Cisco ACI Support for NGINX Rate Limit

Revised: April 4, 2025

New and changed information

This document provides procedures for configuring the NGINX rate limit feature. The following table provides an overview of the significant changes to this guide up to this current release.

Table 1: New and Changed Information In This Document

Release	Feature
Initial release of document	Initial release of document.
APIC release 6.1(2)	Support for custom throttle group.
APIC release 6.1(2)	The global request throttle rate for HTTPS is enforced to be 40 (requests/second) or 2400 (requests/minute) or smaller.
APIC release 6.1(3)	Support for the APIC GUI configuration for Custom Throttle Group.

Understanding the NGINX rate limit feature

NGINX provides REST API services to clients who can read and configure the Cisco ACI fabric. Clients could be the APIC GUI or scripts that have been developed by you or by Cisco.

These services are provided to verify the authentication procedure. If a client sends multiple requests that competes with other applications, then the serviceability of NGINX gets affected for any trusted client as well. For example, if a script sends the same requests in a loop, that will result in the requests being handled in a certain sequence, which will leave other requests waiting for a long period of time where those other requests will eventually time out.

Beginning with Cisco APIC Release 4.2(3), the NGINX rate limit feature has been introduced to avoid this situation. This solution, known as global throttling is applicable across all applications and processes. A new option for the NGINX rate limit feature, called as custom throttle group, prioritizes clients and applications based on the assigned throttle rate. For more details, see the next section, Custom throttle group, on page 2.

Note This procedure describes how to configure the NGINX rate limit (global throttling). For information on configuring HTTP and HTTPS AAA login throttling, see "Configuring HTTP and HTTPS Throttling Using the CLI" in the *Cisco APIC REST API Configuration Guide*.

Custom throttle group

Beginning with APIC release 6.1(2), you can use the custom throttle group feature which allows for prioritization of different clients and applications and adjusts the throttle rates for the applications. Using custom throttle group, you can assign a throttling rate to the required clients using an appropriate tag in their HTTP header of REST API requests. The requests that carry the tag, "Ratelimit-Tag", in the HTTP header are rate-limited accordingly. You can configure a maximum of three throttling groups with the custom tag and rate. Each group can be assigned one tag. Each of these groups are configurable to a maximum of 20 requests per second. This feature is disabled by default.

Read queries through the custom throttle group are given higher priority to access the APIC database compared to the queries throttled by the global throttling or queries without any throttling when the global throttling is disabled. The *read* queries from the APIC GUI are always prioritised over custom or global throttling.

Note The custom rate-limit is supported only for HTTPS.

Configure global throttling using the GUI

Procedure

Step 1 In the menu bar choose **Fabric** > **Fabric Policies**.

Step 2 In the navigation pane, expand **Policies** > **Pod** > **Management Access**, then choose the default management access policy.

Step 3 For releases 6.1(1) and after, choose **Policy** > **Web Access**.

Step 4 Determine if you want to enable global throttling for HTTP or HTTPS requests.

- If you want to enable global throttling for HTTP requests:

  a. Locate the **HTTP** area in the window, then locate the **Request Throttle** field in the **HTTP** area.

  The setting for this field should be set to **Disabled** by default.

  b. Click **Enabled** to enable global throttling for HTTP requests.

  The **Throttle Rate** field appears.

  c. Set the throttle rate for the HTTP requests:

  - Enter a number between 1 and 40 requests per second or 1 and 2400 requests per minute.
  - Select either **Requests/Second** or **Requests/Minute** to set the global throttle unit as either the number of requests per second or the number of requests per minute.

- If you want to enable global throttling for HTTPS requests:

  a. Locate the **HTTPS** area in the window, then locate the **Request Throttle** field in the **HTTPS** area.

  The setting for this field should be set to **Disabled** by default.

  b. The **Throttle Rate** field appears.

  c. Set the throttle rate for the HTTPS requests:

  - Enter a number between 1 and 40 requests per second or 1 and 2400 requests per minute.

    The GUI allows you to enter up to 10,000; however, starting from the 6.1(2) release the GUI rejects a value larger than 40 requests per second or 2400 requests per minute.

- Select either **Requests/Second** or **Requests/Minute** to set the global throttle unit as either the number of requests per second or the number of requests per minute.

Step 5 When you have finished setting the global throttle rate for HTTP or HTTPS requests, click **Submit** in the lower right corner of the window.

Configure custom throttle groups using the GUI

Procedure

Step 1 In the menu bar choose **Fabric** > **Fabric Policies**.

Step 2 In the navigation pane, expand **Policies** > **Pod** > **Management Access**, then choose the default management access policy.

Step 3 For releases 6.1(1) and after, choose **Policy** > **Web Access**.

