Limitation de débit statique et dynamique préemptive avec CPS vDRA

Table des matières

Introduction

Conditions préalables

- Exigences
- Composants utilisés

Informations générales

Problème

Solution

- Limite de débit statique au répartiteur de charge
  - Limite de débit en entrée
  - Limite de débit de sortie
- Limite de débit dynamique

Introduction

Ce document décrit les options de limite de débit dans DRA, un composant de télécommunication qui achemine les messages Diameter et gère le trafic réseau.

Conditions préalables

Exigences

Cisco vous recommande de prendre connaissance des rubriques suivantes :

- Agent de routage de diamètre (vDRA) Cisco Policy Suite (CPS)
- Notions de base et spécifications relatives à Diameter Routing Agent

Composants utilisés

Les informations contenues dans ce document sont basées sur le DRA de Cisco Policy Suite.

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. Si votre réseau est en ligne, assurez-vous de bien comprendre l'incidence possible des commandes.

Informations générales

Le DRA est un composant des réseaux de télécommunications, en particulier dans le contexte

des réseaux basés sur le protocole Diameter. Le DRA achemine efficacement les messages de diamètre entre différents éléments du réseau, tels que les serveurs de politiques, les systèmes de facturation et d'autres périphériques compatibles avec le diamètre. La limitation du débit est une technique de gestion du trafic réseau utilisée pour contrôler la quantité de trafic en provenance ou à destination d'un élément de réseau. Il permet de s'assurer que les ressources réseau ne sont pas épuisées, de maintenir la qualité de service et d'empêcher toute utilisation abusive du réseau.

Problème

Chaque composant du réseau peut gérer la charge de trafic en fonction de sa capacité nominale, mais en temps réel, il peut y avoir des scénarios où le trafic généré est supérieur à ce que le système peut gérer. Certaines d'entre elles sont :

- Comportement de l'utilisateur : activités telles que la diffusion en continu d'événements ou de mises à jour logicielles qui génèrent de grandes quantités de données sur une courte période. Généralement envoyé de la passerelle (Gw) vers DRA.
- Encombrement du réseau : en période de forte utilisation du réseau, l'encombrement peut s'accentuer, ce qui entraîne l'envoi de données en file d'attente par salves lorsque la capacité devient disponible.
- Mécanismes de résilience du réseau : réacheminement du trafic pendant les pannes ou la maintenance, entraînant des pics temporaires. Cela peut affecter le flux de trafic dans les sites associés qui n'ont pas de problème réseau.
- Comportement des éléments du réseau : en cas de surcharge et d'encombrement, vous pouvez commencer à ne voir aucune réponse/expiration d'un ou de plusieurs éléments du réseau, ce qui peut provoquer une reconnexion contribuant à une surcharge supplémentaire du système.
- Vidange de la passerelle : la passerelle peut vider les sessions existantes en raison de modifications de stratégie, de modification de topologie ou de toute activité de maintenance ou de dépannage. Au cours de ces scénarios, les sessions sont effacées et vous pouvez recevoir une rafale de demandes Gx Credit Control Request (CCR)-T.

Solution

DRA peut répartir la charge entre plusieurs serveurs Diameter afin d'assurer un traitement efficace des requêtes et d'éviter la surcharge d'un seul serveur. En cas de panne du serveur, le DRA peut rediriger les messages vers d'autres serveurs, garantissant ainsi une disponibilité et une fiabilité élevées des services réseau.

La limitation du débit sur le DRA protège non seulement le DRA, mais également d'autres entités en assurant un flux contrôlé de messages. Les principaux avantages de la limitation du débit sont les suivants :

- Continuité des services - Maintien d'une disponibilité continue des services en s'assurant que les composants réseau critiques ne sont pas surchargés et évitent les pannes.
- Évolutivité : permet au réseau de gérer des charges variables sans dégradation des performances.

- Conformité aux contrats de niveau de service (SLA) - Garantir que le réseau respecte ses SLA en maintenant des niveaux de performance et de fiabilité cohérents.

Limite de débit statique au répartiteur de charge

Il s’agit d’une approche simple, dans laquelle un seuil fixe est défini en fonction de la capacité nominale de la passerelle DRA/Packet Gateway (pGW) et d’autres éléments du réseau et ne change pas en fonction des conditions du réseau ou des ressources système. En plafonnant le taux de demandes entrantes, vous obtenez un résultat prévisible sur la quantité de trafic que DRA traite.

Les configurations pour la limite de débit statique dépendent du cas d'utilisation pour lequel elle est appliquée.

