Alibaba Cloud
Apsara Stack Enterprise

Technical Whitepaper

Version: 1901

Issue: 20190528

Legal disclaimer

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

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.
2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company, or individual in any form or by any means without the prior written consent of Alibaba Cloud.
3. The content of this document may be changed due to product version upgrades, adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.
4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults" and "as available" basis. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, special, or punitive damages, including lost profits arising from the use or trust in this document, even if Alibaba Cloud has been notified of the possibility of such a loss.
5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified,

reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names , trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.

6. Please contact Alibaba Cloud directly if you discover any errors in this document.

Generic conventions

Table -1: Style conventions

Style	Description	Example
	This warning information indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	 **Danger:** Resetting will result in the loss of user configuration data.
	This warning information indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	 **Warning:** Restarting will cause business interruption. About 10 minutes are required to restore business.
	This indicates warning information, supplementary instructions, and other content that the user must understand.	 **Note:** Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructio ns, best practices, tips, and other contents.	 **Note:** You can use **Ctrl** + **A** to select all files.
>	Multi-level menu cascade.	**Settings** > **Network** > **Set network type**
Bold	It is used for buttons, menus, page names, and other UI elements.	Click **OK**.
`Courier font`	It is used for commands.	Run the `cd /d C:/windows` command to enter the Windows system folder.
`Italics`	It is used for parameters and variables.	`bae log list --instanceid` *`Instance_ID`*
[] or [a|b]	It indicates that it is a optional value, and only one item can be selected.	`ipconfig` *`[-all|-t]`*
{} or {a|b}	It indicates that it is a required value, and only one item can be selected.	`switch` *`{stand | slave}`*

Contents

Legal disclaimer I
Generic conventions I
1 Elastic Compute Service (ECS) 1
1.1 What is ECS 1
1.2 Architecture 2
1.2.1 Overview 2
1.2.2 Virtualization platform and distributed storage 2
1.2.3 Control system 4
1.3 Features 4
2 Auto Scaling (ESS) 7
2.1 What is ESS? 7
2.2 Architecture 7
2.3 Features 8
2.3.1 Typical scenarios 8
2.3.1.1 Overview 8
2.3.1.2 Elastic scale-out 8
2.3.1.3 Elastic scale-in 9
2.3.1.4 Elastic recovery 9
2.3.2 Function components 9
3 Object Storage Service (OSS) 11
3.1 What is OSS 11
3.1.1 Definition 11
3.1.2 Advantages 11
3.1.3 Scenarios 13
3.2 Benefits 13
3.3 Architecture 15
3.3.1 System architecture 15
3.3.2 Data transmission process 17
3.4 Principles 18
3.4.1 Components 18
3.4.2 Features 19
3.4.3 Concepts 21
4 Table Store 24
4.1 What is Table Store 24
4.1.1 Technical background 24
4.1.2 Table Store technologies 25
4.2 Benefits 27
4.3 Architecture 28
4.4 Features 29
4.4.1 Users and instances 29

4.4.2 Data tables....30
4.4.3 Data partitioning....31
4.4.4 Common commands and functions....32
4.4.5 Authorization and permission control....32
5 Network Attached Storage (NAS)....34
5.1 What is NAS....34
5.2 Architecture....34
5.3 Features....35
5.4 Benefits....36
6 Distributed File System (DFS)....38
6.1 What is DFS....38
6.2 Design philosophy....38
6.3 Architecture....39
6.4 Product positioning....40
6.5 Product values....42
6.6 Use cases....42
7 ApsaraDB for RDS....43
7.1 What is ApsaraDB for RDS?....43
7.2 Architecture....44
7.3 Features and principles....46
7.3.1 Data link service....46
7.3.2 High-availability service....48
7.3.3 Backup service....51
7.3.4 Monitoring service....53
7.3.5 Scheduling service....55
8 KVStore for Redis....56
8.1 What is KVStore for Redis....56
8.1.1 Overview....56
8.1.2 Scenarios....56
8.1.3 Benefits....57
8.2 Features....59
8.2.1 Data link service....59
8.2.1.1 DNS....60
8.2.1.2 SLB....61
8.2.1.3 Proxy....61
8.2.1.4 DB engine....61
8.2.2 High-availability service....62
8.2.2.1 Detection....62
8.2.2.2 Repair....63
8.2.2.3 Notice....63
8.2.3 Monitoring service....63
8.2.3.1 Service-level monitoring....63
8.2.3.2 Network-level monitoring....64

8.2.3.3 OS-layer monitoring....64
8.2.3.4 Instance-level monitoring....64
8.2.4 Scheduling service....64
9 ApsaraDB for MongoDB....65
9.1 What is ApsaraDB for MongoDB....65
9.2 System architecture....65
9.3 Functions....66
9.3.1 Data link service....66
9.3.2 High availability service....68
9.3.3 Backup service....69
9.3.4 Monitoring service....70
9.3.5 Scheduling service....71
10 KVStore for Memcache....72
10.1 What is KVStore for Memcache....72
10.1.1 Scenarios....72
10.1.2 Benefits....72
10.2 Functions....73
10.2.1 Data link service....74
10.2.1.1 DNS....75
10.2.1.2 SLB....75
10.2.1.3 Proxy....75
10.2.1.4 DB Engine....75
10.2.2 High availability service....76
10.2.2.1 Detection....77
10.2.2.2 Repair....77
10.2.2.3 Notice....77
10.2.3 Monitoring service....77
10.2.3.1 Service-level monitoring....77
10.2.3.2 Network-level monitoring....78
10.2.3.3 OS-level monitoring....78
10.2.3.4 Instance-level monitoring....78
10.2.4 Scheduling service....78
11 Data Management Service (DMS)....79
11.1 What is DMS?....79
11.2 Architecture....79
11.3 Function module....80
11.4 Benefits....81
11.5 Product value....82
12 Server Load Balancer (SLB)....85
12.1 What is Server Load Balancer?....85
12.2 Architecture....86
12.3 LVS in Layer-4 Server Load Balancer....89
12.4 Tengine in Layer-7 Server Load Balancer....94

13 Virtual Private Cloud (VPC)....95
13.1 What is VPC....95
13.2 Architecture....96
14 Log Service....99
14.1 What is Log Service....99
14.2 Architecture....99
14.3 Components....100
14.4 Features....101
14.5 Benefits....102
15 Apsara Stack Security....103
15.1 What is Apsara Stack Security....103
15.2 Benefits....103
15.3 Architecture....105
15.4 Features....107
15.4.1 Apsara Stack Security Standard Edition....107
15.4.1.1 Traffic Security Monitoring....107
15.4.1.2 Server Intrusion Detection....108
15.4.1.3 Server Guard....109
15.4.1.4 Security Audit....113
15.4.1.5 Web Application Firewall....115
15.4.1.6 Threat Detection Service....117
15.4.1.7 On-premises security operations services....121
15.4.2 Optional security services....124
15.4.2.1 DDoS Traffic Scrubbing....124
16 Key Management Service (KMS)....127
16.1 Product overview....127
16.2 Product architecture....127
16.3 Functions and features....128
16.3.1 Convenient key management....128
16.3.2 Envelope encryption technology....128
16.3.3 Secure key storage....130
17 Domain Name System (DNS)....131
17.1 What is Apsara Stack DNS....131
17.2 System architecture....131
17.3 Features....131
17.4 Benefits....132
17.5 Advantages....133
18 API Gateway....134
18.1 What is API Gateway....134
18.2 System architecture....134
18.3 Basic functions....135
18.3.1 API lifecycle management....135

18.3.2 Multi-protocol access..... 135
18.3.3 Application access control..... 135
18.3.4 Full-link signature verification mechanism..... 136
18.3.5 Anti-replay mechanism..... 136
18.3.6 HTTPS communication based on the SSL certificate of the user..... 136
18.3.7 Support for OpenID Connect..... 136
18.3.8 Bidirectional communication..... 136
18.3.9 Automatic generation of SDKs and API documentation..... 137
18.3.10 Parameter cleaning..... 137
18.3.11 Mappings between frontend and backend parameters..... 137
18.3.12 Request throttling..... 137
18.3.13 Access control based on IP addresses..... 137
18.3.14 Log analysis..... 138
18.3.15 Publish an API in multiple environments..... 138
18.3.16 Online debugging..... 138
18.3.17 Mock mode..... 138
18.3.18 Swagger file import..... 138
18.4 Benefits..... 139

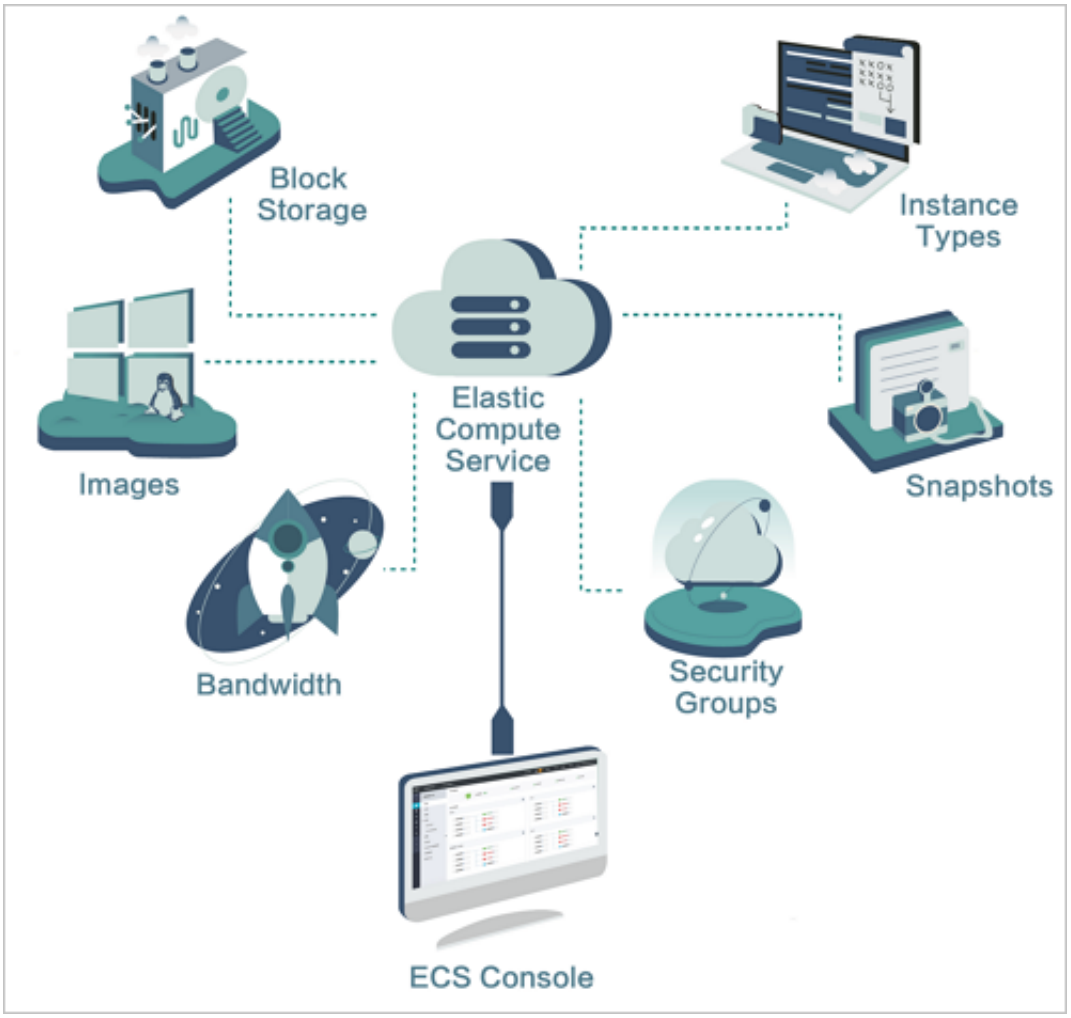
1 Elastic Compute Service (ECS)

1.1 What is ECS

Elastic Compute Service (ECS) is a type of computing service that features elastic processing capabilities. As compared with the physical servers, ECS is more user-friendly and can be managed more efficiently. You can create instances, resize disks, and add or release any number of ECS instances any time according to your business demands.

As a virtual computing environment made up of the basic components such as CPU, memory, and storage, an ECS instance is provided by ECS for you to carry out relevant operations. It is the core concept of ECS and you can perform actions on ECS instances on the ECS console. As for other resources such as block storage, images, and snapshots, they cannot be used until being integraed with ECS instances. *Figure 1-1: Concept of an ECS instance* illustrates the services supported by an ECS instance.

Figure 1-1: Concept of an ECS instance



1.2 Architecture

1.2.1 Overview

ECS is made up of a virtualization platform, a distributed storage and control system, and an O&M and monitoring system.

1.2.2 Virtualization platform and distributed storage

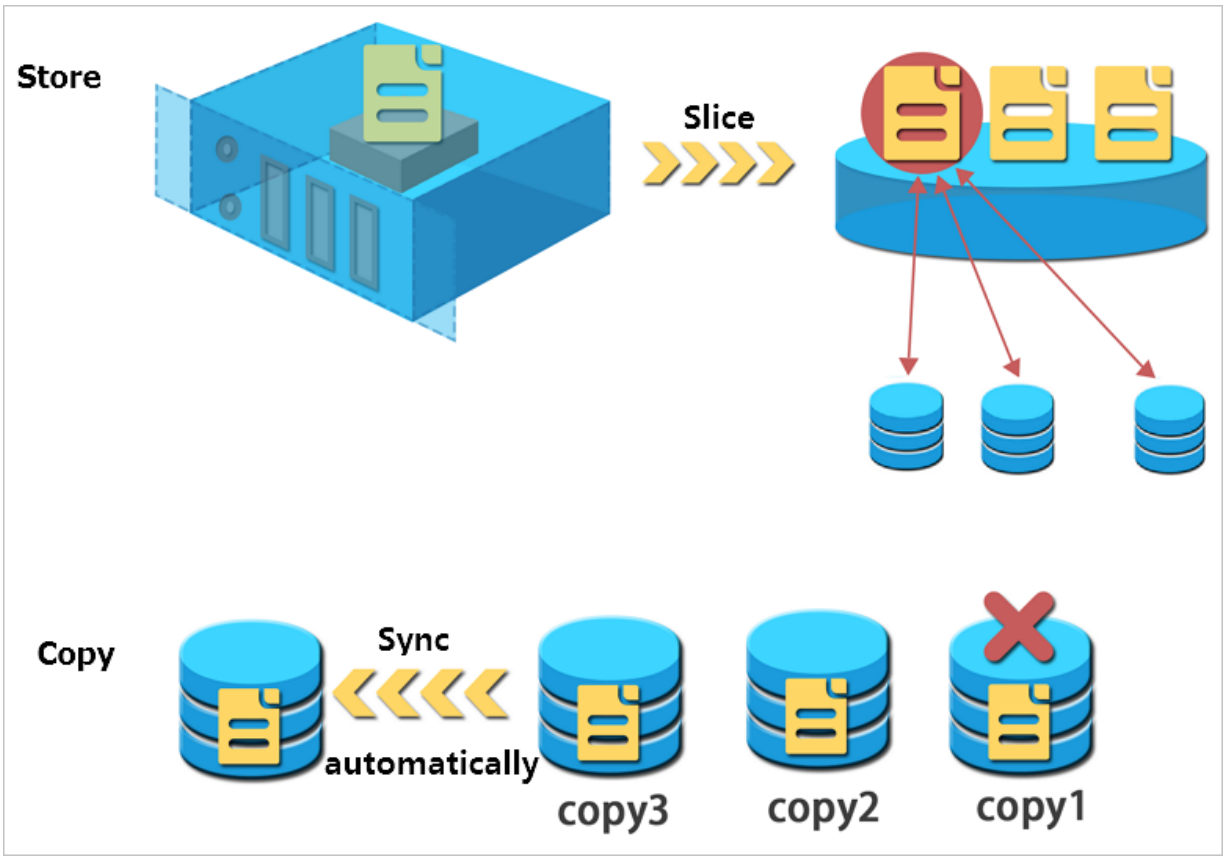
Virtualization is the foundation of ECS. Alibaba Cloud adopts the KVM virtualization technology to virtualize the physical resources, thus providing elastic computing services via the virtualized resources.

ECS contains two key virtualized modules: one for computing resources, and the other for storage resources.

- Computing resources refer to the CPU, memory, bandwidth, and other components of a physical server and are virtualized before being assigned to ECS. The computing resources of an ECS instance can only be located on the same physical server. If the resources of a physical server are used up, you have to create ECS instances on another physical server. With the resource QoS, ECS instances on the same physical server do not affect the running of each other.
- The storage module uses a large-scale distributed storage system. After the storage resources of an entire cluster are virtualized, the resources are bundled together and provided as an external service. Data for a single ECS instance is saved throughout the entire cluster. In the distributed storage system, all data is saved in triplicate. This way, if one copy is damaged, the data can be automatically recovered from another copy.

Triplicate technology is shown in *Figure 1-2: Distributed storage utilizing triplicate technology*

Figure 1-2: Distributed storage utilizing triplicate technology



1.2.3 Control system

As the core of the ECS platform, the control system determines the physical server where an ECS instance starts. In addition, all the information and functions of ECS are processed and maintained through the control center in a centralized way.

The control system is composed of the following four modules:

- **Data collection**

  This module collects data from the entire virtualization platform, including usage information for computing resources, storage resources, and network resources. The data collection module allows you to centrally monitor and manage the usage of cluster resources. Furthermore, it serves as the basis for resource scheduling.

- **Resource scheduling system**

  This module determines where an ECS instance starts. When you create an ECS instance, it rationally schedules the ECS instance based on physical server resource loads. This module can determine where to restart an instance if any fault occurs in an ECS instance.

- **ECS management module**

  This module manages and controls the ECS instances, for example, starting, stopping, and restarting ECS instances.

- **Security control module**

  This module monitors and manages the network security of the entire cluster.

1.3 Features

As the core of the elastic computing products, ECS is designed to provide the computing services for users. It taks only a few minutes to create and start an ECS instance. Moreover, once an ECS instance is created, it has specific system configuration. Compared to the traditional servers, ECS helps you improve the efficiency of delivering services considerably.

ECS instances are used the same way as the traditional hosted physical servers. You have full control over your ECS instances and can perform operations on them through the remote approach or the API approach (console).

The computing capabilities of ECS instances can be expressed in terms of virtual CPUs and virtual memory. ECS disk storage capabilities are measured by the capacity of available cloud disks. Unlike the traditional servers, ECS allows you to make more flexible machine configuration

based on your needs. That is, if the current ECS instance configuration cannot meet the business needs, you can change the configuration at any time.

The ECS life cycle begins with ECS instance creation and ends after you release it. Once an ECS instance is released, all its data is irrevocably deleted.

The ECS console of Apsara Stack offers easy access to the following information areas:

- **Resources**

  In this area, you can view the number of instances created and the number of running instances. You can also view the quantity and distribution of ECS instances in respective zones .

- **Instances**

  In this area, you can do the following:

  - View and manage the created instances.
  - Start, stop, restart, release, and log on to VNC.
  - Replace system disks, reset your password, and change the configuration.
  - View the basic and configuration information of the instances.

- **Disks**

  In this area, you can do the following:

  - View and manage the created disks.
  - Reinitialize a disk, create snapshots, set an automatic snapshot policy for disks, release disks, and attach or detach disks.
  - View the basic information of disks.

- **Images**

  In this area, you can do the following:

  - View and manage information of the created or shared images.
  - Copy, share, and delete images.

- **Snapshots**

  In this area, you can do the following:

  - View and manage the created snapshots.
  - Roll back a disk, create custom images, and delete snapshots.

- **Automatic snapshot policy**

In this area, you can do the following:

- View and manage the configured automatic snapshot policy.
- Configure the automatic snapshot policy in batch.
- Change and delete the automatic snapshot policy.

- **Security group**

  In this area, you can do the following:

  - View and manage the created security groups.
  - Create, change and delete (individually or in batch) security groups.
  - View the instances and rules in a security group.

- **Elastic Network Interface**

  In this area, you can do the following:

  - View and manage the created Elastic Network Interfaces (ENIs).
  - Create, change and delete ENIs.
  - Bind and unbind instances.

- **Deployment set**

  In this area, you can do the following:

  - View and manage the created deployment sets.
  - Create, change and delete the deployment sets.
  - View the basic information of deployment sets.

2 Auto Scaling (ESS)

2.1 What is ESS?

Auto Scaling (ESS) is a management service that automatically adjusts the number of elastic computing resources based on your business demands and strategies. It is suitable for applications with fluctuating business loads, as well as applications with stable business loads.

ESS automatically schedules computing resources based on customer strategies and changing business requirements. It provides support for changing business loads and helps control infrastructure costs within an acceptable range. ESS executes scaling based on user-defined scaling policies and modes. When business loads increase, ESS automatically adds ECS instances to ensure sufficient computing capabilities. When business loads decrease, ESS automatically removes ECS instances to save costs. It also replaces unhealthy ECS instances to ensure service performance and safeguard your business.

In addition, ESS is seamlessly integrated with Server Load Balancer (SLB) and ApsaraDB for Relational Database Service (RDS). This allows ESS to add or remove ECS instances to or from an SLB backend server group, as well as to add or remove IP addresses of ECS instances to or from an RDS whitelist. ESS eliminates the need to manually perform O&M operations, as it adapts to various complex scenarios and automatically processes business loads based on actual requirements.

2.2 Architecture

ESS is a system that orchestrates ECS instances and provides services based on basic components such as ECS. The ESS system consists of the trigger, worker, database, and middleware services.

Table 2-1: Architecture description

Layer	Description
Middleware layer	ZooKeeper: ensures consistency by implementing distributed locks for Server Controller.
	Tair: provides caching services for Server Controller
	Message Queue (MQ): provides message queuing services of VM statuses.
	Diamond: manages persistent configurations.

Layer	Description
Database layer: the business database and workload database	Worker: The core of ESS. After receiving a task, it handles the entire life cycle of the task, including splitting, executing, and returning the execution results.
	Trigger: It obtains information from the health checks of instances and scaling groups, scheduled tasks, and CloudMonitor to perform tasks scheduling.
Public-facing services	Coordinator: serves as the ingress of the ESS architecture . It provides external management and control for services, processes API calls, and triggers tasks.
	OpenAPI Gateway: provides basic services such as authentication and parameter passthrough.

2.3 Features

2.3.1 Typical scenarios

2.3.1.1 Overview

ESS automatically adjusts the number of elastic computing resources to meet fluctuating business demands. Based on user-defined scaling rules, ESS automatically adds ECS instances as business loads increase to ensure sufficient computing capabilities. When your business loads decrease, ESS automatically removes ECS instances to save costs.

2.3.1.2 Elastic scale-out

When business loads surge, ESS automatically increases underlying resources. This helps maintain access speed and ensure that resources are not overloaded.

You can create scheduled tasks to perform automatic scale-out at specified times or configure CloudMonitor to monitor ECS instance usage in real time and perform scale-out based on actual requirements. For example, when CloudMonitor detects that the vCPU utilization of ECS instances in a scaling group exceeds 80%, ESS elastically scales out ECS resources based on user-defined scaling rules. During the scale-out process, ESS automatically creates ECS instances and adds these ECS instances to the SLB instance and RDS whitelist.

2.3.1.3 Elastic scale-in

When loads on services decrease, ESS automatically releases underlying resources to prevent resource wastage and reduce costs.

You can create scheduled tasks to scale in resources automatically at specified points in time. You can also configure CloudMonitor to monitor ECS instance usage in real time and scale in resources based on actual requirements. For example, when CloudMonitor detects that the vCPU utilization of ECS instances in a scaling group falls below a specified threshold, ESS automatically scales in ECS resources based on user-defined rules. During the scale-in process, ESS releases ECS instances and removes these ECS instances from the SLB instance and RDS whitelist.

2.3.1.4 Elastic recovery

ESS provides a health check function and automatically monitors the health of ECS instances inside scaling groups, so that the number of healthy ECS instances in a scaling group does not fall below the user-defined minimum value.

When ESS detects that an ECS instance is not healthy, it automatically releases the unhealthy ECS instance, creates a new ECS instance, and adds the new instance to the SLB instance and RDS whitelist.

2.3.2 Function components

To create a complete automatic scaling solution that performs scale-in and scale-out based on actual requirements, you need to create scaling groups, configurations, rules, and scheduled tasks.

Scaling group

A scaling group is a group of ECS instances that is dynamically scaled based on the configured scenario. You can specify the maximum and minimum number of ECS instances in a scaling group, as well as the SLB and RDS instances associated with the group.

Scaling configuration

A scaling configuration is a template in ESS for creating ECS instances. When creating a scaling configuration, you can specify ECS instance information, such as instance type, image type, storage size, and instance logon key pair. You can also modify an existing scaling configuration as needed.

Scaling rule

A scaling rule defines the specific scaling activity, for example, the number of ECS instances to be added or removed. The following scaling rules are supported:

- Set to N instances: After this scaling rule is executed, the number of instances in service is changed to N.
- Add N instances: After this scaling rule is executed, the number of instances in service is increased by N.
- Decrease N instances: After this scaling rule is executed, the number of instances in service is reduced by N.

Scheduled task

A scheduled task defines execution actions within a scaling group. It can trigger a specific scaling rule at a specific point in time to execute a scaling activity, such as adjusting the number of ECS instances in a scaling group.

3 Object Storage Service (OSS)

3.1 What is OSS

3.1.1 Definition

Object Storage Service (OSS) is a massive, secure, cost-effective, and highly reliable cloud storage service provided by Alibaba Cloud.

OSS can be considered as an out-of-the-box storage cluster with unlimited capacity. Compared with traditional user-created server storage, OSS has many outstanding advantages in reliability, security, cost-effectiveness, and data processing capabilities. You can use OSS to store and retrieve a variety of unstructured data objects, such as texts, images, audios, and videos, over the network at any time.

OSS uploads data files to buckets as objects. OSS stores objects as key-value pairs. You can retrieve the content of an object based on the unique object name (key).

You can perform the following OSS operations:

- Create a bucket and upload objects to the bucket.
- Obtain the URL of an uploaded object, and use the URL to share or download the object.
- Modify the properties or metadata of a bucket or object, such as setting ACLs for the bucket or object.
- Perform basic and advanced operations in the OSS console.
- Perform basic and advanced operations by using SDKs or directly calling RESTful APIs in your application.

3.1.2 Advantages

Advantages of OSS over user-created server storage

Item	OSS	User-created server storage
Reliability	• The capacity is automatically expanded without affecting external services. • Offers automatic redundant data backup.	• Prone to errors due to low hardware reliability. If a disk has a bad sector, data may be irretrievably lost.

Item	OSS	User-created server storage
		• Manual data restoration is complex and requires a lot of time and technical resources.
Security	• Provides hierarchical security protection for enterprises. • User resource isolation mechanisms and local disaster recovery • Provides various authentication and authorization mechanisms, as well as whitelisting, hotlinking protection, and RAM. It also provides Security Token Service (STS) for temporary access.	• Additional scrubbing and black hole equipment is required. • A separate security mechanism is required.
Data processing	Image processing capabilities	Image processing capabilities must be purchased and deployed separately.

More benefits of OSS

- Ease of use

  Provides standard RESTful APIs (some compatible with Amazon S3 APIs), a wide range of SDKs and client tools, and a management console. You can easily upload, download, retrieve, and manage large amounts of data for websites and applications, similar to regular files systems.

  - There is no limit on the number and size of objects. Therefore, you can easily expand your buckets in OSS as required.
  - Supports streaming writing and reading, which is suitable for business scenarios where you need to simultaneously read and write videos and other large objects.
  - Supports lifecycle management. You can delete expired data in batches.
- Powerful and flexible security mechanisms

  Flexible authentication and authorization mechanisms are available. OSS provides STS and URL authentication and authorization, as well as whitelisting, hotlinking protection, and RAM.
- Rich image processing functions

  Supports format conversion, thumbnails, cropping, watermarking, resizing for object formats such as JPG, PNG, BMP, GIF, WEBP, and TIFF.

3.1.3 Scenarios

Massive storage for image, audio, and video applications

OSS can be used to store large amounts of data, such as images, audios, videos, and logs. OSS supports various devices. Websites and mobile applications can directly read or write OSS data. OSS supports file writing and streaming writing.

Dynamic and static content separation for websites and mobile applications

OSS leverages the BGP bandwidth to achieve ultra-low latency of direct data download.

Offline data storage

OSS is cheap and highly available, enabling enterprises to store data that needs to be archived offline for a long time to OSS.

3.2 Benefits

Multifunctionality

- Supports multiple functions: simple upload, form upload, append upload, download, delete, list, replicate, obtain Object Meta, and create multipart upload tasks.
- Supports bucket-based functions: create, delete, and list objects in a bucket as well as obtain Bucket Meta.
- Creates a globally unique bucket and supports cross-region bucket replication.
- Supports lifecycle management, defines and manages lifecycle rules for all objects in a bucket or a part of an object, and changes capacities and ownership.
- Supports IMG. You can obtain image information, convert the image format, resize, crop, and rotate an image, add images, texts, and image-text watermarks to an image, customize an image style, call cascading processing in the first-in, first-out (FIFO) order, and protect a source image.
- Supports zone-disaster recovery. In the zone-disaster recovery mode, buckets with the same name are replicated. Cluster-based disaster recovery is automatically enabled based on configurations made when the cluster is created. In other words, after a primary bucket is created, a secondary bucket with the same name is created automatically. Information stored in the primary bucket is automatically synchronized to the secondary bucket.
- Configures static website hosting for your bucket and allows you to use the bucket endpoint to access this static website.
- Supports hotlink protection based on the referer fields in HTTP headers.

- Supports cross-region access. Supports access logging and log analytics in multiple dimensions. You can view the access initiator.
- Uses the redundant architecture to prevent a single point of failure (SPOF).
- Uploads and downloads large objects, supports multipart upload and range download of large objects, and supports resumable upload, download, and replication.

High performance

Supports the throughput of a cluster that contains tens of thousands of nodes.

Security

Supports an ACL for permission control. You can configure an ACL when creating a bucket and modify it after it is created. Three levels of permissions are included in an ACL: Private, Public (Read-Only), and Public.

Supports Resource Access Management (RAM), Alibaba Cloud accounts, Apsara Stack tenant accounts, and RAM users for employees, applications, and systems based on the department architecture. A separate logon password or AccessKey pair is created for each employee, application, and system. RAM users do not have any permissions on OSS resources by default . You can use RAM to assign permissions to RAM users or use Security Token Service (STS) for temporary access authorization. HTTPS and traffic encryption on the server and client are supported.

Supports APIs, SDKs, or migration tools to easily migrate large amounts of data to or out of Alibaba Cloud.

Supports multiple types of terminals, Web applications, and mobile applications and allows them to write data to or read data from OSS directly. Stream input and object input are supported. You can manage static resources such as images, scripts, and videos on the website in the way you manage folders. After objects are uploaded to OSS, you can apply the features provided by the services in the cloud OS to the objects, such as audio and video processing, IMG, BatchCompute , and offline processing. In this way, you can maximize data values.

Supports hotlink protection to prevent unauthorized access.

Supports Secure Sockets Layer (SSL) to control the read/write permission on each object.

Integrates with the intrusion prevention system to effectively prevent DDoS attacks and CC attacks to ensure that business works properly.

Supports cross-region replication to allow you to synchronize data to a specified region in real time for geo-disaster recovery. In this way, OSS protects important data from the impact of extreme disasters and ensures service stability.

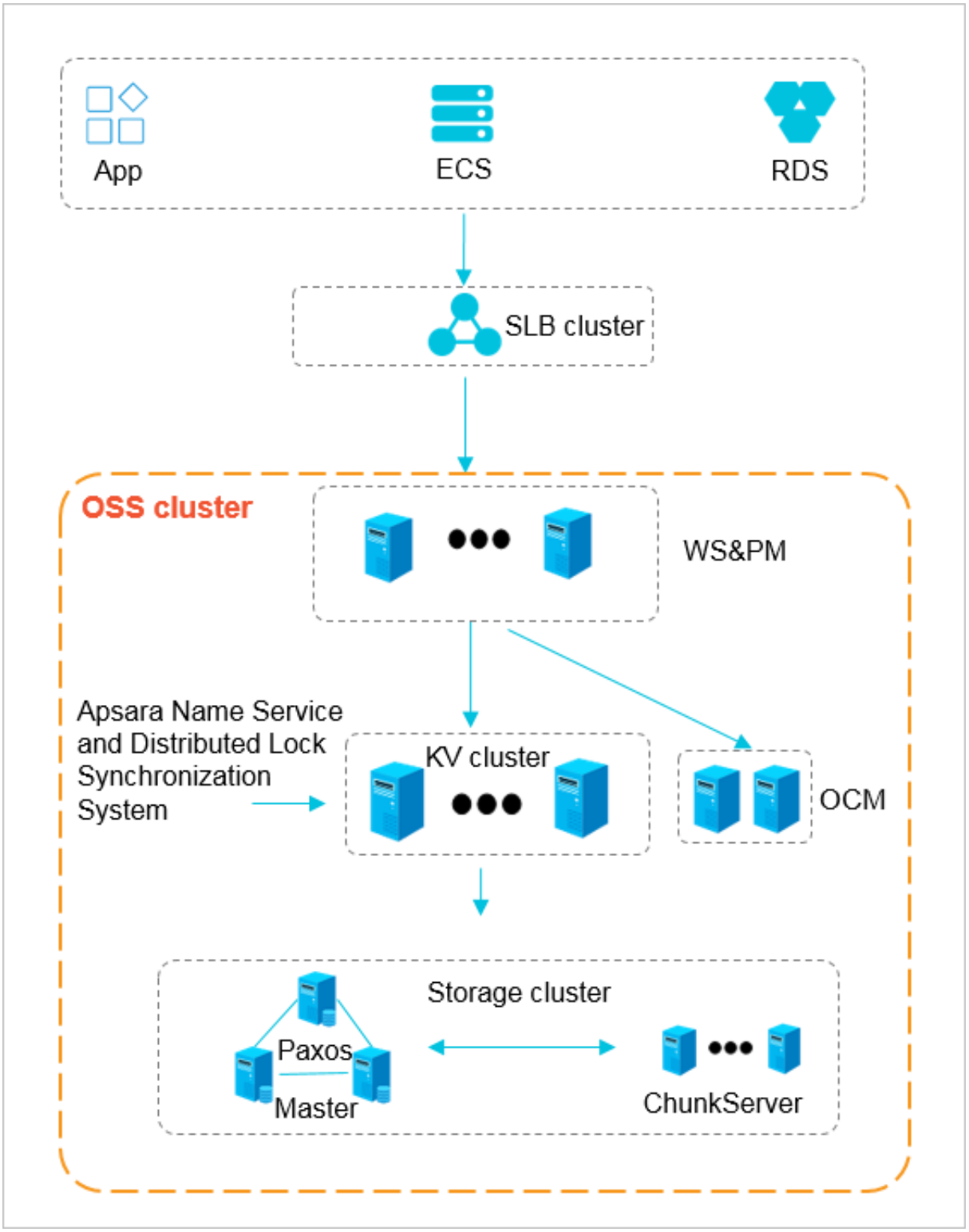
3.3 Architecture

3.3.1 System architecture

OSS is a storage solution that is built on the Apsara system. It is based on the infrastructure such as Apsara Distributed File System and SchedulerX. The infrastructure provides OSS and other Alibaba Cloud services with important features such as distributed scheduling, high-speed networks, and distributed storage.

