CHAPTER 1

Quota Management Overview

Revised: September 09, 2013, OL-24193-08

Introduction

This chapter provides an overview of the Quota Manager. The chapter contains the following sections:

Information About the Quota Management Solution

Quota Management is a type of policing of broadband user traffic that enforces policy actions based on integral and accumulative characteristics. Specifically, the Quota Manager controls the amount of consumed traffic per application and direction. The subscribers who consume various amounts of the data traffic during specified periods of time can help the provider in preventing abusive use of ISP resources shared by ISP subscribers. This is important in ensuring improved user experience for the maximum number of subscribers.

The Cisco Service Control Application for Broadband (Cisco SCA BB) solution provides powerful quota management capabilities that enable simple implementation of quota management for application traffic:

- Managed network resources—Upstream traffic volume, downstream traffic volume, number of sessions.
- External or internal quota replenishment schemes.
- Real-time notifications—Subscriber quota depletion, subscriber quota below threshold.
- Java API for Quota Management integration—Included within the SCE Subscriber Management API.
- Simultaneous different quota management schemes per subscriber group (policy package).
- Various actions upon quota depletion/breach—Bandwidth control, block, subscriber notification, real-time notification.
- Raw Data Record (RDR) based reporting.
- Support for multiple SCEs—Manages the quota consumed by a subscriber when the subscriber is simultaneously logged in from one or more SCEs within a domain.
- Support for multiple bucket quota provisioning with the penalty profile.

Cisco SCA-BB supports three Quota Management operational and integration models that allow gradual investment and trade-off between integration/deployment complexity and functional offering:

- SCE Internal model—Time-based, auto-replenished quota.
- Subscriber Manager Quota Management model—Time-based, auto-replenished quota with preserved state.
- Flexible model—Integration with external QM.

Quota Manager Description

In versions of the Service Control Management Suite (SCMS) Subscriber Manager (SM) earlier than Release 3.0, subscriber quota levels could be maintained across subscriber sessions. This functionality, removed in Release 3.0, has been enhanced and reinstated.

The Quota Manager is now available as a component of the Subscriber Manager, which enables the Service Control solution providers to manage the subscriber quota, with a high degree of flexibility.

This section contains these subsections:

Quota Manager Functionality

The Quota Manager controls SCA-BB quota functionality and acts as an entry-level quota policy repository. The Quota Manager is an event-driven solution, which leverages the functionality of the Service Control Engine (SCE) Subscriber API.

The quota manager provides the following functionality:

- Subscriber quota is preserved across subscriber sessions.
- Aggregation periods and amounts can be set on a per package basis.
- Quota allocation at the beginning of an aggregation period can be spread over time for different subscribers to avoid the buildup of traffic bursts at the start of the aggregation period.
- Subscriber quota is preserved across Quota Manager upgrade.

The Quota Manager supports the following:

- All SCE topologies (1+1 and MGSCP).
- High availability of the Quota Manager server (utilizing a Veritas Cluster Server [VCS]).
- Multiple quota thresholds that allows service providers to move subscribers to penalty packages when a certain quota threshold is breached.

Using the Quota Manager, subscribers can do the following:

- Assign time-based quota for a period called an aggregation period while the consumption during this period is calculated in a sliding window model. For more information, see the "Sliding Window Model" section on page 1-8.
- Assign a one-time quota that can be replenished only manually.
- Move between packages at any time, whether they are logged in or not.
- Purchase additional quota within an aggregation period.

Quota Manager Module

The Quota Manager module runs as a component on the Subscriber Manager. The logic to manage and maintain quotas runs on the Subscriber Manager server; therefore, you should configure the QM on the Subscriber Manager or load the configuration onto the Subscriber Manager. In a cluster setup, you must load the configuration onto each Subscriber Manager in the cluster. Subscriber quotas are stored in the SM database.

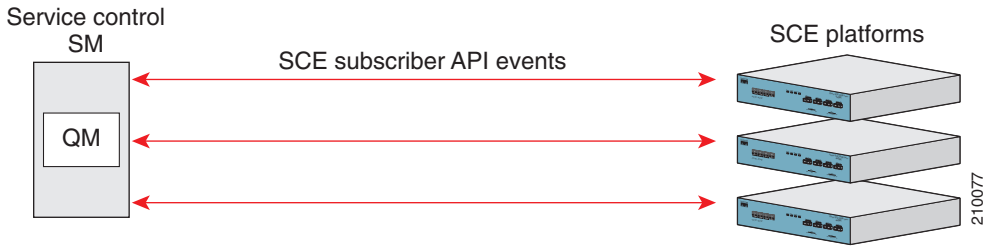
The Quota Manager uses the SCE Subscriber API to provision quota to subscribers upon request by using the existing external quota functionality of the SCE.