Limite de débit en entrée

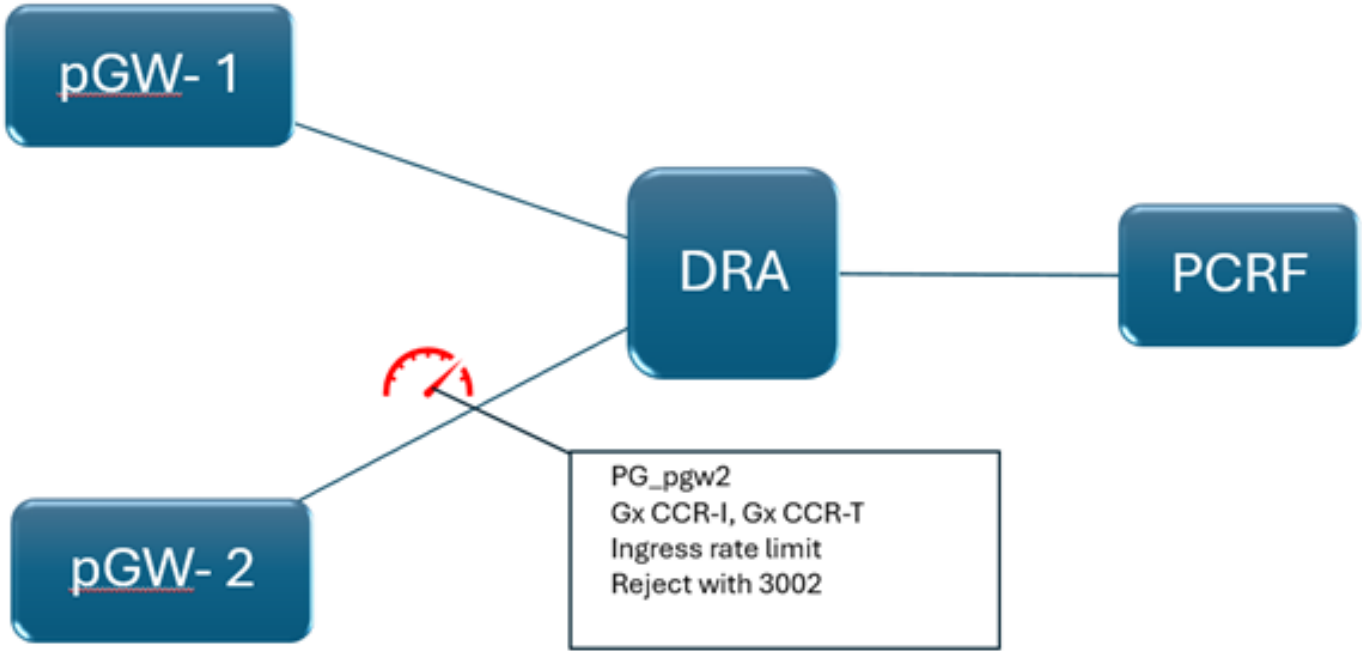
Scénario: Rafales de pGW

Des seuils spécifiques aux pGW qui sont sensibles à ces rafales de trafic sont configurés. La valeur doit être obtenue en fonction des numéros de trafic normal/de trafic de pointe qui peuvent être vus pendant ces rafales.

Les numéros de seuil peuvent être définis spécifiquement pour chaque type de message afin de garantir que seul le trafic en rafale est limité, par exemple, seules les requêtes Gx CCR-I et Gx CCR-T d'une passerelle doivent être limitées, mais le trafic Gx CCR-U ou Gy est transféré tel que reçu.

Dans ce cas, vous pouvez ralentir en entrée, c'est-à-dire que DRA ralentit le message dès qu'il le reçoit, car le but ici est de rejeter en fonction de l'élément de réseau à partir duquel il reçoit la requête et d'éviter de traiter un nombre de requêtes plus élevé que celui que le DRA peut traiter.

Le comportement de régulation peut être de rejeter le message avec un code d'erreur et un message d'erreur particuliers ou de l'abandonner.

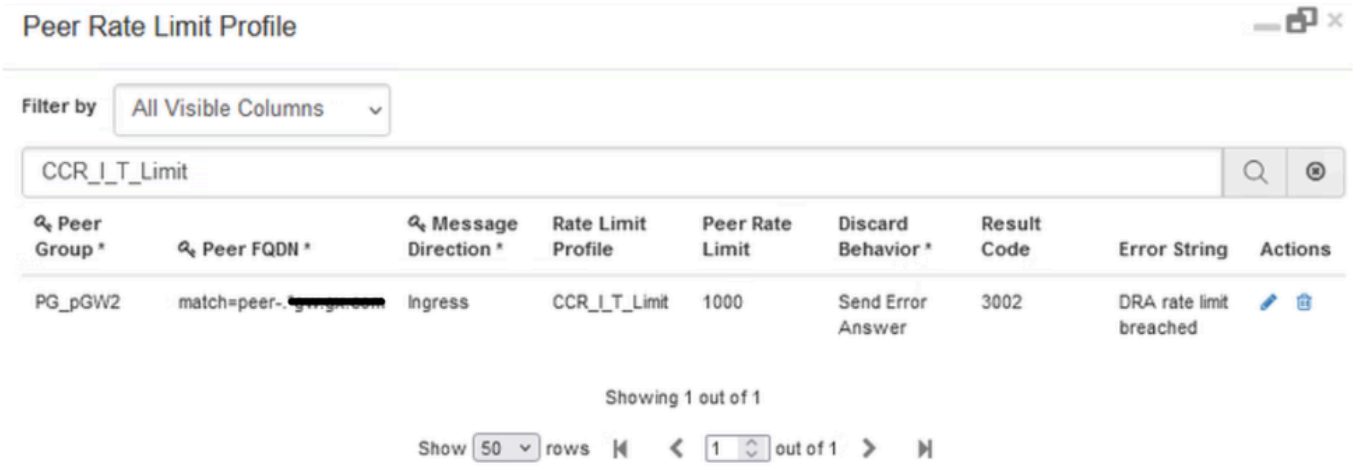


Ce comportement peut être activé dans CPS vDRA en configurant les tables de données de référence personnalisées « Peer Rate Limit Profile » et « Message Rate Limit Profile ». Dans ces tables CRD, vous devez configurer ces valeurs :

Groupe D'Homologues	Un groupe d'homologues est un regroupement logique de noeuds de diamètre basé sur leur domaine et leur hôte. Vous devez configurer le groupe d'homologues qui doit être limité.
Nom de domaine complet de l'homologue (FQDN)	FQDN (correspondance exacte ou regex) pour les homologues du groupe d'homologues dont vous avez besoin pour la limite de taux.
Direction du message	Direction de l'étranglement - Entrée ou Sortie. Dans ce cas - Ingress.
Profil de limite de débit	Nom du profil de limite de débit de messages utilisé pour définir le type de message qui doit être limité.
Limite de débit d'homologue	Taux de demandes autorisées pour ce groupe d'homologues. Cela inclut tous les types de messages de ce groupe d'homologues.
Comportement de rejet	Vous pouvez choisir de supprimer la demande ou de la rejeter avec un code d'erreur.