The following figure shows the OSS architecture.

Figure 3-1: OSS architecture



The OSS architecture consists of three layers: protocol access layer, partition layer, and persistent storage layer.

- Protocol access layer

- WS: uses the open-source Tengine component, and provides HTTP and HTTPS for external services.
- PM: parses the HTTP request as the read/write operation on the backend KV or another module. PM also receives and authenticates the user request sent through a RESTful protocol. If the authentication succeeds, the request is forwarded to KV Engine for further processing. If the request fails the authentication, an error message is returned.

- Partition layer

  The partition layer processes structured data, including querying and storing data based on keys. This layer also supports large amounts of concurrent requests. When a service has to run on a different physical server due to a change to the service coordination cluster, the KV cluster can quickly coordinate data streams to switch the access point. The partition layer manages indexes of objects, and converts objects to the persistent data objects at the persistent storage layer.

  - SchedulerX is responsible for naming services and is based on Apsara Name Service and Distributed Lock Synchronization System.
  - KV consists of KVMaster and KVServer. KVMaster manages and schedules partitions. KVServer stores indexes and actual data of partitions.

- Persistent layer

  The large-scale distributed file system is deployed at the persistent storage layer. Metadata is stored on masters. A distributed message consistency protocol (or Paxos) is adopted between masters to ensure the metadata consistency. In this way, efficient distributed file storage and access are achieved. This method ensures that three copies of data are stored in the system and that the system can recover from any hardware or software fault.

3.3.2 Data transmission process

The data transmission process is as follows from the perspective of user access:

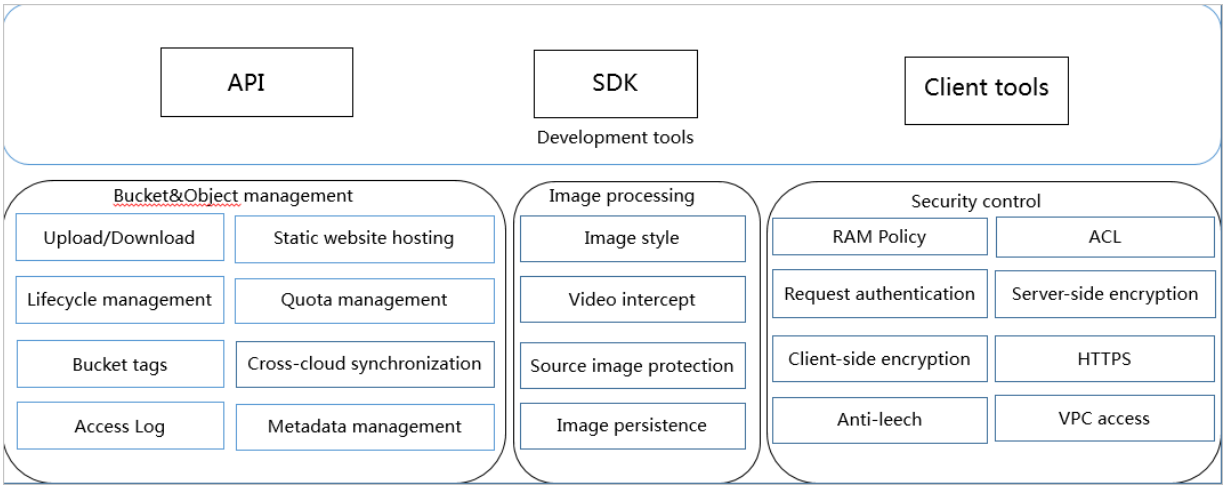
`User→RESTful API→SLB-Web server (WS)→Protocol module (PM)→KV Engine→ Distributed storage:`

- A user uses different clients such as browsers or SDKs to initiate a request that complies with the convention of OSS APIs to the OSS endpoint. The endpoint parses the request and sends it to the LVS VIP of SLB. The backend of the LVS VIP is bound to the actual WS. The request is forwarded to one of the WSs.

- The PM parses the user request.The specific process is as follows: First, the request is authenticated. If the request fails the authentication, the corresponding error code is returned.
- If the authentication succeeds, the request is parsed as the read/write operation on KV Engine and enters the partition layer.
- The partition layer processes structured data, including querying and storing data based on keys. This layer also supports large amounts of concurrent requests. When a service has to run on a different physical server due to a change to the service coordination cluster, the KV cluster can quickly coordinate data streams to switch the access point.
- The data stored in KV Engine of the partition layer is written to the persistent storage layer.
- The large-scale distributed file system is deployed at the persistent storage layer. Metadata is stored on masters. A distributed message consistency protocol (or Paxos) is adopted between masters to ensure the metadata consistency. In this way, efficient distributed file storage and access are achieved. This method ensures that three copies of data are stored in the system and that the system can recover from any hardware or software fault.

3.4 Principles

3.4.1 Components



OSS is a storage solution that is built on the Apsara system.

OSS consists of three modules: access layer, application layer, and infrastructure layer.

- Access layer: APIs, SDKs, and Apsara Stack Management Console
- Application layer: buckets, object management, IMG, and security modules
- Infrastructure layer: Apsara Distributed File System, Job Scheduler, and Apsara Name Service and Distributed Lock Synchronization System

3.4.2 Features

Bucket and object management

- Bucket overview

  All buckets of the request initiator are displayed. If you use HTTP to access the OSS endpoint, all of your buckets are displayed by default.

- Create or delete a bucket

  — You can create a maximum of 10 buckets by default. The name of a new bucket must comply with the bucket naming rules.

  The following scenarios may exist when you create a bucket:

  - If the bucket you want to create does not exist, the system creates a bucket of a specified name and returns a flag, indicating that the bucket is created.
  - If the bucket you want to create exists and the request initiator is the original bucket owner, the original bucket is retained and a flag is returned, indicating that the bucket is created.
  - If the bucket you want to create exists and the request initiator is not the original bucket owner, a flag is returned, indicating that the bucket fails to be created.

  If you want to delete a bucket, ensure that the following conditions are met:

  - The bucket exists.
  - You have the permission to delete the bucket.
  - The bucket contains objects.

- List all objects in a bucket

  To list all objects in a specified bucket, you must have the corresponding operation permissions on the bucket. If the specified bucket does not exist, an error message is returned.

  OSS allows you to search for buckets by prefix and set the maximum number (1,000) of objects that can be returned for each search.

- Upload or delete objects

  You can upload objects to a specified bucket. You can upload objects to a bucket if the bucket exists and you have the corresponding operation permissions on the bucket. If the object you want to upload has the same name with an object that already exists in the bucket, the new object overwrites the original object. You can delete a specified object if you have the corresponding operation permissions on the object.

- Obtain an object or its Object Meta

  To obtain the content of an object or its Object Meta, you must have the corresponding operation permissions on this object.

- Access an object

  OSS allows you to use a URL to access an object.

Image Processing (IMG)

- Custom image styles

  Each change made to an image is added to the URL. Multiple changes result in a long URL that is inconvenient for management and reading. IMG allows you to save common operations as an alias (a style). This style feature combines a series of operations into one operation. This style adds only one segment to the URL instead of multiple segments, which shortens the final image URL.

- Video snapshot

  IMG allows you to process the existing image content and capture the image at a specified point of the video to complete the video frame capturing.

- Source image protection

  To minimize image piracy risks, you must restrict access to the image URLs. Anonymous visitors can obtain only the URLs of thumbnailed or watermarked images. However, source image protection can address this need.

- IMG persistence

  OSS allows you to perform the SaveAs operation for data processing. This feature enables you to save the processed image to a specified bucket as a resource and assign the image with a specified key. After the image is saved, you can specify the bucket to speed up the resource download when you access the resource directly. This feature applies to ultra-large image cropping or other long-latency operations.

Security control

- Set and query the ACL of a bucket

  You can set and view the ACL of a bucket. You can set any one of the following permissions for a bucket:

  - Private: Only the creator or an authorized user of this bucket can read and write objects in the bucket. Other users cannot access the objects in the bucket without authorization.
  - Public (Read-Only): Only the creator of the bucket can perform write operations on the objects in the bucket. Other users (including anonymous users) can perform only read operations on the objects.
  - Public: Any user (including anonymous users) can read and write objects in the bucket.
- Access logging and monitoring

  You can choose to enable access logging for a bucket. After you enable this feature, OSS pushes the access logs on an hourly basis. You can view information such as buckets, traffic, and requests on the Object Storage Service homepage in Apsara Stack Management Console.
- VPC access control

  You can create a single tunnel between OSS and a VPC to access OSS resources over the VPC.
- Hotlink protection

  OSS provides hotlink protection to prevent unauthorized domain names from accessing your data in OSS. You can configure the referer field in the HTTP header to implement hotlink protection. You can configure a referer whitelist through the OSS console for a bucket or configure whether to accept access requests where the referer field is unspecified. For example, you can set the referer whitelist to http://www.aliyun.com for a bucket named oss-example. Then, only requests with a referer of http://www.aliyun.com can access the objects in the oss-example bucket.

3.4.3 Concepts

This topic describes several basic concepts of OSS.

Object

Objects, also known as OSS files, are the basic entities stored in OSS. An object is composed of metadata, data, and key. An object is identified by a unique key in the bucket. Metadata defines the properties of an object, such as the last modification time and the object size. You can also specify custom metadata for an object.

The lifecycle of an object starts when it is uploaded, and ends when it is deleted. During the lifecycle of an object, the metadata cannot be changed. Unlike the file system, OSS does not allow

you to modify objects directly. If you want to modify an object, you must upload a new object with the same name as the existing one to replace it.

Note:

Unless otherwise stated, objects and files mentioned in OSS documents are collectively called objects.

Bucket

A bucket is a container for objects. All objects must be stored in a bucket. You can set and modify the properties of a bucket for object access control and lifecycle management. These properties apply to all objects in the bucket. Therefore, you can create different buckets to perform different management functions.

- OSS does not have the hierarchical structure of directories and subfolders as in a file system. All objects are directly related to their corresponding buckets.
- You can have multiple buckets.
- A bucket name must be globally unique within OSS and cannot be changed once a bucket is created.
- A bucket can contain an unlimited number of objects.

Strong consistency

Object operations in OSS are atomic, which means operations are either successful or failed. There is no intermediate state. OSS will never write corrupted or partial data.

Object operations in OSS are strongly consistent. For example, once you receive a successful upload (PUT) response, the object can be read immediately, and the data has already been written in triplicate. Therefore, OSS avoids the situation where no data is obtained when you perform the read-after-write operation. When you delete an object, the object also has no intermediate state. Once you delete an object, that object no longer exists.

This feature allows OSS to be operated similar to traditional storage devices. Modifications are immediately visible, and consistency is guaranteed.

Comparison between OSS and the file system

OSS is a distributed object storage service that uses a key-value pair format. You can retrieve object content based on unique object names (keys). Although you can use names like test1/test.jpg, this does not necessarily indicate that the object is saved in a directory named test1. In

OSS, test1/test.jpg is only a string, which is no different from a.jpg. Therefore, similar amounts of resources are consumed when you access objects that have different names.

A file system uses a typical tree index structure. Before accessing a file named test1/test.jpg, you must access directory test1 and then locate the file named test.jpg. This makes it easy for a file system to support folder operations, such as renaming, deleting, and moving directories, as these operations are only directory node operations. System performance depends on the capacity of a single device. The more files and directories that are created in the file system, the more resources are consumed, and the lengthier your process becomes.

You can simulate similar functions in OSS, but this operation is costly. For example, if you want to rename the test1 directory test2, the actual OSS operation would be to replace all objects whose names start with test1/ with copies whose names start with test2/. Such an operation would consume a large amount of resources. Therefore, when using OSS, try to avoid such operations.

You cannot modify objects stored in OSS. A specific API must be called to append an object, and the generated object is of a different type from that of normally uploaded objects. Even if you only want to modify a single byte, you must re-upload the entire object. A file system allows you to modify files. You can modify the content at a specified offset location or truncate the end of a file . These features make file systems suitable for more general scenarios. However, OSS supports massive concurrent access, whereas the performance of a file system is subject to the performanc e of a single device.

Therefore, mapping OSS objects to file systems is very inefficient, which is not recommended. If attaching OSS as a file system is required, we recommended that you perform only the operations of writing new files, deleting files, and reading files. We recommend that you take full advantage of OSS features, such as its massive data processing capabilities to store large amounts of unstructured data, such as images, videos, and documents.

4 Table Store

4.1 What is Table Store

4.1.1 Technical background

Data features in the data technology (DT) era

As the mobile Internet becomes more common and widely adopted in various industries and fields , Internet applications present the following significant features and trends:

- The amount of data that needs to be stored and processed increases exponentially. The data includes microblogs, social events, pictures, and access logs.
- With the increase of mobile and IoT devices, the requirements for simultaneous writes for structured data storage also increase.
- Without schema, data tends to be semi-structured and data fields change dynamically.
- User access is characterized by hot spots and peak hours. For example, during promotional activities, user access soars in a few minutes.
- Constant access of users to the mobile Internet and users' HA requirements for Internet applications make failures inevitable, but users cannot bear unstable services caused by failures, even planned service failures.
- Large amounts of data significantly increase the requirements for the performance and scale of compute analysis.

Challenges of traditional IT software solutions

Traditional IT software solutions have the following trends and challenges:

- Scalability

  Traditional software, such as relational databases, is incapable of handling such fast-growing data. It bottlenecks data write throughput and access efficiency. With traditional database solutions, databases and tables are partitioned manually and statically. This method requires laborious maintenance workloads. In particular scenarios where nodes are added to increase the capacity, there is a need to repartition and migrate existing data. During this process, it is difficult to guarantee service performance, stability, and availability. The whole process is complex.
- Data model changes

Data in a traditional database is processed in accordance with a schema. The number of columns in data is fixed without frequent changes. Frequent changes to the table schema and column count affect service availability. Therefore, traditional solutions are incapable of handling the increasingly loosely structured data of Internet applications.

- Quick scaling

  In traditional solutions, business access loads are stable, and the system is not required to quickly scale resources. In this scenario, data needs repartitioning and migration, and the workload is laborious. Once business loads decline, excess hosts need to be removed to avoid low resource usage. Data needs to be migrated again. The entire process is extremely complex and inefficient.

- O&M guarantees

  With traditional software solutions, services are recovered when hardware (network devices or disks) failures occur. Hardware replacement, software upgrades, and configuration tuning and updates need to be performed manually. To ensure that applications are not aware of these processes and avoid service availability deterioration, users need a special engineering team to achieve system O&M. Therefore, workloads caused from recruitment and fund investment bring a huge challenge to fast-developing enterprises.

- Computing bottlenecks

  The current business system uses Online Transaction Processing (OLTP) to process and analyze data in relational databases such as MySQL and Microsoft SQL Server. These relational databases are adept at transaction processing. They maintain strong consistenc y and atomicity in data operations, and support frequent data insertion and modification. However, once the data volume exceeds system processing capability (query and computing) and reaches tens of millions or even billions of data records, or a complex calculation process is needed, OLTP database systems are no longer sufficient.

4.1.2 Table Store technologies

Table Store is a NoSQL data storage service built on Alibaba Cloud's Apsara system. Table Store partitions tables and dispatches data partitions to different nodes to improve scalability. In event of a single hardware failure, Table Store quickly detects the faulty node using the heartbeat mechanism and migrates data partitions from the defective node to a healthy node to continue service, thereby achieving rapid service backup.

Data partitioning and load balancing

The first primary key column in each row of a table is called the partition key. The system partitions a table into multiple partitions based on the range of the partition key. When the data in a partition exceeds a certain size, the partition is automatically split into two smaller partitions. The data and access loads are scattered to two partitions. The partitions are scheduled to different nodes. Eventually, the linear scalability of the single-table data scale and access loads is achieved .

Technical indicator: Table Store stores PBs of data in a single table and allows you to simultaneo usly read/write millions of data.

Automatic recovery of single point of failure

In the storage engine of Table Store, each node serves a number of data partitions in different tables. A master node monitors partition distribution and dispatching, and the health of each service node. If a service node fails, the master node migrates data partitions from this faulty node to other healthy nodes. The migration is logically performed, and does not involve physical entities, so services can recover from the single point of failure (SPOF) within several minutes.

Technical indicator: SPOF affects services of some data partitions only and services can recover within several minutes.

Zone-disaster recovery and geo-disaster recovery

To meet business security and availability requirements, Table Store provides active-standby cluster-based zone-disaster recovery and geo-disaster recovery. Disaster recovery supports instance-based recovery. Any table operation on the primary instance, including insertion, update , or deletion, is synchronized to the table of the same name on the secondary instance. The duration of data synchronization between the primary and secondary instances depends on the network environment of the active and standby clusters. In the ideal network environment, the synchronization latency reaches the millisecond level. Before the manual failover, you must stop resource access to the active cluster and wait for all data to be completely backed up. After the failover, do not perform any failover operation within one hour. Clear original cluster data and reset the standby cluster.

In the active-standby cluster-based zone-disaster recovery scenario, the endpoints remain unchanged when applications access Table Store in the active and standby clusters. In other words, the application endpoints do not need to be changed after the failover. In the active-

standby cluster-based geo-disaster recovery scenario, the endpoints of the active and standby clusters are different. After the failover, endpoints need to be changed for applications.

Technical indicator: The RTO of Table Store is less than 2 minutes, the RPO less than 5 minutes, and the RCO is 1.

4.2 Benefits

Built on Alibaba Cloud's Apsara system, Table Store is a NoSQL database service that enables you to store and access massive amounts of structured data in real time. Table Store organizes data into instances and tables, and achieves seamless scaling by using data partitioning and load balancing. It shields applications from faults and errors that occur on the underlying hardware platform and provides fast recovery capability and high service availability. Additionally, Table Store manages data with multiple data backups to solid state disks (SSDs), enabling quick data access and high data reliability. When using Table Store, you only pay for the resources you reserve and use, and do not need to handle complex issues such as cluster scaling, and upgrades and maintenance of database software and hardware.

Table Store comes with the following features:

- Scalability

  There is no upper limit to the amount of data that can be stored in Table Store tables. As data increases, Table Store adjusts partitions to provide more storage space for tables and improve the capability of handling access request bursts.
- Reliability

  Table Store provides high data reliability. It stores multiple data backups and quickly restore data when some backups become invalid.
- High availability

  Through automatic failure detection and data migration, Table Store shields applications from host- and network-related hardware faults to achieve high availability.
- Ease of management

  Table Store automatically performs complex O&M tasks, such as the management of data partitions, software and hardware upgrades, configuration updates, and cluster scale-out.
- Access security

Table Store provides multiple permission management mechanisms. It performs identity authentication and authorization for each application request to prevent unauthorized data access and ensure the security of data access.

- High consistency

  Table Store ensures high data consistency for data writes. After a write operation succeeds, three replicas are written to a disk. Applications can read the latest data immediately.

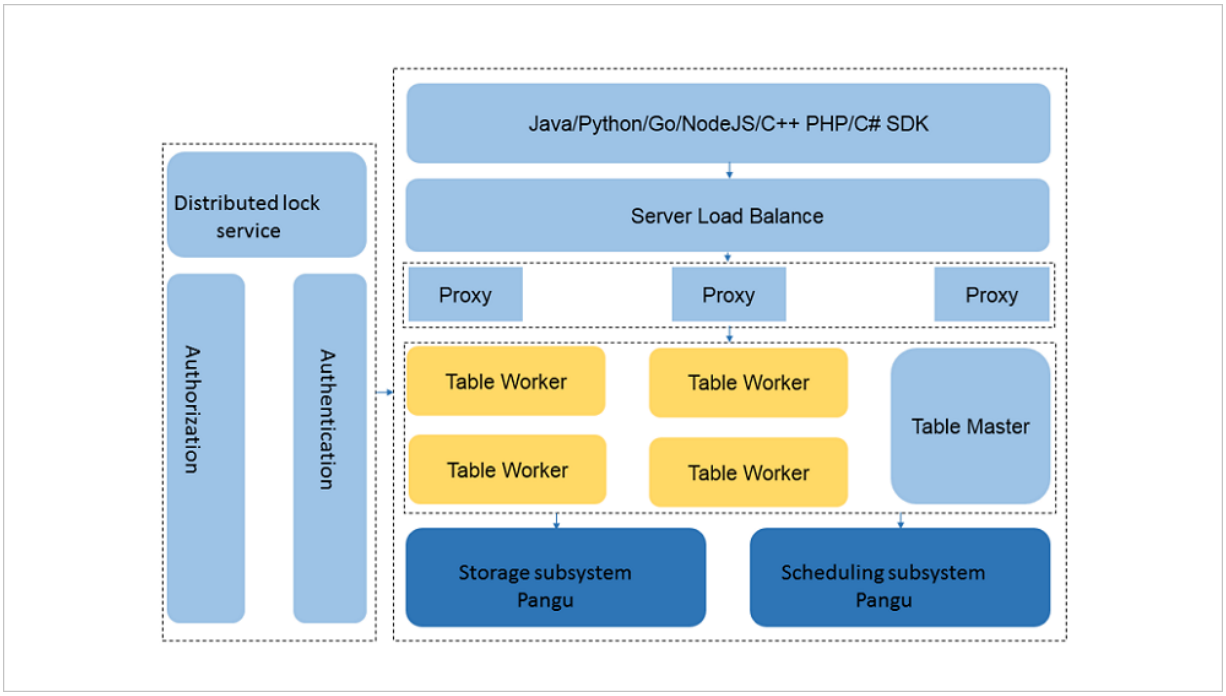
- Flexible data models

  Table Store tables do not require a fixed format. Each row can contain a different number of columns. Table Store supports multiple data types, such as integer, boolean, double, string, and binary.

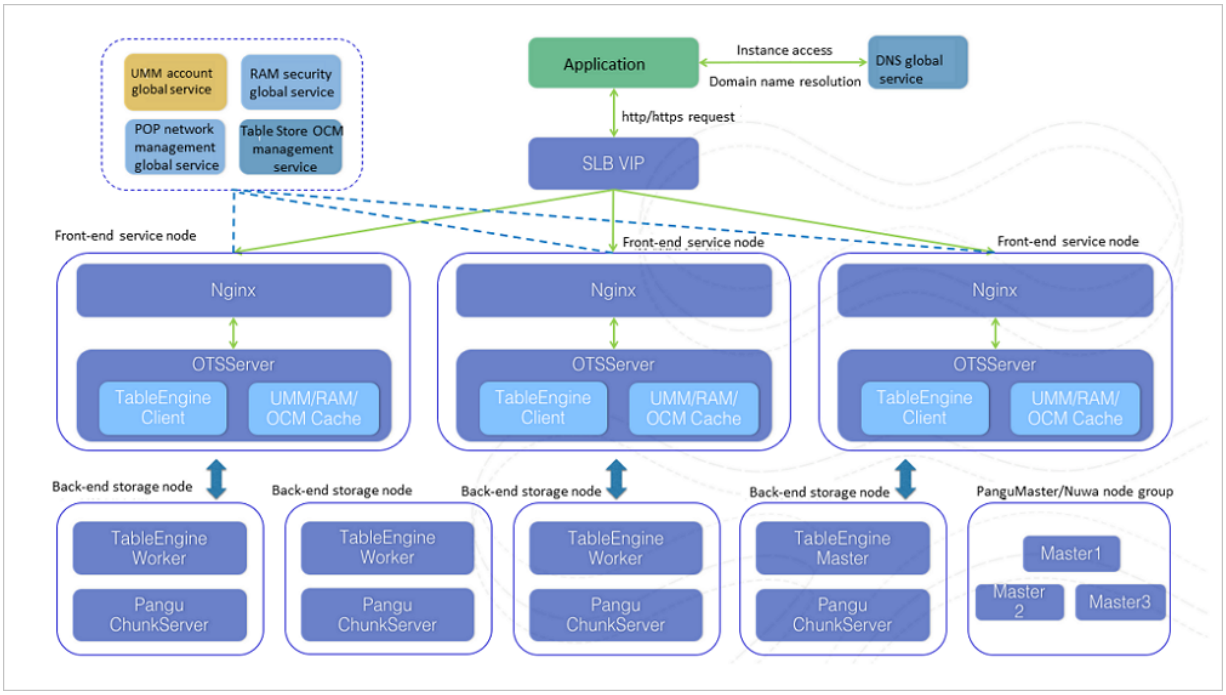
4.3 Architecture

The architecture of Table Store references the BigTable (one of the three core technologies of Google) and uses the log-structured merge-tree (LSM) storage engine to provide high performanc e writes. The performance of primary key-based single-row queries and range queries is stable and predictable. The performance is not affected by the volume of data and access concurrency.

The following figure shows the basic architecture of Table Store.



The following figure shows the detailed architecture of Table Store.



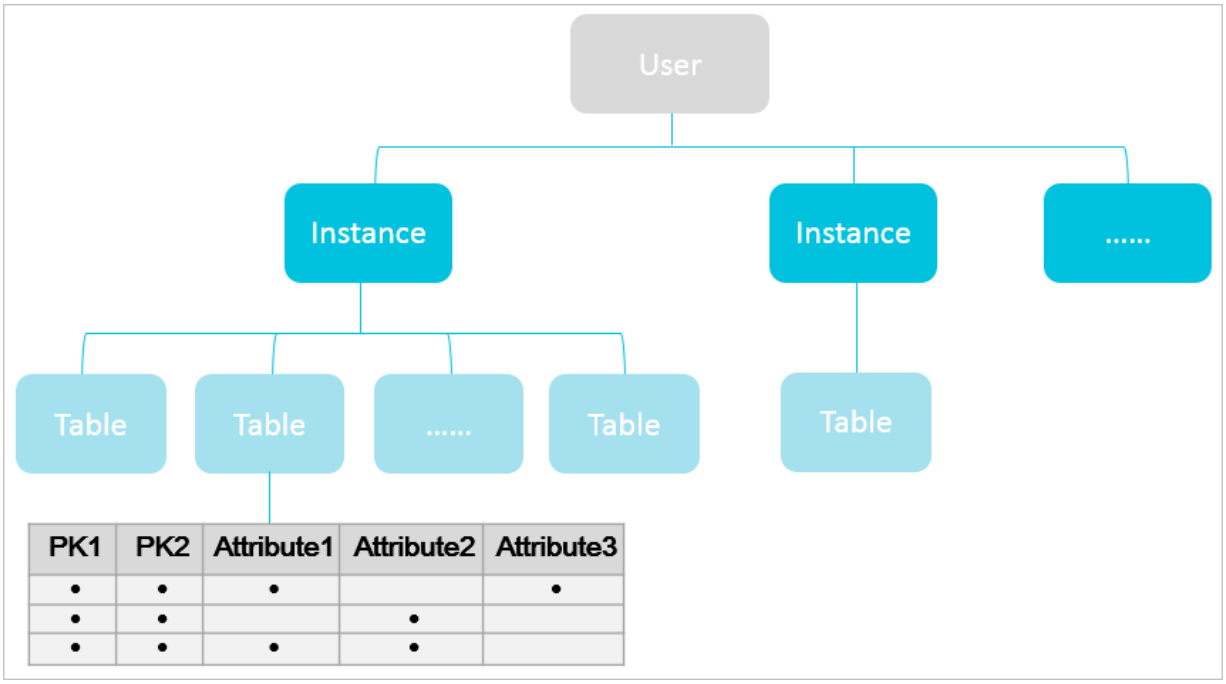
- The top layer is the protocol access layer. SLB distributes user requests to various proxy nodes . The proxy nodes receive requests that are sent through the RESTful protocol and implement security authentication. If the authentication succeeds, the user requests are forwarded to the corresponding data engine based on the value of the first primary key for further operations. If the authentication fails, the error information is directly returned to the user.
- Table Worker is the data engine layer that processes structured data through a primary key to search for or store data. Table Worker supports large-scale access request bursts.
- The bottom layer is the persistent storage layer. At this layer, the large-scale Apsara Distributed File System is deployed. Metadata is stored in Masters. The distributed message consistency protocol Paxos is adopted between Masters to ensure metadata consistency. In this scenario , efficient distributed file storage and access are achieved. This method guarantees three replicas of data stored in the system and system recovery from any hardware or software fault.

4.4 Features

4.4.1 Users and instances

The following figure shows Table Store architecture in relation to a user and instances.

Figure 4-1: User and instance architecture

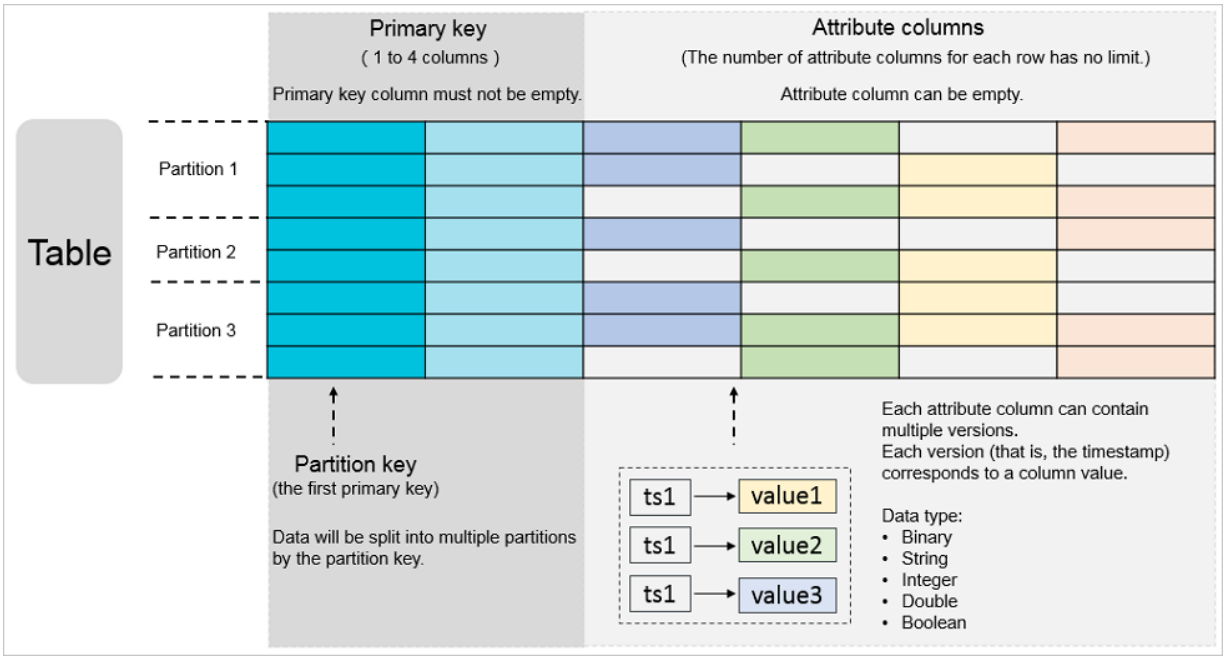


- Users can log on with an Apsara Stack account.
- Users can use fine-grained auditing for operations.
- Users organize resources through instances. A user can create multiple instances and use each instance to create and manage multiple data tables.
- An instance is the basic unit of multi-tenant isolation.
- User permissions can vary with their roles.

4.4.2 Data tables

The following figure shows the data table structure of Table Store.

Figure 4-2: Data table structure



- A data table is the basic unit of resource allocation.
- A table is a set of rows. A row consists of primary keys and attributes.
- A table partitions data based on the size of the first primary key column.
- All rows in a table must have the same quantity of primary key columns with the same names.
- The quantity of attribute columns in a row is variable. So are the names and data types of the attribute columns.
- There is no limit to the number of attribute columns contained in a row. However, the maximum number of attribute columns where each request can write data to is 1024.
- A table can contain hundreds of billions of rows and even more data.
- A table's capacity can store PBs of data.

4.4.3 Data partitioning

- A table partitions data based on the size of the first primary key column.
- The rows whose first primary key column values are within the same partition range are allocated to the same partition.
- To improve load balancing, Table Store splits and merges partitions based on specific rules.
- We recommend that data with the same partition do not exceed 10 GB.

4.4.4 Common commands and functions

Commands

- ListTable: lists all tables in an instance.
- CreateTable: creates a table.
- DeleteTable: deletes a table.
- DescribeTable: obtains attributes of a table.
- UpdateTable: updates the reserved read/write throughput configuration of a table.
- ComputeSplitPointsBySize: logically partitions all table data into several partitions with the specified size; returns the breakpoints between these partitions and the prompt of the hosts where partitions reside.

Functions

- GetRow: reads data in a row.
- PutRow: inserts a row of data.
- UpdateRow: updates a row of data.
- DeleteRow: deletes a row of data.
- BatchGetRow: reads multiple rows in one or more tables simultaneously.
- BatchWriteRow: inserts, updates, or deletes multiple rows in one or more tables simultaneously .
- GetRange: reads data from a table within a range.

4.4.5 Authorization and permission control

Table Store permissions

In addition to access control and private network support, Table Store supports the following permission control:

- Authorization of table-level operations
- API-level permission control
- Authentication of IP address limits, HTTPS, multi-factor authentication (MFA), and access time limits
- Temporary access authorization of STS
- Virtual Private Cloud (VPC) access

Apsara Stack Management Console-based permissions

- Provides account logons and authentication through Apsara Stack Management Console.
- Provides graphical instance creation, management, and deletion functions.
- Provides graphical table creation, management, deletion, and reserved read/write throughput adjustment functions.
- Displays table-level monitoring information.