Note Per package quota is configured on the Subscriber Manager and this adds complexity to the integration or operation. Quota management also requires more management messages on the network.

Quota Manager Network Topology

The Quota Manager can serve more than one SCE. At any point in time, one subscriber is managed by one SCE. Figure 1-1 shows the network topology of a system, which uses quota management.

Figure 1-1 Quota Manager Network Topology



Quota Indications and Quota Responses

The Quota Manager is an event-driven component. All quota logic is performed as a response to quota indications that the SCE initiates. Therefore, the quota state is up-to-date in accordance with the last quota indication received.

Table 1-1 lists the quota indications and the quota responses to the indications.

Table 1-1 Quota Indications and Responses

Indication	Reason Generated	Quota Manager
Quota breach	Generated when a subscriber uses the entire available quota in the SCE.	Responds to this indication by providing more quota for the subscriber if the subscriber quota allocation is not depleted for the current aggregation period.
Remaining quota	Periodically generated to keep the Quota Manager updated with the information about the quota remaining in the SCE. Generated in response to a subscriber logout event.	In most cases, does not respond to this indication, but writes the quota value to the database to be stored until the subscriber logs in next. If the reported quota is below zero, responds by providing more quota for the subscriber if the subscriber quota allocation has not been depleted for the current aggregation period.
Quota below threshold	Generated when the subscriber quota in the SCE drops below a predefined level.	Responds to this indication by providing more quota for the subscriber if the subscriber quota allocation is not depleted for the current aggregation period.
Quota state restore	Generated in response to a subscriber login event.	Responds to this indication by updating the subscriber quota in the SCE.

Note Quota calculation uses the system date and time of the Quota Manager machine to calculate aggregation periods. During Quota Manager operation, if the date and time of the system are changed, we recommend that you delete all quota information from the SM database to recalculate all aggregation periods according to the new date and time. This operation also causes a quota replenish for all subscribers. To remove quota information from the SM database, run the **p3subsdb --clear-all-states** command.

Cisco SCA BB Quota Buckets

The basic building block that the Cisco SCA BB uses to implement a specific quota is called a quota bucket. Each subscriber can be assigned a maximum of 16 quota buckets to maintain the utilization of subscriber traffic over a specific service.

The following network resources can be managed with a quota bucket:

- Traffic volume in units of Layer 3 kilobytes—Subscriber traffic consumption can be monitored separately per traffic direction such as from the subscriber (up) or to the subscriber (down).
- Number of sessions—The total number of network sessions classified to the services associated with the quota bucket.

The SCE provides real-time notifications and reporting of the quota breach, remaining quota periodic, quota state restore, and quota threshold events.

Depletion or breach of a quota bucket occurs when a monitored resource is consumed when the bucket is empty. After the quota is depleted, the quota bucket causes all the service rules associated with that quota bucket to execute breach-state actions. Quota bucket replenishment (automatic or external) can bring the quota to a non-depleted state, and has the service rules associated with the quota bucket execute the normal (nonbreached) actions.

Upon quota depletion, the SCE platform can perform one or more actions, as defined in the service configuration settings. The following actions are:

- Send quota breach RDR.
- Signal a quota breach notification through Java API.
- Activate subscriber notification (HTTP redirect-based notification).

Quota Provisioning

The Quota Manager uses quota provisioning to provide additional quota to the subscriber as a response to the following quota indications: Quota State Restore, Quota Below Threshold, and Quota Breach. Quota provisioning occurs only if quota is still available for the subscriber.

Quota provisioning is split into dosages to ensure that quota consumption is accounted for accordingly in the quota manager database. This ensures that in cases of failures, the amount of quota used, but not accounted for by a subscriber, is limited by the quota dosage size.

The Quota Manager performs the provisioning by adding quota to the SCE so that after the provisioning operation, the available quota in the SCE equals the minimum of either the dosage size or the remaining quota in Quota Manager.

