# Fehlerbehebung bei AAA-Throttling-Funktion

## Inhalt



## Einführung

In diesem Dokument wird die Funktion zur Drosselung von AAA-Protokollen (RADIUS) beschrieben, die die Drosselung des Zugriffs (Authentifizierung und Autorisierung) sowie der an den RADIUS-Server gesendeten Accounting-Datensätze unterstützt.

Mit dieser Funktion kann ein Benutzer die entsprechende Drosselungsrate konfigurieren, um Netzwerküberlastungen und Instabilitäten zu vermeiden, wenn die Bandbreite nicht ausreicht, um eine plötzliche Anhäufung von Datensätzen zu bewältigen, die vom Cisco Router zum RADIUS-Server generiert wurden.

## Voraussetzungen

### Anforderungen

Für dieses Dokument bestehen keine speziellen Anforderungen.

### Verwendete Komponenten

Die Informationen in diesem Dokument basieren auf der ASR5k-Plattform.

Die Informationen in diesem Dokument wurden von den Geräten in einer bestimmten Laborumgebung erstellt. Alle in diesem Dokument verwendeten Geräte haben mit einer leeren (Standard-)Konfiguration begonnen. Wenn Ihr Netzwerk in Betrieb ist, stellen Sie sicher, dass Sie die potenziellen Auswirkungen eines Befehls verstehen.

## Hintergrundinformationen

Wenn ein Administrator die RADIUS-Meldungen mit hoher Geschwindigkeit an den RADIUS-Server sendet (z. B. wenn eine große Anzahl von Sitzungen gleichzeitig abgeschaltet wird, werden für alle Sitzungen gleichzeitig Abrechnungsmeldungen generiert), kann der RADIUS-Server die Meldungen möglicherweise nicht mit so hohen Raten empfangen. Um diese Bedingung

zu erfüllen, benötigen wir einen effektiven Übertragungsratenkontrollmechanismus, der eine optimale Geschwindigkeit für die Nachrichtenübertragung sicherstellt, dass der RADIUS-Server alle Nachrichten empfangen kann und sicherstellt, dass keine Nachrichten aufgrund von Überlastung auf dem RADIUS-Server verworfen werden.

## Arbeitsmechanismus

Wenn ein Administrator Nachrichten mit der konfigurierten Geschwindigkeit an den RADIUS-Server sendet, werden Nachrichten gleichmäßig verteilt gesendet
jede Sekunde, anstatt alle Nachrichten in einem einzelnen Burst zu versenden. Je nach Konfiguration wird jede Sekunde in mehrere Zeitschlitze zur gleichen Zeit unterteilt (mit einer bestimmten Zeitspanne pro Steckplatz). Die Mindestdauer eines Steckplatzes
kann 50 Millisekunden sein.

Die Rate muss unter Berücksichtigung dieser konfiguriert werden.

- Die Rate eingehender Anrufe,
- Anzahl der AMGR-Instanzen
- Die Geschwindigkeit, mit der der RADIUS-Server die Meldungen und
- Intervall der Zeitgeber (für die Buchhaltungskonfiguration)
- Algorithmus für die Serverauswahl

Wenn der konfigurierte Wert für Authentifizierungsserver zu niedrig ist, gibt es einen Flaschenhals, der zu
Engpässe, die dazu führen können, dass Anrufe aufgrund der Zeitüberschreitung bei der Sitzungseinrichtung verworfen werden. Wenn für Accounting-Server ein niedriger Wert konfiguriert ist, wird aufgrund des Überlaufs der Warteschlange eine große Anzahl von Nachrichten gelöscht.

Wenn die Funktion konfiguriert ist, wird die Anzahl der Zeitschlitze in einer Sekunde und einem Zeitraum in einer Sekunde berechnet und auf Radius-Ebene gespeichert. Wenn eine Nachricht an den RADIUS-Server gesendet werden kann, wird überprüft, ob das Kontingent (die Anzahl der Nachrichten für diesen Zeitschlitz) erreicht wurde. Wenn das Limit nicht erreicht ist, wird die Nachricht gesendet, wenn dies der Fall ist, dann wird die Nachricht in der Warteschlange auf Serverebene in Warteschlange gestellt, die in zukünftigen Zeitsteckplätzen gesendet werden soll. Jeder RADIUS-Server enthält Details zur Anzahl der Nachrichten, die im aktuellen Zeitschlitz gesendet wurden, sowie zum Zeitpunkt, zu dem der Zeitschlitz abläuft. Wenn die in die Warteschlange gestellten Meldungen aus der Warteschlange auf Serverebene ausgewählt werden, werden sie in die oberste Warteschlange auf Instanzebene gestellt, sodass ältere Nachrichten als andere neue Nachrichten bevorzugt werden. Nachrichten aus der Warteschlange für die Instanzebene werden für die Wartung ausgewählt.

