Risoluzione dei problemi relativi alla funzionalità AAA Throttling

Sommario

Introduzione

In questo documento viene descritta la funzionalità Limitazione dei record AAA (RADIUS) che supporta la limitazione dell'accesso (autenticazione e autorizzazione) e dei record di accounting inviati al server RADIUS.

Questa funzionalità consente a un utente di configurare la velocità appropriata in modo da evitare congestione e instabilità della rete quando la larghezza di banda non è sufficiente per gestire un improvviso aumento di record generati dal router Cisco al server RADIUS.

Prerequisiti

Requisiti

Nessun requisito specifico previsto per questo documento.

Componenti usati

Le informazioni fornite in questo documento si basano sulla piattaforma ASR5k.

Le informazioni discusse in questo documento fanno riferimento a dispositivi usati in uno specifico ambiente di emulazione. Su tutti i dispositivi menzionati nel documento la configurazione è stata ripristinata ai valori predefiniti. Se la rete è operativa, valutare attentamente eventuali conseguenze derivanti dall'uso dei comandi.

Premesse

Quando amgr invia i messaggi RADIUS al server RADIUS a una velocità elevata, ad esempio quando un numero elevato di sessioni si interrompe contemporaneamente, vengono generati contemporaneamente messaggi di interruzione dell'accounting per tutte le sessioni, il server

RADIUS potrebbe non essere in grado di ricevere i messaggi a velocità così elevate. Per gestire questa condizione è necessario un meccanismo di controllo della velocità efficace in aamgr, in modo che aamgr invii i messaggi a una velocità ottimale in modo che il server RADIUS sia in grado di ricevere tutti i messaggi e garantisca che nessun messaggio venga scartato a causa di un sovraccarico sul server RADIUS.

Meccanismo di lavoro

Quando amgr invia messaggi alla velocità configurata al server RADIUS, questi vengono inviati in modo uniforme in
ogni secondo, invece di inviare tutti i messaggi in una singola sequenza burst. A seconda della configurazione, ogni secondo viene suddiviso in più slot di tempo uguali (con un periodo di tempo specifico per slot). Il periodo di tempo minimo di uno slot
potrebbe essere di 50 millisecondi.

La velocità deve essere configurata tenendo conto di queste

- Frequenza delle chiamate in arrivo,
- Numero di istanze di gestione
- Velocità alla quale il server RADIUS può ricevere i messaggi e
- Intervallo di intervalli (per la configurazione di accounting)
- Algoritmo utilizzato per la selezione del server

Se il valore configurato per i server di autenticazione è troppo basso, si verificherà un errore irreversibile che causerà
congestione, che potrebbe causare l'interruzione delle chiamate a causa del timeout di configurazione della sessione. Se per i server di accounting è configurato un valore basso, verrà rilevata una notevole rimozione dei messaggi di accounting a causa dell'overflow della coda.

Quando la feature è configurata, il numero di intervalli di tempo in un secondo e il periodo di tempo in un secondo vengono calcolati e memorizzati a livello di raggio. Quando un messaggio è pronto per essere inviato al server RADIUS, viene verificato se è stata raggiunta la quota (numero di messaggi per questa fascia oraria). Se il limite non viene raggiunto, il messaggio viene inviato, se lo è, il messaggio viene inserito nella coda a livello di server per essere inviato negli intervalli di tempo futuri. Ogni server RADIUS contiene dettagli sul numero di messaggi inviati nell'intervallo di tempo corrente e sull'ora di scadenza dell'intervallo. Quando i messaggi in coda vengono prelevati dalla coda a livello di server, vengono inseriti nella posizione iniziale della coda a livello di istanza, garantendo una preferenza per i messaggi meno recenti rispetto a qualsiasi altro nuovo messaggio. I messaggi dalla coda a livello di istanza vengono selezionati per la manutenzione.

Code AAMGR

In AAMGR sono disponibili due tipi di code per i messaggi:

1. Code a livello di istanza
2. Code a livello di server

Quando viene generato, un messaggio viene inizialmente inserito nella coda a livello di istanza per la manutenzione.

La coda a livello di istanza viene elaborata per 25 millisecondi ogni 50 millisecondi. Qualsiasi messaggio rimosso dalla coda a livello di istanza verrà inviato al server RADIUS. In alcune condizioni potrebbe non essere possibile inviare i messaggi (larghezza di banda non disponibile o ID non disponibili). In questi casi, i messaggi che non hanno superato il tentativo verranno accodati nelle code a livello di server. Ogni 50 millisecondi si selezionano tutti i messaggi con ID disponibili e larghezza di banda disponibile e li si inserisce all'inizio della coda a livello di istanza (questi messaggi sono più vecchi di qualsiasi altro messaggio presente nella coda a livello di istanza).