Code résultat	Valeur du code de résultat au cas où vous rejetteriez les messages. Ne s'applique pas si les messages sont abandonnés.
Chaîne d'erreur	Chaîne d'erreur utilisée dans le message de réponse de la demande rejetée. Ne s'applique pas au cas où le message est abandonné.
Identificateur d'application	ID d'application du message à limiter.
Code de commande	Code de commande du message à limiter.
Type de message/requête	ID d'application et type de demande des demandes qui doivent être limitées.
Limite de débit des messages	Serveur TelePresence (TPS) de la demande de ce type de message qui doit être traitée par DRA. Les demandes au-delà de ce TPS sont limitées. Cette valeur est par homologue dans le groupe d'homologues.

Voici un exemple, où vous configurez une limite de débit globale de 1000 messages pour pGW2 et une limite de débit de 200 Gx CCR-I et 200 Gx CCR-T. Toute demande au-delà de ce taux est rejetée avec un 3002, et un message d'erreur indiquant que la limite de taux est dépassée.



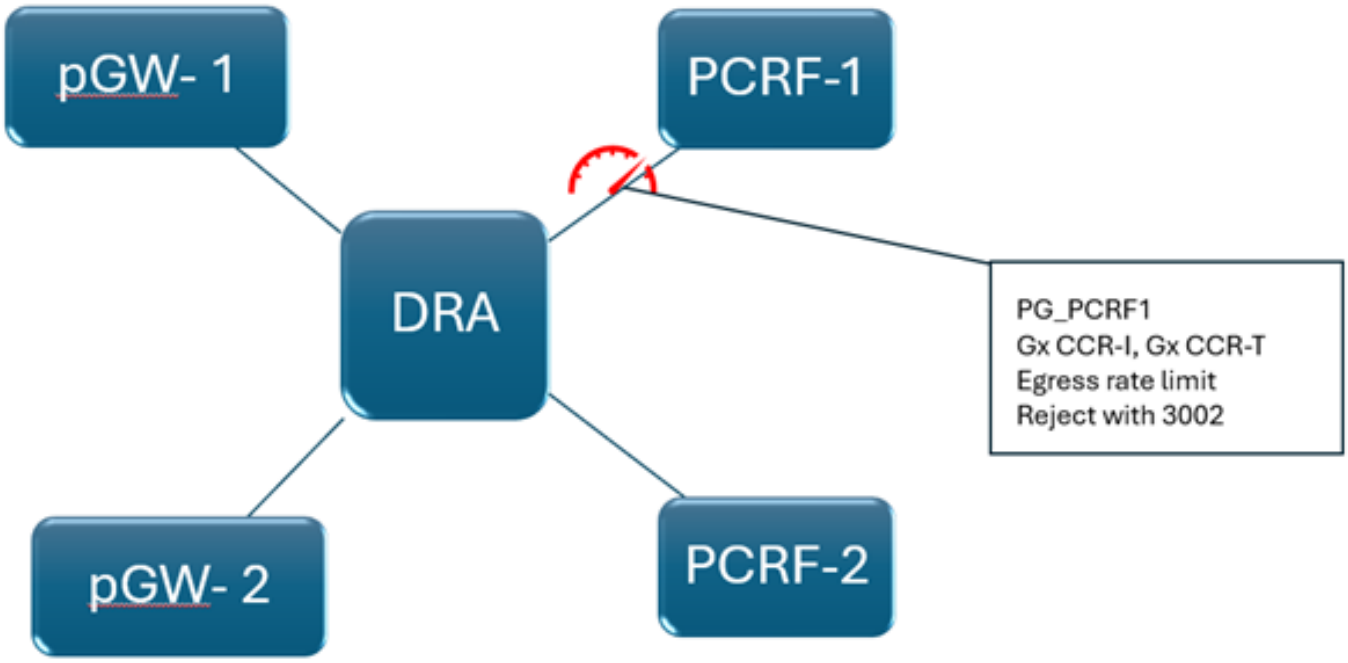
Limite de débit de sortie

Scénario: Protection d'un élément de réseau qui reçoit une requête

Prenons l'exemple d'une transaction Gx dans laquelle la demande est reçue de pGW et transmise à la fonction PCRF (Policy and Charging Rules Function). Si la quantité de données que PCRF peut gérer est limitée, alors même si DRA peut gérer le trafic entrant, vous pouvez utiliser DRA pour limiter le message au niveau de DRA au lieu de transférer la requête vers PCRF et de la surcharger.

Ici, vous devez réguler le côté sortie, c'est-à-dire que DRA régule le message juste avant qu'il ne soit transmis au PCRF, en fonction du groupe d'homologues PCRF identifié sur la base de la logique de routage DRA.

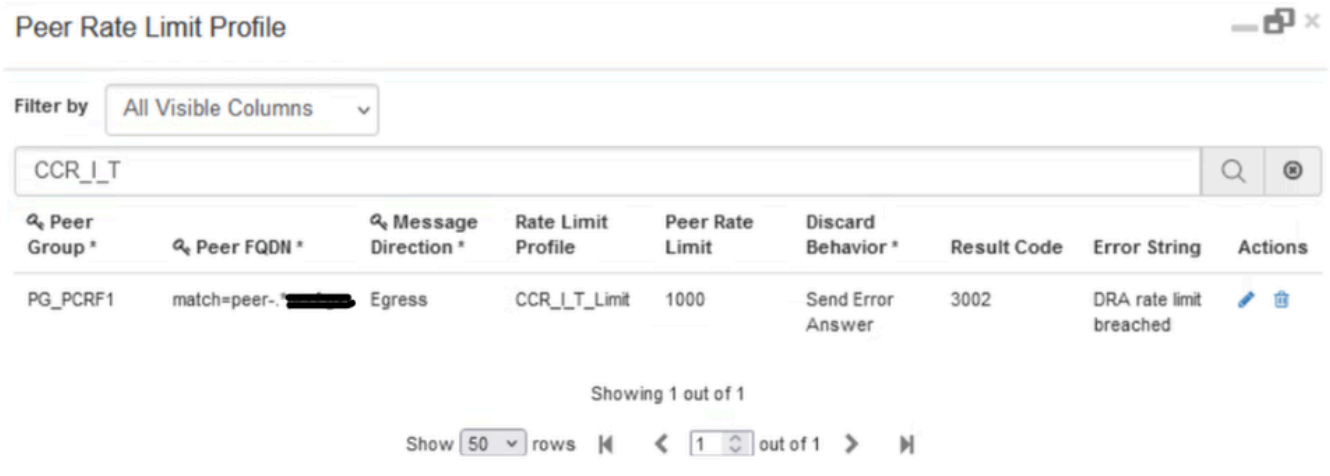
Le comportement de régulation peut être de rejeter le message avec un code d'erreur et un message d'erreur particuliers ou de l'abandonner.