5 Network Attached Storage (NAS)

5.1 What is NAS

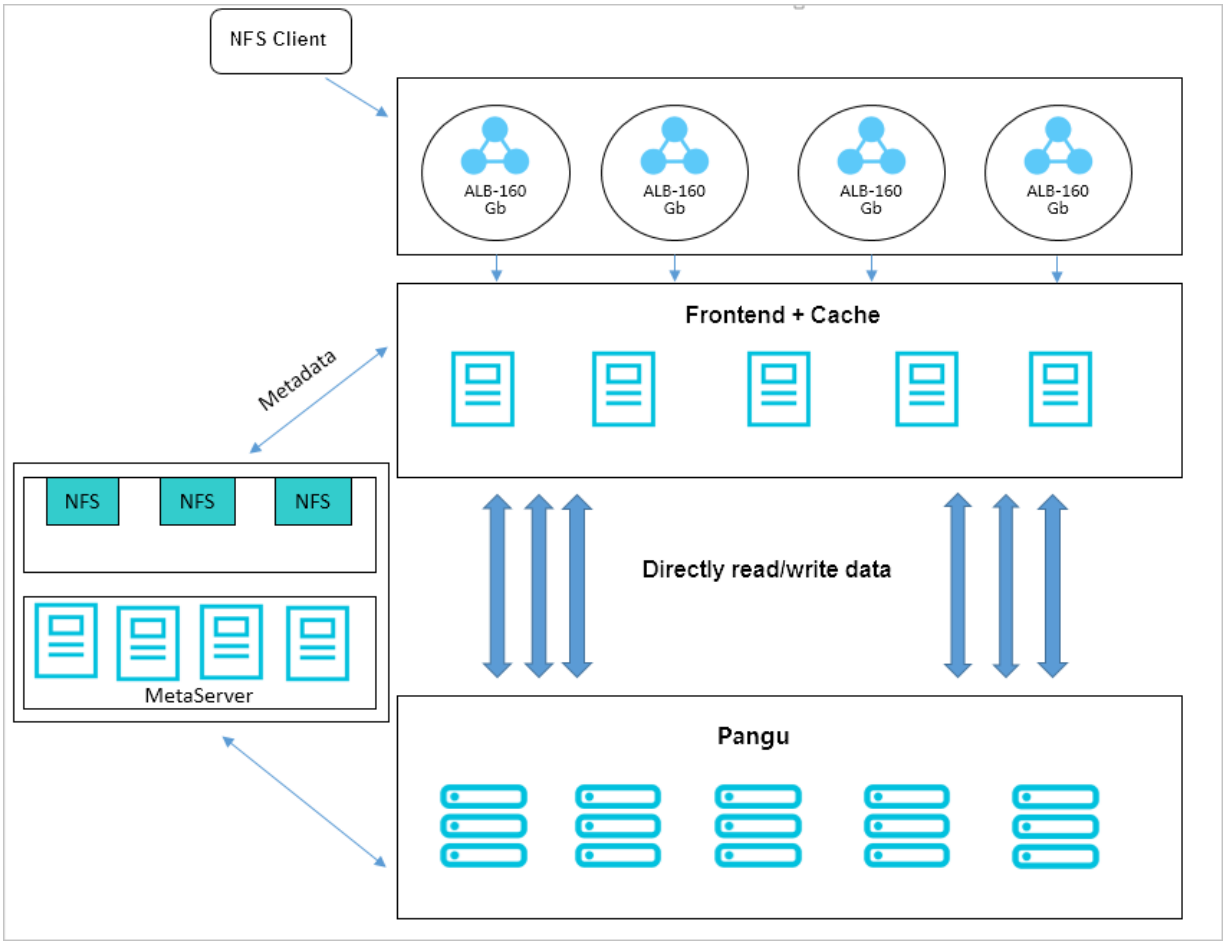
Alibaba Cloud Network Attached Storage (NAS) is a highly reliable, highly available file storage service for Alibaba Cloud ECS, E-HPC, and Container Service. The service features a distributed file system with unlimited capacity and performance scaling ability. It supports a single namespace and allows multiple user access. Additionally, standard file access protocols are supported. You do not need to modify your application to use the service.

5.2 Architecture

Based on Apsara distributed file system, Alibaba Cloud NAS stores and distributes three copies of each data file on multiple storage nodes. The frontend nodes receive connection requests from NFS clients. Deployed in a distributed fashion, these nodes are stateless with cache feature, and ensure frontend high availability. The metadata of the file system is stored on MetaServers. I/O requests from the client can directly access user data stored on backend nodes after obtaining the metadata of the file system from MetaServers.

Both the frontend and backend can expand elastically as demand changes, ensuring high availability, high throughput, and low latency.

Figure 5-1: Architecture



5.3 Features

Alibaba Cloud NAS supports the NFSv3 and NFSv4 protocols. Your applications can use this service without any modifications. Alibaba Cloud NAS can meet various file storage needs, including business file sharing, backend file storage for office automation systems, enterprise database backup and storage, system log storage and analysis, website data storage and distribution, and data storage during system development and testing.

Figure 5-2: Features

5.4 Benefits

Alibaba Cloud NAS provides the following benefits:

- Shared file system

  You can mount the same file system on 10,000 clients using the NFSv3 or NFSv4 protocols to achieve data sharing.

- High performance

  The maximum throughput of the cluster can reach 20 Gbit/s and the IOPS can reach more than 20,000.

- Scalability

You can purchase storage capacity as demand increases. The maximum capacity of a file system can reach 10 PB. Each file system can store a maximum of 1 billion files, and the maximum file size is 32 TB.

- High availability

  Based on Apsara distributed file system, Alibaba Cloud NAS maintains three copies for each data file to achieve high availability and guarantee data reliability.

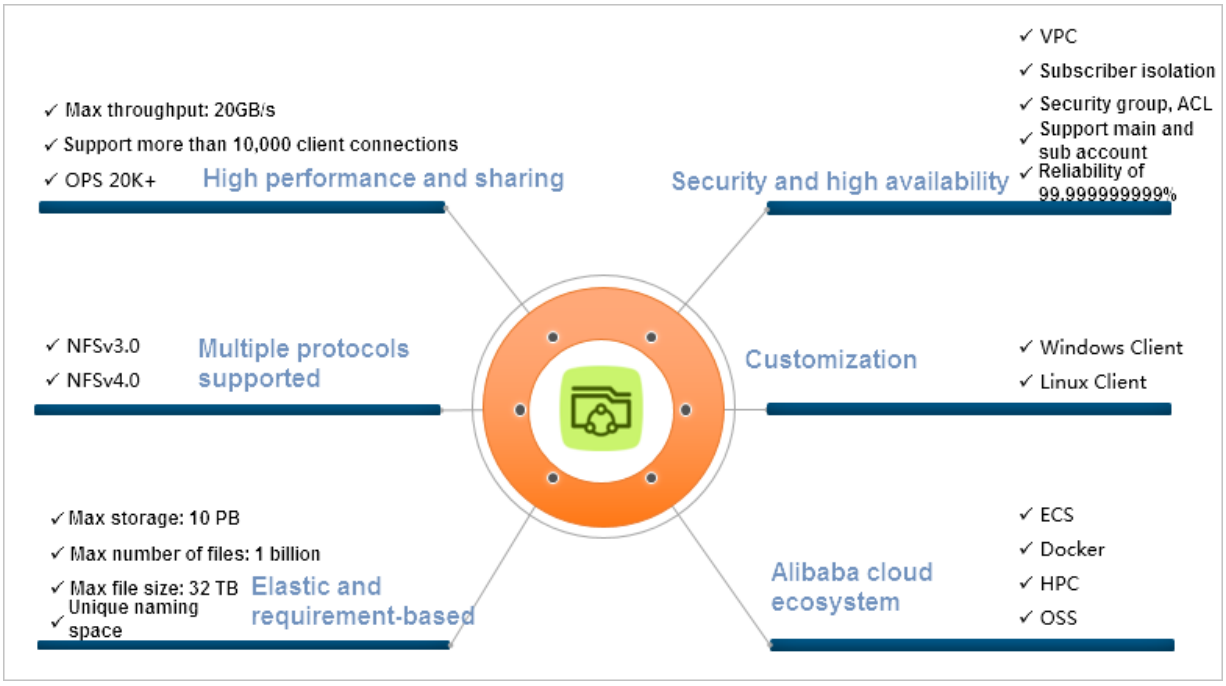
- Security

  Multiple security mechanisms such as VPC, security group, ACL, and account authorization are implemented to safeguard user data.

- Global namespace

  File data is distributed across the whole NAS cluster using a single namespace.

Figure 5-3: Benefits



6 Distributed File System (DFS)

6.1 What is DFS

Apsara Stack Distributed File System (DFS) is a file storage service for computing resources such as Apsara Stack ECS and Container Service instances. It supports standard Hadoop FileSystem interfaces. You can use DFS without the need to modify the existing applications of big data analytics. DFS features unlimited capacity, performance scale-out, single namespace, multi-tenancy, high reliability, and high availability.

After you create a DFS instance, computing resources such as ECS and Container Service instances can access the DFS instance through standard Hadoop FileSystem interfaces. Multiple computing nodes can access the same DFS and share files and directories.

6.2 Design philosophy

Background

With the application of big data and the development of Hadoop technology, there is an increased demand from users for distributed file systems. Compared with the traditional HDFS, a file system should have the following features as required by users.

- Cloud data intercommunication
- Support for calculation of stored data at any time
- Support for disaster recovery
- High performance and low cost

Alibaba has independently developed DFS based on the preceding requirements. DFS is a high-performance cloud-based distributed file system compatible with Hadoop.

Design philosophy

DFS is designed based on the following ideas.

- Thanks to DFSs compatibility with the Hadoop ecology, Hadoop applications can be seamlessly connected to massive cloudified data.
- With the integrated design of software and hardware, the system features extreme end-to-end performance and low cost.

- Due to intercommunication with existing Alibaba storage products (OSS/NAS), data can be accessed from Alibaba Cloud ECS computing resources and MaxCompute instances at any time.
- The system provides users with an excellent experience due to its smart management and O&M capabilities.

Core metrics

The following describes the core metrics of DFS design.

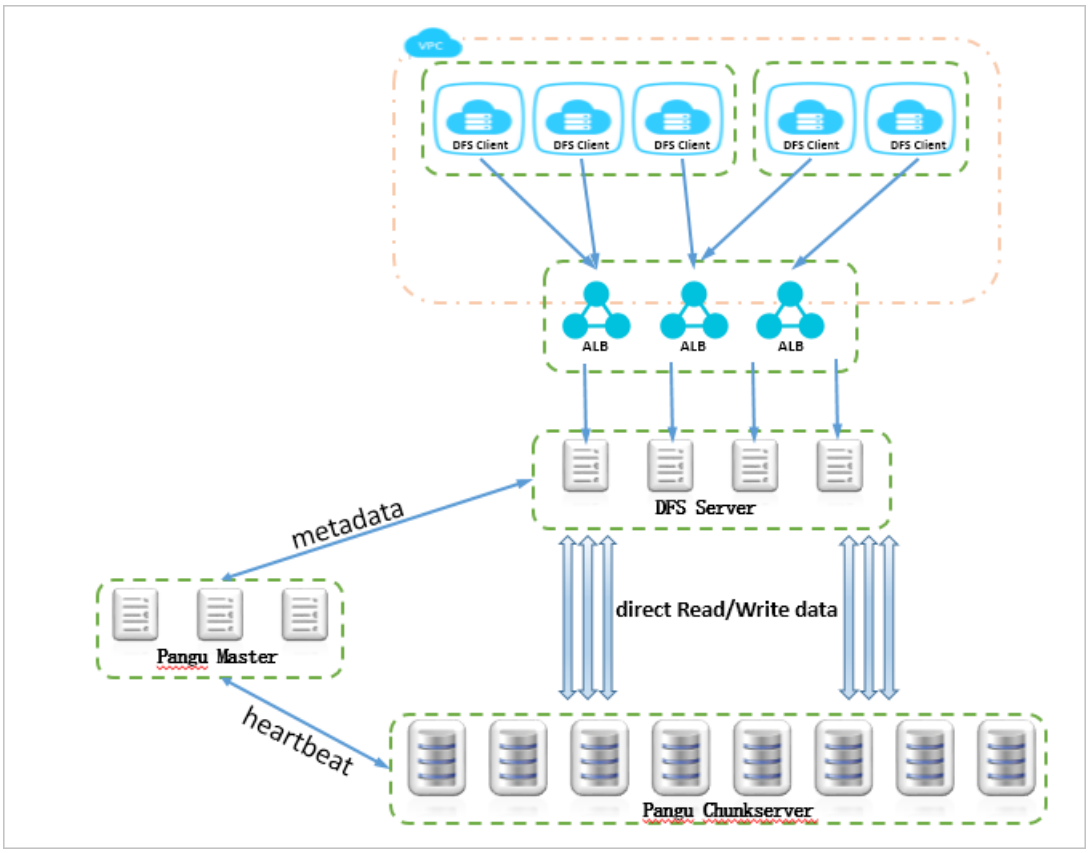
- The system is compatible with HDFS interfaces.
- The system supports disaster recovery for three available zones (AZs) with 99.99% availability.
- A single storage node features a throughput of 20 GB with support for linear expansion.
- A single cluster can be composed of more than 10,000 machines.
- Automatic O&M and management are supported.

6.3 Architecture

The architecture of DFS consists of the frontend and backend.

The backend is based on Apsara Distributed File System. Data is replicated into multiple copies and stored in Apsara Distributed File System. Frontend access nodes of DFS receive and cache connection requests that are sent from computing resources such as ECS instances, Hadoop computing applications (such as MapReduce and Spark), or Container Service instances. Apsara Distributed File System also manages metadata and data of DFS.

The following figure shows the overall DFS architecture.



6.4 Product positioning

Comparison with a traditional HDFS

A traditional HDFS is designed to allow high-throughput data access to support applications on large-scale datasets. DFS has advantages over HDFS in system elasticity, small file storage performance, and support for multiple tenants.

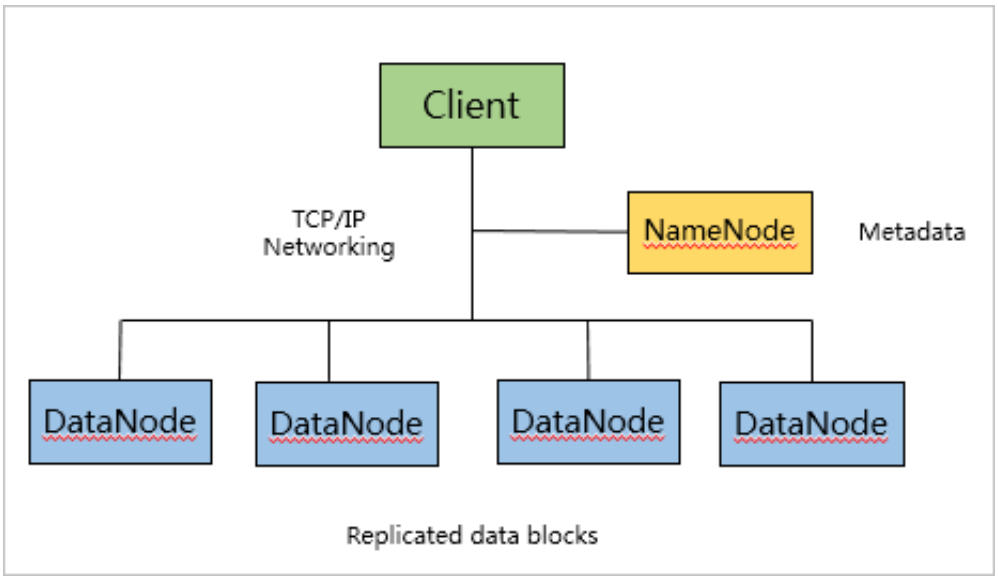
System elasticity

As data nodes of a traditional HDFS need computing capabilities, HDFS is not flexible enough in planning and capacity scaling.

DFS is divided into two clusters during deployment: storage cluster and public service cluster. The storage cluster and computing cluster are independent of each other, and can be planned and scaled separately. After computing, resources can be released and stored data can be retained. In this way, the system elasticity is improved.

Small file storage performance

The following figure shows the architecture of a traditional HDFS.



As shown in the figure, data in HDFS is chunked and copied to multiple data nodes, and metadata of file blocks is saved and maintained in the name node only . For all file I/O requests, clients must obtain metadata from the name node and then write data in each data node in chain mode. In addition, born in the Hard Disk Drive (HDD) era, HDFS cannot adapt to new flash media, such as Solid State Drive (SSD) and NVM Express (NVMe).

The architecture of HDFS allows the system to read large data sets in stream mode. However, HDFS has obvious disadvantages in processing small files because a large amount of small files generate massive metadata, which can exhaust the resources of the name node. In addition, all file operation requests must be sent through the name node, which lengthens the latency of data access.

The architecture of DFS caters to the demand for storage of small files. The system significan tly improves the throughput of small files and storage efficiency through support for hierarchical storage, optimization of metadata, and optimization of write streams.

Support for multiple tenants

A traditional HDFS is designed for the single-cluster/single-tenant mode in non-cloud computing environments. Data nodes are intended for local data with storage and computing capabilities. Therefore, the computing system is closely bound to the storage system.

As compared to HDFS, DFS is based on a cloud computing environment with native support for cloud computing. As a cloud storage system, DFS establishes multiple DFS instances for multiple tenants in the storage system, and each instance supports operations through multiple computing clusters.

6.5 Product values

DFS has the following benefits in terms of convenience, performance, and cost.

Convenient access

- DFS is compatible with HDFS interfaces. The existing Hadoop applications can be accessed without costs.
- Data can be pooled through data intercommunication with Alibaba Cloud storage products.
- MaxCompute can be accessed conveniently.

Excellent performance

- The small file throughput performance is significantly improved through optimization in scenarios of small files.
- The system makes full use of hardware advantages of new-generation flash media.

Cost saving

- The system supports storage cloudification and auto scaling, which saves storage costs.
- The system supports intercommunication with existing stored data and enables computing capabilities, which saves time.
- Massive small files can be stored, which reduces data management costs.

6.6 Use cases

Shared storage and high availability

User data can be imported to the DFS instance in real time or in batches through standard HDFS interfaces. DFS supports high-availability shared access to files.

Big data analysis and machine learning

Data stored in the DFS instance can be directly accessed through MaxCompute, ECS virtual machines, or other computing resources. Hadoop or other machine learning applications deployed on multiple computing resources access data directly through HDFS interfaces to compute online /offline or output results to the DFS instance and store them permanently. DFS supports high-throughput and low-latency access so that you do not need to migrate data to local computing resources.

7 ApsaraDB for RDS

7.1 What is ApsaraDB for RDS?

ApsaraDB for Relational Database Service (RDS) is a stable, reliable, and automatically scalable online database service.

Based on the distributed file system and high-performance storage, ApsaraDB for RDS allows you to easily perform database operations and maintenance with its set of solutions for disaster tolerance, backup, recovery, monitoring, and migration.

ApsaraDB for RDS supports three storage engines including MySQL, PostgreSQL, and PPAS. They help you conveniently and rapidly create database instances suitable for your scenarios.

ApsaraDB RDS for MySQL

Originally based on a branch of MySQL, ApsaraDB RDS for MySQL has proven its performanc e and throughput during the high-volume traffic from concurrent users experienced during the Double 11 Shopping Festival. ApsaraDB RDS for MySQL provides basic functions such as backup and recovery, whitelist configuration for instances, Transparent Data Encryption (TDE), data migration, and management for instances, accounts, and databases. It also provides the following advanced features:

- **Read-only instances:** In scenarios where there are few write requests but a large number of read requests, you can enable read/write splitting to distribute read requests away from the primary instance. Read-only instances allow ApsaraDB RDS for MySQL 5.6 to auto-scale reading capabilities and increase the application throughput when large amounts of data is being read.
- **Read/write splitting:** Read/write splitting provides an extra address that links the primary instance with all of its read-only instances, which can act as a link for automatic read/write splitting. An application can use this method to read and write data by connecting to the same read/write splitting address. Write requests are automatically routed to the primary instance while read requests are routed to the read-only instances based on their weights. To scale up the reading capability of the system, you can add more read-only instances. No application changes are required.
- **Data compression:** ApsaraDB RDS for MySQL 5.6 allows you to compress data by using the TokuDB storage engine. Extensive testing show that the data volume is reduced by 80% to 90% after data tables are transferred from the InnoDB storage engine to the TokuDB storage

engine. For example, 2 TB of data can be compressed to 400 GB or less using TokuDB. In addition to data compression, TokuDB supports transaction and online DDL operations. It is compatible with MyISAM and InnoDB applications.

ApsaraDB RDS for PostgreSQL

ApsaraDB RDS for PostgreSQL is an advanced open source database system with full SQL compliance and support for a diverse range of data formats such as JSON, IP, and geometric data. In addition to excellent support for features such as transactions, subqueries, multi-version concurrency control (MVCC), and data integrity check, ApsaraDB RDS for PostgreSQL integrates a series of important functions including high availability, backup, and recovery to facilitate your operations and maintenance burden.

ApsaraDB RDS for PostgreSQL provides basic functions such as whitelist configuration for instances, backup and recovery, data migration, and management for instances, accounts, and databases.

ApsaraDB RDS for PPAS

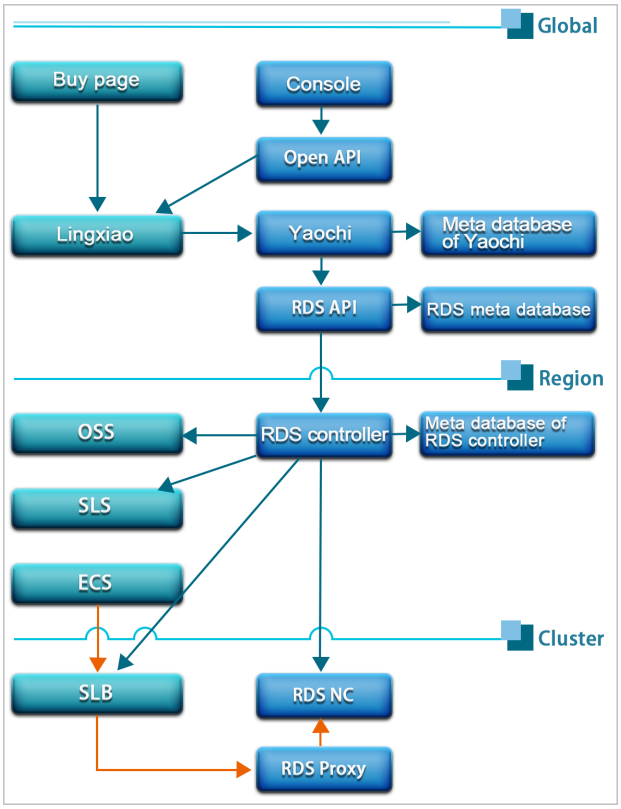
ApsaraDB RDS for Postgres Plus Advanced Server (PPAS) is a stable, secure, and scalable enterprise-class relational database based on PostgreSQL. ApsaraDB RDS for PPAS enhances the performance, application solutions, and compatibility of PostgreSQL and is able to run Oracle applications directly. You can run enterprise-class applications on ApsaraDB RDS for PPAS to implement table and cost-effective services.

ApsaraDB RDS for PPAS provides basic functions such as whitelist configuration for instances, backup and recovery, data migration, and management for instances, accounts, and databases.

7.2 Architecture

The following figure shows the system architecture of ApsaraDB for RDS.

Figure 7-1: RDS system architecture

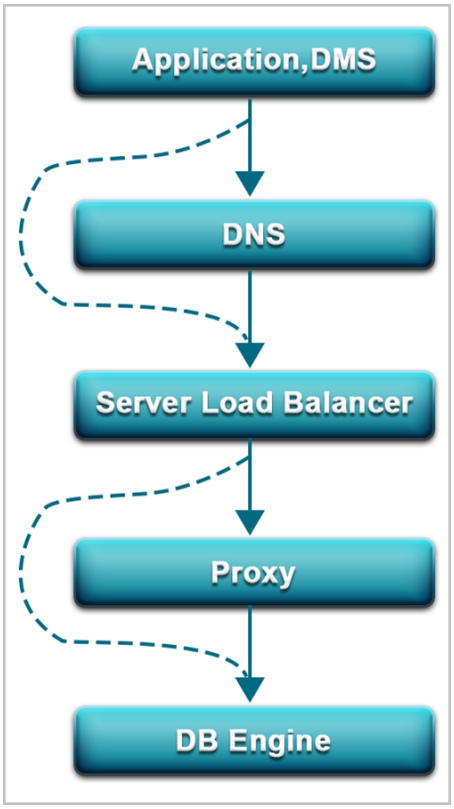


7.3 Features and principles

7.3.1 Data link service

The data link service allows you to carry out operations on data, such as add, delete, modify, and query table structure.

Figure 7-2: RDS data link service



DNS

The DNS module can dynamically resolve domain names into IP addresses. Therefore, IP address changes do not affect the performance of RDS instances, as shown in the following example.

Imagine that the domain name of an ApsaraDB for RDS instance is test.rds.aliyun.com, and the IP address corresponding to this domain name is 10.1.1.1. The instance can be accessed when either test.rds.aliyun.com or 10.1.1.1 is configured in the connection pool of a program.

After a zone migration or version upgrade is performed for this ApsaraDB for RDS instance, the IP address may change to 10.1.1.2. If the domain name test.rds.aliyun.com is configured in the connection pool, the instance can still be accessed. However, if the IP address 10.1.1.1 is configured in the connection pool, the instance will be inaccessible.

SLB

The SLB module provides both the private and public IP addresses of an ApsaraDB for RDS instance. Therefore, server changes do not affect the performance of the instance, as shown in the following example.

Imagine that the private IP address of an ApsaraDB for RDS instance is 10.1.1.1, and the corresponding Proxy or DB Engine runs on 192.168.0.1. The SLB module typically redirects all traffic destined for 10.1.1.1 to 192.168.0.1. If 192.168.0.1 fails, another server in hot standby status with the IP address 192.168.0.2 takes over for the server with the IP address 192.168.0.1 . In this case, the SLB module will redirect all traffic destined for 10.1.1.1 to 192.168.0.2, and the ApsaraDB for RDS instance will continue to provide services normally.

Proxy

The Proxy module provides a number of features including data routing, traffic detection, and session persistence.

- Data routing: aggregates the distributed complex queries found in big data scenarios and provides the corresponding capacity management capabilities.
- Traffic detection: reduces SQL injection risks and supports SQL log backtracking when necessary.
- Session persistence: prevents interruptions to the database connection when faults occur.

DB Engine

The following table describes the mainstream database protocols supported by RDS.

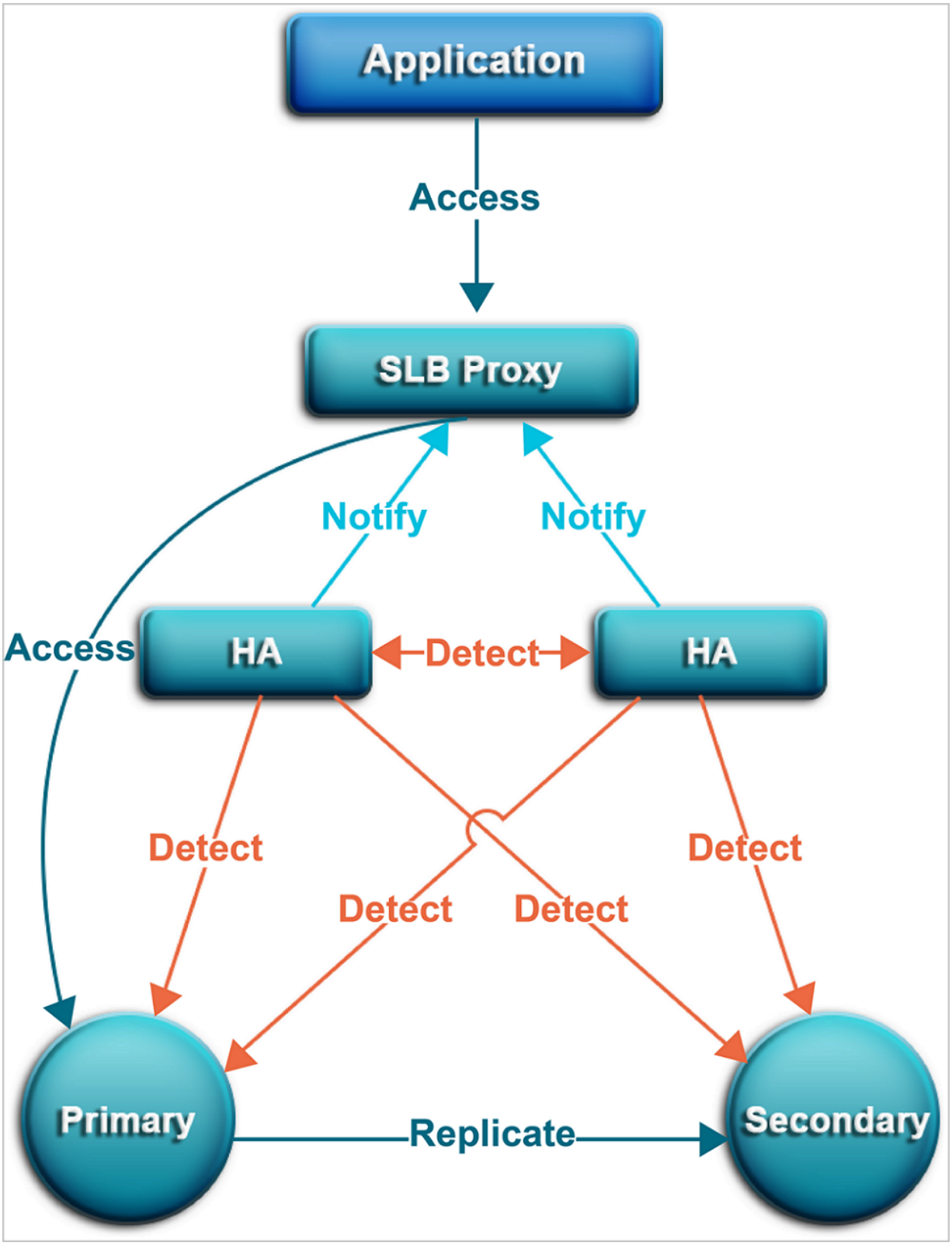
Table 7-1: RDS database protocols

RDBMS	Version
MySQL	5.6 (including read-only instances)
PostgreSQL	9.4
PPAS	9.3/9.6

7.3.2 High-availability service

The high-availability (HA) service ensures the availability of data link services and processes internal database exceptions. The HA service is implemented by deploying multiple HA nodes.

Figure 7-3: RDS HA service



Detection

The Detection module checks whether the primary and secondary nodes of the DB Engine are providing services normally.

The HA node uses heartbeat information taken at 8 to 10 second intervals to determine the health status of the primary node. This information, along with the health status of the secondary node and heartbeat information from other HA nodes, provides a reference for the Detection module. All this information helps the module avoid misjudgment caused by exceptions such as network jitter. Failover can be completed within 30 seconds.

Repair

The Repair module maintains the replication relationship between the primary and secondary nodes of the DB Engine. It can also correct errors that occur on either node during normal operations, as shown in the following examples:

- It can automatically restore primary/secondary replication after a disconnection.
- It can automatically repair table-level damage to the primary or secondary node.
- It can save and automatically repair the primary or secondary node in case of crashes.

Notice

The Notice module informs the SLB or Proxy module of status changes to the primary and secondary nodes to ensure that you always access the correct node, as shown in the following example.

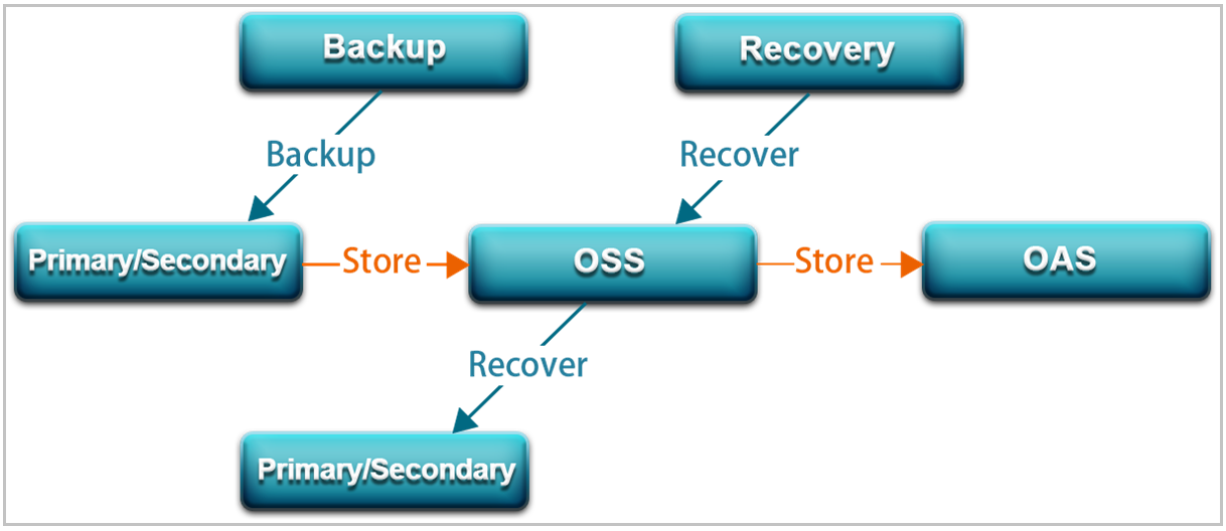
Imagine that the Detection module discovers problems with the primary node and instructs the Repair module to resolve these problems. If the Repair module fails to resolve a problem, it instructs the Notice module to perform traffic switchover. The Notice module forwards the switching request to the SLB or Proxy module, and then all traffic is redirected to the secondary node.

Meanwhile, the Repair module creates a new secondary node on a different host and synchronizes this change back to the Detection module. The Detection module rechecks the health status of the instance to ensure it is healthy.

7.3.3 Backup service

This service supports offline data backup, dump, and recovery functions.

Figure 7-4: RDS backup service



Backup

The Backup module compresses and uploads data and logs on both the primary and secondary nodes. ApsaraDB for RDS uploads backup files to OSS by default and dumps the backup files to a more cost-effective and persistent Archive Storage system. When the secondary node is operating properly, backups are always initiated on the secondary node so as not to affect the services on the primary node. When the secondary node is unavailable or damaged, the Backup module creates backups on the primary node.

Recovery

The Recovery module restores backup files stored on OSS to a destination node. The Recovery module provides the following features:

- Primary node rollback: when an operation error occurs, rolls back the primary node to a specified point in time.
- Secondary node repair: when an irreparable fault occurs on the secondary node, creates a new secondary node to reduce risk.
- Read-only instance creation: creates a read-only instance from backup files.

Storage

The Storage module can upload, dump, and download backup files.

Currently, all backup data is uploaded to OSS for storage. You can obtain temporary links to download this data as needed.

