

Alibaba Cloud Server Load Balancer

Archives

Issue: 20190402

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 <code>cd / d C :/ windows</code> command to enter the Windows system folder.
<i>Italics</i>	It is used for parameters and variables.	<code>bae log list --instanceid <i>Instance_ID</i></code>
[] or [a b]	It indicates that it is an optional value, and only one item can be selected.	<code>ipconfig [-all -t]</code>

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

Contents

Legal disclaimer.....	I
Generic conventions.....	I
1 User Guide (Old Console).....	1
1.1 SLB instances.....	1
1.1.1 Guaranteed-performance instances.....	1
1.1.2 Create an SLB instance.....	8
1.2 Listener.....	10
1.2.1 Layer-4 listeners.....	10
1.2.2 Layer-7 listeners.....	10
1.2.3 Health check.....	10
1.3 Backend servers.....	10
1.3.1 Add default servers.....	10
1.4 Certificate management.....	11
1.4.1 Convert certificate formats.....	11

1 User Guide (Old Console)

1.1 SLB instances

1.1.1 Guaranteed-performance instances

Alibaba Cloud launched the guaranteed-performance instances in May 2017, and charged the capacity fee on guaranteed-performance instances on April 1, 2018.

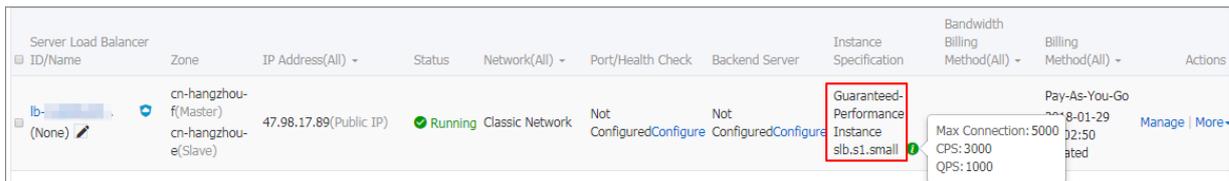
- [1. What are guaranteed-performance instances?](#)
- [2. How are guaranteed-performance instances billed?](#)
- [3. What is the price of each capacity?](#)
- [4. How to select a guaranteed-performance instance?](#)
- [5. Can I modify the capacity after the instance is created?](#)
- [6. When will the guaranteed-performance instances be charged?](#)
- [7. After Alibaba Cloud starts to charge capacity fee on guaranteed-performance instances, will extra fees be charged on shared-performance instances?](#)
- [8. Why sometimes guaranteed-performance instances cannot reach the performance limit as defined in the capacity?](#)
- [9. Can I still buy shared-performance instances?](#)
- [10. Will intranet SLB instances be charged for capacity fee?](#)

1. What are guaranteed-performance instances?

A guaranteed-performance instance provides guaranteed performance metrics (performance SLA). It is opposite to a shared-performance instance. For a shared-performance instance, the performance metrics are not guaranteed and the resources are shared by all instances.

All instances are shared-performance instances before Alibaba launches guaranteed-performance instances. You can view the instance type on the console.

Hover your mouse pointer to the green icon of the target instance to view the performance metrics, as shown in the following figure.



The following are three key performance metrics for guaranteed-performance instances:

- **Max Connection**

The maximum number of connections to a SLB instance. When the maximum number of connections reaches the limits of the capacity, the new connection will be dropped.

- **Connection Per Second (CPS)**

The rate at which a new connection is established per second. When the CPS reaches the limits of the capacity, the new connection will be dropped.

- **Query Per Second (QPS)**

The number of HTTP/HTTPS requests that can be processed per second. When the QPS reaches the limits of the capacity, the new connection will be dropped.

Alibaba Cloud Server Load Balancer provides the following capacities for guaranteed-performance instances:

Type	Type	Max Connection	CPS	QPS
Capacity 1	Small I (slb.s1.small)	5,000	3,000	1,000
Capacity 2	Standard I (slb.s2.small)	50,000	5,000	5,000
Capacity 3	Standard II (slb.s2.medium)	100,000	10,000	10,000
Capacity 4	Higher I (slb.s3.small)	200,000	20,000	20,000
Capacity 5	Higher II (slb.s3.medium)	500,000	50,000	30,000
Capacity 6	Super I (slb.s3.large)	1,000,000	100,000	50,000

If you want to use a larger capacity, contact your customer manager.