Ce comportement peut être activé dans CPS vDRA en configurant les tables CRD « Peer Rate Limit Profile » et « Message Rate Limit Profile ». Dans ces tables CRD, vous devez configurer les valeurs suivantes :

Groupe D'Homologues	Un groupe d'homologues est un regroupement logique de noeuds de diamètre basé sur leur domaine et leur hôte. Vous devez configurer le groupe d'homologues qui doit être limité.
FQDN homologue	FQDN (correspondance exacte ou regex) pour les homologues du groupe d'homologues dont vous avez besoin pour la limite de taux.
Direction du message	Direction de l'étranglement - Entrée ou Sortie. Dans ce cas - Sortie.
Profil de limite de débit	Nom du profil de limite de débit de messages utilisé pour définir le type de message qui doit être limité.
Limite de débit d'homologue	Taux de demandes qui doivent être autorisées pour ce groupe d'homologues. Cela inclut tous les types de messages de ce groupe d'homologues.
Comportement de rejet	Vous pouvez choisir de supprimer la demande ou de la

	rejeter avec un code d'erreur.
Code résultat	Valeur du code de résultat au cas où vous rejetteriez les messages. Ne s'applique pas si les messages sont abandonnés.
Chaîne d'erreur	Chaîne d'erreur utilisée dans le message de réponse de la demande rejetée. Ne s'applique pas au cas où le message est abandonné.
Identificateur d'application	ID d'application du message à limiter.
Code de commande	Code de commande du message à limiter.
Type de message/requête	ID d'application et type de demande des demandes qui doivent être limitées.
Limite de débit des messages	TPS de la demande de ce type de message qui est traitée par DRA. Les demandes au-delà de ce TPS sont limitées. Cette valeur est par homologue dans le groupe d'homologues.



Scénario: Lenteur du réseau entraînant une congestion du trafic, entraînant une défaillance totale ou partielle du DRA

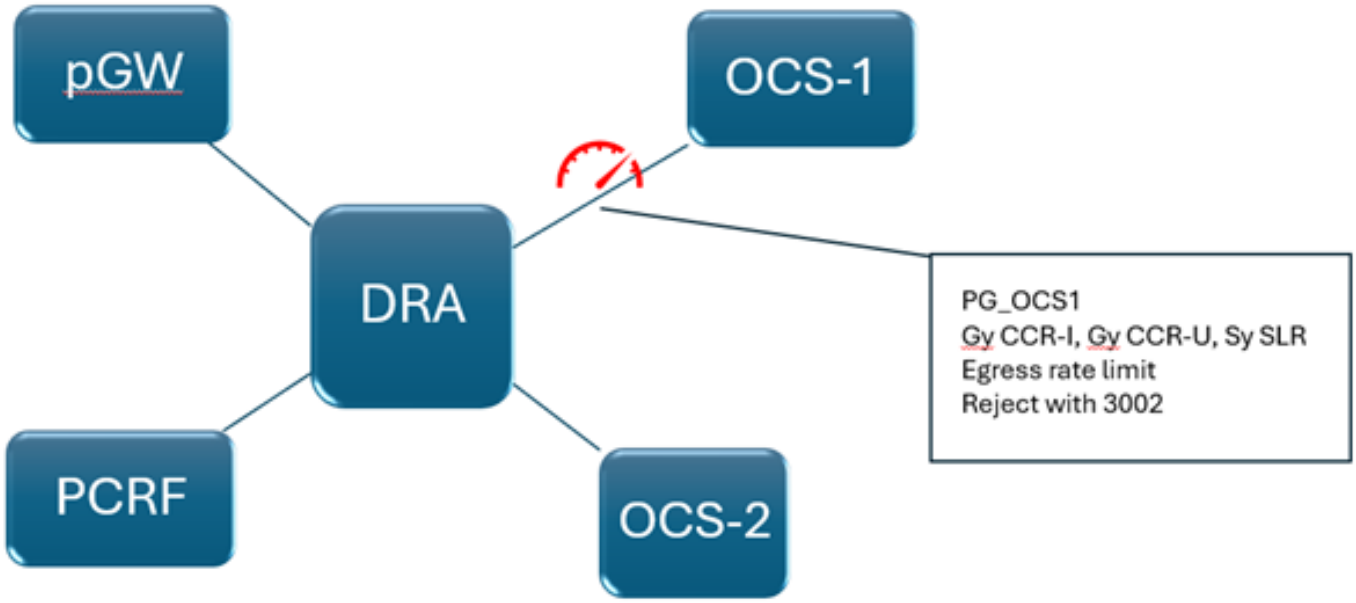
Prenons l'exemple d'une transaction Gy échangée entre pGW et OCS (Online Charging System). En cas de lenteur du réseau sur le canal DRA-OCS (en raison d'un trafic élevé en provenance de pGW ou d'un autre problème réseau), la demande expire en raison d'une violation de SLA. Ces dépassements de délai ont un impact non seulement sur le DRA, mais aussi sur le réseau global.

