Alibaba Cloud
Elastic Compute Service

Quick Start for Enterprise-Level Users

Issue: 20190614

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 content of the Alibaba Cloud website, including but not limited to works, products, images, archives, information, materials, website architecture, website graphic layout, and webpage design, 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 the Alibaba Cloud website, product programs, or content 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.	 Notice: Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	 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]*

Style	Description	Example
{} or {a|b}	It indicates that it is a required value, and only one item can be selected.	`swich {stand | slave}`

Contents

1 Overview

Quick start process

When purchasing and using Elastic Compute Service (ECS) instances, as an enterprise -level user, you must finish the following operations:

- *Select configurations*.
- *Estimate costs*.
- *Plan networks*.
- *Configure a security group*.
- *Automatic snapshot policies*.
- *Image migration*.
- *Implement high availability by using Server Load Balancer*.

Target readers

The Quick Start is a reference for anyone who wants to know how to:

- Select configurations for ECS instances.
- Estimate costs of a large-sized instance and specific configuration.
- Perform network planning for a specific solution.
- Configure the security group information for each instance.
- Select and develop better snapshot policies.
- Complete image migration.

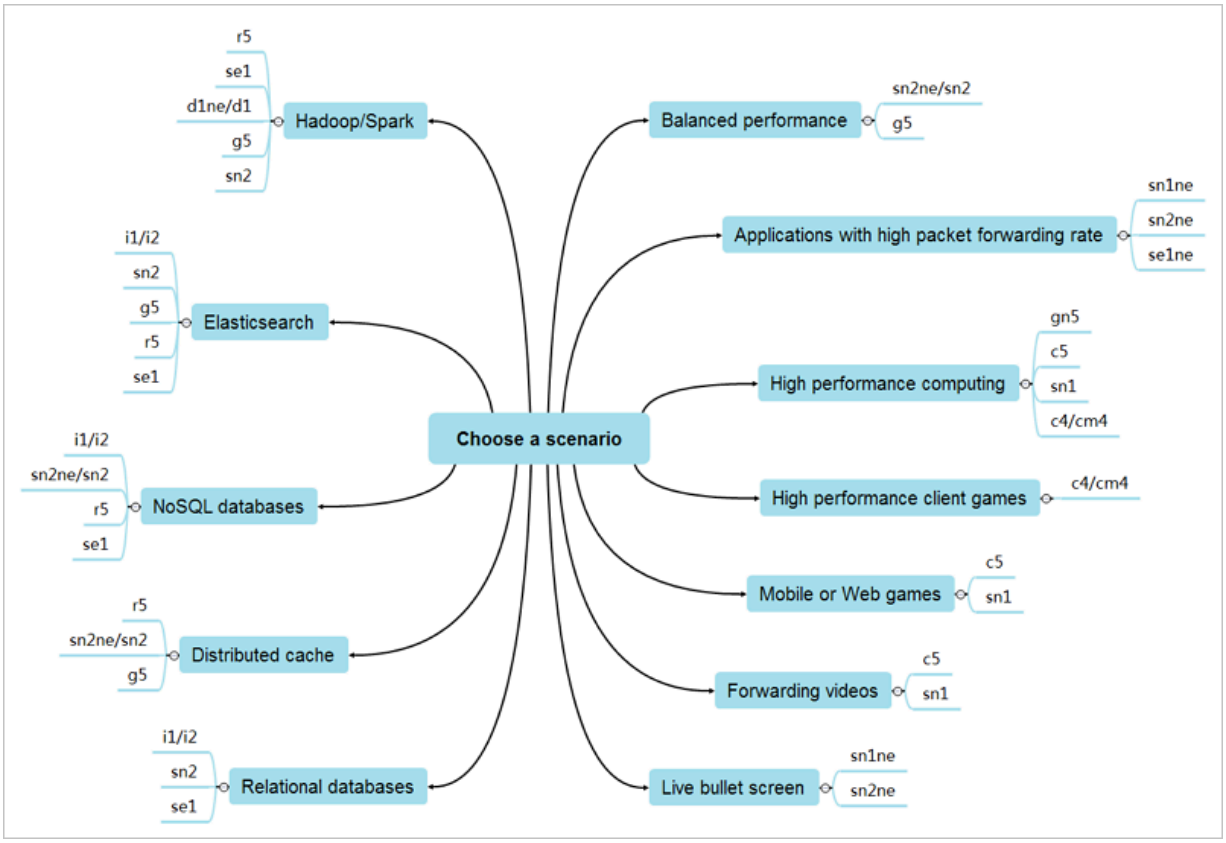
Locate resources

Procedure	Interface	Parameter	Target data
1. Query regions.	*DescribeRegions*	N/A	Region ID (RegionId)
2. Query zones.	*DescribeZones*	Region ID	Zone ID (ZoneId)
3. Determine the billing method.	*DescribeZones*	Billing method/ Bidding policy	*ZoneType*
4. Query resource combinations.	*DescribeZones*	RegionId/Billing method	*ZoneType*

This document is applicable only to operations performed in the console. If you are an API user, see *API references*.

2 Select configuration

An enterprise-level user can select configurations by following the process.



Note:

For more information about the instance types, see *instance generations and type families*.

As an enterprise-lever user of ECS, you may have specific requirements. To meet your requirements, Alibaba Cloud provides you with recommendations for instance configuration in the following scenarios:

- Balanced performance

  A balanced CPU to memory ratio is required to meet the application resource requirements in most scenarios.

- Applications with high packet forwarding rate

