Limitación de velocidad dinámica y estática preventiva con CPS vDRA

Contenido

Introducción

Este documento describe las opciones de límite de velocidad en DRA, un componente de telecomunicaciones que enruta mensajes Diameter y administra el tráfico de red.

Prerequisites

Requirements

Cisco recomienda que tenga conocimiento sobre estos temas:

- Cisco Policy Suite (CPS) Diameter Routing Agent (vDRA)
- Conceptos Básicos y Especificaciones del Agente de Ruteo de Diámetro

Componentes Utilizados

La información de este documento se basa en Cisco Policy Suite DRA.

La información que contiene este documento se creó a partir de los dispositivos en un ambiente de laboratorio específico. Todos los dispositivos que se utilizan en este documento se pusieron en funcionamiento con una configuración verificada (predeterminada). Si tiene una red en vivo, asegúrese de entender el posible impacto de cualquier comando.

Antecedentes

DRA es un componente de las redes de telecomunicaciones, especialmente en el contexto de las redes basadas en el protocolo Diameter. El DRA enruta eficazmente los mensajes Diameter entre diferentes elementos de la red, como servidores de políticas, sistemas de carga y otros dispositivos habilitados para Diameter. La limitación de velocidad es una técnica de administración del tráfico de red utilizada para controlar la cantidad de tráfico hacia o desde un elemento de red. Ayuda a garantizar que los recursos de red no se agotan, mantiene la calidad del servicio y evita el uso indebido o el abuso de la red.

Problema

Cada componente de la red puede gestionar la carga de tráfico en función de su capacidad nominal, pero en tiempo real puede haber situaciones en las que el tráfico generado es más de lo que el sistema puede gestionar. Algunos de ellos son:

- Comportamiento del usuario: actividades como la transmisión de eventos o actualizaciones de software que generan grandes cantidades de datos en un breve período. Normalmente se envía desde la puerta de enlace (Gw) hacia DRA.
- Congestión de red: en períodos de uso elevado de la red, puede acumularse congestión, lo que provoca que los datos en cola se envíen en ráfagas cuando la capacidad esté disponible.
- Mecanismos de flexibilidad de la red: redireccionamiento del tráfico durante interrupciones o mantenimiento, que provoca picos temporales. Esto puede afectar el flujo de tráfico en los sitios acoplados que no tienen ningún problema de red.
- Comportamiento de los elementos de red: en caso de sobrecarga y congestión, puede empezar a ver que no hay respuestas ni tiempos de espera de uno o más elementos de red, lo que puede provocar que la reconexión contribuya a una mayor sobrecarga en el sistema.
- Vaciado del gateway: el gateway puede vaciar las sesiones existentes debido a cambios en las políticas, cambios en la topología o cualquier actividad de mantenimiento o resolución de problemas. Durante estos escenarios, las sesiones se borran y puede recibir una ráfaga de solicitudes Gx Credit Control Request (CCR)-T.

Solución

DRA puede distribuir la carga entre varios servidores Diameter para garantizar una gestión eficiente de las solicitudes y evitar la sobrecarga de un único servidor. En caso de fallo del servidor, el DRA puede redirigir los mensajes a servidores alternativos, lo que garantiza una alta disponibilidad y fiabilidad de los servicios de red.

La limitación de velocidad en el DRA no solo protege el DRA, sino también otras entidades al garantizar un flujo controlado de mensajes. Las ventajas clave de la limitación de velocidad son:

- Continuidad del servicio: mantener la disponibilidad continua del servicio garantizando que los componentes de red críticos no se sobrecarguen y evitando interrupciones.
- Escalabilidad: permite que la red gestione cargas variables sin que el rendimiento se vea afectado.
- Cumplimiento de los acuerdos de nivel de servicio (SLA): garantizar que la red cumple sus

SLA manteniendo unos niveles de rendimiento y fiabilidad uniformes.

Límite de velocidad estática en el equilibrador de carga

Se trata de un enfoque sencillo, en el que se establece un umbral fijo basado en la capacidad nominal de DRA/Packet Gateway (pGW) y otros elementos de la red y que no cambia en función de las condiciones de la red ni de los recursos del sistema. Al limitar la velocidad de las solicitudes entrantes, tendrá un resultado predecible en la cantidad de tráfico que procesa DRA.

Las configuraciones para el límite de velocidad estática dependen del caso práctico al que se aplica.