Les ressources DRA sont bloquées en essayant d'envoyer la requête à OCS sur le réseau lent, ce qui entraîne l'épuisement de ses ressources. Il en résulte que plusieurs demandes sont rejetées par DRA, bien que la capacité nominale de DRA ne soit pas atteinte.

Cela affecte également le trafic qui n'est pas sur le canal DRA-OCS. Ces rejets/délais d'attente et ces abandons déclenchent la reconnexion sur plusieurs éléments du réseau.

Dans ce cas, vous devez ralentir le côté sortie - DRA ralentit le message juste avant qu'il ne soit transmis à l'OCS, en fonction du groupe d'homologues OCS (qui présente des limitations de capacité ou des problèmes réseau).

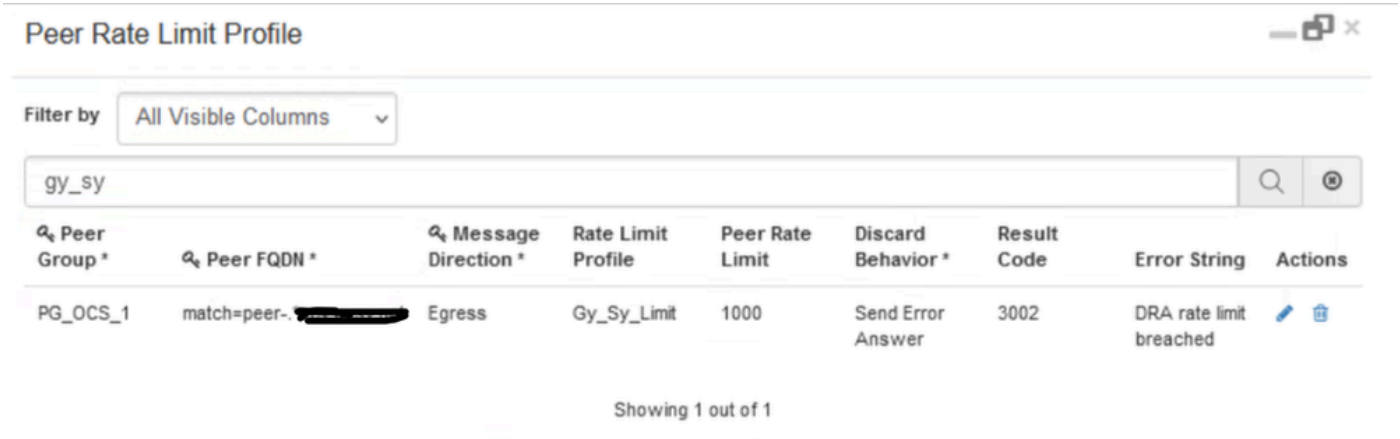
Le comportement de régulation peut être de rejeter le message avec un code d'erreur et un message d'erreur particuliers ou de l'abandonner.



Ce comportement peut être activé dans CPS vDRA en configurant les tables CRD « Peer Rate Limit Profile » et « Message Rate Limit Profile ». Dans ces tables CRD, vous devez configurer ces valeurs :

Groupe D'Homologues	Un groupe d'homologues est un regroupement logique de noeuds de diamètre basé sur leur domaine et leur hôte. Vous devez configurer le groupe d'homologues qui doit être limité.
FQDN homologue	FQDN (correspondance exacte ou regex) pour les homologues du groupe d'homologues dont vous avez besoin pour la limite de taux.
Direction du message	Direction de l'étranglement - Entrée ou Sortie. Dans ce cas - Sortie.
Profil de limite de débit	Nom du profil de limite de débit de messages utilisé pour définir le type de message qui doit être limité.
Limite de débit d'homologue	Taux de demandes autorisées pour ce groupe d'homologues. Cela inclut tous les types de messages de ce groupe d'homologues.

Comportement de rejet	Vous pouvez choisir de supprimer la demande ou de la rejeter avec un code d'erreur.
Code résultat	Valeur du code de résultat au cas où vous rejetteriez les messages. Ne s'applique pas si les messages sont abandonnés.
Chaîne d'erreur	Chaîne d'erreur utilisée dans le message de réponse de la demande rejetée. Ne s'applique pas au cas où le message est abandonné.
Identificateur d'application	ID d'application du message à limiter.
Code de commande	Code de commande du message à limiter.
Type de message/requête	ID d'application et type de demande des demandes qui doivent être limitées.
Limite de débit des messages	TPS de la demande de ce type de message qui est traitée par DRA. Les demandes au-delà de ce TPS sont limitées. Cette valeur est par homologue dans le groupe d'homologues.