  High packet forwarding rate is required. You can select a more reasonable computing capacity to memory resource ratio based on the specific scenarios.

- **High performance computing**

  Many computing resources are required. GPU parallel computing and a high clock speed are typical applications in this scenario.
- **High-performance client games**

  Your services require a high-frequency processor to carry more users. Therefore, a high clock speed is required in this scenario.
- **Mobile and Web games**

  Many computing resources are required for this scenario. A CPU to memory ratio of 1:2 can help achieve optimal cost performance of computing resources.
- **Video forwarding**

  Many computing resources are required for this scenario. A CPU to memory ratio of 1:2 can help achieve optimal cost performance of computing resources.
- **Live bullet screen**

  High packet forwarding rate is required for this scenario. You can select a more reasonable computing capacity to memory resource ratio based on the specific scenarios.
- **Relational databases**

  In this scenario, SSD cloud disks or higher-performance NVMe SSD local disks are required to provide higher IOPS capability and a low read latency. The CPU to memory ratio is balanced (1:4) or the memory proportion is larger (1:8).
- **Distributed cache**

  In this scenario, a balanced CPU to memory ratio (1:4) or a higher memory proportion (1:8), and stable computing performance are required.
- **NoSQL databases**

  In this scenario, SSD cloud disks or higher-performance NVMe SSD local disks are required to provide higher IOPS capacity and a low read latency. The CPU to memory ratio is balanced (1:4) or the memory proportion is larger (1:8).
- **Elastic search**

  In this scenario, SSD cloud disks or higher-performance NVMe SSD local disks are required to provide higher IOPS capacity and a low read latency. The CPU to memory ratio is balanced (1:4) or the memory proportion is larger (1:8).

- Hadoop

  Data nodes require a high disk throughput, high network throughput, and balanced CPU to memory ratio. Computing nodes focus more on the computing performance, network bandwidth, and CPU to memory ratio.

- Spark

  Data nodes require a high disk throughput, high network throughput, and balanced CPU to memory ratio. Computing nodes focus more on the computing performance, network bandwidth, and CPU to memory ratio.

- Kafka

  Data nodes require a high disk throughput, high network throughput, and balanced CPU to memory ratio. Computing nodes focus more on the computing performance, network bandwidth, and CPU to memory ratio.

