Alibaba Cloud Server Load Balancer

Pricing

Issue: 20190415

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 Pay-As-You-Go

SLB instances are billed based on your actual traffic usage.

Billing items

The following table details the items that are billed. Billing items vary by network type and instance type, as shown in the following table.

Note:
"-" means that the corresponding item is not billed, and "#" means that the corresponding item is billed.

Network type	Instance type	Instance fee	Traffic fee	Specification fee
Internet	Shared-performance instances	#	#	-
	Guaranteed-performance instances	#	#	#
Intranet	Shared-performance instances	-	-	-
	Guaranteed-performance instances	-	-	#

Instance fee

SLB instances that communicate through the Internet incur fees for public IP address reservations. SLB instances that communicate through the intranet do not incur such charges. Instance fees for SLB instances that use the Internet are calculated as follows:

- Instance fee = unit price * instance reservation time

  The reservation time is the period from the time at which the instance is created to the time at which the instance is released.

- Instance fees are billed on an hourly basis. If your period of usage is less than one hour, then the bill is rounded up to one hour.

If the price on the purchase page of the console is different from the price listed in the following table, take the price on the purchase page as the standard.

Region	Instance fee (USD/hour)
China (Hangzhou), China (Beijing), China (Shenzhen), China (Shanghai), China (Zhangjiakou)	0.003
China (Qingdao)	0.003
Hong Kong	0.009
US (Virginia), US (Silicon Valley)	0.005
Singapore	0.006
Japan (Tokyo)	0.009
Germany (Frankfurt)	0.006
UAE (Dubai)	0.009
Australia (Sydney)	0.006

Traffic fee

SLB instances that communicate through the Internet incur traffic fees based on your usage. However, SLB instances that communicate through the intranet can be used free of charge. Traffic fees for SLB instances that use the Internet are calculated as follows:

- Internet traffic fee = unit traffic price * time

  Internet traffic is the outbound (downstream) traffic. Inbound (upstream) traffic is not charged.
- Traffic fees are billed on an hourly basis. If your period of usage is less than one hour, then the bill is rounded up to one hour.

  If the price on the purchase page of the console is different from the price listed in the following table, take the price on the purchase page as the standard.

Region	Traffic fee (USD/Gbit/s)
China (Hangzhou)/China (Beijing)/China (Shenzhen)/ China (Shanghai)/China (Zhangjiakou)	0.125
China (Qingdao)	0.113

Region	Traffic fee (USD/Gbit/s)
Hong Kong	0.156
US (Virginia)/US (Silicon Valley)	0.078
Singapore	0.117
Japan (Tokyo)	0.120
Germany (Frankfurt)	0.070
UAE (Dubai)	0.447
Australia (Sydney)	0.130

Capacity fee

The following are three key performance metrics for guaranteed-performance instances. The limits of these metrics are different for instances of different capacities. For more information, see *#unique_4*.

- 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 specification, the new connection will be dropped.
- Query Per Second (QPS)

  The number of HTTP/HTTPS requests that can be processed per second. This metrics is only available for Layer-7 Server Load Balancer. When the QPS reaches the limits of the specification, the new connection will be dropped.

The capacity fee of a guaranteed-performance instance is charged based on usage regardless of the capacity that you choose. If the actual performance metrics is between two capacities, the cost is calculated according to the larger capacity.

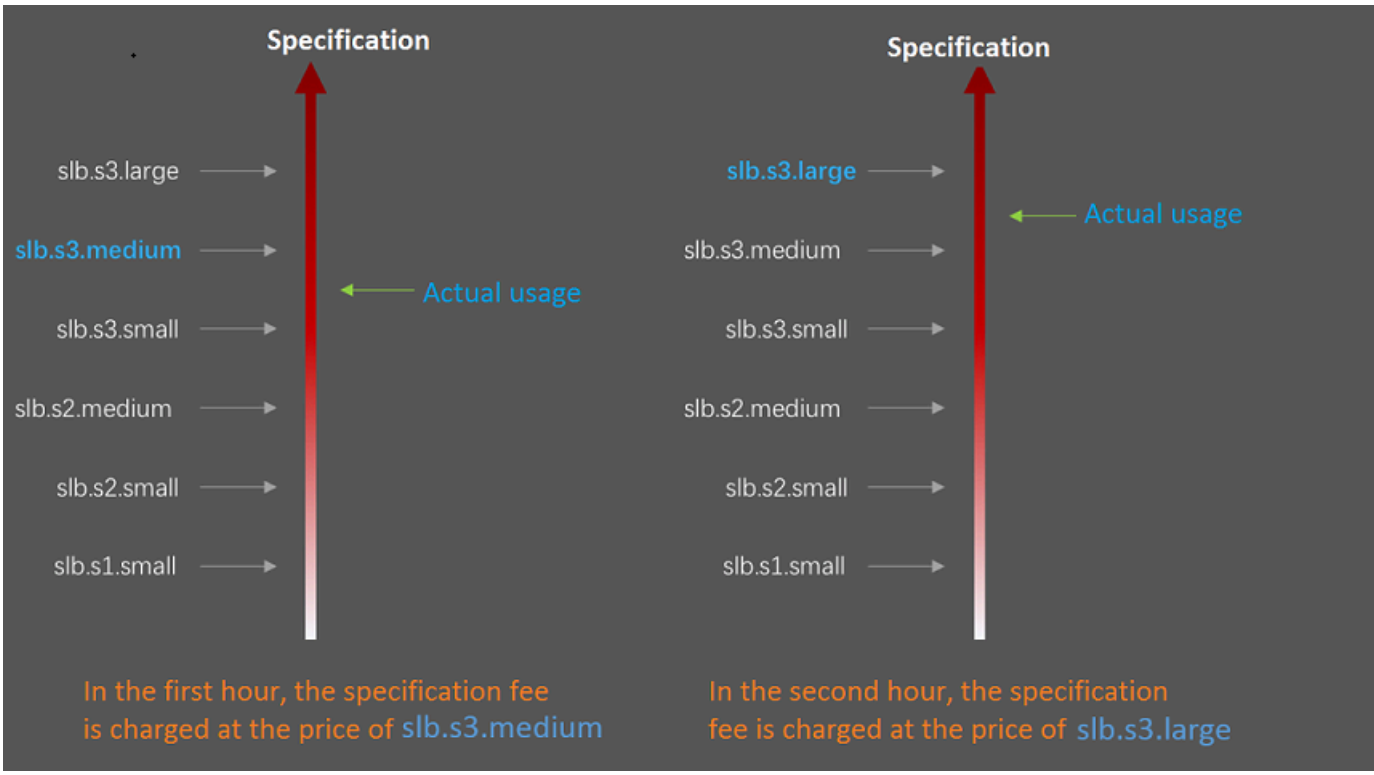
For example, the capacity slb.s3.large (1,000,000; CPS 500,000; QPS 50,000) is selected. The actual usage of your instance in an hour is as follow:

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) specification and the limit 5,000 defined in the Standard I (slb.s2.small) specification. Therefore, the specification 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 specification of the QPS metrics is highest, therefore, the specification fee of the instance in this hour is charged at the price of the Higher I (slb.s3.small) specification.

The following figure is an example showing how the specification fee is billed for an SLB instance:



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.

The price in the following table is only for reference. Take the price on the console as standard.

Region	Capacity	Max Connectio	CPS	QPS	Specification fee (USD/ hour)
China (Hangzhou) China (Zhangjiakou) China (Hohhot) China (Qingdao) China (Beijing) China (Shanghai) China (Shenzhen)	Capacity 1: Small I (slb.s1.small)	5,000	3,000	1,000	Free of charge
	Capacity 2: Standard I (slb.s2. small)	50,000	5,000	5,000	0.05
	Capacity 3: Standard II (slb.s2. medium)	100,000	10,000	10,000	0.10
	Capacity 4: Higher I (slb.s3.small)	200,000	20,000	20,000	0.20
	Capacity 5: Higher II (slb.s3.medium)	500,000	50,000	30,000	0.31
	Capacity 6: Super I (slb.s3.large)	1,000,000	100,000	50,000	0.51
Singapore Malaysia (Kuala Lumpur) Indonesia (Jakarta) India (Mumbai) US (Silicon Valley) US (Virginia) China (Hong Kong)	Capacity 1: Small I (slb.s1.small)	5,000	3,000	1,000	Free
	Capacity 2: Standard I (slb.s2. small)	50,000	5,000	5,000	0.06
	Capacity 3: Standard II (slb.s2. medium)	100,000	10,000	10,000	0.12
	Capacity 4: Higher I (slb.s3.small)	200,000	20,000	20,000	0.24
	Capacity 5: Higher II (slb.s3.medium)	500,000	50,000	30,000	0.37
	Capacity 6: Extra I (slb.s3.large)	1,000,000	100,000	50,000	0.61

2 Overdue instructions

The load balancing service will not be stopped immediately if an SLB bill is overdue. Renew Server Load Balancer in time to avoid service interruption.

The following will happen when a Pay-As-You-Go instance is overdue:

- **After a bill is overdue, the instance will keep running for 15 days. Then, the instance will be locked and stop service.**

  Once the instance stops running, billing is also stopped.
- **If the SLB bill is still overdue 15 days after the instance is locked, the instance will be automatically released.**

  The account owner will receive an email notification one day before the instance is released. The instance configuration and related data will be deleted and cannot be restored after the instance is released.

3 Monitoring data and billing data

Server Load Balancer (SLB) provides a monitoring function that monitors such metrics as the inbound and outbound traffic and the number of connections. You can view real-time monitoring data in the console. Besides monitoring data, billing data is also collected, but it is collected for the calculation of fees to be charged. Monitoring data and billing data differ given the factors described as follows.

Factor	Monitoring Data	Billing data
Calculation method	The SLB system collects monitoring data every minute, and reports the data to CloudMonitor. After every 15 minutes, CloudMonitor calculates the average value of data collected in that time period. The network traffic data displayed in the console is the average value calculated.	Billing data is collected every minute, and the SLB system reports the accumulated value once each hour to the billing system. Monitoring data is the calculated average for a 15-minute time period, but the billing data is the accumulated value in a billing cycle.
Latency	SLB provides real-time monitoring data. However, a short delay may inevitably occur during the process of data collection, calculation, and display. Although this delay is nearly immeasurable, it can create a certain degree of discrepancy between the monitoring and billing data.	Billing data can allow up to a three-hour delay. For example, billing data generated between 01:00-02:00 is normally reported to the billing system before 03:00. However, data may be reported up to three hours later, with the last reporting time being 05:00. As a result, there may be a discrepancy between billing data and monitoring data.
Purpose	The purpose of monitoring is to help you observe if instances are running normally. If not, you can take measures to solve problems in a timely manner.	The purpose of billing is to generate bills. Monitoring data cannot be used as billing data.