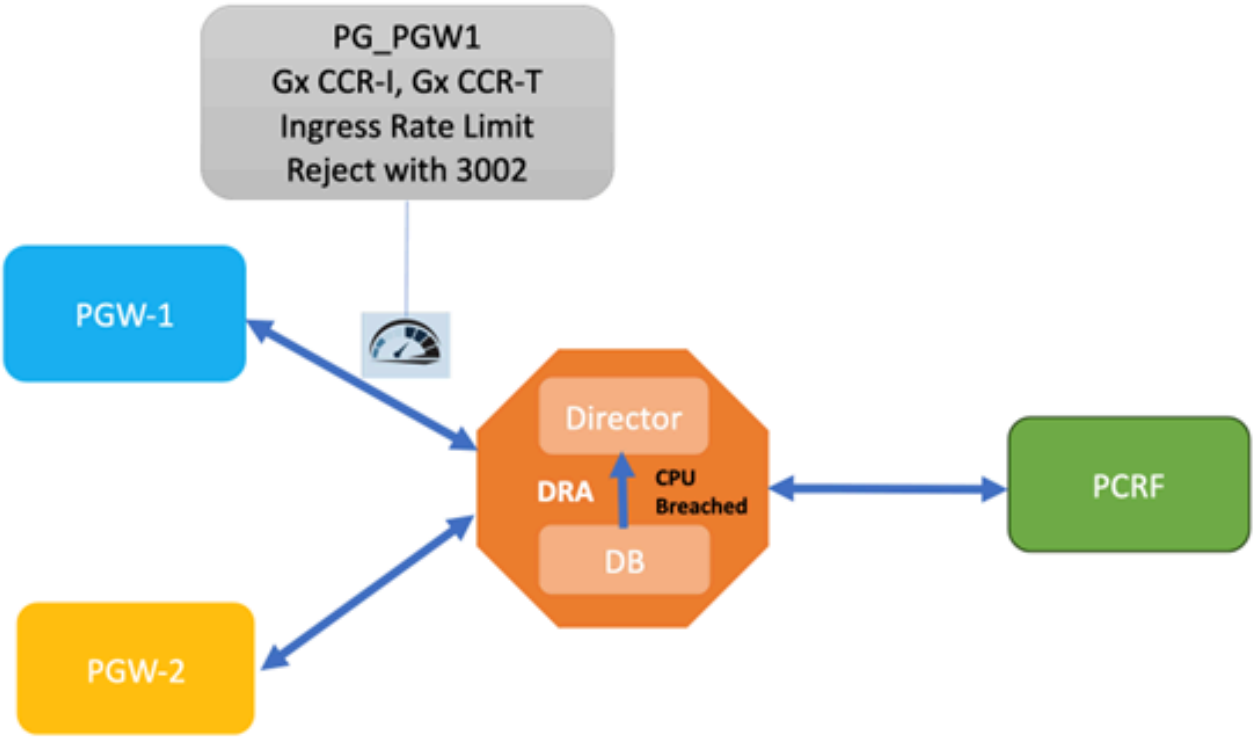


Limite de débit dynamique

Lorsque des rafales CCR-I ou CCR-T se produisent, il peut y avoir une surcharge sur la base de données (DB), ce qui peut entraîner la déstabilisation du système. Pour remédier à ce problème, le DRA prend en charge la limitation dynamique du débit (pour Gx CCR-I et Gx CCR-T uniquement) en fonction de la capacité DB disponible.

DRA surveille l'utilisation du processeur de la base de données et, chaque fois que le seuil est dépassé, il limite les requêtes entrantes. Les seuils du processeur pour la limitation et le trafic entrant à limiter sont configurables.

Différents seuils de CPU avec des pourcentages de limitation correspondants peuvent être configurés. DRA ajuste le niveau de limitation en fonction de l'utilisation actuelle du CPU de la base de données. Lorsque l'utilisation du processeur devient stable, la limitation s'arrête progressivement.



Ce comportement peut être activé dans CPS vDRA en configurant les tables CRD « Peer Rate Limit Profile », « Message Rate Limit Profile », « Dynamic Peer Rate Limit Profile » et « Dynamic Throttling DB CPU Profile ». Dans ces tables CRD, vous devez configurer ces valeurs :

Groupe D'Homologues	Un groupe d'homologues est un regroupement logique de noeuds de diamètre basé sur leur domaine et leur hôte. Dans cet exemple, vous configurez le groupe d'homologues du pGW.
FQDN homologue	FQDN (correspondance exacte ou regex) pour les homologues du groupe d'homologues dont vous avez besoin pour la limite de taux.
Type de message/requête	ID d'application et type de demande des demandes qui doivent être limitées. Dans cet exemple, Gx CCR-I, Gx CCR-T.
Direction du message	Direction de l'étranglement - Entrée ou Sortie. Dans ce cas - Ingress.
Profil de limite de débit	Nom du profil de limite de débit de messages utilisé pour

	définir le type de message qui doit être limité.
Limite de débit d'homologue	Taux de demandes autorisées pour ce groupe d'homologues. Cela inclut tous les types de messages de ce groupe d'homologues.
Comportement de rejet	Vous pouvez choisir de supprimer la demande ou de la rejeter avec un code d'erreur.
Code résultat	Valeur du code de résultat au cas où vous rejetteriez les messages. Ne s'applique pas si les messages sont abandonnés.
Chaîne d'erreur	Chaîne d'erreur utilisée dans le message de réponse de la demande rejetée. Ne s'applique pas au cas où le message est abandonné.
Identificateur d'application	ID d'application du message à limiter.
Code de commande	Code de commande du message à limiter.
Limite de débit des messages	TPS de la demande de ce type de message qui est traitée par DRA. Les demandes au-delà de ce TPS sont limitées. Cette valeur est par homologue dans le groupe d'homologues.
Profil de processeur de base de données de limitation dynamique	Il s'agit du nom du profil de CPU, qui est utilisé pour définir le pourcentage de limitation pour différentes plages de CPU.
Seuil d'utilisation du CPU DB	Vous pouvez choisir la valeur correcte des niveaux de CPU configurés comme limites de brèche selon le modèle de trafic de votre déploiement.
Pourcentage De Commande	% de limite de débit appliqué lorsque le niveau de CPU correspondant est atteint.