Note:

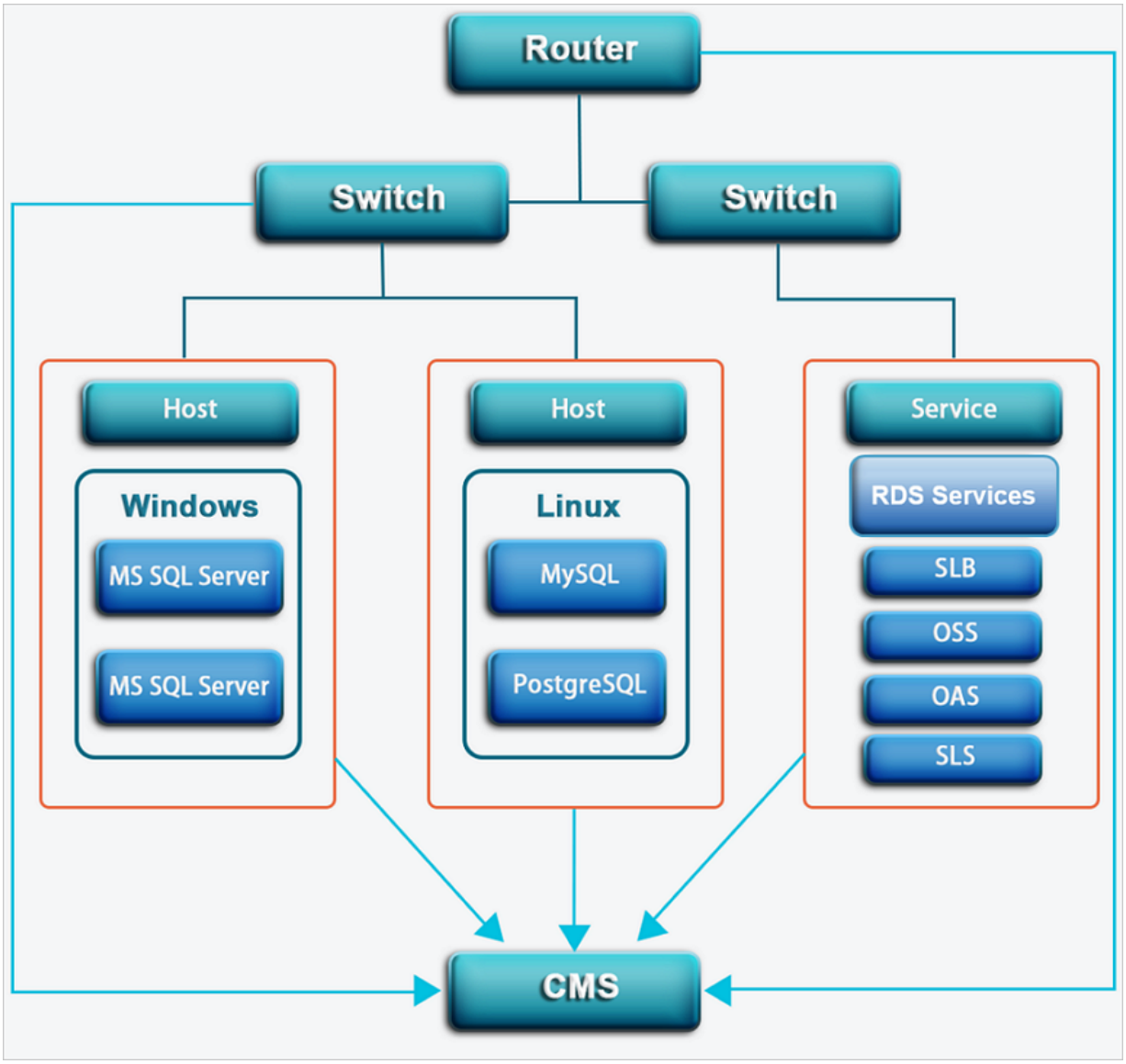
ApsaraDB RDS for PPAS does not support the download of backup files.

In certain scenarios, the Storage module allows you to dump backup files from OSS to Archive Storage for more cost-effective and longer-term offline storage.

7.3.4 Monitoring service

ApsaraDB for RDS provides multilevel monitoring services across the physical, network, and application layers to ensure service availability.

Figure 7-5: RDS monitoring service



Service

The Service module tracks the status of services that RDS depends on, such as Server Load Balancing (SLB), OSS, Archive Storage, and log service, to ensure they are operating properly. Monitored metrics include functionality and response time. The Service module also uses logs to determine whether internal ApsaraDB for RDS services are operating properly.

Network

The Network module tracks statuses at the network layer. The monitored items include:

- The connectivity between ECS and ApsaraDB for RDS
- The connectivity between physical RDS servers
- The rates of packet loss on the VRouter and VSwitch

OS

The OS module tracks the statuses of hardware and the OS kernel. The monitored items include:

- Hardware maintenance: The OS module constantly checks the operating status of the CPU, memory, motherboard, and storage device. It can predict faults in advance and automatically submit repair reports when it determines a fault is likely to occur.
- OS kernel monitoring: The OS module tracks all database calls and analyzes the causes of slow calls or call errors based on the kernel status.

Instance

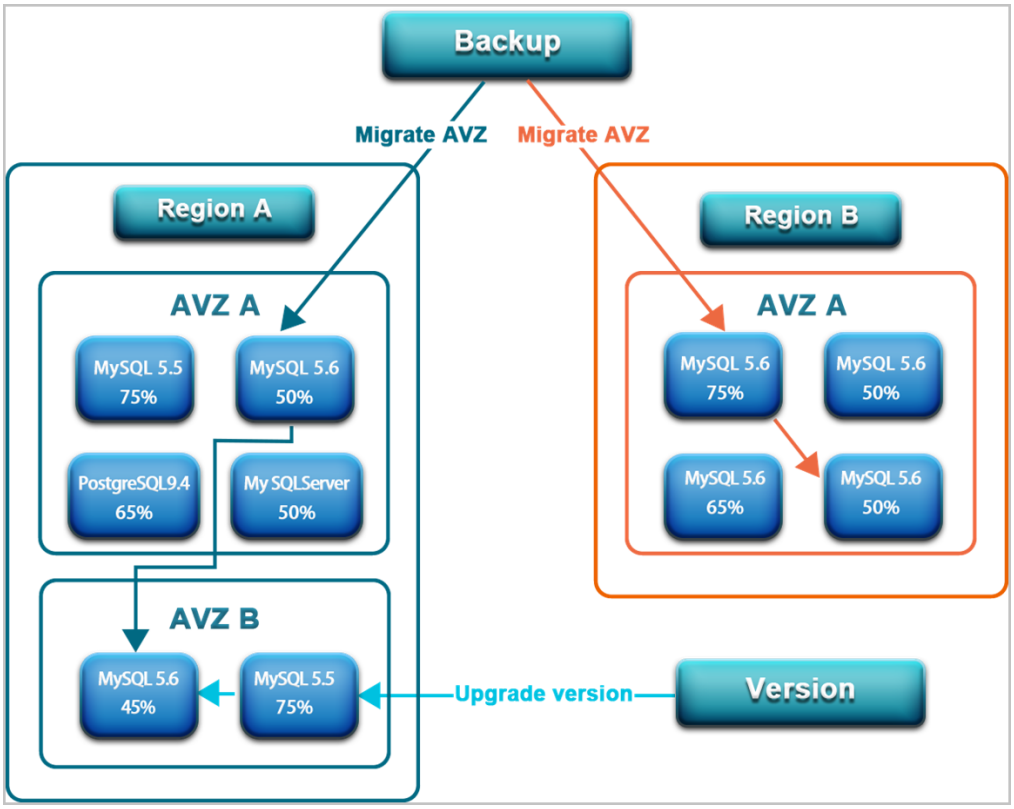
The Instance module collects the following information about ApsaraDB for RDS instances:

- Instance availability information
- Instance capacity and performance metrics
- Instance SQL execution records

7.3.5 Scheduling service

You can use the scheduling service to allocate resources and manage instance versions.

Figure 7-6: RDS scheduling service



Resource

The Resource module implements the scheduling of resources and services. It allocates and integrates underlying RDS resources when you activate and migrate instances. When you use the RDS console or API to create an instance, the Resource module calculates the most suitable host to carry the traffic to and from the instance. A similar process occurs during ApsaraDB for RDS instance migration.

After repeated instance creation, deletion, and migration operations, the Resource module calculates the degree of resource fragmentation and regularly integrates resources to improve service carrying capacity.

8 KVStore for Redis

8.1 What is KVStore for Redis

8.1.1 Overview

KVStore for Redis is an online storage service compatible with the Redis protocol. It supports multiple data types, such as the string, list, set, sorted set, and hash. It also supports advanced features such as transactions and subscribe-publish (Sub/Pub). KVStore for Redis meets persistent storage requirements and provides fast read/write capabilities by using a combined flash memory and hard disk storage architecture.

KVStore for Redis is used as a cloud computing service, with hardware and data deployed on the cloud, providing comprehensive infrastructure planning, network security protection, and system maintenance services.

8.1.2 Scenarios

Game industry scenarios

Game companies can use KVStore for Redis as an important part of their deployment architecture .

Scenario 1: Using KVStore for Redis for data storage

Game deployment architecture is relatively simple. The main program is deployed on ECS, while KVStore for Redis acts as a persistent database to store all business data. KVStore for Redis supports the persistence function, with primary/secondary redundant data storage.

Scenario 2: Using KVStore for Redis as a cache to accelerate application access

KVStore for Redis can be used as a cache layer to accelerate application access. Data is stored in the backend database (RDS).

Reliability is critical to KVStore for Redis. If KVStore for Redis becomes unavailable, business access may overload the backend database. KVStore for Redis provides a hot standby high-availability architecture that ensures extremely high reliability. The primary node provides external services. If this node fails, the system automatically fails over to the secondary node. You do not have to perform any operation for the failover process.

Live streaming scenarios

Live streaming business relies on KVStore for Redis to store user data and relationship informatio n.

Hot standby ensures high availability

KVStore for Redis provides the hot standby mode to maximize service availability.

Cluster instances overcome the performance bottleneck

KVStore for Redis provides cluster instances to overcome the performance bottleneck of the single-thread mechanism. This approach can effectively cope with traffic bursts in the live streaming business and meet high-performance requirements.

Easy scaling helps cope with business peaks

KVStore for Redis supports one-click scaling. The entire upgrade process is completely invisible to users and this helps cope with business impact caused by traffic bursts.

E-commerce industry scenarios

KVStore for Redis is widely used in the E-commerce industry to display items for sale and provide shopping recommendations.

Scenario 1: Time-limited shopping systems

During large-scale time-limited promotions, the shopping system is often overwhelmed by large amounts of traffic. The traffic far exceeds the read/write capabilities of common databases.

The persistence function supported by KVStore for Redis allows you to directly use it as a database system.

Scenario 2: Inventory system with a counter

In such a system, the underlying architecture keeps actual data in RDS and count information in database fields. In contrast, KVStore for Redis reads the counts, while RDS stores the count information. In this scenario, KVStore for Redis is deployed on a physical machine. The underlying architecture is based on SSD high-performance storage that can provide high data storage capabilities.

8.1.3 Benefits

High performance

- KVStore for Redis provides cluster functions. You can create cluster instances of 128 GB or more to meet large capacity and high performance requirements.

- Cluster instances of 32 GB or less can be deployed as primary/secondary instances. These instances are sufficient to meet the requirements for capacity and performance of regular users .

Auto scaling

- One-click storage scaling: You can use the console to adjust the storage capacity of your instances as needed.
- Online scaling without service interruption: Instance storage capacity can be scaled online without having to stop services or affecting your business.

Resource isolation

Instance-level resource isolation provides enhanced stability for individual services.

Data security

- Persistent data storage: Memory and hard disk storage mode meets data persistence requirements while providing high-speed data read/write capabilities.
- Primary/secondary copies for data: All data on the primary node has a backup copy on the secondary node.
- Access control: Password authentication is required for secure access.
- Encrypted data transmission: Secure Sockets Layer (SSL) and Transport Layer Security (TLS) are used to protect data security.

High availability

- Each instance has a primary node and a secondary node: This prevents service interruption caused by SPOF.
- Automatic detection and recovery of hardware failure: KVStore for Redis can automatically detect hardware failures and fail over to the secondary node, recovering the service within seconds.

Ease of use

- Out-of-the-box: This service requires no setup or installation and can be used right after purchase for quick and convenient business deployment.
- Compatible with the Redis protocol: This service is compatible with Redis commands, and any Redis client can easily establish a connection with KVStore for Redis to perform data operations.

8.2 Features

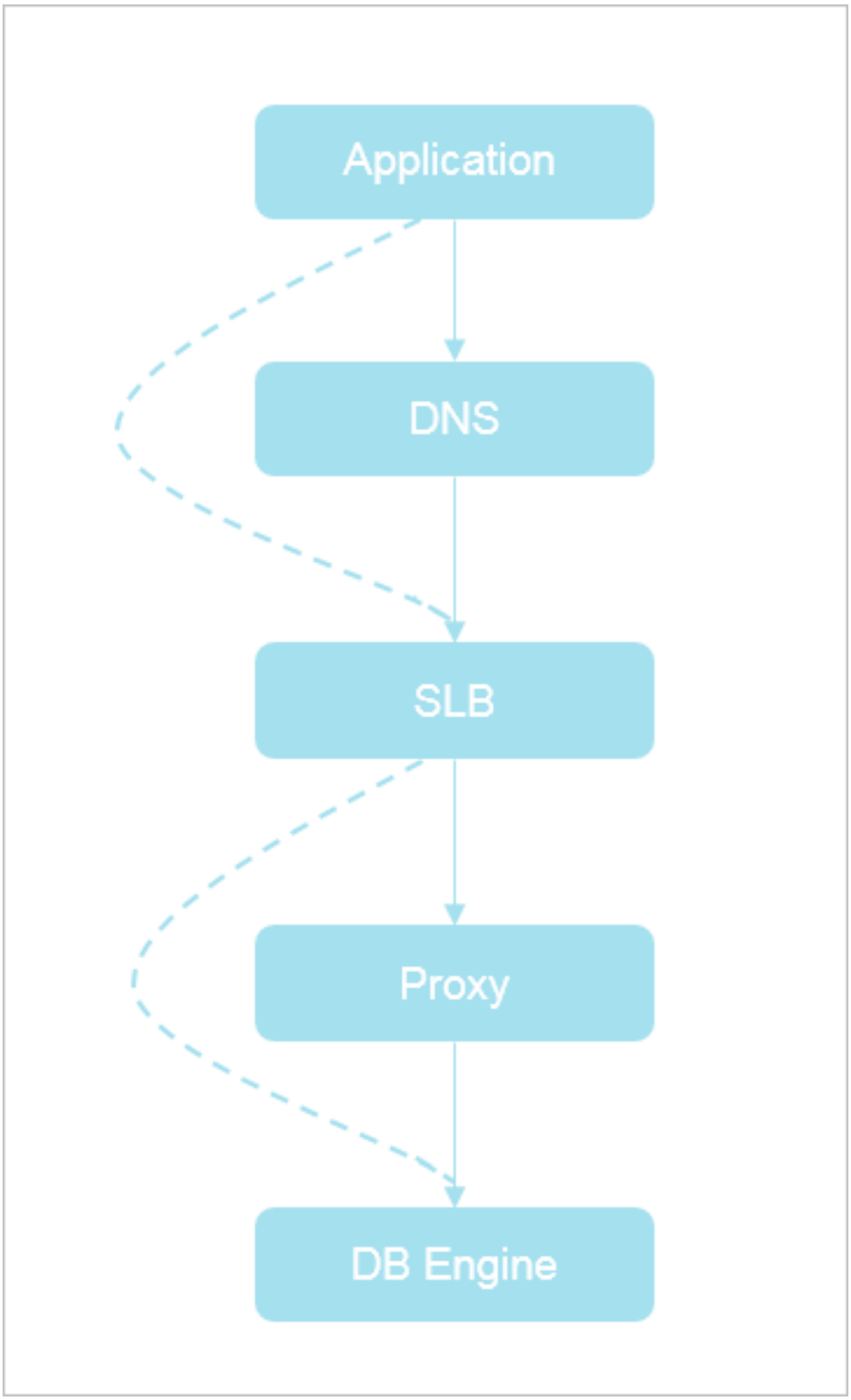
The high-availability KVStore for Redis service provides four core services:

- Data link service
- High-availability service
- Monitoring service
- Scheduling service

8.2.1 Data link service

The data link service allows you to control data, such as adding, deleting, modifying, and querying data.

You can connect to the KVStore for Redis service through an application.



8.2.1.1 DNS

The DNS module supports the dynamic resolution of domain names into IP addresses to avoid KVStore for Redis instances unreachable due to the change of IP addresses.

For example, assume that the domain name of an KVStore for Redis instance is **test.kvstore.aliyun.com**, and the IP address corresponding to this domain name is **10.1.1.1**.

You can access the KVStore for Redis instance if **test.kvstore.aliyun.com** or **10.1.1.1** is added to the connection pool of the application you are using.

If the KVStore for Redis instance is migrated to another host upon failover or its version is upgraded, the IP address may change to **10.1.1.2**.

If the domain name **test.kvstore.aliyun.com** has been added to the application's connection pool, you can still access the instance.

If the IP address **10.1.1.1** has been added, however, the instance is unreachable.

8.2.1.2 SLB

The Server Load Balancer module provides instance IP addresses to avoid KVStore for Redis instances unreachable due to the change of physical servers.

For example, assume that an KVStore for Redis instance has an intranet IP address of **10.1.1.1** and the corresponding Proxy or DB Engine runs on the host at **192.168.0.1**. Normally, the Server Load Balancer module redirects all traffic destined for **10.1.1.1** to **192.168.0.1**.

If **192.168.0.1** fails, its hot-standby host at **192.168.0.2** takes over for **192.168.0.1**. In this case, the Server Load Balancer module redirects the traffic for **10.1.1.1** to **192.168.0.2**, and the KVStore for Redis instance provides services normally.

8.2.1.3 Proxy

The Proxy module provides data routing, traffic detection, and session persistence functions.

- Data routing: KVStore for Redis supports a cluster-based architecture and implements complex query and partition policies for distributed routes.
- Traffic detection: This reduces the risks from cyberattacks that make use of Redis vulnerabil ities.
- Session persistence: This prevents database connection interruptions if any fault occurs.

8.2.1.4 DB engine

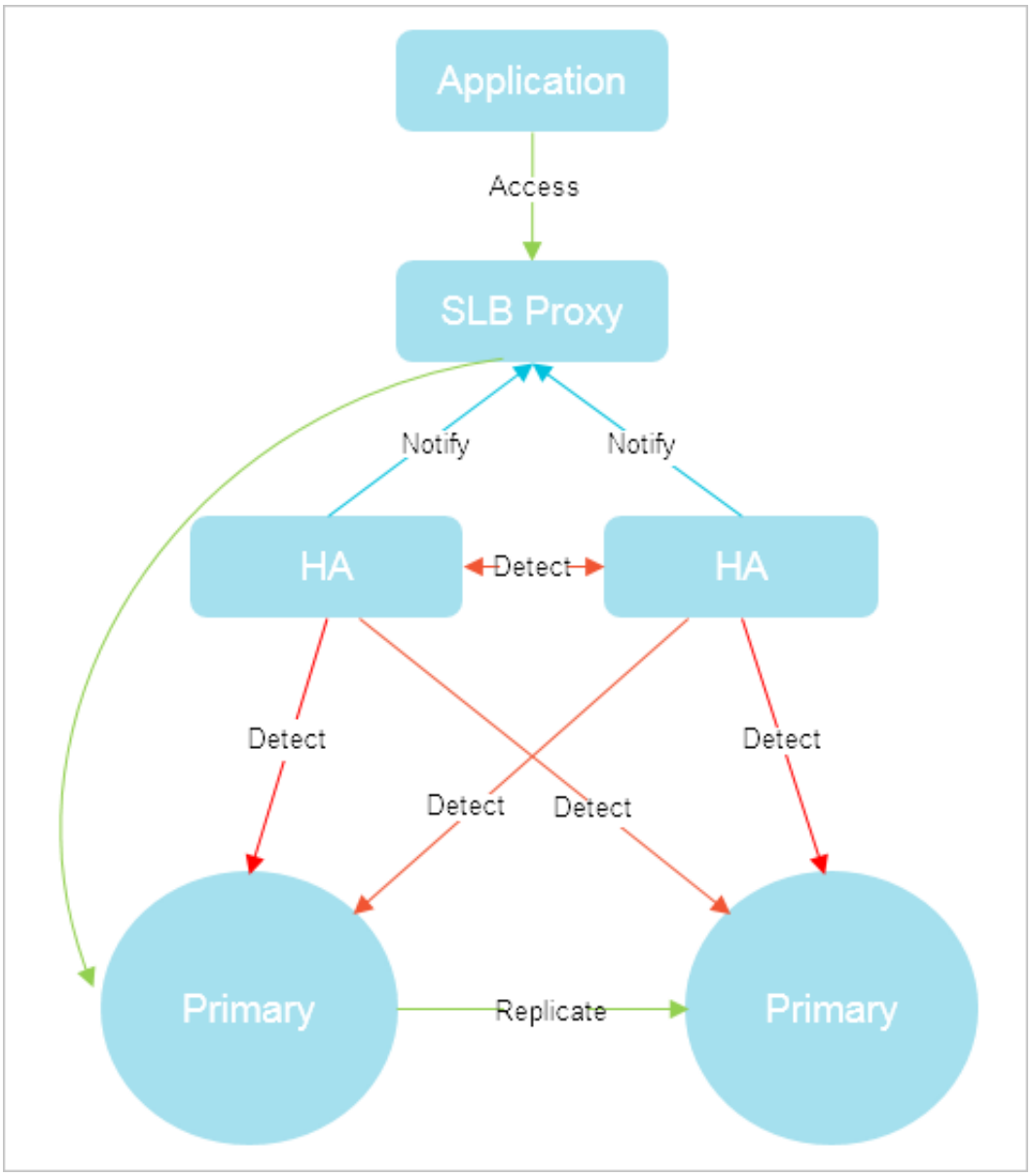
Standard protocols supported by KVStore for Redis:

Engine	Version
Redis	Compatible with V2.8 and V3.0 Geo Edition

8.2.2 High-availability service

The high-availability service ensures the availability of the data link service and processes internal database exceptions.

The high-availability service is provided by multiple HA nodes, ensuring the high availability of the service itself.



8.2.2.1 Detection

The Detection module detects whether the master node and slave node of the DB Engine are providing services normally.

The HA node uses heartbeat information, acquired at an interval of 8 to 10 seconds, to determine the health status of the master node. This information, combined with the health status of the slave node and heartbeat information from other HA nodes, allows the Detection module to eliminate any risk of misjudgment caused by exceptions such as network jitter. As a result, switchover can be completed within 30 seconds.

8.2.2.2 Repair

The Repair module maintains the replication between the master and slave nodes of the DB Engine. It also repairs any errors that may occur in either node during daily operations, for example:

- It can automatically restore master/slave replication in case of disconnection.
- It can automatically repair table-level damage to the master or slave node.
- It can save and automatically repair crashes of the master or slave node.

8.2.2.3 Notice

The Notice module informs the Server Load Balancer or Proxy of status changes to the master and slave nodes to ensure that you can continue to access the correct node.

For example, the Detection module discovers that the master node encounters an exception and instructs the Repair module to fix it. If the Repair module fails to resolve the problem, it directs the Notice module to initiate traffic switching. The Notice module then forwards the switching request to the Server Load Balancer or Proxy, which begins to redirect all traffic to the slave node.

At the same time, the Repair module creates a new slave node on another physical server and synchronizes this change back to the Detection module. The Detection module starts to recheck the health status of the instance and discovers it is healthy.

8.2.3 Monitoring service

The monitoring service tracks the status of KVStore for Redis instances in terms of services, networks, operating systems, and instances.

8.2.3.1 Service-level monitoring

The independent Service module monitors KVStore for Redis instances in terms of services.

For example, the Service module monitors KVStore for Redis-dependent Alibaba Cloud services such as Server Load Balancer, including function implementation and response time.

8.2.3.2 Network-level monitoring

The Network module monitors KVStore for Redis instances in terms of networks, for example:

- Connectivity between ECS and KVStore for Redis
- Connectivity between KVStore for Redis hosts
- Packet loss rates for routers and VSwitches

8.2.3.3 OS-layer monitoring

The OS (operating system) module monitors KVStore for Redis instances in terms of hardware and kernel of the operating system, for example:

- Hardware overhaul: The OS module constantly checks the operation statuses of the CPU, memory, motherboard, and storage. It predicts the possibilities of a fault and automatically submits a repair request in advance.
- OS kernel monitoring: The OS module tracks all database calls and uses the kernel status to analyze the causes of call slowdowns or errors.

8.2.3.4 Instance-level monitoring

The Instance module collects information of KVStore for Redis instances, for example:

- Instance availability
- Instance capacity

8.2.4 Scheduling service

The scheduling service allocates resources. It integrates and allocates the underlying KVStore for Redis resources. To you, this is the same as instance activation and migration.

For example, when you create an instance using the console, the scheduling service will calculate the most suitable physical server to carry the traffic.

After lengthy instance creation, deletion, and migration operations, the scheduling service calculates the degree of resource fragmentation in a zone and initiates resource integration regularly to improve the service carrying capacity of the zone.

9 ApsaraDB for MongoDB

9.1 What is ApsaraDB for MongoDB

ApsaraDB for MongoDB is a stable, reliable, and automatically scalable database service that is fully compatible with MongoDB protocols. The service offers a full range of database solutions, such as disaster recovery, backup, restore, monitoring, and alarms.

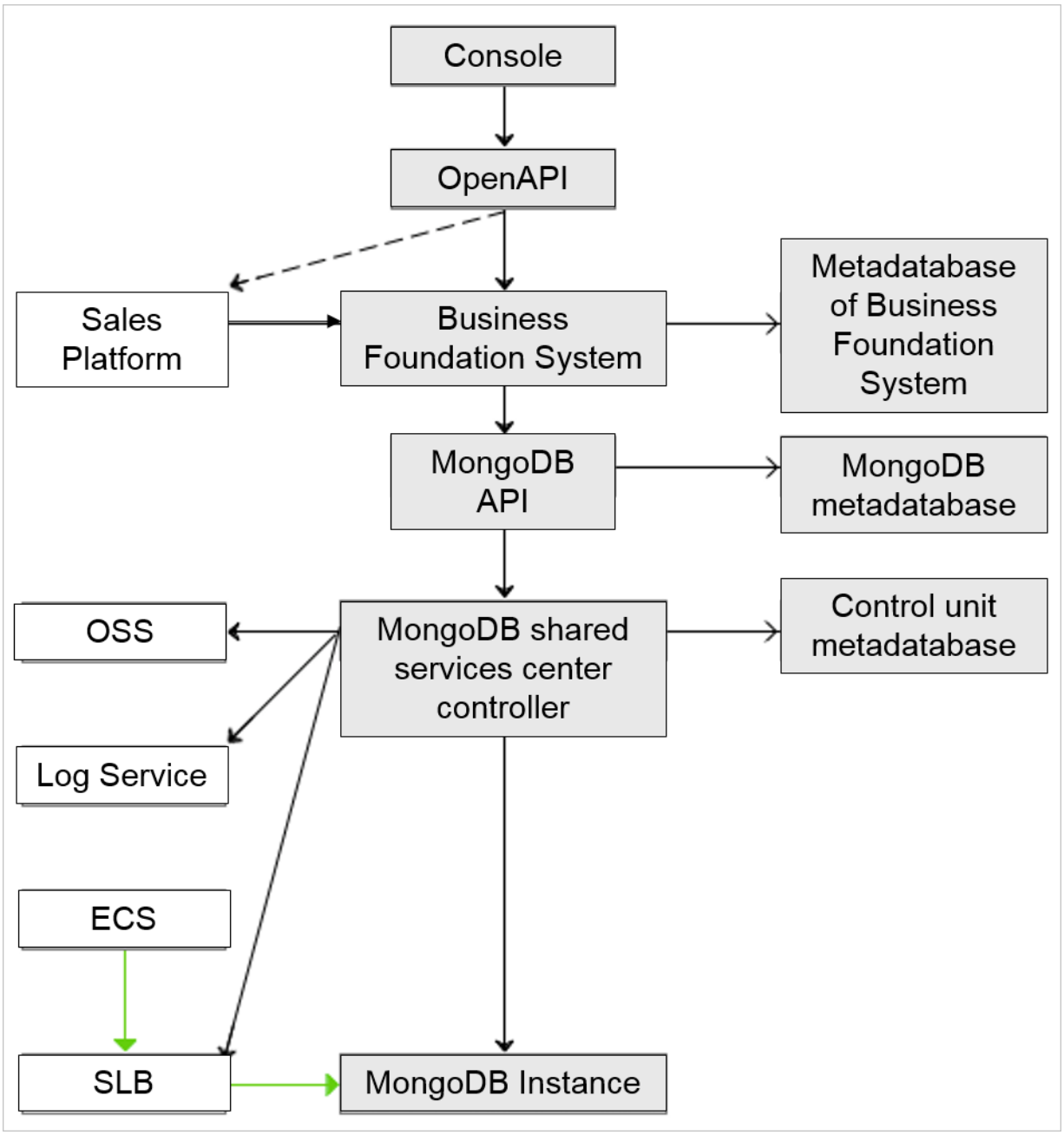
The three-node replica set architecture is deployed for ApsaraDB for MongoDB by default. The primary node supports read/write access, the secondary node provides routine read-only operations, and the other standby node is hidden and ensures high availability.

ApsaraDB for MongoDB provides solutions to ensure secure and reliable services in multiple aspects, including but not limited to the following:

- Access control, including database account and password management and IP address whitelist
- Network isolation
- Data backup
- Version maintenance
- Service authorization

9.2 System architecture

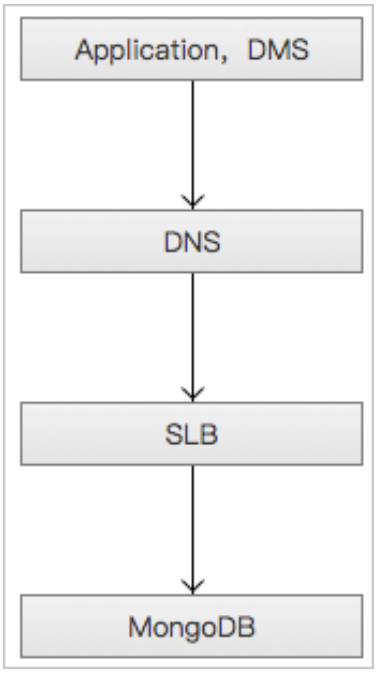
ApsaraDB for MongoDB supports six core services: data link service, scheduling service, backup service, high availability service, monitoring service, and migration service.



9.3 Functions

9.3.1 Data link service

The data link service provides support for data operations.



DNS

For example, the domain name of an ApsaraDB for MongoDB instance is mongodb.aliyun.com , and the corresponding IP address is 10.1.1.1. If either mongodb.aliyun.com or 10.1.1.1 is configured in the connection pool of a program, the instance is accessible from both the domain name and IP address.

However, the IP address may change to 10.1.1.2 if the instance is scaled up or migrated. If the domain name configured in the connection pool is mongodb.aliyun.com, the instance is still accessible. If the IP address configured in the connection pool is 10.1.1.1, the instance is no longer accessible.

SLB

The SLB module provides instance IP addresses (including both internal and public IP addresses) to prevent host changes from affecting instance performance.

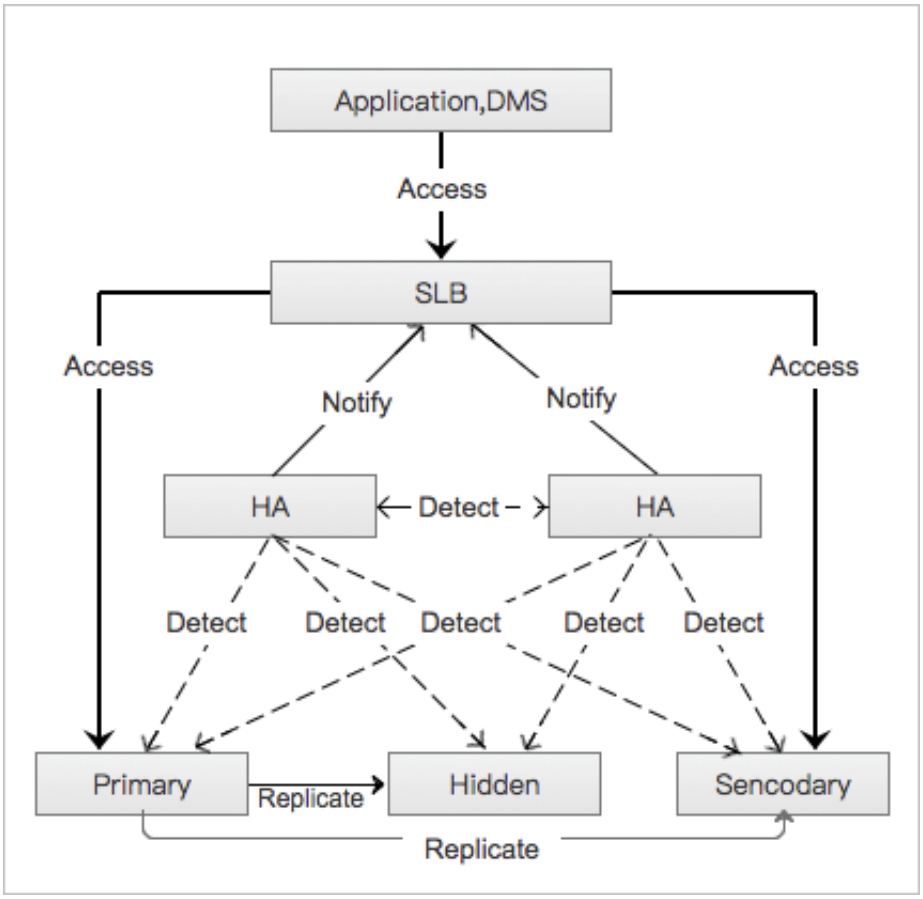
For example, the internal IP address of an ApsaraDB for MongoDB instance is 10.1.1.1, and the corresponding instance runs on 192.168.0.1. Normally, the SLB module redirects all traffic destined for 10.1.1.1 to 192.168.0.1.

When 192.168.0.1 fails, another address, 192.168.0.2 in hot standby status, takes over for 192. 168.0.1. In this case, the SLB module redirects all traffic destined for 10.1.1.1. to 192.168.0.2, and the instance continues to provide its service normally.

9.3.2 High availability service

The high availability (HA) service guarantees the availability of data link services and handles internal database exceptions.

In addition, this service is provided by multiple HA nodes which are also highly available.



Detection

The Detection module checks whether the primary, secondary, and hidden nodes of ApsaraDB for MongoDB are providing their services normally. An HA node uses heartbeat information, which is acquired at an interval of 8 to 10 seconds, to determine the health status of the primary node. This information, combined with the heartbeat information of the secondary and hidden nodes, allows the Detection module to eliminate any risk of false negatives and positives caused by exceptions such as network jitter. As a result, switchover can be completed within 30 seconds.

Repair

The Repair module maintains the replication relationship among the primary, secondary, and hidden nodes, and fixes or recreates faulty nodes.

