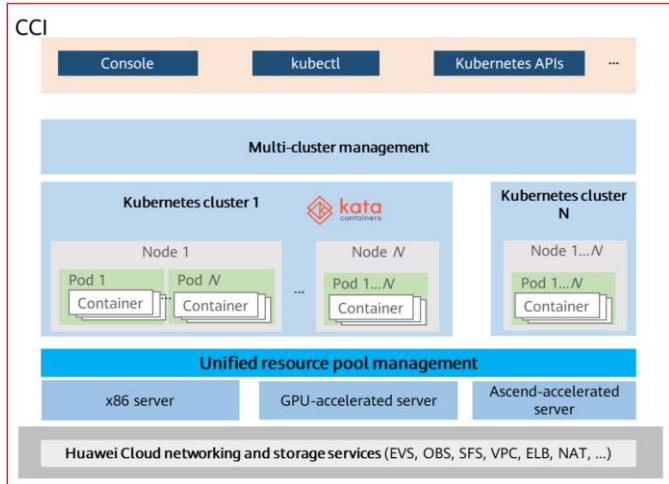


38 Huawei Confidential



- In the serverless model, a cloud provider runs servers and dynamically allocates resources so that you can build and run applications without having to create, manage, or maintain servers. This model helps you improve development efficiency and reduce IT costs.
- CCE provides semi-hosted clusters, while CCI provides fully-hosted clusters that do not need manual management.
- Functions:
 - CCI provides one-stop container lifecycle management, allowing you to run containers without creating or managing server clusters.
 - CCI supports multiple types of compute resources, including CPUs, GPUs, and Ascend chips, to run containers.
 - Various network access modes and layer-4 and layer-7 load balancing are available to meet scenario-specific needs.
 - CCI can store data on various Huawei Cloud storage volumes, including EVS, SFS, and OBS.
 - CCI supports fast auto scaling. Users can customize scaling policies and combine multiple scaling policies to cope with traffic surge during peak hours.
 - The comprehensive container status monitoring of CCI monitors the resources consumed by containers, including the CPU, memory, GPU, and GPU memory usage.
 - CCI provides dedicated container instances, which run Kata containers on high-performance physical servers, enabling VM-level security isolation without performance deterioration.

CCI Architecture

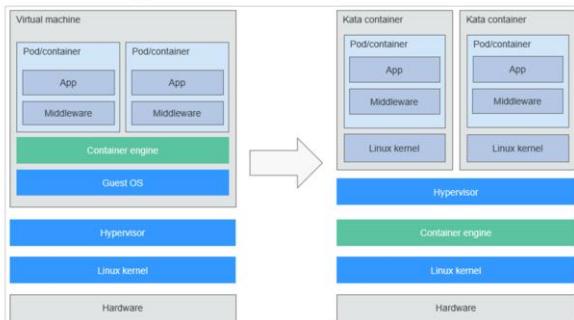


- CCI is integrated with Huawei Cloud services, including network (VPC/ELB/NAT) and storage (EVS/OBS/SFS) services.
- High-performance heterogeneous infrastructure allows containers to run directly on physical servers.
- CCI uses Kata Containers and its own virtualization acceleration technology to isolate at the VM level for higher performance and security.
- Centrally manage multiple clusters and schedule container loads without needing to perform cluster operations.
- The Kubernetes-based workload model enables fast workload deployment, elastic load balancing, auto scaling, and blue-green deployment.

- With CCI, you can stay focused on your own services, instead of underlying hardware and resources. CCI is billed by the second for convenient use anytime.
- Dedicated container instances allow you to exclusively use physical servers and support service isolation among departments. They run Kata Containers on high-performance physical servers, enabling VM-level security isolation without performance loss. Huawei Cloud performs O&M, allowing you to completely focus on your services.

High Security

- CCI provides dedicated container instances, which run Kata Containers on high-performance physical servers and use Huawei-developed hardware virtualization acceleration technology. It enables VM-level security isolation without performance loss.



The containers are isolated by lightweight VMs. Each container has an independent OS kernel.

Advantages of Kata Containers

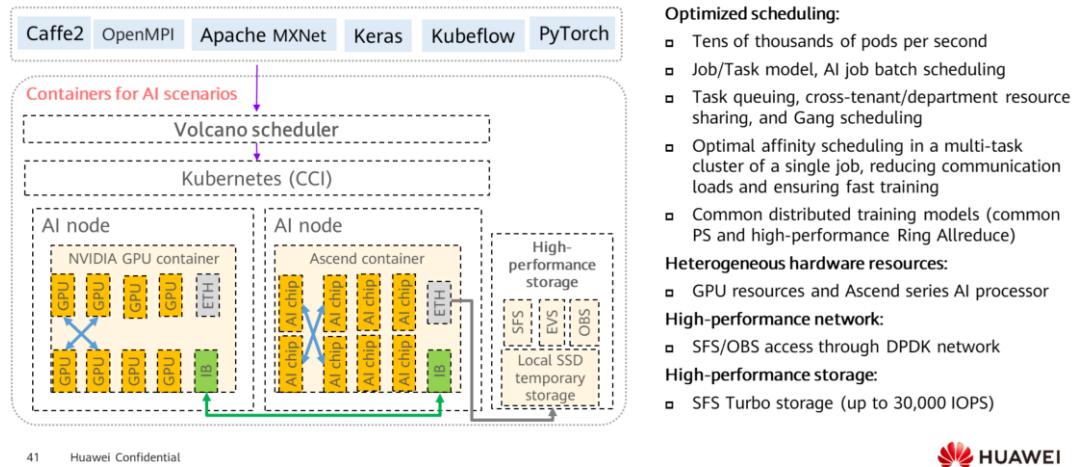
- Security: Lightweight virtualization enables VM-level isolation.
- Performance: Huawei-developed hardware virtualization acceleration technology delivers higher performance (kernel overhead less than 64 MB) at the startup speed of traditional containers (up to 300 ms).
- Compatibility: Kata Containers are compatible with mainstream container API specifications, such as Open Container Initiative (OCI) and Kubernetes Container Runtime Interface (CRI). They run on various hardware platforms and virtualization environments, enabling container images to be built once and run anywhere.



- CCI provides VM-level isolation without compromising the startup speed, offering you better container experience. It has the following features:
 - Native support for Kata containers
 - Kata-based kernel virtualization, providing comprehensive security isolation and protection
 - Huawei-developed virtualization acceleration technologies for higher performance and security

Fast Scaling

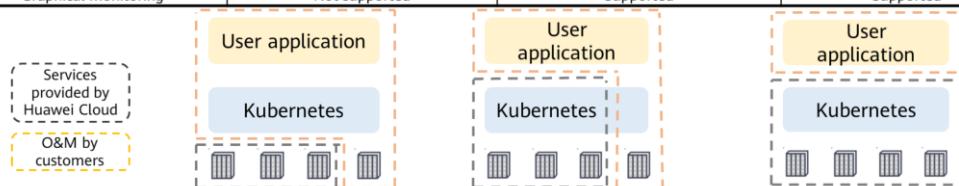
- Kubernetes clusters in CCI are created in advance. Their resources are unlimited from a single user's perspective and scalable in seconds.



- Currently, most big data and AI training and inference applications (such as TensorFlow and Caffe) run in containers. These applications are GPU intensive and require high-performance network and storage. As these applications are task-based, resources must be quickly allocated upon task creation and released upon task completion, and powerful compute and network resources as well as high I/O storage are required for high-density computing.
- CCI resources are billed on demand by second, reducing costs.
- Volcano is a batch processing platform based on Kubernetes. It provides a series of features required by machine learning, deep learning, bioinformatics, genomics, and other big data applications, as a powerful supplement to Kubernetes capabilities. Volcano provides general-purpose, high-performance computing capabilities, such as job scheduling, heterogeneous chip management, and job running management, serving end users through computing frameworks for different industries, such as AI, big data, gene sequencing, and rendering. (Volcano has been open-sourced in GitHub.)

O&M Free

Type	Self-Managed Kubernetes Cluster	Public Cloud Kubernetes Clusters	Using CCI
Applying for nodes	Required	Required	Not required
Establishing network and storage connections	Required	Not required	Not required
Creating Kubernetes clusters	Manual	Automatic	No attention required
Managing nodes in Kubernetes clusters	Manual	Automatic	No attention required
Deploying applications	Automatic	Automatic	Automatic
Scaling in or out clusters	Manual node application and management	Automatic node application and management	No attention required
Upgrading Kubernetes version	Manual	Automatic	No attention required
Locating/Fixing Kubernetes bugs	Manual	No attention required	No attention required
Maintaining cluster usage	Manual (with possibility of resource waste)	Manual (with possibility of resource waste)	No attention required
Upgrading hardware	Limited live migration	Live migration of nodes	Live migration of Kata Containers
Graphical monitoring	Not supported	Supported	Supported

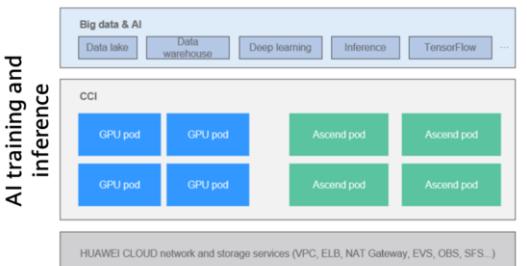


42 Huawei Confidential

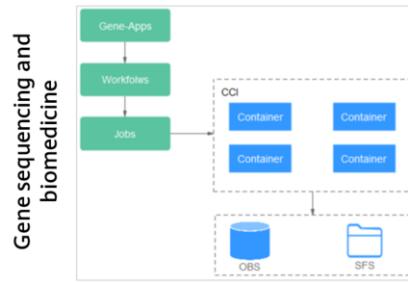


- No O&M is required for clusters and servers, which greatly reduces costs.

Application Scenarios



- Currently, most big data and AI training and inference applications (TensorFlow and Caffe) run in containers. These applications are GPU intensive and require high-performance network and storage. As these applications are task-based, resources must be quickly allocated upon task creation and released upon task completion.
 - Heterogeneous computing: Powered by GPU-/Ascend-based acceleration.
 - Faster training: No more virtualization layer overheads and better GPU linear acceleration ratio and distributed training.
 - Lower costs: Pay-per-use billing of training and inference resources.
 - O&M-free: An AI algorithm engineer is fully qualified.

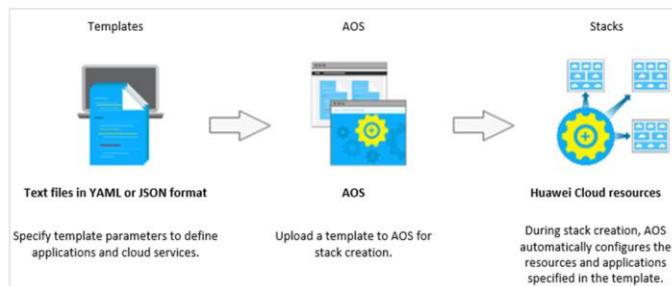


- Fields such as genomics and drug R&D require high-performance, high-density and O&M-free computing at low cost. As most scientific computing is task-based, resources must be quickly allocated for new tasks, and released once the tasks are completed.
 - High-performance computing: CCI provides powerful compute and networking resources and high-I/O storage.
 - Fast scaling: Auto scaling in seconds minimizes resource consumption.
 - O&M-free: Greatly reduces costs.
 - Pay-per-use: Containers start on demand and are billed by specification and duration.

- CCI is tailored for task-based scenarios.
 - These scenarios include heterogeneous hardware-based AI training and inference, training tasks can be hosted on CCI.
 - It also works in HPC scenarios, such as gene sequencing.
 - Third, burst scale-out in a long-term stable running environment, such as e-commerce flash sales and hot topic-based marketing.
- The main advantages of CCI are on-demand use for lower costs, and full hosting for O&M-free. It also enables consistency and scalability based on standard images.
- CCI supports pay-per-use or package-based billing. A core-hour indicates the number of cores multiplied by time. For example, 730 core-hours indicate that you can use 730 cores for one hour or one core for 730 hours.
 - In pay-per-use mode, you will be charged by second for each instance and the billing statistics are presented by hour.
 - In package-based billing mode, if your resource usage exceeds the quota of the package within the package validity period, you will be billed for the excess usage on a pay-per-use basis. If you buy multiple packages, resources in the package with the earliest expiration time will be used first.

Application Orchestration Service (AOS)

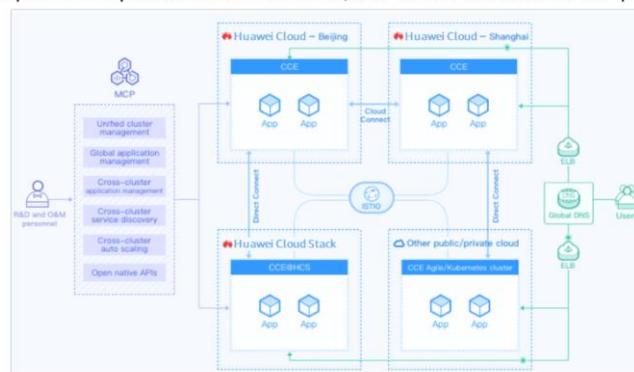
- Automate application migration to the cloud: Create, replicate, and migrate your applications and provision required resources in just a few clicks.
- AOS templates are text files that are easy to read and write, and edit them directly in YAML or JSON format.



- To work with AOS, you only need to create a template describing the applications and the required cloud resources, including their dependencies and references. AOS will then set up these applications and provision the resources as specified in the template. For example, when creating an ECS, together with a VPC and a subnet on which the ECS runs, you only need to create a template defining an ECS, VPC, subnet, and their dependencies. AOS will then create a stack, namely, a collection of resources you specified in the template. After the stack has been successfully created, the ECS, VPC, and subnet are available to use.
- Product functions:
 - AOS provides automatic orchestration of mainstream Huawei Cloud services. For details, see [Cloud Services and Resources that Can Be Orchestrated in AOS](#). AOS also provides lifecycle management including resource scheduling, application design, deployment, and modification, to reduce O&M costs through automation.
 - Standard languages (YAML and JSON) can be used to describe required basic resources, application systems, upper-layer services, and their relationships. Automatic resource provision, application deployment, and service loading can be implemented in a few clicks based on uniform description and defined dependency relationships. You can manage deployed resources and applications in a unified manner.
 - AOS Template Market provides abundant templates for free, including basic resource templates, service combination templates, and industry templates, covering common application scenarios. You can use public templates directly to deploy services in the cloud in a few clicks.

Multi-Cloud Container Platform (MCP)

- MCP is developed by Huawei Cloud after years of experience in cloud container fields and cluster federation technology (Karmada). It provides multi-cloud and hybrid cloud solutions to unify cluster management across clouds and deployment/traffic distribution of applications across clusters. It not only resolves multi-cloud disaster recovery, but also plays an important role in traffic sharing, decoupling of data storage and service processing, decoupling of development and production environments, and flexible allocation of computing resources.



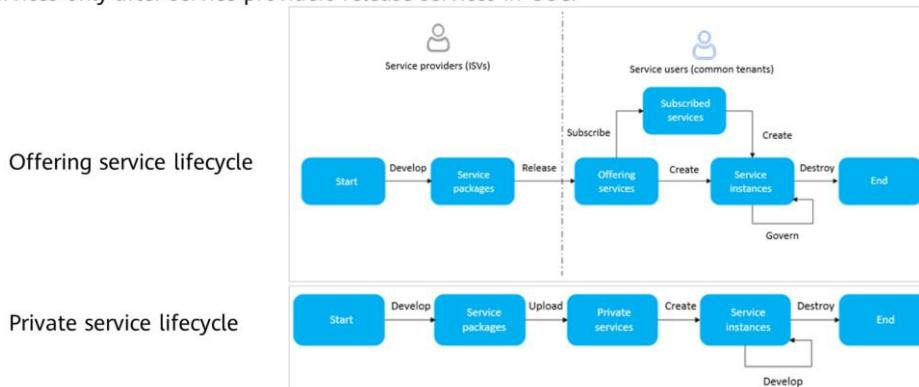
45 Huawei Confidential



- Karmada is a multi-cluster management system built on Kubernetes native APIs. It provides automated multi-cluster management capabilities in a pluggable manner for multi-cloud and hybrid cloud applications. Karmada enables centralized management, high availability, fault recovery, and traffic scheduling.
- MCP leverages cluster federation to implement unified management of clusters of different cloud service providers. As a unified entry for multiple clusters, MCP supports dynamic cluster access and global cluster monitoring dashboard.
- Based on the multi-cluster and federation technologies, MCP manages Kubernetes clusters across regions or clouds and supports full lifecycle management of applications across clusters, including deployment, deletion, and upgrade, by using standard cluster federation APIs in Kubernetes.
- MCP supports cross-cluster auto scaling policies to balance the pod distribution in each cluster and implement global load balancing.
- You can create federated Services for cross-cluster service discovery. MCP enables service region affinity based on the proximity access principle, reducing network latency.
- MCP is compatible with the latest Kubernetes-community federation architectures, Kubernetes native APIs and Karmada APIs.
- MCP supports application federation, which allows you to deploy an application from only one cluster to multiple clusters across clouds in just a few clicks. In this way, cross-cloud DR and traffic sharing are implemented.
- You can clone or migrate your applications to other clusters or across clouds/regions in just a few clicks without re-writing or modifying your service code.

Operator Service Center (OSC)

- A cloud native service lifecycle management platform for service providers and users: service development, release, subscription, deployment, upgrade, and update.
- The two major OSC roles are service providers and service users. Service users can subscribe to and use services only after service providers release services in OSC.



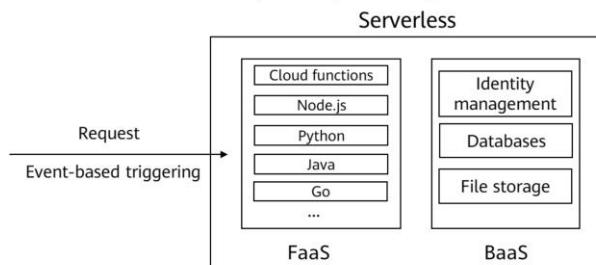
- Service release: Service providers upload a service package, verify the lifecycle and features of the service in the OSC, and release the service as an offering for other tenants to subscribe to.
- Service subscription: OSC contains Huawei-developed services, services published by ecosystem partners, and open source services. All services can be subscribed to by users. Instances can be deployed only after successful subscription.
- Service unsubscription: Users can unsubscribe from a service at any time. Upon unsubscription, the system automatically deletes the deployed services and instances.
- Private service uploading: Users can upload services developed based on Helm, Operator Framework, or OSC service specifications to OSC as private services for management.
- Service upgrade: When a provider publishes the updated version of a service, the subscribers will receive an upgrade notification and can decide whether to upgrade the service to the latest version.
- Instance deployment: After subscribing to a service, users can deploy an instance, specifying the region, container cluster, and running parameters.
- Instance O&M: OSC provides the O&M view of instances. Users can view the monitoring and logs of instances and switch from the O&M view to the corresponding cloud service for in-depth data analysis.
- Instance update: Users can modify the running configurations of an instance.
- Instance deletion: When the lifecycle of a service running in an instance ends, users can delete the instance to reclaim related resources.

Contents

1. Cloud Native Concepts and Background
2. Open Source Container Technologies
3. Huawei Cloud Container Services
- 4. Serverless Overview**

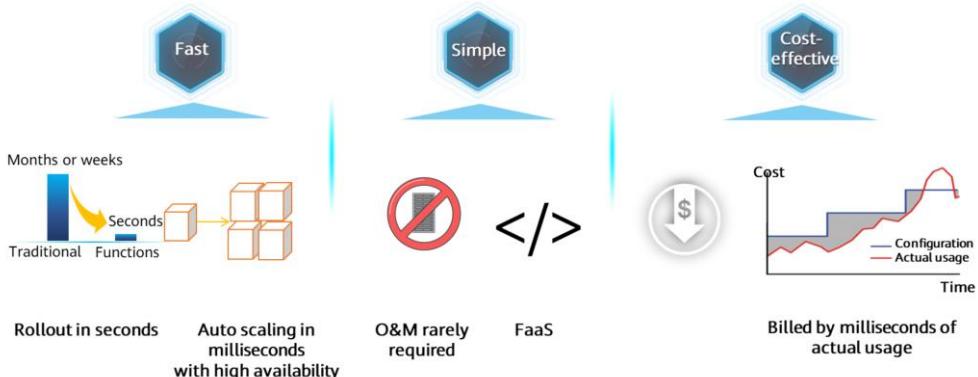
What Is Serverless?

- **CNCF:** Serverless computing refers to the concept of building and running applications that do not require server management. It describes a finer-grained deployment model where applications, bundled as one or more functions, are uploaded to a platform and then executed, scaled, and billed in response to the exact demand needed at the moment.
- **Wikipedia:** Serverless computing is a cloud computing execution model. The cloud provider allocates machine resources on demand, taking care of the servers on behalf of their customers, who are not involved in capacity planning, configuration, management, maintenance, fault tolerance, or scaling of containers, VMs, or physical servers.
- Serverless is a collection of cloud capabilities, not a single cloud service.



- Serverless computing does not mean that we no longer use servers to host and run code, nor does it mean that O&M engineers are no longer needed. Conversely, it means that consumers no longer need to spend time and resources on configuring, maintaining, updating, or expanding servers, or planning capacity. All these are handled by a serverless platform, enabling developers to focus on service logic and O&M engineers to process key service tasks.
- There are two serverless architectures:
 - Functions-as-a-service (FaaS): provides event-driven computing. Developers use functions triggered by events or HTTP requests to run and manage application code. They deploy small units of code to FaaS, where the code is executed as discrete actions on request, and can be expanded without managing servers or any other underlying infrastructure.
 - Backend-as-a-service (BaaS): an API-based third-party service that can replace the core function subset in applications. Because these APIs are provided as services that can be automatically expanded and transparently operated, they are serverless for developers.
- FaaS executes function code, and BaaS only uses APIs to provide backend services on which applications depend.

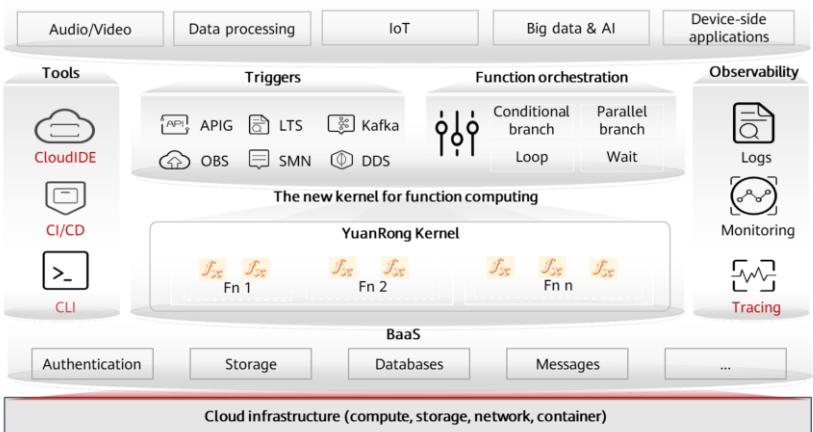
Why Serverless?



- Generally, serverless is recommended for workloads in the following scenarios: Asynchronous, concurrent, easy to be parallelized into independent units. Infrequent requests with huge and unpredictable expansion requirements. Stateless and transient, without instant cold start requirements. Highly dynamic service requirement changes.
- Serverless products or platforms have the following benefits:
 - No need to configure, update, or manage servers. Managing servers, VMs, and containers involves personnel, tools, training, and time.
 - FaaS and BaaS products can be scaled flexibly and precisely to process each request. For developers, a serverless platform does not need capacity planning or auto scaling triggers or rules.
- No cost for idle resources: For consumers, a major benefit of serverless products is that idle resources do not incur any cost. For example, idle VMs and containers will not be charged. However, the costs for stateful storage, functions, and feature sets will be charged.

Huawei Cloud Serverless Service – FunctionGraph

- This service hosts and computes event-driven functions while maximizing scalability and efficiency. Simply write your code and set running conditions without provisioning or managing servers.

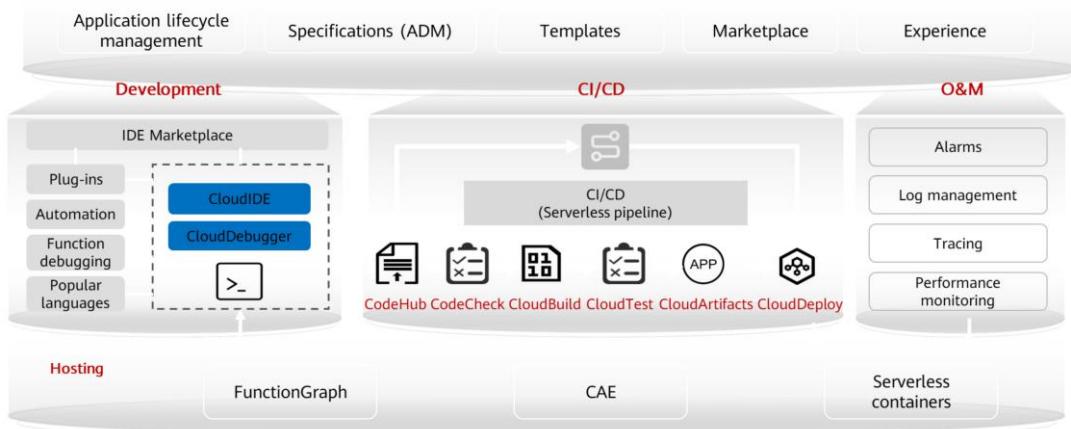


50 Huawei Confidential



- When using FunctionGraph, you do not need to apply for or pre-configure any compute, storage, or network services. You only need to upload and run code in supported runtimes. FunctionGraph provides and manages underlying compute resources, including CPUs, memory, and networks. It also supports configuration and resource maintenance, code deployment, automatic scaling, load balancing, secure upgrade, and resource monitoring.
- FunctionGraph supports Node.js, Java, Python, Go, and C#, allowing you to edit code inline, import OBS files, and upload ZIP and JAR packages. It uses SMN, APIG, and OBS triggers. It collects and displays real-time metrics and logs, and enables you to query logs online, making it easy to view function status and locate problems. Function flows orchestrate and coordinate multiple distributed functions. FunctionGraph provides unified plug-ins for on-/off-cloud development and debugging. HTTP functions can be triggered for web service optimization by sending HTTP requests to specific URLs. In addition, you can enable tracing on the function configuration page so that you can view Java virtual machine (JVM) and tracing information on the APM console. Currently, this feature is only available for Java functions. You can package and upload container images to FunctionGraph for running.
- FunctionGraph 2.0 is a next-generation function computing and orchestration service. It has the following features:
 - Deep integration with CloudIDE, concurrent function debugging, tracing, wizard-based building, and full lifecycle management
 - Six programming languages and custom runtime, cold startup and auto scaling in 100 milliseconds
 - First to support stateful functions in China, visualized function orchestration
 - Serverless web applications with zero reconstruction

Serverless Lifecycle Management



51 Huawei Confidential

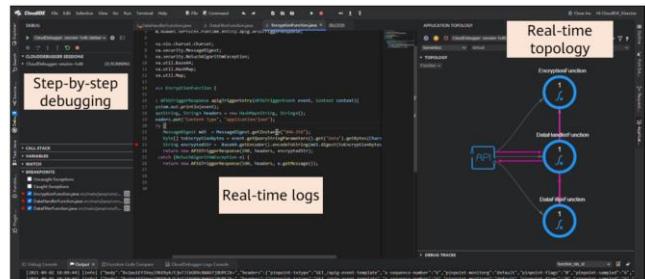


- Application development: out-of-the-box CloudIDE, debugging and tracing of clustered serverless applications, code breakpoints, stack viewing, call topologies, and hot code replace (HCR)
- CI/CD: deep integration with serverless runtimes; lightweight DevOps with O&M tools
- Application hosting: lifecycle management with unified specifications; templates and marketplace for experience and reuse
- Cloud application engine (CAE): a one-stop serverless application hosting service that enables ultra-fast deployment at low cost with simple O&M. It releases applications from source code, software packages, and image packages, with seconds of auto scaling, pay-per-use billing, no infrastructure O&M, and multiple observable metrics.

On-/Off-Cloud Development and Debugging with Unified Plug-ins

(On-cloud) CloudIDE

- Create function using template
- View function on the cloud and download to CloudIDE for debugging
- Push function to the cloud
- Supports Java and Node.js

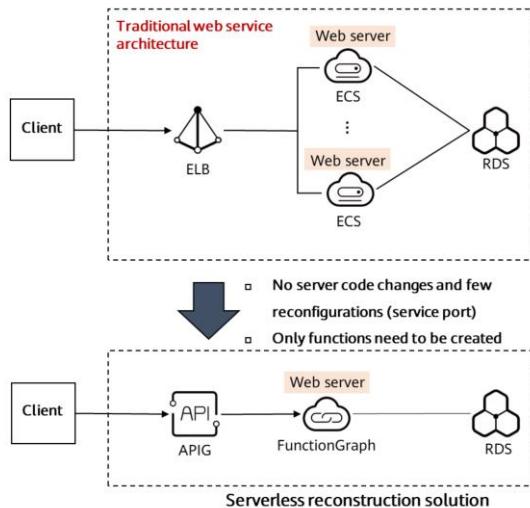


(Off-cloud) VSCode plug-in

- Create function using template
- View function on the cloud and download to local host for debugging
- Push local function to the cloud
- Supports Node.js and Python

- (On-cloud) CloudIDE: Create a function using a template, view the function and download it to the cloud, debug it using CloudIDE, and push it to the cloud.
- (Off-cloud) VSCode plug-in: Create a function using a template, view the function on the cloud, download it to a local host, debug it using VSCode plug-in, and push it to the cloud.

Low Reconstruction Costs

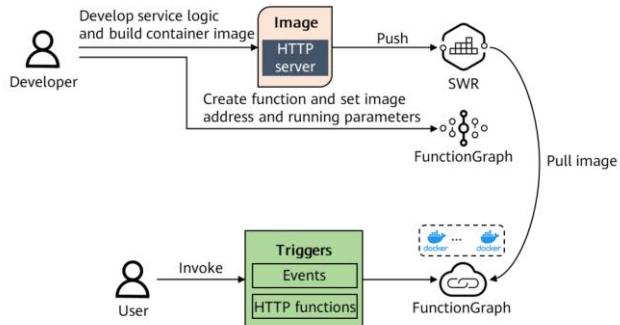


- Package the original web service code into an HTTP function using APIG and FunctionGraph for serverless reconstruction.
- Reconstruct applications developed using frameworks like Java - Spring Boot and Node.js - Express with minimal modification.
- Use your familiar development frameworks and test tools with low learning curve.
- Enable auto scaling and gray upgrade with simple built-in mechanisms.

- HTTP functions are better for optimizing web services and can be triggered by sending HTTP requests to specific URLs. You can specify this type when creating a function. HTTP functions only support APIG and APIC triggers.

Container Images

- Package container images and upload them to FunctionGraph for running. Customize your own code package with high flexibility and low migration costs.
- Reuse the container ecosystem (open-source CI/CD tools and standard version management). No need for refactoring code or re-compiling binary dependencies.



54 Huawei Confidential

HUAWEI

- The following challenges may exist when you shift from the traditional development mode to the serverless mode:
 - Different runtimes and deliverable formats: The runtime provided by serverless function vendors may be Docker or microVM. The deliverable formats and function signatures are different. You have to make adaptations.
 - Immature ecosystem: Popular open-source tools (such as CI/CD pipeline) are not supported.
- The container ecosystem is mature and does not have portability and agile delivery issues. Container images are standard deliverables in the cloud native era. However, containers still involve O&M and idle resource costs.
- You can create custom images for both event and HTTP functions.

Scenarios

Web applications

Use FunctionGraph with other cloud services to quickly build cloud-native web applications by simply writing code, for better rollout efficiency and lower O&M costs.

- Applet backends
- Web backends
- Q&A bots
- Backends for frontends (BFF)
- Microservice applications

Event-driven applications

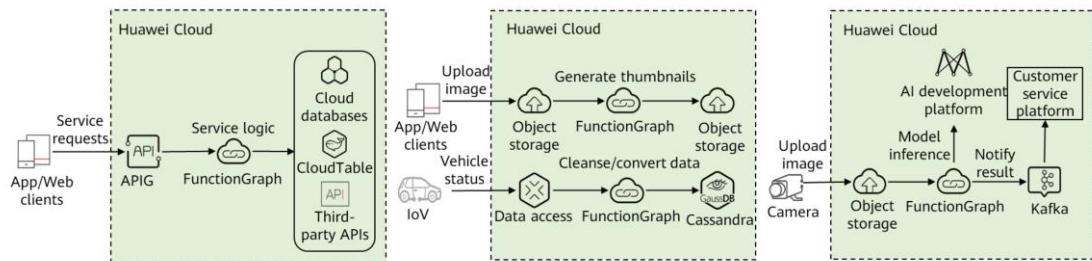
Services are driven by events, and resources are provided on demand, so service peaks/troughs are not concerns. Idle resources are not billed, so O&M costs are reduced.

- Real-time image processing
- Real-time stream processing
- IoT event processing
- O&M alarm handling

AI applications

Intelligence evolution requires various services to be integrated for quick rollout.

- Live streaming
- AI inference
- Facial recognition
- License plate recognition



Quiz

1. (True or false) In Kubernetes, pods are the minimum deployment units that can be created, scheduled, and managed. You can use ConfigMaps to label and classify pods.
 - A. True
 - B. False
2. (Multiple-answer question) Which of the following statements are true about serverless FunctionGraph?
 - A: It is O&M-free.
 - B: It hosts and computes event-driven functions in a serverless context.
 - C: It has two forms: BaaS and PaaS.
 - D: Currently, it supports only online editing in Java.

- Answer 1: False
 - Label instead of ConfigMaps
- Answer 2: ABC
 - Supports multiple languages, such as Node.js, Java, Python, Go, and C#

Quiz

1. (Discussion) What are the advantages of cloud container services (such as CCE) over on-premises, self-built Kubernetes clusters?