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-.*[redacted]	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

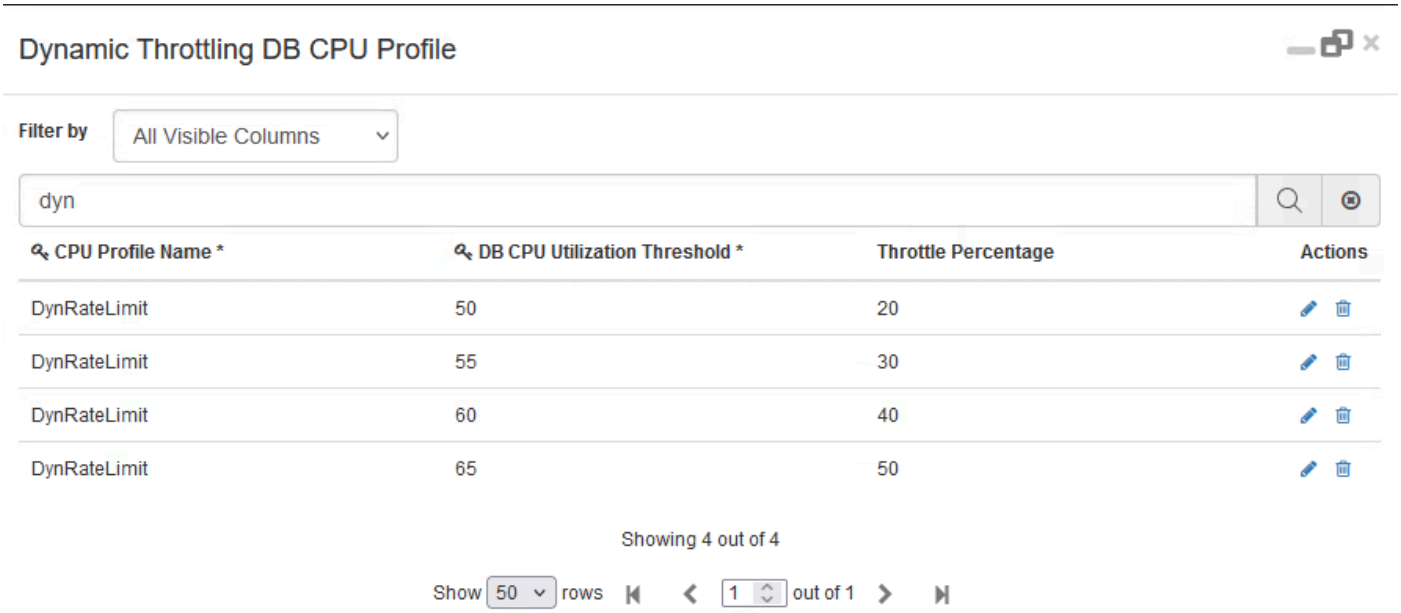
Filter by: All Visible Columns

DynRateLimit

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

Showing 1 out of 1

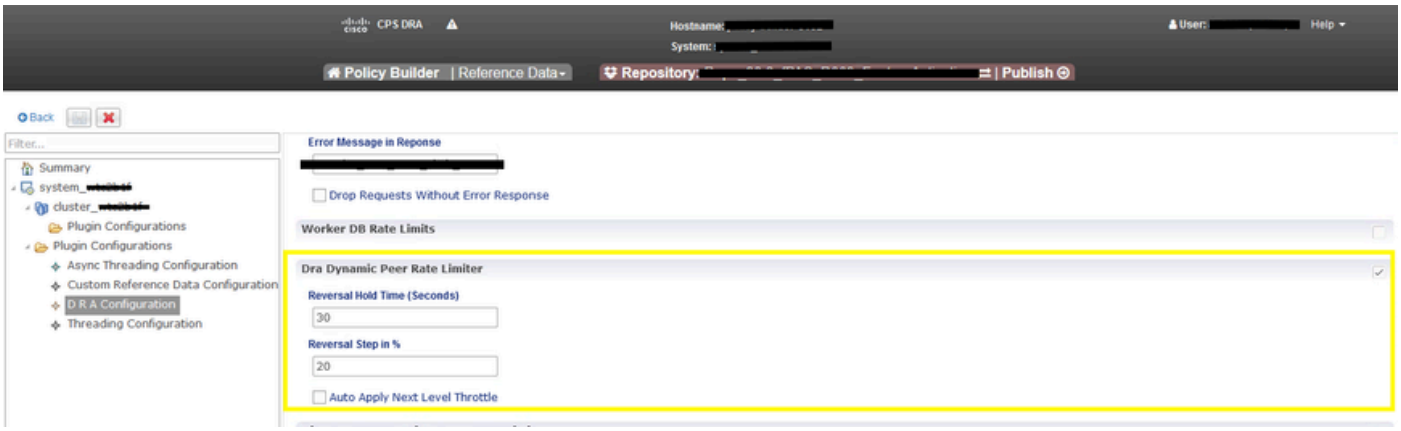
Show 50 rows 1 out of 1



En outre, ce comportement doit être activé, en cochant la case dans Policy Builder sous 'DRA Configuration Plugin', sous la section 'DRA Dynamic Peer Rate Limiter'.

Durée d'attente de contrepassation : période pendant laquelle l'utilisation du processeur est surveillée avant l'application de la contrepassation.

Etape de contrepassation en % - Le pourcentage de limitation qui est contrepassé.



Scénario: Limitation dynamique du débit basée sur l'utilisation du CPU

Considérez cette configuration au niveau de DRA :