For example, when the bucket size is 100 MB, the dosage size is 10 MB, and the threshold size is 1 MB, the following quota provisioning takes place:

- When the subscriber logs in for the first time, the SCE initiates a Quota State Restore indication for this subscriber. This indication triggers the following actions:
  - Quota of 100 MB is added to the Quota Manager database for the first aggregation period.
  - The Quota Manager provisions only 10 MB to the SCE, as configured by the quota dosage value.
- After the subscriber consumes 9 MB of quota, the quota threshold (1 MB) is reached and the SCE initiates a Quota Below Threshold indication for this subscriber. This indication triggers the following actions:

- The QM provisions a further dosage of 10 MB to the SCE.
- At the same time, the QM updates its database to indicate that 9 MB of quota was consumed by the subscriber. After the provisioning operation, the remaining quota in the QM is 91 MB (for example, 100 - 9).

Aggregation Period, Slices, and Bucket

To measure the consumption of each subscriber, Quota Manager uses these three concepts in the sliding window model:

- Aggregation period—Time-based quota for a period.
- Slice—Aggregation period is further divided into time-based units called slices. Slices are valid only when the aggregation period is set to minutes, hourly, daily or weekly. Minimum value is 10. When the aggregation period is set to monthly or none, the slice period is not used or should be set to -1. Here -1 is the default value and it means that the period equal to the aggregation period.
- Bucket—Quota limits within an aggregation period.

Table 1-2 provides details of aggregation periods, allowed slice periods, and maximum slice period.

Table 1-2 Aggregation Periods, Allowed Slice Periods, and Maximum Slice Periods

Aggregation Period	Allowed Slice Period	Maximum Slice Period
60 minutes	10, 20, 30, 60	60
180 minutes	10, 20, 30, 60, 90, 180	180
1440 minutes	10, 20, 30, 60, 90, 180...1440	1440
hourly	10, 20, 30, 60	60
daily	10, 20, 30, 60, 90, 180...1440	1440
weekly	420...10080	10080
monthly	-1	-1
none	-1	-1

Slice quota must always be a whole number. If it is a fractional value, the fractional part is ignored during quota calculation. This validation is done on loading the configuration file. A warning is displayed. Modify the values so that the slice quota will be a whole number. You can use '--ignore-warnings' option to complete the task.

For example, if bucket is 1000, aggregation period is 30 minutes, and slice period is 10 minutes, the number of slices will be 3 and the slice quota will be 333.3333. The fractional part is ignored. So the total remaining quota in Quota Manager is 999 (333+333+333). In this scenario, adjust the bucket size to 1200 or number of slices to 2 to get a whole number as the slice quota.

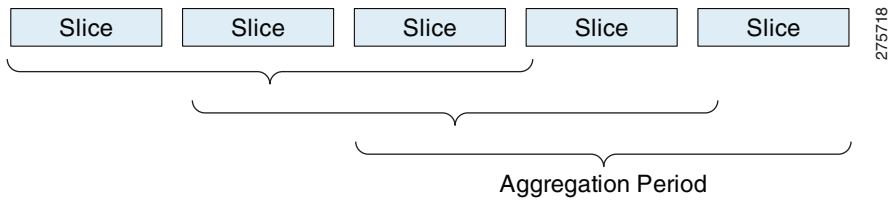
Sliding Window Model

Before Release 3.5.5, the Quota Manager measured the consumption of each subscriber for the fixed aggregation periods. At the end of the aggregation period, a new period starts with no memory about the consumption of the previous period (in other words, starting each aggregation period from scratch).

Starting in Release 3.5.5, the aggregation period is divided into multiple slices. Using the sliding window module, the SCE calculates the bandwidth consumption for each time slice or time period separately and saves the calculated bandwidth consumption in the Quota Manager. The SCE calculates the bandwidth consumption during the aggregation period over a configured number of time slices (*N* slices) or time periods. Calculating bandwidth consumption over the configured number of time periods ensures that the bandwidth consumption calculation is based on the average bandwidth consumption.

Figure 1-2 illustrates the aggregation time over multiple slices.

Figure 1-2 Aggregation Time Over Multiple Time Slices



Note The sliding history data may be inaccurate if the subscriber consumption is not spread across the aggregation period. If you expect that a switching of packages may occur, we recommend that you enable the **reset_quota_on_profile_switch** tunable and the **reset_quota_on_penalty_profile_switch** tunable. The tunables are in the Quota Manager section of the Cisco Service Control Subscriber Manager configuration file. To enable the tunables, change the value of the tunables to *true*.

