# Solución de problemas de la función de aceleración AAA

## Contenido



## Introducción

Este documento describe la función Throttling of AAA (RADIUS) Records que admite la limitación del acceso (autenticación y autorización) y los registros de contabilidad que se envían al servidor RADIUS.

Esta función permite que un usuario configure la velocidad de regulación adecuada para evitar la congestión e inestabilidad de la red cuando no hay ancho de banda suficiente para acomodar una ráfaga repentina de registros generados desde el router Cisco al servidor RADIUS.

## Prerequisites

### Requirements

No hay requisitos específicos para este documento.

### Componentes Utilizados

La información en este documento se basa en la plataforma ASR5k.

The information in this document was created from the devices in a specific lab environment. All of the devices used in this document started with a cleared (default) configuration. If your network is live, make sure that you understand the potential impact of any command.

## Antecedentes

Cuando un administrador envía los mensajes radius al servidor RADIUS a una velocidad alta (por ejemplo, cuando se interrumpe un gran número de sesiones al mismo tiempo, se generan mensajes de detención de contabilización para todas las sesiones al mismo tiempo), es posible que el servidor RADIUS no pueda recibir los mensajes a velocidades tan altas. Para manejar esta

condición necesitamos un mecanismo de control de velocidad efectivo en aaamgr, de modo que aamgr envíe mensajes a una velocidad óptima de tal manera que el servidor RADIUS sea capaz de recibir todos los mensajes y se asegure de que no se descarte ningún mensaje debido a la sobrecarga en el servidor RADIUS.

## Mecanismo de trabajo

Cuando un administrador envía mensajes a la velocidad configurada al servidor RADIUS, envía mensajes de manera uniforme a través de
cada segundo en lugar de enviar todos los mensajes en una sola ráfaga. Dependiendo de la configuración, cada segundo se divide en varias ranuras de tiempo iguales (con un período de tiempo específico por ranura). El período de tiempo mínimo de una ranura
puede ser de 50 milisegundos.

La velocidad debe configurarse teniendo en cuenta

- La velocidad de las llamadas entrantes,
- Número de instancias de aaamgr
- La velocidad a la que el servidor RADIUS puede recibir los mensajes y
- Intervalo de intervalos (para la configuración de contabilidad)
- Algoritmo utilizado para la selección del servidor

Si el valor configurado para los servidores de autenticación es demasiado bajo, habrá un cuello de botella que llevará a
congestión, que puede provocar que las llamadas se pierdan debido al tiempo de espera de configuración de la sesión. Si se configura un valor bajo para los servidores de contabilidad, se observará una gran cantidad de depuración de mensajes de contabilidad, debido al desbordamiento de la cola.

Cuando se configura la función, el número de ranuras de tiempo en un segundo período de tiempo de un segundo se computa y almacena en el nivel de radio. Cuando un mensaje está listo para ser enviado al servidor RADIUS, se verifica si se ha alcanzado la cuota (número de mensajes para esta ranura de tiempo). Si no se alcanza el límite, se envía el mensaje, si es así, entonces el mensaje se coloca en la cola de nivel de servidor para ser enviado en futuras ranuras de tiempo. Cada servidor RADIUS contiene detalles sobre el número de mensajes enviados en la ranura de tiempo actual y la hora en la que caduca la ranura de tiempo. Cuando los mensajes en cola se eligen de la cola de nivel de servidor, se colocan en la cabecera de la cola de nivel de instancia, lo que garantiza la preferencia por los mensajes más antiguos que cualquier otro mensaje nuevo. Los mensajes de la cola de nivel de instancia se seleccionan para el servicio.

## colas AAAMGR

Hay dos tipos de colas en AAAMGR para mensajes:

1. Colas de nivel de instancia
2. Colas de nivel de servidor

Cuando se genera un mensaje, se coloca inicialmente en la cola de nivel de instancia para la prestación de servicios.

La cola de nivel de instancia se procesa durante 25 milisegundos cada 50 milisegundos. Cualquier mensaje que se elimine de la cola de nivel de instancia se intentará enviar al servidor RADIUS. En algunas condiciones, es posible que no podamos enviar los mensajes (sin ancho de banda disponible o sin ID disponibles). En estos casos, los mensajes que fallaron en el intento se pondrán en cola en las colas de nivel de servidor. Por cada 50 milisegundos puede seleccionar tantos mensajes que tengan ID disponibles y también ancho de banda disponibles y colocarlos en la cabecera de la cola de nivel de instancia (estos mensajes son más antiguos que cualquier otro mensaje que esté presente en la cola de nivel de instancia).

Cuando hay un control de velocidad para los mensajes de contabilización, y si hay muchos mensajes de contabilización en la cola de nivel de instancia, cualquier mensaje de autenticación nuevo va a la cola de la cola de nivel de instancia. Para que se procese, debe esperar a que todo el mensaje de contabilización (anterior al nuevo mensaje de autenticación) se envíe al servidor RADIUS o se mueva a la cola de nivel de servidor. Se trata de un comportamiento existente y no se modifica. Por lo tanto, puede causar un pequeño retraso para que se procese el nuevo mensaje de autenticación.