## AAAMGR-Warteschlangen

Es gibt zwei Arten von Warteschlangen bei AAAMGR für Nachrichten:

1. Warteschlangen auf Instanzebene
2. Warteschlangen auf Serverebene

Wenn eine Nachricht generiert wird, wird sie zunächst in der Warteschlange für die Instanzenebene für die Wartung in die Warteschlange gestellt.

Die Warteschlange auf Instanzebene wird für 25 Millisekunden alle 50 Millisekunden verarbeitet. Alle Nachrichten, die aus der Warteschlange auf Instanzebene dewartetet werden, werden an den RADIUS-Server gesendet. Unter bestimmten Umständen können wir die Nachrichten nicht senden (keine verfügbare Bandbreite oder keine verfügbaren IDs). In solchen Fällen werden die Meldungen, die den Versuch nicht bestanden haben, in die Warteschlange auf Serverebene gestellt. Für jede 50 Millisekunde wählen Sie so viele Nachrichten aus, die IDs haben und auch über eine Bandbreite verfügen, und stellen sie an den Kopf der Warteschlange auf Instanzebene (diese Nachrichten sind älter als jede andere Nachricht, die in der Warteschlange auf Instanzebene vorhanden ist).

Wenn es eine Ratenkontrolle für Accounting-Meldungen gibt und wenn sich in der Warteschlange für Instanzenebene viele Accounting-Meldungen befinden, wird jede neue Authentifizierungsmeldung an den Tail der Warteschlange für Instanzenebene gesendet. Um verarbeitet zu werden, muss die Accounting-Nachricht (vor der neuen Authentifizierungsmeldung) entweder an den RADIUS-Server gesendet oder in die Warteschlange auf Serverebene verschoben werden. Es handelt sich um ein vorhandenes Verhalten, das nicht geändert wird. So kann es zu einer kleinen Verzögerung kommen, bis die neue Authentifizierungsnachricht verarbeitet wird.

### Beispiel

Basierend auf der maximalen Rate von 5 Nachrichten können Sie fünf Nachrichten in einer Sekunde senden und haben ausstehende 256 RADIUS-Authentifizierungsmeldungen (max-ausstehende Standardkonfiguration) pro AMAGR zum Radius-Authentifizierungsserver unbeantwortet. Wenn mehr als 5 Nachrichten vorhanden sind, werden die Nachrichten in einer Sekunde in die Warteschlange gestellt, bis der AAA-Server auf die bestehenden Anfragen reagiert.

Wenn Sie 256 RADIUS-Authentifizierungsmeldungen erreichen, die von einem Administrator an den Server gesendet werden, werden die verbleibenden Anforderungen in die Warteschlange gestellt, bis der AAA-Server auf die bestehenden Anforderungen reagiert. Sie wird wieder in dieselbe Warteschlange gestellt wie die maximale Rate. Die Nachricht wird nur dann aus der Warteschlange übernommen, wenn Sie einen freien Steckplatz haben. Der freie Steckplatz wird angezeigt, wenn Sie eine Antwort für die Nachricht erhalten oder eine Zeitüberschreitung feststellen.

## Einschränkungen

Da es sich bei dem Cisco ASR5K um ein verteiltes System mit unabhängigen Sessmgr/Aamgr-Paaren handelt, die die Anrufe verarbeiten, konnte die Ratendrosselung nur für unabhängige Instanzen von Amateure implementiert werden. Es ist theoretisch erforderlich, die Rate einer einzelnen Instanz auf die gesamte Cisco ASR5K-Box zu erweitern, indem die Gesamtzahl der Instanzen mit der maximalen Rate multipliziert wird.
der einzelnen Instanzen.

