

Limitazione della velocità statica e dinamica pre-vuota con CPS vDRA

Sommario

[Introduzione](#)

[Prerequisiti](#)

[Requisiti](#)

[Componenti usati](#)

[Premesse](#)

[Problema](#)

[Soluzione](#)

[Limite di velocità statico nel servizio di bilanciamento del carico](#)

[Limite di velocità in ingresso](#)

[Limite velocità in uscita](#)

[Limite di velocità dinamica](#)

Introduzione

Questo documento descrive le opzioni di limite di velocità in DRA, un componente di telecomunicazione che instrada i messaggi di Diameter e gestisce il traffico di rete.

Prerequisiti

Requisiti

Cisco raccomanda la conoscenza dei seguenti argomenti:

- Cisco Policy Suite (CPS) Diameter Routing Agent (vDRA)
- Nozioni di base e specifiche dell'agente di routing del diametro

Componenti usati

Le informazioni fornite in questo documento si basano su Cisco Policy Suite DRA.

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

DRA è un componente delle reti di telecomunicazione, in particolare nel contesto delle reti basate sul protocollo del diametro. Il DRA indirizza in modo efficiente i messaggi di Diameter tra diversi elementi di rete, come i server delle policy, i sistemi di ricarica e altri dispositivi abilitati al Diameter. La limitazione della velocità è una tecnica di gestione del traffico di rete utilizzata per controllare la quantità di traffico da o verso un elemento di rete. Garantisce che le risorse di rete non siano esaurite, mantiene la qualità del servizio e previene abusi o abusi della rete.

Problema

Ciascun componente della rete può gestire il carico di traffico in base alla sua capacità nominale, ma in tempo reale possono verificarsi scenari in cui il traffico generato è superiore a quello che il sistema è in grado di gestire. Alcuni di essi sono:

- Comportamento dell'utente: attività come eventi di streaming o aggiornamenti software che generano grandi quantità di dati in un breve periodo. Generalmente inviato da Gateway (Gw) a DRA.
- Congestione della rete: nei periodi di elevato utilizzo della rete, è possibile che si verifichi un aumento della congestione e che i dati in coda vengano quindi inviati in burst quando la capacità diventa disponibile.
- Meccanismi di resilienza della rete: reindirizzamento del traffico durante interruzioni o interventi di manutenzione, che causano picchi temporanei. Ciò può influire sul flusso del traffico nei siti accoppiati che non hanno alcun problema di rete.
- Comportamento degli elementi di rete: in caso di sovraccarico e congestione, è possibile iniziare a visualizzare l'assenza di risposte/timeout da uno o più elementi di rete che possono causare un ulteriore sovraccarico del sistema a causa della riconnessione.
- Scaricamento del gateway: il gateway può svuotare le sessioni esistenti a causa di modifiche delle policy, della topologia o di qualsiasi attività di manutenzione o risoluzione dei problemi. Durante questi scenari le sessioni vengono cancellate ed è possibile ricevere una frammentazione di richieste Gx Credit Control Request (CCR)-T.

Soluzione

DRA può distribuire il carico tra più server Diameter per garantire una gestione efficiente delle richieste ed evitare di sovraccaricare un singolo server. In caso di guasto del server, DRA può reindirizzare i messaggi su server alternativi, garantendo un'elevata disponibilità e affidabilità dei servizi di rete.

Limitazione della velocità sul DRA, non solo protegge il DRA ma anche altre entità garantendo un flusso controllato di messaggi. I vantaggi principali della limitazione delle velocità sono:

- Continuità del servizio: mantenimento della disponibilità continua del servizio garantendo che i componenti critici della rete non siano sovraccarichi e prevengano le interruzioni.
- Scalabilità: consente alla rete di gestire carichi variabili senza compromettere le prestazioni.
- Conformità agli accordi sui livelli di servizio (SLA): garanzia che la rete soddisfi gli SLA mantenendo livelli coerenti di prestazioni e affidabilità.

Limite di velocità statico nel servizio di bilanciamento del carico

Si tratta di un approccio diretto, in cui una soglia fissa viene impostata in base alla capacità nominale di DRA/Packet Gateway (pGW) e altri elementi di rete e non cambia in base alle condizioni di rete o alle risorse di sistema. Limitando il numero di richieste in arrivo si ottiene un risultato prevedibile sulla quantità di traffico elaborata da DRA.

Le configurazioni per il limite di velocità statica dipendono dal caso di utilizzo a cui viene applicato.