2. (Discussion) What are the differences between deploying applications on ECSs and deploying services on CCE?

- Discussion 1:
 - Construction cost, including equipment, site, and prices
 - O&M costs, including manpower, power, and network costs
 - Security
 - Convenience
- Discussion 2:
 - Response speed
 - Performance
 - Security
 - Maintainability
 - Cost
 - Convenience

Summary

- This course walks you through open-source container technologies such as Docker and Kubernetes, and help you understand related products and functions. You can learn about how Huawei Cloud Container Engine (CCE), Cloud Container Instance (CCI), and other container services can help you migrate to the cloud, and how FunctionGraph, Huawei's serverless solution, implements serverless cloud computing.

Acronyms and Abbreviations

- AOM: Application Operations Management
- AOS: Application Orchestration Service
- API: Application Programming Interface
- APM: Application Performance Management
- AS: Auto Scaling
- BMS: Bare Metal Server
- CCE: Cloud Container Engine
- CCI: Cloud Container Instance
- CI/CD: Continuous Integration/Continuous Delivery
- CNCF: Cloud Native Computing Foundation
- DDoS: Distributed Denial of Service
- DevOps: Development and Operations
- DataArts Studio: Data Lake Governance Center
- DIS: Data Ingestion Service
- DLI: Data Lake Insight
- DNS: Domain Name Service
- ECS: Elastic Cloud Server
- EIP: Elastic IP
- ELB: Elastic Load Balance
- EVS: Elastic Volume Service
- GSLB: Global Server Load Balance
- HA: High Availability

Acronyms and Abbreviations

- IAM: Identity and Access Management
- IDC: Internet Data Center
- IMS: Image Management Service
- ISV: independent software vendor
- MCP: Multi-Cloud Container Platform
- MRS: MapReduce Service
- NAT: Network Address Translation
- NAT: Network Address Translation
- OBS: Object Storage Service
- OCI: Open Container Initiative
- OCR: Optical Character Recognition
- RDS: Relational Database Service
- SMN: Simple Message Notification
- SWR: SoftWare Repository for Container
- VM: virtual machine
- VPC: Virtual Private Cloud
- VPN: Virtual Private Network
- WAF: Web Application Firewall

Recommendations

- Huawei iLearning
 - <https://e.huawei.com/en/talent/portal/#/>
- Huawei Cloud Help Center
 - <https://support.huaweicloud.com/intl/en-us/index.html>
- HUAWEI CLOUD Developer Institute
 - <https://edu.huaweicloud.com/intl/en-us/>
- Huawei Talent Online
 - <https://e.huawei.com/en/talent/portal/#/>

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©2022 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.



Cloud Native Application Architecture



Foreword

- Evolving IT technologies mean more digital enterprises, which migrate their services to the cloud. More applications are delivered via API and microservice instead of software packages.
- This course describes the development and evolution of microservice architectures: Spring Cloud, Istio, and Huawei Cloud services, such as Application Service Mesh (ASM), DevCloud, and ServiceStage.

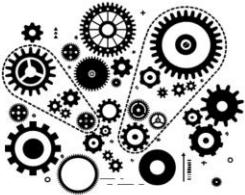
Objectives

- Upon completion of this course, you will understand:
 - Microservice evolution
 - Mainstream open-source architectures
 - Cloud native services from Huawei Cloud

Contents

1. Cloud Native Applications and Microservices
2. Mainstream Frameworks of Cloud Native Applications
3. Huawei Cloud Native Application Solutions

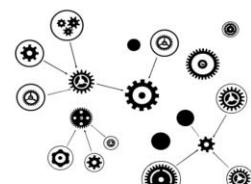
Architecture Evolution



1st generation: monolithic architecture



2nd generation: service-oriented architecture (SOA)



3rd generation: microservice architecture

- Tightly coupled
- Interdependence: A single change affects the entire system.
- Recurring investment in OSs, databases, and middleware
- Closed architecture

- Loosely coupled
- Popular with large enterprises
- System integration through enterprise service buses (ESBs)
- Large teams, each with 100 to 200 members
- Time to market (TTM): one year, half a year, or months
- Centralized scaling with planned downtime

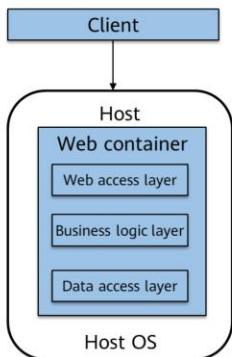
- Decoupled
- Favored by small- and medium-sized enterprises (SMEs), startups, and Internet companies
- Small team: a few to a dozen people
- TTM: upgrading in days or weeks
- DevOps: fully automated continuous integration (CI) and continuous delivery (CD)
- Auto scaling
- High reliability: hitless upgrade and scaling

Microservices improve team R&D efficiency, are compatible with new technologies, and shorten service rollout periods.

- Enterprises need to make a trade-off between rapid service development and exquisite application architecture. Microservice architecture is the future trend. The microservice architecture has abundant features, including fault tolerance, quick rollout, more complex functions, high availability, requirement response, manageability, and independent module release.
- On the monolithic architecture, all functions are integrated in one project. The architecture is simple, the development cost in the early phase is low, and the development period is short. Therefore, this architecture is ideal for small-scale projects. However, as the small projects grow larger, it is difficult to develop, expand, and maintain the monolithic architecture.
- Projects using the monolithic architecture are vertically divided, so small projects cannot become too large.
- On the SOA, repeated common functions are extracted as components to provide services for each system. Projects (or systems) communicate with services through WebService or remote procedure call (RPC). SOA improves the development efficiency and system reusability and maintainability.
- But the SOA has disadvantages. The boundary between systems and services is blurred, which is not conducive to development and maintenance. The granularity of the extracted services is too large, and systems are highly coupled with the services.
- The microservice architecture is an approach to developing a single application as a suite of small services, each running in its own process, and communicating with lightweight mechanisms, often an HTTP resource API. Services are split at a finer granularity, which facilitates resource reuse and improves development efficiency. In this way, optimization solutions for each service can be formulated more accurately, improving the system maintainability.

Monolithic Architecture

- These applications have a single layer. User interface and data access functions are combined into a single program on a single platform. Monolithic applications operate independently and perform all steps required for a specific function.



Advantages

- Simple
- Easy to test
- Easy to deploy and upgrade

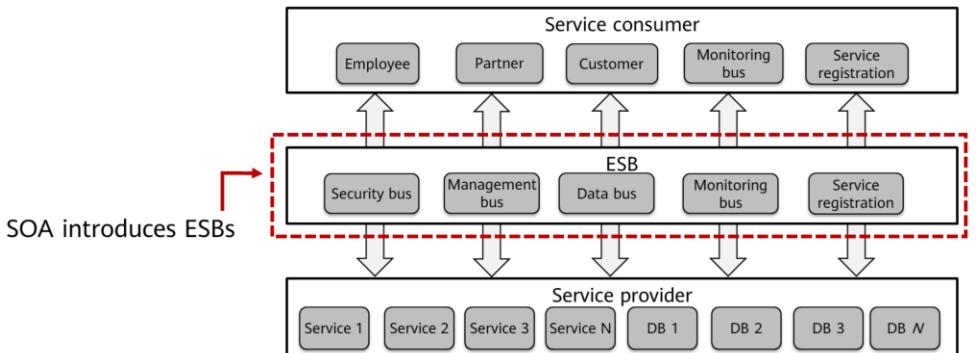
Disadvantages

- High complexity
- Performance bottlenecks
- Poor scalability
- No fault isolation, low reliability
- High maintenance costs

- The monolithic architecture is an archive package. The package contains applications with all functions. In the early stage of software development, the monolithic architecture is popular because it is easy to deploy, the technologies are simple, and the labor cost is low. However, in the Internet era, the complexity of service requirements and the delivery frequency increase. The traditional monolithic architecture cannot meet the requirements of developers:
 - The monolithic architecture is complex as all modules are coupled. They have blurred boundaries and complex dependencies. Function adjustment may bring unknown impacts and potential bugs.
 - When a monolithic system encounters a performance bottleneck, the system can only scale out horizontally and add service instances to balance the load. Vertical expansion and module decoupling are not supported.
 - The monolithic architecture has poor scalability. A monolithic application can be scaled only as a whole. Scaling of a single module cannot be performed.
 - The monolithic architecture cannot isolate faults. The entire system may break down even when a small module is faulty (for example, a request is blocked) as all function modules are aggregated.
 - On the monolithic architecture, the release impact is large. The entire system is released each time and the system restarts upon each release. This poses a great challenge to a large-scale integrated system. If we decouple each module, only the modified module needs to be released.
 - The deployment slows down. The build and deployment duration increases as the code size increases.
 - Technological innovation is hindered. A monolithic application solves all problems using a unified technical platform or solution. Each team member must use the same development language and architecture.

SOA

- This component model links an application's functional units (services) through well-defined application programming interfaces (APIs) and service protocols. APIs are defined neutrally and must be independent of service hardware platform, OS, and programming language. This allows services built in multiple such systems to interact in the same way.



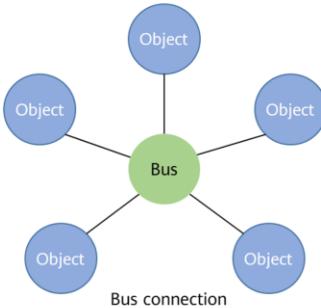
6 Huawei Confidential



- SOA decouples applications, modularizes them, and builds functions into independent units to provide services.
- SOA contains multiple services. The services communicate with each other through mutual dependency or communication mechanisms to provide a series of functions. A service independently exists in an OS process. Services invoke each other through networks.
- SOA solves the following problems:
 - First, system integration. SOA sorts out the mesh structure between scattered and unplanned systems into a regular and governable star structure. Some products, such as the ESB, technical specifications, and service management specifications, need to be introduced.
 - Second, system as a service. SOA abstracts service logic into reusable and assemblable services and orchestrates the services to quickly regenerate services. This transforms inherent functions into common services to quickly reuse business logic.
 - Third, business as a service. SOA abstracts enterprise functions into reusable and assemblable services. It transforms the enterprise architecture into a service-oriented one to provide better services. SOA solves the problems of system invoking and system function reuse from the technical perspective.

ESB

- The most common component of SOA and a software architecture built on by middleware infrastructure. It serves more complex service-oriented architectures through event-driven and XML-based messaging engines.
- An ESB typically provides an abstraction layer on the enterprise messaging system. Integration architects can use messages for integration without coding.



- Functions
 - Monitor and route communication messages
 - Control service version and deployment
 - Monitor communication quality and rectify faults
 - Handle troubleshooting
- Disadvantages
 - Larger systems have more calls and put more access pressure on the ESB.

- The term bus is an extension of a physical bus that transports bits between different devices of a computer. An ESB provides similar functions at a higher abstraction level. In an enterprise architecture that uses an ESB, applications interact with each other through the bus, and the bus schedules information between applications. ESB reduces the number of point-to-point connections required for interaction between applications, making it easier and more intuitive to analyze the impact of major software changes. Reconstruction of a component in the system also becomes simple.
- An ESB provides reliable message transmission, service access, protocol conversion, data format conversion, and content-based routing regardless of physical locations, protocols, and data formats.
- In the future, the enterprise integration architecture will integrate application APIs, messages, devices, data, and multiple clouds, and connect all applications, big data, cloud services, devices, and partners of enterprises. The traditional "integration factory" mode controlled by IT teams will be transformed to the self-service integration mode that is supported by business lines, subsidiaries, application development teams, and end users, that is, the "unified hybrid integration platform."

Microservice Description

- Microservice is a software architecture style based on small building blocks that focus on single responsibilities and functions. It combines complex, large-scale applications as modules. Functional blocks communicate with each other using language-independent/-agnostic APIs.
 - A microservice provides service functions with open, language-agnostic APIs (most commonly HTTP). An application consists of one or more microservices.

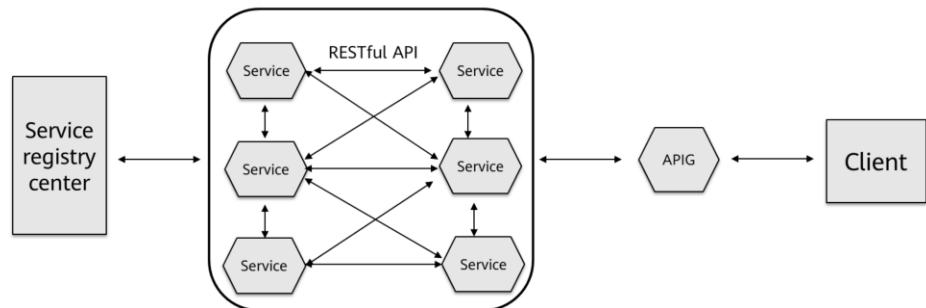


8 Huawei Confidential

 HUAWEI

- The origin of microservices was Micro-Web-Service proposed by Dr. Peter Rodgers at the 2005 Cloud Computing Expo. Juval Lowy had a similar idea, that is, to turn categories into granular services. The core was that services are used by Unix-like pipelines. In 2014, Martin Fowler and James Lewis jointly proposed the concept of microservices. The concept defines microservices as small services composed of single applications and has its own schedule and lightweight processing. Services are designed according to functions and automatically deployed. They communicate with other services using HTTP APIs. In addition, services are managed at the minimum scale (such as Docker), and implemented using different programming languages and components such as libraries.
 - Microservice is an architecture and organization method for developing software. Software consists of small independent services that communicate with each other through clearly defined APIs.
 - The microservice architecture emerged because any small change in a monolithic architecture affects all other modules. Any small change on the cloud needs to be compiled and released in a unified manner. If a module needs to be extended, the whole architecture needs to be extended. Therefore, a series of microservices are used to build applications. Microservices can be independently deployed and expanded, and can be developed using different languages. Moreover, modular boundaries are provided.

Microservice Architecture

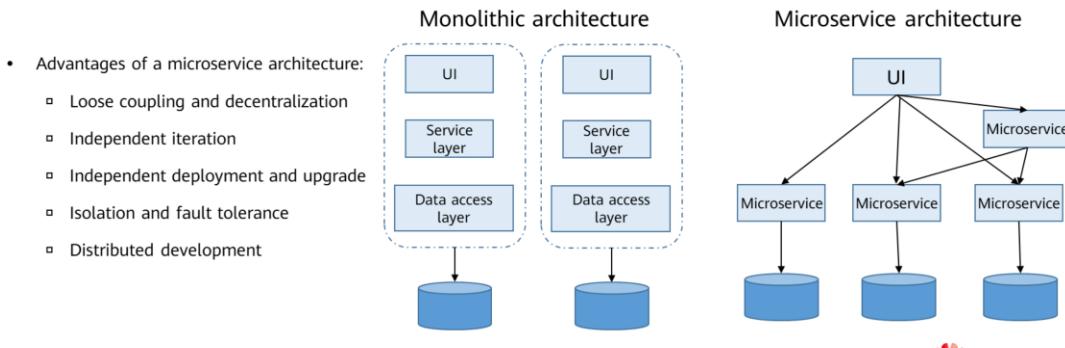


- The microservice architecture breaks down applications into **groups of small services** called "microservices", which cooperate with each other to provide business capabilities.
- Each service **runs as an independent process**. **Lightweight** communication mechanisms (usually based on RESTful APIs) are used between services.
- Each microservice is **designed, developed, and tested by an independent team**, and can be deployed independently and automatically in the production environment.

- In a microservice architecture, the entire web application is organized into a series of small web services. These small web services can be compiled and deployed independently and communicate with each other through their exposed APIs. They cooperate with each other to provide functions for users as a whole, but can be expanded independently.
- Microservice is a software architecture style based on small building blocks that focus on single responsibilities and functions. It combines complex, large-scale applications as modules. Functional blocks communicate with each other using language-independent or language-agnostic APIs.
- Microservices follow the design concept that focus on functions. During application design, an application can be divided based on functions or processes. Each function is independently implemented as a service that can be independently executed. Then, the same protocol is used to combine all the services to form an application. If a specific function needs to be extended, you only need to operate that function, not the entire application.
- The API gateway is generally located on the execution path of each API request. It belongs to the data plane, receives requests from clients, reversely proxies the requests to underlying APIs and execute traffic control and user policies before that. Before proxying the request back to the original client, it can also respond to the instructions of the underlying API and execute the corresponding policy again.
- RESTful APIs are REST-styled. RESTful is a development and design style of network applications and can be defined in XML or JSON format.

Microservice Features

- Different from traditional monolithic architectures, in a microservice architecture, an application can be split into multiple core functions. Each function is called a service and can be built and deployed independently. This means that services do not affect each other when they are running normally or if they are faulty.
- In microservice, an application consists of many small, loosely coupled services, opposite to the overall approach where an application is monolithic, tightly coupled.



10 Huawei Confidential

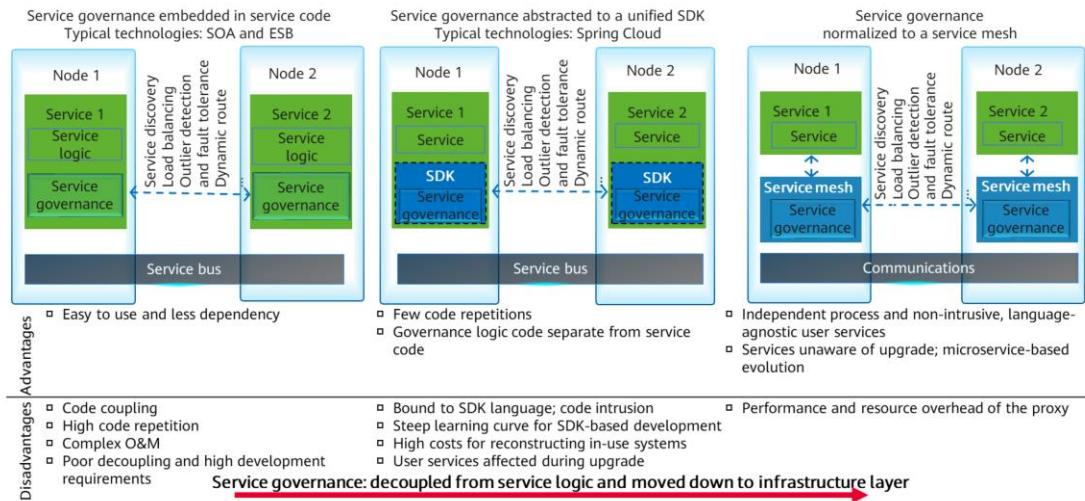


- Complexity is solved by decomposing huge monolithic applications into multiple services. An application is divided into multiple manageable branches or services without changing functions. Each service is defined through APIs.
- The microservice architecture provides a modular solution for functions that are impossible in monolithic encoding. A single service is easy to develop, understand, and maintain.
- Microservices are independently implemented and deployed, that is, they run in independent processes. Therefore, they can be independently monitored and expanded.
- In the microservice architecture, each microservice is independently deployed. Developers do not need to worry about whether other services will impact the microservices.
- The microservice architecture enables each service to be developed by a dedicated development team. Developers can freely choose development technologies to provide API services.

Contents

1. Cloud Native Applications and Microservices
2. **Mainstream Frameworks of Cloud Native Applications**
3. Huawei Cloud Native Application Solutions

Architecture Evolution



12 Huawei Confidential



- In the early stage, communication between computers required a physical layer to transmit bytecodes and electronic signals at the bottom layer. In addition to service logic, services also needed to handle a series of network transmission problems such as packet loss, disorder, and retry.
- In the 1980s, TCP was published, solving the common traffic control problems in network transmission. The technology stack was moved downwards and extracted from services to become a part of the network layer in the OS.
- In the 1990s, network communication between computers was no longer a problem. Distributed systems represented by GFS, BigTable, and MapReduce developed rapidly. Communication semantics specific to distributed systems emerged, such as circuit breaker policies, load balancing, service discovery, authentication and authorization, quota limit, and monitoring. Each service needed to implement the required semantics.
- To address this issue, microservice-oriented development frameworks with common semantic functions are developed, including Finagle from Twitter, Proxygen from Facebook, and Spring Cloud.
- However, developers still need to handle the complex frameworks, and track and solve framework problems. In addition, the frameworks usually support only one or several languages. Services that are not written with the framework-supported languages are difficult to integrate into the frameworks. Therefore, the proxy (sidecar) mode represented by Linkerd and Envoy emerges. This is the first-generation Service Mesh.
- To provide a unified upper-layer O&M portal, a centralized control panel is introduced. All single-node agent components interact with the control panel to update network topology policies and report data. In this model, each service is paired with a sidecar proxy. Services communicate with each other only through sidecars. This is the second-generation Service Mesh, represented by Istio, a joint

project launched by Google, IBM, and Lyft.

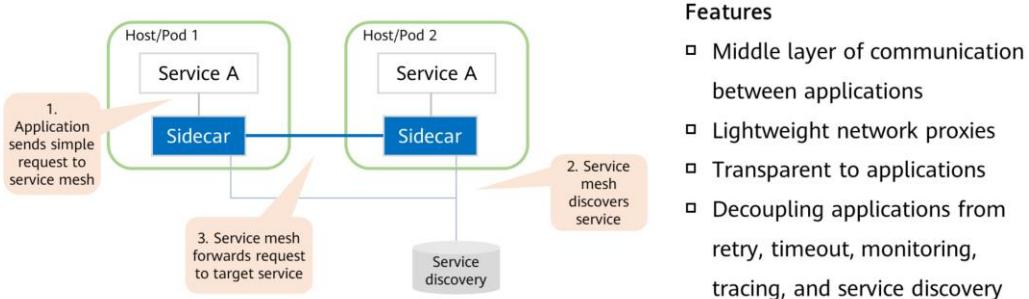
Spring Cloud Overview

- Spring Cloud is a complete framework for microservice building. Built on Spring Boot, it provides the necessary components for configuration management, service discovery, circuit breakers, intelligent routing, micro-proxy, control bus, global locks, leadership election, distributed sessions, and cluster state management during microservice development. It makes it easier to develop cloud services in the microservice architecture.
- Spring Boot streamlines the creation of Spring applications and services as products thanks to simplified configuration files, embedded web servers and out-of-the-box microservices.

- Spring Cloud is an ordered set of microservice solutions or frameworks for building distributed microservice systems. Based on Spring Boot, it encapsulates mature and verified microservice frameworks in the market to shield complex configurations and implementation principles.
- Spring Cloud sub-projects can be classified into two types. Most sub-project are for encapsulating and abstracting mature frameworks by using Spring Boot. The other type is for implementing infrastructure with certain distributed systems developed. For example, Spring Cloud Stream works similarly to Kafka or ActiveMQ.
- Spring Boot is a new framework provided by the Pivotal team. It simplifies the initial setup and development process of Spring applications. The framework is configured in a specific way so that developers no longer need to define a templated configuration. In this way, Spring Boot is committed to becoming a leader in the rapid application development field.
- Spring Cloud has the following features:
 - It has strong support from companies such as Netflix, and active contributions from the Spring open source Community.
 - By combining mature microservice products and frameworks in a standardized manner, Spring Cloud delivers a complete set of microservice solutions to reduce development costs and risks.
 - Thanks to Spring Boot, Spring Cloud features simple configuration, quick development, easy deployment, and convenient testing.
 - Spring Cloud can call REST services. Compared with RPC, REST is more lightweight and flexible. REST services depend on a contract rather than code, facilitating cross-language implementation, release and deployment.
 - Spring Cloud is compatible with Docker and Kubernetes microservice orchestration.

Service Mesh Features

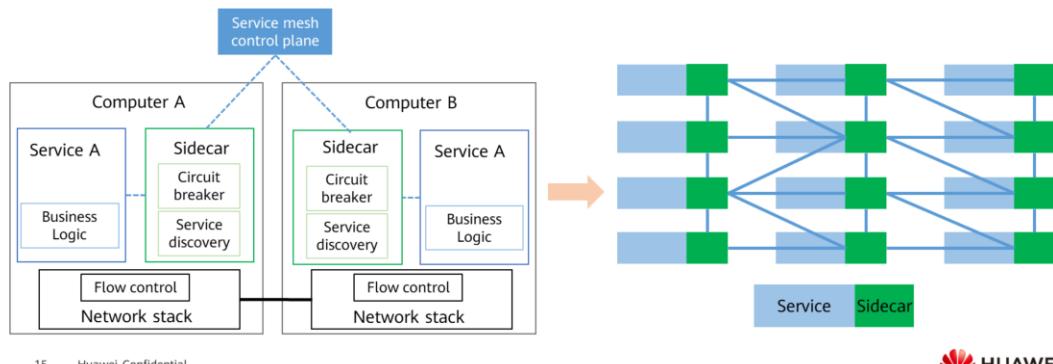
- This dedicated infrastructure layer processes service communication and transmits requests over complex cloud native application topology. In practice, it is a group of lightweight network proxies that are deployed with application services and remain transparent to them.



- The term service mesh was first proposed by Buoyant and first publicly used in 2016. In 2017, Buoyant released its first Service Mesh product, Linkerd, and an article *What's a service mesh? And why do I need one?* The article provides an authoritative definition of service mesh.
- Without a service mesh layer, the logic governing communication can be coded into each service. However, as the communication between microservices and for logical governing becomes more complex, service meshes are required to integrate a large number of discrete services into one functional application.
- Similar to a TCP/IP layer between applications or microservices, a service mesh is responsible for network invoking, rate limiting, outlier detection, and monitoring between services. Developers usually do not need to pay attention to the TCP/IP layer when developing applications. Similarly, service meshes free developers from what service frameworks like Spring Cloud and Netflix OSS can implement.
- Without a service mesh, each microservice is coded with logic to govern service-to-service communication, which means developers are less focused on service development. It is also more difficult to diagnose communication failures because the logic that governs inter-service communication is hidden within each service.
- Service Mesh describes the network of microservices that make up applications and the interactions between applications. As a service mesh grows in size and complexity, it can become harder to understand and manage. You need to take care of basic operations, such as service discovery, load balancing, failure recovery, metrics, and monitoring. Advanced O&M includes blue-green deployment, canary release, rate limiting, access control, and end-to-end authentication.

Service Mesh Architecture

- Each service is paired with a reverse proxy server (service proxy or sidecar). A service and sidecar share one container managed by an orchestration tool. The sidecar communicates with other services for service discovery, load balancing, authentication/authorization, and secure communication.
- Sidecars abstract the communication of distributed services as an independent layer. They are deployed together with services to take over service traffic and communicate through proxies. Sidecars also provide load balancing, service discovery, monitoring, tracing, and traffic control to manage distributed systems.



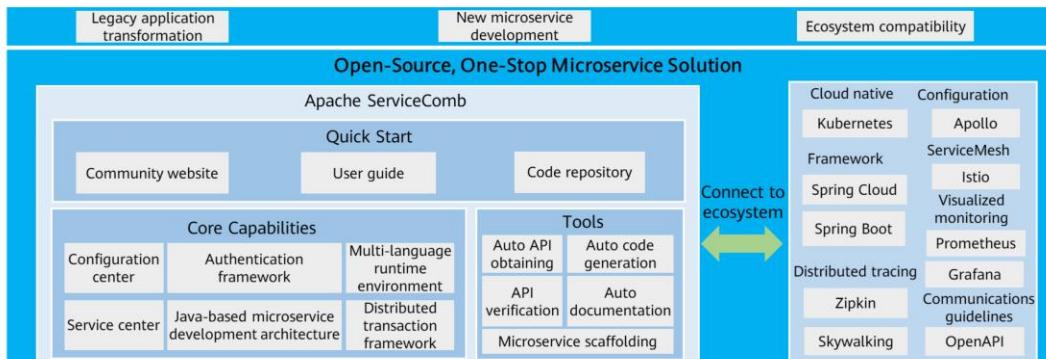
15 Huawei Confidential



- Rules must be defined in the logical governing layer of each service for communication between microservices. A service mesh extracts the rules from each service and abstracts them as an infrastructure layer. The service mesh does not add new functions to the runtime environment of each microservice. When microservices communicate with each other, requests are routed through the proxies of the service mesh. The proxies are also called sidecars. They run independently alongside services and form a mesh network. Sidecar proxies work with microservices to route requests to other proxies.
- A sidecar is a design pattern which separates certain functionality or a set of functionalities from the application into a separate process. It can add functionality non-intrusively to the application without adding additional code to meet third-party requirements. In software architecture, a sidecar is loosely coupled with a main or parent application to extend and enhance functionality.
- In a service mesh, services and their sidecar proxies constitute a data plane for data management, request processing and response. A service mesh also includes a control plane for managing interactions between services that are coordinated by sidecar proxies.
- In the service mesh workflow, the control plane pushes service configurations in the entire mesh to the sidecar proxy of each node. The routing information can be dynamically configured for all or certain services. After confirming the destination address, the sidecar sends the traffic to the corresponding service discovery endpoint.

ServiceComb

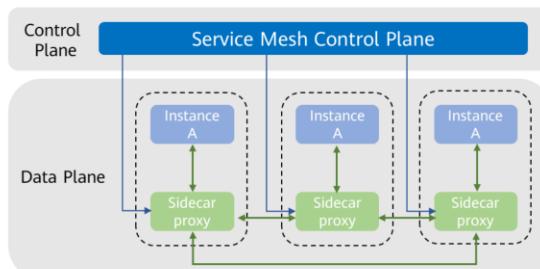
- ServiceComb is a prominent microservice project contributed by Huawei and incubated by Apache in 2017. It is the first Apache-incubated microservice project.
- It fully connects to the open source ecosystem, allowing enterprises, individuals, and developers to deploy applications as microservices for efficient O&M. ServiceComb delivers diverse products that flexibly combine for different scenarios, facilitating cloud migration.



- ServiceComb, the top microservice project, was contributed by Huawei and incubated by Apache in 2017. It is the first Apache-incubated microservice project.

What Is Istio?

- As microservice deployment grows in size and complexity, developers demand better service discovery, load balancing, failure recovery, monitoring, canary release, blue-green deployment, rate limiting, access control, and end-to-end authentication. Istio reduces complexity while easing the strain on development teams.
- Istio is an open platform for connection, protection, control, and observation. It provides complete non-intrusive microservice governance for better cloud native service management, network connection, and security.



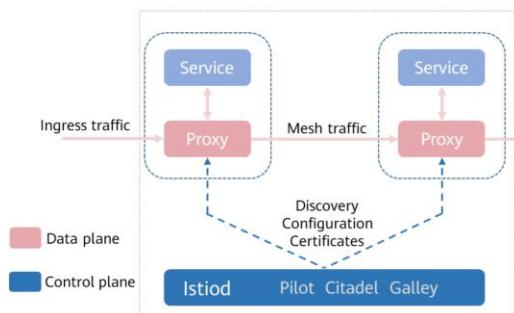
17 Huawei Confidential



- Istio is an open source service mesh layered transparently onto existing distributed applications. It is also a platform that has APIs for connecting log record platform, telemetry, or policy system. Istio provides a uniform and more efficient way to run the distributed microservice architecture and secure, connect, and monitor microservices.
- Earlier, the data plane SideCar proxy is unstable and traffic-intensive as Service Mesh puts too many functions, including the inter-service communications and related governance into it. As a solution to these problems, the second-generation Service Mesh emerges and separates the configuration policy and decision logic from the proxy servers to form an independent control plane.
- Istio has two components: the data plane and the control plane.
 - The data plane is the communication between services. Without a service mesh, the network cannot figure out the type, source, and destination of the traffic.
 - The control plane takes your desired configuration, and its view of the services, and dynamically programs the proxy servers, updating them as the rules or environments change.

Istio Key Features

- Istio has two planes: The data plane, for deploying the Sidecar proxy to enable Istio. The Sidecar proxy runs alongside microservices to route requests to or from other proxies in a mesh that intercepts communications requests between microservices. The control plane, for managing and configuring proxies to route traffic, and collecting telemetry data via policy components.



18 Huawei Confidential



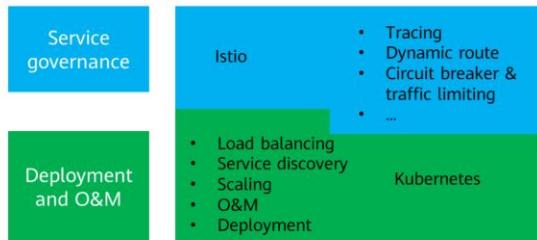
Features

- Traffic management: traffic routing rules control traffic and API calls between services, simplifying configuration of service-level properties.
- Observability: application-level monitoring provides observability of service behaviors for application troubleshooting, maintenance, and optimization.
- Security: identity authentication, authorization, and TLS encryption for large-scale service communication.

- Istio service mesh has two components: the data plane (Envoy) and the control plane (Istiod).
 - Envoy is a high-performance proxy developed in C++ to mediate all inbound and outbound traffic for all services in the service mesh. Envoy proxies are deployed as sidecars to services and are the only Istio components that interact with data plane traffic. In addition to load balancing, circuit breakers, and fault injection, Envoy also supports a pluggable extension model built on WebAssembly (Wasm) that allows for custom policy enforcement and telemetry generation for mesh traffic.
 - Istiod is a control plane component that provides service discovery, configuration, and certificate management. Istiod converts advanced rules written in YAML into Envoy-specific configurations and propagates them to the sidecars. Pilot abstracts platform-specific service discovery mechanisms and synthesizes them into a standard format that sidecars can consume. Citadel enables strong service-to-service and end-user authentication with built-in identity and credential management. You can also use Istio's authorization feature to control who can access your services.
- As a core component of Istio, Pilot manages and configures all sidecar proxies deployed in a specific Istio service mesh. As a component responsible for configuration management, Galley verifies the format and content of the configuration information and provides it for the Pilot on the control plane. Citadel consists of the CA server, security discovery server, and certificate key controller.
- Core concepts of Istio:
 - Data plane components are injected as non-intrusive sidecars into service containers, with transparent traffic hijacking.
 - Upper-level APIs are implemented based on Kubernetes CRDs, fully declarative and standardized.
 - The data plane and control plane communicate with each other through standard protocols, allowing pub/sub messaging.

Istio + Kubernetes

- Istio is a complete solution that provides behavior insights and operation control for the entire service mesh for any microservice requirement.
- Kubernetes provides deployment, upgrade, and limited traffic management, but not circuit breaker, rate limiting, service degradation, and trace governance. Istio is an open platform built to complement Kubernetes in microservice governance.