- Machine learning

  In this scenario, a high performance Nvidia GPU computing processor is required, and the memory size must be at least twice the video memory.

- Video encoding

  In this scenario, a high performance GPU computing processor or a high performance CPU is required for encoding and decoding.

- Rendering

  In this scenario, a high performance GPU computing processor is required for rendering.

3 Plan networks

We recommend Virtual Private Clouds (VPCs), which are logically isolated from each other. VPCs have the following characteristics:

- Isolated network environment
- Controllable network configurations

Determine the number of VPCs

- We recommend that you use multiple VPCs if any of the following requirements is involved:
  - The system is deployed in multiple regions.

    VPCs are region-specific resources, which cannot be deployed across regions. To deploy a system across regions, you must use multiple VPCs. The *Express Connect* product built on the Alibaba backbone network can easily achieve communication between VPCs across regions or countries.
  - Multiple business systems are isolated from each other.

    To strictly isolate multiple business systems within the same region by using VPCs, such as strict isolation between the production environment and the test environment, you must use multiple VPCs.
- Use a single VPC.

  A single VPC is recommended if you do not have to deploy a system in multiple regions or isolate systems from each other by using VPCs. Now, up to 15,000 cloud product instances can run in a single VPC. Such a capacity can meet basic requirements.

Determine the number of VSwitches

We recommend that you use at least two VSwitches even if you use only one VPC. In addition, deploy the two VSwitches in two different zones to achieve cross-zone disaster tolerance.

The network communication latency between different zones in the same region is low, but the time is still needed for business system adaptation and verification. A latency higher than expected may be introduced due to complicated system calls,

system processing, and cross-zone calls. We recommend that you optimize and adapt the system to balance between high availability and low latency.

The number of VSwitches used is related to the system size and planning. If the front-end systems can access Internet or can be accessed by Internet, you can deploy different front-end systems under different VSwitches to achieve disaster tolerance, and deploy backend systems under other switches.

Select network blocks

You can use the standard private CIDR blocks listed in the following table and their subnets as the private IP address ranges of VPCs.

CIDR blocks	Number of available IP addresses	Remarks
192.168.0.0/16	65532	Addresses occupied by systems are excluded.
172.16.0.0/12	1048572	Addresses occupied by systems are excluded.
10.0.0.0/8	16777212	Addresses occupied by systems are excluded.

Note:

- After a VPC is created, you cannot modify its CIDR blocks.
- If you have requirements for other special CIDR blocks, *open a ticket* or contact your customer manager to activate the CIDR blocks.
- If multiple VPCs exist, or you have to build a hybrid cloud composed of VPCs and offline IDCs, we recommend that you use subnets of the preceding standard CIDR blocks and that the netmask length does not exceed 16.
- If only one VPC is on the cloud and the VPC does not have to communicate with offline IDCs, you can select any of the preceding CIDR blocks or their subnets.
- You must also consider whether the class network is used. If you are using a classic a classic network on the cloud and plan to connect ECS instances of the classic network to a VPC by using the *ClassicLink* feature, we recommend that you use a CIDR block other than 10.0.0.0/8 for the VPC, because the CIDR block of a classic network is 10.0.0.0/8.

For more information, see *plan and design VPC*.

4 Estimate costs

As an enterprise-level user, once you select the instance type and have completed the network planning, you can estimate the costs based on the following factors: billing methods, regions, images, networks, and quantity.

Billing methods

Now, the following billing methods are available to an enterprise-level user, including *Subscription, Pay-As-You-Go*, and *preemptible instances*.