2. How are guaranteed-performance instances billed?

Guaranteed-performance instances are billed as follows:

Total fee (per instance) = instance fee + traffic fee + capacity fee



Note:

For intranet SLB instances, you can also choose to use guaranteed-performance instances. If guaranteed-performance instances are selected, capacity fee is collected and billed as the Internet SLB instance, but no traffic fee and instance fee are collected.

The performance guarantee instance specification fee is charged by usage, no matter what kind of specification you choose, the instance specification fee will be charged according to the specifications you actually use.

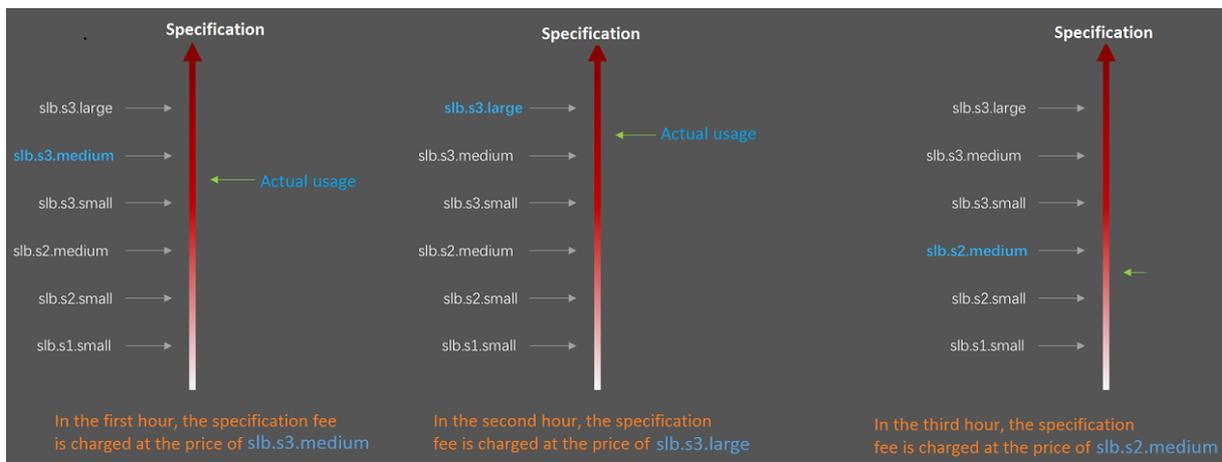
For example, if you purchase the slb.s3.large capacity (1,000,000; CPS 500,000; QPS 50,000) and the actual usage of your instance in an hour is as follow:

Max Connection	CPS	QPS
90,000	4,000	11,000

- From the perspective of Max Connection, the actual metrics 90,000 occurs between the limit 50,000 defined in the Standard I (slb.s2.small) capacity and the limit 100,000 defined in the Standard II (slb.s2.medium) capacity. Therefore, the capacity of the Max Connection metrics in this hour is Standard II (slb.s2.medium).
- From the perspective of CPS, the actual metrics 4,000 occurs between the limit 3,000 defined in the Small I (slb.s1.small) capacity and the limit 5,000 defined in the Standard I (slb.s2.small) capacity. Therefore, the capacity of the CPS metrics in this hour is Standard I (slb.s2.small).
- From the perspective of QPS, the actual metrics 11,000 occurs between the limit 10,000 defined in the Standard II (slb.s2.medium) capacity and the limit 20,000 defined in the Higher I (slb.s3.small) capacity. Therefore, the capacity of the QPS metrics in this hour is Higher I (slb.s3.small)

Comparing these three metrics, the capacity of the QPS metrics is highest, therefore, the capacity fee of the instance in this hour is charged at the price of the Higher I (slb.s3.small) capacity.

The following figure is an example showing how the capacity fee is billed for an SLB instance in the first three hours:



The billing of the guaranteed-performance instances is flexible. The capacity you select when purchasing an instance is the performance limitation of the instance. For example, if slb.s3.medium is selected, the new connections are dropped when the HTTP requests in one second reach 30,000.