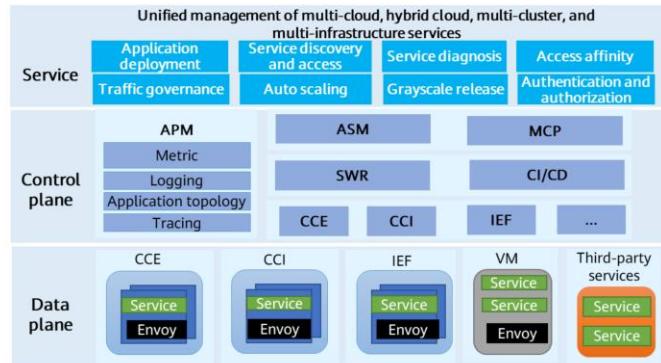
- Istio extends Kubernetes to establish a programmable, application-aware network using the powerful Envoy service proxy. Working with both Kubernetes and traditional workloads, Istio simplifies deployment with standard, universal traffic management, telemetry, and security.
- Istio aims to achieve scalability and meet various deployment requirements. Istio control plane runs on Kubernetes. In this way, applications deployed in a cluster can be added to your mesh. In addition, the mesh can be extended to other clusters, and even connected with VMs or other endpoints running outside Kubernetes.
- To enable Istio, you only need to deploy a special sidecar proxy in the environment and use the Istio control plane to configure and manage the proxy to intercept all network communication between microservices. You can use Istio to achieve:
 - Automatic load balancing for HTTP, gRPC, WebSocket, and TCP traffic
 - Fine-grained control of traffic behavior with rich routing rules, retries, failovers, and fault injection
 - A pluggable policy layer and configuration API supporting access control, rate limits and quotas
 - Automatic metrics, logs, and traces for all traffic within a cluster, including cluster ingress and egress
 - Secure service-to-service communication in a cluster with strong identity-based authentication and authorization
- Google Remote Procedure Call (gRPC) is a high-performance open source RPC software framework built on the HTTP 2.0 transport layer protocol. It provides an API design method for managing and configuring network devices. gRPC supports multiple programming languages, such as C, Java, Golong, and Python.

Contents

1. Cloud Native Applications and Microservices
2. Cloud Native Application Frameworks
3. **Huawei Cloud Native Application Solutions**
 - Application Service Mesh (ASM)
 - Application Middleware
 - DevCloud
 - ServiceStage

Application Service Mesh (ASM)

- A fully-managed service mesh with high performance and reliability. Easy to use and supports multiple infrastructures (VMs and containers) and unified governance of multi-cluster services across regions. Enables you to manage and monitor service traffic, ensure secure access, and grayscale release for stable service iteration based on infrastructure.
- Fully compatible with Istio (control and data planes), and interconnected with CCE (a hosted Kubernetes service on Huawei Cloud).



21 Huawei Confidential



- ASM supports smooth access and unified governance of multiple applications, such as containers, traditional microservices, and third-party services. It enables hybrid management of cross-cluster traffic under various network conditions in multi-cloud and hybrid cloud scenarios. Large-scale meshes are provided for intelligent O&M and scaling to help you automatically and transparently manage application access.
- ASM provides a high-performance, low-loss, lightweight, and multi-form mesh data plane and supports uninstallation by pod and node, accelerating sidecar forwarding. Flexible topology learning optimizes configurations and resources on the mesh control plane.
- ASM can well resolve application network governance issues such as challenges in cloud native application management, network connection, and security management.
- ASM is deeply integrated with CCE to manage application traffic and lifecycle in a non-intrusive manner. ASM enhances the full-stack capabilities of Huawei Cloud container services with better usability, reliability, and visualization.

ASM Benefits

- Important part of the cloud native ecosystem. A complete non-intrusive microservice governance solution for cloud native application management, network connection, and security management.

Connection



Security



- Intelligently controls service traffic and cross-service API calls.
- Upgrades services through grayscale release and continuous testing.

Control



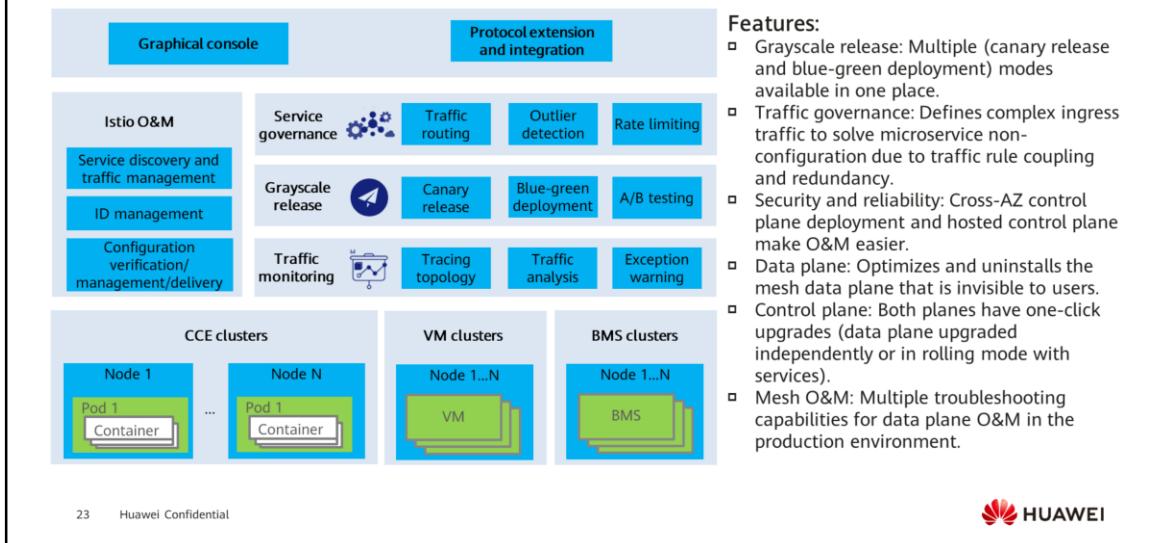
Telemetering



- Manages authentication, authorization, and encryption for communication between services.

- Delivers and executes traffic control policies for fair customer allocation of network link resources.
- Provides automatic link tracing and log monitoring for services to visualize their status.

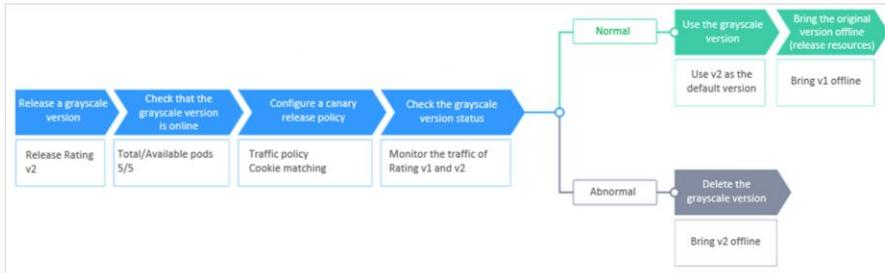
ASM Architecture



- Hybrid deployment: Unified governance of hybrid deployment of VM applications and containerized applications
- Observability: Out-of-the-box usability and end-to-end intelligent monitoring, logs, topologies, and tracing
- Unified service governance in the multi-cloud and hybrid cloud scenarios, unified service governance of multiple infrastructure resources (multi-container cluster/container-VM/VM-PM), and cross-cluster grayscale release, topology, and tracing
- Protocol extension: Solution of integrating with microservice SDKs for Spring Cloud
- Community and open source: No. 3 in the world by contribution to Istio community; quick response to community version issues and requirements

ASM Grayscale Release

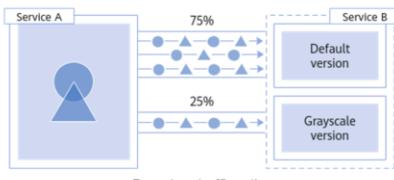
- Also known as canary release. It deploys both old and new versions of an application in the environment. Service requests are sent to either version. When releasing a new version, configure traffic policies to gradually bring the new version online, minimizing risks and fault impacts. Supports quick rollback.



- Grayscale release policies:
 - Grayscale policies based on request content: You can set criteria based on request content, such as header and cookie. Only requests meeting the criteria will be distributed to the grayscale version.
 - Grayscale policies based on traffic ratio: You can set specific ratio for the traffic to be distributed to the grayscale version.
 - Canary release: Guidance will be provided to help you perform canary release on a service, including rolling out a grayscale version, observing the running and traffic of the grayscale version, configuring grayscale release policies, and diverging the traffic.
 - Blue-green deployment: Guidance will be provided to help you perform blue-green deployment on a service, including rolling out a grayscale version, observing the running and traffic of the grayscale version, and switching the traffic.

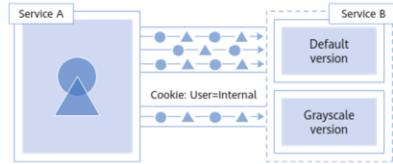
Grayscale Policy

- Grayscale policy: Before releasing a new service version in the production environment and letting it serve all the live traffic, you can add a grayscale version and configure grayscale policies to serve just a proportion of the traffic. After the grayscale version has run stably for a period, it can serve as the default version to take over all traffic in place of the original version in the production environment.
- Grayscale version: Only one grayscale version can be released for a service. You can configure grayscale policies for the version.



Based on traffic ratio

- You can set the traffic ratio for the original version and grayscale version. The system distributes traffic to the two versions based on the specific traffic ratio.
- For example, 75% of the traffic is directed to the original version, and 25% is directed to the grayscale version.



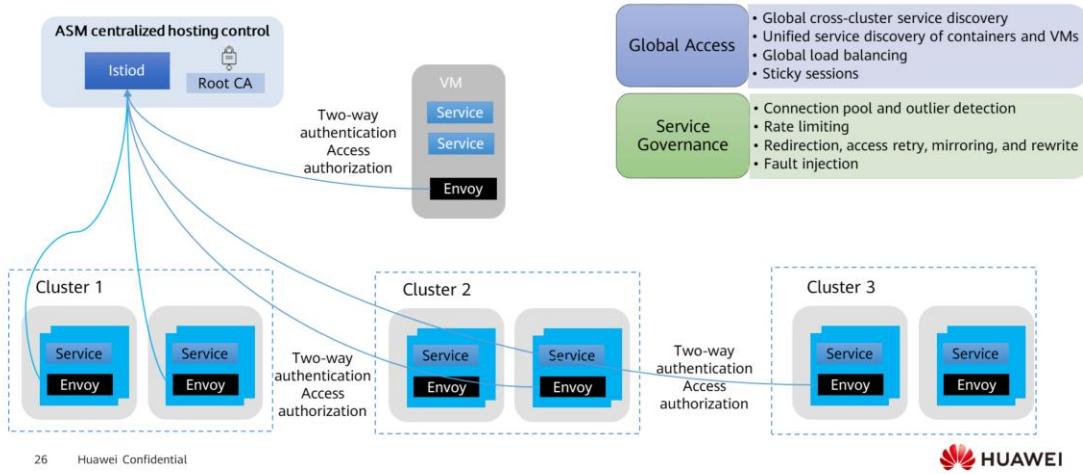
Based on request content

- The grayscale version can be accessed only when the traffic meets the rules based on the cookies, custom headers, queries, operating systems, and browsers.
- For example, only HTTP requests whose cookies meet User=internal can be forwarded to the grayscale version. Other requests are still received by the original version.



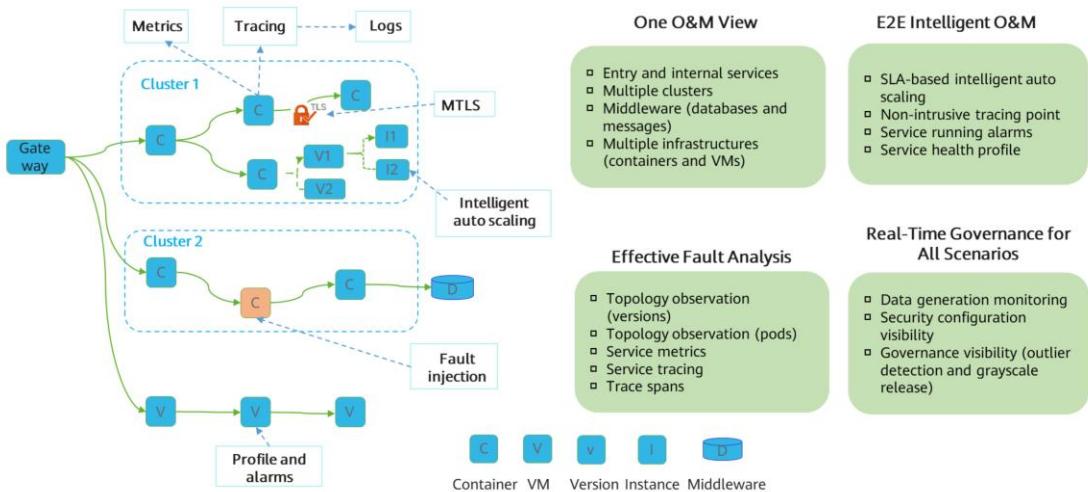
Infrastructure Discovery and Multi-Cluster Management

- ASM supports global service discovery and governance across clusters and infrastructure.



- An O&M-free hosting control plane is provided. Unified service governance, grayscale release, security, and service running monitoring capabilities for multiple clouds and clusters are supported. Unified service discovery and management of multiple infrastructure resources such as containers and VMs are provided.
- The meshes of multiple clusters share a set of root certificates. They distribute keys and certificate pairs to service pods in the data plane, and periodically change key certificates. Key certificates can be revoked as required. When a service calls another service, the mesh data plane envoy performs two-way authentication and channel encryption. These two services can come from two different clusters. Transparent end-to-end two-way authentication across clusters is supported.
- Load balancing, service routing, fault injection, outlier detection, and fault tolerance policies can be intuitively configured using an application topology. Microservice traffic management can be real-time, visualized, intelligent, and automated, requiring no modifications on your applications.
 - Routing rules based on weight, content, and TCP/IP implements flexible grayscale release of applications.
 - HTTP sticky session achieves service processing continuity.
 - Rate limiting and outlier detection ensure stable and reliable links between services.
 - Network persistent connection management saves resources and improves network throughput.
 - Service security certification, authentication, and audit lay a solid foundation for service security assurance.

Unified Governance Policies and Real-Time Traffic Monitoring



27 Huawei Confidential



- Load balancing, service routing, fault injection, outlier detection, and fault tolerance policies can be intuitively configured using an application topology. Microservice traffic management can be real-time, visualized, intelligent, and automated, requiring no modifications on your applications.
 - Routing rules based on weight, content, and TCP/IP implements flexible grayscale release of applications.
 - HTTP sticky session achieves service processing continuity.
 - Rate limiting and outlier detection ensure stable and reliable links between services.
 - Network persistent connection management saves resources and improves network throughput.
 - Service security certification, authentication, and audit lay a solid foundation for service security assurance.
- Requests can be distributed based on the request content (browsers or OSs).
- Requests can be distributed based on traffic ratio.

Application Scenarios



Service traffic governance

Istio non-intrusive traffic governance. Policy- and scenario-based network connection management suits different service protocols. Different governance rules for APIs on the topology meet different service requirements.

No code refactoring is required when managing traffic.

Service monitoring

Detailed telemetry for all service communications within the mesh. Observability of service behaviors enables application troubleshooting, maintenance, and optimization. With ASM, operators can better understand how services interact with other services and their components.

Grayscale release

Built-in, flexible, and automated release.

Multiple functions for application governance to detect and fix issues at the early stage and ensure smooth iteration.

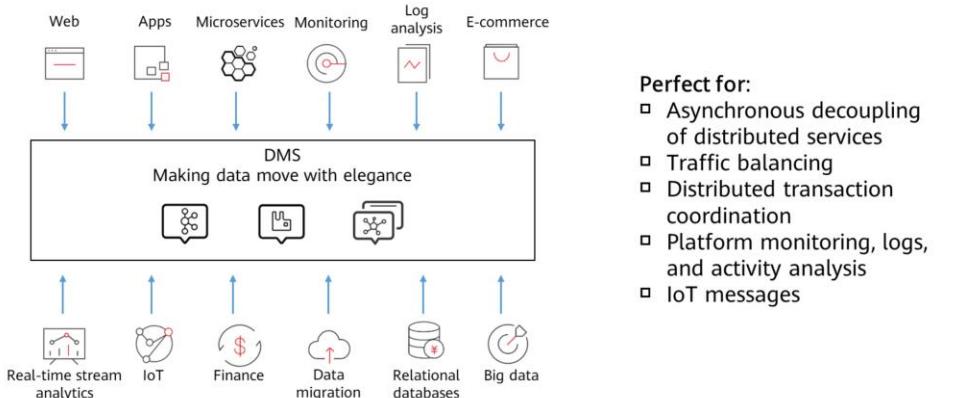
- Container-based infrastructure brings a series of new challenges. It is necessary to evaluate and enhance the performance of API endpoints and identify potential risks of infrastructure. ASM enables you to enhance API performance with no code refactoring and service delay.
- In traditional iterations, a new service version is directly released to all users at a time. This is risky, because once an online accident or a bug occurs, the impact on users is great. It could take a long time to fix the issue. Sometimes, the version has to be rolled back, which severely affects user experience. Grayscale release is a smooth iteration mode for version upgrade. During the upgrade, some users use the new version, while other users continue to use the old version. After the new version is stable and ready, it gradually takes over all the live traffic.

Contents

1. Cloud Native Applications and Microservices
2. Cloud Native Application Frameworks
3. Cloud Native Solutions from Huawei Cloud
 - ASM
 - Application Middleware
 - DevCloud
 - ServiceStage

Distributed Message Service (DMS)

- Provides system decoupling, cross-system and cross-region data transmission, and distributed transaction coordination for distributed architectures. DMS is compatible with open-source Kafka, RabbitMQ, and RocketMQ.



Perfect for:

- Asynchronous decoupling of distributed services
- Traffic balancing
- Distributed transaction coordination
- Platform monitoring, logs, and activity analysis
- IoT messages

30 Huawei Confidential

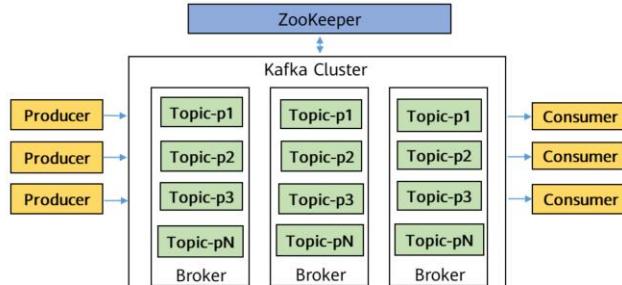


- Main features:

- Ease of use: Instances created in minutes; out of the box with visual operations and real-time monitoring
- Reliability: Cross-AZ deployment, automatic fault detection, alarms, and failover; fixes for open-source availability issues (split brains or multiple controllers)
- Proven success: Widely deployed in customer cloud; major e-commerce events (VMALL 11.11 Shopping Festival); open-source community links; customer trusted choice

DMS for Kafka

- A message queuing service using open-source Apache Kafka. It provides Kafka instances with isolated computing, storage, and bandwidth resources.
- Kafka is distributed pub/sub messaging middleware with high throughput, data persistence, horizontal scalability, and stream data processing. Common applications include log collection, data streaming, online/offline system analytics, and real-time monitoring.



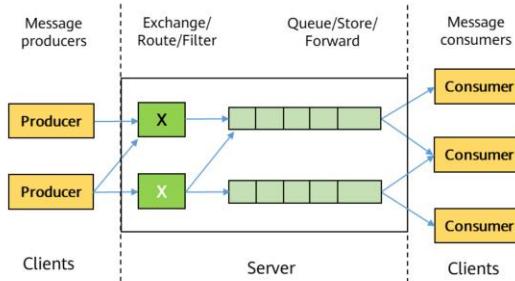
31 Huawei Confidential



- ZooKeeper: a distributed coordination application that stores Kafka metadata.
- Clients:
 - Producer: a client application that continuously publishes messages to one or more topics.
 - Consumer: a client that subscribes to one or more topics.
- Server: consists of service processes called brokers. A Kafka cluster consists of multiple brokers.
- Kafka: distributed message stream processing middleware.
- Broker: receives and processes requests from clients and persists messages.
- Topic: a publish/subscription object in Kafka. Dedicated topics can be created for each service, application, or even each category of data. Topics are divided into partitions.
- High availability mechanism of Kafka:
 - Different brokers run on different machines. If one broker is down, other brokers can still provide services for external systems.
 - The same data is replicated to multiple machines.

DMS for RabbitMQ

- Built on open-source RabbitMQ and provides rich messaging with flexible routing, high availability, monitoring, and alarm functions. Applications include flash sales, flow control, and system decoupling.



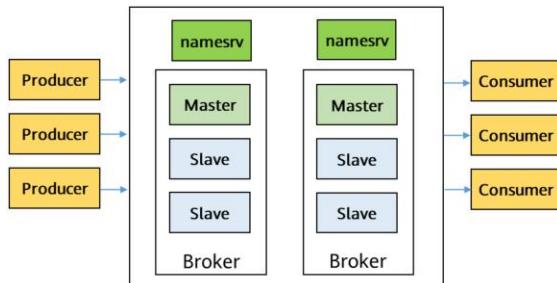
32 Huawei Confidential



- Immediate use: DMS for RabbitMQ provides single-node and cluster instances with a range of specifications for you to choose from. Instances can be created with just a few clicks on the console, without requiring you to prepare servers.
- Rich features: DMS for RabbitMQ supports Advanced Message Queuing Protocol (AMQP) and a variety of messaging features such as message broadcast, delayed delivery, and dead letter queues.
- Flexible routing: In RabbitMQ, an exchange receives messages from producers and pushes the messages to queues. RabbitMQ provides direct, topic, headers, and fanout exchanges. You can also bind and customize exchanges.
- High availability: In a RabbitMQ cluster, data is replicated to all nodes through mirrored queues, preventing service interruption and data loss in case of a node breakdown.
- Monitoring and alarm: RabbitMQ cluster metrics are monitored and reported, including broker memory, CPU usage, and network flow. If an exception is detected, an alarm will be triggered.
- AMQP is an advanced message queue protocol at the application layer of the unified messaging service. It is an open standard application layer protocol for message-oriented middleware. A client and message middleware developed based on this protocol can exchange messages without product or programming language barriers.

DMS for RocketMQ

- Message-oriented middleware with low latency, high flexibility, high throughput, dynamic expansion, easy management, and abundant messaging functions. Compatible with the open-source RocketMQ client, it provides functions (ordered message delivery, intentional delivery delay, message retry, dead letter messages, transactional messages, and session messages) for adaptability to various service scenarios such as e-commerce and finance.



33 Huawei Confidential



- Supported message types:
 - Normal messages: Messages that do not have any features of delayed messages, ordered messages, or transactional messages.
 - Delayed/Scheduled messages: Messages that are delivered to consumers after a specific period after being sent from producers to DMS for RocketMQ.
 - Ordered messages: Messages that are retrieved in the exact order that they are created.
 - Transactional messages: Messages that achieve eventual consistency, delivering distributed transaction processing similar to X/Open XA.
- Producer: a program that delivers messages.
- Consumer: a program that receives messages.
- namesrv: stores topic routing information. Clients must access namesrv to obtain topic routing information before production and consumption.
- Master: receives production and consumption requests from clients.
- Slave: functions as a replica node and receives replicated data from master.
- Raft consensus algorithm ensures data consistency between the master and slave nodes. Automatic failover is performed between these nodes in the same group.
- Broker: receives and processes client requests and persists messages. The three nodes in a broker work in master/slave mode.

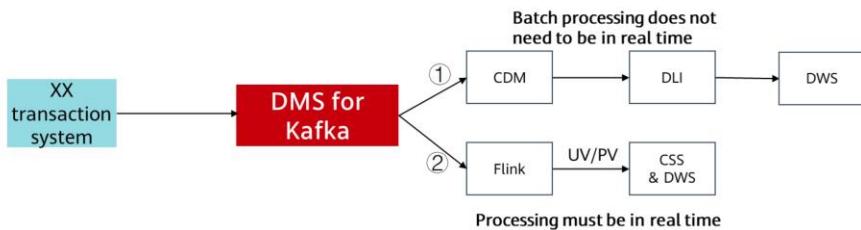
Comparing DMS for Kafka, RabbitMQ, and RocketMQ

Feature	Kafka	RabbitMQ	RocketMQ
Priority queue	Not supported	Supported	Not supported
Delayed queue	Not supported	Supported	Supported
Message retry	Not supported	Not supported	Supported
Retrieval mode	Pull-based	Pull-based and push-based	Pull-based and push-based
Message tracking	Supported	Not supported. Acknowledged message retrieval notifies RabbitMQ to delete the message	Supported
Persistence	Supported	Supported	Supported
Message tracing	Not supported	Supported	Supported
Message filtering	Supported	Not supported, but can be encapsulated	Supported
Multi-tenancy	Not supported	Supported	Supported
Multi-protocol	Only supports Apache Kafka	RabbitMQ is based on AMQP and supports MQTT and STOMP	Compatible with RocketMQ
Throttling	Supports throttling on producer or consumer clients	Supports credit-based throttling on producers (internal protection mechanism)	Supported
Transactional messages	Supported	Supported	Supported

- RabbitMQ supports persistence with the firehose feature or the rabbitmq_tracing plugin. However, rabbitmq_tracing reduces performance and should be used only for troubleshooting.
- The performance of message-oriented middleware is measured by throughput. While RabbitMQ provides tens of thousands of QPS, Kafka provides millions. However, if idempotency and transactions are enabled for Kafka, its performance will be compromised.

Case: Using Kafka to Build a Real-Time Transaction Analysis Platform

- Kafka is popular message-oriented middleware for highly reliable, asynchronous message delivery. It is widely used for transmitting data across systems in the enterprise application, payment, telecommunications, e-commerce, social networking, instant messaging, video, IoT, and IoV industries.

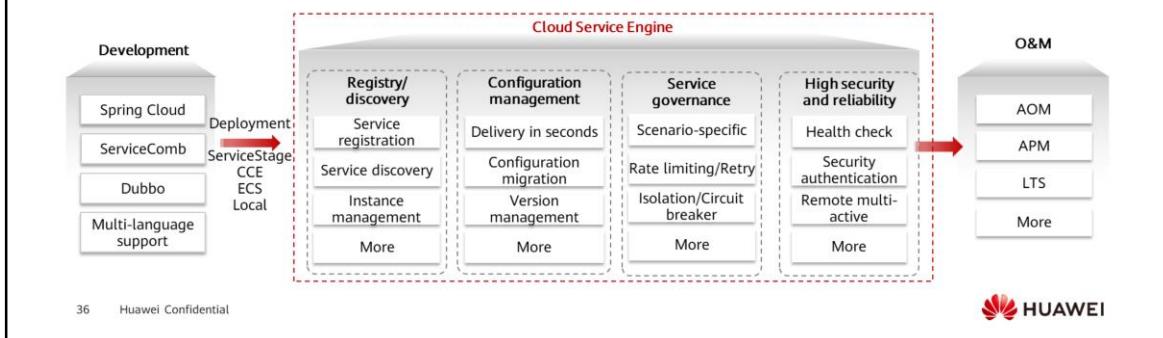


A top e-commerce player uses DMS to assist in decision-making.

- (1) Batch processing
- (2) Real-time transactions and user data tracking for analysis

Cloud Service Engine (CSE)

- This cloud middleware for microservice applications gives you powerful, resilient enterprise-class cloud service capabilities, such as registry and discovery, service governance, and configuration management. CSE is seamlessly compatible with open-source ecosystems such as Spring Cloud and ServiceComb. Combine CSE with other cloud services into a cloud-native microservice system.
- CSE is a one-stop management platform for microservice solutions. It enables developers to focus on service development and improve product efficiency and quality.

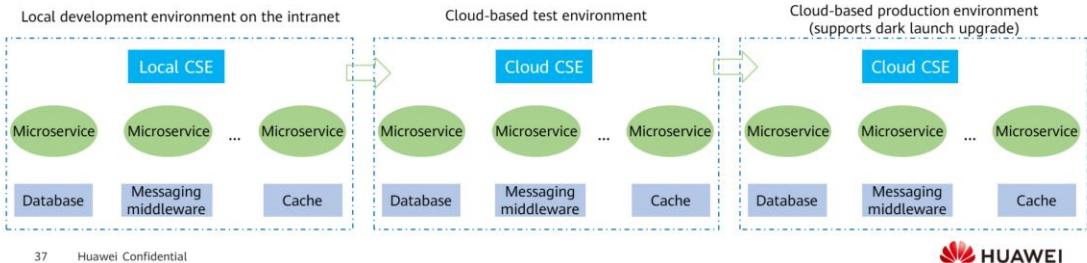


36 Huawei Confidential

- The microservice architecture includes remote procedure call (RPC) communication between microservices, distributed microservice instances and service discovery, external and dynamic configurations, centralized configuration management, microservice governance capabilities (such as circuit breaker, isolation, and load balancing), tracing, and log collection and retrieval.
- The microservice architecture consists of the following:
 - RPC communication between microservices. Using RPC for communication reduces coupling between microservices and makes the system more open with fewer technological restrictions.
 - Distributed microservice instances and service discovery. The microservice architecture focuses on resilience and the microservice design is generally stateless. Increasing stateless microservice instances lets you improve processing performance. When there are a large number of instances, a middleware that supports service registry and discovery is required for microservice calling and addressing.
 - Dynamic and centralized configuration management. Configuration management is increasingly complex as the number of microservices and instances increases. The configuration management middleware provides a unified view for all microservices, simplifying their configuration management. These governance capabilities can mitigate the impact of some common faults of the microservice architecture on the services.
 - Tracing and centralized log collection and retrieval. Viewing logs remains the most commonly used method for analyzing system faults. Tracing information helps locate faults and analyze performance bottlenecks.

Application Scenarios

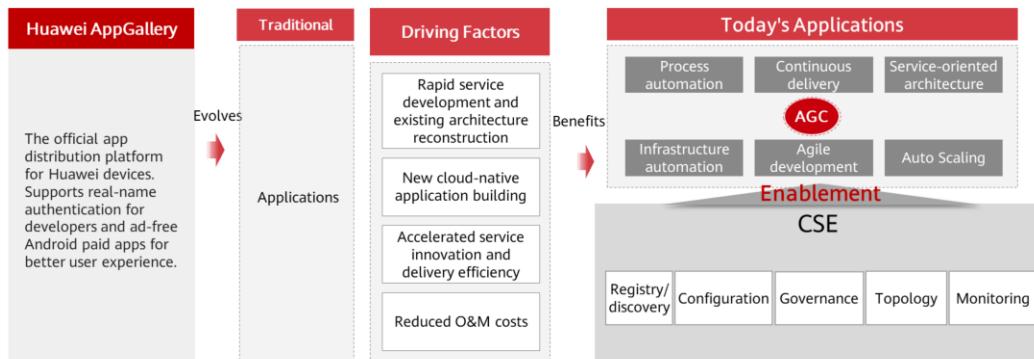
- Host Spring Cloud applications to replace open-source components with highly reliable commercial middleware for better application management and maintenance, while minimizing the impact on services.
 - Application systems developed on Spring Boot do not have basic microservice capabilities, so they integrate Spring Cloud Huawei for service registration/discovery and dynamic configuration management.
 - Application systems developed on Spring Cloud open-source technology systems (Eureka for registration/discovery and Nacos for dynamic configuration) integrate Spring Cloud Huawei and replace open-source middleware with highly reliable commercial middleware for lower maintenance costs.
 - Cloud native applications built on other Spring Cloud development systems (Spring Cloud Alibaba and Spring Cloud Azure) use Spring Cloud Huawei to migrate to Huawei Cloud.



- The purpose of planning the development environment is to ensure that developers can better work in parallel, reduce dependencies, reduce the workload of environment setup, and reduce the risks of bringing the production environment online.
 - Set up a local development environment on the intranet. The advantage of the local development environment is that each service domain or developer can set up a minimum function set environment that meets their requirements to facilitate log viewing and code debugging. The disadvantage of local development environment is the low integration. When the integration and joint commissioning are required, it is difficult to ensure environment stability.
 - The cloud-based test environment is a relatively stable integration test environment. After the local development and test are complete, each service domain deploys their own services in the cloud test environment and can invoke services in other domains for integration tests. These test environments are integrated in ascending order.
 - The production environment is a formal service environment. It needs to support dark launch upgrades, online joint commissioning, and traffic diversion to minimize the impact of upgrade faults on services.
 - In the cloud-based test environment, the public IP addresses of CSE and middleware can be opened, or network interconnection can be implemented. In this way, the middleware on the cloud can be used to replace the local environment, reducing the time for developers to install the environment.

Case Study: The Foundation of Huawei Mobile Cloud Services

- 10,000+ microservice instances run on Huawei Device services with hundreds of millions of users



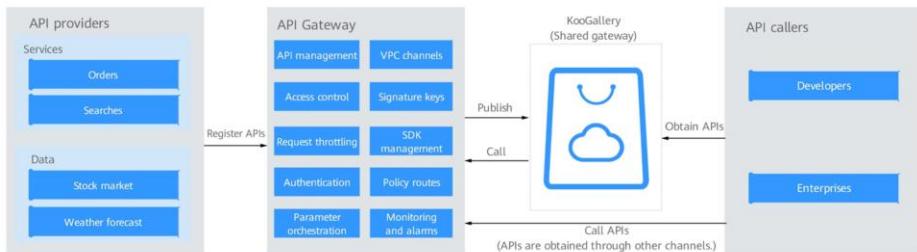
38 Huawei Confidential



- Microservices and components provide a technical basis for large-scale collaborative development and a unified framework for internal sharing. By the beginning of 2021, AppGallery already had more than 300 microservices available, with more than 10,000 instances deployed on the live network. More than 500 dynamic layout cards have been developed on the client, and more than 100 components have been built.
- AppGallery Connect: provides developers with full-lifecycle mobile app services, covering all devices and scenarios, reducing development costs, improving operation efficiency, and facilitating business success.