Notice

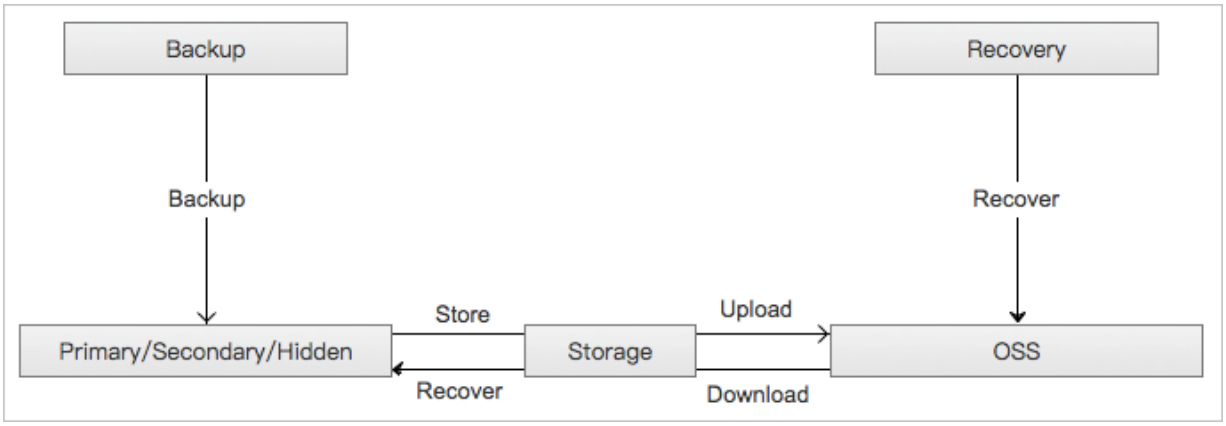
The Notice module informs the SLB of node status changes to ensure that you can continue to access the correct node.

For example, the Detection module discovers that the primary node has an exception and instructs the Notice module to switch traffic. The Notice module then forwards the switching request to the SLB, which begins to redirect traffic destined for the primary node to the secondary node and also redirect traffic destined for the secondary node to the hidden node. In this case, the secondary node becomes the primary node and the hidden node becomes the secondary node.

Simultaneously, the Repair module attempts to fix the original primary node and turn it into a new hidden node. If the fix process fails, the Repair mode creates a new hidden node on another host and synchronizes this change to the Detection module. The Detection module then incorporates this new information and starts to recheck the health status of the instance.

9.3.3 Backup service

The backup service supports offline data backup, transfer, and restore.



Backup

The Backup module backs up and compresses data and logs of an instance, and uploads the compressed files to OSS. Data backup in ApsaraDB for MongoDB is always performed on the hidden node to avoid affecting services on the primary and secondary nodes.

Recovery

The Recovery module restores backup files in OSS to a specified node.

Primary node rollback: You can roll back the settings on the primary node to a specific time point in case of data-related misoperations.

Secondary and hidden node restore: The system automatically selects a new secondary node to reduce risks when an irreparable failure occurs on the original secondary node.

Storage

The Storage module supports the upload, transfer, and download of backup files. Currently, all backup data is uploaded to OSS for storage. You can obtain temporary links to download data as needed.

9.3.4 Monitoring service

The monitoring service tracks the status of services, networks, operating systems, and instances.

Service

The Service module supports status tracking at the service level. For example, the Service module monitors SLB, OSS, and SLS services on which ApsaraDB for MongoDB depends. The module checks whether functions and response time of those services are within normal ranges. The module also uses corresponding logs to check whether the internal ApsaraDB for MongoDB services are running properly.

Network

The Network module supports status tracking at the network level. For example, the module monitors the connectivity between ECS and ApsaraDB for MongoDB and between ApsaraDB for MongoDB hosts. It also monitors packet loss rates of VRouters and VSwitches.

OS

The OS module supports status tracking at hardware and OS kernel levels.

Examples:

- Hardware inspection: The OS module regularly checks the running status of devices such as CPUs, memory modules, motherboards, and storage devices. If the module detects any potential hardware failures, it automatically submits a repair ticket.
- OS kernel monitoring: The OS module tracks all kernel invocations of databases and analyzes the cause of a slow or faulty invocations based on the kernel status.

Instance

The Instance module supports the following features:

- Collecting instance-level information for ApsaraDB for MongoDB

- Providing instance availability information
- Monitoring instance capacity and performance metrics
- Recording statement executions for instances

9.3.5 Scheduling service

The scheduling service provides resource allocation and instance version management.

Resource

The Resource module allocates and integrates underlying ApsaraDB for MongoDB resources. This module allows you to create and modify instances.

For example, when you use the ApsaraDB for MongoDB console or Open APIs to create an instance, the Resource module selects the optimal host to handle traffic. After instances are created, deleted, and migrated, the module calculates the resources of available zones such as disk space, and then periodically integrates the resources in the available zone to handle more traffic.

Version

The Version module allows you to upgrade versions of ApsaraDB for MongoDB instances. For example, you can upgrade an ApsaraDB for MongoDB instance between major versions such as from Version 3.2 to Version 3.4. You can also upgrade an instance between minor versions, fix a bug in the ApsaraDB for MongoDB source codes, or optimize the ApsaraDB for MongoDB kernel as needed.

10 KVStore for Memcache

10.1 What is KVStore for Memcache

10.1.1 Scenarios

Frequently-accessed businesses

Some examples of frequently-accessed businesses are social networks, e-commerce, games, and advertisements. Frequently-accessed data can be stored in KVStore for Memcache and the underlying data in KVStore for RDS.

Large promotion business

Large promotion or flash sales systems are usually under high access pressure. The average database usually cannot undertake such read or write stress. In such cases, KVStore for Memcache can turn out to be a viable option.

Inventory systems with counters

KVStore for RDS and KVStore for Memcache can be used in combination. KVStore for RDS stores the specific data information, while the database fields store the specific statistics. KVStore for Memcache reads the statistics, while KVStore for RDS stores the statistics.

Data analysis business

KVStore for Memcache can be used in combination with the open data processing service MaxCompute to implement distributed analysis and processing of big data. It is suitable for the big data processing scenarios such as data mining and business analysis. The data integration service can synchronize data between KVStore for Memcache and MaxCompute, simplifying the data operations.

10.1.2 Benefits

Ease of use

- Immediate availability: An instance is immediately available after it is created, facilitating fast business deployment.
- Compatibility with open-source Memcached: KVStore for Memcache is compatible with Memcached Binary Protocol. All clients that support this protocol and SASL can connect to KVStore for Memcache.

- Visualized management and monitoring panel: The console provides multiple monitoring metrics for your convenience to manage KVStore for Memcache instances.

Cluster features

KVStore for Memcache supports super large capacity and provides super high performance. The default cluster output utilizes super large cluster instances to meet demands for large capacity and high performance.

Elastic scalability

- Scale-out of storage capacity with a single click: You can adjust the storage capacity of an instance in the console based on business requirements.
- Online scale-out without service interruption: You can adjust the instance capacity online without suspending your services or affecting your business.

Resource isolation

Instance-level resource isolation provides enhanced stability for individual services.

High security and reliability

- Password authentication is supported to ensure secure and reliable access.
- Persistent data storage: The use of memory and hard disks meets data persistence demands while high-speed data reading and writing are provided.

High availability

- Each instance has a primary node and a secondary node. This prevents service interruption caused by single point of failures (SPOFs).
- Automatic detection and recovery of hardware faults: KVStore for Memcache automatically detects hardware faults and fails services over within seconds to recover services.

10.2 Functions

The 6 core services of Memcache include:

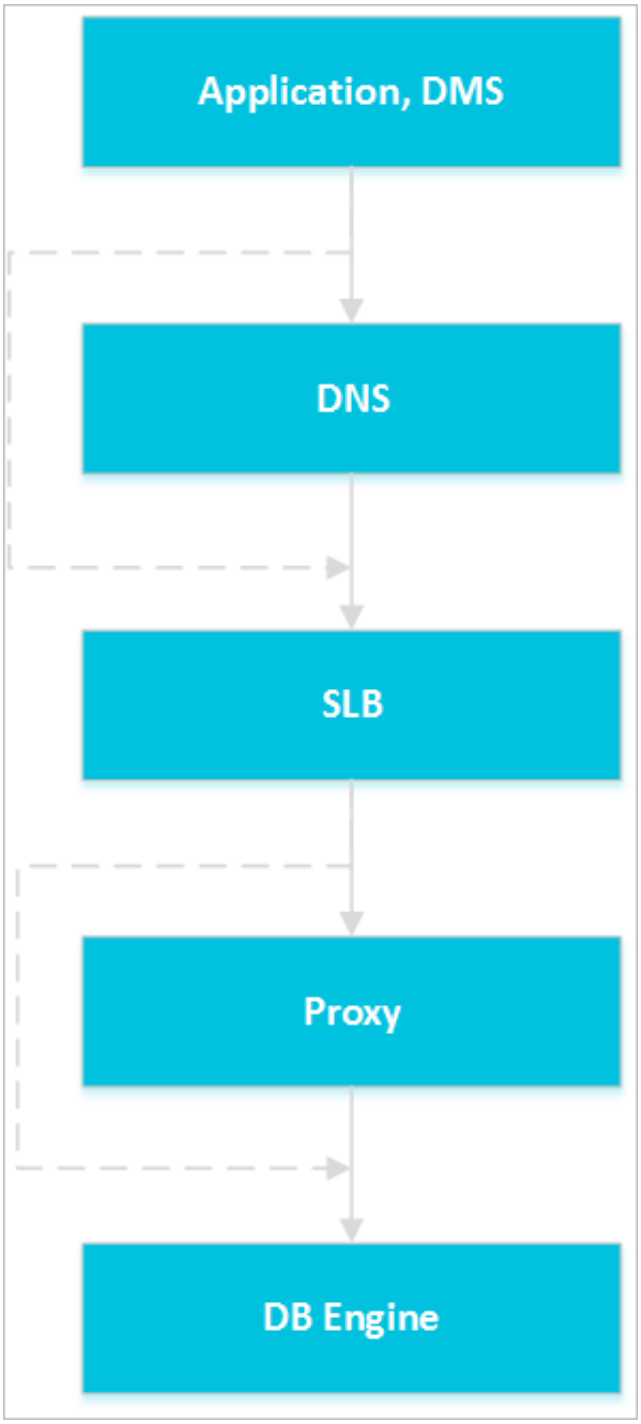
- Data link service
- Scheduling service
- Backup service
- High availability service
- Monitoring service
- Migration service

10.2.1 Data link service

Data link service offers data operations, such as adding, deleting, modifying and querying data.

You may connect to Memcache using applications, or you can use a data management tool (DMS) provided by Memcache for gui-based data management.

Figure 10-1: Data link service



10.2.1.1 DNS

The DNS module supports the dynamic resolution of domain names to IP addresses. It prevents IP address changes from affecting the performance of Memcache instances.

For example, consider a Memcache instance with an intranet IP of 10.1.1.1, and a correspond ing Proxy or DB Engine running on 192.168.0.1: Normally, the SLB module redirects all traffic destined for 10.1.1.1 to 192.168.0.1.

If 192.168.0.1 fails, another hot standby address, 192.168.0.2, takes over for 192.168.0.1. The SLB module redirects all traffic destined for 10.1.1.1 to 192.168.0.2, and the Memcache instance continues to offer its services normally.

10.2.1.2 SLB

The SLB module provides instance IP addresses to prevent physical server changes from affecting the performance of Memcache instances.

For example, consider an KVStore for Memcache instance with an intranet IP of **`10.1.1.1`**, and a corresponding Proxy or DB Engine running on **`192.168.0.1`**: Normally, the SLB module redirects all traffic destined for **`10.1.1.1`** to **`192.168.0.1`**.

If **`192.168.0.1`** fails, another hot standby address, **`192.168.0.2`**, takes over for **`192.168.0.1`**. The SLB module redirects all traffic destined for **`10.1.1.1`** to **`192.168.0.2`**, and the Memcache instance continues to offer its services normally.

10.2.1.3 Proxy

The Proxy module provides data routing, traffic detection, and session persistence. However, its functions continue to expand.

- Data routing: Memcache data routing supports cluster architectures and allows complex query and partitioning strategies for distributed routing.
- Traffic detection: this reduces the risk of network attacks directed against Memcache.
- Session persistence: this prevents database connection interruptions if any failures occur.

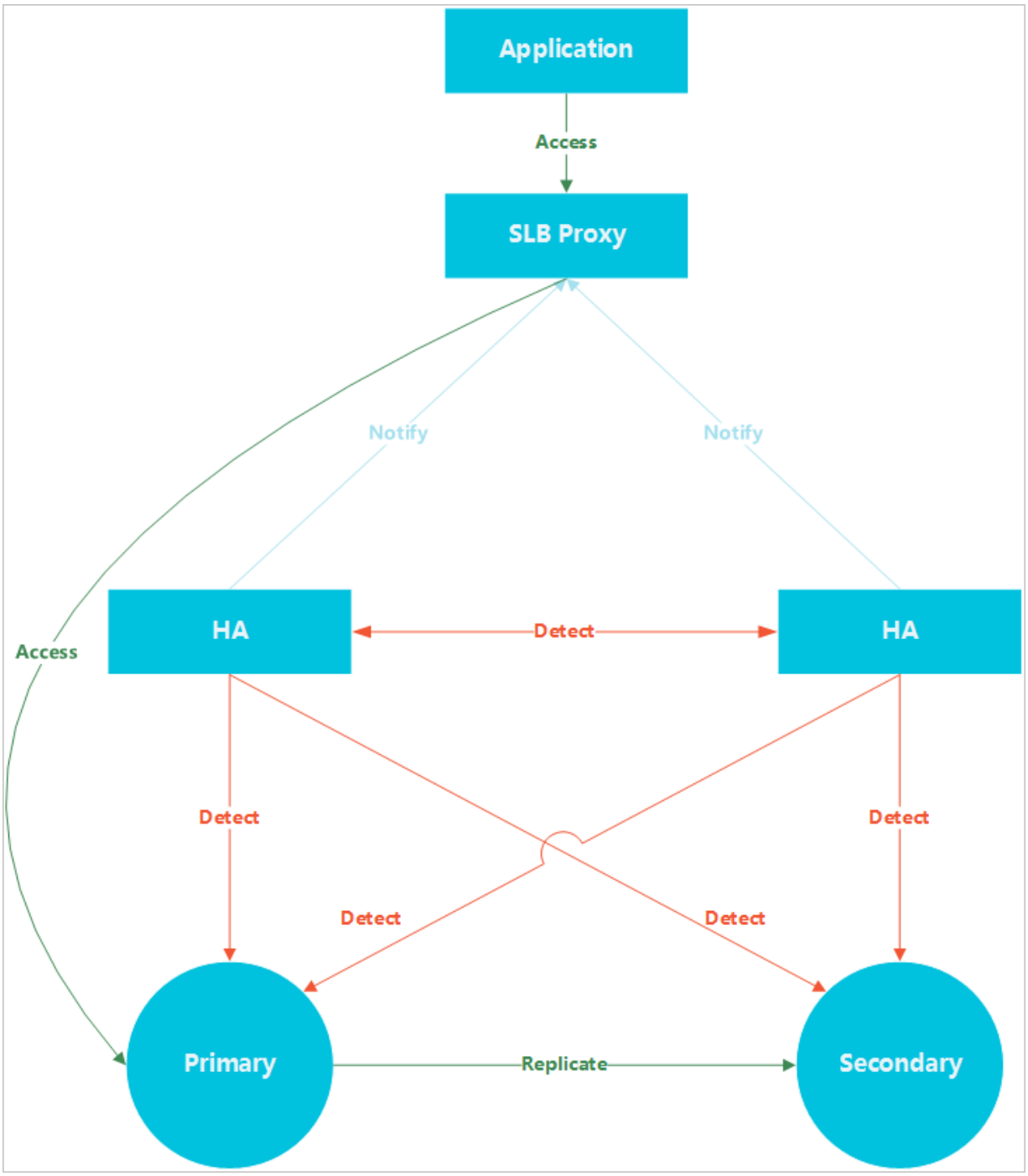
10.2.1.4 DB Engine

KVStore for Memcache supports mainstream protocols and direct connections with a variety of clients.

10.2.2 High availability service

The high-availability service guarantees the availability of the data link services and processes any internal database exceptions. In addition, the high availability service is provided by multiple HA nodes which are highly available.

Figure 10-2: High availability service



10.2.2.1 Detection

The Detection module checks whether the master and slave nodes of the DB Engine offer their services normally.

The HA (High Available) node uses heartbeat information, acquired at an interval of 8 to 10 seconds, to check the health status of the master node. This information is combined with the health status of the standby node and heartbeat information from other HA nodes. It allows the Detection module to eliminate any risk of misjudgment caused by exceptions, such as network jitter, and allows the exception switchover to be completed within 30 seconds.

10.2.2.2 Repair

The Repair module maintains the replication relationship between the master and slave nodes of the DB Engine. It can also repair any errors that may occur on either node. For example:

- Automatic restoration of master/slave replication in case of disconnection
- Automatic repair of table-level damage to a master or slave node
- On-site saving and automatic repair if a master or slave node crash

10.2.2.3 Notice

The Notice module informs the SLB or Proxy of status changes to the master and slave nodes to guarantee that you can continue to access the correct node.

For example, the Detection module discovers that the master node has an exception and instructs the Repair module to fix it. If the Repair module fails to resolve the problem, it directs the Notification module to initiate traffic switching. The Notification module then forwards the switching request to the SLB or Proxy, which begins to redirect all traffic to the slave node.

Simultaneously, the Repair module creates a new slave node on another physical server and synchronizes this change back to the Detection module. The Detection module then incorporates this new information and starts to recheck the health status of the instance.

10.2.3 Monitoring service

KVStore for Memcache provides multilevel monitoring services across the service, network, operating systems, and instance layers to ensure status tracking.

10.2.3.1 Service-level monitoring

The Service module tracks the service-level status.

For example, it monitors whether other cloud products, such as SLB, on which Memcache depends, are normal. This includes their functionality and response time.

10.2.3.2 Network-level monitoring

The Network module tracks the network level status.

For example, the connectivity between ECS and KVStore for Memcache; the connectivity between KVStore for Memcache and physical machines; and the packet loss rate of VRouters and VSwitches.

10.2.3.3 OS-level monitoring

The OS (operating system) module tracks status at the hardware and OS kernel layer, for example:

- Hardware maintenance: The OS module constantly checks the operational status of the CPU , memory, main board, and storage; evaluates whether a fault will occur; and automatically submits a repair report in advance.
- OS kernel monitoring: The OS module tracks all database calls and uses the kernel status to analyze the reasons for slowdowns or call errors.

10.2.3.4 Instance-level monitoring

The Instance module collects KVStore for Memcache instance-level information, for example, available information for instances, instance capacity, and performance indicators.

10.2.4 Scheduling service

The scheduling service implements resource allocation, such as allocating and integrating the underlying Memcache resources. For you, this includes creating and migrating instances.

For example, when you create an instance through the console, the scheduling module determines which physical server is best suited to carry the traffic.

After lengthy instance creation, deletion, and migration operations, the scheduling service calculates the degree of resource fragmentation in a zone, and initiates resource integration regularly to improve the service carrying capacity of the zone.

11 Data Management Service (DMS)

11.1 What is DMS?

Data Management Service (DMS) provides unified management of relational databases and OLAP databases. It is built on Alibaba's iDB database service platform and has been providing database development support for tens of thousands of R&D engineers since it was brought online eight years ago. Enterprises can use DMS to build their own database DevOps, which improves database R&D efficiency through better self-service and ensures employee database access security and high database performance.

DMS is used to manage relational databases such as MySQL, SQL Server, and PostgreSQL, as well as OLAP databases. It integrates data management with structure management.

11.2 Architecture

Alibaba Cloud Data Management Service (DMS) consists of the service layer, scheduling layer, and connection layer. It processes real-time data access and schedules data-related background tasks for relational databases.

Business layer

- The DMS business layer provides online GUI-based database operations. The business layer can be extended linearly to improve the general service capabilities of DMS.
- DMS supports stateless failovers to ensure a 24/7 availability.

Scheduling layer

- The scheduling layer allows you to import and export tables and compare table structures. This layer schedules tasks by using the thread pool in two modes: real-time scheduling and background periodic scheduling.
- Real-time scheduling allows you to quickly schedule and execute a task in the frontend. After you submit a task, you do not need to wait for the execution result. The DMS backend automatically executes the task. After the task has been executed, you can download or view the execution result.
- Background periodic scheduling allows you to periodically obtain specified data, such as data trends. DMS collects business data in the background based on scheduled tasks, allowing you to query and analyze the collected data.

Connection layer

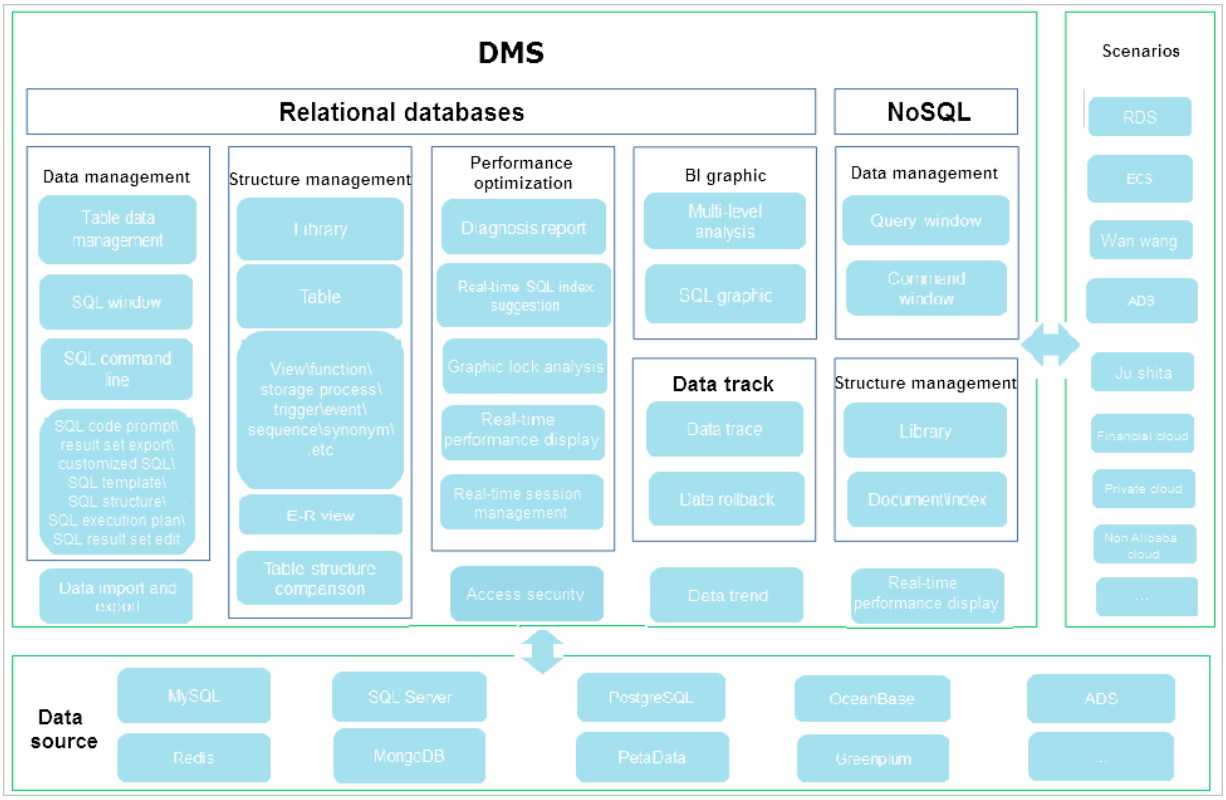
The connection layer is the core component for data access in DMS. It has the following features:

- It can process requests from MySQL, SQL Server, and PostgreSQL databases.
- It isolates sessions and provides session persistence. You can open multiple SQL windows using DMS, and the SQL window sessions are isolated from each other. In addition, the session in each SQL window is persistent to simulate the client experience.
- It controls the number of instances to avoid establishing a large number of connections to a single instance.
- Different functions have different connection release policies. This improves user experience and reduces the number of connections to the databases.

11.3 Function module

The following figure shows the functions of DMS.

Figure 11-1: Function module



- **Functions of relational databases**

  - Data management: provides the ability to manage SQL windows, SQL command lines, table data, SQL prompt, SQL formatting, custom SQL statements, SQL templates, SQL execution plans, and import and export operations.
  - Structure management: provides table structure comparison capabilities and management for objects such as libraries, tables, views, functions, stored procedures, triggers, events, sequences, and synonyms.
  - Performance optimization: provides features such as real-time performance visualization , real-time SQL index recommendations, graphical lock management, instance session management, and diagnostic reports.
  - Access security: provides a four-layer authentication system, logon and action trail, and fine -grained permissions such as cloud account permission, access permission, and feature enabling.
- **Functions of NoSQL databases**
  - Data management: provides query window and command window functions.
  - Structure management: provides management capabilities for objects such as databases, files, and indexes.
  - Real-time performance visualization: provides real-time visualization of key performance metrics.

11.4 Benefits

Improved R&D efficiency

- Table structure comparison.
- Intelligent SQL prompt.
- Reuse of custom SQL statements and SQL templates.
- Automatic recovery for working environments.
- Export of dictionary files.

Real-time optimization of database performance

- Useful session management.
- Monitoring of core metrics in seconds.
- Graphical lock management.
- Real-time SQL index recommendations.
- Diagnostic reports for overall performance.

Comprehensive access security protection

- 4-layer authentication system.
- Fine-grained authorization.
- Logon and action trail.

Extensive options for data sources

- Relational databases such as MySQL, SQL Server, PostgreSQL, and PPAS.
- NoSQL databases such as Redis, MongoDB, and Memcache.

11.5 Product value

DMS provides you with a convenient and secure database access and management platform . Visualized data services enable you to use databases on browsers, eliminating the need to install various database clients. When you edit data online, you can easily perform operations on table data and change table structures, without having to write complex SQL statements. DMS provides users with advanced functions that common clients do not offer, such as table structure synchronization, database cloning, result visualization, and real-time monitoring.

To use DMS, you first need to log on to the Apsara Stack Management Console, then use your database account and password to log on to the DMS console. This can prevent your database account and password from being stolen. DMS supports HTTPS and SSL for data transmission, and prevents data from being intercepted or tampered with during transmission.

DMS also supports RAM and STS for permission verification to prevent unauthorized actions.

DMS supports access to VPC instances. It provides you with an interface for data access and ensures the network security of database instances. Common clients do not provide this feature.

You can benefit from the following functional advantages that are bundled with DMS:

- Convenient data operations
  - Pain point: You need a convenient and full-featured product to complete SQL operations, save common operations, and apply common operations to specific services.
  - Solution: You can open a table in DMS and perform operations on table data as you would in an Excel worksheet. You can add, delete, change, query, and make statistical analysis of table data without understanding SQL. You can customize SQL operations and save common business-related SQL operations in DMS. Then you can apply these operations directly when managing other databases or instances.

- Visualization of database table structures
  - Pain point: When you design a new business table or operate an existing business table , you often need to understand the structures of all the tables in a database. You can execute SQL commands one by one to display the table structure, but this method is neither intuitive nor convenient.
  - Solution: Through the document generation function of DMS, you can generate the table structures of an entire database with a single click. Then you can browse these structures online or export them to other formats such as Word, Excel, and PDF.
- Real-time optimization of database performance
  - Pain point: Detailed monitoring logs over a long period of time are required for database performance optimization. You need to make a detailed analysis of the logs and locate exceptions to better improve the database performance.
  - Solution: DMS provides second-level monitoring of database performance metrics, such as SELECT, INSERT, UPDATE, and DELETE operations, the number of active connections, and network traffic volume, and helps keep you aware of any performance variations. DMS enables you to view and kill a database session.
- Creating charts of SQL result sets
  - Pain point: In the past, users used to use SQL statements to find data, and import the data into Excel to create static charts such as line charts and pie charts. This process takes a lot of time.
  - Solution: With DMS, you can directly create charts from SQL result sets. You can also create many advanced charts, such as dynamic charts, chain indexes, and personalized tooltips. This helps you to produce high quality work.
- SQL reuse
  - Pain point: When you access a database, there is always a need to run SQL statements. Simple queries are easy to master, while complex analytical queries or queries with certain business logics are not. The cost of rewriting SQL statements each time is too high, and even if the statements are saved to text files, they require constant maintenance and cannot be used flexibly.
  - Solution: You can save commonly used SQL statements to DMS through the "My SQL" function of DMS. The statements are not limited by local storage, and can be used in a wide variety of situations, such as the current database or instance, or all instances.
- Monitoring of changes to the table data volume

Big data is the latest trend in data analysis, and everyone is talking about it. However, taking full advantage of the values provided by big data analysis is not an easy task. The core idea of DMS is to start analyzing data when data becomes available.

DMS performs operations through a custom RDS kernel, which allows it to quickly collect row count changes of each instance, database and table. DMS provides real-time monitoring, historical trend, and detailed data through professional data analysis and interaction.

- Table structure synchronization
  - Pain point: Within enterprises, database environments are categorized into production environment and test environment. After a database is verified in the test environment, it is released in the production environment. If some table structures in the test environment are not synchronized to the production environment, major faults can occur during the release.
  - Solution: You can use the structural comparison function of DMS to detect inconsistencies in database table structures between the production and test environments. You can also obtain a DDL statement for table structure correction to ensure table structure consistency between the production and test environments.

12 Server Load Balancer (SLB)

12.1 What is Server Load Balancer?

Server Load Balancer (SLB) is a traffic distribution control service that distributes the incoming traffic among multiple ECS instances according to the configured forwarding rules. SLB expands the service capabilities of the application and enhances application availability.

By setting a virtual service address, SLB virtualizes the added ECS instances into an application service pool with high-performance and high availability, and distributes client requests to ECS instances in the server pool based on forwarding rules.

SLB also checks the health status of added backend servers, and automatically isolates abnormal ECS instances to eliminate single point of failure (SPOF), thus improving the overall service capability of your application. Additionally, working with Alibaba Anti-DDoS, SLB is able to defend DDoS attacks.

Components

Server Load Balancer consists of the following components:

- SLB instances

  An SLB instance is a running load balancing service that distributes incoming traffic to backend servers. To use the load balancing service, you must create an SLB instance, and then add at least one listener and two backend servers to the instance.

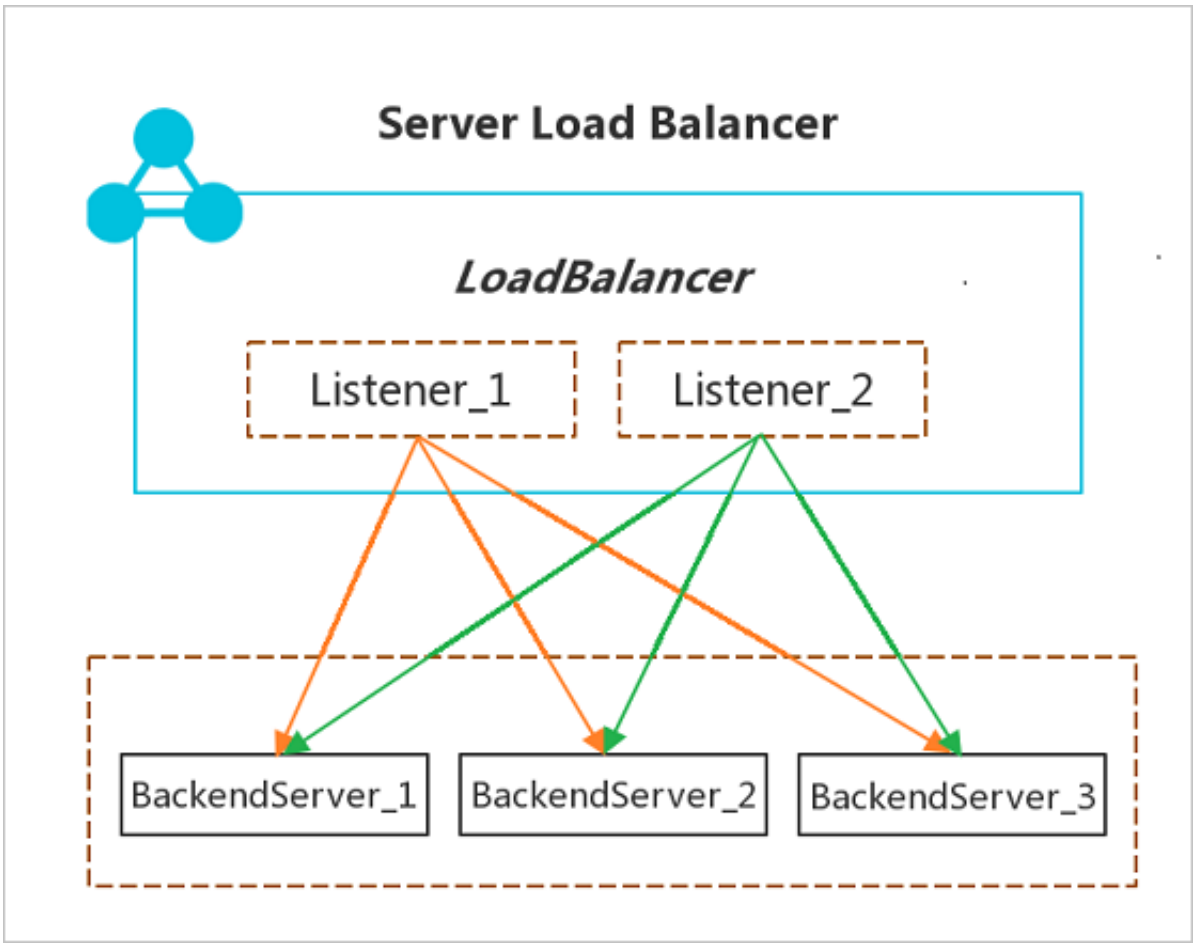
- Listeners

  A listener checks client requests and forwards the requests to backend servers according to the configured rules. It also performs health check on backend servers.

- Backend servers

  Backend servers are the ECS instances added to an SLB instance to receive and process distributed requests. You can classify ECS instances running different applications or playing different roles by creating server groups.

  As shown in the following figure, once a request arrives at an SLB instance, Server Load Balancer will distribute the request to the corresponding backend server according to the listener configurations.