**Ejemplo:**

Basándose en la velocidad máxima con un valor de 5, puede enviar cinco mensajes en 1 segundo y tener 256 mensajes de autenticación radius pendientes (configuración máx. pendiente predeterminada) sin responder por aamgr hacia el servidor de autenticación Radius. En caso de que haya más de 5 mensajes, en 1 segundo los mensajes se ponen en cola hasta que el servidor AAA responda a las solicitudes existentes.

En caso de que alcance 256 mensajes de autenticación de RADIUS enviados desde un administrador hacia el servidor, las solicitudes restantes se pondrán en cola hasta que el servidor AAA responda a las solicitudes existentes. Volverá a entrar en la misma cola que la velocidad máxima. El mensaje se recoge de la cola sólo cuando tiene una ranura libre. La ranura libre entra cuando recibe una respuesta para el mensaje o cuando se agota el tiempo de espera.

# Limitaciones

Dado que Cisco ASR5K es un sistema distribuido con pares de sessmgr/aamgr independientes que procesan las llamadas, la limitación de velocidad se puede implementar solamente para instancias de aamgr independientes. Es teórico ampliar la velocidad de una única instancia a todo el cuadro Cisco ASR5K en su totalidad multiplicando el número total de instancias con la velocidad máxima
de cada instancia.

Este número es sólo el límite máximo absoluto en un escenario de día soleado. No puede tratar a Cisco ASR5K como una caja negra y no puede suponer que todas las llamadas deberían tener éxito si el valor calculado que se ve en el sistema no supera el límite superior.

La velocidad máxima de radio está vinculada con otros parámetros internos y externos relacionados con el sistema. Consulte el impacto esperado si no se cumple una de las condiciones.

| Condiciones | Impacto si no se cumple |
| --- | --- |
| Distribución uniforme de llamadas de demuxmgr a todos los sessmgrs | Si la distribución de la llamada no es uniforme, los mensa radius pueden<br>estar en cola para algunas instancias. Por lo tanto, aunqu |

| | |
|---|---|
| | se alcance el límite teórico de velocidad máxima, se elimi las llamadas para las instancias en las que se ponen en c los mensajes. |
| Distribución uniforme de IMSI (esto es justo en el caso de contabilidad de mediación de ordenamiento) | El ordenamiento cíclico basado en la contabilidad de mediación se basa en el routing basado en IMSI. En este caso, según la distribución IMSI, es posible que s prefiera un conjunto de servidores en lugar de otros segú lógica de ruteo; es posible que la cola esté integrada para aquellos servidores que conduzcan a la llamada descarta |
| No hay ráfagas repentinas de llamadas entrantes | Si hay una ráfaga de llamadas nuevas, los mensajes radi recién generados se pondrán en cola en el sistema. Para momento en que se procesen las nuevas solicitudes de RADIUS. Es posible que el tiempo de configuración de la sesión haya caducado, lo que dará lugar a caídas de llam |
| Los servidores Radius deben responder a tiempo | Cuando el radio solicita un tiempo de espera agotado deb problemas del servidor, volverá a haber acumulación de c porque las nuevas solicitudes no se enviarán a menos qu elimine del sistema la que espera una respuesta. La veloc a la que se eliminarán los mensajes de tiempo de espera sistema también depende de las configuraciones de tiemp espera máximo y de tiempo de espera. |

En muchos casos, podemos ver que las solicitudes de acceso no son procesadas por todas las tareas de administración activas. Esto significa que estamos teniendo una distribución desigual de llamadas dentro de las tareas de sessmgr y más adelante, no todas las instancias de aamgr están involucradas en el procesamiento de llamadas.

La distribución de llamadas no se basa en el mecanismo estricto de ordenamiento cíclico, es decir, si hay 10 llamadas entrantes, se dirigirán a 10 sessmgrs en un algoritmo monotónico.

La distribución de llamadas se basa en estos cuatro factores principales

- active_session_count
- cpu_load
- Round_trip_delay (demuxmgr - sessmgr - demuxmgr)
- pending_add_request (demux to sessmgr)

Esta es la implementación actual. La velocidad máxima es sólo un límite superior, pero debido a la naturaleza distribuida de nuestra arquitectura, no puede extrapolarla directamente a la carga del chasis. El comportamiento depende de la carga de un AAAmgr dado en un momento dado.

La cola de velocidad máxima de RADIUS debe utilizarse para monitorear el estado del sistema. Si hay una acumulación de cola, significa que una de estas 4 condiciones (consulte la tabla) no se cumple y debe identificar la causa raíz para la misma.

**el umbral de cola de velocidad máxima se puede implementar y monitorear constantemente.