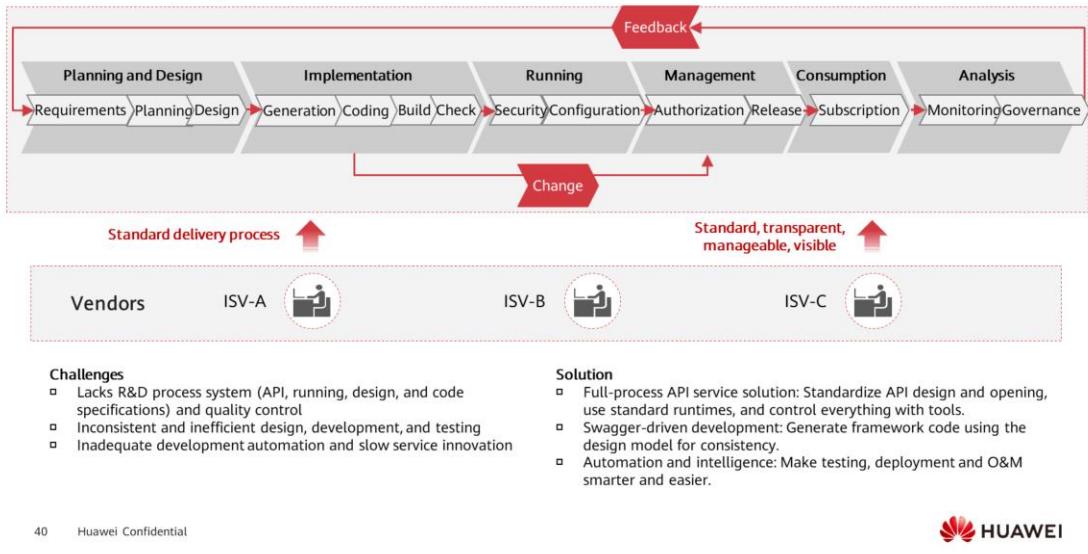
APIG

- API Gateway (APIG) is your fully managed API hosting service. Build, manage, and deploy APIs at any scale to integrate your internal systems, open your enterprise capabilities to partners, and monetize at minimal cost and risk.
- APIG uses standard RESTful APIs to simplify service architecture, decouple internal systems, and separate frontend from backend.



- You can open your services and data by directly providing open APIs to API callers or releasing them on KooGallery for monetization.
- You can also obtain and call open APIs from APIG to reduce your development time and costs.
- By using APIG, you can monetize services while reducing R&D investment for more business focus and higher operational efficiency. For example, enterprise A has created a mobile number location lookup API in APIG and released it on KooGallery. Enterprise B obtains and calls the API from KooGallery and pays for the fee incurred. In this way, enterprise A monetizes its services and enterprise B reduces its development time and costs, achieving shared success.

High-Quality, Efficient Control with E2E API R&D



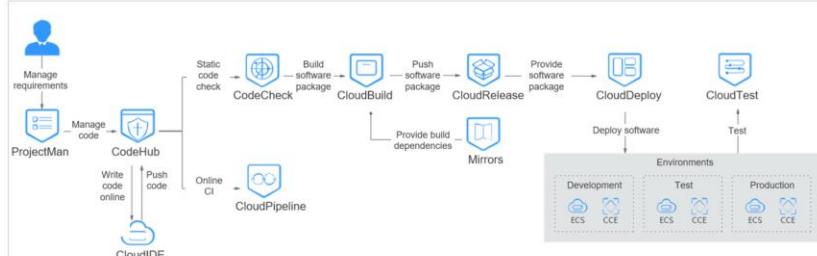
- Swagger is a standard, complete framework for generating, describing, invoking, and visualizing RESTful web services. It aims to define standard, language-independent RESTful APIs. It enables people and computers to discover and understand services without accessing source code or documentation or monitoring network traffic.

Contents

1. Cloud Native Applications and Microservices
2. Cloud Native Application Frameworks
3. **Huawei Cloud Native Application Solutions**
 - ASM
 - Application Middleware
 - DevCloud
 - ServiceStage

DevCloud

- Huawei Cloud is a one-stop DevSecOps platform for developers. It streamlines the entire software delivery process – from requirement breakdown, code commit and build, testing and verification, to deployment and O&M.
- DevSecOps stands for Development, Security, and Operations. It is an approach to culture, automation, and platform design that integrates security as a shared responsibility throughout the entire IT lifecycle. You can use the built-in CI/CD capabilities of Huawei Cloud DevCloud to continuously deliver value.



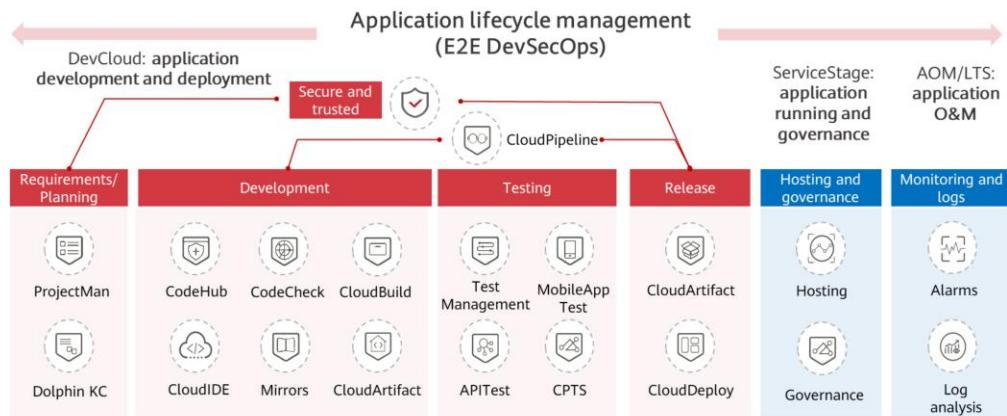
42 Huawei Confidential



- DevCloud consists of the following services:
 - ProjectMan: provides agile project management and collaboration, supports management of sprints, milestones, and requirements across projects, tracks bugs, and provides multi-dimensional statistics reports.
 - CodeHub: a Git-based online code hosting service for software developers. It is also a code repository for security management, member and permission management, branch protection and merging, online editing, and statistics. The service addresses issues such as cross-region collaboration, multi-branch concurrent development, and code version management.
 - CloudPipeline: provides visualized, customizable pipelines to shorten the delivery period and improve efficiency.
 - CodeCheck: manages code quality in the cloud. You can easily perform static checks and security checks on code in multiple programming languages and obtain comprehensive quality reports. CodeCheck also allows you to view grouped defects with fix suggestions provided, effectively controlling quality.
 - CloudBuild: provides an easy-to-use hybrid language build platform to implement cloud-based build, and supports continuous and efficient delivery. With CloudBuild, you can create, configure, and execute build tasks with a few clicks to obtain, build, and package code automatically and monitor build status in real time.
 - CloudDeploy: provides visualized, one-click deployment. It supports deployment on VMs or containers by using Tomcat, Spring Boot, and other templates or by flexibly orchestrating atomic actions. It also supports parallel deployment and seamless integration with CloudPipeline, providing standard deployment environments and implementing automatic

deployment.

Application Full-lifecycle Management



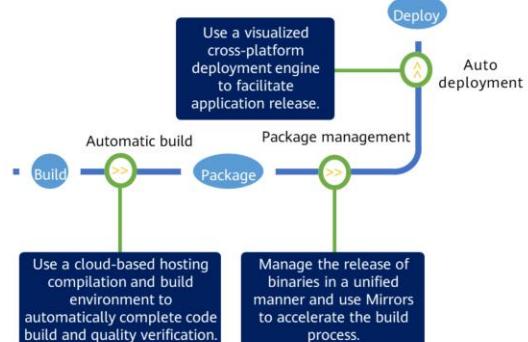
44 Huawei Confidential



- E2E process: One platform covers common functions in software development. These functions are embedded and integrated for governance and O&M.
- Over 20 mainstream programming languages, development frameworks, and running environments, for seamless application migration.
- Secure and trustworthy: DevCloud provides security testing, trustworthiness building, high security standards, and 7,000+ code check rules.

CI/CD Full-process Implementation

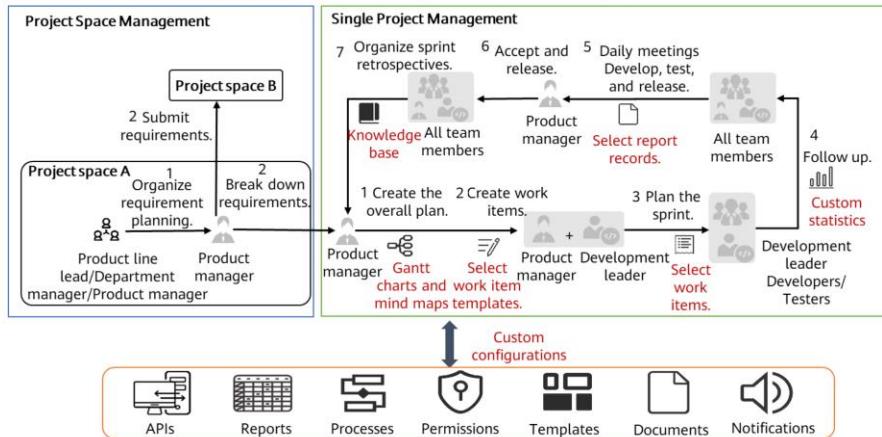
- In software engineering, CI/CD or CICD is the combined practices of continuous integration (CI) and (more often) continuous delivery or (less often) continuous deployment (CD). CI/CD bridges gaps between development and operation activities and teams by enforcing automation when building, testing and deploying of applications.
- You can use the built-in CI/CD capabilities of Huawei Cloud DevCloud to continuously deliver value.
 - Continuous integration (CI): Developers can commit content to repositories multiple times, at any time they want. There is no need to developing each functional module separately and then wait till the end of a development cycle to commit their code.
 - Continuous delivery (CD): Verified code can be released to a repository such as GitLab. O&M teams can quickly obtain code from the repository and deploy applications in the related environment.
 - Continuous deployment (CD): In addition to continuous delivery, software builds are automatically deployed after all tests are passed. Such automatic deployments can be configured to quickly install patches or distribute components or functional modules.



- Evolving from waterfall, agile, to DevOps is the technical route for modern developers to build excellent products. New CI/CD methods rise with DevOps. Traditional software development and delivery methods are rapidly becoming outdated. In the agile era, most companies released software every month, every quarter, or even every year. In the DevOps era, it is normal that software is released every week, every day, or even multiple times a day. This is especially true when SaaS becomes popular in the industry. Applications can be updated dynamically without forcing users to download updated components. Many times, users are not even aware of changes.
- CI focuses on integrating the work of each developer into a code repository, which is performed several times a day. The main purpose is to detect integration errors as early as possible, so that the team can collaborate better. Continuous delivery (CD) aims to minimize the inherent team friction during deployment or release. It automates each build and deployment step so that code release can be securely completed at any time (ideally). Continuous deployment (CD) is more automated. Whenever the code is greatly changed, the build/deployment is automatically performed.

ProjectMan

- ProjectMan provides simple and efficient team collaboration services, including multi-project management, Scrum projects, requirement management, defect tracking, and statistical analysis.



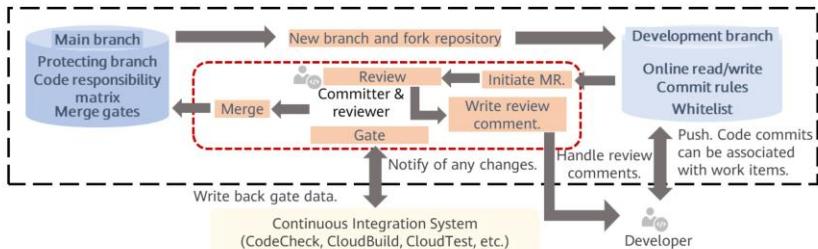
46 Huawei Confidential



- A project consists of a series of coordinated and controlled activities in a certain process. The objective of a project is to meet specific requirements and is restricted by time and resources. Project management covers the project process and results to achieve project objectives. Kanban project is a special type. Kanban depends on projects. It displays work items, their levels, and their types.
- Professional agile project management: agile-based project set management, single-project Scrum, and lean Kanban
- Professional product planning: Gantt charts, mind maps, and overall product plans
- Multi-dimensional and professional reports: multi-project Kanban, dashboard, and reports
- R&D knowledge management: Structured knowledge and accumulated innovations.
- Trusted audit logs: 1000+ audit events, comprehensive tracing, and high security and reliability
- Typical scenario:
 - Collaborative operations of product, development, and test personnel
 - Requirement management
 - Project health (progress, quality, risk, and personnel) management
 - Defect management

CodeHub

- Huawei's CodeHub extends and hardens Git kernel capabilities a lot and uses a loosely coupled architecture to solve performance and capacity problems during large-scale development.
- CodeHub is based on Huawei's years of experience in security and trustworthiness technologies. It provides a solid protection chain that makes your core code more secure, trusted, and tamper-proof.
- It stores your code and code repository templates in multiple programming languages, such as Java, C, C++, C#, PHP, Python, Go, and Node.js. It also allows you to share private repositories as templates with other users.



47 Huawei Confidential



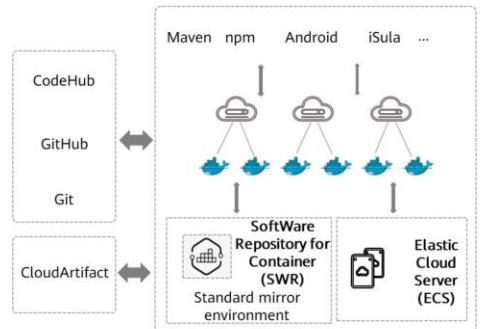
- Access security control: CodeHub provide authentication tools such as branch protection and IP address whitelist to ensure that only accounts with specific permissions and IP addresses can access code repositories.
- Remote backup: Authorized users can back up repositories to other regions, physical hosts, and cloud hosts on Huawei Cloud by one click.
- Repository locking: You can manually lock a repository to disable any changes or commits, preventing the stable version to be released from being compromised.
- SSH deployment key: Use the SSH key to control read and write permissions of a repository. Use the deployment key to enable the read-only permission of a repository.
- Misoperation tracing and recovery: Code and branches that are deleted by mistake can be accurately rolled back or retrieved. For deleted repositories, backups are kept in the physical storage for a specific retention period.
- Operation logs: All operations have tokens. Key operations are audited and recorded.
- Rule setting: CodeHub allows you to configure commit rules, merge requests, and gates to ensure that the code quality is controllable.
- Notification setting: When an important change occurs in a repository, a notification such as an email or SMS message can be sent to the preset role.

CodeCheck and CloudBuild

- Driven by data, CodeCheck provides Huawei in-house platform for out-of-the-box, highly scalable code checks integrated with development processes.
- CodeCheck allows developers to use intelligent technologies to identify potential defects in code in a timely and accurate manner and fix defects automatically.
- With CloudBuild, developers can create, configure, and execute build tasks with one click and obtain, build, and package code automatically.
- The secure and trustworthy CloudBuild service implements build environment traceability and build tool chain selection.



48 Huawei Confidential

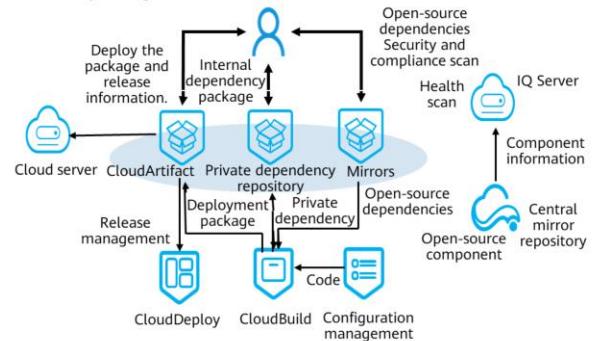
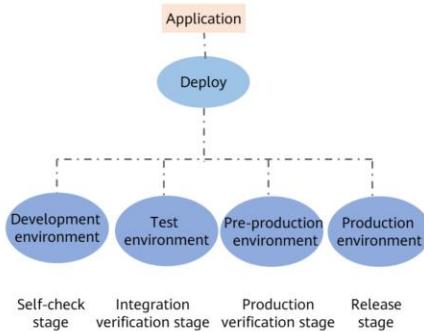


HUAWEI

- In-house development
 - Huawei-developed cross-process code check engine based on the syntax tree and context-free grammar (CFG), supporting code check in 10 languages, such as C, C++, Java, Python, and Go.
- High-quality code check rule set based on Huawei's 30-year R&D experience
 - 3000+ code check rules and 20+ scenarios, covering programming styles/coding security/memory management/input verification/unsafe functions/thread synchronization/code repetition rate.
 - Compatible with more than 5 secure coding standards, such as CWE/OWASP TOP 10/SANS TOP 25/MISRA/CERT.
- Automatic auxiliary defect fixing
 - CodeHub offers intelligent fix suggestions and fixes defects by automatic code changes, improving fix efficiency.
 - Provides Java and C/C++ programming guidelines for defect fixing. Provides automatic fixing of Go code.

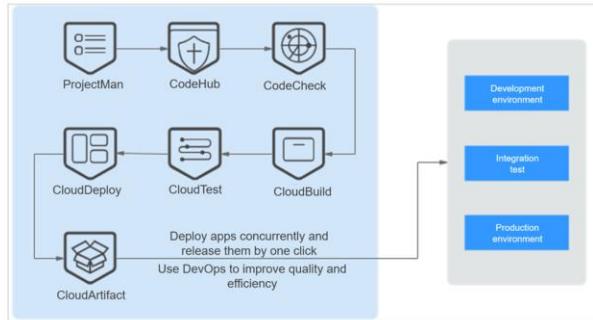
CloudDeploy and CloudArtifact

- CloudDeploy allows you to deploy applications in multiple environments, such as physical machines, virtual machines (VMs), and containers. You can use various parameters and methods to deploy the same application in different environments.
- CloudArtifact manages software packages generated during development. It is an important link between CI and CD.
- CloudArtifact also requires operations such as release review, tracing, and security control of software packages.



Scenario: Software and Solution Operation Enterprises

- Challenge: Communication is difficult for developers working in different locations and using different tools and environments. Customer requirements change fast, which requires fast responses and often causes rework. Enterprises urgently need tools for automated CI.



DevCloud provides multiple services such as ProjectMan, CodeHub, and CloudPipeline, and supports mainstream R&D scenarios such as Internet app, mobile app, microservice, and embedded app development. DevCloud provides tools for the entire software lifecycle to simplify cloud-based application development, deployment, release, and cloudification.

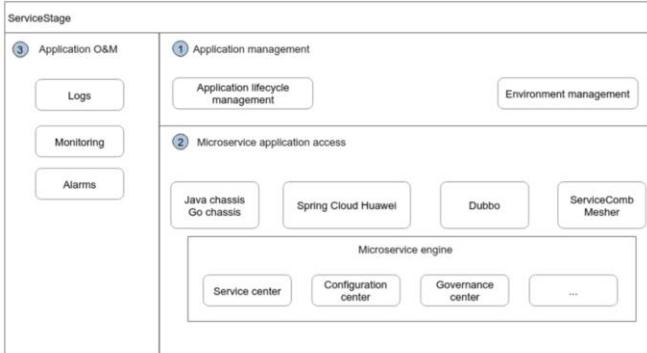
Contents

1. Cloud Native Applications and Microservices
2. Cloud Native Application Frameworks
3. **Huawei Cloud Native Application Solutions**
 - ASM
 - Application Middleware
 - DevCloud
 - **ServiceStage**

ServiceStage

- This application management and O&M platform lets you deploy, roll out, monitor, and maintain applications all in one place. Supported running environments include Java, Go, PHP, Node.js, Python, and Docker. Web, microservice (Apache ServiceComb, Spring Cloud, Dubbo, and service mesh), and common applications make enterprise cloud migration easier.

- Functions**
- Manages application lifecycle and environment.
 - Microservice application access, working with the microservice engine for service registry and discovery, configuration, and governance
 - Application O&M through logs, monitoring, and alarms



52 Huawei Confidential



- ServiceStage provides application hosting, monitoring, alarms, and log analysis for enterprise developers, test personnel, O&M personnel, and project managers. The platform is compatible with mainstream application technology stacks, including multiple languages, microservice frameworks, and running environments in the industry. It helps enterprises improve the management and O&M efficiency of traditional, web, and microservice applications, focus on industry-oriented application innovation, and improve enterprise competitiveness.
- Spring Cloud: mainstream open-source microservice development framework in the industry.
- spring-cloud-huawei: Spring Cloud applications can be hosted on Huawei Cloud using spring-cloud-huawei.
- ServiceComb: open-source microservice framework contributed by Huawei to Apache.

Application Management and Microservice Application Access

Application management

- A developed application is hosted on ServiceStage for complete lifecycle management.
 - Creation and deployment from source code, software packages (JAR, WAR, or ZIP), and container images.
 - Full process management (creation, deployment, start, upgrade, rollback, scaling, stop, deletion, logout)
- An environment is managed as a collection of compute, storage, and network infrastructures used for application deployment and running.

Microservice application access

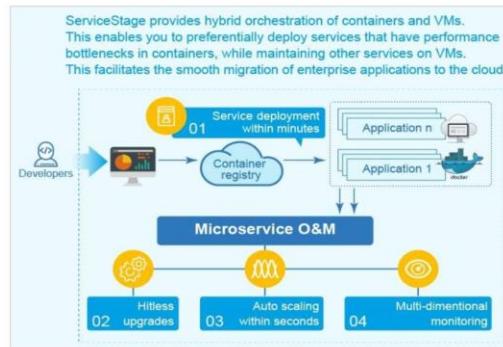
Mainstream microservice frameworks are accessible and manageable. Flexibly select the most suitable microservice technologies to quickly develop cloud applications as you adapt to complex and changing service requirements.

- Supports the native ServiceComb microservice framework.
- Compatible with mainstream open-source microservice frameworks, simplifying Spring Cloud and Dubbo microservice access.
- Provides microservice governance capabilities. After applications developed using the microservice framework are hosted on ServiceStage, microservices are registered with the service center when application instances are started.

- ServiceStage combines basic resources (such as CCE and ECS) and optional resources (such as ELB, RDS, and DCS) in the same VPC into an environment, such as a development environment, testing environment, pre-production environment, or production environment. The resources within an environment can be networked together. Managing resources and deploying services by environment simplifies O&M.
- Dubbo is an open-source, high-performance, and lightweight Java RPC service framework developed by Alibaba. It can be seamlessly integrated with the Spring framework.

Application O&M

- Multi-dimensional metrics monitoring for application components display the running status of online applications.
- GUI-based log query and search help you quickly locate faults.



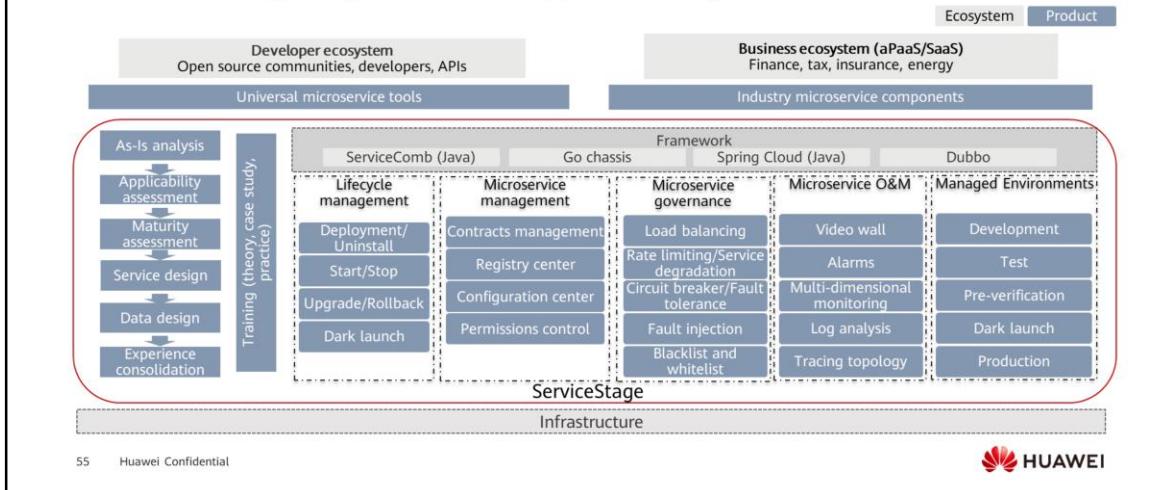
54 Huawei Confidential



- ServiceStage:
 - Graphically displays application monitoring metrics in real time, including CPU usage, alarms, node exceptions, run logs, and key events.
 - Supports microservice API-level SLA metrics (throughput, latency, and success rate) monitoring and governance in real time (in seconds), ensuring continuous service running.

ServiceStage Hybrid Cloud Microservice Solution

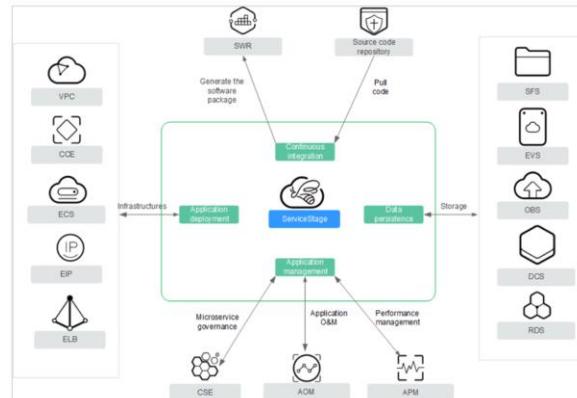
- This solution embraces the open-source ecosystem and provides full-scenario capabilities: consulting, environment management, and microservice application management and O&M.



- Solution value:
 - Application hosting: Full-lifecycle hosting of traditional, web, and microservice applications is supported, supporting dark launch and scaling of applications.
 - Application monitoring: Application running status can be observed, monitored, and controlled, ensuring easy O&M.
 - Application alarms: Alarm information is delivered through multiple channels in real time so enterprises can respond to system faults as quickly as possible.
 - Application logs: A massive number of logs are stored, supporting second-level search and facilitating fault locating and operations analysis.

ServiceStage and Other Cloud Services

- ServiceStage is a one-stop cloud application platform integrating knowledge and experience in cloud transformation and technology innovation. It offers core functions for infrastructure, storage, database, software repository, monitoring and O&M, and middleware services.



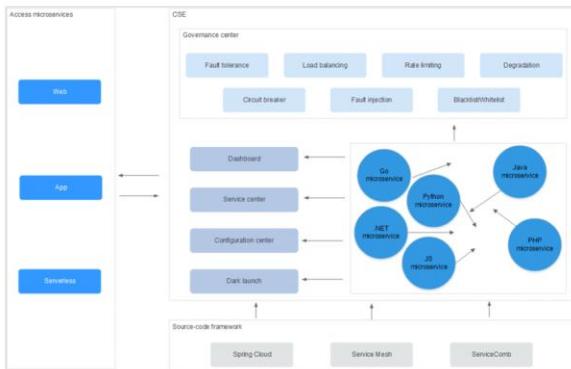
56 Huawei Confidential



- ServiceStage:
 - Interconnects with source code repositories, such as DevCloud, GitHub, Gitee, GitLab, and Bitbucket. After it is bound, you can directly pull up the source code from source code repositories for building.
 - Integrates the software center and archives the built software packages (or image packages) to the corresponding repositories and organizations.
 - Integrates related infrastructure, such as VPC, CCE, ECS, EIP, and ELB. When deploying applications, you can directly use existing or new infrastructures.
 - Integrates the Cloud Service Engine (CSE). You can perform operations related to microservice governance on the ServiceStage console.
 - Integrates Application Operations Management (AOM) and Application Performance Management (APM) services. You can perform operations related to application O&M and performance monitoring.
 - Integrates storage, database, and cache services and implements persistent data storage through simple configuration.

Application Scenario: Using ServiceStage to Build Microservices

- With the increasing complexity of enterprise services, many applications built on traditional monolithic architecture are stacked, bloating their service systems. There is also increasing complexity from interconnection and integration of legacy and new applications, causing rollout delays of new applications. The solution is loosely coupled and distributed microservice architecture.



57 Huawei Confidential

- Microservice-based applications allow enterprises to divide a bulky system into smaller service components. These components intercommunicate through lightweight protocols, decoupling their individual lifecycle management.
- ServiceStage supports full lifecycle management of microservice applications. It supports runtime environments (Java, Go, Node.js, Docker, and Tomcat) and manages microservice applications (Apache ServiceComb Java chassis, Spring Cloud, Dubbo, and Service Mesh) without intrusion. In addition, it provides functions (configuration management, monitoring and O&M, and service governance) for easier cloud migration.



- Ever growing services may encounter various unexpected situations, such as instantaneous and large-scale concurrent access, service errors, and intrusion. The microservice architecture implements fine-grained service management and control to meet service requirements.
- ServiceStage provides superior microservice application solutions and has the following advantages:
 - Supports multiple microservice frameworks, such as native ServiceComb, Spring Cloud, Dubbo, and Service Mesh, and supports the dual-stack mode (SDK and Service Mesh interconnection). The service code can be directly managed on the cloud without modification.
 - Supports API management based on Swagger.
 - Supports multiple languages, such as Java, Go, .Node.js, PHP, and Python.
 - Provides functions such as service center, configuration center, dashboard, and dark launch.
 - Provides complete microservice governance policies, including fault tolerance, rate limiting, service degradation, circuit breaker, fault injection, and blacklist and whitelist. GUI-based operations can be performed in different service scenarios, greatly improving the availability of service governance.

Quiz

1. (True or False) Microservice architecture starts to replace the SOA architecture thanks to its nature.
 - A. True
 - B. False
2. (Multiple answers) Which of the following are ServiceStage functions?
 - A. Application management
 - B. Microservice application access
 - C. Distributed transaction management
 - D. Application O&M

- Answer 1: False. The microservice architecture features decoupling and DevOps.
- Answer 2: ABCD

Quiz

1. (Discussion) What are the advantages of microservice architecture over SOA?
2. (Discussion) What are the advantages and precautions when building microservice applications on the cloud?

- Discussion 1: Discuss the architecture, development, release, and O&M.
- Discussion 2: Discuss the advantages of microservices, precautions for cloudification, and O&M management.

Summary

- This course walks you through the microservice architecture, technologies, and related services, including open-source Spring Cloud, ServiceComb, Istio, and Huawei Cloud ASM, common middleware, ServiceStage, and DevCloud. After learning this section, you will be familiar with the common microservice architecture and related functions.

Acronyms and Abbreviations

- AOM: Application Operations Management
- AOS: Application Orchestration Service
- API: Application Programming Interface
- APM: Application Performance Management
- AS: Auto Scaling
- BMS: Bare Metal Server
- CCE: Cloud Container Engine
- CCI: Cloud Container Instance
- CI/CD: Continuous Integration/Continuous Delivery
- CNCF: Cloud Native Computing Foundation
- DDoS: Distributed Denial of Service
- DevOps: Development and Operations
- DataArts Studio: Data Lake Governance Center
- DIS: Data Ingestion Service
- DLI: Data Lake Insight
- DNS: Domain Name Service
- ECS: Elastic Cloud Server
- EIP: Elastic IP

Acronyms and Abbreviations

- ELB: Elastic Load Balance
- EVS: Elastic Volume Service
- GSLB: Global Server Load Balance
- HA: High Availability
- IAM: Identity and Access Management
- IDC: Internet Data Center
- IMS: Image Management Service
- ISV: independent software vendor
- MCP: Multi-Cloud Container Platform
- MRS: MapReduce Service
- NAT: Network Address Translation
- NAT: Network Address Translation
- OBS: Object Storage Service
- OCI: Open Container Initiative
- OCR: Optical Character Recognition
- RDS: Relational Database Service
- SMN: Simple Message Notification
- SWR: SoftWare Repository for Container
- VM: virtual machine
- VPC: Virtual Private Cloud
- VPN: Virtual Private Network
- WAF: Web Application Firewall

Recommendations

- Huawei iLearning
 - <https://e.huawei.com/en/talent/portal/#/>
- Huawei Cloud Help Center
 - <https://support.huaweicloud.com/intl/en-us/index.html>
- HUAWEI CLOUD Developer Institute
 - <https://edu.huaweicloud.com/intl/en-us/>
- Huawei Talent Online
 - <https://e.huawei.com/en/talent/portal/#/>

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©2022 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.



Huawei Cloud O&M Solution



Foreword

- O&M personnel need to ensure that services and systems can run smoothly on the cloud and meet the requirements for IT resource and service system O&M.
- Cloud O&M services provide service resource monitoring and alarm reporting to improve O&M efficiency and ensure services run smoothly.
- This course covers some common O&M tools and Huawei Cloud O&M services.

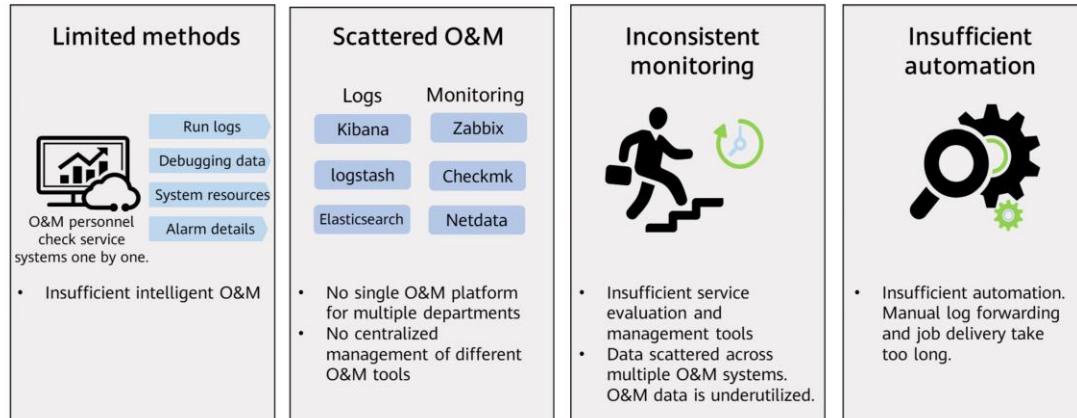
Objectives

- Upon completion of this course, you will:
 - Understand Huawei Cloud O&M.
 - Understand open source O&M tools.
 - Understand the features and functions of Huawei Cloud O&M services.
 - Be familiar with the Huawei Cloud O&M solution.