Límite de velocidad de ingreso

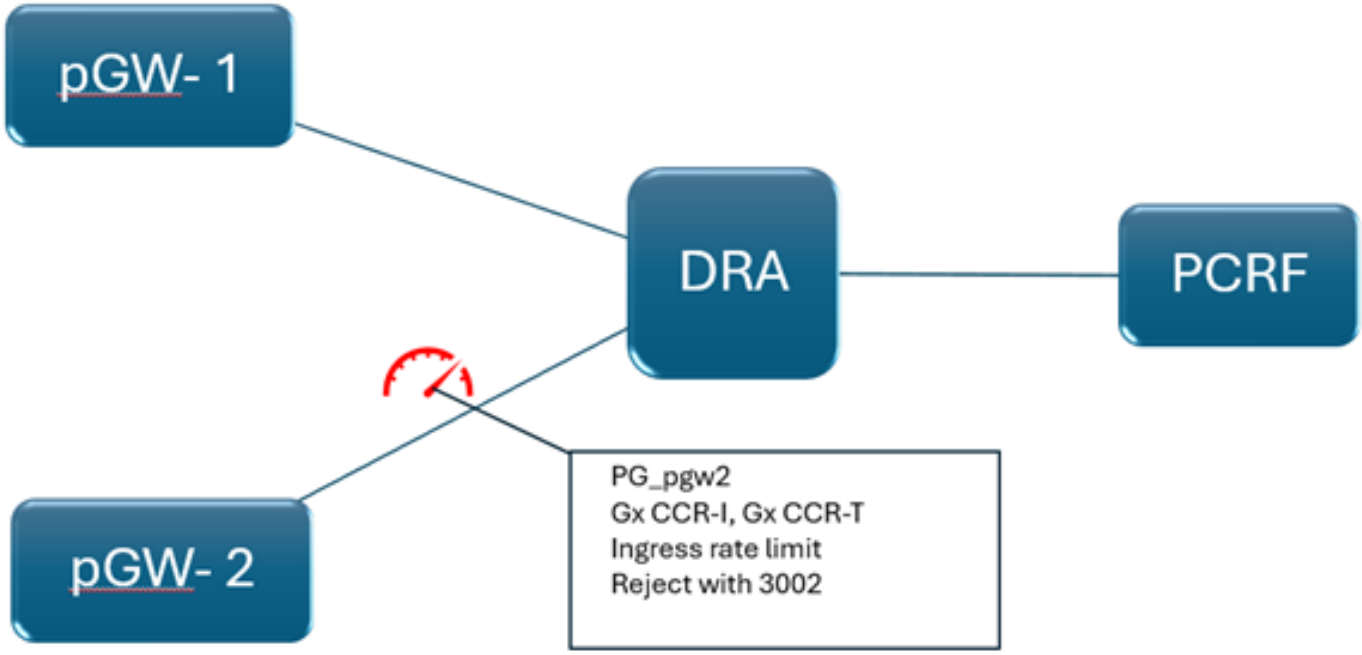
Situación: Ráfagas de pGW

Se configuran los umbrales específicos de los pGW que son susceptibles a estas ráfagas de tráfico. El valor debe obtenerse en función de los números de tráfico normal/tráfico pico que se pueden ver durante estas ráfagas.

Los números de umbral se pueden definir específicamente para cada tipo de mensaje para garantizar que solo se limite el tráfico de ráfaga, como, solo se deben limitar las solicitudes Gx CCR-I y Gx CCR-T de un GW pero el tráfico Gx CCR-U o el tráfico Gy se reenvía como se recibe.

En este caso, puede acelerar en el lado de entrada, es decir, DRA acelera el mensaje tan pronto como lo recibe, ya que el propósito aquí es rechazar en función del elemento de red desde el que recibe la solicitud y evitar el procesamiento de un número mayor de solicitudes que el que DRA puede manejar.

El comportamiento del acelerador puede ser rechazar el mensaje con un código de error y un mensaje de error concretos o descartarlo.



Este comportamiento se puede habilitar en CPS vDRA configurando las tablas de datos de referencia personalizados (CRD) 'Perfil de límite de velocidad de par' y 'Perfil de límite de velocidad de mensaje'. En estas tablas de CRD, debe configurar estos valores:

Grupo de pares	Un grupo de peers es una agrupación lógica de nodos Diameter basada en su rango y host. Necesita configurar el grupo de peers que necesita ser acelerado.
Nombre de dominio completamente calificado (FQDN) del mismo nivel	FQDN (coincidencia exacta o regex) para los pares del grupo de pares que necesita para limitar la velocidad.
Dirección del mensaje	Dirección del límite: entrada o salida. En este caso, Ingress.
Perfil de límite de velocidad	Nombre de perfil de límite de velocidad de mensajes que se utilizó para definir el tipo de mensaje que debe acelerarse.
Límite de velocidad de peer	Tasa de solicitudes permitidas para este grupo de pares. Esto incluye todos los tipos de mensajes de ese grupo de pares.
Descartar comportamiento	Puede optar por rechazar la solicitud o rechazarla con un código de error.