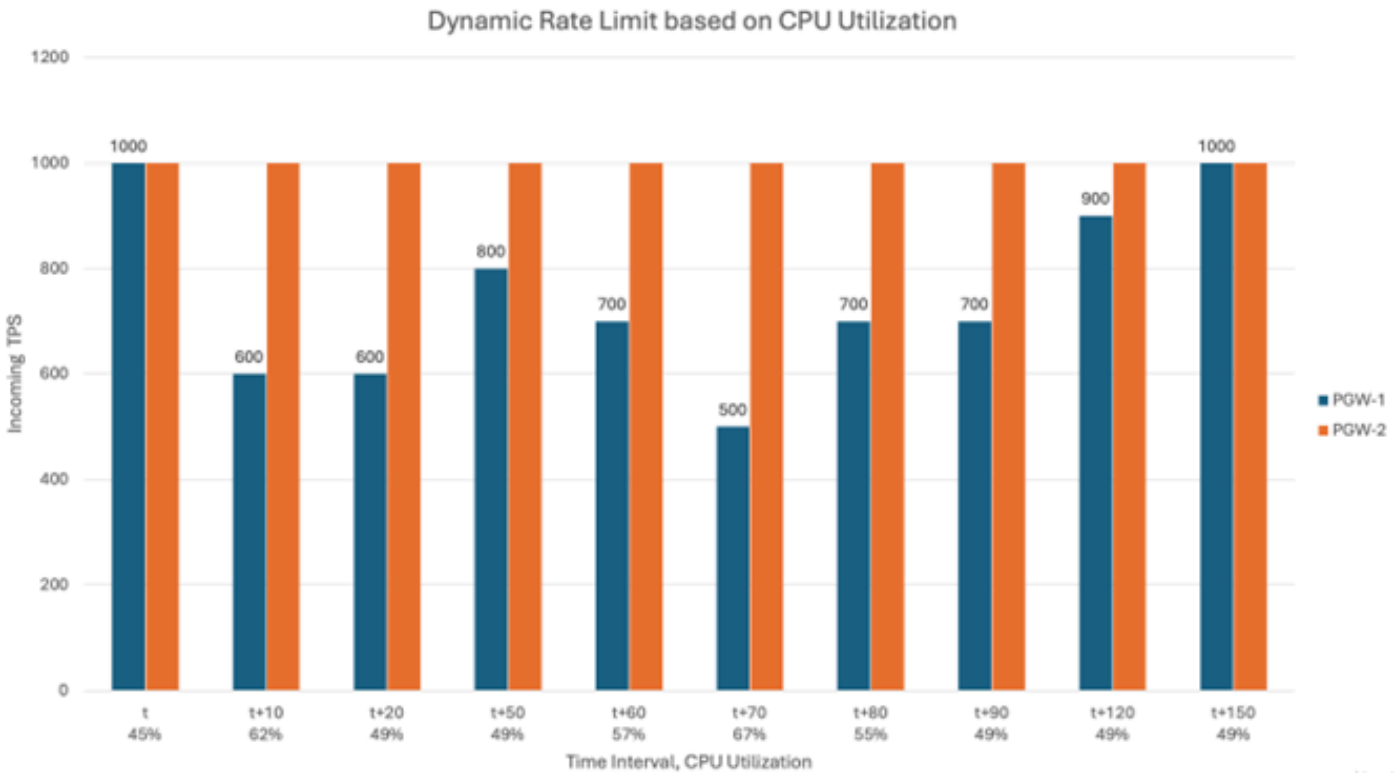
Limite du taux de messages statiques : 1000 (c'est donc la valeur du TPS entrant)

Durée d'attente de contrepassation : années 30

Étape de contrepassation en % : 20 %

Chaque fois que l'utilisation du CPU de la base de données dépasse le seuil, elle se réfère à la configuration « Dynamic Throttling DB CPU Profile » et étrangle les TPS entrants en conséquence en notifiant le directeur. Étant donné qu'elle limite en fonction des valeurs d'utilisation du

processeur en constante évolution, vous pouvez dire que son débit dynamique limite le trafic.



- Initialement, l'utilisation du CPU de la base de données est inférieure à la limite, donc aucune limitation n'a lieu. En outre, PGW-2 n'a pas la configuration de Dynamic Rate Limiting et donc aucun étranglement n'y a lieu, quelle que soit l'utilisation du CPU.
- Lorsque l'utilisation du CPU de la base de données est de 62 %, le trafic est limité de 40 % et la limite de débit effective est de 600 (le TPS entrant est de 1 000, le DRA n'autorise que 600).
- Si l'utilisation de l'UC reste entre 60 et 65 %, la limitation de 40 % continue d'être appliquée sur la limite de débit configurée de 1 000 et la limite de débit effective est de 600 (TPS entrant est de 1 000, DRA n'autorise que 600).
- L'utilisation du CPU est réduite à 49 %, l'inversion de la limitation commence à pGW-1.
- Si l'utilisation du processeur reste à 49 % ou moins pendant 30 secondes, la limitation est réduite de 20 % à 20 %. Désormais, la limite de débit effective est de 800 (TPS entrant est de 1000, DRA n'autorise que 800). Lors de l'inversion, selon la configuration, elle est effectuée par étapes de 20 %.
- Lorsque l'utilisation du CPU de la base de données augmente à 57 %, le trafic est limité de 30 % et la limite de débit effective est de 700 (le TPS entrant est de 1 000, le DRA n'autorise que 700).
- Lorsque l'utilisation du CPU de la base de données augmente à 67 %, le trafic est limité de 50 % et la limite de débit effective est de 500 (le TPS entrant est de 1 000, le DRA n'autorise que 500).
- Lorsque l'utilisation du CPU de la base de données diminue à 55 %, le trafic est limité à 30 % et la limite de débit effective est de 700 (le TPS entrant est de 1 000, le DRA n'autorise que 700).
- Si le processeur tombe à 49 % ou moins pendant l'intervalle de 30 secondes suivant, la limitation est diminuée de 20 % à 10 % et la limite de débit effective est de 900 (TPS entrant

est de 1 000, DRA n'autorise que 900).

- Si le processeur reste en outre à 49 % ou moins pendant l'intervalle de 30 secondes suivant, l'étranglement est réduit de 20 % à 0 et aucune limite de débit n'est appliquée lorsque l'inversion est terminée (TPS entrant est 1000, DRA autorise 1000.)

À propos de cette traduction
Cisco a traduit ce document en traduction automatisée vérifiée par une personne dans le cadre d'un service mondial permettant à nos utilisateurs d'obtenir le contenu d'assistance dans leur propre langue.

Il convient cependant de noter que même la meilleure traduction automatisée ne sera pas aussi précise que celle fournie par un traducteur professionnel.