Contents

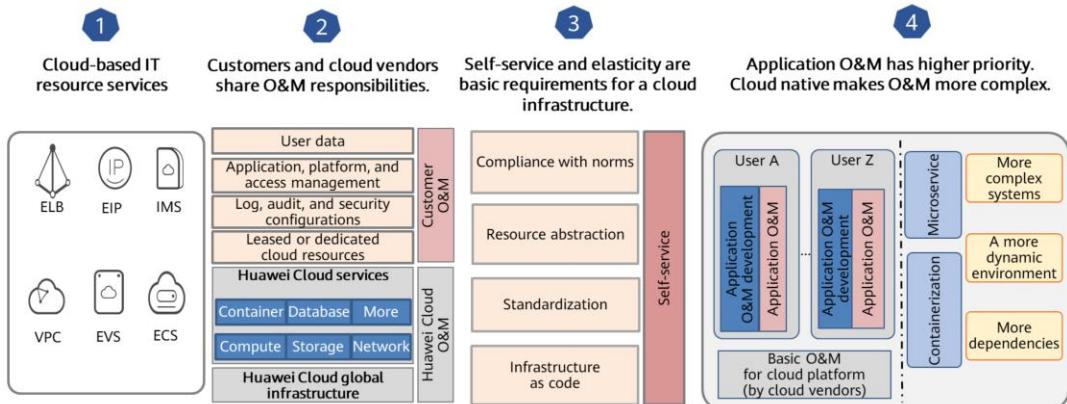
1. Cloud O&M Overview
2. Open Source O&M Tools
3. Huawei Cloud O&M Services

Challenges Facing Traditional O&M



- O&M personnel have to master professional skills, make complicated configurations, and maintain multiple systems.
- Metrics cannot be associated for analysis. Therefore, O&M personnel need to check metrics one by one based on their experience.
- Distributed tracing systems are complicated, expensive, and unstable.

O&M Requirements of the Cloud Architecture



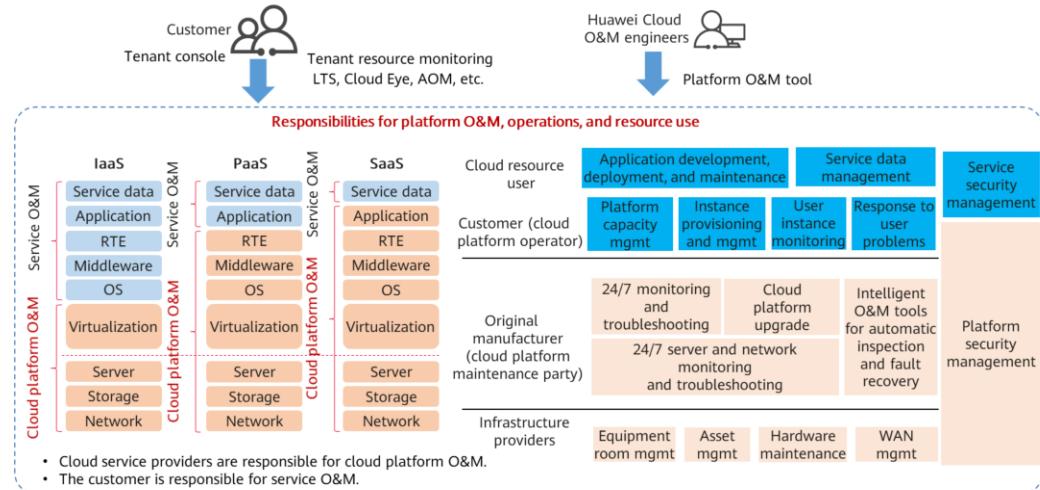
- Although O&M methods depend on the IT architecture evolution, fundamental O&M requirements do not change with cloud migration. Such requirements include service stability and reliability, shorter mean time to repair (MTTR), and higher routine O&M efficiency.

5 Huawei Confidential



- The IT architecture becomes more and more complex, and there are obvious differences between cloud O&M and traditional IT O&M. O&M personnel face many challenges.
- Many enterprises opt to have development and O&M departments with different goals. However, department miscommunication may hinder projects and lower efficiency. Therefore, the entire system architecture needs to evolve continuously, moving from traditional O&M to automated O&M. This will help break down the barriers between O&M engineers, development engineers, and quality assurance engineers, and form an efficient work system.

Huawei Cloud O&M and Service O&M

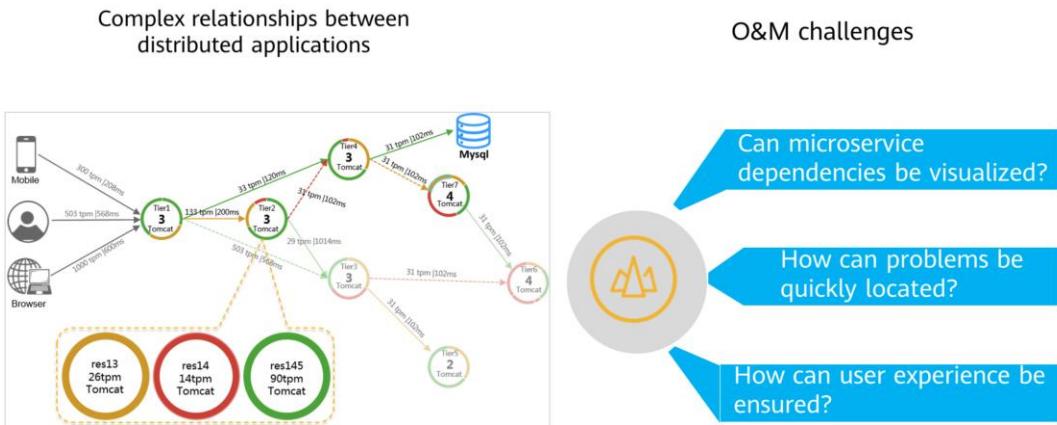


6 Huawei Confidential



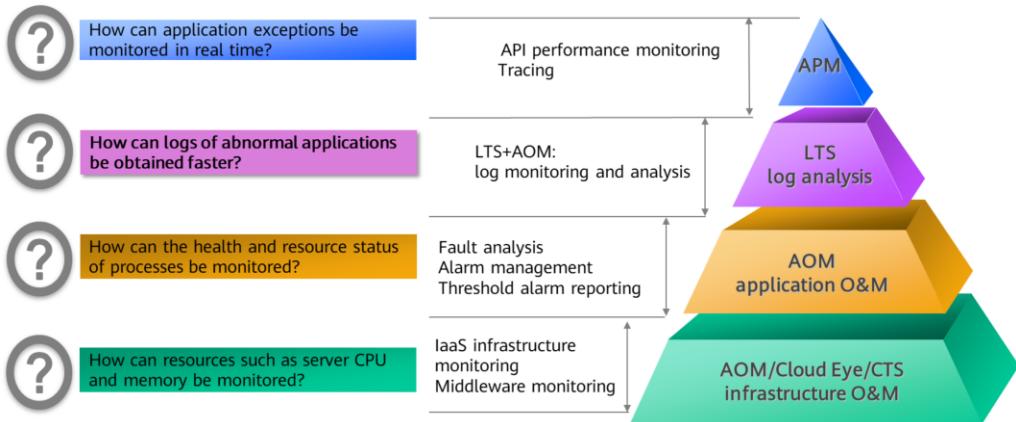
- To help users focus on service O&M and reduce the workload in routine platform maintenance, Huawei is responsible for platform O&M and provides users with a stable and reliable cloud platform.
- Console is a visualized entry for cloud resource users to manage and provision resources.
- Cloud Eye, Application Operations Management (AOM), and Application Performance Management (APM) are multi-dimensional monitoring platforms that allow users to monitor cloud resource usage and service status, set alarm rules, quickly respond to exceptions, thereby ensuring smooth service running.
- Users can use the cloud O&M service console and tools to support service O&M.

The Challenges of Diverse Flexible Cloud Applications



- With the popularization of microservices, the relationship between applications is increasingly complex. O&M personnel cannot handle it anymore. Professional tools are required to comprehensively monitor application calls, and display service execution traces and statuses, thereby helping users quickly demarcate performance bottlenecks and faults.
- After applications are migrated to the cloud, users still want microservice dependency visualization, better end user experience, fast problem tracing, association analysis on scattered logs. To meet these requirements, Huawei Cloud provides diverse O&M services to improve O&M efficiency.

Application O&M Solution Portfolios



- Huawei Cloud launched a dimensional cloud application O&M solution that integrates AOM and APM. This solution monitors infrastructure, applications, and services in real time, and supports association analysis of application and resource alarms, log analysis, intelligent threshold, distributed tracing, and mobile app exception analysis, enabling users to quickly diagnose and rectify faults within minutes, and ensure stable application running.
 - Resource monitoring: AOM monitors applications and cloud resources in real time, collects metrics, logs, and events to analyze application health status, and supports alarm reporting and data visualization.
 - Log management: LTS provides log collection, real-time query, and storage, helping users easily cope with routine O&M.
 - Locating of performance problems: APM provides professional distributed application performance analysis capabilities, enabling O&M personnel to quickly locate problems and resolve performance bottlenecks in a distributed architecture.

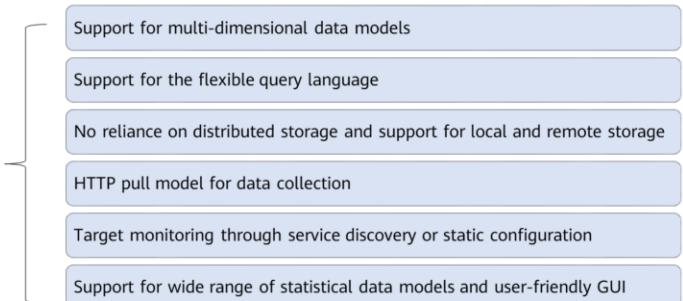
Contents

1. Cloud O&M Overview
2. **Open Source O&M Tools**
3. Huawei Cloud O&M Services

Prometheus Overview

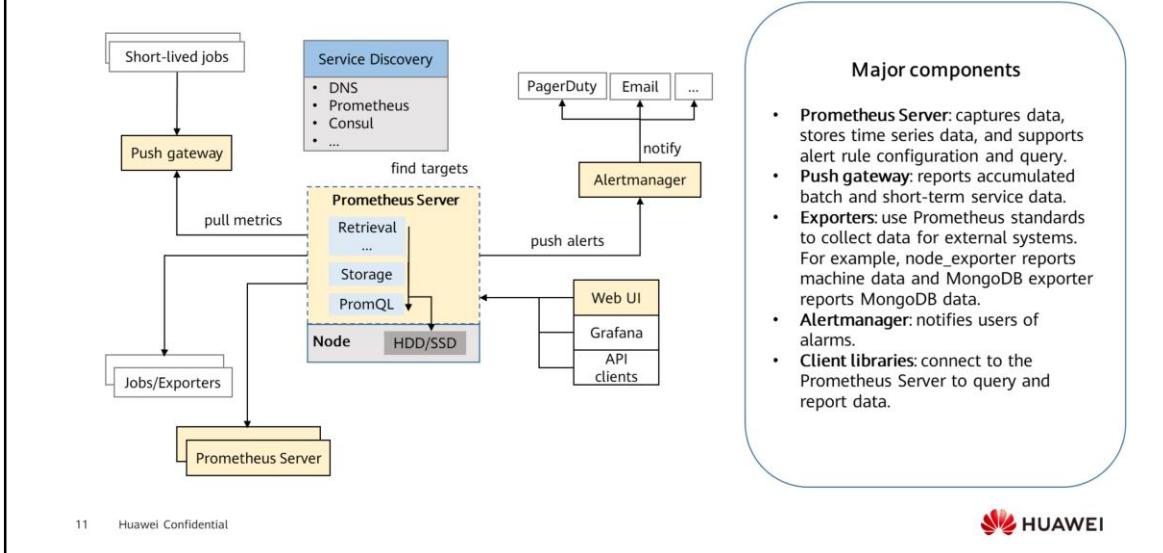
- Prometheus is an open source monitoring and alarming framework of the Cloud Native Computing Foundation (CNCF). It provides common data models and APIs for data collection, storage, and query. Prometheus can display data and alarms in graphs by working with data visualization tools such as Grafana. Go-based implementation also slashes server O&M costs.
- Perfect for time series data, machine-centric monitoring and monitoring for highly dynamic service-oriented architectures, and collection and query of multidimensional microservice data

Main functions



- Prometheus is an open source monitoring tool. It is derived from Google's borgmon monitoring system, which was created by former Google employees working at SoundCloud in 2012. Prometheus was developed as an open source community project and officially released in 2015. In 2016, Prometheus officially joined the Cloud Native Computing Foundation, after Kubernetes.
- As a key part of observability practices (monitoring, logging, and tracing), monitoring has changed a lot in the cloud native era compared with previous system monitoring. Microservice and containerization lead to the exponential increase of monitoring objects and metrics. Short lifecycles of monitoring objects greatly increase monitoring data volumes and complexity.
- Therefore, Prometheus is developed to unify monitoring metrics and data query languages. Prometheus can be easily integrated with many open source tools to monitor systems and services. It also analyzes vast volumes of data, facilitating system optimization and decision-making. It can be used in any scenarios where metrics need to be collected.
- PromQL is a query language for labeled time series data. It is totally different from the SQL query statements for relational databases.
- Prometheus is not only a time series database. It provides functions of integrated tools in the entire ecosystem.
- Prometheus is mainly used to monitor infrastructures, including servers (such as CPU and memory), databases (such as MySQL and PostgreSQL), and web services. It pulls data based on the configuration and connection with data sources.

Prometheus Architecture



- Prometheus is designed for reliability and allows users to quickly diagnose problems. Each Prometheus server is standalone, not depending on network storage or other remote services.
- Prometheus pulls data from exporters or through a gateway. (If it is deployed in Kubernetes, service discovery can be used). It stores scraped data locally, runs rules to cleanse and sort data, and stores processed data in new time series.
- Prometheus components:
 - The Prometheus server periodically scrapes data from targets via service discovery or static configuration.
 - When the size of the newly scraped data is larger than the configured cache, the data is persisted to disks. (If remote storage is used, the data will be persisted to the cloud).
 - Prometheus periodically queries data. When conditions are met, Prometheus pushes alerts to the configured Alertmanager.
 - When receiving an alert, the Alertmanager performs aggregation, deduplication, and noise reduction based on the configuration, and then sends the alert.
 - APIs, the Prometheus console, or Grafana can be used to query and aggregate data.
- Data can be pulled by and pushed to Prometheus.
 - Pull: Existing exporters are installed on the client and run as a daemon process. Exporters collect data, respond to HTTP requests, and return metrics.
 - Push: The client (or server) with the official pushgateway plug-in installed can organize monitoring data into metrics and send them to the pushgateway using a script. Then, the pushgateway pushes the metrics to Prometheus as an intermediary forwarding medium.

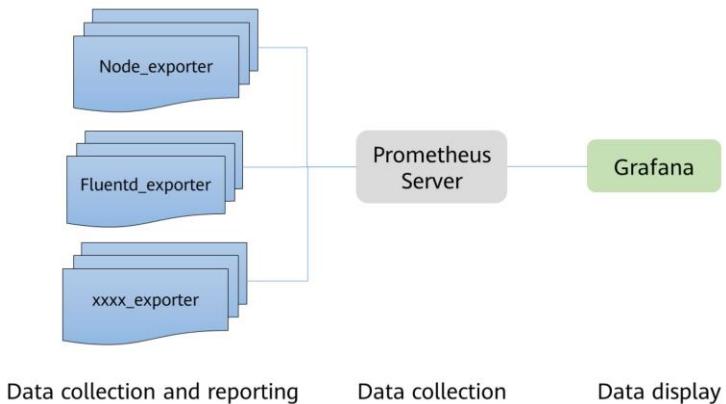
Grafana Overview

- Grafana is a multi-platform, open source web application for data analytics and visualization. After connecting to a data source, Grafana can display data charts and alerts in web browsers. It presents the data from a time series database (TSDB) in intuitive chart form. It is commonly used to visualize infrastructure time series data and application analysis and is widely used in domains including industrial sensors and device automation.
- Grafana supports many data sources, such as Graphite, InfluxDB, OpenTSDB, Prometheus, and Elasticsearch. Each data source has its own query editor that is customized to include the features and capabilities of the data source.

- It has the following six features:
 - Display mode: It provides fast and flexible visualization and supports extensive dashboard plug-ins, such as heatmaps and line charts.
 - Data sources: It supports diverse data sources, such as Graphite, InfluxDB, OpenTSDB, Prometheus, Elasticsearch, CloudWatch, and KairosDB.
 - Notifications: Rules are defined based on different metrics to determine whether to trigger an alarm and send a notification.
 - Transformation: Different data sources can be used in the same chart. Data sources can be specified based on each query or even customized.
 - Annotations: Users can annotate graphs with rich events from different data sources and hover over events to show full event metadata and tags.
 - Filters: Ad hoc filters allow users to add key/value filters that are automatically added to all metric queries that use the specified data source.
- A TSDB is a database optimized for time-stamped or time series data. It is built specifically for handling measurements and events that are time-stamped. Time series data can be measurements or events that are tracked, monitored, downsampled, and aggregated over time. It includes server metrics, application performance, network data, sensor data, and many other types of analytics data.
- Grafana components:
 - filebeat: collects Fault Tracing & Diagnosing System (FTDS) data.
 - metricbeat: collects system resource data.
 - logstash: cleanses logs.
 - influxdb: distributed time series database.
 - grafana: displays data.

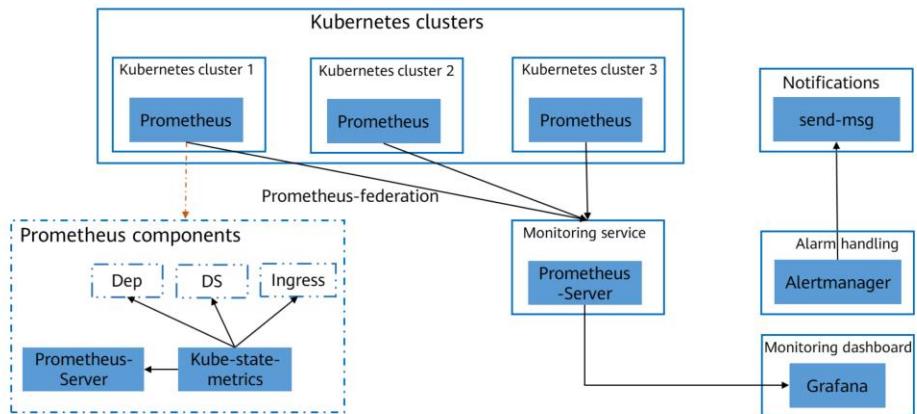
Prometheus + Grafana

- Prometheus supports Grafana and can quickly display monitoring data through Grafana.



Open Source O&M Solution Architecture

- Prometheus or the combination of Prometheus and Grafana is used to monitor Kubernetes clusters.



14 Huawei Confidential



- Prometheus is used to monitor Kubernetes clusters, including:
 - Node metrics, such as CPUs, load, fdisk, and memory.
 - Status of internal components, such as kube-scheduler, kube-controller-manager, and kubedns or coredns.
 - Application metrics, such as the Deployment status, resource requests, scheduling, and API latency.

Contents

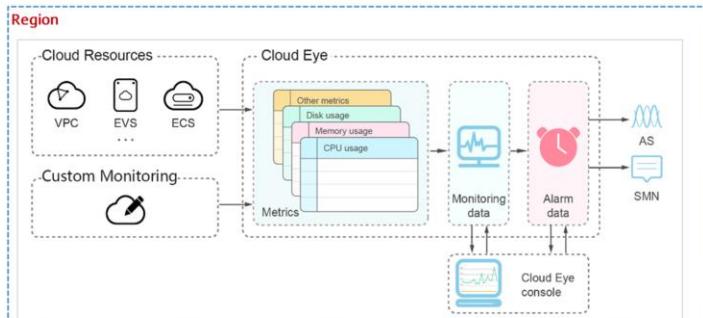
1. Cloud O&M Overview
2. Open Source O&M Tools
- 3. Huawei Cloud O&M Services**

- Cloud Eye
- CTS
- LTS
- AOM
- APM
- CPTS

Cloud Eye

- Cloud Eye is a **multi-dimensional resource monitoring service**. It monitors resource utilization, track the status of cloud services, and includes configurable alarm rules and notifications, to help users quickly respond to any changes when they happen.

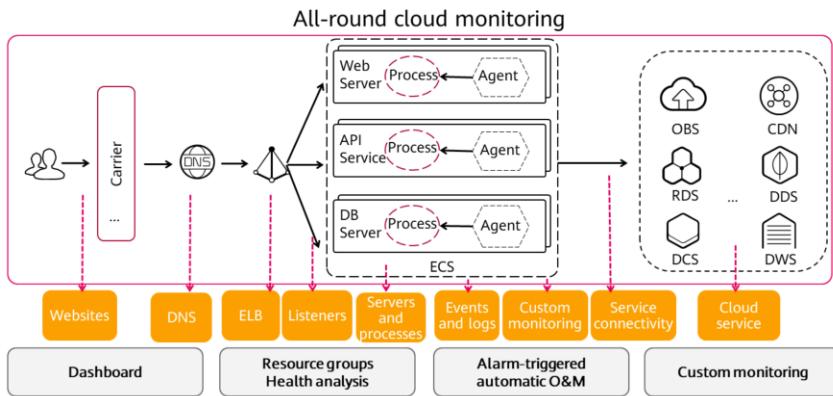
- Main functions and features**
- Cloud resource monitoring
 - Flexible alarm rules
 - Multiple alarm notifications
 - Refined server monitoring
 - Visualized data panels
 - Resource groups
 - Website monitoring
 - Long-term data retention
 - Event monitoring



- Cloud Eye provides the following functions:
 - Automatic monitoring: Cloud Eye automatically starts after resources such as ECSs are created. On Cloud Eye console, Users can view the resource status and create alarm rules.
 - Server monitoring: After installing the Agent on an ECS or Bare Metal Server (BMS), users can collect minute-level ECS or BMS monitoring data in real time.
 - Flexible alarm rule configuration: Users can create alarm rules for multiple resources at the same time. After an alarm rule is created, users can flexibly manage it, for example at any time users can modify, enable, disable, or delete it.
 - Real-time notification: Users can enable Alarm Notification when creating alarm rules. When the cloud service status changes and the monitoring data of the metric reaches the threshold specified in an alarm rule, Cloud Eye notifies users by sending messages, emails, or HTTP or HTTPS requests to server IP addresses. In this way, users can monitor the cloud resource status and changes in real time.
 - Monitoring panel: allows users to view cross-service and cross-dimension monitoring data on a monitoring panel and centrally displays metrics of key services that users care about. This not only provides an overview of the status of cloud services, but also allows users to view monitoring details during troubleshooting.
 - Monitoring data transfer to OBS: The retention period of raw data of each metric is two days. After the retention period expires, the raw data will not be saved. Users can dump raw data to OBS buckets for longer storage.

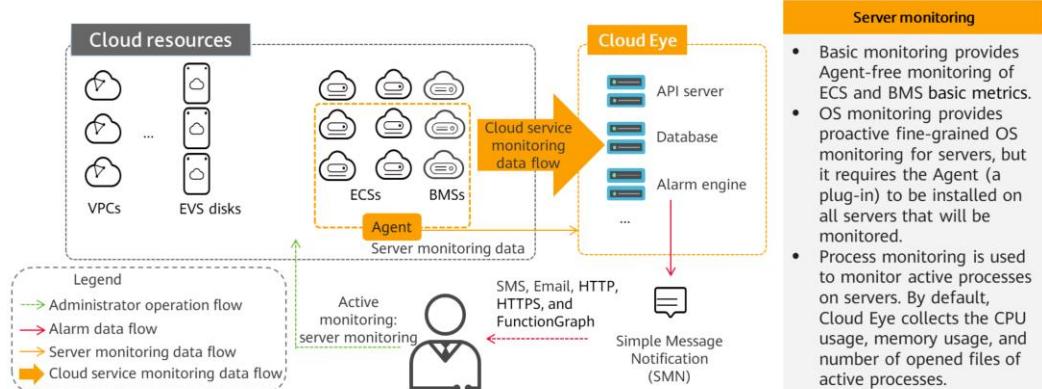
Cloud Eye

- Cloud Eye provides multiple built-in metrics based on the attributes of different services. After a service is enabled for Huawei Cloud, Cloud Eye automatically associates its built-in metrics. Users can track the cloud service status by monitoring these metrics.



Server Monitoring

- Server Monitoring includes basic monitoring, OS monitoring, and process monitoring. Basic monitoring monitors built-in basic metrics reported by servers. OS monitoring and process monitoring provide server monitoring that is proactive and fine-grained, but they require Agents to be installed on all servers that will be monitored. The data is collected every minute.



19 Huawei Confidential

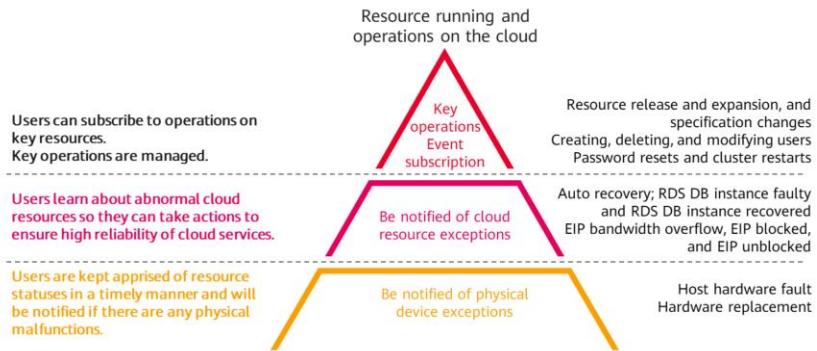


Server monitoring:

- 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 users a better understanding of the ECS or BMS usages.
- Basic monitoring covers metrics automatically reported by ECSs. The data is collected every 5 minutes.
- OS monitoring provides proactive and fine-grained OS monitoring for ECSs or BMSs, and it requires the Agent to be installed on all servers that will be monitored. The data is collected every minute. OS monitoring supports metrics such as CPU usage and memory usage (Linux).
- Process monitoring is used to monitor active processes on servers. By default, Cloud Eye collects the CPU usage, memory usage, and number of opened files of active processes.

Event Monitoring

- In event monitoring, users can query system events and custom events reported to Cloud Eye through an API. Users can create alarm rules for both system events and custom events. When specific events occur, Cloud Eye generates alarms for them.
- Events, stored and monitored by Cloud Eye, are key operations on cloud service resources. Users can view events to see operations performed by specific users on specific resources, such as deleting or rebooting an ECS.



- 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.

Alarm Management

- Users can create alarm rules for key metrics of cloud services. When the conditions in the alarm rule are met, Cloud Eye reports the alarms using emails, SMS, or HTTP/HTTPS requests, enabling users to quickly respond to resource changes.



Flexible policies

Users can select monitoring objects and configure thresholds and triggering policies of specified metrics for various alarm scenarios.



Diversified methods

Alarm notifications can be reported by HTTP, HTTPS, FunctionGraph, SMS, and email. Cloud Eye can work with Auto Scaling (AS) to trigger the system to automatically add or remove servers.



Efficient configuration

Users can use custom and default alarm templates and preset. Alarm rules can be created for resources in batches.



Rich templates

Alarm templates for multiple services are preset in the system.

- Alarm rules can be created for all monitoring items of Cloud Eye.
- Users can configure the effective time of alarm rules.
- Notifications can be reported by multiple methods, such as email, SMS, HTTP, or HTTPS.
- Service invoking based on alarm rules are supported. For example, when a certain type of alarm is triggered, other cloud services (such as FunctionGraph) can be triggered to perform configured operations.

Monitoring Panel

- Cloud Eye monitoring panels allow users to compare and view metrics in a customized manner, especially key metrics.



Monitoring overview



Instance display



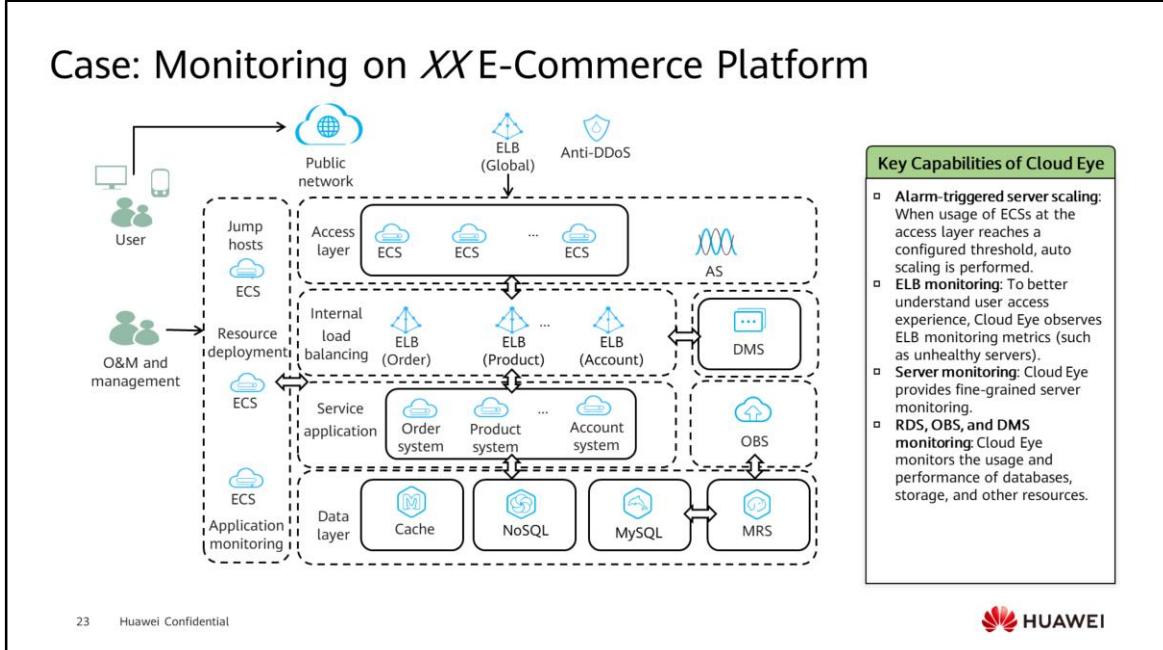
Item comparison



Custom dashboards

- Dashboards allow users to compare performance data of different services from different dimensions. Users must create a dashboard before adding graphs.

Case: Monitoring on XX E-Commerce Platform



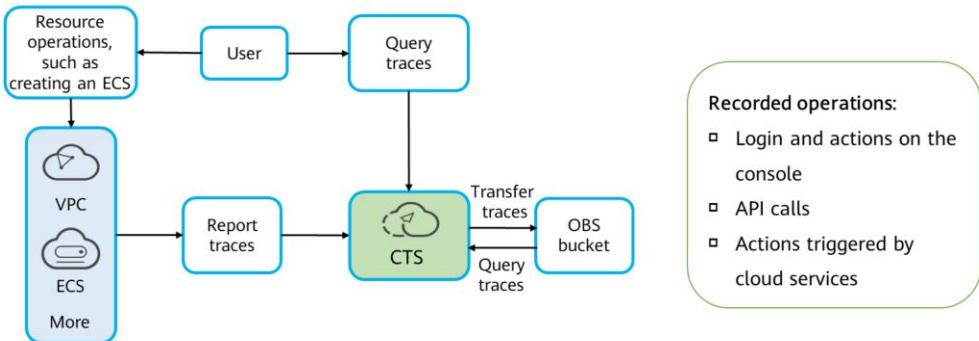
- E-commerce services feature large data volume and large data access, which requires large memory, fast data exchange and processing, and extremely strict monitoring.
- ECS is a core service in e-commerce scenarios. Therefore, a comprehensive and three-dimensional ECS monitoring system plays an important role in service stability. Proactive fine-grained server monitoring of Cloud Eye helps ensure that e-commerce services run smoothly.
- People access the websites of e-commerce platforms and make transactions. During grand annual shopping festivals, the websites are often hit by various problems like slow page loading and long network latency when people access from different networks. Website monitoring can perform continuous dialing tests on websites or ECS elastic IP addresses (EIPs) to monitor the availability and response time of the websites.
- For services used by an e-commerce platform, such as Relational Database Service (RDS), Elastic Load Balance (ELB), and Virtual Private Cloud (VPC), cloud service monitoring allows users to track the status of each cloud service and usage of each metric. After setting alarm rules for cloud service metrics, users can get a more accurate picture of the health of cloud services.
- An e-commerce platform involves many Huawei Cloud services, such as ECS, Content Delivery Network (CDN), AS, Web Application Firewall (WAF), RDS, ELB, and Object Storage Service (OBS). With resource groups, users can view resource usages, alarms, and health status and manage alarm rules, relating to a specific service. This greatly reduces O&M complexity and improves O&M efficiency.

Contents

1. Cloud O&M Overview
2. Open Source O&M Tools
- 3. Huawei Cloud O&M Services**
 - Cloud Eye
 - CTS
 - LTS
 - AOM
 - APM
 - CPTS

CTS

- Cloud Trace Service (CTS) is Huawei Cloud's log audit service. It tracks user activities and changes to cloud resources. It helps users collect, store, and query operational records for security analysis, audit and compliance, and fault location.



- Log auditing is the core of information security audit. They are essential for the security risk control of information systems in both private and public sectors.
- CTS directly connects to other Huawei Cloud services, records operations on cloud resources and the results, and transfers these records in the form of trace files to OBS buckets in real time.
- CTS provides the following functions:
 - Trace recording of operations performed, including system-triggered operations, operations on the management console, and API-calling operations.
 - Trace query on the CTS console from the last seven days by multiple dimensions: trace type, trace source, resource type, filter, operator, trace status.
 - Trace transfer to OBS buckets periodically for later query to meet compliance and persistent storage requirements.
 - Trace file encryption using keys provided by the Data Encryption Workshop (DEW) during the transfer.