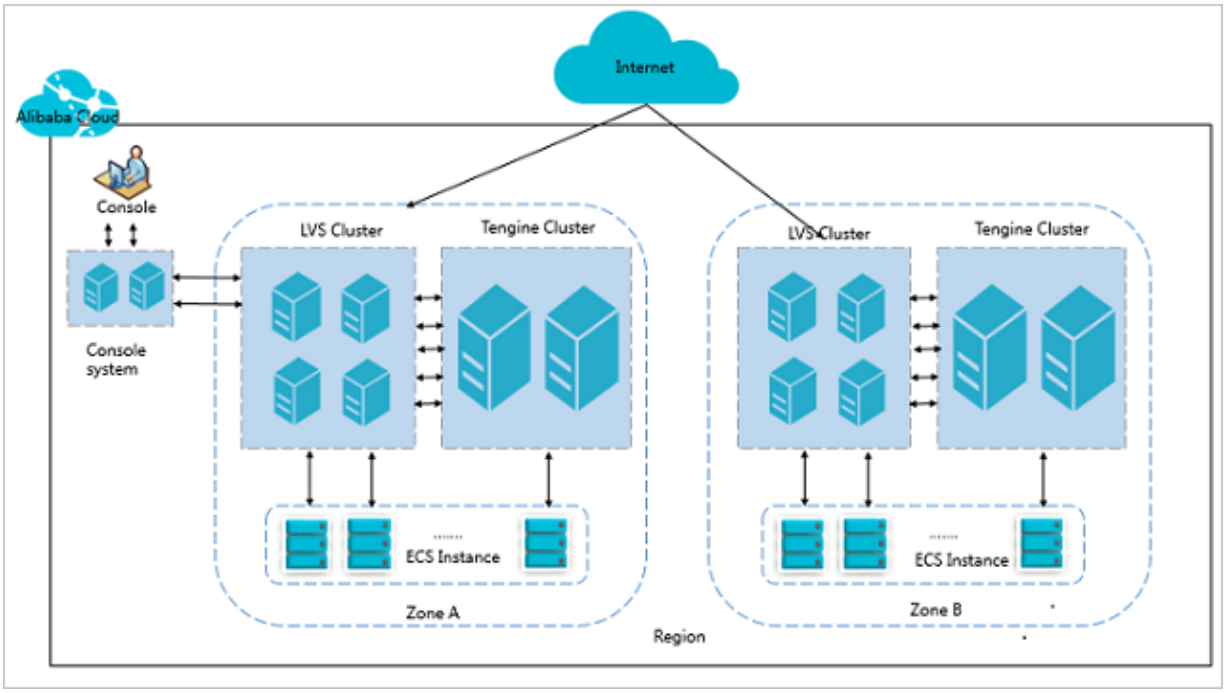
12.2 Architecture

Server Load Balancer is deployed in clusters.

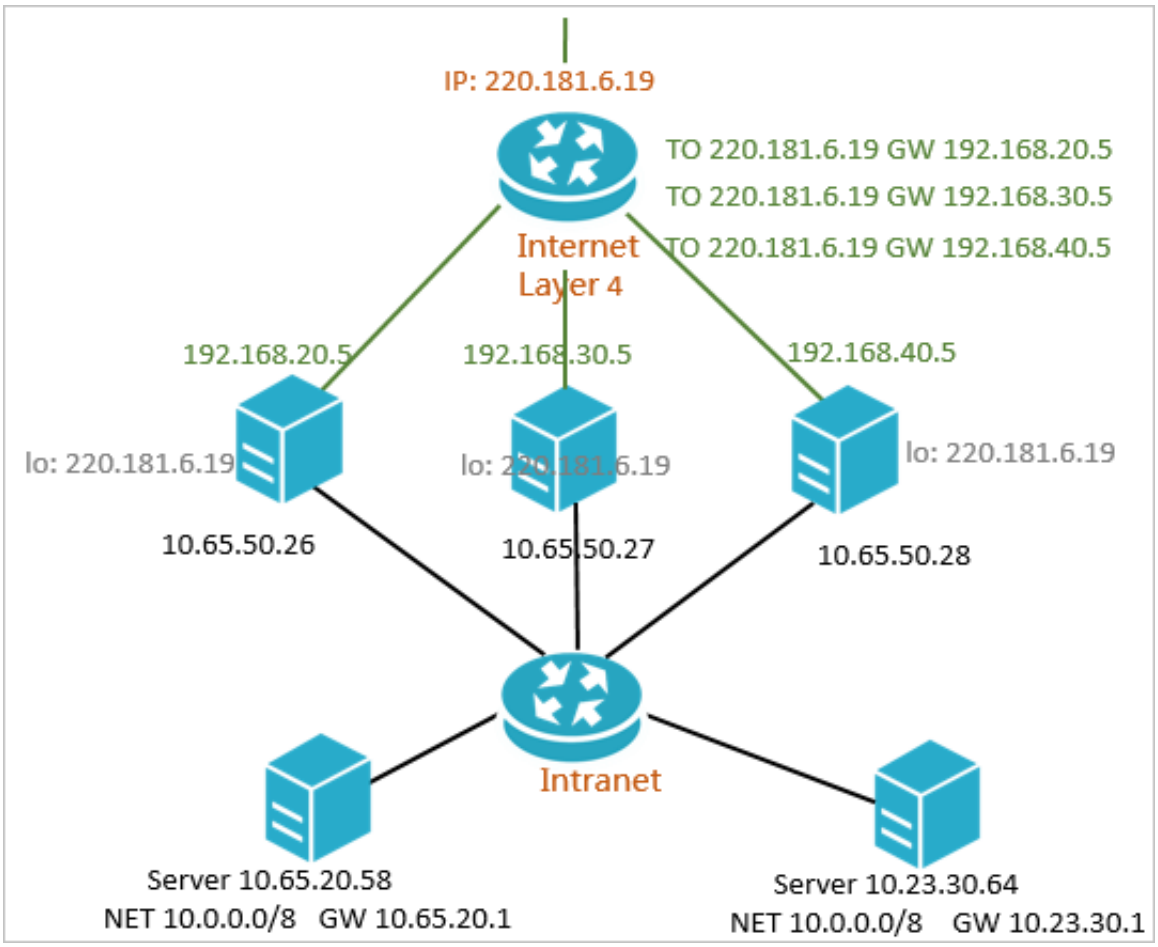
Apsara Stack provides the layer-4 (TCP protocol and UDP protocol) and layer-7 (HTTP protocol and HTTPS protocol) load balancing services. Deployed in clusters, Server Load Balancer can synchronize sessions to protect the ECS instances from single points of failure (SPOFs). This improves redundancy and guarantees the service stability.

- Layer-4 uses the open source software Linux Virtual Server (LVS) with keepalived to achieve load balancing, and also makes some customization to it according to the cloud computing requirements.
- Layer-7 uses Tengine to achieve load balancing. Tengine is a Web server project based on Nginx that adds a wide range of advanced features dedicated for high-traffic websites.

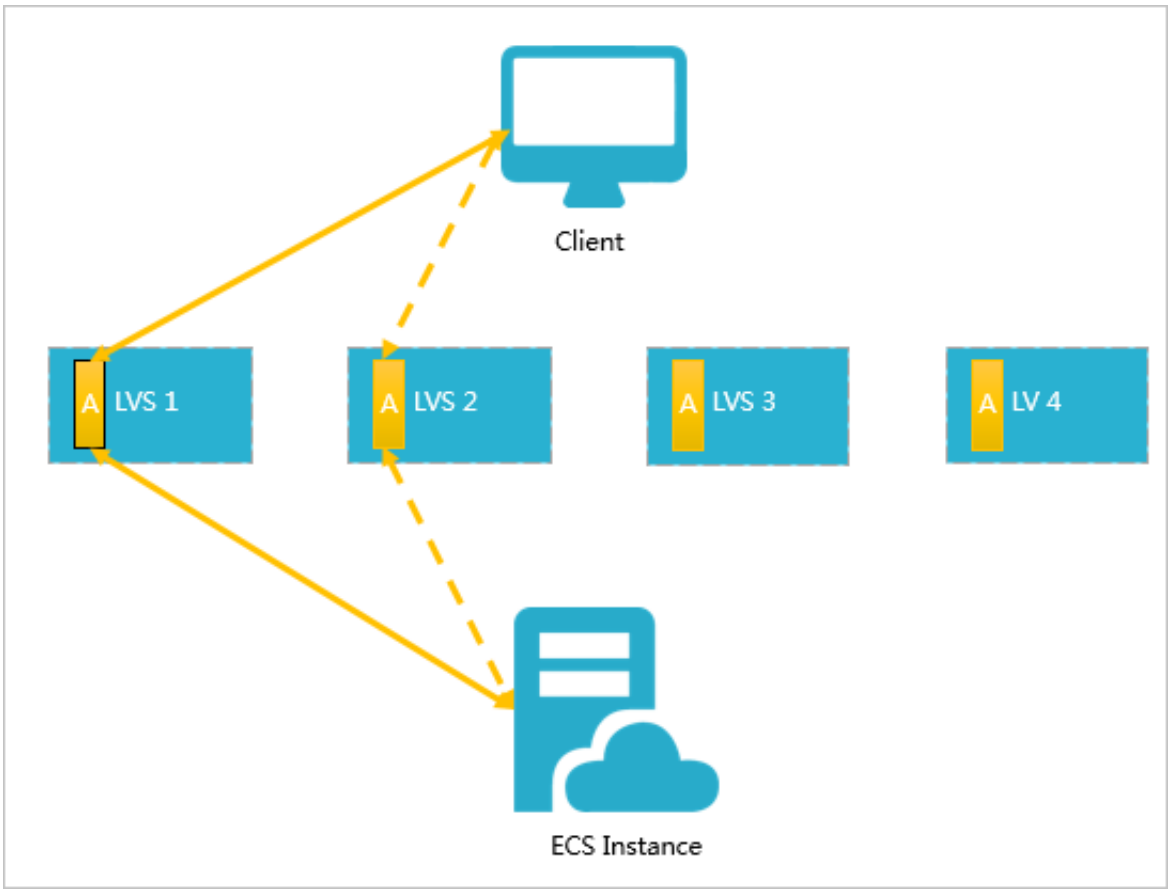
Figure 12-1: SLB architecture



As shown in the following figure, the layer-4 load balancing in each region is actually run in a cluster of multiple LVS machines. The cluster deployment model strengthens the availability, stability, and scalability of the load balancing services in abnormal circumstances.



Additionally, the LVS machine in the LVS cluster uses multicast packets to synchronize sessions to other LVS machines. As shown in the following figure, session A established on LVS1 is synchronized to other LVS machines after three packets are transferred. In normal situations, the session request is sent to LVS1 as the solid line shows. If LVS1 is abnormal or being maintained, the session request will be sent to other machines working normally, as the dotted line shows. In this way, you can perform hot upgrades, machine failure maintenance, and cluster maintenance without affecting business applications.



12.3 LVS in Layer-4 Server Load Balancer

Problems in standard LVS

LVS is the most popular open-source Layer-4 load balancing software on the world, founded by Dr. Zhang Wensong in May 1998. It achieves load balancing on the Linux platform. LVS is a kernel module implemented on the basis of netfliter framework of Linux (same as iptables), which is known as *IPVS (IP Virtual Server)*. It hooks into netfilter at the NF_IP_LOCAL_IN and NF_IP_FORWARD points.

In a large-scale cloud computing network, standard LVS has the following drawbacks:

- Drawback 1: LVS supports three packet forwarding methods: NAT, DR, and TUNNEL. When these forwarding modes are deployed in a network with multiple VLANs, the network topology becomes complex and poses high O&M costs.
- Drawback 2: LVS lacks the DDoS defense compared with commercial load balancing equipment, for example, F5.

- Drawback 3: LVS uses PC servers and the Virtual Router Redundancy Protocol (VRRP) of Keepalived to do the master-slave deployment. Therefore, its performance cannot be extended .
- Drawback 4: The configurations and health check performance of the keepalived software (widely used in LVS) are insufficient.

LVS customized features

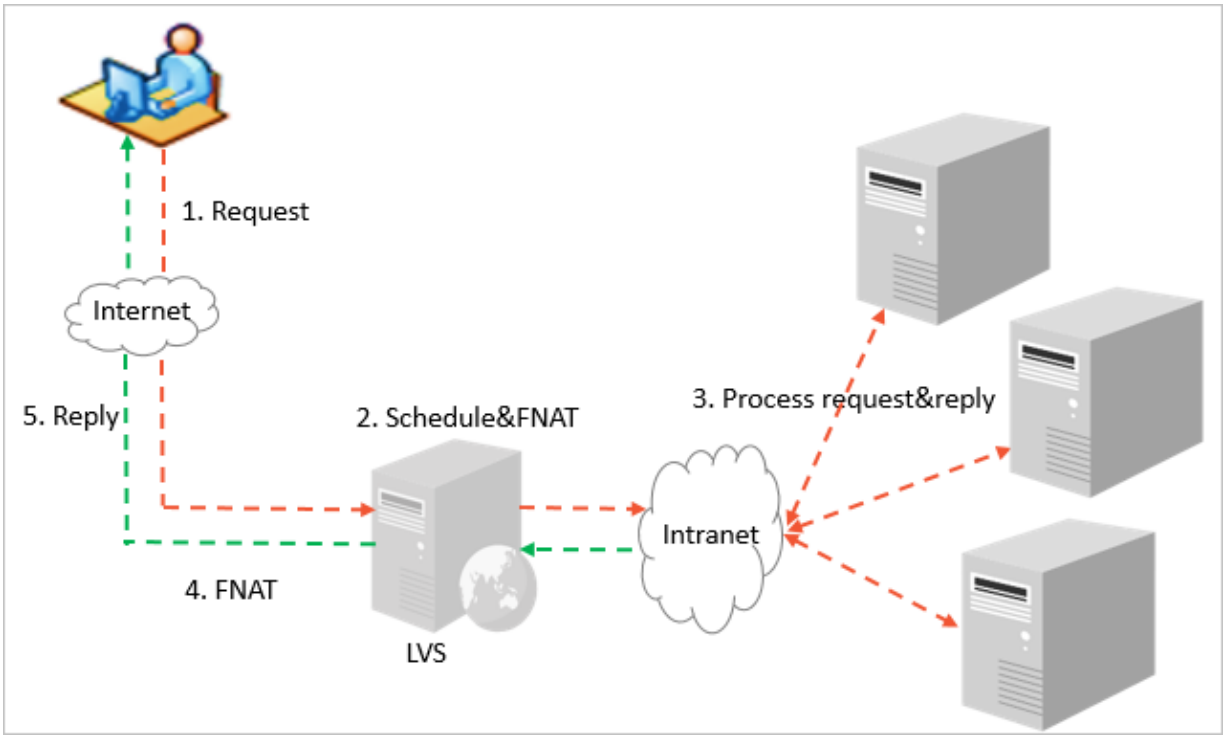
To to solve these problems, Alibaba Cloud added to following customized features to LVS. The URL for Alibaba Cloud LVS is *https://github.com/alibaba/LVS*.

- Customization 1: A new packet forwarding method, FULLNAT, so that LVS load balancer and real servers can be in different vlans.
- Customization 2: Defense modules such as SYNPROXY against synflooding attack.
- Customization 3: Support for LVS cluster deployment.
- Customization 4: Optimization of keepalived performance.

FULLNAT technology

- The main principle is as follows: The module introduces local address (internal IP address), IPVS translates cip-vip to lip-rip, in which lip and rip both are internal IP addresses. This means that the load balancers and real servers can communicate across vlans.
- All inbound and outbound data flows are transferred through LVS. 10 GB NIC is used to ensure the network bandwidth.
- Only TCP protocol is supported by the FULLNAT method.

Figure 12-2: FULLNAT forwarding



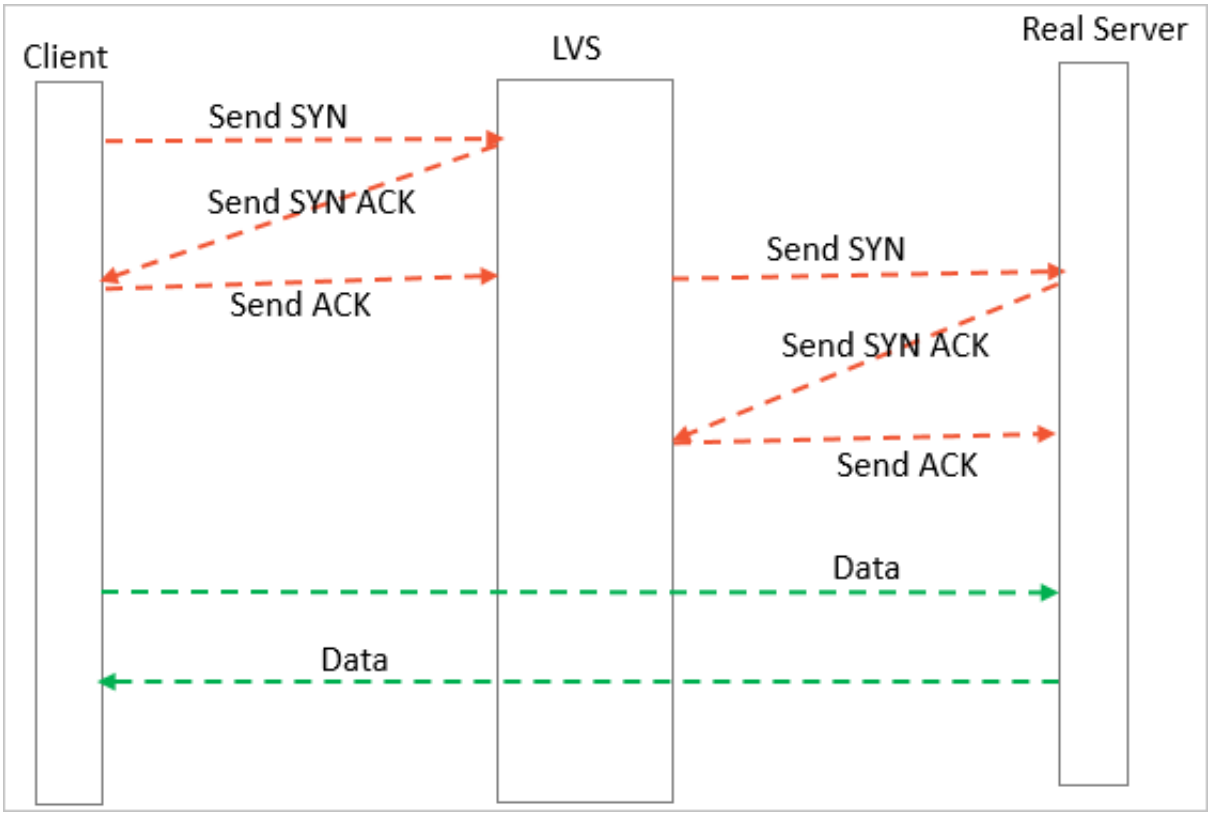
SYNPROXY technology

The main principle is as follows: Based on TPC syncookies, LVS uses a proxy to initiate a TCP three-way handshake.

The proxy process is as follows:

1. A client sends an SYN packet to LVS.
2. LVS constructs an SYN+ACK packet with a special sequence number and sends this packet to the client.
3. The client sends back an ACK response to LVS. LVS checks whether the ack_seq value in the ACK response is valid. If so, LVS establishes a three-way handshake with the real server

Figure 12-3: LVS proxy of three-way handshake



To defend against ACK, FIN, and RST flood attacks, LVS checks the connection table and discards any requests for connections which are undefined in the table.

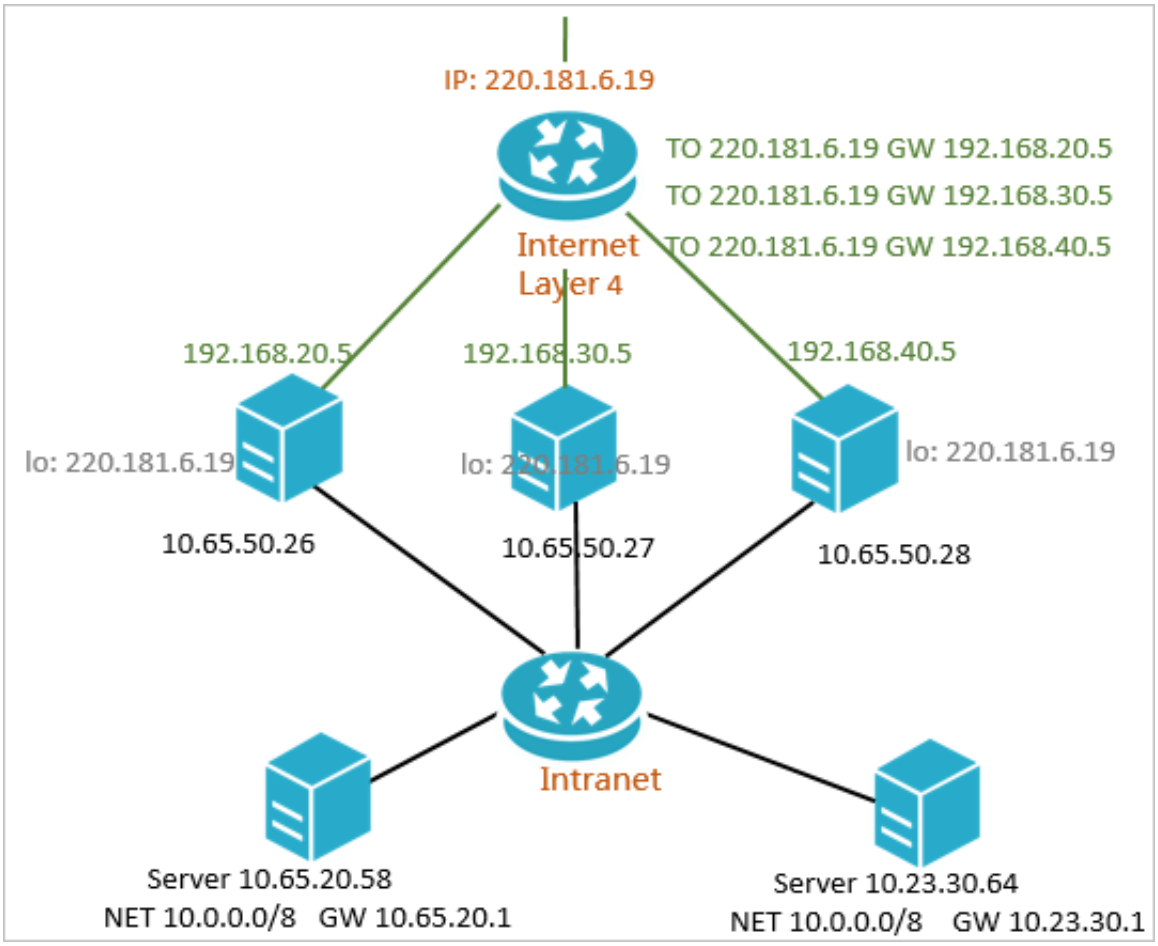
Cluster deployment

The main principle is as follows: An LVS cluster communicates with the uplink switches over the OSPF protocol. The uplink switches use equal-cost multi-path (ECMP) routing to route traffic to the LVS cluster. Then, the LVS cluster forwards the traffic to the servers.

The cluster deployment ensures the stability of Layer-4 Server Load Balancer by supporting the following characteristics:

- Robustness: The LVS and the uplink switches use OSPF as the heartbeat protocol. A VIP is added to all LVS nodes in the cluster. The switches can discover the failure of any LVS node and remove it from the ECMP routing list.
- Fexibility: If the traffic from a VIP exceeds the capacity that the current LVS cluster supports, you can scale up the cluster horizontally.

Figure 12-4: Cluster deployment



keepalived optimization

Improvements made by Alibaba Cloud to the Keepalived software include:

- Changing the asynchronous network model from select to epoll.
- Optimizing the reload process.

Benefits of Layer-4 Server Load Balancer

As described in the preceeding sections, Layer-4 Server Load Balancer has following characteristics:

- High availability: The LVS cluster ensures redundancy and prevents SPOF.
- Security: Together with Alibaba Cloud Security, LVS's intrinsic defenses provide near real-time defensive capabilities.
- Health check: LVS performs health checks on ECS instances and automatically blocks abnormal ones. Once the faulty ECS instance recovers, LVS unblocks it automatically.

12.4 Tengine in Layer-7 Server Load Balancer

Tengine is a Web server project initiated by Alibaba. Based on Nginx, Tengine adds a wide range of advanced features dedicated for high-traffic websites. Nginx is one of the most popular open-source Layer-7 load balancing software.

The URL for Alibaba Cloud Tengine is *http://tengine.taobao.org/*.

Customized features

For cloud computing scenarios, Tengine customizes the following features:

- Inherits all features of Nginx 1.4.6 and is fully compatible with Nginx configurations.
- Supports the dynamic shared object (DSO) module. This means you do not need to recompile the wholeTengine to add a module.
- Provides enhanced load balancing capabilities, including a consistent hash module and session persistence module. In addition, it can actively perform health checks on backend servers and automatically enable or disable the servers based on their status.
- Monitors system loads and resource usage to protect the system.
- Provides an enhanced attack protection (access speed limiting) module.
- Provides user-friendly error messages to help find the abnormal servers.

Using Tengine as its basic load balancing module, Layer-7 Server Load Balancer has the following features:

Benefits of Layer-7 Server Load Balancer combined with Tengine

Using Tengine as the basic module, Layer-7 Server Load Balancer has the following characteristics:

- High availability: The Tengine cluster ensures redundancy and prevents SPOF.
- Security: Tengine provides multidimensional protection against HTTP flooding attacks.
- Health check: Tengine performs health checks on ECS instances and automatically blocks abnormal ones. Once the faulty ECS instance recovers, Tengine automatically recovers the ECS instance.
- Supports session persistence.
- Supports consistent hash scheduling.

13 Virtual Private Cloud (VPC)

13.1 What is VPC

Virtual Private Cloud (VPC) is a private network established in Apsara Stack. VPCs are logically isolated from other virtual networks in Apsara Stack.

You have full control over your VPC. For example, you can select its IP address range and configure route tables and gateways. You can also use Alibaba Cloud resources such as ECS, RDS, and SLB in your own VPC. Additionally, you can connect a VPC to other VPCs or a local network to form an on-demand customizable network environment. This allows you to smoothly migrate applications to the cloud with little effort.

Components

Each VPC consists of a private CIDR block, a VRouter and at least a VSwitch.

- CIDR block

  When creating a VPC or a VSwitch, you must specify the private IP address range in the form of Classless Inter-Domain Routing (CIDR) block. For more information, see *Classless Inter-Domain Routing*.

  You can use any of the following standard CIDR blocks and their subnets as the IP address range of the VPC.

CIDR block	Number of available private IPs (system reserved ones not included)
192.168.0.0/16	65,532
172.16.0.0/12	1,048,572
10.0.0.0/8	16,777,212

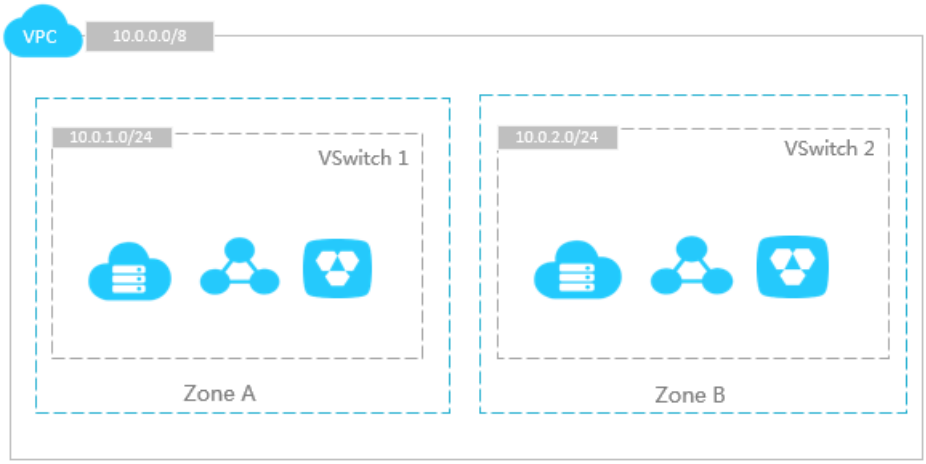
- VRouter

  VRouter is the hub of a VPC. As an important component of a VPC, it connects VSwitches in a VPC and serves as the gateway connecting the VPC with other networks. After you successfully create a VPC, the system automatically creates a VRouter, which is associated with a route table.

- VSwitch

VSwitch is a basic network device of a VPC and used to connect different cloud product instances. After creating a VPC, you can further segment your virtual private network to one or more subnets by creating VSwitches. The VSwitches within a VPC are interconnected. Therefore, you can deploy an application in VSwitches of different zones to improve the service availability.

Figure 13-1: Virtual Private Cloud



13.2 Architecture

Based on mainstream tunneling technologies, VPC isolates virtual networks. Each VPC has a unique tunnel ID, and each tunnel ID corresponds to only one VPC.

Backgroud information

With the continuous development of cloud computing, virtual network requirements are getting higher and higher, such as scalability, security, reliability, privacy, and connection performance. This gives a rise to a variety of network virtualization technologies.

The earlier solutions combined the virtual machine's network with the physical network to form a flat network architecture, such as the large layer-2 network. With the increase of virtual network scalability, problems are getting more serious for the earlier solutions. These problems include ARP spoofing, broadcast storms, host scanning, and more. Various network isolation technologies emerged to resolve these problems by completely isolating the physical networks from the virtual networks. One technology isolates users with VLAN, but VLAN only supports up to 4096 nodes. It cannot support the huge amount of users in the cloud.

VPC theory

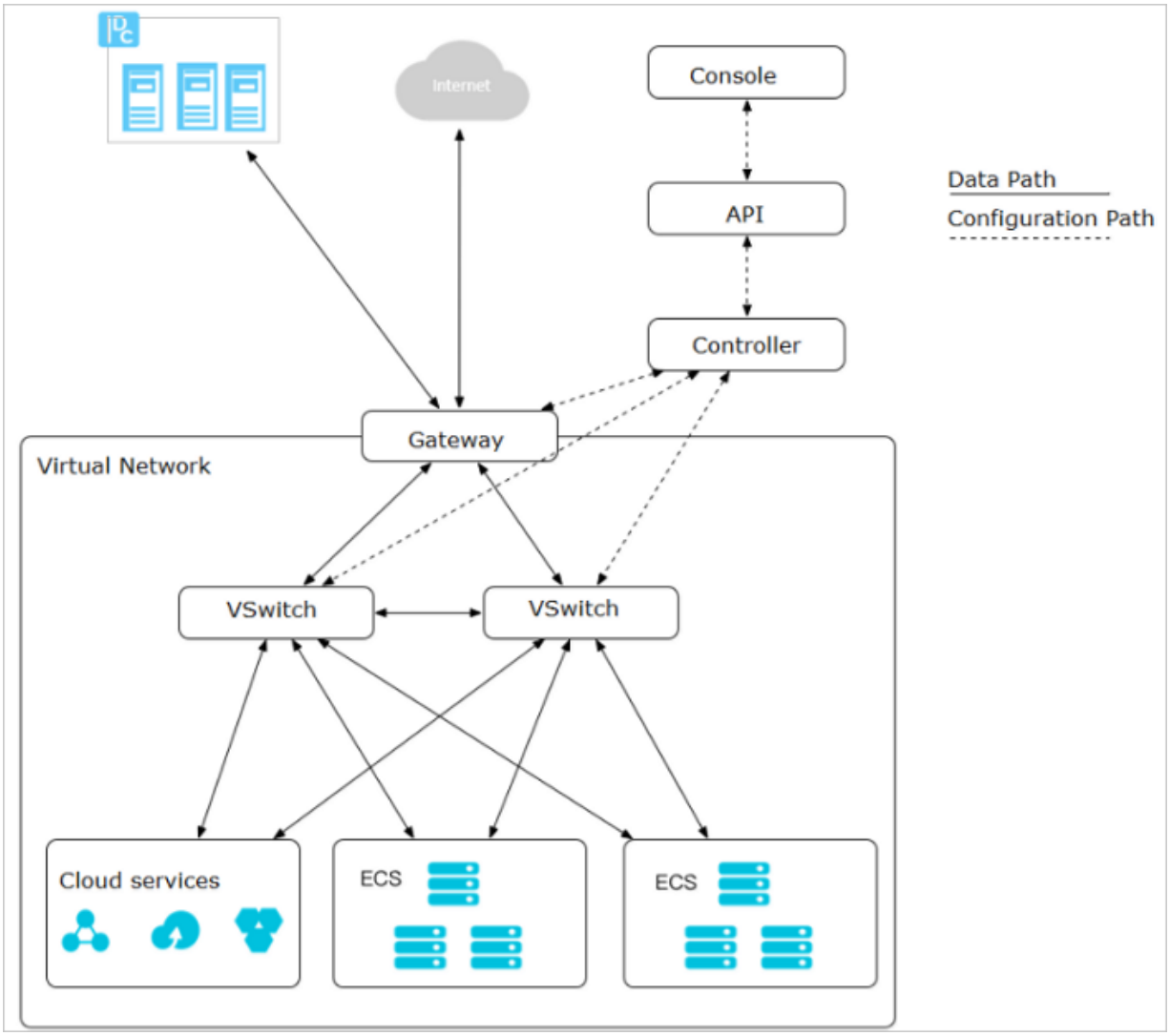
Based on tunneling technologies, VPCs isolate virtual networks. Each VPC has a unique tunnel ID , and a tunnel ID corresponds to only one VPC. A tunnel encapsulation carrying a unique tunnel ID is added to each data packet transmitted between the ECS instances within a VPC. Then, the data packet is transmitted over the physical network. Because the tunnel IDs are different for ECS instances in different VPCs and the ECS instances are located on two different routing planes, the ECS instances from different VPCs cannot communicate with each other and are isolated by nature.

With the tunneling technology, Alibaba Cloud has developed VSwitch, Software Defined Network (SDN) and hardware gateway and thus created VPC.

Logical architecture

As shown in the following figure, the VPC architecture contains three main components: VSwitches, gateway, and controller. VSwitches and gateways form the key data path. Controller s use the self-developed protocol to forward the forwarding table to the gateway and VSwitches, completing the key configuration path. In the overall architecture, the configuration path and data path are separated from each other. VSwitches are distributed nodes, the gateway and controller are deployed in clusters, and all links have redundant disaster recovery. This improves the overall availability of the VPC.

Figure 13-2: Logical architecture



14 Log Service

14.1 What is Log Service

Log Service is a one-stop solution designed for log data scenarios, providing functions such as collection, subscription, dump, and query of large volumes of log data.