Limite di velocità in ingresso

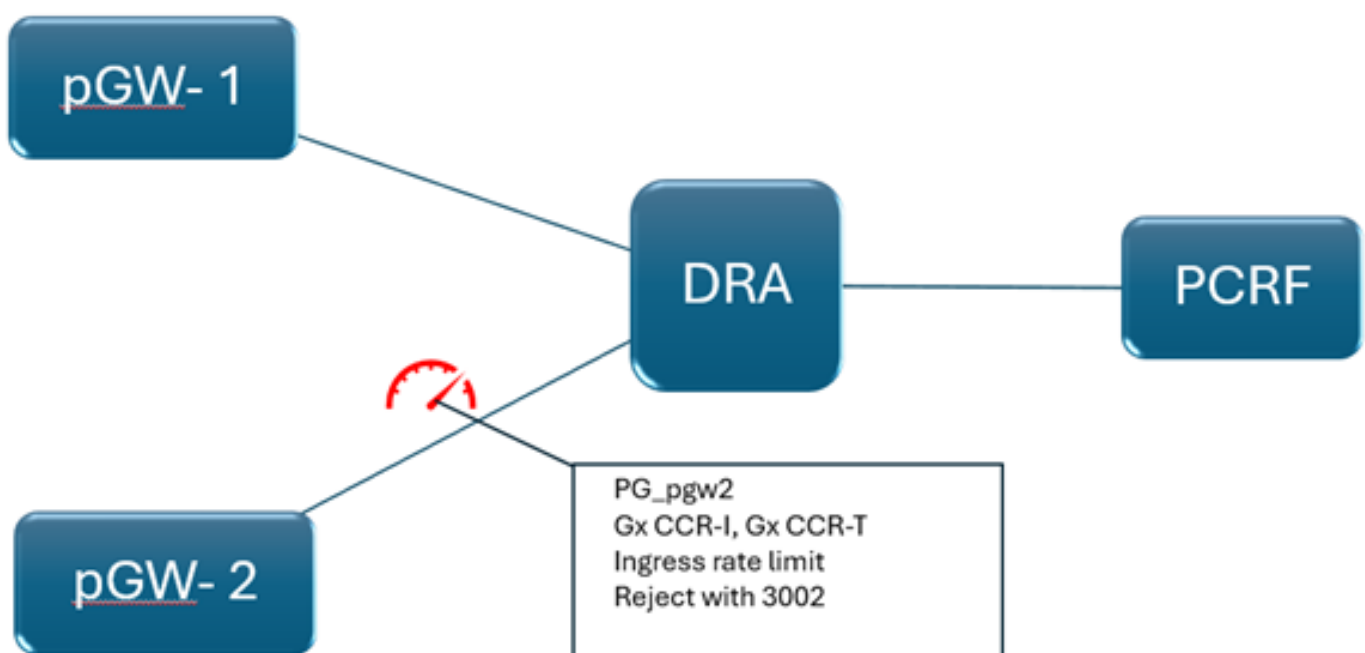
Scenario: Burst da pGW

Vengono configurate soglie specifiche per i pGW suscettibili a questi picchi di traffico. Il valore deve essere stabilito in base ai numeri di traffico regolari/di picco che possono essere rilevati durante i picchi di traffico.

I numeri di soglia possono essere definiti in modo specifico per ciascun tipo di messaggio in modo da garantire che venga limitato solo il traffico burst, ad esempio, devono essere limitate solo le richieste Gx CCR-I e Gx CCR-T provenienti da un GW, ma il traffico Gx CCR-U o Gy viene inoltrato come ricevuto.

In questo caso, è possibile impostare la limitazione sul lato entrata, ovvero DRA limita il messaggio non appena lo riceve, poiché lo scopo è quello di rifiutare in base all'elemento di rete da cui riceve la richiesta e di evitare l'elaborazione di un numero di richieste superiore a quello che DRA può gestire.

Il comportamento della limitazione può consistere nel rifiutare il messaggio con un codice di errore e un messaggio di errore specifico oppure nel eliminarlo.



Questo comportamento può essere abilitato in CPS vDRA configurando le tabelle CRD (Custom Reference Data) 'Profilo limite velocità peer' e 'Profilo limite velocità messaggi'. In queste tabelle CRD, è necessario configurare i seguenti valori:

Gruppo peer	Un gruppo peer è un raggruppamento logico di nodi Diameter basato sul relativo realm e host. È necessario configurare il gruppo peer da limitare.
Nome di dominio completo (FQDN) peer	FQDN (corrispondenza esatta o regex) per i peer nel gruppo peer per cui è necessario determinare il limite.
Direzione messaggio	Direzione della limitazione - In entrata o In uscita. In questo caso, Ingress.
Profilo Rate Limit	Nome profilo limite di velocità messaggi utilizzato per definire il tipo di messaggio da limitare.
Limite di velocità peer	Frequenza di richieste consentite per questo gruppo peer. Sono inclusi tutti i tipi di messaggi di tale gruppo peer.
Comportamento di eliminazione	È possibile scegliere di eliminare la richiesta o di rifiutarla con un codice di errore.
Codice risultato	Valore del codice risultato se si rifiutano i messaggi. Non applicabile nel caso in cui i messaggi vengano eliminati.
Stringa di errore	Stringa di errore utilizzata nel messaggio di risposta della richiesta rifiutata. Non applicabile nel caso in cui il messaggio venga eliminato.
Identificatore applicazione	ID applicazione del messaggio da limitare.
Codice comando	Codice comando del messaggio da limitare.
Tipo messaggio/richiesta	ID applicazione e tipo di richiesta delle richieste da limitare.
Limite di velocità messaggi	TelePresence Server (TPS) della richiesta del tipo di messaggio che deve essere elaborata da DRA. Le richieste

oltre questo TPS sono limitate. Questo valore è per peer nel gruppo peer.