CTS Functions

- Traces are cloud resource operation records captured and stored by CTS. They identify the time and operator of each operation (such as API calls).
- There are two types of traces: management traces and data traces.
- The trace list displays traces generated in the last seven days. These traces record operations on cloud service resources, including creation, modification, and deletion, but query operations are not recorded.
 - Management traces: details of the creation, configuration, or deletion of any cloud resources in the tenant account.
 - Data traces: details of data operations, such as when data is uploaded or downloaded.

Management traces

- Traces reported by cloud services
- Operations on cloud services, such as creating and deleting ECSSs and VPCs

Data traces

- Read and write traces reported by OBS
- Operations on OBS buckets, such as deleting files

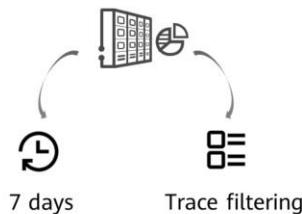
- A trace file is a collection of traces. CTS generates trace files based on services and transfer cycle and send these files to the specified OBS bucket in real time. In most cases, all traces of a service generated in a transfer cycle are compressed into one trace file. However, if there are a large number of traces, CTS will adjust the number of traces contained in each trace file. Trace files are in JSON format.

CTS - Tracker

- A tracker named system is automatically created when users enable CTS. This tracker identifies and associates with all cloud services a tenant account is using, and records all operations of the account.

Tracker management:

- Query traces generated within seven days for free
- There are two types of trackers: management trackers and data trackers.



Trace list management:

- Query traces in the last seven days
- Filter traces by multiple dimensions (time range, trace status, and trace type)



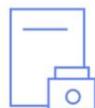
- Management trackers record operations on all cloud resources, such as creation, login, and deletion.
- Data trackers record operations on data, such as upload and download.

Application Scenarios



Compliance audit

- CTS helps users obtain certifications for industry standards, such as DJCP MLPS and PCI DSS, for user service systems.
- It detects unnoticed issues, thereby helping users improve processes and provide evidence for service audits.



Key event notifications

- CTS works with FunctionGraph to send notifications to natural persons or service APIs when any key operation is performed.



Data mining

- CTS mines data in traces to facilitate service health analysis, risk analysis, resource tracking, and cost analysis. Users can also obtain the data from CTS and explore its value on their own.



Fault locating and analysis

- If there is a fault, filters help users pinpoint unusual operations for faster and make troubleshooting easier.

- Compliance audit:
 - Users need to ensure the compliance of their own service systems, and the cloud vendors they choose need to ensure the compliance of users' service systems and resources.
- Key event notifications:
 - Users can configure HTTP or HTTPS notifications targeted at their independent audit systems and synchronize CTS logs to these systems for auditing.
 - Users can also select a certain type of logs (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 preventing risks.
- Data mining:
 - A trace contains up to 24 fields, recording when an operation was performed by a specific user on a specific resource and the IP address from which the operation was performed.
- Fault locating and analysis:
 - CTS provides the following search dimensions: trace type, trace source, resource type, filter, operator and trace status. Each trace contains the request and response of an operation. Querying traces is one of the most efficient methods for locating a fault.

Contents

1. Cloud O&M Overview
2. Open Source O&M Tools
3. **Huawei Cloud O&M Services**

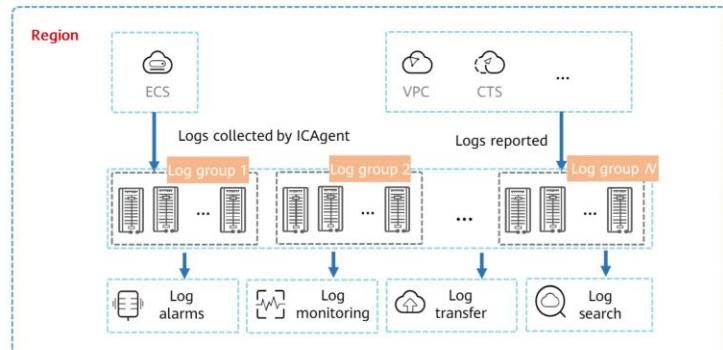
- Cloud Eye
- CTS
- LTS
- AOM
- APM
- CPTS

LTS

- 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 users with the insights needed to optimize the availability and performance of cloud services and applications. It helps users make faster data-driven decisions, perform device O&M, and analyze service trends.

Main functions:

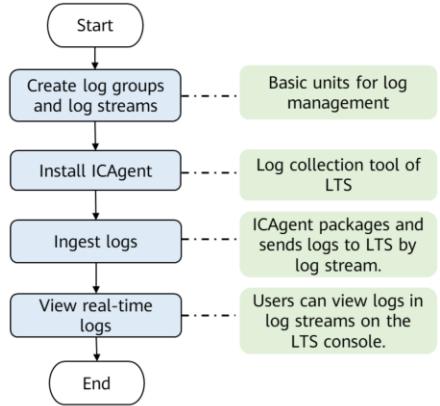
- Real-time log collection
- Log query and real-time analysis
- Log monitoring and alarms
- Log transfer to OBS



- 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. Users can query logs or transfer logs for long-term storage.
- Log query and real-time analysis: Collected logs can be quickly queried by keyword or fuzzy match. Users can analyze logs in real time to perform security diagnosis and analysis, or obtain operations statistics, such as cloud service visits and clicks.
- Log monitoring and alarm reporting: 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 of hosts and cloud services are retained in LTS for seven days by default. Users can also set the retention duration to a value ranging from 1 to 30 days. Retained logs are deleted once the duration is over. For long-term storage, users can transfer logs to OBS and Data Ingestion Service (DIS).
- A dashboard is composed of multiple charts and allows users to view SQL analysis results of logs in real time.

Concepts and Operations

- Log groups and log streams are basic units for log management. The first step to use LTS is to create a log group.
 - A log group is a collection of log streams and the basic unit for LTS to manage logs. Users can set log retention duration for a log group.
 - A log stream is the basic unit for log read and write. Users can create log streams in a log group for finer log management.
 - ICAgent is used to collect logs.



- Log groups can be created in two ways. They are either automatically created when other Huawei Cloud services are connected to LTS, or manually created by users on the LTS console.
- Users can configure logs of different types, such as operation logs and access logs, to be written into different log streams. ICAgent will package and send the collected logs to LTS by log stream. In this way, users can quickly find the target logs in the corresponding log streams. The use of log streams greatly reduces the number of log reads and writes and improves efficiency.
- If ICAgent has been installed on the host for other cloud services, skip the installation. The time and time zone of the local browser must be consistent with those of the host before the installation. Users can install ICAgent on the Host Management page of the LTS console. When ICAgent is installed, users need to configure log collection paths, which are paths of the host logs to be collected.
- During log structuring, logs with fixed or similar formats are extracted from a log stream based on the defined structuring method and irrelevant logs are filtered out. Users can then use SQL syntax to query and analyze the structured logs.

Log Collection and Analysis

- LTS allows users to ingest logs from hosts and cloud services in real time through various means, including the ICAgent or APIs. Ingested logs are displayed on the LTS console in an intuitive and orderly manner. Users can quickly query logs or dump them to OBS for long-term storage.
 - Collected logs can be structured for analysis. LTS extracts logs that are in a particular format or share a similar pattern based on user-configured extraction rules. Then users can use SQL syntax to query the structured logs.

Log data can be divided into structured data and unstructured data. Structured data refers to data that can be described by numbers or a defined data model, with strict length and format. Unstructured data refers to data that is not convenient to be represented by two-dimensional logical tables or databases. The data structure is irregular, or ad-hoc, and there is no predefined data model.

You can choose one of the following five methods to structure logs:

--	--	--	--	--	--

Format the log body by specifying separators (such as spaces and commas).

Step 1 Select a sample log event.

Enter an event or select one from existing.

Select from existing log events.

Step 2 Define logon log format.

Space Tab | + Custom

Step 3 Extract fields.

Intelligent Extraction

Content Fields **Content Fields**

Field	Type	Type	Example Value
-------	------	------	---------------

Quick A... Operation

32 Huawei Confidential



- Collected logs can be quickly queried by keyword or fuzzy match. Users can analyze logs in real time to perform security diagnosis and analysis, or obtain operations statistics, such as cloud service visits and clicks.

Log Transfer and Visualization

- Logs collected from hosts and cloud services are retained in LTS for seven days by default, but users can set the retention period to be one to thirty days. Once this retention period is over, the logs are deleted. For long-term storage or persistent logging, users can transfer logs to other cloud services, such as OBS, DIS, and DMS.
- Log transfer means copying the logs to another cloud service, but the original copies are not deleted from LTS.



LTS can display log data in tables, bar charts, line charts, and pie charts.

- Log transfer:
 - Logs can only be transferred to OBS buckets that are deployed in the same region as LTS.
 - Logs cannot be written to an encrypted OBS bucket.

Application Scenarios



- When logs are scattered across different hosts and cloud services and are periodically deleted, it is hard to obtain the desired information. Logs collected by LTS can be quickly queried by keyword or fuzzy search. Users can analyze logs in real time to perform security diagnosis and analysis.
- For example, they can obtain operational statistics, such as page visits or click counts.



- The performance and quality of website services play an important role in customer satisfaction. By analyzing the network congestion logs, users can identify performance bottlenecks, and take measures such as improving website caching policies or network transmission policies to improve performance.
- For example, users can analyze historical website data as a traffic benchmark, detect performance bottlenecks in a timely manner, scale capacity up or down as needed, and optimize network security policies based on network traffic analysis.



- Network quality is the cornerstone of service stability. LTS centralizes logs from different sources, helping users quickly detect and locate faults, backtrack easily.
- For example, users can quickly locate a problematic ECS that is using too much bandwidth. Users can also judge whether there are ongoing attacks, unauthorized hot-linking, or malicious access requests by analyzing access logs, and locate and rectify faults as soon as possible.

- Log collection and analysis:
 - Logs of hosts and cloud services are difficult to query and will be cleared regularly. LTS collects logs for unified management and displays them on the LTS console in an intuitive and orderly manner for fast query. LTS also supports long-term log storage. Collected logs can be quickly queried by keyword or fuzzy match. Users can analyze logs in real time to perform security diagnosis and analysis, or obtain operations statistics, such as cloud service visits and clicks.
- Service optimization:
 - The performance and quality of website services play an important role in customer satisfaction. By analyzing the network congestion logs, users can identify performance bottlenecks, and take measures such as improving website caching policies or network transmission policies to improve performance.
- Network troubleshooting:
 - Network quality is the cornerstone of service stability. LTS centralizes logs from different sources, helping users quickly detect and locate faults, backtrack easily. For example, users can quickly locate a problematic ECS that is using too much bandwidth. Users can also judge whether there are ongoing attacks, unauthorized hot-linking, or malicious access requests by analyzing access logs, and locate and rectify faults as soon as possible.

Contents

1. Cloud O&M Overview
2. Open Source O&M Tools
3. **Huawei Cloud O&M Services**

- Cloud Eye
- CTS
- LTS
- AOM
- APM
- CPTS

AOM

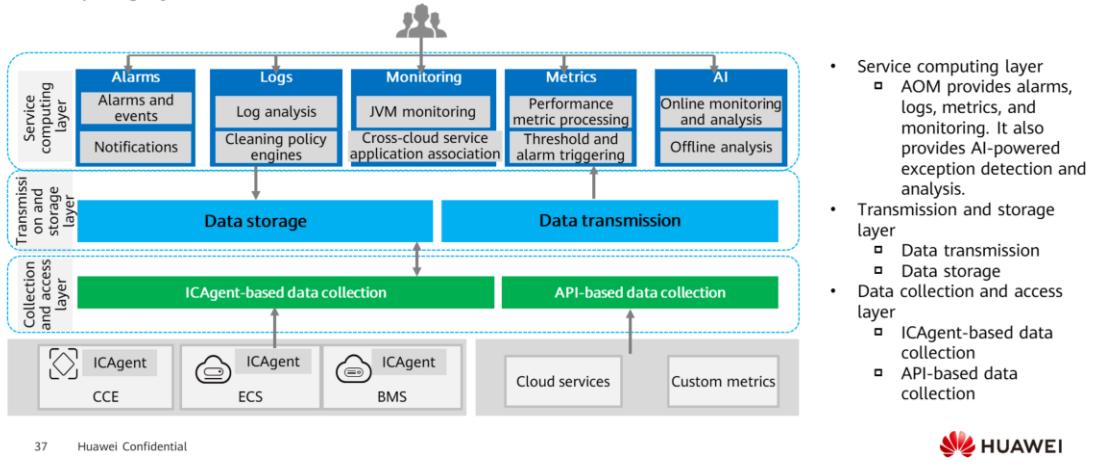
- Application Operations Management (AOM) is a one-stop, multidimensional O&M management platform for cloud applications. It monitors applications and related cloud resources in real time, analyzes application health status, and provides flexible data visualization functions. It helps detect faults in a timely manner and monitors the status of applications, services, and other resources in real time.
- It monitors hundreds of O&M metrics of mobile apps, browsers, networks, application services, middleware, and cloud resources in real time. It quickly detects and diagnoses exceptions by leveraging an O&M knowledge base and AIOps engines. This improves the reliability and quality of IT applications, improves user experience, and reduces the total cost of ownership (TCO).



- With the popularization of container technologies, more and more enterprises develop applications using microservice frameworks. As the number of cloud services increases, enterprises gradually turn to cloud O&M. However, they face the following O&M challenges:
 - O&M personnel have to master professional skills, make complicated configuration, and maintain multiple systems at the same time. Distributed tracing systems are complicated, expensive, and unstable.
 - Distributed applications face analysis difficulties such as how to visualize the dependency between microservices, improve user experience, associate scattered logs for analysis, and quickly trace problems.
- Advantages of AOM:
 - Management of massive quantities of logs: High-performance search and service analysis are supported. Logs are automatically associated and can be filtered by application, host, file, or instance.
 - Association analysis: AOM finds correlations between metrics and alarm data from applications, components, instances, hosts, and transactions, allowing users to quickly locate faults.
 - Open ecosystem: AOM opens O&M data query APIs and collection standards, and supports independent development.

Service Architecture

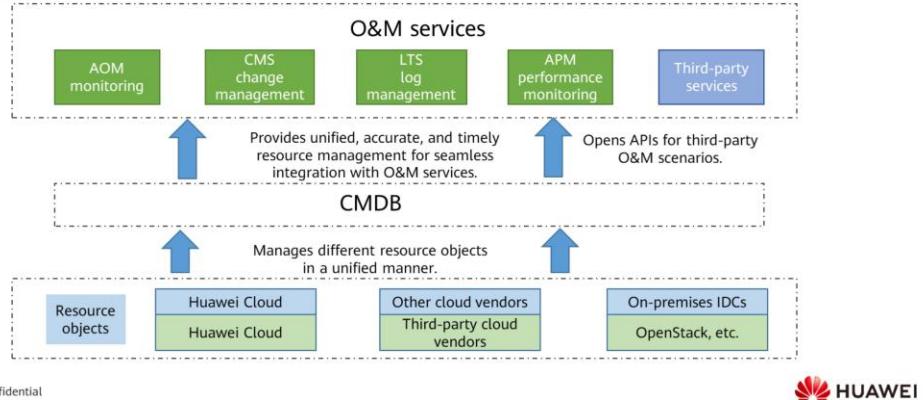
- AOM is a multi-dimensional O&M platform focused on resource data. It associates logs, metrics, resource data, alarms, and events. It consists of a data collection and access layer, transmission and storage layer, and service computing layer.



- Data collection and access layer:
 - ICAgent-based data collection: ICAgent plug-ins are installed on hosts to report O&M data.
 - API-based data collection: Custom metrics can be connected to AOM by using open APIs or Exporter APIs.
- Transmission and storage layer:
 - Data transmission: AOM Access is a proxy for receiving O&M data. Received data will be placed in a Kafka queue. Kafka then transmits the data to the service computing layer in real time using its high-throughput capability.
 - Data storage: After being processed by the AOM backend, O&M data is written into a database. Cassandra stores time series data, Redis is used for cache query, etcd stores AOM configuration data, and Elasticsearch stores resources, logs, alarms, and events.
- Service computing layer:
 - AOM provides basic O&M services such as alarm reporting, logging, and metric monitoring, and AI services such as exception detection and analysis.

Configuration Management Database (CMDB)

- Combining services from different cloud vendors has become commonplace for enterprises. CMDB can manage resource objects from Huawei Cloud, third-party cloud vendors, and on-premises IDCs in a unified manner. It also enables unified, accurate, and timely resource configuration for upper-layer O&M services.



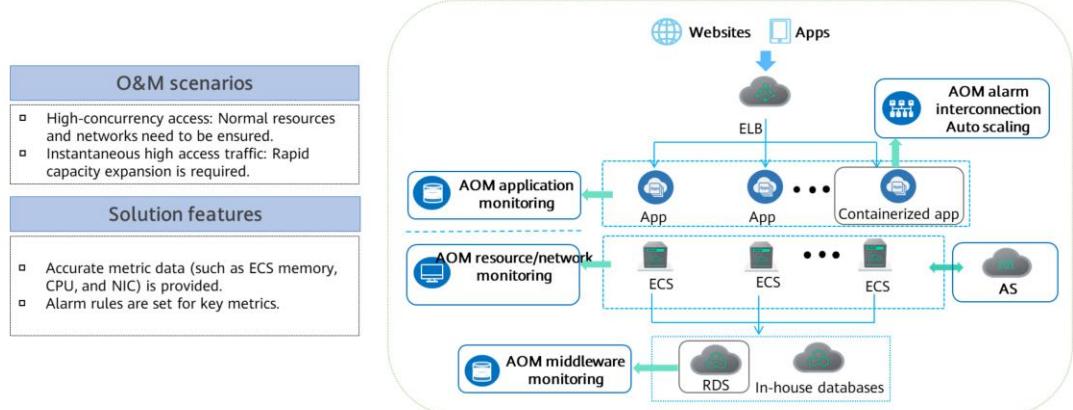
38 Huawei Confidential



- As cloud migration becomes popular, enterprises are facing the challenge of managing diverse resources from different cloud vendors. Configuration management database (CMDB) is a DevOps-based resource management platform for the entire application lifecycle. As a fundamental service for automated O&M, it centrally manages the relationships between applications and resource objects of Huawei Cloud as well as other cloud vendors.
- CMDB functions:
 - Resource search: Users can search for resources (such as applications and hosts) by ID, keyword, or name.
 - Application management: CMDB manages the relationships between cloud services and applications (especially those running on ECS, CCE, and RDS).
 - Resource management: CMDB manages all cloud services of users in a unified manner. Users can view the relationships between applications and all cloud service resource objects (including those that have not been bound to applications) for resource analysis and management.
 - Environment tags: Users can add tags to application environments to filter environments with the same attribute.

Application Monitoring

- Application monitoring includes resource usage, trends, and application alarms in real time, so that users can quickly respond to exceptions and ensure that applications run smoothly.



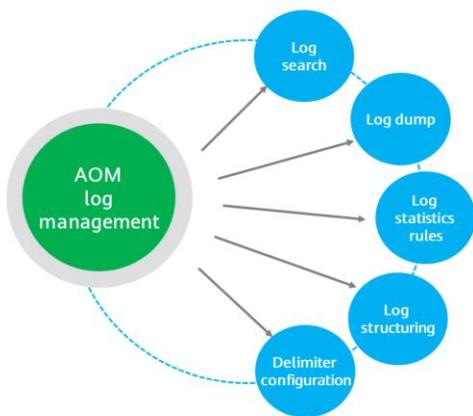
39 Huawei Confidential



- Application monitoring adopts the hierarchical drill-down design. The hierarchy is as follows: Application list > Application details > Component details > Instance details > Container details > Process details. That is, applications, components, instances, containers, and processes are associated and their relationships are directly displayed on the console.

Log Management

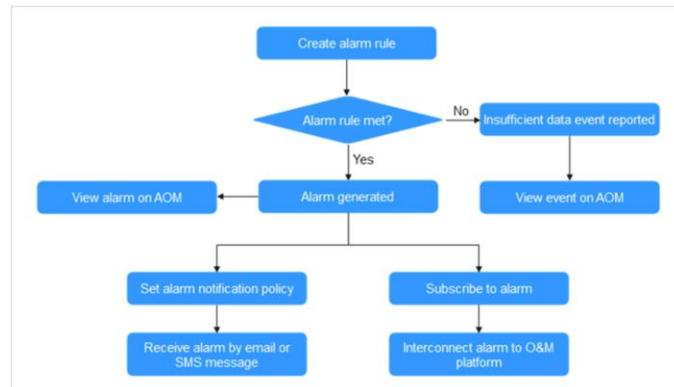
- AOM collects and displays logs of container services and cloud server services for search.



- Log search: AOM quickly searches vast numbers of logs to find the ones users need fast.
- Log dump: AOM dumps logs to OBS for long-term storage.
- Log statistics rules: AOM periodically counts keywords and generates metrics that users can use to monitor individual services or overall system performance in real time.
- Log structuring: Logs in log buckets are structured based on extraction rules. Similar logs or fixed-format logs are extracted and irrelevant logs are filtered out.
- Delimiter configuration: Users can separate log contents into multiple words by using delimiters, and then search for logs based on these words.

Alarm Management

- Alarms are reported when AOM or an external service is abnormal or may cause exceptions. Measures need to be taken accordingly. Otherwise, service exceptions may occur.



- The alarm center enables users to manage alarms and events. It supports custom notification actions, allowing users to obtain alarm information by email or SMS message. In this way, users can detect and handle exceptions at the earliest time. Before using the alarm management function, ensure that the ICAgent has been installed on the host.
- With a dashboard, different graphs can be displayed on the same screen. Various graphs, such as line graphs, digital graphs, and top N resource graphs allow users to comprehensively monitor resource data.
- Log search enables users to quickly search for required logs from massive quantities of logs. Log dump enables users to store logs for a long period of time. After users create statistical rules, AOM can periodically count keywords in logs and generate metric data, so that users can monitor system performance and services in real time. By configuring delimiters, users can divide log content into multiple words and use these words to search for logs.

Case: Using AOM for Routine Inspection and Fault Locating

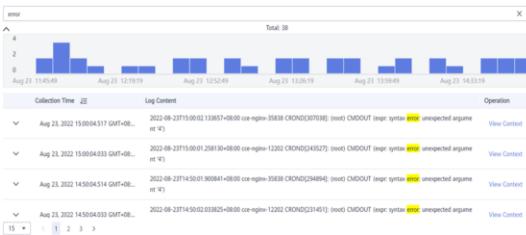
- Scenario:

The customer is a short video and livestream service provider. It needs to monitor resources and services in real time, and detect system exceptions through analysis of Nginx logs.

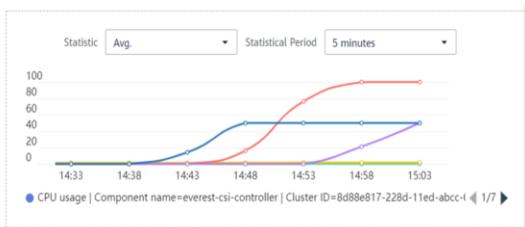
- Solution:

Use AOM to monitor the status, trends, and relationships of resources and services, set threshold rules for metrics, count exception keywords in Nginx logs, and report alarms when a threshold is exceeded.

1. Obtain exception keyword statistics.



2. Configure thresholds.



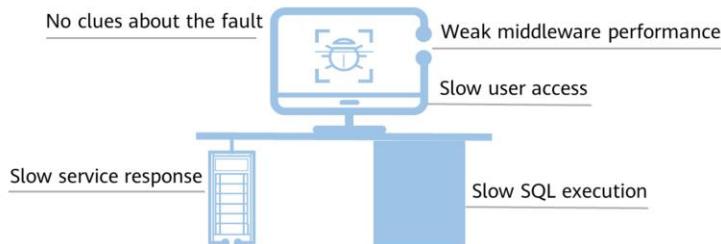
Contents

1. Cloud O&M Overview
2. Open Source O&M Tools
3. **Huawei Cloud O&M Services**

- Cloud Eye
- CTS
- LTS
- AOM
- **APM**
- CPTS

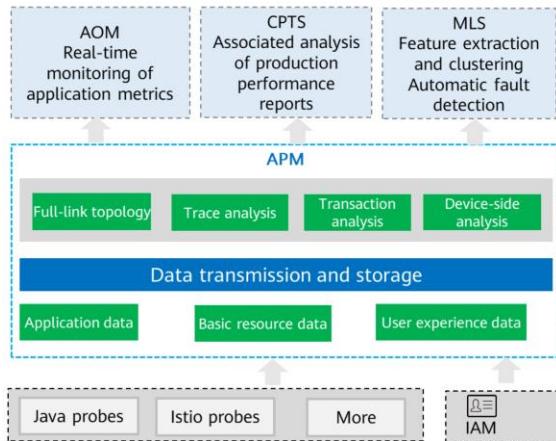
APM

- Application Performance Management (APM) monitors and manages the performance of cloud applications in real time. It provides performance analysis of distributed applications, helping O&M personnel quickly locate and resolve faults and performance bottlenecks.
- APM 2.0 consists of application and frontend monitoring. It manages the performance of distributed applications, container environments, browsers, applets, and apps. Full-stack performance monitoring and E2E full-link tracing/diagnosis make application management easy and efficient.



- In the cloud era, more and more applications are deployed in the distributed microservice architecture. As the number of users increases rapidly, many application exceptions occur. In traditional O&M, metrics cannot be associated for analysis, so they need manual and subjective processing. This results in low efficiency, high maintenance costs, and non-ideal performance.
- When there are massive quantities of services, O&M personnel face two major challenges:
 - Large distributed applications have complex relationships, making it difficult to analyze and locate problems. O&M personnel face problems such as how to ensure normal application running, and quickly locate faults and performance bottlenecks.
 - Users choose to leave due to poor experience. O&M personnel fail to detect and track services with poor experience in real time, and cannot quickly diagnose application exceptions, greatly affecting user experience.
- APM helps O&M personnel quickly identify application performance bottlenecks and locate root causes of faults, ensuring experience.

Service Architecture

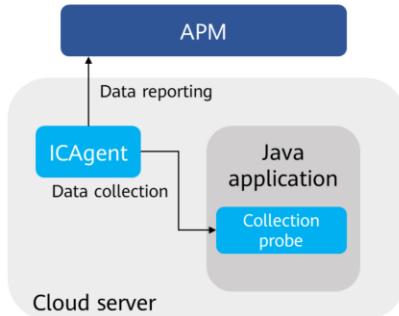


Main functions:

- Application metric monitoring: APM measures the overall health of applications. APM Agents collect metrics (such as JVM, service calls, and database access) from Java applications, to help users comprehensively monitor applications.
- Tracing: APM comprehensively monitors API calls, and displays service execution traces and statuses, helping users quickly demarcate performance bottlenecks and faults.
- Application topology: The topology displays the call relationship between services for a given period of time. Users can also view a graph of the call activity. The statistics can be collected from the calling or called party.
- Intelligent alarming: When an application connected to APM meets a preset alarm condition, an alarm is triggered and reported in a timely manner. In this way, users can quickly rectify faults and ensure services do not suffer.

APM Probes

- APM uses probes to collect application data. Probes use bytecode enhancement to trace applications and generate call data. ICAgents obtain and process call data, which is then reported to and displayed on APM.
- APM collects service tracing data, resource information, resource attributes, memory monitoring information, and call request metric data, but does not collect personal data.

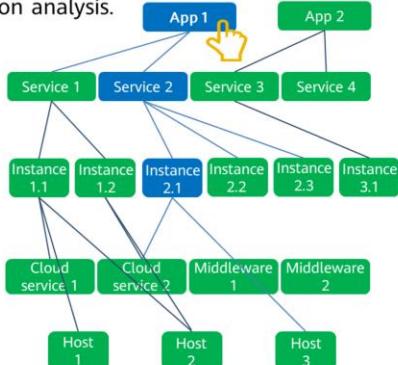


- APM probes are Java agents that collect application metrics in a non-intrusive way based on the Pinpoint open source project.
- APM probes inject the trace code into distributed transactions and performance information during class loading.

- APM probes inject the trace code into distributed transactions and performance information during class loading.
- APM transactions are HTTP transactions. When a user purchases a mobile phone from VMALL, the user's PC sends an HTTP request to the VMALL backend. This HTTP request is an HTTP transaction. As the HTTP request URL is unique, it is used as the transaction name. After a service (Java application) with a probe (pinpoint) deployed receives an HTTP transaction, APM extracts the transaction information and displays it on the console.

Application and Resource Association Analysis

- Users just need to install Agents for applications, and then APM can provide comprehensive monitoring. APM can quickly locate malfunctioning APIs and slow APIs, reproduce calling parameters, and detecting system bottlenecks to facilitate online diagnosis.
- APM is a cloud application diagnosis service that supports full-link topology display, tracing, and transaction analysis.



Application-centric, it connects services, instances, hosts, and middleware for analysis.

- Robust metrics: More than 200 different metrics and support for CNCF open source API standards such as those used for Prometheus and Zipkin.
- Association analysis: Metrics and alarm data from applications, services, instances, hosts, and transactions are associated for analysis.

- Full-link topology:
 - Visible topology: APM displays application call and dependency relationships in topologies. Application Performance Index (Apdex) is used to quantify user satisfaction with application performance. Different colors indicate different Apdex value ranges, helping users quickly detect and locate performance problems.
 - Inter-application calling: APM can display call relationships between application services on the topology. When services are called across applications, APM can collect inter-application call relationships and display application performance data.
 - SQL analysis: APM can count and display key metrics about databases or SQL statements on the topology.
 - JVM metric monitoring: APM can count and display JVM metric data of instances on the topology. APM monitors the memory and thread metrics in the JVM running environment in real time, enabling users to quickly detect memory leakage and thread exceptions.
- Tracing: APM comprehensively monitors calls and displays service execution traces and statuses, helping users quickly demarcate performance bottlenecks and faults.
 - In the displayed trace list, click the target trace to view its basic information.
 - On the trace details page, users can view the trace's complete information, including the local method stack and remote call relationships.
- Transaction analysis: APM analyzes service flows on servers in real time, displays key metrics (such as throughput, error rate, and latency) of transactions, and uses Apdex to evaluate users' satisfaction with applications. If transactions are abnormal, alarms are reported. For transactions with poor user experience, locate problems through topologies and tracing.

Transaction Session Monitoring

- Users can quickly obtain service statuses and diagnose abnormal applications.



No.	Transaction URL	Transaction Name	Throughput	Average Latency	Health Metric	Error Rate	Performance Analysis
1	hwid1.vmall.com/CAS/portal/login.html?validated=true	Login	52	323 ms	Normal	0%	Trace analysis
2	hwid1.vmall.com/portal/search?id=34211223411	Product search	234	721 ms	Slow	1%	Trace analysis
3	hwid1.vmall.com/portal/buy?id=34211223411	Purchase	3	432 ms	Normal	0%	Trace analysis
4	hwid1.vmall.com/portal/pay?id=34211223411	Payment	1	2.1s	Abnormal	100%	Trace analysis

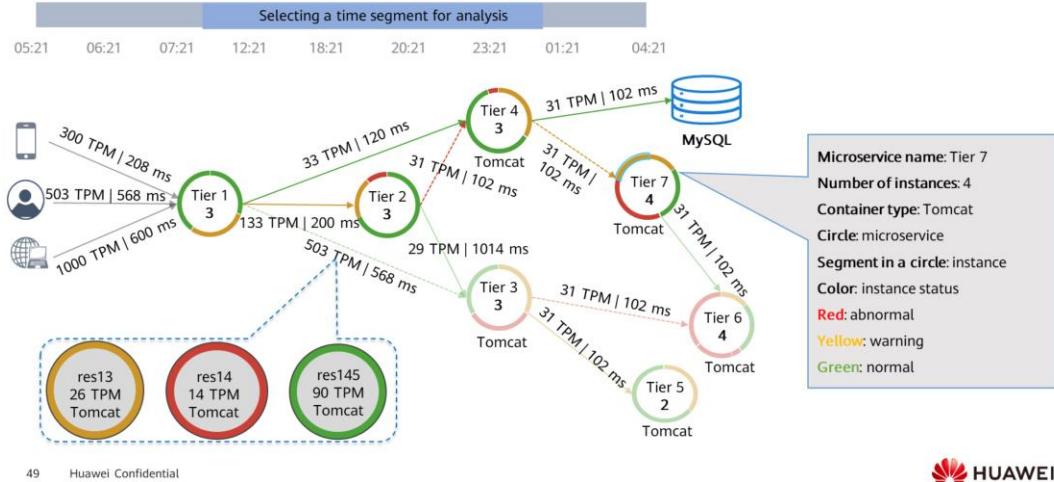
Key technologies

- Transaction customization:** Users can define transaction names based on URLs for convenience.
- Health rule matching:** Each transaction matches a health rule. If a transaction takes more than 1s to complete, an error message is displayed.
- Performance tracing:** APM accurately collects abnormal performance data and compares current data with historical baseline data to find application exceptions.

- APM traces each service transaction in real time, quickly analyzes the transaction status, and diagnoses problems.
 - Transaction customization:** Users can define transaction names based on URLs for better understanding.
 - Health rule configuration:** A health rule can be configured for each transaction. If the threshold is exceeded, an error message is displayed.
 - Performance tracing:** APM accurately collects abnormal performance data and compares current data with historical baseline data to find application exception methods and improve O&M efficiency.

Precise Fault Locating

- APM displays application relationships and exceptions clearly and locates faults precisely.

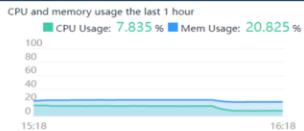


- Application discovery and dependency: APM monitors application metric data in a non-intrusive way and automatically generates dependencies through APIs between services.
- Application metric aggregation: Key metrics of microservice instances are automatically aggregated to applications.