Multiple Thresholds of Subscriber Quota

The Multiple Thresholds of Subscriber Quota allows differentiating quota use among the groups of subscribers.

The Quota Manager manages a bucket per subscriber and when a defined threshold is crossed, the Quota Manager changes the profile of the subscriber to the penalty profile according to the configuration. Subscribers who breach their quota are put in the penalty profile for a period of time. After the penalty time expires, Quota Manager verifies subscriber quota consumption and moves the subscriber to the appropriate profile based on the configuration. Quota Manager also identifies the quota consumption under a certain threshold during the penalty period.

For multiple penalty quota profile configuration example, see the "Configuring the Quota Manager—Example" section on page 3-8.

Note On Cisco SCE Releases earlier than Release 3.7.2, penalty profiles support only a single bucket for each subscriber.

Configuring the sliding window algorithm to *n* slices allows the sliding window algorithm to account for the penalty calculation for the *N* – 1 last slices and the current slice.

For example, if $N = 3$ and the quota threshold is configured to x bytes over the aggregation time. The algorithm will take the consumption for the last two time slices and the consumption report received for the current time slice to calculate the average consumption for three time slices. If the subscriber consumed more than the average x bytes over the time slices, the subscriber is switched to the penalty package.

If a subscriber exceeds one of the thresholds, the subscriber is switched to the package for a penalty time. If during the penalty time the subscriber again exceeds one of the thresholds, the subscriber is switched to the penalty package according to the configuration and the penalty timer is reset (counting the penalty time from zero).

Multiple Quota Thresholds—Example

In the following example, the requirement is to distinguish between the three groups of the subscribers over the aggregation time. The three subscriber groups are, namely:

- Regular subscribers.
- Heavy subscribers—Subscribers who consumed more than 500 KB in the last 30 minutes.
- Very heavy subscribers—Subscribers who consumed more than 1000 KB in the last 30 minutes.

Regular subscribers are untouched, heavy subscribers are lightly controlled when the interface is close to congestion, and very heavy subscribers are forcefully controlled when the interface is close to congestion.

The SCE sends RDRs to the Quota Manager. The Quota Manager detects heavy subscribers based on the amount of bandwidth (BW) these subscribers consumed in the last x minutes. Heavy subscribers are divided into two groups based on their consumption (those who crossed first threshold and those who crossed first and second thresholds) and each group is assigned a different package. The Quota Manager manages a bucket for each subscriber and when a threshold is crossed, the Quota Manager changes the subscriber package to either the heavy package or very-heavy package.

The finite state machine (FSM), as shown in Figure 1-3, illustrates the new quota algorithm. Immediately after the subscriber is moved to a new package, the current status is verified. As a result of this, the subscriber might immediately move two steps in a row.