3. What is the price of each capacity?

The following table lists the capacity price of each capacity. In addition to the capacity fee, you are also charged for instance fee and traffic fee. For more information, see [Pay-As-You-Go](#).

Region	Type	Max Connectio	CPS	QPS	Capacity fee (USD/ Hour)
China (Hangzhou) China (Zhangjiakou) China (Hohhot) China (Qingdao) China (Beijing) China (Shanghai) China (Shenzhen)	Small I (slb.s1. small)	5,000	3,000	1,000	Free
	Standard I (slb.s2. small)	50,000	5,000	5,000	0.05
	Standard II (slb.s2. medium)	100,000	10,000	10,000	0.10
	Higher I (slb.s3. small)	200,000	20,000	20,000	0.20
	Higher II (slb.s3. medium)	500,000	50,000	30,000	0.31
	Super I (slb.s3.large)	1,000,000	100,000	50,000	0.51

Region	Type	Max Connectio	CPS	QPS	Capacity fee (USD/ Hour)
Singapore Malaysia (Kuala Lumpur) Indonesia (Jakarta) India (Mumbai) US (Silicon Valley) US (Virginia) China (Hong Kong)	Small I (slb.s1.small)	5,000	3,000	1,000	Free
	Standard I (slb.s2.small)	50,000	5,000	5,000	0.06
	Standard II (slb.s2.medium)	100,000	10,000	10,000	0.12
	Higher I (slb.s3.small)	200,000	20,000	20,000	0.24
	Higher II (slb.s3.medium)	500,000	50,000	30,000	0.37
	Super I (slb.s3.large)	1,000,000	100,000	50,000	0.61

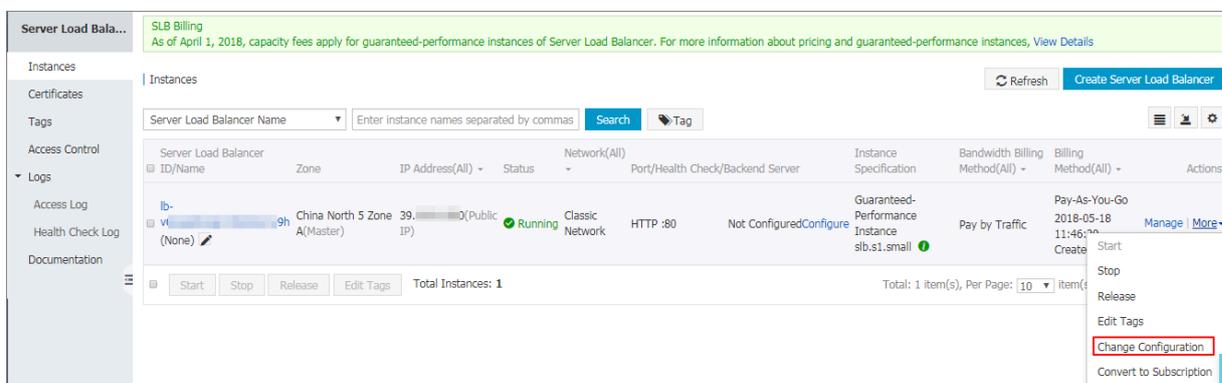
Capacity fees of guaranteed-performance instances in the international regions enjoy an 83% discount.

4. How to select a guaranteed-performance instance?

Because the capacity fee is billed based on the actual usage, we recommend that you select the largest capacity (slb.s3.large). This guarantees the business flexibility (flexibility) and will not cause extra costs. If your traffic does not reach the largest capacity, you can select a more reasonable capacity, such as slb.s3.medium.

5. Can I modify the capacity after the instance is created?

Yes. You can change the capacity at any time and the change takes effect immediately.



Configuration upgrade

Network and instance type

Instance type
Internet

Instance Spec
Small I (slb.s1.small) ▾

Max connection: 5000, CPS: 3000, QPS: 1000

Bandwidth
By traffic

Note:

- After you change a shared-performance instance to a guaranteed-performance instance, you cannot change it back.
- Some SLB servers are deployed in old clusters. If you change a shared-performance instance to a guaranteed-performance instance, a brief disconnection of service may occur for 10 to 30 seconds. We recommend that you change the specification when the business is not busy.
- The IP of the SLB instance will not be changed after you changing the instance type or the capacity.