Case: AOM (Monitoring) + APM (Locating) Solution

- Scenario:
The customer was using a microservice architecture and routine inspection and O&M was becoming burdensome. Because there were so many services, it was difficult to demarcate faults in a timely manner.
- Solution:
By using AOM and APM, the customer now can see the relationships between resources and services, search for error calls and drill down to locate root causes based on logs.

Routine inspection 1: Monitoring application resources



Routine inspection 2: Viewing alarm center information



50 Huawei Confidential

1. Check error calls using transaction APIs.

Search	Origin	Total Calls	Total Latency (ms)	Total Errors	Apdex	Apdex Threshold	Current Apdex T1	Operation
Transaction Type: GET_{product}searchAll	30	29	0	1	501	501	501	VMware-vcenter-service
Transaction Type: POST_{product}buy{id}	43	43	0	0.2	600	600	600	VMware-vcenter-service
Transaction Type: POST_{user}login	3	1023	0	0.5	500	500	500	VMware-vcenter-service

2. Drill down to abnormal traces and view their details to locate faults.



3. Search for logs on AOM to locate root causes.

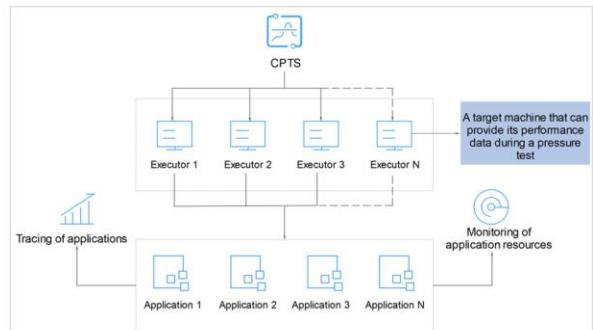


Contents

1. Cloud O&M Overview
2. Open Source O&M Tools
- 3. Huawei Cloud O&M Services**
 - Cloud Eye
 - CTS
 - LTS
 - AOM
 - APM
 - **CPTS**

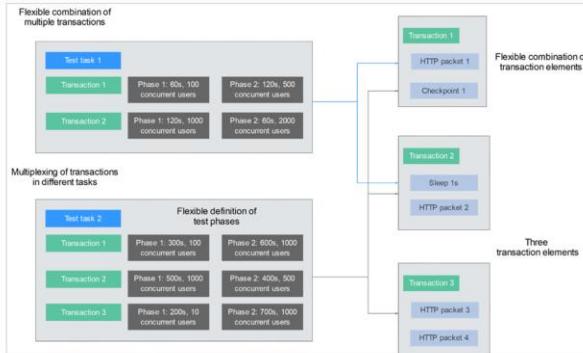
CPTS

- Cloud Performance Test Service (CPTS) provides performance testing services for cloud applications built based on HTTP, HTTPS, TCP, or UDP. CPTS performs rapid simulation of service peaks with up to millions of concurrent users. It allows users to define the contents and time sequences of packets and supports flexible combinations of multiple transactions for complex scenario testing.



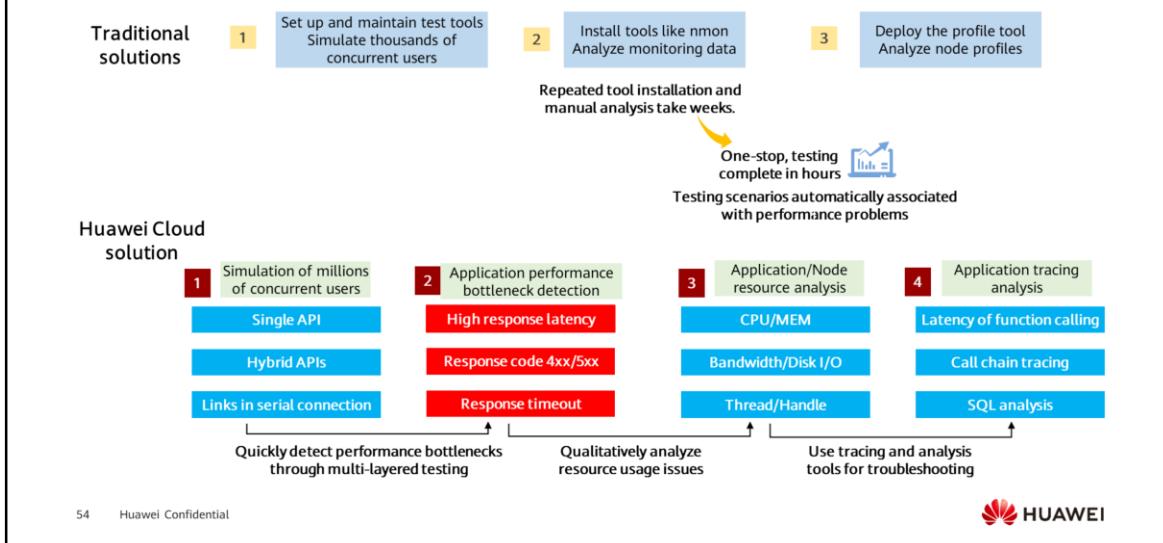
CPTS Functions

- CPTS provides tests for HTTP/HTTPS/TCP/UDP-based applications with high concurrency. It supports flexible definition of multi-protocol packet contents, transactions, and test task models. CPTS also allows users to view real-time performance statistics offline, such as concurrency, RPS, and response time. Users can also create private test clusters or scale in or out resource groups based on performance test scales.



- Multi-protocol and high-concurrency performance tests:
 - Users can quickly define standard HTTP, HTTPS, TCP, or UDP packet contents and simply adjust loads for different tested applications. CPTS allows users to define any fields in HTTP, HTTPS, TCP, or UDP packets based on their requirements.
 - Different behaviors of virtual users defined for different test scenarios: The number of requests initiated by each user per second can be set by think time, which is the interval for the same user to send, or by defining multiple request packets in a transaction.
 - Customizing the response result verification provides more accurate standards of successful requests. CPTS allows users to configure check points for requests. After obtaining response packets, CPTS verifies their response code and header. Only response packets meeting the specified conditions are regarded as normal.
- Test task model customized for complex scenarios:
 - With multiple flexible combinations of transaction elements and test task phases, CPTS helps users test application performance in scenarios with different operation types and concurrent operations.
 - A transaction can be used by multiple test tasks, and multiple test phases can be defined for a transaction. In each test phase, users can define the test duration, number of concurrent users and tests, as well as simulate complex scenarios with different traffic peaks and troughs.
- Two types of costs will be generated when users run performance tests in CPTS: the cost for using CPTS and the cost for using resources of other cloud services, such as ECS. CPTS is billed by package on a pay-per-use or yearly/monthly basis.

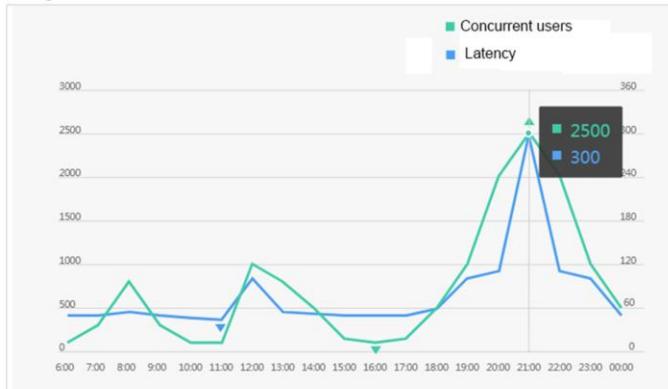
One-Stop Cloud Performance Test Solutions



- Engines for millions of concurrent users and capability of full-link bottleneck analysis accelerate testing from weeks to just hours.

Application Scenarios: Traffic Surge Testing

- Gaming and e-commerce services have obvious peaks and troughs and need to be scalable. During traffic surges, the scalability and various KPIs, like network latency during traffic surges, need to be tested.



55 Huawei Confidential

- Multi-scenario simulation: CPTS simulates real-world scenarios by combining multiple transactions, which include diverse elements, and customizable packets.
- Peak and trough simulation: CPTS develops a pressure test curve for each transaction within a defined period to simulate traffic peaks and troughs.
- KPI measurement: Users can verify game KPIs during traffic surges based on a custom response timeout interval.



Quiz

1. (True or false) LTS allows users to transfer logs only to Object Storage Service (OBS).
 - A. True
 - B. False
2. (Single-answer question) Which of the following statements about Cloud Eye is false?
 - A. It checks site health.
 - B. It provides different types of alarm notifications.
 - C. It tests cloud application performance.
 - D. It monitors cloud resources.

- 1. False
 - LTS allows user to transfer logs to OBS and DIS.
- 2. C
 - Cloud Eye does not support performance tests.

Quiz

1. (Discussion) What are the differences between cloud-based O&M and traditional O&M?

2. (Discussion) Can Cloud Eye monitor ECSs without Agents installed? What are the differences in the monitoring metrics of Cloud Eye with and without Agents?

- Discussion 1: The differences lie in the sites, equipment, routine inspections, troubleshooting, and software tools.
- Discussion 2: Basic monitoring can be performed without Agents. Data is collected every 5 minutes. With Agents installed, Cloud Eye can provide advanced monitoring. For example, system-level, proactive, and fine-grained monitoring is provided, and data is collected every minute. Host processes cannot be monitored unless Agents are installed.

Summary

- This course describes O&M, including open source O&M tools and Huawei Cloud O&M services (including Cloud Eye, CTS, LTS, AOM, APM, and CPTS). Upon completion of this section, you are expected to understand the Huawei Cloud O&M work routine and be able to use O&M tools to optimize services.

Acronyms and Abbreviations

- AOM: Application Operations Management
- AOS: Application Orchestration Service
- API: Application Programming Interface
- APM: Application Performance Management
- AS: Auto Scaling
- BMS: Bare Metal Server
- CCE: Cloud Container Engine
- CCI: Cloud Container Instance
- CI/CD: Continuous Integration/Continuous Delivery
- CNCF: Cloud Native Computing Foundation
- DDoS: Distributed Denial of Service
- DevOps: Development and Operations
- DataArts Studio: Data Lake Governance Center
- DIS: Data Ingestion Service
- DLI: Data Lake Insight
- DNS: Domain Name Service
- ECS: Elastic Cloud Server
- EIP: Elastic IP
- ELB: Elastic Load Balance
- EVS: Elastic Volume Service

Acronyms and Abbreviations

- GSLB: Global Server Load Balance
- HA: high availability
- IAM: Identity and Access Management
- IDC: Internet data center
- IMS: Image Management Service
- ISV: independent software vendor
- MCP: Multi-Cloud Container Platform
- MRS: MapReduce Service
- NAT: Network Address Translation
- OBS: Object Storage Service
- OCI: Open Container Initiative
- OCR: Optical Character Recognition
- RDS: Relational Database Service
- SMN: Simple Message Notification
- SWR: SoftWare Repository for Container
- VM: virtual machine
- VPC: Virtual Private Cloud
- VPN: Virtual Private Network
- WAF: Web Application Firewall

Recommendations

- Huawei iLearning:
 - <https://e.huawei.com/en/talent/portal/#/>
- Huawei Cloud Help Center
 - <https://support.huaweicloud.com/intl/en-us/index.html>
- HUAWEI CLOUD Developer Institute
 - <https://edu.huaweicloud.com/intl/en-us/>
- Huawei Talent Online
 - <https://e.huawei.com/en/talent/portal/#/>

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©2022 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.



Huawei Cloud Innovations and Solutions



Foreword

- With the development of cloud technologies, companies have been paying more attention to digital transformation. Many are aiming to innovate their services, and are seeking better development paths. Cloud service vendors provide abundant PaaS and SaaS resources to help companies achieve digital transformation.
- This course describes Huawei Cloud innovations and solutions in enterprise intelligence (EI), IoT, and application data development.

Objectives

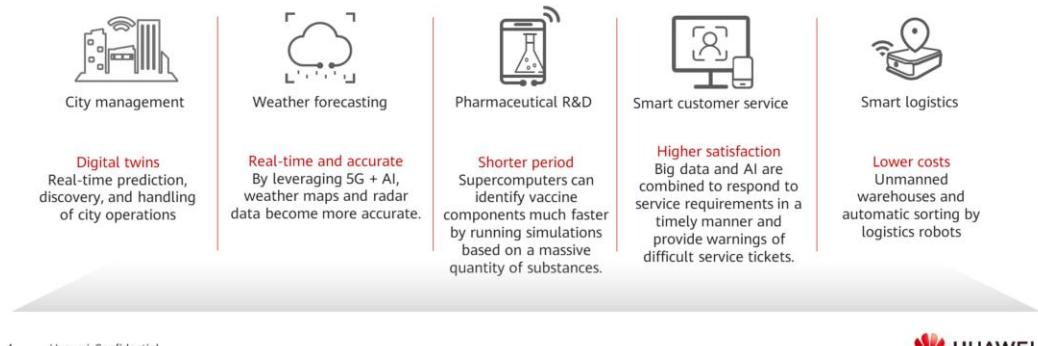
- Upon completion of this course, you will:
 - Have a working knowledge of Huawei Cloud's innovative products and solutions in mainstream application scenarios, such as EI and IoT, and be able to identify the solutions that best suit your needs.

Contents

- 1. Overview of Huawei Cloud Innovations and Solutions**
2. Huawei Cloud EI Solution
3. Huawei Cloud IoT Solution
4. Huawei Cloud Application and Data Solution

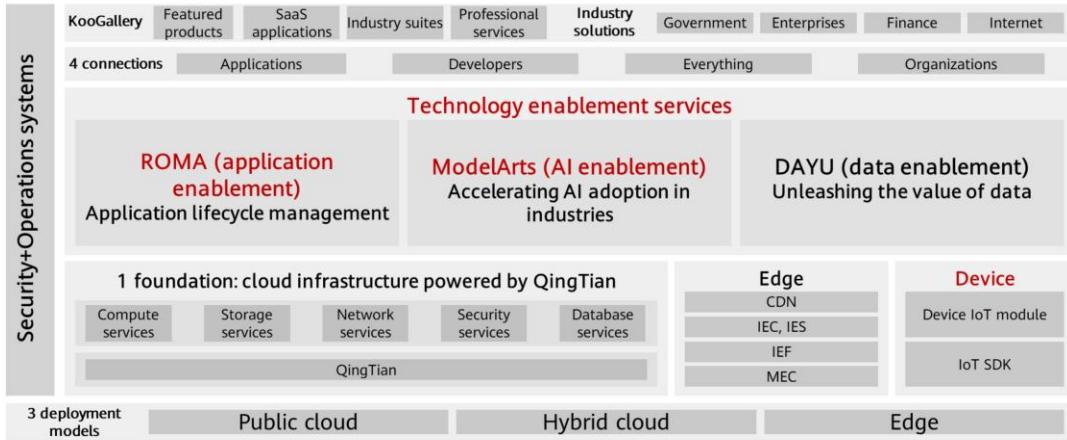
Digital Transformation: Companies' Core Concerns

- With the rise of next-gen information technologies, companies have been generating ever-increasing amounts of data. If that data can be digitally processed and used in intelligent application systems, then service operations, products, and services can be innovated, and user experience improved, which ultimately helps companies create a new competitive edge.



- As companies' digital construction gradually enters the intelligent upgrade phase, companies need to fully enjoy the dividends brought by cloud computing. The value of cloud to services is no longer simple resource provision, but also application-centric for service enablement.
- Digital twins: Fully utilize the simulation process and completes mapping in the virtual space to reflect the entire lifecycle of the corresponding physical equipment, effectively reducing the actual production cost.

Huawei Cloud Full-Stack Technologies Help Enterprises Go Digital



Contents

1. Overview of Huawei Cloud Innovations and Solutions
- 2. Huawei Cloud EI Solution**
3. Huawei Cloud IoT Solution
4. Huawei Cloud Application and Data Solution

Enterprise Intelligence (EI)

EI allows enterprises to utilize intelligent technologies to process massive amounts of data, and extract value from that data, to enable business development.



EI application scenarios

Improve efficiency by eliminating repetitive operations



Customer service
(intelligent Q&A)



Video surveillance
(automatic detection)

Employ experts with professional knowledge



Automatic design
(knowledge management)



Medical review
(image recognition)

Institute multi-domain collaboration



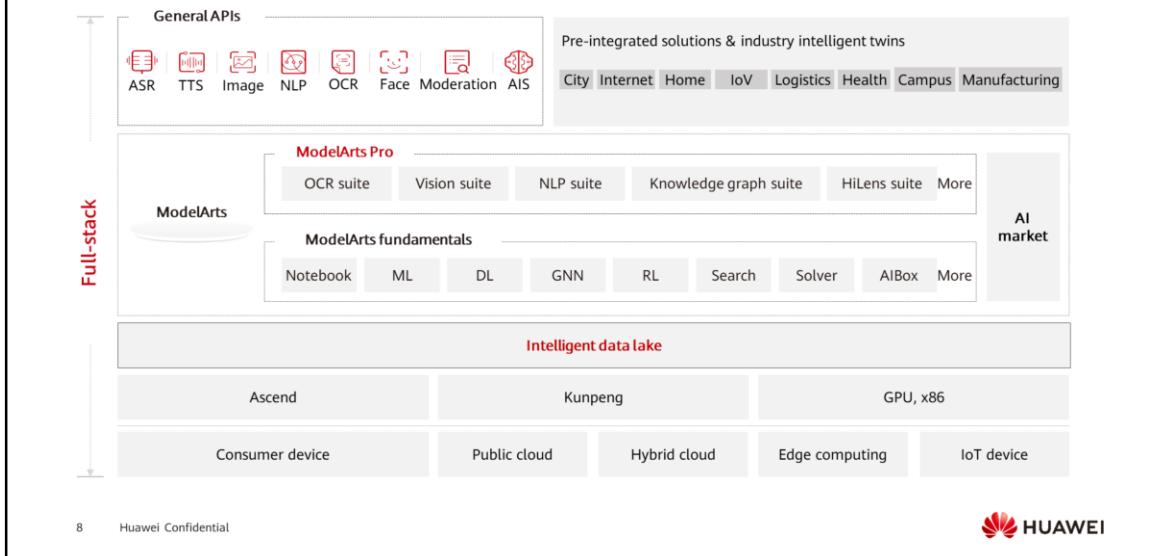
City governance
(smart scheduling)



Industrial production
(intelligent control)

- Huawei Cloud EI consists of big data and AI solutions.

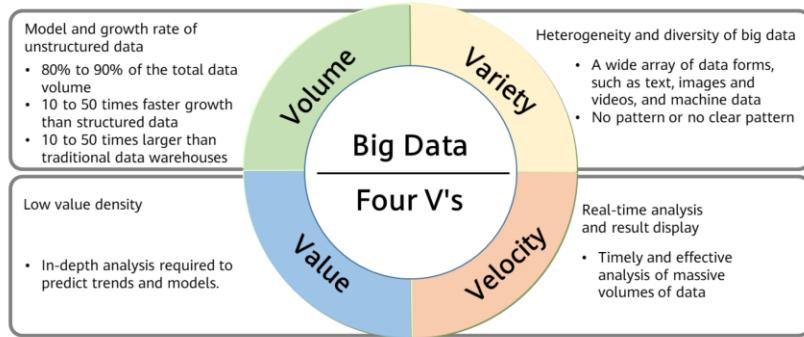
Huawei Cloud EI Solution: Full-Stack, All-Scenario



- The content in red will be further learned.
- Huawei Cloud DAYU:
 - It is dedicated to transforming enterprise data from resources to assets. By importing data from all domains into a lake, data can be transmitted across isolated systems, service awareness can be implemented, and data resources can be intelligently managed. Data value is mined from multiple perspectives, layers, and granularities to implement data-driven digital transformation.
- Huawei Cloud AI-enabled ModelArts:
 - ModelArts: a one-stop AI development platform for developers
- Training framework:
 - MindSpore is an open-source AI framework developed by Huawei. It is a deep learning training and inference framework that supports all-scenario device-edge-cloud scenarios and is mainly applied to AI fields such as computer vision and natural language processing.
- Computing power:
 - NPU: a new type of processor based on neural network algorithms and acceleration.

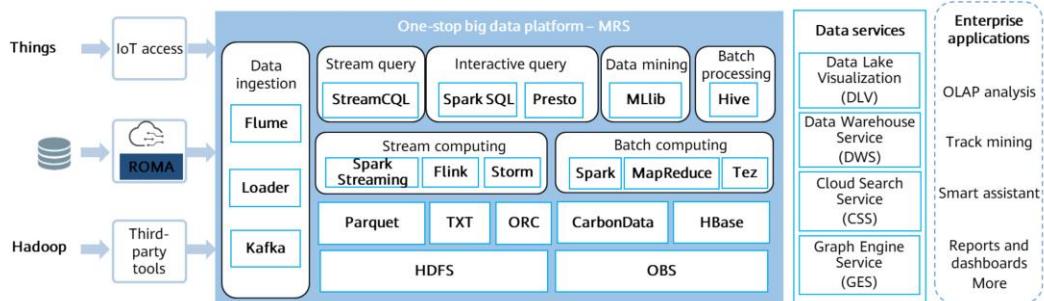
What Is Big Data?

- McKinsey: Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. Generally, the size ranges from several terabytes to petabytes.
- National Institute of Standards and Technology (NIST), USA: Big Data means the data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies.



MRS

- MapReduce Service (MRS) provides Hadoop-based high-performance big data components, such as Hudi, ClickHouse, Spark, Flink, Kafka, and HBase for data lakes, data warehouses, business intelligence (BI), AI, and more.



- MRS helps customers build a unified big data platform for data access, data storage, data analysis, and value mining. Furthermore, it interconnects with Huawei Cloud IoT, ROMA platform, DLF, and DLV to help customers easily resolve difficulties in data channel cloudification, big data job development and scheduling, and data display.
- MRS provides different big data analysis and processing components for different scenarios. You can select stream computing components such as Flink for real-time processing, and Spark or MapReduce for offline batch computing.
- CarbonData is a new local file format of Apache Hadoop. It uses advanced column-based storage, indexing, compression, and encoding technologies to improve computing efficiency and accelerate PB-level data query. Therefore, it can be used for faster interaction query. CarbonData is also a high-performance analysis engine that integrates data sources with Spark.

MRS Application Scenarios

Massive data analysis

Typically, an enterprise has multiple data sources. After data sources are connected, data is extracted, transformed, and loaded to generate modeled data for each service module to analyze and sort out data. This type of service has the following characteristics:
Data can be stored in OBS and periodically dumped to HDFS for batch analysis. In the environmental protection industry, MRS can analyze up to 10 TB weather data within one hour.

Massive data storage

After a user has a large amount of structured data, the index-based quasi-real-time query capability is required. MRS can respond to massive data queries within seconds.
In the IoV scenario, users need to query vehicle maintenance information based on the vehicle ID. However, such data may have been stored for one to three years, and the data volume is large. MRS supports HBase. Vehicle information is indexed based on vehicle IDs, achieving second-level query response.

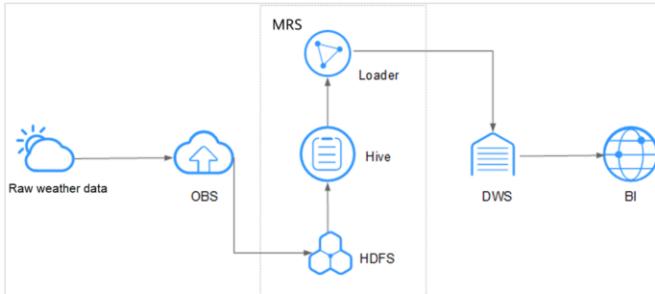
Real-time data processing

Real-time data processing is typically used in scenarios such as anomaly detection, fraud detection, rule-based alarms, and service process monitoring. Data is processed while being input. MRS can implement real-time ingestion of massive amounts of data. In the Internet of Elevators (IoE) industry, data of tens of thousands of smart elevators and escalators is imported to MRS streaming clusters in real time, facilitating real-time alarm reporting.

- Advantages of MRS in massive data analysis scenarios (environmental protection industry):
 - Low costs: Enjoy the cost-effective storage of OBS.
 - Analysis of mass data: Analyze TB or PB of data with Hive.
 - Visualized data import and export tool: Use Loader to export data to Data Warehouse Service (DWS) for business intelligence (BI) analysis.
- Advantages of MRS in massive data storage scenarios (IoV industry):
 - Real-time: With Kafka, you can access massive amounts of vehicle messages in real time.
 - Storage of mass data: With HBase, you can store a large volume of data and query data in milliseconds.
 - Distributed data query: With Spark, you can analyze and query a large volume of data.
- Advantages of MRS in real-time data processing scenarios (elevator industry):
 - Real-time data ingestion: With Flume, you can achieve real-time data ingestion and enjoy various data collection and storage access methods.
 - Data source access: Use Kafka to access the data of tens of thousands of elevators and escalators in real time.

Large-Scale Data Analysis in the Environmental Protection Industry

- **Background and pain points:** In the environmental protection industry, a large amount of weather data is analyzed to learn about the distribution and orientation of pollutants in an area. Pollution source information is often obtained from multiple sources and has many formats, involving massive data analysis and complex processing.

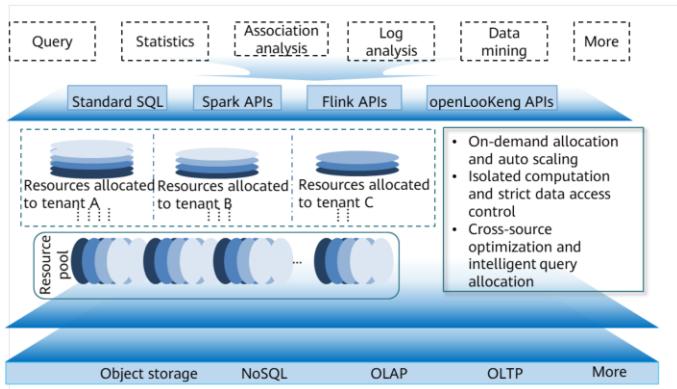


- **MRS advantages:**
 - **Low costs:** Enjoy the cost-effective storage of OBS.
 - **Analysis of mass data:** Analyze TB or PB of data with Hive.
 - **Visualized data import and export tool:** Use Loader to export data to Data Warehouse Service (DWS) for business intelligence (BI) analysis.
- **MRS can analyze up to 10 TB weather data in one hour.**

- Weather data can be stored in OBS and periodically dumped to HDFS for batch analysis.

Data Lake Insight (DLI)

- DLI is a serverless big data computing and analysis service that is fully compatible with the Apache Flink, Apache Spark, and openLooKeng (based on Apache Presto) and integrates batch data, streaming data, and interactive analysis.

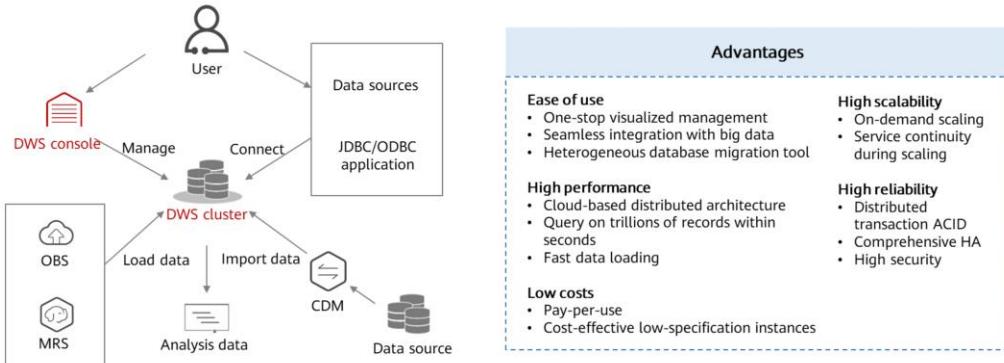


- **SQL queries with AI:** You do not need to have a background in big data to do big data analysis. If you know SQL, you are good to go. The SQL syntax is fully compatible with the standard ANSI SQL 2003.
- **Serverless Spark/Flink:** Seamlessly migrate your offline applications to the cloud with serverless technology. DLI is fully compatible with Apache Spark, Apache Flink, and Presto ecosystems and APIs. A set if your data is available for both streaming and batch processing for diverse computing types.
- **Cross-source analysis:** Analyze your data across databases. No migration required. A unified view of your data gives you a comprehensive understanding of your data and helps you innovate faster. There are no restrictions on data formats, cloud data sources, or whether the database is created online or off.
- **Enterprise multi-tenant:** Manage compute or resource related permissions by project or by user. Enjoy fine-grained control that makes it easy to maintain data independence for separate tasks.

- DLI frees you from managing any servers. DLI supports standard SQL and is compatible with standard SQL and Spark and Flink SQL. It also supports multiple access modes and mainstream data formats. You can use SQL applications to query mainstream data formats without data ETL. DLI supports SQL statements for heterogeneous data sources, including CloudTable, RDS, DWS, CSS, OBS, custom databases on ECSs, and offline databases.
- DLI is applicable to large-scale log analysis, federated analysis of heterogeneous data sources, and big data ETL processing.

GaussDB(DWS)

- GaussDB Data Warehouse Service (DWS) is an online data processing database using the public cloud infrastructure to provide scalable, fully-managed analytic databases that are immediately available upon purchase, freeing you from database management and monitoring.



- DWS is a cloud-native service based on Huawei converged data warehouse GaussDB. Based on the shared-nothing distributed architecture, GaussDB(DWS) uses a massively parallel processing (MPP) engine and consists of multiple independent logical nodes that do not share system resources, such as CPUs, memory, and storage. In such an architecture, data is distributed on multiple nodes. Data analytics tasks can be quickly executed in parallel on the nodes where data is stored.
- DWS provides a web-based service management platform, that is, the management console. You can also manage DWS clusters using HTTPS-based APIs.
- DWS is often used together with Cloud Data Migration (CDM) and Data Ingestion Service (DIS). CDM is used for batch data migration, and DIS is used for stream data ingestion.

GaussDB(DWS) Application Scenarios

- GaussDB(DWS) is compatible with ANSI SQL-99 and SQL 2003, and with PostgreSQL and Oracle database ecosystems, providing competitive solutions for PB-level big data analysis in diverse industries.

Data warehouse migration	Converged big data analysis	Enhanced ETL + Real-time BI analysis	Real-time analytics
<p>Data warehouses are important for corporate data analytics. As the business volume grows, customer-built data warehouses no longer suffice – due to low scalability and high costs. As an enterprise-grade data warehouse on the cloud, GaussDB(DWS) features high performance, low costs, and high scalability, supporting corporate business growth in the big data era.</p>	<p>Predictive analysis heavily depends on the technologies to consolidate data and explore value from massive amounts of data. GaussDB(DWS) has scalable data storage, query, and analysis capabilities, able to store and analyze up to petabytes of data. Analysts can obtain information from the big data platform in real time.</p>	<p>A company often has multiple data sources and needs to perform extract, transform, load (ETL) and BI analysis on them, generating model-based data for each service module. DWS supports diverse data sources and can efficiently import data in batches in real time. It can store petabytes of at a low cost and respond to correlation analysis on trillions of data records within seconds.</p>	<p>The mobile Internet and IoT generate a large amount of data every second, which has to be processed in real time. The strong data import and query capabilities of DWS support real-time data analysis.</p>

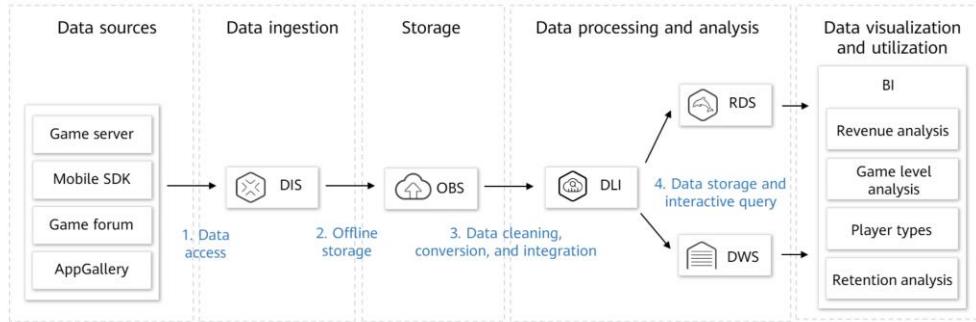
Massive Gaming Operations Data Analysis (1)

- **Background and pain points:** A gaming company had a large number of data sources that were difficult to access or join for query. Gaming data analysis usually had a latency of T+1. Data generated in the previous day was calculated only in the early morning, and many compute resources were idle for the most of the time in a day. Sensitive user data was used for different purposes, such as operations, notification push, and planning. The company needed to design permissions to allow specific department roles to view only the part of the data they require.



- **Solution:**
 - **Data access:** DIS provides various access tools, such as Agent, Flume Plugin, and Logstash Plugin, to access different data sources. DIS is compatible with open source Kafka APIs.
 - **Data storage:** Data is stored offline to OBS, connecting data silos and reducing data storage costs.
 - **Data processing and analysis**
 - DLI-based Spark cleans and converts gaming data in OBS, and integrates and associates logs with database data for modeling analysis, including retention analysis and funnel analysis. DLI also supports IAM fine-grained authorization, facilitating permissions management.
 - DLI writes common data to RDS and writes complex data, such as OLAP data, to DWS. DLI works with BI for visualized processing and data utilization.

Massive Gaming Operations Data Analysis (2)



Advantages:

- **Low costs, higher efficiency:** OBS and DLI decouple storage and computing resources. The increase in data volume does not lead to an increase in computing costs. Computing and storage can be used on demand, reducing the total costs of the big data platform.
- **Ease of use:** Data can be collected without programming by configuring SDKs/Agents. The platform is compatible with standard SparkSQL, and data analysts can easily get started without learning from scratch.
- **Easy O&M:** DLI uses the serverless architecture. Data analysts can be freed from O&M workloads and focus on their jobs.

