

Alibaba Cloud Server Load Balancer

Pricing

Issue: 20190306

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 Pay-As-You-Go.....	1
2 Overdue instructions.....	7
3 Monitoring data and billing data.....	8

1 Pay-As-You-Go

SLB is billed based on traffic usage.

Billing items

The cost of an SLB instance is the sum of the following billing items. The billing items vary by the 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

The instance fee is the public IP reservation fee of Internet SLB instances.



Note:

Intranet SLB instances are free of instance fee.

Instance fees of Internet SLB instances are billed as follows:

- Instance fee = unit instance price x instance reservation time

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

- Instance fees are billed on an hourly basis. In a billing cycle, partial hours are billed as full hours.

If the price on the purchase page is different from the price listed in the 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

Traffic fees are charged based on the traffic usage of Internet SLB instances.



Note:

Intranet instances are free of traffic fee.

Traffic fees of Internet SLB instances are billed as follows:

- Internet traffic fee = unit traffic price x usage time

Internet traffic is the outbound (downstream) traffic. Inbound (upstream) traffic is not included in the cost.

- Traffic fees are billed on an hourly basis. In a billing cycle, partial hours are billed as full hours.

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

Region	Traffic fee (USD/Gbps)
China (Hangzhou)/China (Beijing)/China (Shenzhen)/ China (Shanghai)/ China (Zhangjiakou)	0.125

Region	Traffic fee (USD/Gbps)
China (Qingdao)	0.113
China (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)	0.096

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 [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 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:

Region	Capacity	Max Connectio	CPS	QPS	Specificat ion fee (USD/ hour)
	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 provides a function that monitors the inbound and outbound traffic, number of connections, and more. You can view real-time monitoring data on the console. You are charged for the network traffic consumed by the Server Load Balancer instance. However, monitoring data is different from billing data, which is caused by factors as described in the following table.

Factor	Monitoring Data	Billing data
Calculation method	Monitoring data is collected every one minute by the Server Load Balancer system, and reported to the cloud monitoring system. Then, the cloud monitoring system calculates the average value of all collected data in each 15 minutes. The displayed network traffic data is the calculated average value.	Billing data is collected at the same granularity and the Server Load Balancer system reports the accumulated value in each hour to the billing system. The monitoring data is the calculated average value, but the billing data is the accumulation value. These two data sets are incomparable because they are calculated and generated differently.
Latency	Server Load Balancer provides real-time monitoring data. However, a short delay may inevitably occur in the data collection, calculation, and display process. Although this delay is almost insignificant, it can create a certain degree of discrepancy between the monitoring and billing data.	Billing data tolerates a maximum delay of three hours. For example, billing data generated between 01:00-02:00 is normally reported to the billing system at 03:00, but is allowed to be reported to the billing system at 05:00. As a result, there are differences between billing data and monitoring data.
Purpose	The purpose of monitoring is to help users observe if the instance is in abnormal conditions. If so, users can take specific measures to solve the problem as soon as possible.	The purpose of billing is to generate bills. Monitoring data cannot be used as the billing data.