Di seguito è riportato un esempio, in cui si sta configurando un limite di velocità complessivo di 1000 messaggi per pGW2 e un limite di velocità di 200 Gx CCR-I e 200 Gx CCR-T. Qualsiasi richiesta che superi questa velocità viene rifiutata con un 3002 e un messaggio di errore indica che il limite di velocità viene superato.

Peer Rate Limit Profile

Filter by All Visible Columns

CCR_I_T_Limit

Peer Group *	Peer FQDN *	Message Direction *	Rate Limit Profile	Peer Rate Limit	Discard Behavior *	Result Code	Error String	Actions
PG_pGW2	match=peer- xxxxxx	Ingress	CCR_I_T_Limit	1000	Send Error Answer	3002	DRA rate limit breached	 

Showing 1 out of 1

Show 50 rows 1 out of 1

Message Rate Limit Profile

Filter by All Visible Columns

CCR_I_T_Limit

Rate Limit Profile Name *	Application Identifier *	Command Code *	Message/Request Type *	Message Rate Limit *	Actions
CCR_I_T_Limit	16777238	272	1	200	 
CCR_I_T_Limit	16777238	272	3	200	 

Showing 2 out of 2

Show 50 rows 1 out of 1

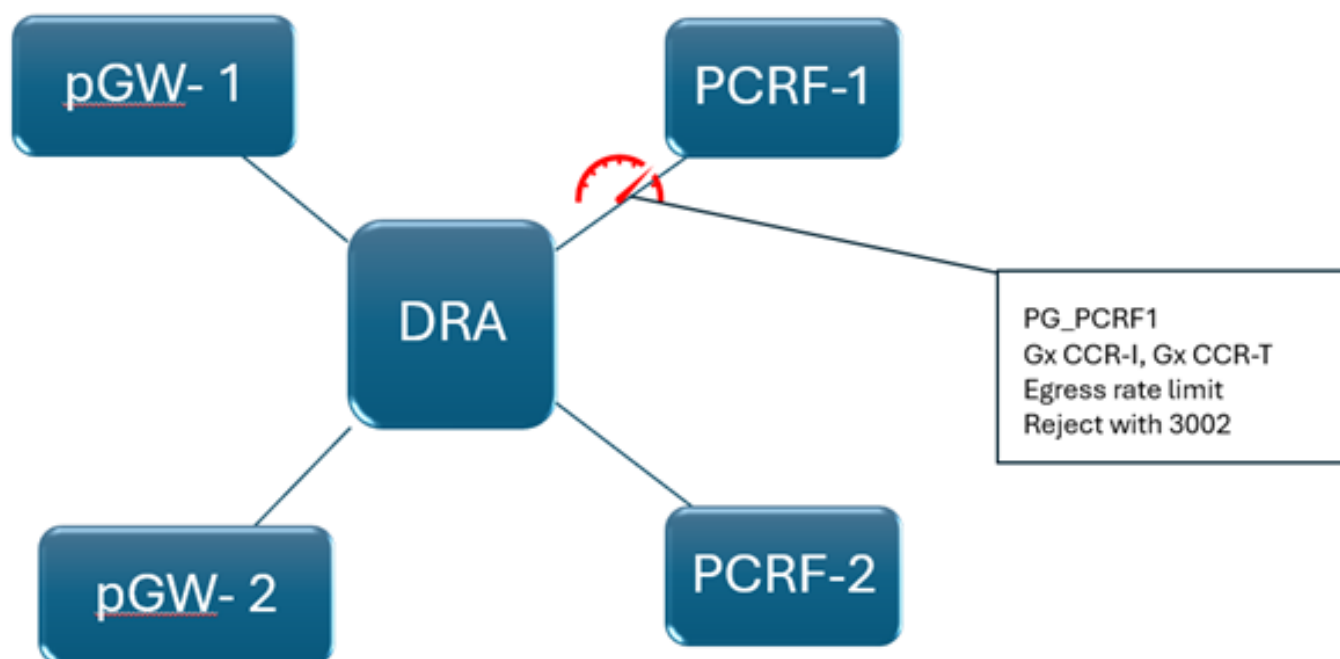
Limite velocità in uscita

Scenario: Protezione dell'elemento di rete che riceve la richiesta

Si consideri ad esempio una transazione Gx in cui la richiesta viene ricevuta da GpW e inoltrata alla funzione PCRF (Policy and Charging Rules Function). Se la quantità di dati che PCRF è in grado di gestire è limitata, anche se DRA è in grado di gestire il traffico in entrata, è possibile utilizzare DRA per limitare il messaggio a DRA invece di inoltrare la richiesta a PCRF e sovraccaricarla.

In questo caso, è necessario eseguire la limitazione sul lato in uscita, ovvero DRA limita il messaggio prima di inoltrarlo al PCRF, in base al PCRF Peer Group identificato in base alla logica di routing DRA.

Il comportamento della limitazione può consistere nel rifiutare il messaggio con un codice di errore e un messaggio di errore specifico oppure nel eliminarlo.