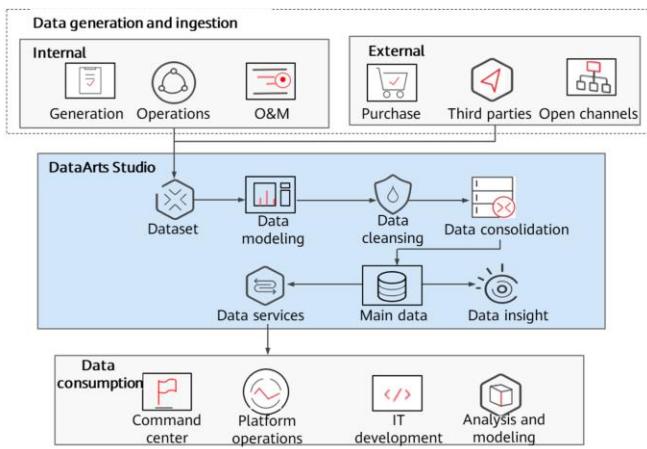
DataArts Studio

- DataArts Studio is a one-stop development and operations platform that provides data lifecycle management. It provides data integration, data development, data governance, and data services. It also supports intelligent construction of industrial knowledge libraries and incorporates data foundations such as big data storage, computing, and analytics engines, helping enterprises quickly build data operations capabilities.



- DataArts Migration: Based on the big data cloud migration and intelligent data lake solution, DataArts Migration provides easy-to-use migration capabilities and can integrate a broad set of data sources into the data lake more easily and efficiently.
- DataArts Architecture can be used to create entity-relationship (ER) models and dimensional models to standardize and visualize data development and output data governance methods that can guide development personnel to work with ease.
- DataArts Factory is a one-stop collaborative big data development platform that provides fully managed big data scheduling capabilities.
- DataArts Quality can monitor metrics and data quality, and screen out unqualified data in a timely manner.
- DataArts Catalog provides enterprise-class metadata management to clarify information assets. It uses a data map to display a data lineage and panorama of data assets for intelligent data search, operations, and monitoring.
- DataArts DataService enables you to manage APIs centrally and control the access to subjects, profiles, and metrics. It improves data access, query, and retrieval efficiency and data consumption experience, and monetizes data assets. It also allows you to quickly generate new APIs based on data tables, register your legacy APIs, and centrally manage and publish them.
- DataArts Security provides all-round security assurance to safeguard network security and control user permissions. It provides a review mechanism for key processes in DataArts Architecture and DataArts DataService. Data is managed by level and category throughout the lifecycle, ensuring data privacy compliance and traceability.

DataArts Studio Application Scenarios



One-stop data operations and governance platform

DataArts Studio provides one-stop intelligent data operations — covering data collection, standard design, quality monitoring, data cleansing, data modeling, data connection, data consolidation, data consumption, and intelligent analysis — and helps enterprises quickly build data operations capabilities.

Quick building of cloud data platforms

DataArts Studio allows you to migrate offline data to the cloud and integrate the data into big data services. On the DataArts Studio console, you can use the integrated data to quickly develop jobs and easily build a data system.

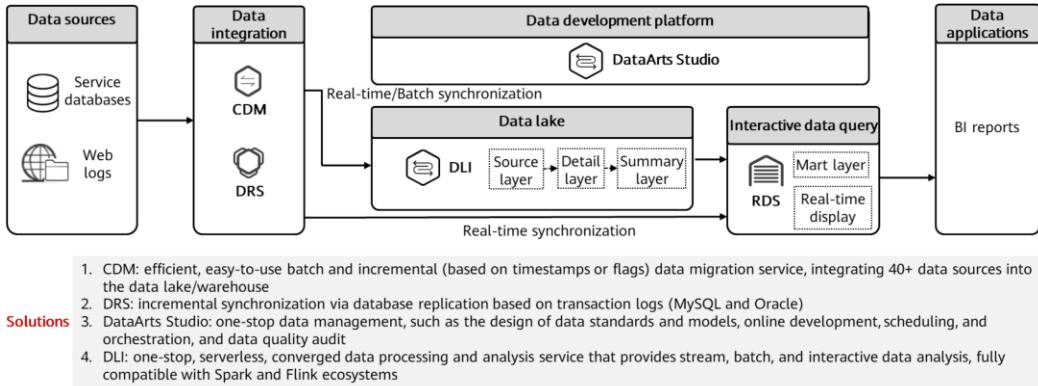
Quick building of data mid-ends

Huawei's models and algorithms accumulated in the enterprise business field help enterprises build data mid-ends, improving their data operations capabilities. Data mid-ends apply to a wide range of industries, such as government, taxation, smart city, smart transportation, and smart campus.



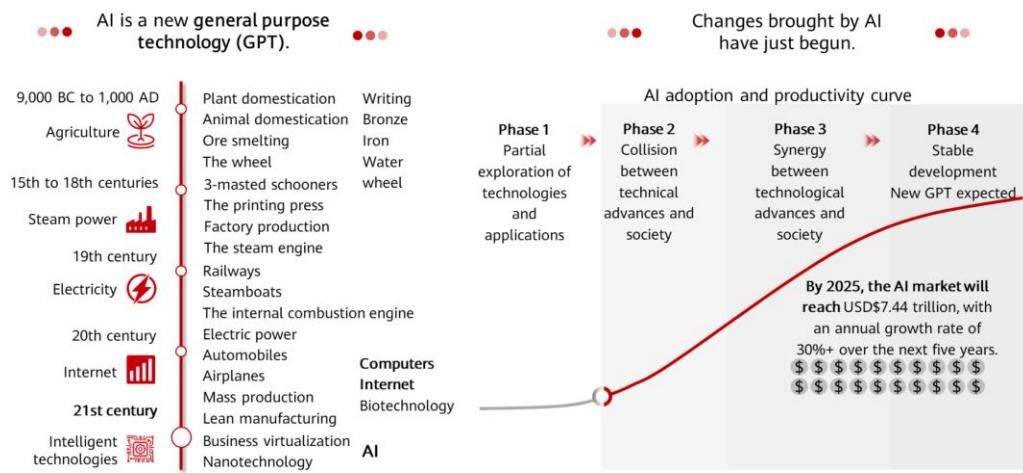
DataArts Studio Application — Medium- and Long-Tail Enterprise Data Processing

- Background and pain points: Medium- and long-tail enterprises need data technologies, talent, and capital for digital transformation. A universal digitalization method can reduce costs and lower the barrier to data use for enterprises.



- The long tail is a business strategy that allows companies to realize significant profits by selling low volumes of hard-to-find items to many customers, instead of only selling large volumes of a reduced number of popular items.
- To achieve digital transformation, medium- and long-tail enterprises need advanced data technologies, professionals, and large amounts of capital investment. Therefore, they urgently need a universal model offered by a leader in the big data industry to reduce digitalization costs and lower the barrier to data use.
- Based on Huawei's IT process data governance methodology, Huawei Cloud launched a lightweight big data solution. This Serverless solution uses Huawei Cloud assets to enable quick data governance, requiring less resources and development, deployment, and O&M workloads. It frees medium- and long-tail enterprises from worrying about technology stacks and cloud resources, and allows them to use resources on demand, reducing operational costs.
- Huawei Cloud big data services provide one-stop management and development throughout the entire data lifecycle and significantly simplify the data governance process for medium- and long-tail enterprises. With these services, medium- and long-tail enterprises can analyze a large amount of data more quickly and efficiently, use data more easily, monetize data in a shorter time, and digitize their business smoothly.

AI Development History

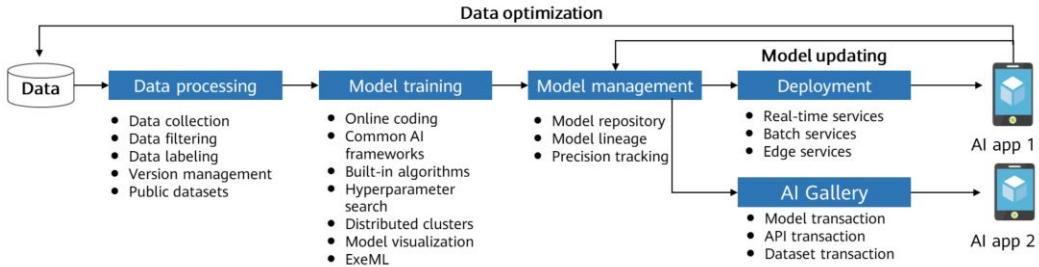


21 Huawei Confidential



- As big data has grown, there has been a corresponding growth in the power of AI. AI has been constantly changing methods of production and how we live.

ModelArts: A One-Stop AI Development Platform

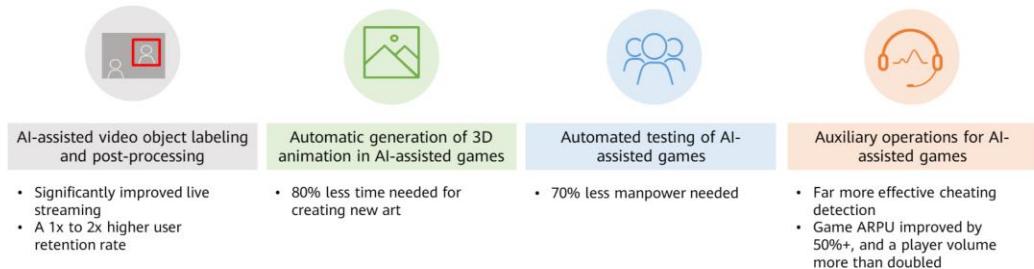


- AI engineers face many challenges when they are installing and configuring various AI tools, preparing data, and training models. ModelArts, a one-stop AI development platform is designed to address these challenges. ModelArts integrates data preparation, algorithm development, model training, and model deployment into the production environment, allowing AI engineers to perform one-stop AI development.
- ModelArts supports the entire development process, including data processing, and model training, management, and deployment. It also includes AI Gallery, a place where models can be shared.
- Data processing: All data formats are supported, as well as team labeling.
- Training: Pre-trained models accelerate the implementation of AI applications. Huawei-developed inference frameworks hide underlying hardware and software differences from the upper layer software to improve performance. Multi-vendor, multi-framework, and multi-function models are centrally managed.
- Deployment: High-concurrency model deployment, low-latency access, auto scaling, grayscale release, and rolling upgrade are provided. Models can be deployed in different production environments, for example, deployed as in-cloud real-time or batch inference services, or on devices and edge devices.
- AI Gallery: Common algorithms are preconfigured in AI Gallery, and models can be shared publicly or within an enterprise.

Model Training for Games (1)

- **Pain points:** Intelligent gaming is the future of the industry. However, AI algorithms are too complex to be handled by many gaming companies. Massive expensive AI computing power is required, and extensive investment into AI would drive up OPEX.

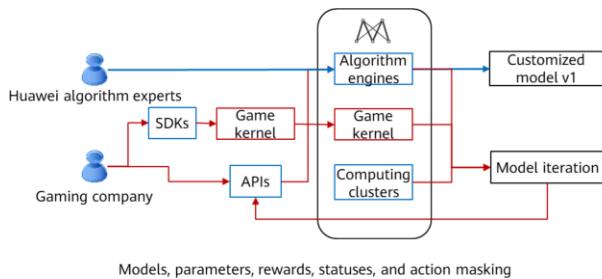
AI-empowered games



- ARPU = Total revenue/Number of active users

Model Training for Games (2)

- Huawei Cloud ModelArts makes AI work fast and easy.

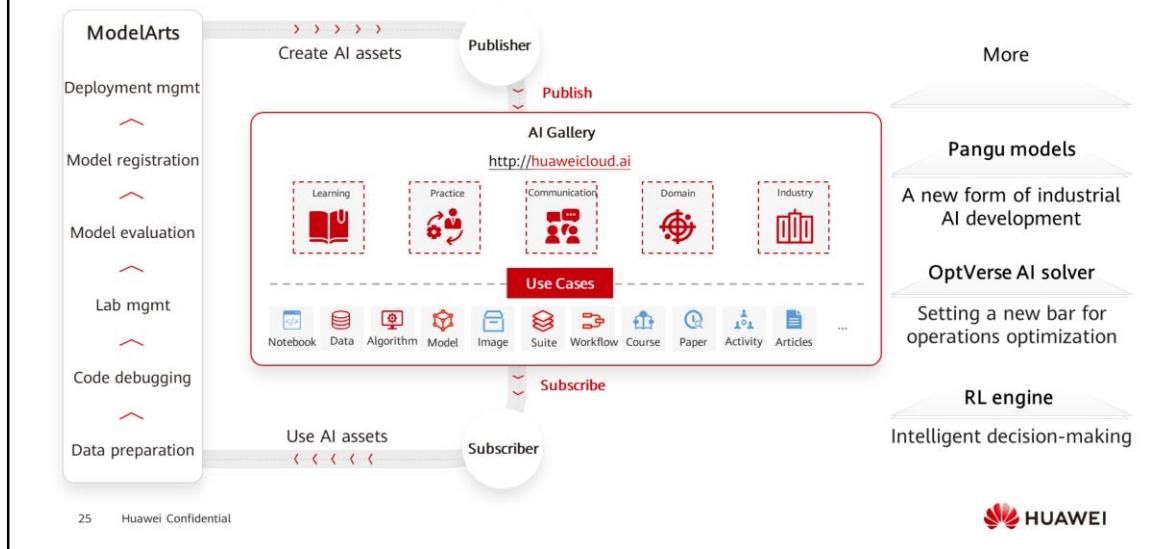


Highlights of Huawei Cloud AI solution

- Over 10 full-code open AI tutorials
- End-to-end AI solution for popular game types
- On-demand, cost-effective heterogeneous clusters with thousands of cores
- One-stop development and end-to-end inference service deployment in one click

- Huawei algorithm experts provide resources such as algorithm engines, SDKs, and APIs for gaming companies to call for training. Furthermore, models can be customized on Huawei Cloud to significantly improve algorithm development efficiency.

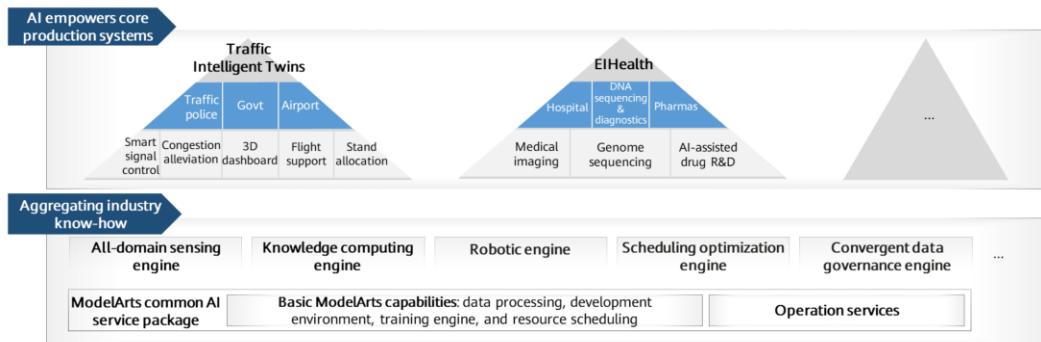
AI Gallery: An AI Development Community



- AI Gallery is a developer ecosystem community built on ModelArts. In this community, scientific research institutions, AI application developers, solution integrators, enterprises, and individual developers can share and purchase algorithms, models, and datasets. This accelerates the development and implementation of AI assets and enables every participant to create business value in the AI development ecosystem.
- Pangu models: There are multiple foundation models, including the NLP, CV, multi-modal, and scientific computing models. Through model generalization, the Pangu models enable large-scale industrialized AI that could not be supported in traditional AI development. This enables brand-new industrial AI development.
- OptVerse AI solver: integrates AI with operations research to break through the optimization limit of operations research in the industry and find the optimal solution for linear and integer models, helping enterprises make quantitative decisions and refine their operations.

Intelligent Twins: Pre-integrated Intelligent Solutions

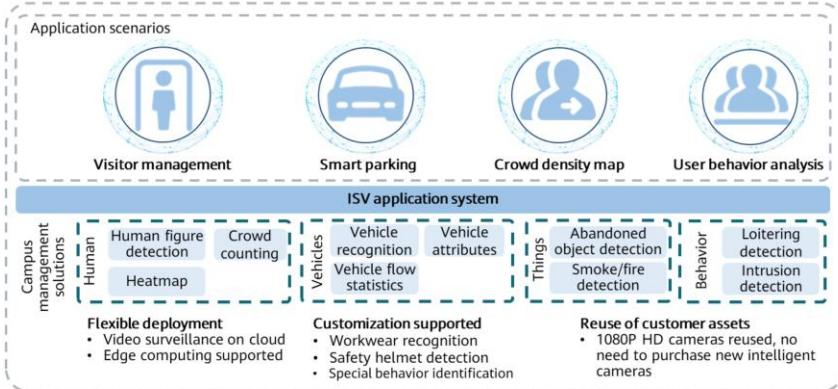
- AI has become the new engine powering the evolution of Smart Cities. In aggregating the capabilities of Huawei and its partners, Huawei Cloud offers a platform for application enablement, data enablement, and full-stack, all-scenario AI services. This platform allows us to build intelligent applications tailored to the needs of specific scenarios based on a unified city data foundation.



- Government Intelligent Twins, Traffic Intelligent Twins, EIHealth, GeoGenius, Campus Intelligent Twins, Water Intelligent Twins, Heating Intelligent Twins, Industrial Intelligent Twins, Network Intelligent Twins

Campus Intelligent Twins

- Pain points in campus management: high labor costs that come with a large staff, slow coordination, slow response with manual dispatching, and slow situation awareness and warnings



Solution:

- Campus Intelligent Twins integrates AI and big data capabilities to provide customizable intelligent solutions.
- Image processing techniques are used for admission management, and big data is used for visitor management and tracking.

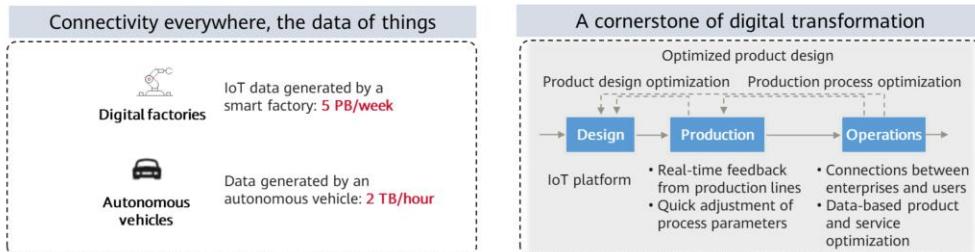


Contents

1. Overview of Huawei Cloud Innovations and Solutions
2. Huawei Cloud EI Solution
- 3. Huawei Cloud IoT Solution**
4. Huawei Cloud Application and Data Solution

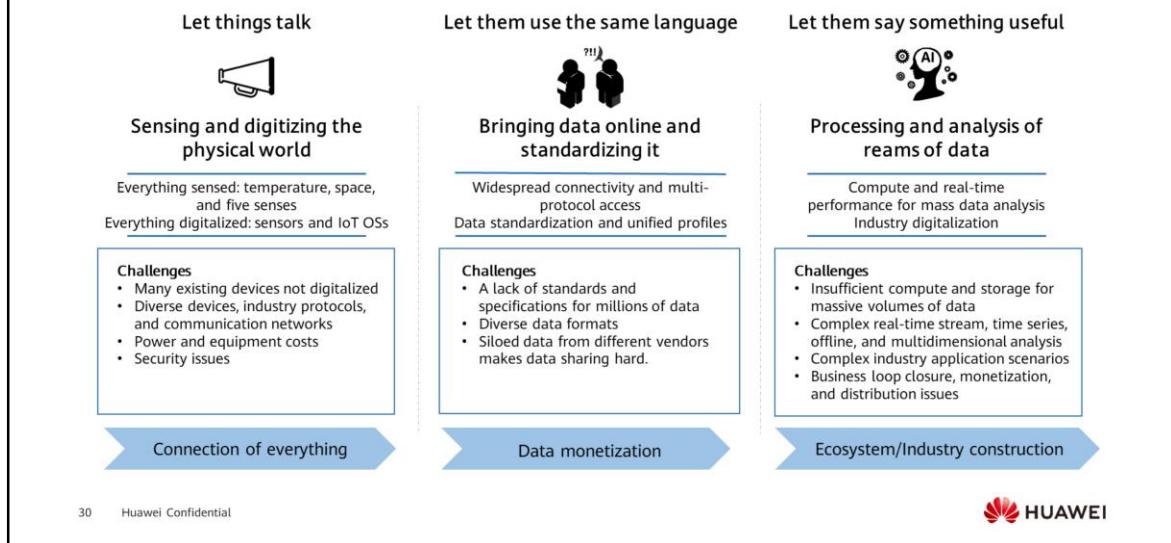
What Is IoT?

- The Internet of Things (IoT) is a next-generation information technology, a milestone of the information era. IoT is known as the third wave of the global information industry development after computers and the Internet. It is widely used for network convergence, integrating communications and sensing technologies, such as intelligent sensing, identification, and pervasive computing.



- Digital factory:** By connecting production line devices to an IoT platform for real-time monitoring, analysis, and alarm management, efficiency is improved and power saved.
- Product design optimization:** Enterprises can connect their products to the Huawei Cloud IoT platform to improve product design and provide personalized services based on collected user and product data.

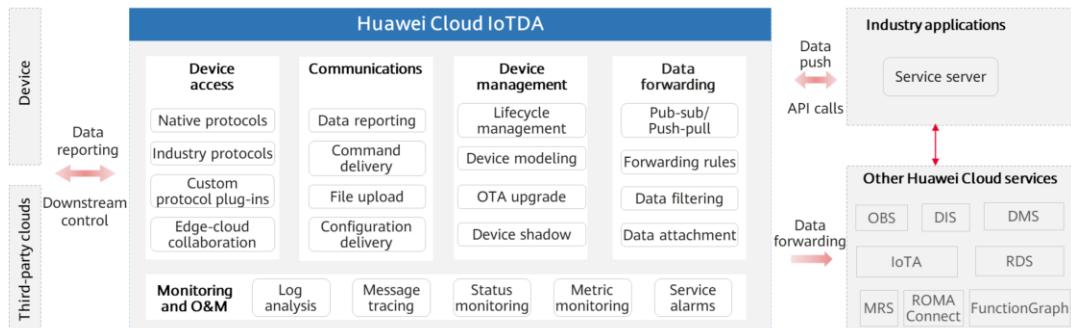
Challenges Facing the IoT Industry



- We need to connect everything, monetize data, and build an ecosystem to develop the IoT industry.

IoTDA for Device Access in All Scenarios

- IoT Device Access (IoTDA) allows you to connect device fleets to the cloud, enable device-cloud intercommunications, manage devices in batches, remotely control and monitor devices, perform over-the-air (OTA) upgrades, configure device linkage rules, and forward device data to other Huawei Cloud services. Using IoTDA, you can quickly connect devices and integrate applications.



IoTDA Application Scenarios

Scenario 1: Integration of IoT and PaaS components for migrating IoT data to the cloud

- Seamless IoT and PaaS integration makes solution development more efficient.
- Seamless integration with six cloud services in three scenarios for on-demand device data forwarding.
- Plan in 2021: SQL-like rules can be used to customize data and cover six types of integration scenarios.

Scenario 2: A cost-effective platform for high-concurrency that you can count on

- IoTDA supports high concurrency access for half the price of an enterprise-managed platform.
- Dedicated cloud instances provide access for tens of millions of devices and enable concurrent collection of 100,000 device data records per second.
- Plan in 2021: It supports three-AZ deployment, cross-region DR, dynamic KPI monitoring, and alarm reporting.

Scenario 3: Simplified device connection using a unified platform

- SDKs are pre-integrated with popular modules. All it takes is two AT commands to connect devices to the cloud.
- IoTDA supports 10+ native protocols and 30+ industry protocols. There are 12 access modes to cover diverse range of scenarios.
- Device connection to the cloud is simplified and accelerated. Applications can be easily integrated on the same platform.
- Plan in 2021: Devices can be connected to the platform without code changes in non-intrusive mode. 60+ industry protocols are supported.

AIoT

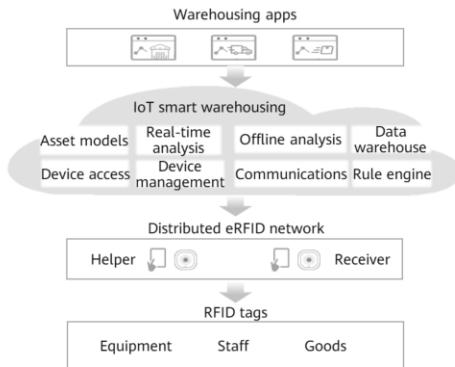
- In the IoT era, everything needs to be intelligently connected to unleash the value of connectivity. IoT bridges the physical and digital worlds. AI technologies can help achieve industry-specific objectives.
- AIoT enables a system to collect data from different sensors in real time and implement computing and intelligence transformation on the cloud, edge, and devices.

$$\text{AIoT} = \text{AI} + \text{IoT}$$



Smart Warehousing

- **Service overview and pain points:** Warehousing is an important part of logistics. Efficient warehousing can significantly improve the logistics and supply chain systems of an enterprise. Warehousing in traditional enterprises is not mechanized or intelligent, which results in too many errors and excessive labor costs.



Solution

- Use RFID for regional networks and automatic inventory. Real-time positioning is accurate to the meter, so goods can be tracked digitally, in real time, and checked against the manifest.
- Connect devices to IoTDA so developers can quickly build asset models and manage assets more easily.
- Use big data technologies to build a data warehouse to store and process vast amounts of data in real time.
- Use AI to evaluate warehousing capabilities and risks for intelligent operations and management.

Benefits

- Automatic batch identification of stored goods **reduces labor costs by 20%**. Automatic analysis of transport data **improves efficiency by 30%**. Subsystem integration and service flow streamlining simplify application deployment.

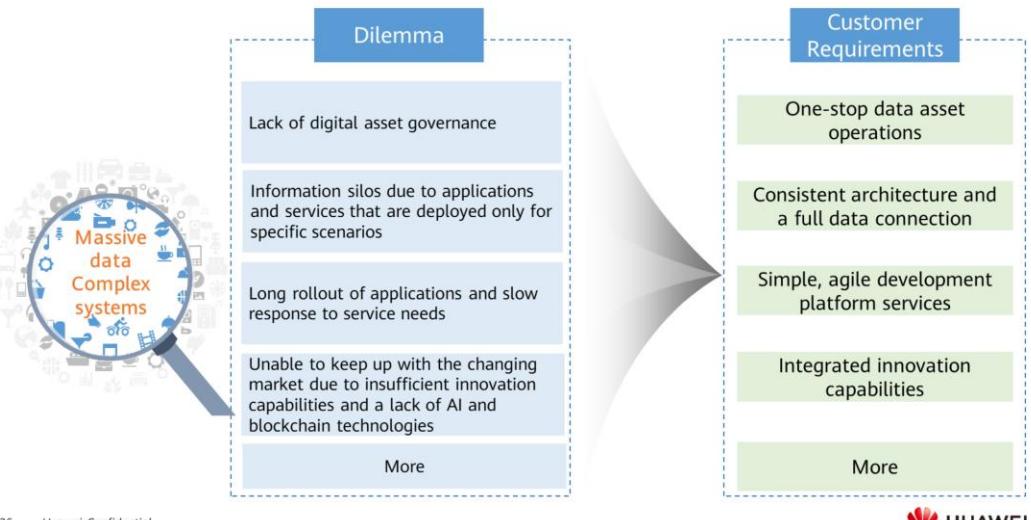


- As there are mappings between boxes and RFID tags and mappings between boxes and warehouse gates, a large amount of RFID data is generated during the inbound and outbound processes. By leveraging the stream computing capability of Flink, IoTDA can detect inbound and outbound goods under a gate in seconds. Then, the system checks goods against the goods list and informs warehouse staff of the goods status in real time.

Contents

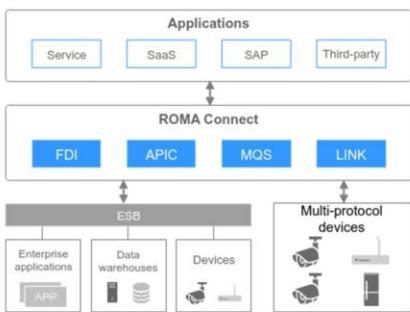
1. Overview of Huawei Cloud Innovations and Solutions
2. Huawei Cloud EI Solution
3. Huawei Cloud IoT Solution
- 4. Huawei Cloud Application and Data Solution**

Application and Data Development Dilemma



ROMA Connect: an Integration Platform

- ROMA Connect is a full-stack application and data integration platform. It integrates data, APIs, messages, and devices to allow interconnection between cloud and on-premises applications, helping enterprises achieve digital transformation.

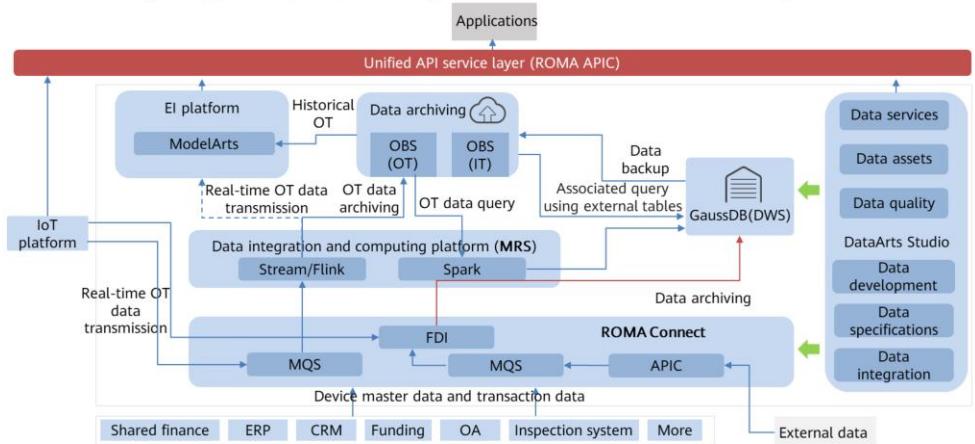


- Hybrid integration:** one-stop application/data/API/device integration capabilities
 - FDI for data integration: flexibly converts many heterogeneous data sources, incorporating data import capabilities of DRS and CDM.
 - APIC for service integration: implements lifecycle management of APIs.
 - MQS for message integration: connects to message middleware based on Kafka protocol, which is unnoticeable on frontend and backend applications.
 - LINK for device integration: implements IoT device access, device management, data collection, and device control based on the MQTT protocol.
 - FDI, APIC, MQS, and LINK can be combined to form multiple integration solutions.
- Business object connection**
 - Build business models, and build processes using these models through certain rules.
 - Graphical business flow orchestration allows configuration in less than 5 minutes.
- Multi-level interconnection across clouds and networks**
- Open ecosystem of industry suites**

- FDI:** If an enterprise and its partners use different data sources, information transmission will be ineffective. FDI can convert multiple mainstream data formats, such as MySQL, Kafka, and APIs. It can also work with other services, such as Gauss200, to store, convert, and analyze big data.
- APIC:** If a corporate group integrates its IT system with those of its branches in different regions, direct access to each other's databases can be very complex and cause information leaks. Open access through APIs and enhanced API call security ensure collaboration across networks and regions.
- MQS:** If an enterprise and its partners use different message systems, interconnection between their message systems is costly, and message transmission may not be reliable or secure. To address these issues, the Kafka protocol can be used for communication between the enterprise and its partners, while MQS functions as a message transfer station to provide secure and reliable message transmission. The enterprise can create multiple topics, authorize each partner to subscribe to these topics, and publish messages to the topics. Then, partners can subscribe to these topics to obtain messages.
- LINK:** In industrial scenarios, device information and production parameters are scattered. If a fault occurs in a production line, it requires a long time to manually collect information and parameters from each device. LINK connects devices to IT systems or big data platforms, and uploads information such as device running status to these platforms so that enterprises can view information about all devices graphically and therefore quickly locate faults.

Integrating Systems for Smart City Management

- ROMA Connect integrates applications, data, and messages to connect services across the board for digital transformation.



38 Huawei Confidential



- Using ROMA Connect, connections can be secure, reliable, and efficient for safe cross-organization collaboration of APIs, data, and messages.
- ROMA Connect provides hybrid integration capabilities to connect service systems, devices, and heterogeneous data sources. Beyond that, developing new applications costs half of the time by connecting IT and OT data through ROMA Connect.
- ROMA Connect provides API gateways and custom backends for simplified and quick API openness. Various data tables can be directly opened as RESTful APIs for service systems to call.

Quiz

1. (Single-answer question) An enterprise wants to connect scattered mobile devices to the cloud platform for unified management. Which of the following is the most suitable product? ()
 - A. MRS
 - B. ROMA Connect
 - C. IoTDA
 - D. DLI
2. (Multiple-answer question) ModelArts improves AI development efficiency throughout the entire process. Which of the following are steps in ModelArts? ()
 - A. Model evaluation
 - B. Inference monitoring
 - C. Model deployment
 - D. Application deployment

- 1. C
- 2. ABC. ModelArts focuses on model deployment rather than application deployment.

Summary

- This course introduced Huawei Cloud innovations and solutions in EI, IoT, and application and data development, and the implementation of these solutions in actual business scenarios.

Acronyms and Abbreviations

- AI: Artificial Intelligence
- AIoT: Artificial Intelligence of Things
- AMQP: Advanced Message Queuing Protocol
- API: Application Programming Interface
- APIC: API Connect
- BCS: Blockchain Service
- BI: Business Intelligence
- CDM: Cloud Data Migration
- CSS: Cloud Search Service
- DataArts Studio: Data Lake Governance Center
- DLI: Data Lake Insight
- DLV: Data Lake Visualization
- DRS: Data Replication Service
- DWS: Data Warehouse Service
- EI: Enterprise Intelligence
- ESB: Enterprise Service Bus
- ETL: Extract, Transform, Load
- FDI: Fast Data Integration
- GES: Graph Engine Service

Acronyms and Abbreviations

- HDFS: Hadoop Distributed File System
- IoTDA: IoT Device Access
- IoTDP: IoT Device Provisioning
- MQS: Message Queue Service
- MRS: MapReduce Service
- NPU: Neural Network Processing Unit
- RDS: Relational Database Service
- RFID: Radio Frequency Identification
- RPA: Robotic Process Automation
- TICS: Trusted Intelligent Computing Service

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©2022 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.