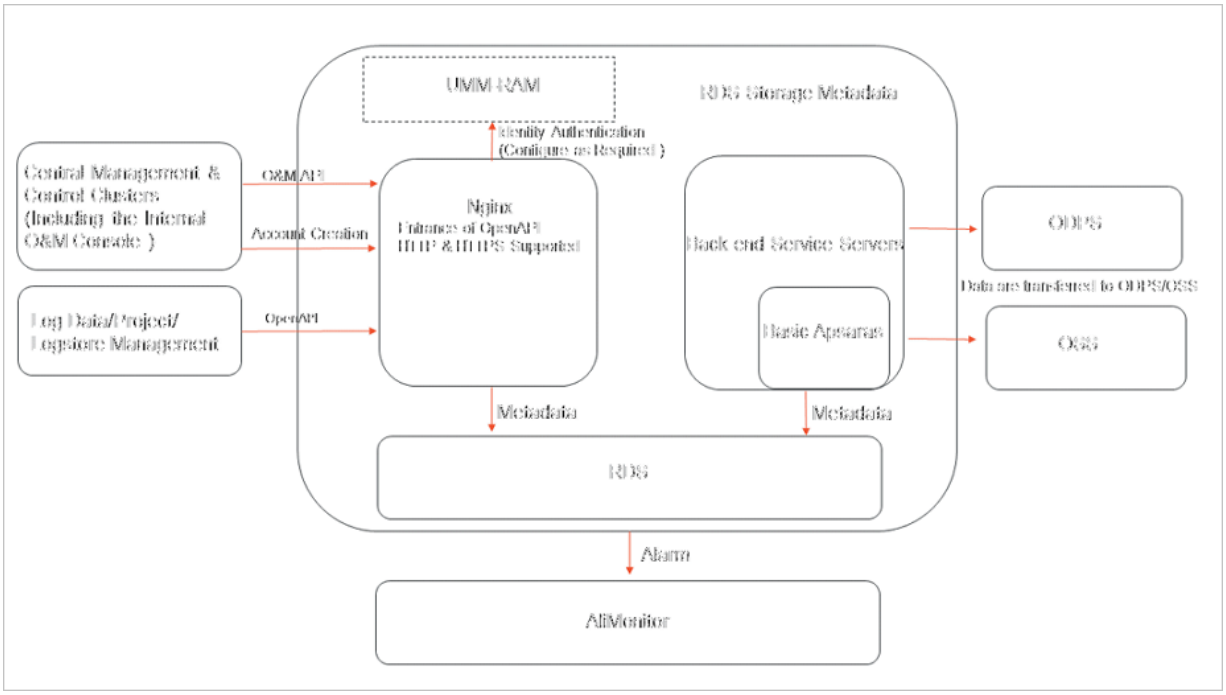
- Real-time collection and consumption: Real-time collection of data from multiple channels, including the client, API, Tracking JS, Library and other methods. After the data is written, it can be read in real time. Spark Streaming, Storm, Consumer Library and other interfaces can be used for real-time processing of data.
- Log data indexing and querying: Log Service creates indexes for log data in real time, provides real-time and powerful storage and query engines, and allows you to retrieve logs by time, keyword, context, and other dimensions.

Log Service supports automatic scaling, as well as horizontal scale-out to process PB-level data.

14.2 Architecture

The Log Service system architecture is shown in the following diagram.

Figure 14-1: Architecture



- Consoles and OpenAPI are on the left of the diagram and can interact with external modules.

- The following core modules are in the middle of the diagram:
  - UMM and RAM - the account module
  - RDS - metadata storage
  - Nginx - the frontend server
  - Log Service background - backend service servers

14.3 Components

Logtail

Logtail helps you quickly collect logs through the following features：

- Non-invasive log collection based on log files
  - It only reads files.
  - It is not invasive during the reading process.
- Secure and reliable
  - It supports file rotation, preventing data lost.
  - It supports local caching.
  - It provides a network exception retry mechanism.
- Convenient management
  - It offers a Web client.
  - It allows visualization configuration.
- Comprehensive self-protection
  - It allows real-time monitoring of process CPU and memory.
  - It has consumption and restrictions on CPU and memory usage.

Frontend servers

Frontend machines are built using LVS+Nginx. The features are as follows:

- HTTP and REST protocols
- Horizontal scaling
  - Frontend machines can be quickly added when traffic increases to improve processing capabilities.
- High throughput and low latency
  - Asynchronous processing: a single request exception will not affect other requests.

- LZ4 compression: The processing capabilities of individual machines are increased and network bandwidth consumption is reduced.

Backend servers

The backend is a distributed process deployed on multiple machines. It provides real-time Logstore data persistence, indexing, and query. The features of the overall backend service are as follows:

- High data security
  - Each log you write is saved in three copies.
  - Data is automatically recovered in case of any disk damage or machine downtime.
- Stable service
  - Logstores are automatically migrated in case of a process crash or machine downtime.
  - Automatic server load balancing ensures that traffic is distributed evenly among different machines.
  - Strict quota restrictions prevent abnormal behavior of a single user from affecting other users.
- Horizontal scaling
  - Horizontal scaling is performed using shards as the basic unit.
  - You can dynamically add shards as needed to increase throughput.

14.4 Features

Real-time log collection (LogHub)

Real-time collection. It uses over 30 methods to collect large volumes of data for real-time downstream consumption.

- Using Logtail to collect logs: stable, reliable, secure, available for all platforms (Linux, Windows , and Docker), high performance, and low resource utilization.
- Using APIs or SDKs to collect logs: flexible, convenient, scalable, and available in more than 10 languages and mobile terminals.
- Cloud product log collection: support for logs from Elastic Compute Service (ECS). It is convenient and efficient to integrate with these products.
- Other methods: Syslog, Unity3D, Logstash, Log4j, and Nginx.

Real-time log consumption (LogHub)

Stream computing, collaborative consumption library, and multiple-language support.

- Comprehensive functions: It is compatible with all Kafka functions while offering ordering, automatic scaling, time-frame-based retrieval, and other functions.
- Stable and reliable: It supports data consumption immediately after being written, multiple data copies, automatic scaling, and low cost.
- Easy to use: It supports Spark Streaming, Storm, Consumer Library (an automatic load balancing programming mode), SDK subscriptions, and more.

LogSearch

Real-time data indexing and querying. It creates indexes for LogHub data and support search by time and keyword.

- Large capacity: real-time indexing of PB-level data volumes (data can be queried within 1 second of writing) and query over a billion log entries per second.
- Flexible queries: support for search by fuzzy match, keyword, cross-topic, and context.

14.5 Benefits

It is a quick solution for large volumes of log data.

Typical Log Service application scenarios include: data collection, real-time computing, data warehousing and offline analysis, product operation and analysis, and O&M and management.

- Data collection and consumption
- ETL/Stream processing
- Event sourcing/tracing
- Log management

15 Apsara Stack Security

15.1 What is Apsara Stack Security

Apsara Stack Security is a solution that provides Apsara Stack with a full suite of security features, such as network security, server security, application security, data security, security management, and security operations services.

In today's cloud computing environment, new technologies are developed every day. Border security protection methods that use traditional detection technologies are insufficient to secure cloud businesses. Apsara Stack Security combines the powerful data analysis capabilities of Alibaba Cloud with the expertise of the Alibaba Cloud security operations team. It provides integrated security protection services at the network layer, application layer, and server layer.

Apsara Stack Security protects core business applications that provide services for the Internet. It provides real-time protection capabilities, including DDoS detection and prevention, Web attack detection and prevention, Web vulnerability detection and fix, server vulnerability detection and fix , and server intrusion prevention. Using a large amount of local security data and the intelligence collected from the cloud, this service performs centralized security big data analysis in the security data analysis engine cluster. It then presents security administrators with the overall security situation and intrusion tracing results, including targeted attack detection, user information leak alerts, and intrusion cause analysis. Based on this core security information, security administra tors can understand the security status and use the custom analysis interface provided by the security data analysis engine to perform scenario-based analysis on security data for flexible customization of security analysis capabilities.

15.2 Benefits

Since the enforcement of China Internet Security Law, Regulations on Critical Information Infrastructure Security Protection and Cloud Security Classified Protection Standard 2.0 have been published. As a result, private cloud platforms must pass the classified protection evaluation to ensure the security of cloud systems. Increasing security threats such as attacker intrusions and ransomware have led to the rising needs for security issue detection and prevention.

At the network perimeter of Apsara Stack, Apsara Stack Security uses a traffic security monitoring system to detect and block network-layer attacks in real time. This service detects and removes trojans and malicious files on servers to prevent attackers from exploiting the servers. Apsara Stack Security can also block brute-force attacks and send alerts on unusual logons. This

prevents attackers from stealing or destroying business data after logging on with weak passwords.

In-depth defense system

Apsara Stack Security comprises multiple function modules. These modules work together to provide in-depth defense on the Apsara Stack network perimeter, within the Apsara Stack network, and on the ECS instances in Apsara Stack. To help you manage cloud platform security risks in real time, Apsara Stack Security provides a unified security management system. This system allows you to manage the security policies in all security protection modules and perform association analysis on the logs.

Apsara Stack Security comprises multilevel security protection modules, including network security, server security, application security, and threat analysis. It provides an in-depth defense system on the cloud network perimeter, in the cloud network, and on the ECS instances. Using a management center that can integrate the security information from all modules, this service can accurately detect and block attacks. Apsara Stack Security effectively protects your business systems in the cloud against intrusions.

Security solutions completely integrated with the cloud platform

Apsara Stack Security is a product born from ten years of protection experience. With a decade of experience in providing security operations services for Alibaba Group's internal businesses and six years of safeguarding the Alibaba Cloud security operations, Alibaba has obtained considerable security research achievements, security data, and security operations methods, and has built a professional cloud security team. Developed based on the rich experience of Alibaba security experts, Apsara Stack Security is an attack prevention product specifically designed for cloud platforms. This product can effectively protect the security of the cloud network environments and business systems of users in Apsara Stack.

The components of Apsara Stack Security are software-defined, with a full hardware compatibility. With these components, you can implement elastic cloud computing services based on quick deployment, expansion, and implementation. The protection modules on the cloud network perimeter or in the cloud network adopt the bypass architecture, which completely fits the cloud businesses and has the minimal adverse impacts. To fit the flexibility of ECS instances, the protection modules running on them are all virtualized.

User security situation awareness

The cloud platform provides services for users. In Apsara Stack Security Center, a user can view the security protection data, generate security reports, and enable SMS and email alerts by configuring external resources.

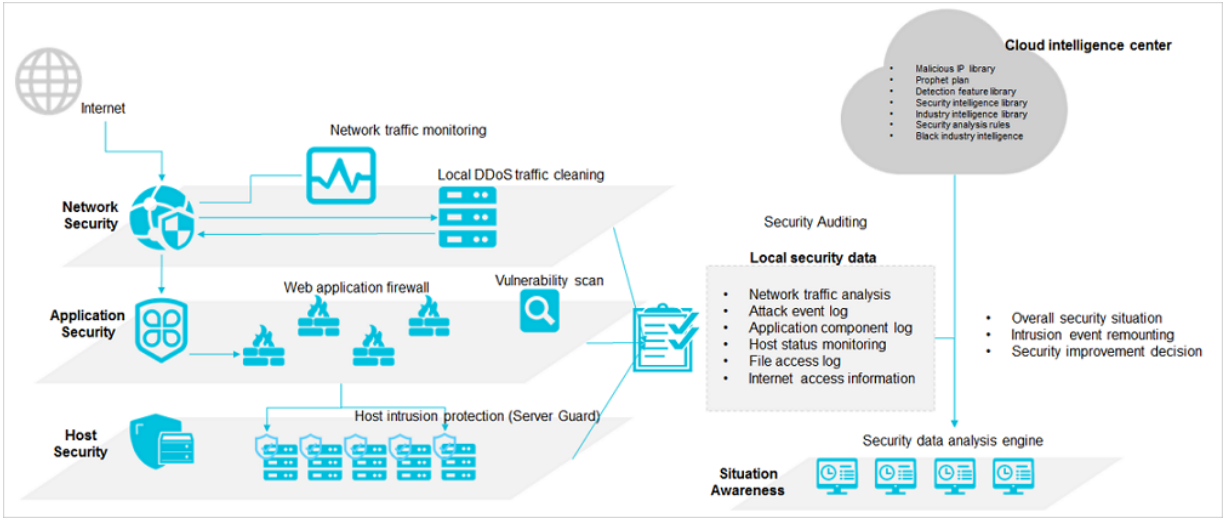
Security capability output

Apsara Stack Security has accumulated a large amount of protection policies over the last several years. The service has protected millions of users from hundreds of thousands of attacks every day. This has generated a large amount of security protection data. Apsara Stack Security analyzes over 10 TB of this data every day. The analysis results are used to enhance the fundamental security capabilities, such as the malicious IP library, malicious activity library, malicious sample library, and vulnerability library. These capabilities are then applied in the protection modules of Apsara Stack Security to enhance your business security.

15.3 Architecture

The architecture of Apsara Stack Security Standard Edition is shown in *Figure 15-1: Apsara Stack Security Standard Edition architecture*.

Figure 15-1: Apsara Stack Security Standard Edition architecture



- **Traffic Security Monitoring**: This module is deployed on the network perimeter of Apsara Stack. It allows you to inspect and analyze each inbound or outbound packet of an Apsara Stack network by traffic mirroring. The analysis results are used by other Apsara Stack Security modules.

- **Server Intrusion Detection**: This module collects information and performs detection through the client deployed on physical servers. It detects file tampering, suspicious processes, suspicious network connections, suspicious port listening, and other suspicious activities on all servers in the Apsara Stack environment. This helps you detect server security risks in time.
- **Server Guard**: This module provides security protection features such as vulnerability management, baseline check, intrusion detection, and asset management for ECS instances using log monitoring, file analysis, and signature scanning.
- **Security Audit**: This module collects database logs, server logs, user console operation logs, IT administrator console operation logs, and network device logs in Apsara Stack. This module can store and analyze these logs and trigger alerts on unusual events.
- **Web Application Firewall (WAF)**: This module protects Web applications against common Web attacks reported by OWASP, such as SQL injections, XSS, exploitation of Web server plugin vulnerabilities, trojan uploads, and unauthorized access. It blocks a large number of malicious visits to avoid website data leaks. This ensures the security and availability of your websites.
- **Threat Detection Service**: This service collects traffic data and server information and detects potential intrusions or attacks through machine learning and data modeling. It detects vulnerability exploitation and new virus attacks launched by advanced attackers, and shows you the information of ongoing attacks, enabling business security visualization and awareness.

Apsara Stack Security Basic Edition also provides on-premises security operations services. On-premises security operations services help users make good use of the features of Apsara Stack products and Apsara Stack Security products to ensure the user application security. Security operations services include pre-release security assessment, access control policy management , Apsara Stack Security product configuration, periodic security check, routine security inspection , and urgent event handling. These services cover the entire lifecycle of the user businesses in Apsara Stack. On-premises security operations services help users create a security operations system for cloud businesses. This system enhances the security of application systems and ensures the security and stability of user businesses.

You can also choose the following optional services based on your own business needs to enhance your system security.

- **DDoS Traffic Scrubbing**: This service detects and filters out DDoS attack traffic to block DDoS attacks.

15.4 Features

15.4.1 Apsara Stack Security Standard Edition

15.4.1.1 Traffic Security Monitoring

The Traffic Security Monitoring module is an Apsara Stack Security product that can monitor attacks within milliseconds. By performing in-depth analysis on the traffic packets mirrored from the Apsara Stack network ingress, this module can detect various attacks and unusual activities in real time and coordinate with other protection modules to implement defenses. This module also outputs a large amount of data to support the Apsara Stack Security defense system.

Features

The Traffic Security Monitoring module provides the following features:

Feature	Description
Traffic statistical analysis	Collects inbound and outbound traffic of the interconnection switch (ISW) using a bypass in traffic mirroring mode and generates a traffic diagram.
Suspicious traffic detection	Uses traffic mirroring to detect suspicious traffic that exceeds the scrubbing threshold and forwards the malicious traffic to the DDoS scrubber. The traffic rate (Mbps), packet rate (PPS), HTTP request rate (QPS), or number of new connections can be set as the threshold.
Malicious server detection	Detects attacks that are launched by malicious servers within the Apsara Stack network.
Web application protection	Based on Web attack detection rules, this service uses a bypass to block common Web application attacks, including SQL injection, code and command execution, script Trojan, file inclusion, and exploitation of upload vulnerabilities and common CMS vulnerabilities.
Suspicious TCP connection blocking	Uses a bypass to send TCP RST packets to the server and the client to block layer-4 connections.
Network log recording	Records UDP and TCP traffic logs and the Request and Response logs of HTTP queries. These logs are used by Threat Detection Service for big data analysis.

How it works

The Traffic Security Monitoring module performs in-depth packet analysis on traffic and detects attacks and unusual activity in real time. This module also reports security events to Apsara Stack Security Center as a data input for other protection modules. This module outputs a large amount of data to support the Apsara Stack Security defense system.

The Traffic Security Monitoring module collects data, processes the data, and then outputs data processing results. It uses sockets for data exchange.

- Collection: The module collects traffic data using multiple high-performance PCs with dual-port 10GE network interface cards.
- Processing: Traffic from an IP address may pass through multiple collectors. Therefore, traffic data must be consolidated to generate usable information.
- Output: The module stores and outputs the consolidated traffic data.

15.4.1.2 Server Intrusion Detection

The Server Intrusion Detection module collects information through the client program installed on a physical server. It detects file tampering, suspicious processes, suspicious network connections, suspicious port listening, and other activities on all servers in the Apsara Stack environment. This helps you detect server security risks in time.

Features

The Server Intrusion Detection module provides the following features:

Feature	Description
Key directory integrity check	Detects file tampering in the `/etc/init. d` path in the server system and generates alerts.
Suspicious process alert	Detects suspicious processes such as XOR DDoS trojans, Bill Gates DDoS trojans, and Minerd mining processes, and generates alerts.
Suspicious port listening alert	Detects new port listening tasks in time, and generates alerts.
Suspicious network connection alert	Detects connections to the public network actively initiated by internal network servers, and generates alerts.

How it works

The client program is installed on a physical server and collects information in real time, including the key directories, processes, open ports, and network connections. Then it detects suspicious events using rule and signature matching, and reports the suspicious events to the user.

15.4.1.3 Server Guard

Server Guard provides security protection features such as vulnerability management, baseline check, intrusion detection, and asset management for your servers. These security features are based on logging and monitoring, file analysis, and signature scanning.

Server Guard includes a client and a server. The client monitors attacks and vulnerabilities at the server layer and the application layer, and sends the data to the server for real-time server protection.

Features

Server Guard provides the following features:

Category	Feature	Description
Vulnerability management	Linux software vulnerability detection and fixes	Detects CVE vulnerabilities in the software on your servers based on the software versions. The software include SSH, OpenSSL, and MySQL. Provides vulnerability information and fixes.
	Windows vulnerability detection and fixes	Bases on the latest vulnerability information released by Microsoft, this feature detects critical Window vulnerabilities on your servers, and provides Windows patches to fix vulnerabilities such as the SMB remote code execution vulnerability. **Note:** By default, only critical vulnerabilities are reported. You can manually check for security updates and detect low-risk vulnerabilities.
	Web CMS vulnerability detection and fixes	Based on the security intelligence provided by Alibaba Cloud, this feature detects Web CMS vulnerabilities by scanning directories and files. It also provides patches developed by Apsara Stack Security to fix vulnerabilities in software such as WordPress and Discuz!, and allows you to undo vulnerability fixes.
	Configuration and component vulnerability detection	Detects vulnerabilities that cannot be detected by software version comparison or file vulnerability scanning, and identifies critical configuration vulnerabilities in software, such

Category	Feature	Description
		as configuration and component vulnerabilities including the Redis unauthorized access and ImageMagick vulnerabilities.
Baseline check	Account security baseline check	• Detects SSH, RDP, FTP, MySQL, PostgreSQL, and SQLServer accounts with weak passwords. • Detects the at-risk accounts of your servers , such as suspicious hidden accounts and cloned accounts. • Checks the password policy compliance of Linux servers. • Detects accounts without passwords on the servers.
	System configuration check	Checks the system group policies, logon baseline policies, and registry configuration risks, including: • Suspicious auto-startup items in the scheduled tasks of Linux servers. • Auto-startup items on Windows servers. • The sharing configurations of the system. • The SSH logon security policies of Linux servers. • The account-related security policies on Windows servers.
	Database security baseline check	Checks whether the Redis service on the server is exposed to the public network, whether unauthorized access vulnerabilities exist, and whether suspicious data is written to important system files.
	Benchmark compliance check	Checks whether the system baseline complies with the latest CIS Centos Linux 7 benchmark.
Intrusion detection (suspicious logons)	Disapproved logon location alert	Automatically records all logons, and determines the approved logon cities based on the usual logon locations. Generates alerts on logons in disapproved locations. You can customize the approved logon cities.

Category	Feature	Description
	Disapproved logon IP alert	Generates alerts on logons using IP addresses that are not whitelisted after a logon IP whitelist is created.
	Disapproved logon time alert	Generates alerts on logons within disapproved time ranges after the approved logon time range is set.
	Disapproved logon account alert	Generates alerts on logons with disapproved accounts after the approved logon account list is created.
	Brute-force attack detection and blocking	Detects and blocks brute-force attacks in real time. Both SSH and RDP brute-force attacks can be detected.
Intrusion detection (webshells)	Webshell detection	Uses a webshell detection engine developed by Apsara Stack Security to detect and remove webshells on the servers. Both scheduled and real-time scans are supported. Detects webshells written in ASP, PHP, and JSP, and allows you to manually quarantine these webshell files.
Intrusion detection (suspicious processes)	Suspicious process activity detection	Detects suspicious process activities such as reverse shells, Java processes running CMD commands, and unusual Bash file downloads.
Asset management	Asset grouping	Allows you to group the servers into four groups, filter assets by region, online status, or other features, and manage the asset tags.
	Asset fingerprints	• **Ports**: Collects and displays the port listening information and records changes. This allows you to easily check the port listening status. • **Accounts**: Collects the account and permission information to discover privileged accounts and detect privilege escalations. • **Processes**: Collects and displays the information of processes using snapshots, to list normal processes and detect suspicious processes. • **Software**: Lists the software installation information so that the affected assets

Category	Feature	Description
		can be quickly located when a critical vulnerability is exploited.
Server logs	Log retrieval	• Previous logons logs: Records successful system logons. • Brute-force attacks logs: Records failed system logons. • Process snapshot logs: Records the information of running processes on the server at a specific time. • Port listening snapshot logs: Records the port listening information on the server at a specific time. • Account snapshot logs: Records the logon accounts information on the server at a specific time. • Process startup logs: Records the process startup information on the server. • Network connection logs: Records the network connections started by this server.

How it works

Server Guard includes a client and a server. The client is installed on servers. The client communicates with the server using TCP keepalive and obtains scripts, rules, and installer packages from the server using HTTP.

The client can be used in Windows or Linux. It can automatically connect to the server for online updates.

The key modules of Server Guard work as follows:

- **Vulnerability management**: First, the client collects the server information, including component information, software versions, file information, and registry information. Then, the client checks whether these information match with the vulnerability detection rules provided by the server. The information that matches the rules will be sent to the server for further analysis. The detected vulnerabilities will be displayed in the Server Guard console. Users can fix vulnerabilities in the console or by calling APIs. After receiving the vulnerability patches from the server, the client on the vulnerable server automatically fixes the vulnerabilities and synchronizes the vulnerability status to the server.

- **Baseline check**: When the user manually starts a scan or a periodic scan is triggered, the Server Guard server sends a baseline check request to the client. The client then collects the server information according to the scan policy and compares the information with the security baseline. Scanned items that do not comply with the baseline are labeled as at-risk items and reported to the server.
- **Unusual logon detection**: The client monitors the logon log of the server system in real time. In a Linux system, the `/var/log/secure` and `/var/log/auth.log` files are also monitored. All failed and successful logons are recorded. Unusual logons or brute-force attacks will be reported to the server.
- **Webshell detection**: The client uses an Alibaba-developed dynamic webshell detection engine to detect complex webshells. It then restores these webshells to an identifiable status to analyze the hidden webshell activities. This prevents webshells from bypassing the detection due to the use of static detection rules.
- **Suspicious process detection**: The Server Guard server uses a data analysis rules engine to analyze the server process data collected by the client. By doing so, the server can detect suspicious processes such as reverse shells, mining processes, DDoS trojans, worms, viruses, and hacking tools.
- **Log collection**: The client collects logs such as processes logs and network logs.

Scenarios

Server Guard is applicable to server security protection in the following scenarios:

- **Common software is used for website building.**

  In this scenario, attackers may intrude into servers by exploiting vulnerabilities in common software. You can use Server Guard to detect and fix vulnerabilities.
- **Web application services are used.**

  Attackers may steal website data through both internal and external Web services. You can use Server Guard to prevent attackers from launching attacks or controlling your servers.

15.4.1.4 Security Audit

The Security Audit module is an integrated audit solution based on the cloud computing platform. This module meets the basic requirements for classified security protection of information systems . It provides behavior log collection, storage, analysis, and alerts at the physical server layer, network device layer, and cloud computing platform application layer.

This module collects logs from data sources such as cloud services, network devices, servers, and databases. It then uses audit rules and an audit rules engine to perform an Apsara Stack security audit based on the collected data and generates alerts on activities that meet the rule conditions. This module also allows you to query logs and configure rules.

Features

The Security Audit module provides the following features:

Feature	Description
Audit overview	Supports queries by time, database, network, server, user operation, maintenance operation, and other metrics. Generates reports on the audit operation details, such as raw logs, audit events, audit risks, log usage, and storage usage.
Raw logs	Allows you to query all raw logs within 7 days by audit target, audit type, risk severity, time, keyword, and other metrics.
Audit query	Allows you to query the audit logs that record events meeting the audit rule conditions within 30 days by audit target, audit type, risk severity, time, keyword, and other metrics.
Policy configuration	• **Audit policy settings**: Allows you to query the audit rules on cloud services, servers, network devices, and database configurations. You can also add, modify, or delete these rules. • **Audit type settings**: Allows you to query and add audit types. • **Alert settings**: Allows you to set the alert recipients based on the audit rules and audit risks. • **Log archive management**: Allows you to query and download all raw log files within 185 days. • **Log export management**: Allows you to query and manage log export tasks. • **System settings**: Allows you to configure global parameters for the audit system, including the amount of daily alerts, daily audited logs, server logs, network device logs, user operations logs, and maintenance operation logs.

How it works

The Security Audit module collects logs from multiple data sources and stores the logs in Log Service. This module audits these logs by product or log type, and detects invalid operations using the audit rules engine. The information of operations that meet the audit rule conditions is sent as

alerts to the responsible engineers and archived by the module. The logs that do not contain any invalid operations are archived for later queries.

Benefits

The Security Audit module has the following features and benefits:

- **Full-coverage behavior logs**

  This module can collect behavior logs from multiple businesses in Apsara Stack and physical servers from various perspectives. This ensures the full coverage of audit. The log collection center supports centralized and synchronized collection of behavior logs in quasi-real time.

- **Reliable log storage**

  Audit logs are stored based on cloud computing storage services and clustered in three copies . This ensures secure and stable storage. The storage space can be quickly expanded.

- **Real-time query of large amounts of data**

  This module creates a global index for large amounts of log data. This enables fast data retrieval.

15.4.1.5 Web Application Firewall

Web Application Firewall (WAF) protects the Web applications of cloud users against common Web attacks.

These attacks can be common Web application attacks such as SQL injections, XSS, or the attacks that affect website availability by consuming resources, such as HTTP flood. WAF allows you to customize protection policies based on the businesses on your website to block malicious Web requests.

WAF protects the traffic of businesses on HTTP and HTTPS websites. In the WAF console, you can import certificates and private keys to enable end-to-end encryption. This prevents the interception of business data on the links.

Features

The WAF module provides the following features:

Feature	Description
Protection against common Web attacks	Protects your website against common Web attacks, such as SQL injections, XSS, file uploads, file inclusions, common directory traversal, common CMS vulnerability exploits, code injections, webshells, and scanner attacks.

Feature	Description
	The detection and prevention modes are provided to handle Web attacks: • In detection mode, WAF generates alerts for the attacks, but does not immediately block the traffic. You can determine whether the alert is a false positive. • In prevention mode, WAF blocks attack-related requests.
HTTP flood mitigation	Collects information including the URL request frequency, source IP address distribution, and unusual response codes, and blocks unusual activities. The normal and urgent modes are provided to block HTTP flood attacks. You can enable the urgent mode for enhanced protection if HTTP flood attacks have caused the website to become inaccessible. This service provides default protection rules and allows you to customize rules on the URL requests initiated by a source IP address.
Precise access control	Provides an easy-to-use interface for access control policy configurat ion. You can combine common HTTP fields, such as IP, URL, Referer , and User-Agent, to create precise access control policies that are applicable to scenarios such as hotlinking protection and website back -end protection.
Automatic blocking of malicious IP addresses	Automatically blocks an IP address for a period of time if this IP address continuously initiates Web attacks to the domain name.
Region blocking	Blocks source IP addresses from specific provinces or regions outside China.

How it works

WAF configures domain forwarding to redirect the website access requests to the WAF cluster . It then checks, filters, and scrubs the request traffic, and uses a reverse proxy to forward valid requests to the origin server. This enables application-layer protection for the origin server.

Scenarios

WAF can be used for Web application protection in fields such as government, finance, insurance, e-commerce, O2O, Internet Plus, and games. It provides the following features:

- Prevents website data leaks caused by SQL injections.
- Mitigates HTTP flood attacks by blocking large numbers of malicious requests. This ensures the availability of your website.
- Prevents website defacement based on trojans to ensure the credibility of your website.

- Provides virtual patches that enable quick fix for newly discovered vulnerabilities.

15.4.1.6 Threat Detection Service

Threat Detection Service is a big data security analysis system developed by the Alibaba Cloud security team.

Based on machine learning and data modeling, Threat Detection Service analyzes the server traffic and network traffic in Apsara Stack to detect threats, attacks, suspicious visits, and other suspicious activities. It can also detect vulnerability exploitation and new virus attacks launched by advanced attackers from the perspective of attackers. The service displays the data of ongoing attacks to enable business security visualization and awareness.

Features

Threat Detection Service provides the following features:

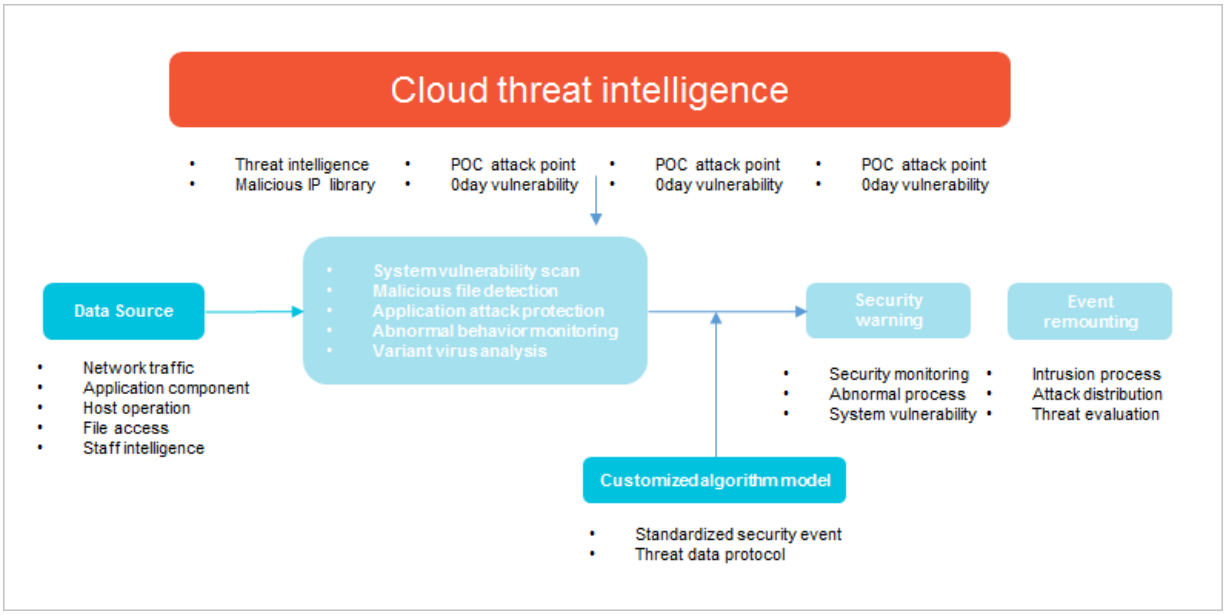
Feature	Description
Security situation overview	Provides the overall security information, including the number of emergencies, the current day's attacks, the current day's flaws, attack trend, latest threat analysis, latest intelligence, and protected assets information.
Access analysis	Analyzes all information about the access to the protected Web services, including the top 10 accessed services, number of normal source IP addresses, number of malicious source IP addresses, number of crawler source IP addresses, and detailed access samples.
Screens	Provides map-based traffic data screens and server security screens.
Security event analysis	Uses big data algorithms and models to detect the following security events in the cloud: • Zombie activities: A server becomes a zombie that is controlled by an attacker and launches DDoS attacks on other servers. • Brute-force attacks: An attacker logs on to a server through brute-force attacks. • Backdoors: The WannaCry ransomware, unusual MySQL scripts, or webshells are detected. • DDoS attacks: A server encounters DDoS attacks. • Hacking tools: The logon credentials of a server is stolen, and hacking tools and attacks are detected. • Suspicious network connections: An attacker uses PowerShell to download suspicious files, runs suspicious VBScript commands , downloads malicious files or scripts to Linux servers, or runs commands in reverse shells.

Feature	Description
	• Suspicious network traffic: A mining process is running.
Traffic statistical analysis	Collects the traffic data in the monitored IP range, including the current day's traffic, traffic in the last 30 days, traffic in the last 90 days, and the QPS. Displays the traffic data of a specific IP address.
Malicious server identification	Detects attacks launched by internal malicious servers, such as HTTP flood and DDoS attacks, and identifies the controlled malicious servers.
Web attack detection	Detects Web vulnerability exploitation, malicious scanning tools, webshell uploads and connections, SQL injections, XSS, local and remote file inclusion attacks, code or command execution, and other attacks.
Server vulnerability exploitation detection	Converts packet feature characters into binary strings and matches these strings with signatures to detect security events such as the exploitation of Redis server vulnerabilities.
Application vulnerability analysis	Detects Web application vulnerabilities and provides advice on vulnerability fixes and vulnerability fix verification. Periodically and automatically scans NAT assets and servers for application vulnerabil ities, verifies the detected vulnerabilities, and updates the vulnerability status.
Weak password analysis	Detects weak passwords of accounts in common systems such as Web , SSH, FTP, MySQL, and SQL Server and allows users to customize weak password policies. Automatically scans NAT assets and servers at a scheduled time each day, verifies the detected weak passwords, and updates the weak password detection time.
At-risk configuration detection	Scans the access to external service pages, generates alerts on leaks of Web page configuration items, verifies the detected leaks of configuration items at a scheduled time every day, and updates the detection time.

How it works

How Threat Detection Service works is shown in *Figure 15-2: How Threat Detection Service works*.

Figure 15-2: How Threat Detection Service works



- **Big data security analysis platform**
  - **Network**: Threat Detection Service uses the HTTP requests and responses collected by the traffic security monitoring module to create HTTP logs, and uses big data models to analyze the logs and discover security events and threats.
  - **Server**: Threat Detection Service uses the rules engine to analyze the server process data collected by Server Guard and discover security events and threats.
- **Security events**: Threat Detection Service shows you all security events collected from the following three sources:
  - Security events reported by Server Guard.
  - Server security events discovered by server process analysis based on the rules engine.
  - Network security events discovered by HTTP log analysis based on big data models.
- **Flaw analysis**: This feature involves the cactus-batch and cactus-keeper modules.
  - The cactus-batch module processes the URLs to be scanned and sends the URLs to cactus -keeper through message queues.
  - The cactus-keeper module is integrated with a scan engine, Alibaba's extensive scanning rules, and plugin libraries accumulated over many years. It can scan for vulnerabilities, weak passwords, and at-risk configuration items. After detecting system flaws, this module sends reports to the user. This allows users to handle the flaws in time.