Step 4 For **Custom Throttle Groups**, click **Enabled**.

Step 5 For each of the three custom **Request tag for group** fields and the respective **Throttle Rate** fields, enter the desired values.

- **Request tag for group X**—The tag value to be used for the HTTP header "Ratelimit-Tag" for the custom throttle group X where X is 1, 2, or 3. API requests with such a tag are throttled with the configured rate. If you leave a custom throttle group field empty, then that group will not be active.
- **Throttle Rate (requests/second)**—Request throttle rate to be applied to the given custom group.

Step 6 Click **Submit**.

Configuring global throttling using the NX-OS Style CLI

Prior to Cisco APIC Release 4.2(3), the following throttling commands were only available through the NX-OS style CLI:

- **enable-throttle**: Used to enable HTTP or HTTPS AAA login or refresh throttling.
- **throttle**: Used to set the throttle rate used for HTTP or HTTPS communication service after enabling throttling using the **enable-throttle** command.

Beginning with Cisco APIC Release 4.2(3), the following throttling command is now also available:

- **global-throttle**: Used to enable global throttling for HTTP or HTTPS requests.

Note the following behaviors, depending on which throttling command is enabled or disabled:

- When **enable-throttle** is disabled and **global-throttle** is enabled, the login or login refresh is counted as one of the requests in global rate-limiting, but is not counted as login-specific rate-limiting.
- When **enable-throttle** is enabled and **global-throttle** is disabled, only the login or login refresh is affected.

Procedure

Step 1 Navigate to the area in the CLI where you can configure the default communication policy:

Example:

```
apic1# config
apic1(config)# comm-policy default
apic1(config-comm-policy)#
```

Step 2 Determine if you want to enable global throttling for HTTP or HTTPS requests.

- If you want to enable global throttling for HTTP requests, enter `http` to configure the HTTP communication policy group:

  ```
  apic1(config-comm-policy)# http
  apic1(config-http)#
  ```

- If you want to enable global throttling for HTTPS requests, enter `https` to configure the HTTPS communication policy group:

  ```
  apic1(config-comm-policy)# https
  apic1(config-https)#
  ```

Note
The commands for the remaining steps are the same, whether you are configuring an HTTP or an HTTPS communication policy group. The following steps show how to configure an HTTP communication policy group as an example.

Step 3 Enable global throttling for the HTTP or HTTPS requests.

Example:

```
apic1(config-http)# global-throttle
apic1(config-http)#
```

Step 4 Set the global throttling rate for the HTTP or HTTPS requests:

```
apic1(config-http)#  global-throttle-rate <1-40>
```

Example:

```
apic1(config-http)# global-throttle-rate 40
apic1(config-http)#
```

Step 5 Set the global throttling unit for the HTTP or HTTPS requests.

- To set the global throttling unit as number of requests per second:

  ```
  apic1(config-http)# global-throttle-unit r/s
  ```

- To set the global throttling unit as number of requests per minute:

  ```
  apic1(config-http)# global-throttle-unit r/m
  ```

Step 6 To disable global throttling for the HTTP or HTTPS requests:

Example:

```
apic1(config-http)# no global-throttle
apic1(config-http)#
```

Step 7 Exit the configuration area for the default communication policy in the CLI.

Example:

```
apic1(config-http)# exit
apic1(config-comm-policy)# exit
apic1(config)# exit
apic1#
```

Configuring global throttling using the REST API

Procedure

Step 1 Configure the NGINX rate limit feature through the REST API.

The following configurable properties are added to the communication policy, where:

- `globalThrottleSt` is used to enable or disable the feature
- `globalThrottleRate` is used to set the global throttling rate
- `globalThrottleUnit` is used to set the global throttling unit

```
<type name="RateUnitType" base="string:Basic">
    <allowed name="uname" type="include" regex="[r]/[ms]"/>
</type>

<property name="globalThrottleSt"
    label="Throttle state for all clients without tag0 in header"
    type="AdminState"
    owner="management"
    mod="explicit"
    >
    <default value="disabled"/>
</property>

<property name="globalThrottleRate"
    type="scalar:Uint32"
    owner="management"
    mod="explicit"
    label="The maximum MO api calls allowed per unit time"
    >
    <default value="40"/>
    <range min="1" max="40"/>
</property>

<property name="globalThrottleUnit"
```

```
        type="RateUnitType"
        owner="management"
        mod="explicit"
        label="Unit of rate limit"
        >
        <default value="r/s"/>
    </property>
```

Step 2 To enable the NGINX rate limit feature:

```
POST:
                  {{url}}/api/node/mo/uni/fabric/comm-default/http.xml
BODY:
                 <commHttp globalThrottleSt="enabled"  dn="uni/fabric/comm-default/http"
 globalThrottleRate="1" globalThrottleUnit="r/m" ></commHttp>

 OPTIONS:
              globalThrottleSt: "enabled" or "disabled"
              globalThrottleRate: "1" to "40"
              globalThrottleUnit="r/s" or "r/m"
```

The rate can be configured using a range of 1 to 40, which could be rate per second or rate per minute.