Questo comportamento può essere abilitato in CPS vDRA configurando le tabelle CRD 'Profilo limite velocità peer' e 'Profilo limite velocità messaggio'. Nelle tabelle CRD è necessario configurare i seguenti valori:

Gruppo peer	Un gruppo peer è un raggruppamento logico di nodi Diameter basato sul relativo realm e host. È necessario configurare il gruppo peer da limitare.
FQDN peer	FQDN (corrispondenza esatta o regex) per i peer nel gruppo peer per cui è necessario determinare il limite.
Direzione messaggio	Direzione della limitazione - In entrata o In uscita. In questo caso - Egress.
Profilo Rate Limit	Nome profilo limite di velocità messaggi utilizzato per definire il tipo di messaggio da limitare.
Limite di velocità peer	Frequenza di richieste consentite per questo gruppo peer.

	Sono inclusi tutti i tipi di messaggi di tale gruppo peer.
Comportamento di eliminazione	È possibile scegliere di eliminare la richiesta o di rifiutarla con un codice di errore.
Codice risultato	Valore del codice risultato se si rifiutano i messaggi. Non applicabile nel caso in cui i messaggi vengano eliminati.
Stringa di errore	Stringa di errore utilizzata nel messaggio di risposta della richiesta rifiutata. Non applicabile nel caso in cui il messaggio venga eliminato.
Identificatore applicazione	ID applicazione del messaggio da limitare.
Codice comando	Codice comando del messaggio da limitare.
Tipo messaggio/richiesta	ID applicazione e tipo di richiesta delle richieste da limitare.
Limite di velocità messaggi	TPS della richiesta del tipo di messaggio elaborato da DRA. Le richieste oltre questo TPS sono limitate. Questo valore è per peer nel gruppo peer.

Peer Rate Limit Profile



Filter by All Visible Columns

CCR_I_T



Peer Group *	Peer FQDN *	Message Direction *	Rate Limit Profile	Peer Rate Limit	Discard Behavior *	Result Code	Error String	Actions
PG_PCRF1	match=peer- XXXXXXXXXX	Egress	CCR_I_T_Limit	1000	Send Error Answer	3002	DRA rate limit breached	

Showing 1 out of 1

Show 50 rows 1 out of 1

Message Rate Limit Profile



Filter by All Visible Columns

CCR_I_T_Limit



Rate Limit Profile Name *	Application Identifier *	Command Code *	Message/Request Type *	Message Rate Limit *	Actions
CCR_I_T_Limit	16777238	272	1	200	
CCR_I_T_Limit	16777238	272	3	200	

Showing 2 out of 2

Show 50 rows 1 out of 1

Scenario: Rallentamento della rete che provoca congestione del traffico e un errore completo/parziale di DRA

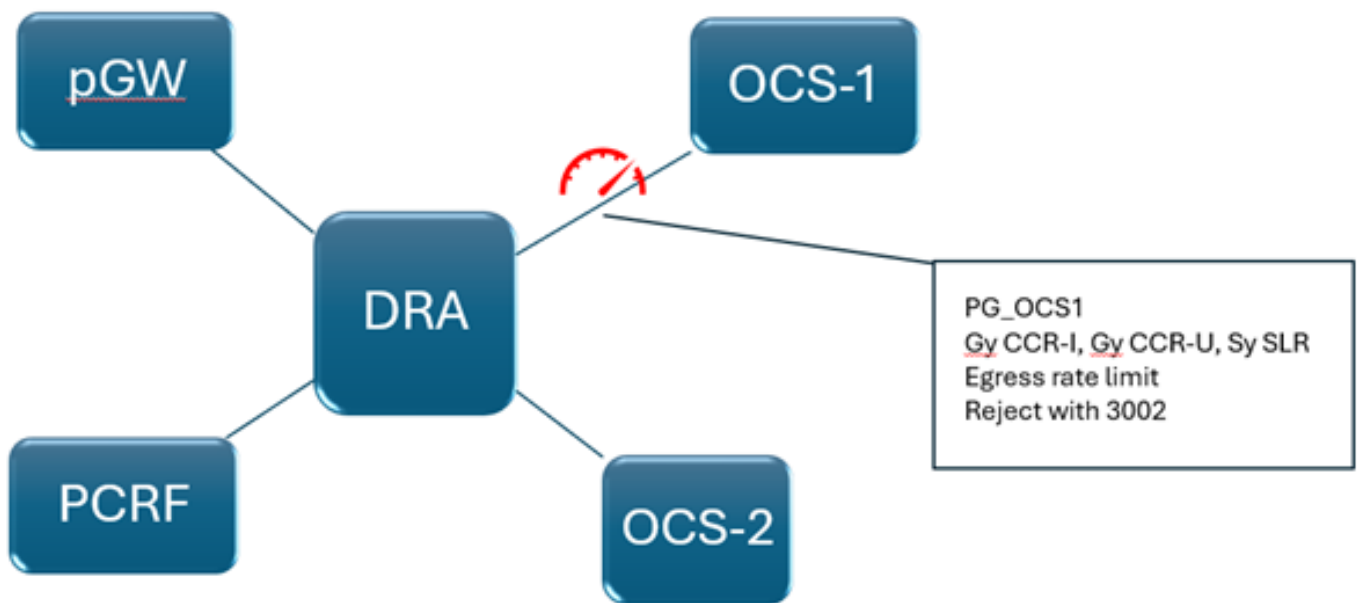
Si consideri ad esempio una transazione Gy scambiata tra pGW e Online Charging System (OCS). In caso di lentezza nella rete sul canale DRA-OCS (a causa di traffico elevato da PGW o di qualsiasi altro problema di rete), la richiesta scade a causa di violazione dello SLA. Questi timeout hanno un impatto non solo su DRA, ma sull'intera rete.