Quando è disponibile un controllo della velocità per i messaggi di accounting e se nella coda a livello di istanza sono presenti molti messaggi di accounting, qualsiasi nuovo messaggio di autenticazione viene inserito nella coda a livello di istanza. Per essere elaborato, deve attendere che tutti i messaggi di accounting (che precedono il nuovo messaggio di autenticazione) vengano inviati al server RADIUS o spostati nella coda a livello di server. Si tratta di un comportamento esistente che non viene modificato. Ciò può causare un lieve ritardo nell'elaborazione del nuovo messaggio di autenticazione.

Esempio

In base alla velocità massima con valore 5, è possibile inviare cinque messaggi in 1 secondo e disporre di 256 messaggi di autenticazione radius in attesa (configurazione massima in attesa predefinita) senza risposta per ogni amministratore verso il server di autenticazione Radius. Se sono presenti più di 5 messaggi, in 1 secondo i messaggi vengono messi in coda finché il server AAA non risponde alle richieste esistenti.

Se si raggiungono i 256 messaggi di autenticazione Radius inviati da un amministratore al server, le richieste rimanenti verranno messe in coda finché il server AAA non risponderà alle richieste esistenti. Verrà nuovamente inserita nella stessa coda di quella di max-rate. I messaggi vengono prelevati dalla coda solo quando si dispone di uno slot libero. Lo slot libero entra quando si riceve una risposta per il messaggio o quando scade.

Limitazioni

Poiché Cisco ASR5K è un sistema distribuito con coppie di sessmgr/aamgr indipendenti che elaborano le chiamate, la limitazione della velocità può essere implementata solo per istanze di asamgr indipendenti. In teoria, è possibile estendere la velocità di una singola istanza all'intera casella Cisco ASR5K semplicemente moltiplicando il numero totale di istanze per la velocità massima
di ogni istanza.

Questo numero è solo il limite massimo assoluto in una giornata di sole. Non è possibile trattare Cisco ASR5K come una black box e non si può presumere che tutte le chiamate abbiano esito positivo se il valore calcolato visualizzato nel sistema non supera il limite superiore.

La velocità massima del raggio è legata ad altri parametri interni ed esterni relativi al sistema. Verificare l'impatto previsto se una delle condizioni non viene soddisfatta.

Condizioni	Impatto se non viene soddisfatto
Distribuzione uniforme delle chiamate da demuxmgr a tutti i sessmgr	Se la distribuzione delle chiamate non è uniforme, i messa radius possono per alcune istanze. Pertanto, anche se non viene raggiunt limite massimo teorico della velocità, le chiamate verranno

	interrotte nei casi in cui i messaggi vengono accodati.
Distribuzione uniforme degli IMSI (solo nel caso della contabilità di mediazione round robin)	Round robin di Mediation Accounting basato su routing IM In questo caso, in base alla distribuzione IMSI, alcuni serv possono essere preferiti rispetto ad altri in base alla logica routing, la coda potrebbe essere costruita per quei server portano alla chiamata di drop.
Nessun aumento improvviso di chiamate in arrivo	Se si verifica una frammentazione di nuove chiamate, i messaggi radius appena generati verranno nuovamente accodati nel sistema. Al momento dell'elaborazione delle richieste radius. È possibile che il tempo di impostazione sessione sia scaduto e ciò potrebbe causare interruzioni chiamate.
I server Radius devono rispondere in tempo	Quando si verifica il timeout delle richieste radius a causa problemi del server, si verifica nuovamente la creazione d coda, in quanto le nuove richieste non verranno inviate a che non venga rimossa dal sistema la richiesta corrente c prevede una risposta. La velocità con cui i messaggi di tim verranno rimossi dal sistema dipende anche dalle configurazioni di timeout e di attesa massime.

In molti casi è possibile notare che le richieste di accesso non vengono elaborate da tutte le attività di gestione attive. Ciò significa che la distribuzione delle chiamate non è uniforme all'interno delle attività di sessmgr e, più avanti, non tutte le istanze di gestione sono coinvolte nell'elaborazione delle chiamate.

La distribuzione delle chiamate non si basa sul meccanismo di round robin, ovvero se ci sono 10 chiamate in arrivo andranno a 10 sessioni in un algoritmo monotono.

La distribuzione delle chiamate si basa su questi quattro fattori principali

- conteggio_sessioni_attive
- cpu_carica
- Round_trip_delay (demuxmgr - sessmgr - demuxmgr)
- standing_add_request (demux in sessmgr)

Questa è l'implementazione corrente. La velocità massima è solo un limite superiore, ma a causa della natura distribuita della nostra architettura, non è possibile estrapolarla direttamente al carico dello chassis. Il comportamento dipende dal carico su un determinato AAAmgr in un determinato momento.

Utilizzare la coda velocità massima Radius per monitorare lo stato del sistema. Se è presente una creazione di code,
indica che una delle 4 condizioni (fare riferimento alla tabella) non è soddisfatta ed è necessario identificare la causa principale per la stessa.

**la soglia della coda di velocità massima può essere implementata e monitorata costantemente.