- Subscription: Subscription is a type of prepayment whereby instances can be used only after you make the payment. Instances are charged on a monthly basis, and the unit is USD/month. This billing method is applicable to fixed 24/7 services, such as the Web service.
- Pay-As-You-Go: Pay-As-You-Go is a type of post payment whereby payment can be made after you use the resources. Instances are billed by second and settled by hour. The unit is USD/hour . This billing method is applicable to scenarios with traffic volume spikes, such as temporary scaling, interim testing, and scientific computing.
- Preemptible instance: To reduce the ECS computing cost, you can select a preemptible instance. As an instance on demand, you must set the maximum hourly rate you want to pay for your expected instance type. When your bid is higher than the current market price, your instance runs. The final price you pay for the instance type is based on the current market price. For more information, see *preemptible instances*.

Advantages of each billing method vary according to the instance type. For more information, see *ECS product pricing*.

Regions

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

- Region of the instance and the geographical location of your target users
- Relationship between ECS and other Alibaba Cloud products
- Resource price
- Special requirements of some regions. For example, to use an ECS instance as a Web server in mainland China, your business license must be filed for record.

> **Note:**
> The price for the same instance type may vary according to the region. For more information about the specific prices, see *ECS product pricing*.

Images

The price varies according to the image type:

- Public images: The price varies with the region.
- Shared images: The price of the custom image shared with you by other accounts is based on the source image.
- Custom images: The price is based on the source image.

Networks

After completing network planning of the general solution, you must select a specified VPC and switch.

> **Note:**
> Before purchasing an instance, you must create a VPC and a switch in advance.

When selecting a specific network, the selected VPC and switch are free of charge.

- If you want to allocate a public IP address, you can set the peak bandwidth . Different fees are charged based on the selected bandwidth. If the selected public bandwidth is 0 Mbit/s, no public IP address is allocated by default.

  > **Note:**
  > The allocated public IP addresses cannot be unbound from their ECS instances.

- If you choose not to allocate public IP addresses and you need a static public IP address solution with greater flexibility, we recommend that you do not allocate a public IP address. Then, configure and bind an EIP address. For more information about the fees if you choose to configure an EIP, see EIP billing.

Quantity

If you purchase more ECS instances, the cost rises.

5 Configure a security group

A security group is a virtual firewall that is used to control the ECS outbound and inbound traffic.

Within the same VPC, ECS instances in the same security group can communicate with each other over the intranet. By default, ECS instances under different VSwitches in a VPC can access each other by using system routes. You can configure the security group rules to isolate the instances from one another. For more information, see *isolate the subnets in a VPC*.

Default security group rules

When you create a VPC ECS instance, you can add the instance into a default security group provided by the system, or select other existing security groups in the VPC. For more information about the rules of a default security group, see *default security group rules*.

Configure security groups

After you have the network plan, you can configure the security groups. For more information about how to add a security group, see *Add security group rules*.

For more information about the scenarios of security groups, see *Scenarios*.

6 Automatic snapshot policies

Snapshots are the state of system data at a certain time point, and are used for data backup or image creation. You can create automatic snapshot policies to automatically take snapshots of disks. Alternatively, you can manually take snapshots of disks. The creation time of automatic snapshot is determined by the automatic snapshot policies, whereas manually created snapshots are irrelevant to the automatic snapshot policies.

Parameter settings

One account can create up to 100 automatic snapshot policies in one region. For more information about how to create automatic snapshot policies, see *create and delete an automatic snapshot policy*. Parameters for creating automatic snapshot policies are described as follows:

- `Policy name` : Name of an automatic snapshot policy, which is a string of 2–128 characters. The name must start with an uppercase or lowercase letter or a Chinese character, and can contain numbers and special characters, including periods (.), underscores (_), and hyphens (-).
- `Creation time` : You can create a snapshot at any of the 24 time points each day. The value ranges from 00:00 to 23:00.
- `Repeated day` : You can set up to seven days of repetition day each week. The value ranges from Monday to Sunday.
- `Retention time` : Number of days snapshots can be retained. The value can be 1–65,536 days or permanently, and the default value is 30 days.