Le risorse DRA vengono bloccate nel tentativo di inviare la richiesta a OCS sulla rete lenta, con conseguente esaurimento delle risorse. Ciò comporta il rifiuto di più richieste da parte di DRA, anche se la capacità nominale di DRA non viene violata.

Questo influisce anche sul traffico che non è sul canale DRA-OCS. I rifiuti, i timeout e le interruzioni causano la riconnessione di più elementi di rete.

In questo caso, è necessario eseguire la limitazione sul lato uscita - DRA limita il messaggio prima di inoltrarlo a OCS, in base al gruppo peer OCS (che ha limitazioni di capacità o problemi di rete).

Il comportamento della limitazione può consistere nel rifiutare il messaggio con un codice di errore e un messaggio di errore specifico oppure nel eliminarlo.



Questo comportamento può essere abilitato in CPS vDRA configurando le tabelle CRD 'Profilo limite velocità peer' e 'Profilo limite velocità messaggio'. Nelle tabelle CRD riportate di seguito è necessario configurare i seguenti valori:

Gruppo peer	Un gruppo peer è un raggruppamento logico di nodi Diameter basato sul relativo realm e host. È necessario configurare il gruppo peer da limitare.
FQDN peer	FQDN (corrispondenza esatta o regex) per i peer nel gruppo peer per cui è necessario determinare il limite.
Direzione messaggio	Direzione della limitazione - In entrata o In uscita. In questo caso - Egress.
Profilo limite di velocità	Nome profilo limite di velocità messaggi utilizzato per definire il tipo di messaggio da limitare.
Limite di velocità peer	Frequenza di richieste consentite per questo gruppo peer. Sono inclusi tutti i tipi di messaggi di tale gruppo peer.
Comportamento di eliminazione	È possibile scegliere di eliminare la richiesta o di rifiutarla con un codice di errore.

Codice risultato	Valore del codice risultato se si rifiutano i messaggi. Non applicabile nel caso in cui i messaggi vengano eliminati.
Stringa di errore	Stringa di errore utilizzata nel messaggio di risposta della richiesta rifiutata. Non applicabile nel caso in cui il messaggio venga eliminato.
Identificatore applicazione	ID applicazione del messaggio da limitare.
Codice comando	Codice comando del messaggio da limitare.
Tipo messaggio/richiesta	ID applicazione e tipo di richiesta delle richieste da limitare.
Limite di velocità messaggi	TPS della richiesta del tipo di messaggio elaborato da DRA. Le richieste oltre questo TPS devono essere limitate. Questo valore è per peer nel gruppo peer.

Peer Rate Limit Profile

Filter by All Visible Columns

gy_sy

Peer Group *	Peer FQDN *	Message Direction *	Rate Limit Profile	Peer Rate Limit	Discard Behavior *	Result Code	Error String	Actions
PG_OCS_1	match=peer- 	Egress	Gy_Sy_Limit	1000	Send Error Answer	3002	DRA rate limit breached	

Showing 1 out of 1

Message Rate Limit Profile

Filter by All Visible Columns

Gy_Sy_Limit

Rate Limit Profile Name *	Application Identifier *	Command Code *	Message/Request Type *	Message Rate Limit *	Actions
Gy_Sy_Limit	16777302	8388635	1	300	
Gy_Sy_Limit	4	272	1	500	

