Server Load Balancer

Pricing

MORE THAN JUST CLOUD | Alibaba Cloud

Pricing

Billing method

This document introduces the billing method for Server Load Balancer (SLB).

Billing items

SLB is billed based on traffic usage. The total cost of an SLB instance is the sum of the billing items. The billing items vary by network type and instance type as shown in the following table:

> **Note**: “√” indicates that the corresponding item is billed. “—” indicates that the corresponding item is not billed.

Network type	Instance type	Billing items		
		Instance fee	Traffic fee	Capacity fee
Internet	Shared-performance instances	√	√	—
	Guaranteed-performance instances	√	√	√
Intranet	Shared-performance instances	—	—	—
	Guaranteed-performance instances	—	—	√

Instance fee

Instance fees are charged for Internet SLB instances to reserve a public IP. Intranet SLB instances are not charged an instance fee.

Instance fees of Internet SLB instances are billed as follows:

> Total cost for each instance = unit price (USD/Hour) of the instance x instance reservation time
>
> The reservation time is the time range from when the instance is created to when the instance is released.
>
> Instance fees are billed on an hourly basis. Partial hours are billed as full hours.

Traffic fee

Traffic fees are charged based on the traffic usage of Internet SLB instances. The intranet SLB instances are not billed a traffic fee.

Traffic fees of Internet SLB instances are billed as follows:

> Total cost for each instance = unit price (USD/Hour) of the public network traffic x instance reservation time
>
> Public network traffic is the outbound traffic (downstream). Inbound traffic (upstream) is not included in the cost.
>
> Instance fees are billed on an hourly basis. Partial hours are billed as full hours.

Capacity fee

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

Guaranteed-performance instances provide guaranteed performance metrics (performance SLA). Different capacities are provided to meet unique business requirements. For more information, see Guaranteed-performance instances.

The corresponding capacity fee is billed for each guaranteed-performance instance no matter the network type of the instance, and is billed based on the actual usage depending on the capacity selected. If the actual performance metrics of an instance occurs between two capacities, the capacity fee is charged at the higher capacity fee.

For example, if you purchase the **Super I (slb.s3.large)** capacity, 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)**.

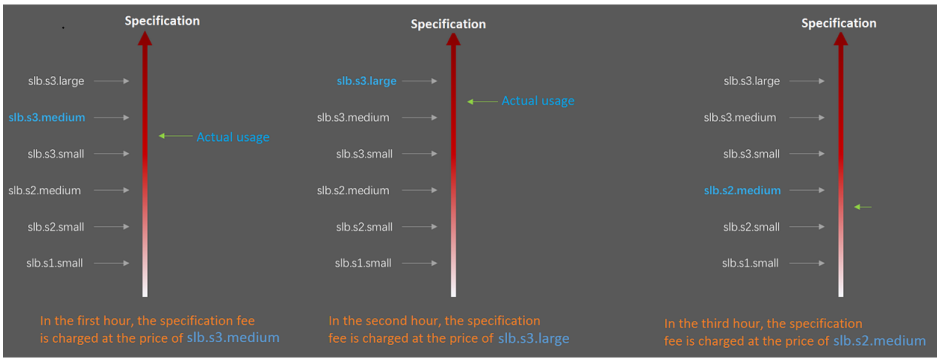
From the perspective of CPS, the actual metrics 4,000 occurs between the limit 3,000 defined in the **Small I (slb.s1.small)** capacity and the limit 5,000 defined in the **Standard I (slb.s2.small)** capacity. Therefore, the capacity of the CPS metrics in this hour is **Standard I (slb.s2.small)**.

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.

Capacity		**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

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 performance capacity selected when purchasing an SLB instance limits the performance. For example, if **slb.s3.medium** is selected, the new connections are dropped when the HTTP requests in one second reach 30,000.

Pricing

Table 1: Instance and traffic fee

Region	Instance fee (USD/Instance/Hour)	Traffic fee (USD/Gbps)
East China 1 (Hangzhou)/North China 2 (Beijing) /South China 1 (Shenzhen) /East China 2 (Shanghai) /North China 3 (Zhangjiakou)	0.003	0.125
North China 1 (Qingdao)	0.003	0.113
Hong Kong	0.009	0.156
US East 1 (Virginia)/US West 1 (Silicon Valley)	0.005	0.078
Singapore	0.006	0.117
Asia Pacific NE 1 (Japan)	0.009	0.120
Central Europe 1 (Frankfurt)	0.006	0.070
Middle East 1 (Dubai)	0.009	0.447
Asia Pacific SE 2 (Sydney))	0.006	0.130

Table 2: Capacity price

Capacity		Max Connection	CPS	QPS	Capacity price (USD/Hour)
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	0.30
Capacity 6	Super I (slb.s3.large)	1,000,000	100,000	50,000	0.50

Overdue payments

The following are overdue payments policies:

> If an instance payment becomes overdue, you can still use the service for another 15 days. You will receive an email that reminds you to make a payment and get services renewed. The service will not be affected if the payment is made within 15 days.
>
> If you do not renew the service after the payment is overdue for 15 days, the service will be stopped. No fees are charged to the Server Load Balancer instance when the service is stopped. Instance configuration and related data will be reserved for another 15 days after the service is stopped.
>
> If you recharge your delinquent account within 15 days after the service is stopped, the instance will be automatically restarted.
>
> If you do not recharge your delinquent account within 15 days after the service is stopped, the instance will be released.
>
> An email reminder will be sent to you one day before the instance is released. Once the instance is released, the configuration data will be permanently deleted and cannot be restored.

Differences between 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.

Factors	Monitoring data	Billing data
Calculation methods	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. Therefore, the billing data is different from the monitoring data.
Purpose	The purpose of monitoring is to help users observe if the instance is in abnormal conditions. If so, users can resolve the problem as soon as possible.	The purpose of billing is to generate bills. Monitoring data cannot be used as the billing data.