Código de resultado	Valor del código de resultado en caso de que rechace los mensajes. No se aplica en caso de que se descarten los mensajes.
Cadena de error	Cadena de error que se utiliza en el mensaje de respuesta de la solicitud que se rechaza. No se aplica en caso de que el mensaje se descarte.
Identificador de aplicación	ID de aplicación del mensaje que se va a limitar.
Código de comando	Código de comando del mensaje que se va a limitar.
Tipo de mensaje/solicitud	ID de aplicación y Tipo de solicitud de las solicitudes que deben acelerarse.
Límite de velocidad de mensajes	TelePresence Server (TPS) de la solicitud de ese tipo de mensaje que procesa DRA. Las solicitudes más allá de este TPS son aceleradas. Este valor es por peer en el grupo de peers.

A continuación se muestra un ejemplo en el que se configura un límite de velocidad global de 1000 mensajes para pGW2 y un límite de velocidad de 200 Gx CCR-I y 200 Gx CCR-T. Cualquier solicitud superior a esta tasa se rechaza con un 3002 y un mensaje de error que indica que se ha superado el límite de tasa.

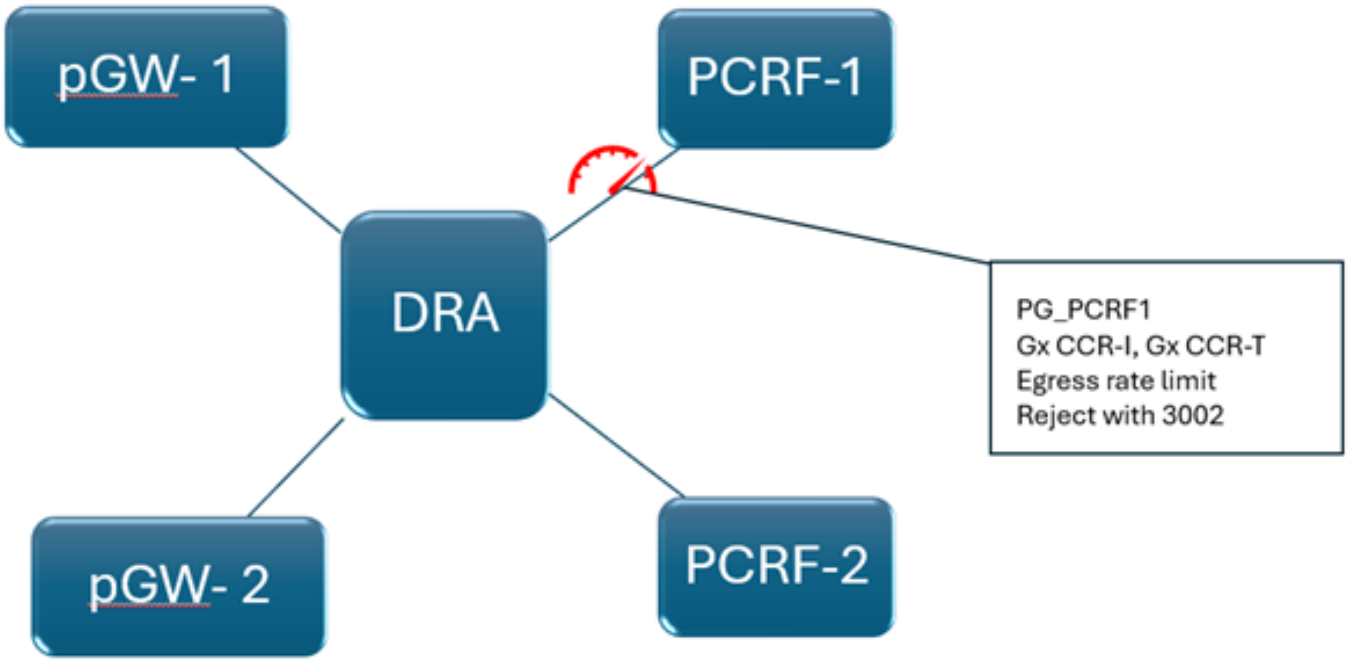
Límite de velocidad de salida

Situación: Protección del elemento de red que recibe la solicitud

Considere un ejemplo de una transacción Gx en la que la solicitud se recibe de pGW y se reenvía a la función de reglas de cobro y políticas (PCRF). Si hay limitaciones en la cantidad de datos que puede manejar PCRF, aunque DRA pueda manejar el tráfico entrante, puede utilizar DRA para regular el mensaje en DRA en lugar de reenviar la solicitud a PCRF y sobrecargarlo.