Showing 2 out of 2

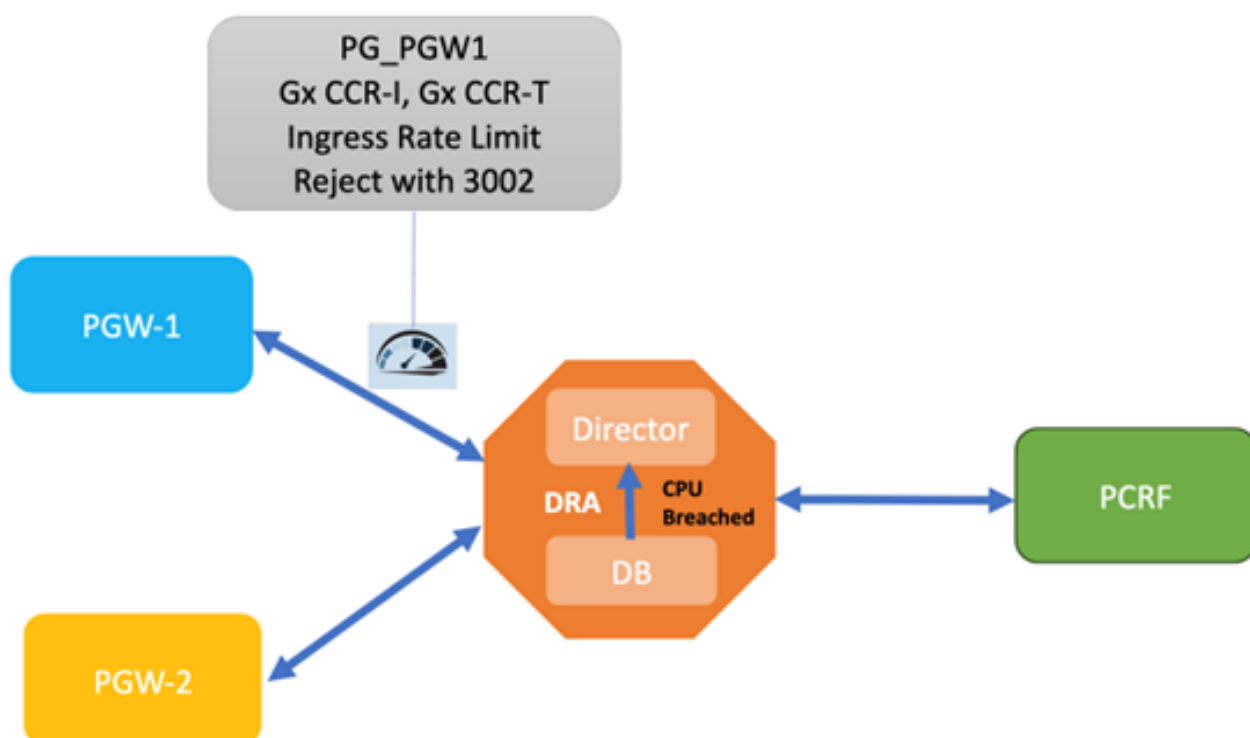
Show 50 rows 1 out of 1

Limite di velocità dinamica

Quando si verificano burst CCR-I o CCR-T, può verificarsi un sovraccarico sul database (DB), che può provocare la destabilizzazione del sistema. Per superare questo problema, DRA supporta la limitazione dinamica della velocità (solo per Gx CCR-I e Gx CCR-T) basata sulla capacità DB disponibile.

DRA controlla l'utilizzo della CPU del database e, ogni volta che la soglia viene superata, limita le richieste in arrivo. È possibile configurare le soglie della CPU per la limitazione e il traffico in ingresso da limitare.

È possibile configurare soglie CPU diverse con percentuali di velocità corrispondenti. DRA regola il livello di limitazione in base all'utilizzo corrente della CPU del database. Quando l'utilizzo della CPU diventa stabile, la limitazione si interrompe gradualmente.



Questo comportamento può essere abilitato in CPS vDRA configurando le tabelle CRD 'Profilo limite velocità peer', 'Profilo limite velocità messaggi', 'Profilo limite velocità peer dinamica' e 'Profilo CPU database limitazione dinamica'. Nelle tabelle CRD riportate di seguito è necessario configurare i seguenti valori:

Gruppo peer	Un gruppo peer è un raggruppamento logico di nodi Diameter basato sul relativo realm e host. In questo esempio viene configurato il gruppo peer del pGW.
-------------	--

FQDN peer	FQDN (corrispondenza esatta o regex) per i peer nel gruppo peer per cui è necessario determinare il limite.
Tipo messaggio/richiesta	ID applicazione e tipo di richiesta delle richieste da limitare. In questo esempio Gx CCR-I, Gx CCR-T.
Direzione messaggio	Direzione della limitazione - In entrata o In uscita. In questo caso, Ingress.
Profilo limite di velocità	Nome profilo limite di velocità messaggi utilizzato per definire il tipo di messaggio da limitare.
Limite di velocità peer	Frequenza di richieste consentite per questo gruppo peer. Sono inclusi tutti i tipi di messaggi di tale gruppo peer.
Comportamento di eliminazione	È possibile scegliere di eliminare la richiesta o di rifiutarla con un codice di errore.
Codice risultato	Valore del codice risultato se si rifiutano i messaggi. Non applicabile nel caso in cui i messaggi vengano eliminati.
Stringa di errore	Stringa di errore utilizzata nel messaggio di risposta della richiesta rifiutata. Non applicabile nel caso in cui il messaggio venga eliminato.
Identificatore applicazione	ID applicazione del messaggio da limitare.
Codice comando	Codice comando del messaggio da limitare.
Limite di velocità messaggi	TPS della richiesta del tipo di messaggio elaborato da DRA. Le richieste oltre questo TPS sono limitate. Questo valore è per peer nel gruppo peer.
Profilo CPU database con limitazione dinamica	Si riferisce al nome del profilo della CPU, utilizzato per definire la percentuale di limitazione per un intervallo di CPU diverso.
Soglia utilizzo CPU	È possibile scegliere il valore corretto dei livelli di CPU

database	configurati come limiti di violazione in base al modello di traffico nella distribuzione.
Percentuale di limitazione	% del limite di velocità applicato quando viene superato il livello di CPU corrispondente.