Benefits

Threat Detection Service has the following features and advantages:

- **Fast scan covering all vulnerabilities**

  The flaw analysis feature adopts the stateless scanning technology, and can concurrently scan 10,000 IP addresses per second when the bandwidth is 5 MB. With its patented technology in third-party software vulnerability scanning, this service can rapidly scan a third-party CMS using fingerprint recognition.

  Working with the traffic security monitoring module, this feature can monitor URL requests on the network in real time and scan all target URLs and interfaces. It also supports scanning by binding to the proxy, HTTPS, and DNS.

  This service scans for the following flaws:

  - More than 30 common Web vulnerabilities and more than 150 common Web application vulnerabilities, which cover all vulnerability types collected by OWASP, WASC, and CNVD.
  - Common weak passwords of systems and databases.
  - Malicious data tampering in environments such as Web 2.0, Ajax, PHP, ASP, .NET, and Java.
  - Malicious tampering of the encoding mode of complex characters.
  - Malicious changes in compression methods such as Chunk, Gzip, and Deflate.
  - Malicious changes in authentication methods such as Basic, NTLM, Cookie, and SSL.

  Backed by Alibaba Cloud's big data computing capabilities, Apsara Stack Security can perform daily data mining and analysis on a large number of attacks on the Alibaba Cloud platform. This service can obtain samples and intelligence of the latest attacks in real time and detect zero-day vulnerabilities. Then vulnerability libraries are generated to ensure more timely and comprehensive flaw analysis.

- **Threat analysis based on big data**

  Threat Detection Service provides analysis and computing of petabyte-level big data and collects all security data and threat intelligence of the entire network. It also uses machine learning technologies to create complete and smart security threat models that can be used in the application scenarios of millions of users.

  This service focuses on the security trends and new threats that are faced by users of cloud computing in data centers, such as targeted Web application attacks, system brute-force

attacks, system intrusions, and application-layer and server-layer vulnerabilities. It defends user systems against these threats from different fields.

- **Screens**

  Based on Internet visualization technologies, Threat Detection Service displays the results of big data threat analysis in graphs on screens to support security decision making on Apsara Stack.

15.4.1.7 On-premises security operations services

To ensure the stability, reliability, security, and regulatory compliance of the cloud platform, Apsara Stack Security Standard Edition provides multiple security products and on-premises security operations services to ensure the availability, confidentiality, and integrity of the systems and data of users. Security operations services are indispensable in the security system. The combination of security products and security operations services gives full play to the security features of both Apsara Stack products and Apsara Stack Security products, and enhances the security of the Apsara Stack network environment from both technology and management aspects.

On-premises security operations services aim to help users use the security features of both Apsara Stack products and Apsara Stack Security products to protect the user applications. Security operations services include services that cover the entire security lifecycle of Apsara Stack user businesses, such as pre-release security assessment, access control policy optimizati on, periodic security assessment, routine security inspection, and emergency response. These services help users create a cloud security operations system to enhance the application system security and ensure secure and stable businesses.

Services

On-premises security operations services are as follows:

Table 15-1: On-premises security operations services

Category	Service	Description
User business security operations	User asset research	With the authorization of a user, this service periodically researches the cloud businesses of the user and develops a business list containing information such as the business system name, ECS, RDS, IP address, domain name, and owner.

Category	Service	Description
	New business security assessment	• Before a user migrates a new business system to the cloud, this service detects system vulnerabilities and application vulnerabilities in the new business system using both automation tools and manual operations. • Provides advice and verification on vulnerability fixes.
	Periodic business security assessment	• Periodically uses automation tools to detect system vulnerabilities, application vulnerabilities, and security risks in running businesses. • Provides advice on handling detected risks , including but not limited to security policy settings, patch updates, and application vulnerability handling.
	Access control management	Provides inspection and guidance on applying access control policies when a new business is migrated to the cloud.
	Access control routine inspection	Periodically checks for access control risks of user businesses.
	Security risk routine inspection	Monitors and inspects security events in Apsara Stack Security. Informs the user of verified events and provides advice on event handling.
Apsara Stack Security operations	Rule update	Periodically updates the rule libraries of Apsara Stack Security products.
	Product integration	• Provides support for integrating Apsara Stack Security products with the applicatio n systems of users. • Helps users customize and optimize security policies.
Security event response	Event alerts	Synchronizes recent security events information from Alibaba Cloud, and helps users remove the risks.
	Event handling	Handles urgent events such as attacker intrusions.

Service output

On-premises security operations services output the following documents:

- Weekly, monthly, and yearly service reports
- Asset lists
- System security check reports

SLA

The SLA terms of on-premises security operations services are as follows:

- **Asset management**: Update the asset list once a month.
- **Security event response**: Respond within 30 minutes during work hours.
- **Security check**:
  - Complete a pre-release security check within two workdays.
  - Perform a periodic security check once a quarter.

Duties

Partners authorized by Alibaba Cloud provide on-premises security operations services, and Alibaba Cloud provides service quality management and technical support.

Owner	Duties
Alibaba Cloud	• Assign and manage tasks of service providers and on-premises engineers. • Assess the services provided by service providers and on-premises engineers. • Train service providers and on-premises engineers and provide technical support. • Provide project coordination and process and quality management.
Service provider	• Perform security check and routine inspection on the system of the user. • Provide advice on fixing vulnerabilities. • Maintain the access control policies of the user resources. • Update and maintain the security rules and policies of Apsara Stack Security. • Respond to security events. • Provide security technical support for users.
User	• Authorize service providers to perform security operations.

Owner	Duties
	• Follow the security advice to carry out the security plans on businesses. • Improve the security system.

Risk control

The following measures are taken to control risks in on-premises security operations services:

Category	Risk Item	Measure
Engineer and organization qualification	Organization	Only Alibaba Cloud and authorized enterprises can provide security services.
	Engineers	All engineers must be assessed and trained by the Alibaba Cloud security team.
Confidentiality	Confidentiality agreements	All enterprise and individual service providers must sign a confidentiality agreement.
Service tool security	Tool selection	Only security tools specified by Alibaba Cloud are allowed.
	Tool use	Apply standard configurations to avoid risks in using the tools.
Operation security	Operation procedure	Perform at-risk operations, such as scanning, in batches.
	Risk notification	Inform the users of risks in the operations, and provide risk avoidance and control methods. Perform operations only with the consent of the users.

15.4.2 Optional security services

In addition to the security services provided by Apsara Stack Security Standard Edition, multiple optional security services are also provided to meet various security needs. We recommend that you choose optional security services based on your business needs.

15.4.2.1 DDoS Traffic Scrubbing

Supported by a large-scale distributed operating system developed by Alibaba Cloud and more than a decade of experience in security protection, Alibaba Cloud has developed a DDoS Traffic Scrubbing module based on the cloud computing architecture. This module can protect user businesses on the cloud platform against large numbers of DDoS attacks.

Features

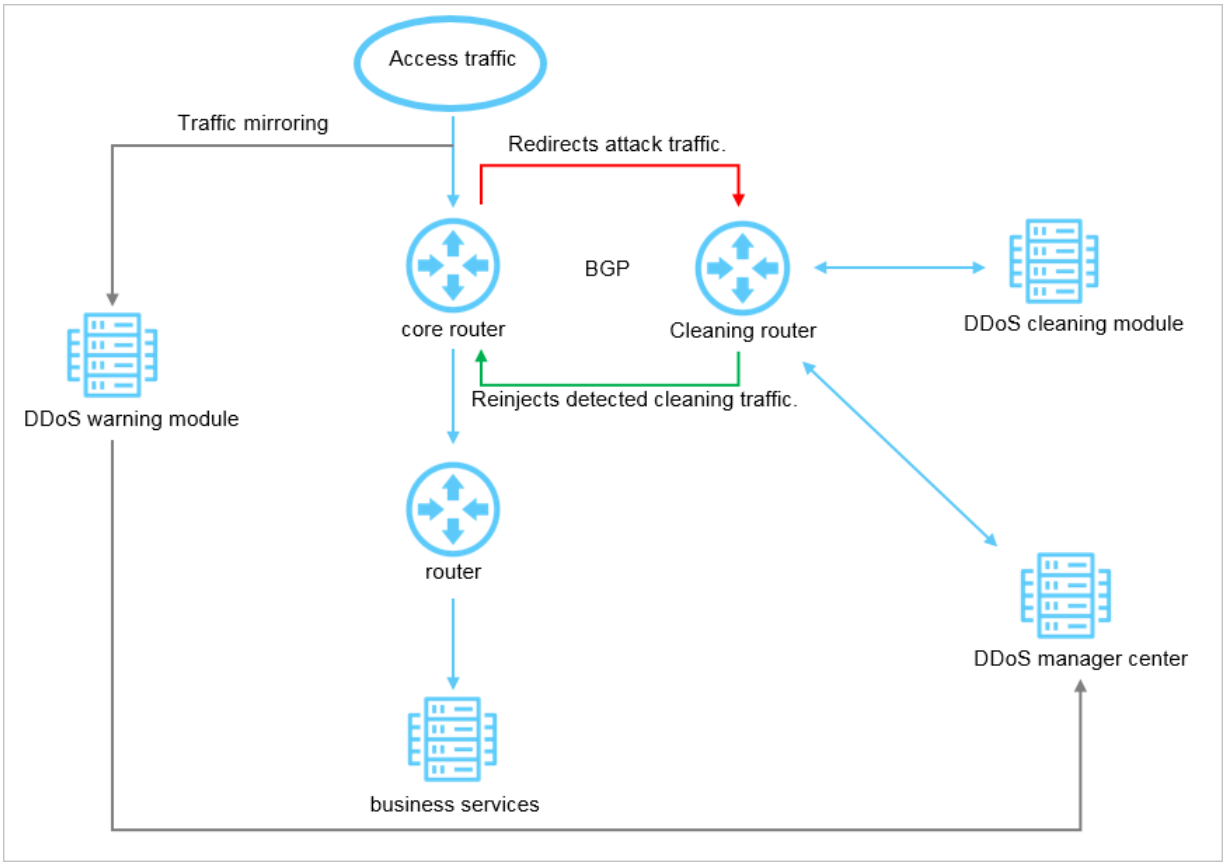
The DDoS Traffic Scrubbing module provides the following features:

Feature	Description
DDoS traffic scrubbing	Detects and prevents attacks such as SYN flood, ACK flood, ICMP flood, UDP flood, NTP flood, DNS flood, and HTTP flood.
DDoS attack display	Allows you to view DDoS attacks in the console and search for DDoS attacks by IP address, status, and event information.
DDoS traffic analysis	Allows you to monitor and analyze the traffic of a DDoS attack, and view the attack traffic protocol and the 10 IP addresses that have launched most attacks.

How it works

After the Traffic Security Monitoring module detects unusual traffic, the DDoS Traffic Scrubbing module reroutes, scrubs, and reinjects the traffic, as shown in *Figure 15-3: Traffic scrubbing*. This mitigates DDoS attacks and ensures normal running of businesses.

Figure 15-3: Traffic scrubbing



The Traffic Security Monitoring module sends information about the detected DDoS attacks to the DDoS Traffic Scrubbing module. The DDoS Traffic Scrubbing module is connected with the border gateway device. When a DDoS attack is detected, this module configures a BGP path for the border gateway to reroute the attack traffic to the traffic scrubbing system. The DDoS Traffic Scrubbing module then scrubs the traffic based on the configured scrubbing policies, filters out unusual traffic, and reinjects the normal traffic to the border gateway.

Benefits

The DDoS Traffic Scrubbing module has the following features and benefits:

- **Detection of all common DDoS attacks**

  This module prevents all DDoS attacks at the network layer, transport layer, and application layer, such as HTTP flood, SYN flood, UDP flood, DNS query flood, stream flood, ICMP flood, and HTTP GET flood. It also informs the user of the protection status with SMS notifications in real-time.

- **Automatic response to attacks within one second**

  This module uses world leading attack detection and prevention technologies. It can complete the protection process within one second, covering attack discovery, traffic rerouting, and traffic scrubbing. This module triggers traffic scrubbing when the traffic scrubbing thresholds are violated or when DDoS attacks are detected during network behavior analysis. This reduces network jitter and ensures the availability of your businesses in the case of DDoS attacks.

- **High scalability and high redundancy of anti-DDoS capabilities**

  With high scalability and high redundancy of the cloud computing architecture, this module can be easily scaled up to realize high scalability of anti-DDoS capabilities.

- **Bidirectional protection that avoids abuse of cloud resources**

  The DDoS Traffic Scrubbing module protects your system against external DDoS attacks and detects invalid behaviors of your cloud resources. If your ECS instance in Apsara Stack is used to launch DDoS attacks, the traffic security monitoring module will cooperate with Server Guard to restrict the network access of the hijacked ECS instance and generate an alert.

16 Key Management Service (KMS)

16.1 Product overview

Key Management Service (KMS) is a secure and easy-to-use management service provided by Alibaba Cloud. The confidentiality, integrity, and availability of keys are guaranteed at a low cost.

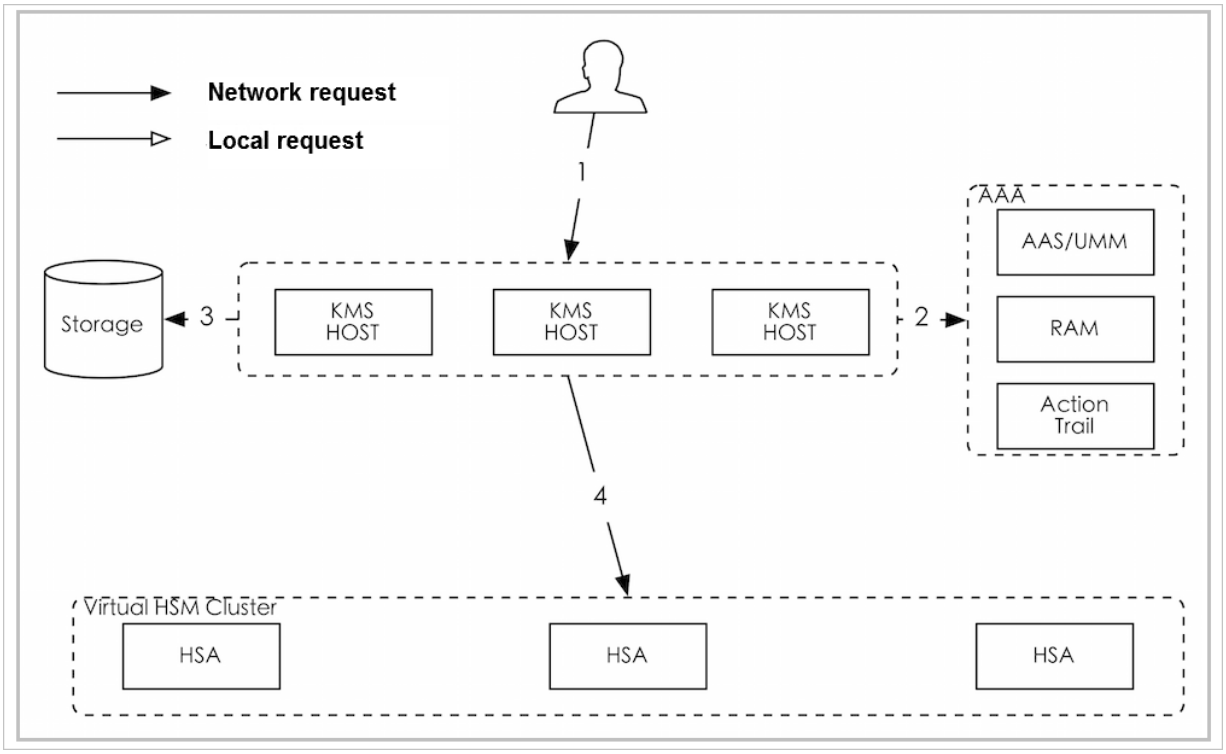
With the help of KMS, you can use keys securely and conveniently, and focus on developing encryption/decryption scenarios.

16.2 Product architecture

KMS is deployed in different regions. Each region provides the same functions, but the data is mutually independent. In a single region, KMS adopts a distributed architecture composed of multiple equivalent nodes. All the nodes in a single region provide the same level of availability, allowing you to resize the service based on your actual access needs.

The KMS architecture is shown in *Figure 16-1: Product architecture*.

Figure 16-1: Product architecture



KMS is divided into four modules:

- Storage

This module is for storing exported key tokens (EKTs) and other metadata.

- AAA

  This modole is for authentication, authorization, and auditing.

- KMSHOST

  This module is for processing user API requests.

- HSA (Hardware Security Appliance)

  The Hardware Security Module (HSM) is for processing the cryptographic logic of KMS powered by RAFT protocol distributed storage and trusted computing technology.

16.3 Functions and features

16.3.1 Convenient key management

You can use the APIs provided by KMS or the KMS console to conveniently manage your keys.

- You can disable and enable user keys at any time. After a key is disabled, the data encrypted using this key cannot be decrypted.
- A pre-deletion policy is used to delete keys. You can cancel key pre-deletion at any time, reducing the potential impact of accidental operations.
- You can use Resource Access Management (RAM) to manage key permissions and separate key encryption and decryption permissions.
- You can use EncryptionContext to enhance control over keys and ciphertext data.

16.3.2 Envelope encryption technology

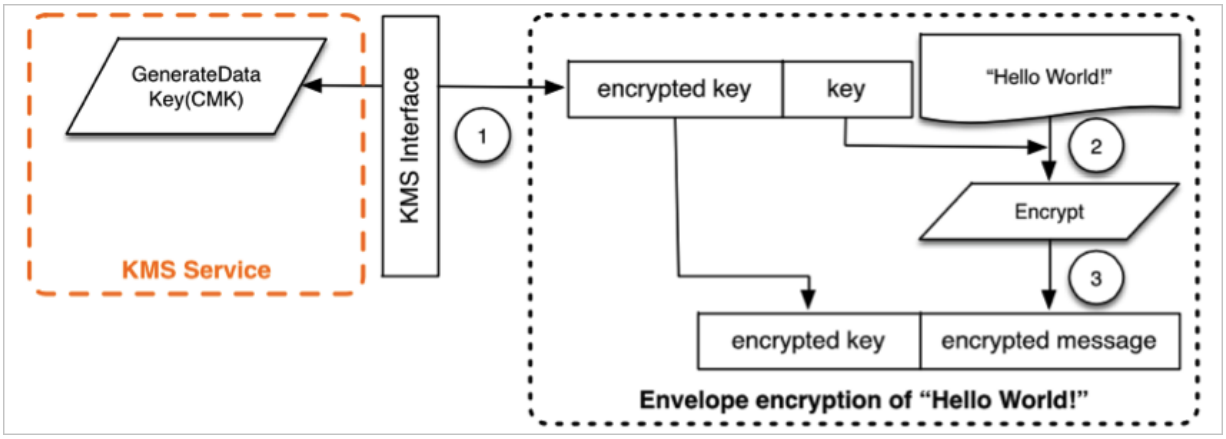
Although KMS provides the Encrypt API, it does not actually encrypt data. KMS provides a Customer Master Key (CMK) management service and data key encryption/decryption service. You have to use the data keys to encrypt data yourself.

You can encrypt data using your own data key and then use the Encrypt interface to protect your data key. Or, you can obtain a data key from the KMS GenerateDataKey API.

Encryption process

For the envelop encryption process, see *Figure 16-2: Encryption flowchart*.

Figure 16-2: Encryption flowchart



As you can see in *Figure 16-2: Encryption flowchart*, the encryption process is as follows:

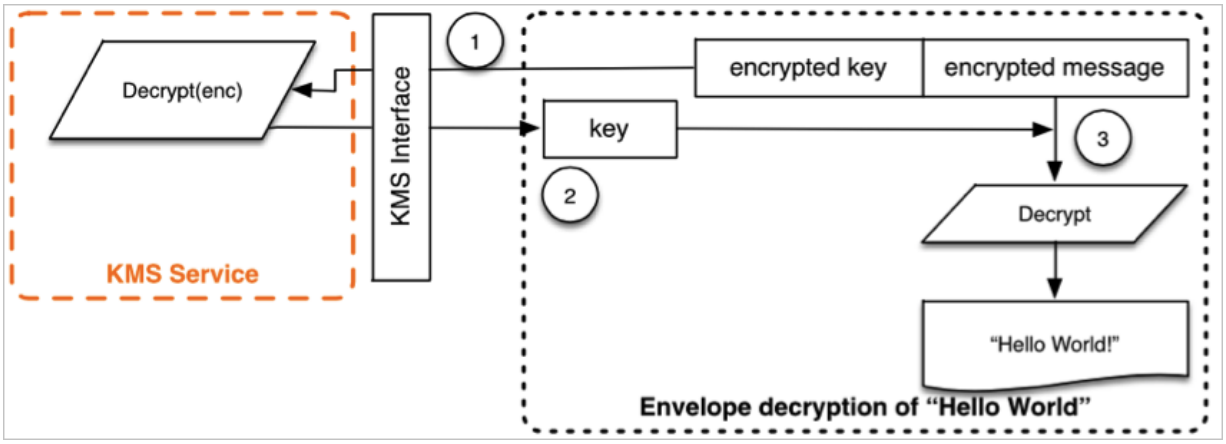
1. Use the specified CMK to generate a data key and obtain the key and encrypted key.

   Or, you can generate your own data key and use the Encrypt interface to obtain the correspond ing encrypted key.
2. Use the data key to encrypt the data and obtain the ciphertext data.
3. Store the ciphertext data together with the encrypted key.

Decryption process

For the envelop decryption process, see *Figure 16-3: Decryption flowchart*.

Figure 16-3: Decryption flowchart



As you can see in *Figure 16-3: Decryption flowchart*, the decryption process is as follows:

1. Use KMS to decrypt the encrypted key.
2. Obtain the plaintext key.

3. Use the plaintext key to decrypt the ciphertext data and obtain the plaintext data.

16.3.3 Secure key storage

KMS guarantees the security of keys during storage in the following ways:

- Customer Master Key (CMK) plaintext only appears in the memory of the HSA module. The KMS storage module only stores the CMK ciphertext.
- CMKs are encrypted using the HSA module's domain key. This domain key is rotated once per day.
- The domain key is encrypted for storage using trusted computing technology and stored based on a distributed storage protocol. This guarantees the high reliability of the domain key.

17 Domain Name System (DNS)

17.1 What is Apsara Stack DNS

Apsara Stack DNS is a service that runs on Apsara Stack and translates domain names. Based on the rules you have set, Apsara Stack DNS translates domain names that you have requested and direct requests from the client to the corresponding cloud services, business systems in enterprise internal networks, and services provided by Internet service providers.

Apsara Stack DNS provides basic domain name translation and scheduling services for VPC environments. You can perform the following operations through Apsara Stack DNS in your VPC:

- Access other ECS servers deployed in VPCs.
- Access cloud service instances provided by Apsara Stack.
- Access custom enterprise business systems.
- Access Internet services and business.
- Establish network connections between Apsara Stack DNS and user-created DNS through a leased line.

17.2 System architecture

The architecture of Apsara Stack DNS

- Deploys two physical servers for network connections and allows you to add more servers based on your needs.
- Uses two control interfaces for bond, which is uplinked to ASW. The gateway is the default gateway of the internal network.
- Two service interfaces are uplinked to LSW (ECMP is supported). These interfaces support the OSPF routing protocol to advertise Anycast VIP routes, and are connected to the Internet.
- Clusters can be deployed in one zone, multiple zones, or multiple regions.
- The control system is deployed in a container in the control area.

17.3 Features

Internal domain name management

Apsara Stack DNS provides data management for internal domain names. You can register, search, and delete internal domain names and add remarks. You can also add, delete, and modify

DNS records. Supported DNS record types include A, AAAA, CNAME, NS, MX, TXT, SRV, and PTR.

Internal domain name management can translate internal domain names for servers deployed in a VPC. The DNS server addresses are deployed based on anycast, which ensures the continuity of services if errors occur.

Domain name forwarding management

Apsara Stack DNS can forward a specific domain name to other DNS servers for translation.

The domain name forwarding feature includes two forwarding modes: forward all requests (with recursion) and forward all requests (without recursion).

- Forward all requests (without recursion): Uses the target DNS server to translate domain names. If the domain names cannot be translated, or the request is timed out, a message is returned to the DNS client indicating that the query fails.
- Forward all requests (with recursion): Uses the target DNS server to translate domain names. If the domain names cannot be translated, then uses the local DNS server to translate them.

Recursive query management

Apsara Stack DNS supports recursive queries, which enables your servers to access the Internet.

Option configuration

You can enable, modify, or disable global default forwarding for Apsara Stack DNS.

17.4 Benefits

Domain name management for enterprise domains

Apsara Stack DNS provides domain name management and translation services for enterprise domains.

- Apsara Stack DNS supports DNS resolution and reverse DNS resolution for domain names of cloud service instances, including ECS instance domain names.
- It also supports DNS resolution and reverse DNS resolution for your internal domain names.
- You can add, modify, and delete DNS records, including A, AAAA, CNAME, NS, MX, TXT, SRV, and PTR.
- You can add multiple DNS records, including A, AAAA, and PTR, for one host. By default, the resolution finds all matching records. The records can be randomly rotated to balance the load.

Flexible networking

Apsara Stack DNS provides the domain name forwarding service for enterprise domains, which allows you to flexibly create or combine networks.

- Supports forwarding all domain names.
- Supports forwarding specific domain names.

Access the Internet from your server

When the public network is accessible, Apsara Stack DNS supports recursive queries for public domain names and Internet domain names. This service allows your servers to access the Internet.

A unified management platform

The management system of Apsara Stack DNS is built on the unified management platform of Apsara Stack. You can use one account to manage all services. Apsara Stack DNS has the following benefits:

- Data management and service management support Web actions, which are easy to learn and operate.
- Apsara Stack DNS is deployed on clusters. You can add more clusters based on your needs.
- You can deploy Apsara Stack DNS in multiple zones. Apsara Stack DNS supports active-active deployment in the same city, disaster recovery deployment in the same city, and disaster recovery deployment across two cities.
- Apsara Stack DNS is deployed based on anycast. High availability and disaster recovery can be automatically enabled.

17.5 Advantages

As a key network service, Apsara Stack DNS controls traffic that goes through Apsara Stack, translates domain names, balances traffic for applications, and connects Apsara Stack to on-premises data centers. Apsara Stack provides you with multiple solutions for cloud environment deployment, zone high availability, server load balancing, and disaster recovery to support your IT business.

18 API Gateway

18.1 What is API Gateway

API gateway is a complete API hosting service. It helps you use APIs to provide capabilities, services, and data to your partners. You can also publish APIs to the API marketplace for other developers to purchase and use.

- API Gateway provides a range of mechanisms to enhance security and reduce risks arising from APIs. These mechanisms include attack prevention, replay prevention, request encryption , identity authentication, permission management, full-link signature verification, parameter cleaning, and request throttling.
- API Gateway provides a range of API lifecycle management functions to create, test, publish, and unpublish APIs. It also allows you to generate API documentation by importing Swagger files, improving API management and iteration efficiency.
- API Gateway allows end users to access the system by using different methods and protocols . It provides a simple function to generate SDKs in commonly used languages and supports most terminals available on the market.
- API Gateway can work with other Alibaba Cloud services to provide convenient O&M functions such as monitoring, alarms, and log analysis. These functions reduce API O&M costs.

API Gateway maximizes capability multiplexing. It allows enterprises to share capabilities and focus more on their core businesses, which benefits all parties involved.

18.2 System architecture

API Gateway consists of three components:

- API Gateway

  API Gateway is the system core that implements all service logics.

  It provides multi-protocol access for all clients, manages client connections, and throttles requests.
  It loads user-defined APIs to the memory and processes received client requests based on the API definitions.

- OpenAPI

  OpenAPI is a group of standardized management APIs provided to manage API definitions . OpenAPI manages groups, metadata, and authorization for APIs. When OpenAPI receives

an API change request, it synchronizes the change to all API Gateway services. System administrators can use APIs to manage the APIs running in API Gateway in real time.

System administrators can manage their own APIs in the API Gateway console. They can also call APIs in their management systems to manage their own APIs.

- Console

  The console can implement all API Gateway features for system administrators to manage their APIs.

  The API Gateway console calls APIs to provide Web-based operations.

18.3 Basic functions

18.3.1 API lifecycle management

It provides a range of lifecycle management functions to publish, test, and unpublish APIs.

It provides maintenance functions such as routine API management, API version management, and quick API rollback.

18.3.2 Multi-protocol access

Clients and API Gateway can communicate through the following protocols:

- HTTP: is the most popular Internet text protocol.
- HTTP2: is similar to HTTP but supports multiplexing and header compression for higher efficiency.
- WebSocket: uses keepalive connections for binary communications.

18.3.3 Application access control

API Gateway provides an application-based authentication mechanism. This mechanism ensures that only authorized clients can send requests to the backend service. Applications are used to call APIs. Each application has a key pair made up of an AppKey and an AppSecret. The AppKey is attached to the header of a request as a parameter, while the AppSecret is used to calculate the request signature.

The AppKey and AppSecret pair has all permissions on an application, and therefore must be kept confidential. If AppKey or AppSecret is leaked, you must reset it in the API Gateway console. You can own multiple applications, to which different APIs can be assigned based on your business requirements.

System administrators can manage applications in the API Gateway console. They can create, modify, delete, or query applications, manage keys, or view authorized applications.

18.3.4 Full-link signature verification mechanism

API Gateway provides a full-link signature verification mechanism for communication between the client and API Gateway or between API Gateway and the backend service. This mechanism prevents data tampering during request transmission. When a client calls an API, the client must convert the key request data into a signature string based on API Gateway signature algorithms. The client must attach the signature string to the request header. API Gateway performs symmetric calculation to parse the signature and verify the identity of the request sender. HTTP, HTTPS, and WebSocket requests must have a signature in their header.

18.3.5 Anti-replay mechanism

API Gateway provides an anti-replay mechanism to protect against data tampering used in replay attacks.

18.3.6 HTTPS communication based on the SSL certificate of the user

A system administrator can upload an SSL certificate corresponding to the domain name in the API Gateway console. Data transmitted between clients and API Gateway will then be encrypted based on the certificate. This prevents data tampering during transmission.

System administrators can update SSL certificates in real time in the API Gateway console.

18.3.7 Support for OpenID Connect

API Gateway supports OpenID Connect authentication, allowing API providers to verify requests based on their own user systems. OpenID Connect is a lightweight authentication standard based on OAuth 2.0. It provides a framework for identity interaction through APIs. Compared with OAuth, OpenID Connect not only authenticates a request, but also specifies the identity of the requester.

18.3.8 Bidirectional communication

API Gateway provides bidirectional communication capabilities. It maintains keepalive connections between itself and clients and automatically ensures that the client stays online. The backend service can access the API Gateway port to query the online status of clients or push in-application notifications.

API Gateway implements bidirectional communication capabilities based on the WebSocket protocol. Bidirectional communication is supported on Android, Objective-C, and Java SDKs .

18.3.9 Automatic generation of SDKs and API documentation

API Gateway can automatically generate Java, Objective-C, and Android SDKs for the APIs customized by providers. It can also generate API documentation.

18.3.10 Parameter cleaning

System administrators can define the data type, regular expression, and enumeration of all API parameters. API Gateway forwards API requests that match the API definition to the backend service, while rejecting the requests that do not match the definition. This ensures that the backend service only receives standard requests that match API definitions.

18.3.11 Mappings between frontend and backend parameters

API Gateway provides parameter mapping capabilities to relocate parameters within a request before sending the request to the backend service. For example, a parameter in a request sent to API Gateway is defined in Query. API Gateway can map the parameter to Form and then send the request to the backend service. This function ensures that users can access complicated backend functions by calling well-organized APIs.

18.3.12 Request throttling

System administrators can set a traffic threshold based on the maximum processing capabilities of the backend service. If the total number of requests exceeds the threshold, API Gateway directly rejects excess requests. This prevents the backend service from being overloaded. System administrators can set request throttling based on applications, APIs, or users.

18.3.13 Access control based on IP addresses

API Gateway provides access control based on IP addresses to enhance the security of APIs. This function controls the source IP addresses or address segments that can call APIs. System administrators can add an IP address to the whitelist or blacklist of an API to permit or deny API requests from that IP address.

18.3.14 Log analysis

API Gateway sends called logs to Log Service. System administrators can use Log Service to query or download logs, or perform statistical analysis in real time. Logs can also be sent to OSS or MaxCompute.

18.3.15 Publish an API in multiple environments

API Gateway allows you to publish an API group in three different environments: test, pre-release , and release environments. The test and pre-release environments are used by testers to test or debug APIs. The release environment is where the APIs can be used.

You can use the environment management function for API groups to set environment parameters for the test, pre-release, and online environments. The environment parameter is a common constant that can be customized for each environment. When you call an API, you can place the environment parameter in any location of the request. API Gateway identifies the environment based on the environment parameter in your request.

18.3.16 Online debugging

API Gateway provides the online API debugging function for system administrators and client developers.

18.3.17 Mock mode

A project is typically developed by multiple partners working together toward a specific goal. The interdependence among various stakeholders often restricts individual members during the process, and misunderstandings may affect the development process or even delay the project schedule. You can mock expected responses to be returned to API callers during the project development process. This can greatly reduce misunderstanding among partners and greatly improve the development efficiency.

API Gateway supports the Mock mode.

18.3.18 Swagger file import

Swagger is a widely-used specification to define and describe backend service APIs. You can create APIs in the API Gateway console by importing Swagger 2.0 files.

The API Gateway Swagger extension is based on Swagger 2.0. You can create the Swagger definition for API entities, and import the Swagger file to API Gateway for bulk creation or updating

of API entities. API Gateway supports Swagger 2.0 by default, which is compatible with most Swagger specifications.

18.4 Benefits

- **Easy maintenance**

  After you create APIs in API Gateway, API Gateway performs all the other API management functions, such as documentation maintenance and version management for APIs, and SDK maintenance. This significantly reduces routine maintenance costs.

- **High performance**

  API Gateway provides efficient client access capabilities based on HTTP2 or WebSocket keepalive connections.

  API Gateway uses a distributed deployment and automatic scaling model to respond to a large number of API access requests at very low latencies. It provides highly secure and efficient gateway functions for your backend services.

- **Stability**

  API Gateway has provided services for Alibaba Cloud users for over two years, and has a proven record of performance. API Gateway can maintain stable operation even in special cases where oversized packets are received, or the backend service is unstable and does not respond in a timely manner.

- **Security**

  API Gateway uses full-link SSL communication to prevent eavesdropping during transmission.

  API Gateway uses full-link signature verification to prevent data tampering during transmission.

  API Gateway provides effective authorization management, replay prevention, parameter cleaning, access control based on IP addresses, and request throttling functions. These functions help you ensure security, stability, and controllability of your services.