Snapshot creation time and repeated day

If you are an enterprise-level user, when setting the snapshot creation time and repeated day, avoid traffic peak periods based on characteristics of your own industry to avoid affecting normal service running.

Note:

Creating snapshots may slightly degrade the disk performance and the disks may slow down temporarily.

Snapshot retention time

You can set the snapshot retention time appropriately based on characteristics of your own industry and the data update cycle. The snapshot retention time is related to the number of snapshots. After the snapshot quota is reached, the earliest automatic snapshots are automatically deleted.

 Note:
Manually created snapshots are retained unless you delete them. If they are no longer needed, manually delete them.

Costs

Now, snapshots are free of charge.

7 Image migration

Feasibility

Before migrating an image, you must conduct a survey on the server to be migrated and evaluate whether image migration is used and whether this operation is feasible.

- If the number of servers to be migrated is large, most servers contain system disks , and network conditions are unfavorable, image migration is not recommended. Image files are large, and image migration performed under such conditions may increase the migration time and labor costs.
- If the application configurations on the servers to be migrated are complicated and not manually maintained, and network conditions are favorable, image migration is recommended. Although data disks do not support image migration, you can migrate images of system disks to Alibaba Cloud, and then synchronize data of data disks to data disks of Alibaba Cloud by means of file synchronization.

Migration tools

Alibaba Cloud Migration Tool, or Cloud Migration Tool for short, is a proprietary resource migration tool of Alibaba Cloud.

It supports resource migration, such as the operating system, applications, and application data in a computer disk, of physical machines, VMs (virtual machines), or cloud hosts to the image list in the ECS. Being lightweight and handy, Alibaba Cloud Migration Tool helps you balance the workload between your local and cloud hosts, or cloud hosts from different cloud platforms. For more information, see *what is Alibaba Cloud Migration Tool*.

Procedure

To migrate an image, follow these steps:

1. Prepare for the migration.
   - Prepare for the migration.
   - Prepare the image format conversion tools and platforms.
   - Check and prepare the operating system before exporting image files.
2. Migrate your on-premises server. For more information, see *migrate to Alibaba Cloud by using Cloud Migration Tool*.

3. Start ECS instances based on images. You can go to the *ECS console* page, and create instances based on the new image files.

8 Implement high availability by using Server Load Balancer

Server Load Balancer is a traffic distribution control service that distributes traffic to multiple backend ECS instances based on the forwarding policies.

Scenarios

In the following scenarios, you can use Server Load Balancer to greatly improve high availability of ECS instances.

- High-traffic services

  If traffic of an application is high, you can configure monitoring rules to distribute traffic to different ECS instances. In addition, you can use the session persistence function to forward requests from the same client to the same backend ECS instance to improve access efficiency.

- System extension

  Based on the business development requirements, you can add or remove ECS instances at any time to extend the service capability of application systems and adapt to various Web servers and application servers. For more information, see *Backend server overview*.

- Eliminate a single-point of failures

  You can add multiple ECS instances behind a Server Load Balancer instance. When some ECS instances are faulty, Server Load Balancer automatically shields faulty ECS instances and distributes requests to healthy ECS instances, guaranteeing that the application systems can still work normally.

- Intra-city disaster tolerance (multi-zone disaster tolerance)

  You can deploy the Server Load Balancer instances in a region with multiple zones to implement intra-city disaster tolerance. With this feature, when one data center fails, Server Load Balancer can quickly switch front-end traffic to other zones in the same region to restore the service capability.

  > Note:
  > If at least one ECS instance is added for each zone, the efficiency of Server Load Balancer is the highest in this deployment mode.

- Cross-region disaster tolerance

  You can deploy the Server Load Balancer instances in different regions, and attach ECS instances of different zones in the corresponding regions. The upper layer uses Alibaba Cloud DNS as intelligent DNS, and resolves domain names to service addresses of Server Load Balancer instances in different regions, thus implementi ng global load balancing. When the system becomes unavailable in a region, you can temporarily stop DNS so that access of all users is not affected.