Figure 1-3 Finite State Machine

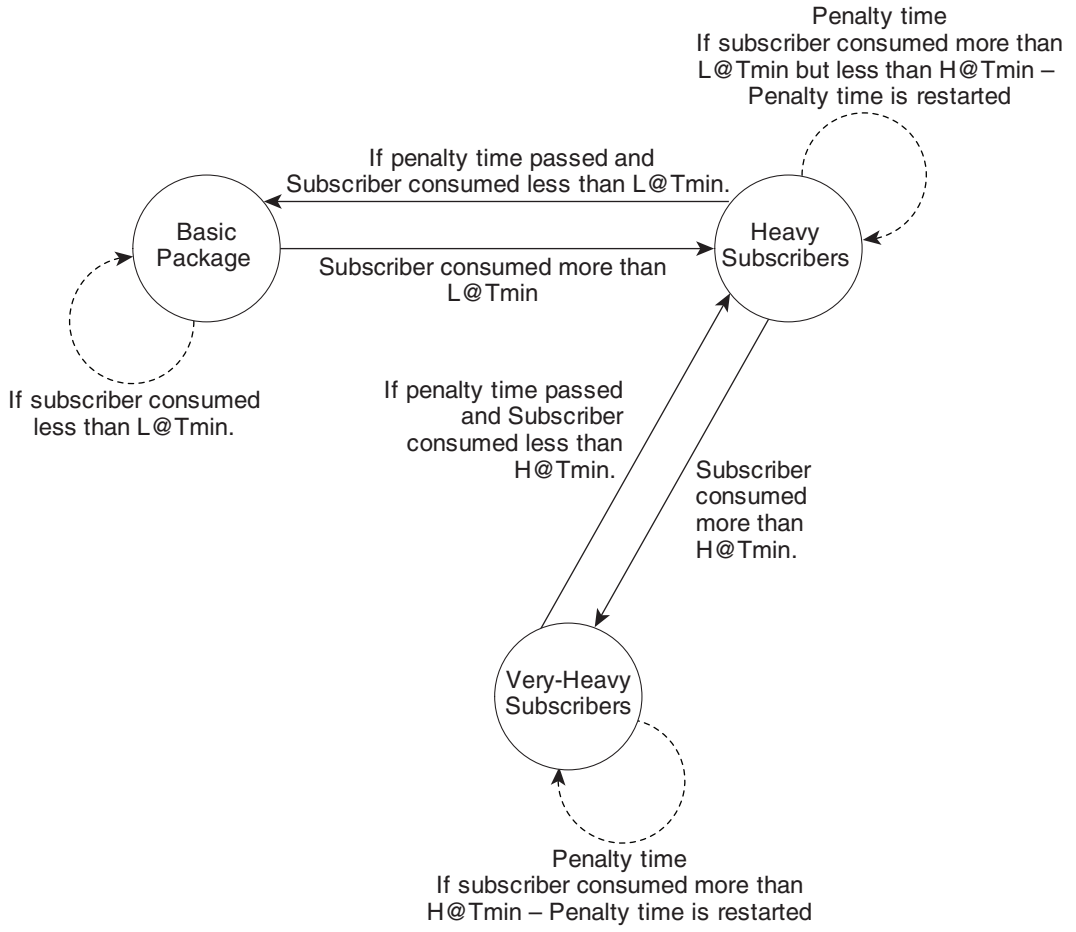


Table 1-3 lists the parameters relevant to each threshold and quota package.

Table 1-3 Quota Threshold Parameters

	Quota Profile/ Package 1	Quota Profile/ Package 2	Quota Profile/ Package 3
Threshold on quota greater than/Move to	L/2	H/3	H/3
Threshold on quota less than/Move to	L/1	L/1	H/2
Assurance level	10	6	2

During the aggregation period, if the subscriber exceeds the high threshold, the subscriber is moved to another package and the aggregation period start is reset to be the package change time. In addition, a penalty timer is reset.

On the first RDR after the aggregation period, the QM checks if the penalty time has passed and if the quota was below the low threshold. During the last aggregation period, if the quota was below the low threshold and the penalty time has passed, the package is changed.

Support for Multiple Bucket Quota Provisioning with Penalty Profile

Starting with Cisco SCMS Quota Manager Release 3.7.2, Cisco SCE supports multiple bucket quota provisioning with the penalty profile.

Upstream flows, downstream flows, or both flows, can be used as quota buckets with the penalty profile. Cisco Service Control treats all flows as individual buckets.

Based on the bucket usage, the penalty profile is defined. Penalty switch occurs on breach of at least one of the configured buckets, rather than all the configured buckets. Each bucket in the quota profile must be associated with the penalty profile. The number of quota buckets and the number of penalty profiles must always be equal in the quota profile. If the subscriber breaches more than one bucket simultaneously, the penalty switch moves the first bucket to the penalty profile.

Even if a subscriber has breached only one bucket, and has not breached or used any other bucket, penalty switch occurs based on the first bucket. If a subscriber who is present in the penalty profile reaches the threshold, the subscriber moves to the presently configured postpenalty profile.

While configuring postpenalty thresholds, use the logical or OR (|) separators to configure different thresholds for different buckets. If the consumption is less than the configured threshold within a penalty period, the subscriber will be switched to the corresponding postpenalty profile.

post_penalty.[10|20]=QP11

The subscriber will be moved to QP11 if the consumption of quota is less than 10 percentage of bucket 1 or less than 20 percentage of bucket 2.

post_penalty.[90]=QP22

The subscriber will be moved to QP22 if the consumption of quota is less than 90 percentage of bucket 1 or less than 90 percentage of bucket 2.

Note The number of quota buckets should always be equal in all the quota profiles in a penalty profile chain.

Support for Multiple Cisco SCEs

When a subscriber is connected from two SCEs—SCE fixed and SCE mobile—both SCEs request quota for the subscriber according to the configured service. When the Quota Manager receives a quota request from SCE mobile, it replenishes the previous dosage given to SCE fixed and associates the subscriber to SCE mobile instead of SCE fixed. This may mislead the quota calculation on the Quota Manager and the subscriber may be dissociated from SCE fixed.