Diese Zahl ist nur die absolute Obergrenze in einem sonnigen Tagesszenario. Sie können Cisco ASR5K nicht als Blackbox behandeln und können nicht davon ausgehen, dass alle Anrufe erfolgreich sein sollten, wenn der im System angezeigte berechnete Wert die Obergrenze nicht überschreitet.

Die maximale Radius-Rate wird mit anderen internen und externen Parametern im Zusammenhang mit dem System verknüpft. Bitte sehen Sie die erwarteten Auswirkungen, wenn

eine der Bedingungen nicht erfüllt ist.

| Bedingungen | Auswirkungen, wenn sie nicht erfüllt werden |
|---|---|
| Einheitliche Verteilung von Anrufen von demuxmgr auf alle Sitzungen | Wenn die Anrufverteilung nicht einheitlich ist, können Rad Meldungen für einige Instanzen in die Warteschlange gestellt werden Obwohl die theoretische maximale Durchsatzrate nicht er wird, werden Anrufe bei Instanzen, in denen Nachrichten Warteschlange gestellt werden, verworfen. Mediation Accounting Round-Robin basiert auf IMSI-basie Routing. |
| Einheitliche Verteilung von IMSIs (dies ist der Fall der Round-Robin-Mediation Accounting) | In diesem Fall können, basierend auf der IMSI-Verteilung, einige Server basierend auf der Routing-Logik gegenüber anderen bevorzugt werden. Für diese Server kann eine Warteschlange aufgebaut werden, die zu einem Abbruch Anrufe führt. |
| Keine plötzlichen Anrufspitzen bei der Anmeldung | Wenn es zu einer Anrufübernahme kommt, werden die ne generierten RADIUS-Meldungen erneut in die Warteschla des Systems gestellt. Zum Zeitpunkt der Verarbeitung der neuen RADIUS-Anfragen. Die Sitzungseinrichtungszeit ka abgelaufen sein, was zu einem Anrufverfall führt. |
| Radius-Server sollten rechtzeitig reagieren | Wenn RADIUS-Anfragen wegen Serverproblemen ein Tim erfordern, wird erneut eine Warteschlangeneinrichtung eingerichtet, da neue Anforderungen nur gesendet werde wenn die aktuelle Anforderung, die eine Antwort erwartet, dem System entfernt wird. Die Geschwindigkeit, mit der d Timeout-Meldungen aus dem System entfernt werden, hä auch von den Konfigurationen für max. ausstehende und Timeout ab. |

In vielen Fällen können wir feststellen, dass Zugriffsanfragen nicht von allen aktiven Verwaltungsaufgaben verarbeitet werden. Das bedeutet, dass wir eine ungleichmäßige Anrufverteilung innerhalb der sessmgr-Aufgaben haben und darüber hinaus nicht alle Instanzen von AMAGR an der Anrufverarbeitung beteiligt sind.

Die Anrufverteilung basiert nicht auf dem strikten Round-Robin-Mechanismus, d. h. wenn 10 eingehende Anrufe eingehen, werden diese in einem monotonen Algorithmus an 10 Sitzungen weitergeleitet.

Die Anrufverteilung basiert auf den folgenden vier Hauptfaktoren.

- **Active_Session_Count**
- **cpu_load**
- **Round_trip_delay** (demuxmgr - sessmgr - demuxmgr)
- **ausstehend_add_request** (demux an sessmgr)

Dies ist die aktuelle Implementierung. Die maximale Rate ist nur eine Obergrenze, aber aufgrund der verteilten Architektur können Sie sie nicht direkt auf die Chassis-Last hochrechnen. Das Verhalten hängt von der Last eines bestimmten AAAmgr ab.
zu einem bestimmten Zeitpunkt.

Zur **Überwachung** des **Status** des Systems sollte eine Radius-Warteschlange mit max. Rate verwendet werden. Wenn eine **Warteschlangeneinrichtung** vorhanden ist,

Dies bedeutet, dass eine dieser vier Bedingungen (siehe Tabelle) nicht erfüllt ist und Sie die Ursache für diese Bedingung identifizieren müssen.

**Schwellenwert für Warteschlangen mit max. Rate könnte implementiert und kontinuierlich überwacht werden.