Peer Rate Limit Profile

Filter by

All Visible Columns

DRA rate

Peer Group *	Peer FQDN *	Message Direction *	Rate Limit Profile	Peer Rate Limit	Discard Behavior *	Result Code	Error String	Actions
PG_pGW_1	match=peer-*	Ingress	CCR_I_T	1000	Send Error Answer	3002	DRA rate limit breached	

Showing 1 out of 1

Message Rate Limit Profile

Filter by

All Visible Columns

CCR_I_T

Rate Limit Profile Name *	Application Identifier *	Command Code *	Message/Request Type *	Message Rate Limit *	Actions
CCR_I_T	16777238	272	1	1000	
CCR_I_T	16777238	272	3	1000	

Showing 2 out of 2

Show

50

rows

1

out of 1

Dynamic Peer Rate Limit Profile

Filter by

All Visible Columns

DynRateLimit

Peer Group *	Peer FQDN *	Dynamic Throttling DB CPU Profile	Actions
PG_pGW_1	*	DynRateLimit	

Showing 1 out of 1

Show

50

rows

1

out of 1

Dynamic Throttling DB CPU Profile



Filter by

All Visible Columns



dyn



CPU Profile Name *	DB CPU Utilization Threshold *	Throttle Percentage	Actions
DynRateLimit	50	20	
DynRateLimit	55	30	
DynRateLimit	60	40	
DynRateLimit	65	50	

Showing 4 out of 4

Show 50 rows



1



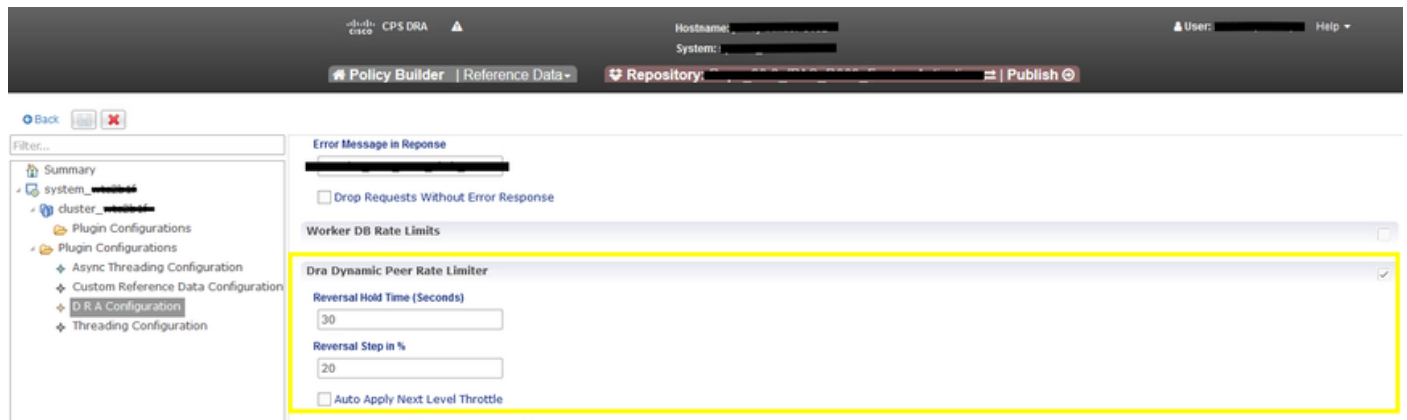
out of 1



È inoltre necessario attivare questo comportamento selezionando la casella di controllo in Policy Builder in 'DRA di configurazione DRA' nella sezione 'DRA Dynamic Peer Rate Limiter'.

Tempo di attesa storno - Periodo di tempo per il quale viene monitorato l'utilizzo della CPU prima dell'applicazione dello storno.

Passo di storno in %: la percentuale di limitazione stornata.



Scenario: Limitazione dinamica della velocità in base all'utilizzo della CPU

Considerare questa configurazione in DRA:

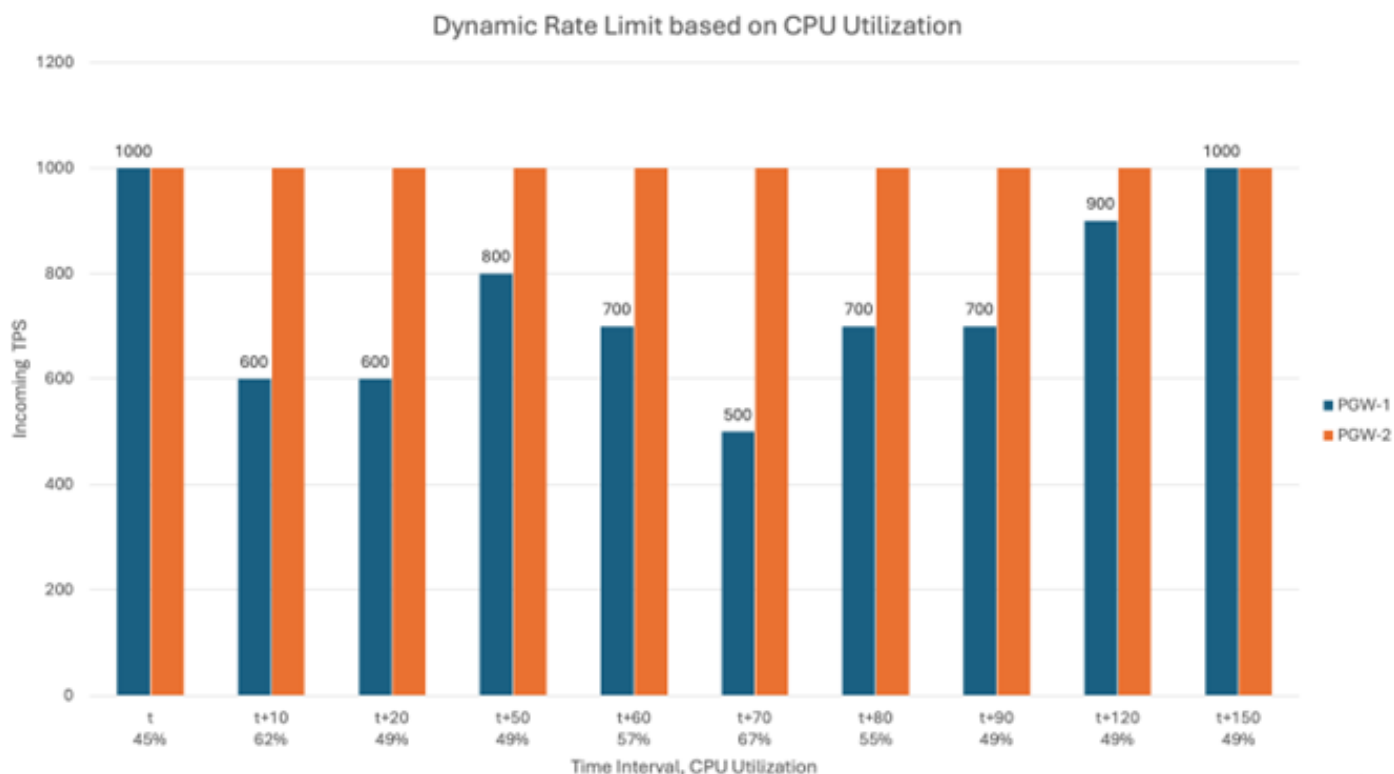
Limite di velocità messaggi statici: 1000 (è quindi il valore di TPS in ingresso)

Tempo di attesa storno: 30

Fase di storno in %: 20%

Ogni volta che l'utilizzo della CPU del database supera la soglia, fa riferimento alla configurazione 'Profilo CPU database limitazione dinamica' e limita di conseguenza i TPS in ingresso notificando la directory. Poiché la limitazione si basa su valori di utilizzo della CPU in continuo cambiamento,

è possibile pronunciare la velocità che limita dinamicamente il traffico.



- Inizialmente, l'utilizzo della CPU del database è al di sotto del limite, quindi non si verifica alcuna limitazione. Inoltre, PGW-2 non dispone della configurazione di Dynamic Rate Limiting e pertanto non viene applicata alcuna limitazione indipendentemente dall'utilizzo della CPU.
- Quando l'utilizzo della CPU del database è pari al 62%, il traffico viene limitato del 40% e il limite di velocità effettiva è 600 (il valore TPS in ingresso è 1000, il valore DRA è di solo 600).
- Se l'utilizzo della CPU rimane tra il 60 e il 65%, la limitazione del 40% continua ad essere applicata al limite di velocità configurato di 1000 e il limite di velocità effettiva è 600 (il valore TPS in ingresso è 1000, il valore DRA è solo 600).
- L'utilizzo della CPU si riduce al 49%, l'inversione della limitazione inizia a pGW-1.
- Se l'utilizzo della CPU rimane pari o inferiore al 49% per 30 secondi, la limitazione viene ridotta del 20% al 20%. Ora il limite di velocità effettiva è 800 (TPS in ingresso è 1000, DRA ne consente solo 800). Mentre lo storno, come per configurazione, viene fatto in passaggi del 20%.
- Quando l'utilizzo della CPU del database aumenta fino al 57%, il traffico viene limitato del 30% e il limite di velocità effettiva è 700 (il valore TPS in entrata è 1000, il valore DRA è di solo 700).
- Quando l'utilizzo della CPU del database aumenta fino al 67%, il traffico viene limitato del 50% e il limite di velocità effettiva è 500 (il valore TPS in entrata è 1000, il valore DRA è di solo 500).
- Quando l'utilizzo della CPU del database diminuisce al 55%, il traffico viene limitato del 30% e il limite di velocità effettiva è 700 (il valore TPS in entrata è 1000, il valore DRA è di solo 700).
- Se la CPU scende al 49% o meno per il successivo intervallo di 30 secondi, la limitazione

viene ridotta dal 20% al 10% e il limite di velocità effettiva è 900 (TPS in ingresso è 1000, DRA ne consente solo 900).

- Se inoltre la CPU rimane al 49% o meno per i successivi 30 secondi, la limitazione viene ridotta del 20% a 0 e non viene applicato alcun limite di velocità al completamento dell'inversione (il valore TPS in ingresso è 1000, il valore DRA è 1000).

Informazioni su questa traduzione

Cisco ha tradotto questo documento utilizzando una combinazione di tecnologie automatiche e umane per offrire ai nostri utenti in tutto il mondo contenuti di supporto nella propria lingua. Si noti che anche la migliore traduzione automatica non sarà mai accurata come quella fornita da un traduttore professionista. Cisco Systems, Inc. non si assume alcuna responsabilità per l'accuratezza di queste traduzioni e consiglia di consultare sempre il documento originale in inglese (disponibile al link fornito).