Starting with Cisco SCMS Quota Manager Release 3.7.0, the Quota Manager supports quota consumption monitoring from multiple SCEs. The quota consumed is calculated from the moment the subscriber simultaneously logs in from one or more SCEs. Dosage is allocated to all the SCEs until the subscriber moves to a depleted or breached state.

The Quota Manager keeps track of the quota allocated to the SCEs and the quota consumed by the SCEs for a specific subscriber to compute the quota calculations accurately.

Definitions for aggregation period and slice period remain same even if multiple SCE support is enabled.

Quota Allocation Modes

When a subscriber is allowed to log in from multiple SCEs, because of the asynchronous nature of quota consumption, quota consumed may be greater than or less than the bucket size.

When the quota usage approaches the bucket value, the SCE decides whether the subscriber be allowed to consume the quota greater than or less than the bucket size. The additional quota or lesser quota that is provisioned is equal to the value of dosage size multiplied by the number of SCEs from which the subscriber is logged in.

To minimize the amount of quota divergence consumed by a subscriber, you must configure the quota profile in the Quota Manager. We recommend that the dosage size be as minimum as possible to minimize the quota loss/gain consumption. Configure the dosage size as a fraction of the bucket size and based on the number of Cisco SCEs that are configured. Make sure you configure the dosage in such a way that the Cisco Service Control Quota Manager can allocate the required quota to all the configured Cisco SCEs upon login.

Two quota allocation modes are supported, namely:

- Consumption
- Provisioned

Consumption Mode

In this mode, the dosage allocation to the SCEs is based on the actual quota consumption. The subscriber may consume more quota than the bucket size per aggregation period, because the actual quota consumption for a subscriber is calculated based on the Quota Status RDRs and QUOTA_BELOW_THRESHOLD RDRs received from the SCEs.

The potential gain for each bucket for each aggregation period can be up to the value equal to the dosage value multiplied by the number of SCEs from which the subscriber is logged in.

The available quota is considered to be equal to the quota available in the Quota Manager. If the available quota is greater than 0, a dosage is provisioned to the SCE. Other wise, the request is discarded.

For the configuration example, see the "Consumption Quota Allocation Mode with Multiple Cisco SCEs" section on page 2-14.

Provisioned Mode

In this mode, the dosage allocation is based on the dosages provisioned to the SCEs when a Quota Request RDR is sent to the Quota Manager. After a dosage is provisioned to the SCE, the Quota Manager assumes that this dosage has been consumed irrespective of whether the SCE consumes the dosage or not.

In this mode, there might be a potential loss to the subscriber if the provisioned dosage is not consumed by that particular SCE. The amount of this loss per bucket for each aggregation period can be up to a value equal to the dosage value multiplied by the number of SCEs from which a subscriber is logged in.

The available quota is considered to be equal to the quota available in the Quota Manager minus the sum of remaining quota from all the SCEs. If the available quota is greater than 0, a dosage is provisioned to the SCE. Otherwise, the request is discarded.

For the configuration example, see the "Provisioned Quota Allocation Mode with Multiple Cisco SCEs" section on page 2-16.

Limitations

There are some limitations to providing support for multiple Cisco SCEs:

- This feature is applicable only to the Cisco SCEs within a domain.
- A subscriber can be in breached state in one Cisco SCE and in normal mode in the other Cisco SCE.

  This implies that, a subscriber might continue to be on a base package without getting penalized even after the subscriber breaches the quota on one Cisco SCE. The subscriber is penalized only after the subscriber breaches the quota on all configured Cisco SCEs. To overcome this limitation control the quota consumption using appropriate breach actions available through Cisco SCA BB.
- If there are frequent logouts by a subscriber, there might be additional loss or gain in the quota consumption.
- All Cisco SCEs should have similar profiles. The number of buckets in each profile is limited to one and no slices should be configured on these buckets.
- The configured dosage size is treated as the dosage allocated to the single Cisco SCE per subscriber.
- This feature supports only volume-based quota management.
- This feature supports simultaneous subscriber login from a maximum of eight Cisco SCEs.
- There is no provision to add or set quota for the breached subscriber with multiple Cisco SCE support.

Recommendations

We recommend that:

- If the probability of concurrent logins from different SCEs is high for a subscriber, configure a lower dosage value to minimize the additional factor or the quota consumption.
- If the subscriber logouts from different SCEs are more frequent, use actuals quota mode.