Preparations

Before using Server Load Balancer, make the following preparations.

- Plan regions where the Server Load Balancer instances are deployed

  Alibaba Cloud provides Server Load Balancer for all the regions.

  > Note:
  >
  > - To reduce the latency and increase the download speed, we recommend that you select the region closest to your customers.
  > - As Server Load Balancer does not support cross-region deployment, plan regions and select the same region as the backend ECS instance.

- Select the type (public or private network) of the Server Load Balancer instance

  Determine the Server Load Balancer instance type based on your service type. After a Server Load Balancer instance is created, the system allocates a public or private IP address based on the type of the Server Load Balancer instance.

  - Public network Server Load Balancer instances are only assigned public IP addresses. These Server Load Balancers are accessible over the Internet.
  - Private network Server Load Balancer instances are only assigned private IP addresses. Server Load Balancer is accessible only over the Alibaba Cloud intranet, but not over the Internet. No fee is charged for private network Server Load Balancer instances.

- Select listening protocols

  Server Load Balancer supports Layer-4 (TCP and UDP) and Layer-7 (HTTP and HTTPS) listening.

  - During Layer-4 listening, requests are directly forwarded to the backend server and headers are not modified. After requests of clients arrive at the Server Load

Balancer listener, the Server Load Balancer servers establish TCP connections with backend ECS instances through the backend ports configured for listening.

- In principle, Layer-7 listening is a way of implementing reverse proxy. After requests of clients arrive at the Server Load Balancer listener, the Server Load Balancer servers establish TCP connections with the backend ECS instances, namely to access the HTTP backend servers through the new TCP connections again instead of directly forwarding packets to the backend ECS instances.

Compared with Layer-4 listening, Layer-7 listening requires an extra step of Tengine processing. Therefore, the performance of Layer-7 listening is not so good as that of Layer-4 listening. In addition, Layer-7 listening performance may be deteriorated by such factors as insufficient number of client ports and too many backend server connections. If high performance requirements are raised, Layer-4 listening is recommended.

· Prepare backend servers

Add ECS instances behind a Server Load Balancer instance to process requests forwarded by the front-end listener. Before creating a Server Load Balancer instance, create ECS instances and deploy related applications. When creating an ECS instance, note the following items:

- Region and zone of the ECS instance

  Make sure that the region of the ECS instance is the same as that of the Server Load Balancer instance. In addition, we recommend that you deploy ECS instances in different zones to improve the local availability.

- ECS configuration

  After deploying applications on the ECS instances, you do not have to perform special configuration. owever, to configure a Layer-4 listener (TCP or UDP) for Linux ECS instances, make sure that values of the following parameters in the net.ipv4.conf file are set 0s:

```
net . ipv4 . conf . default . rp_filter  =  0
net . ipv4 . conf . all . rp_filter  =  0
net . ipv4 . conf . eth0 . rp_filter  =  0
```

- ECS instance deployment

  Now, no quota is set for backend ECS instances that can be configured for a Server Load Balancer instance. However, to guarantee stability and efficiency of your external services, we recommend that you add ECS instances of applicatio

n servers that provide different services or perform different tasks to different Server Load Balancer instances based on service types or application modules.

Procedure

1. Create ECS instances. You must have at least two ECS instances for the Server Load Balancer service. For more information, see *Create an instance by using the wizard*.
2. After the ECS instances are created, deploy related applications on the ECS instances.
3. Create a Server Load Balancer instance. A Server Load Balancer instance can be mapped to multiple listeners and backend servers. For more information, see *Create an SLB instance*.
4. After the Server Load Balancer instance is created, add at least one listener and one group of backend servers behind it. For more information, see *Configure an SLB instance*.

Billing description

Server Load Balancer supports PayByTraffic. The specific charging items vary according to the instance type and performance type. For more information, see *Pricing*.