Caution

When you change the configuration of an SLB instance or change a shared-performance instance to a guaranteed-performance instance, a brief disconnection of service may occur for 10 to 30 seconds. We recommend that you perform this operation when the service is not busy or after the service migrates to another SLB instance by using [Global Server Load Balancer](#). (Changes made to the billing method and network bandwidth of the SLB instance will not affect the service.)

I Agree
No, Not Now

6. When will the guaranteed-performance instances be charged?

Alibaba Cloud launched the guaranteed-performance instances in May 2017, and charged the capacity fee on guaranteed-performance instances on April 1, 2018.

The capacity fee takes effect in batches as follows:

- **The first batch:**

Time: From April 1st to April 10th

Regions: Singapore, Malaysia (Kuala Lumpur), Indonesia (Jakarta), India (Mumbai), US (Silicon Valley), US (Virginia)

- **The second batch:**

Time: From April 11th to April 20th

Regions: China (Hangzhou), China (Zhangjiakou), China (Hohhot), China (Hong Kong)

- **The third batch:**

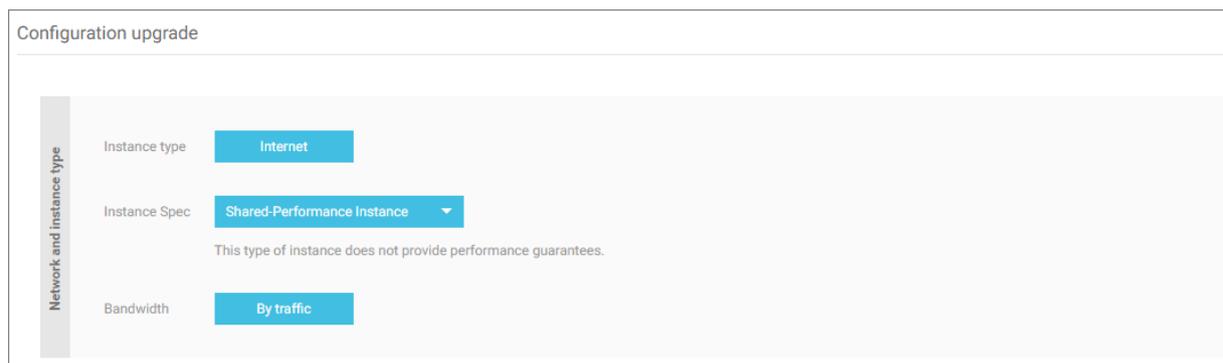
Time: From April 21th to April 30th

Regions: China (Qingdao), China (Beijing), China (Shanghai), China (Shenzhen)

7. After Alibaba Cloud starts to charge capacity fee on guaranteed-performance instances, will extra fees be charged on shared-performance instances?

No.

The billing of the original shared-performance instances is the same if you do not change it to a performance-guaranteed instance. You can change a shared-performance instance to a guaranteed-performance instance. After changing to the guaranteed-performance instance, capacity fees are collected.



8. Why sometimes guaranteed-performance instances cannot reach the performance limit as defined in the capacity?

The cask theory.

Guaranteed-performance instances do not guarantee that the three metrics can reach the capacity limits at the same time. The limitation is triggered as long as a metric first reaches the limitation defined in the capacity.

For example, you have purchased a guaranteed-performance instance of the Higher I (slb.s3.small) capacity. When the QPS of the instance reaches 20,000 but the number of maximum connections does not reach 200,000, the new connections are still dropped because the QPS has reached the limitation.

9. Can I still buy shared-performance instances?

Yes.

However, shared-performance instances will be phased out in the future. Please pay attention to the official announcement.

10. Will intranet SLB instances be charged for capacity fee?

If the intranet SLB instance is a shared-performance instance, no capacity fee is charged. If the intranet SLB instance is a guaranteed-performance instance, corresponding capacity fee is charged, and no other fees are charged.

1.1.2 Create an SLB instance

Prerequisites

Before creating an SLB instance, make sure that you have properly prepared the environment. For more information, see [Plan and prepare](#).

Procedure

1. Log on to the [SLB console](#).
2. On the Instances page, click Create Server Load Balance.
3. Configure the SLB instance according to the following information.