Step 3 To disable the NGINX rate limit feature:

```
POST:
                  {{url}}/api/node/mo/uni/fabric/comm-default/http.xml
BODY:
                 <commHttp globalThrottleSt="disabled"  dn="uni/fabric/comm-default/http"
  globalThrottleRate="1" globalThrottleUnit="r/m" ></commHttp>
```

Configure custom throttle group using the REST API

Use this procedure to configure a custom throttle group. The configurable properties are:

- adminST: enables the custom throttle groups. Only the groups for which the tag is configured are enabled. Each group can be assigned only one tag.
- tag: tag value to be used for the HTTP Header "Ratelimit-Tag". API requests with such a tag are throttled with the configured rate.

  You can configure a maximum of three tags for custom throttle groups, one for each group. For example, "tag1" and "ratePerSec1" are for custom group 1. If the value for "tagX" is left empty, "ratePerSecX" is not activated.
- ratePerSec: throttle rate (requests per second).

You can either use the `.json` or `.xml` files.

Setting the custom rate limit using JSON:

```
POST - {{apic}}/api/mo/uni/fabric/comm-default/https/customRl.json
BODY -
{
    "commCustomRl": {
        "attributes": {
         "adminSt":"enabled",
         "tag1":"<tag1>"
         "ratePerSec1":"<rate>"
         "tag2":<tag2>"
         "ratePerSec2":"<rate>"
         "tag3":<tag3>"
         "ratePerSec3":"<rate>"
 }
    }
}
```

Setting the custom rate limit using XML:

```
POST - {{apic}}/api/mo/uni/fabric/comm-default/https/customRl.xml
BODY -
<commCustomRl
 adminSt="enabled"
 tag1="<tag>"
 ratePerSec1="<rate>"
 tag2="<tag>"
 ratePerSec2="<rate>"
/>
```

THE SPECIFICATIONS AND INFORMATION REGARDING THE PRODUCTS REFERENCED IN THIS DOCUMENTATION ARE SUBJECT TO CHANGE WITHOUT NOTICE. EXCEPT AS MAY OTHERWISE BE AGREED BY CISCO IN WRITING, ALL STATEMENTS, INFORMATION, AND RECOMMENDATIONS IN THIS DOCUMENTATION ARE PRESENTED WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED.

The Cisco End User License Agreement and any supplemental license terms govern your use of any Cisco software, including this product documentation, and are located at: https://www.cisco.com/c/en/us/about/legal/cloud-and-software/software-terms.html. Cisco product warranty information is available at https://www.cisco.com/c/en/us/products/warranty-listing.html. US Federal Communications Commission Notices are found here https://www.cisco.com/c/en/us/products/us-fcc-notice.html.

IN NO EVENT SHALL CISCO OR ITS SUPPLIERS BE LIABLE FOR ANY INDIRECT, SPECIAL, CONSEQUENTIAL, OR INCIDENTAL DAMAGES, INCLUDING, WITHOUT LIMITATION, LOST PROFITS OR LOSS OR DAMAGE TO DATA ARISING OUT OF THE USE OR INABILITY TO USE THIS MANUAL, EVEN IF CISCO OR ITS SUPPLIERS HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Any products and features described herein as in development or available at a future date remain in varying stages of development and will be offered on a when-and if-available basis. Any such product or feature roadmaps are subject to change at the sole discretion of Cisco and Cisco will have no liability for delay in the delivery or failure to deliver any products or feature roadmap items that may be set forth in this document.

Any Internet Protocol (IP) addresses and phone numbers used in this document are not intended to be actual addresses and phone numbers. Any examples, command display output, network topology diagrams, and other figures included in the document are shown for illustrative purposes only. Any use of actual IP addresses or phone numbers in illustrative content is unintentional and coincidental.

The documentation set for this product strives to use bias-free language. For the purposes of this documentation set, bias-free is defined as language that does not imply discrimination based on age, disability, gender, racial identity, ethnic identity, sexual orientation, socioeconomic status, and intersectionality. Exceptions may be present in the documentation due to language that is hardcoded in the user interfaces of the product software, language used based on RFP documentation, or language that is used by a referenced third-party product.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA 95134-1706
USA

Asia Pacific Headquarters
CiscoSystems(USA)Pte.Ltd.
Singapore

Europe Headquarters
CiscoSystemsInternationalBV
Amsterdam,TheNetherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.