Configurat ion	Description
Region	Select the region where the SLB instance is located.  Note: Make sure that the region of the SLB instance is the same as that of backend ECS instances.

Configurat ion	Description
Zone type	<p>Display the zone type of the selected region. The zone of a cloud product refers to a set of independent infrastructure and is usually represented by Internet data centers (IDCs). Different zones have independent infrastructure (network, power supply, air-conditioning and so on). Therefore, an infrastructure fault in one zone will not affect other zones. A zone belongs to a specific region, however, a single region may have one or more zones. SLB has deployed multi-zone in most regions.</p> <ul style="list-style-type: none"> · Single zone: The SLB instance is deployed only in one zone. · Multi-zone: The SLB instance is deployed in two zones. By default, the instance in the primary zone is used to distribute traffic. If the primary zone is faulty, the instance in the backup zone will automatically take over the load balancing service.
Primary zone	Select the primary zone for the SLB instance. The primary zone carries traffic in normal conditions.
Backup zone	Select the backup zone for the SLB instance. The backup zone only takes over traffic when the primary zone is unavailable.
Instance spec	Select a performance capacity for the instance. The performance metrics varies based on different capacities. For more information, see Guaranteed-performance instances .
Instance Type	<p>Select the instance type based on your business needs. A public or a private IP address is allocated to the SLB instance based on the instance type.</p> <ul style="list-style-type: none"> · Internet: Internet SLB instances distribute requests from Internet clients to backend ECS instances according to your listening rules. · Intranet: Intranet SLB instances only distribute requests from clients that have access to the private network of the SLB instance.
Network type	<p>If the selected instance type is Intranet, you have to select a network type for the instance:</p> <ul style="list-style-type: none"> · Classic network: The IP of the instance is allocated and managed by Alibaba Cloud in a unified manner. · VPC: The IP of the instance is allocated from your specified VSwitch CIDR block.
Bandwidth	Display the billing method.

Configurat ion	Description
Quantity	Select the number of instances to create.

4. Click **Buy Now** and complete the payment.

1.2 Listener

1.2.1 Layer-4 listeners

1.2.2 Layer-7 listeners

1.2.3 Health check

1.3 Backend servers

1.3.1 Add default servers

Before using the SLB service, you must add at least one default server.

Prerequisites

- You have [created an SLB instance](#).
- You have created ECS instances and deploy applications to process distributed requests.

Procedure

1. Log on to the [SLB console](#).
2. On the Server Load Balancer page, select a region.
3. Click the ID of the target SLB instance.
4. Click the Default Server Group tab.
5. Click Add.
6. On the Servers Not Added page, click Add.
Then the Available Servers page is displayed.
7. Click Add next to the target ECS instance, or select multiple ECS instances and then click Add to Selected Server List.

8. In the displayed Available Servers dialog box, specify the weight of the added ECS instance and then click OK.

An ECS instance with a higher weight will receive a larger number of connection requests. You can set the weight based on the service capabilities of the ECS instances.

**Notice:**

If the weight is set to 0, no requests will be sent to the ECS instance.

The added ECS instances are listed on the Default Server Group page. You can remove or change the weights of the added ECS instances.

1.4 Certificate management

1.4.1 Convert certificate formats

Server Load Balancer supports PEM certificates only. Certificates in other formats must be converted to PEM before they can be uploaded to Server Load Balancer. We recommend that you use Open SSL for conversion.

Convert DER to PEM

DER: This format is usually used on a Java platform.

- Run the following command to convert the certificate format.

```
openssl x509 -inform der -in certificate .cer -out certificate .pem
```

- Run the following command to convert the private key.

```
opensslrsa -inform DER -outform PEM -in privatekey .der -out privatekey .pem
```

Convert P7B to PEM

P7B: This format is usually used in Windows Server and Tomcat.

Run the following command to convert the certificate format.

```
openssl pkcs7 -print_certs -in incertificate .p7b -out outcertificate .cer
```

Convert PFX to PEM

PFX: This format is usually used in Windows Server.

- **Run the following command to extract the certificate format.**

```
openssl pkcs12 - in certname . pfx - nokeys - out cert .  
pem
```

- **Run the following command to extract the private key.**

```
openssl pkcs12 - in certname . pfx - nocerts - out key .  
pem - nodes
```