Aquí necesita acelerar en el lado de salida, es decir, DRA acelera el mensaje justo antes de reenviarlo a PCRF, en función del grupo de pares PCRF que se identifica en función de la lógica de routing DRA.

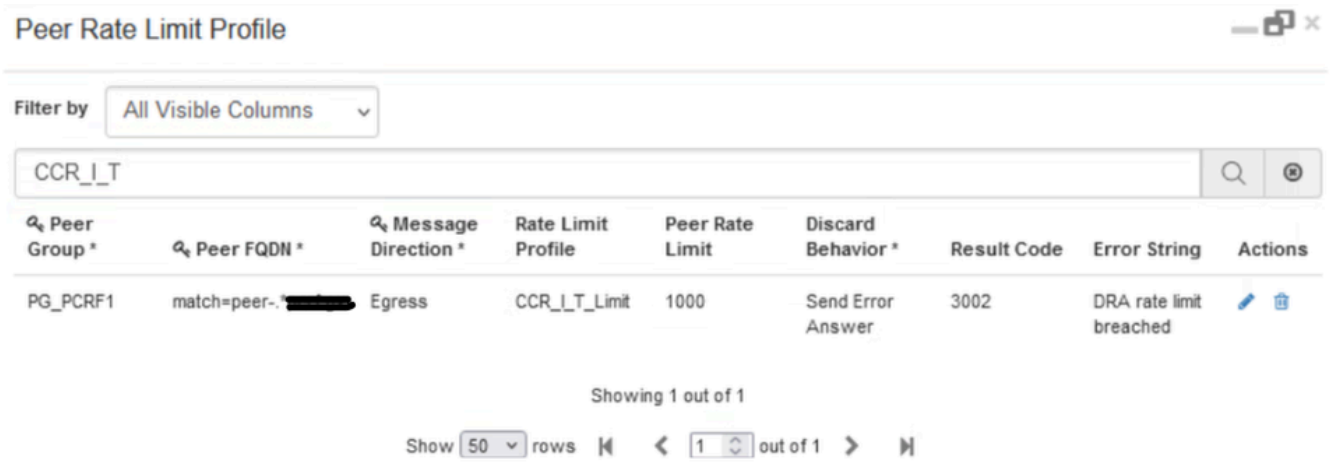
El comportamiento del acelerador puede ser rechazar el mensaje con un código de error y un mensaje de error concretos o descartarlo.



Este comportamiento se puede habilitar en CPS vDRA configurando las tablas CRD 'Peer Rate Limit Profile' y 'Message Rate Limit Profile'. En estas tablas de CRD, debe configurar estos valores:

Grupo de pares	Un grupo de peers es una agrupación lógica de nodos Diameter basada en su rango y host. Necesita configurar el grupo de peers que necesita ser acelerado.
FQDN de peer	FQDN (coincidencia exacta o regex) para los pares del grupo de pares que necesita para limitar la velocidad.
Dirección del mensaje	Dirección del límite: entrada o salida. En este caso, Egress.
Perfil de límite de velocidad	Nombre de perfil de límite de velocidad de mensajes que se utilizó para definir el tipo de mensaje que debe acelerarse.
Límite de velocidad de peer	Tasa de solicitudes que se permite para este grupo de pares. Esto incluye todos los tipos de mensajes de ese grupo de pares.
Descartar comportamiento	Puede optar por rechazar la solicitud o rechazarla con un código de error.
Código de resultado	Valor del código de resultado en caso de que rechace los

	mensajes. No se aplica en caso de que se descarten los mensajes.
Cadena de error	Cadena de error que se utiliza en el mensaje de respuesta de la solicitud que se rechaza. No se aplica en caso de que el mensaje se descarte.
Identificador de aplicación	ID de aplicación del mensaje que se va a limitar.
Código de comando	Código de comando del mensaje que se va a limitar.
Tipo de mensaje/solicitud	ID de aplicación y Tipo de solicitud de las solicitudes que deben acelerarse.
Límite de velocidad de mensajes	TPS de la solicitud de ese tipo de mensaje que procesa DRA. Las solicitudes más allá de este TPS son aceleradas. Este valor es por peer en el grupo de peers.



Situación: Lentitud en la red que provoca congestión del tráfico, lo que provoca que DRA tenga un error parcial o completo

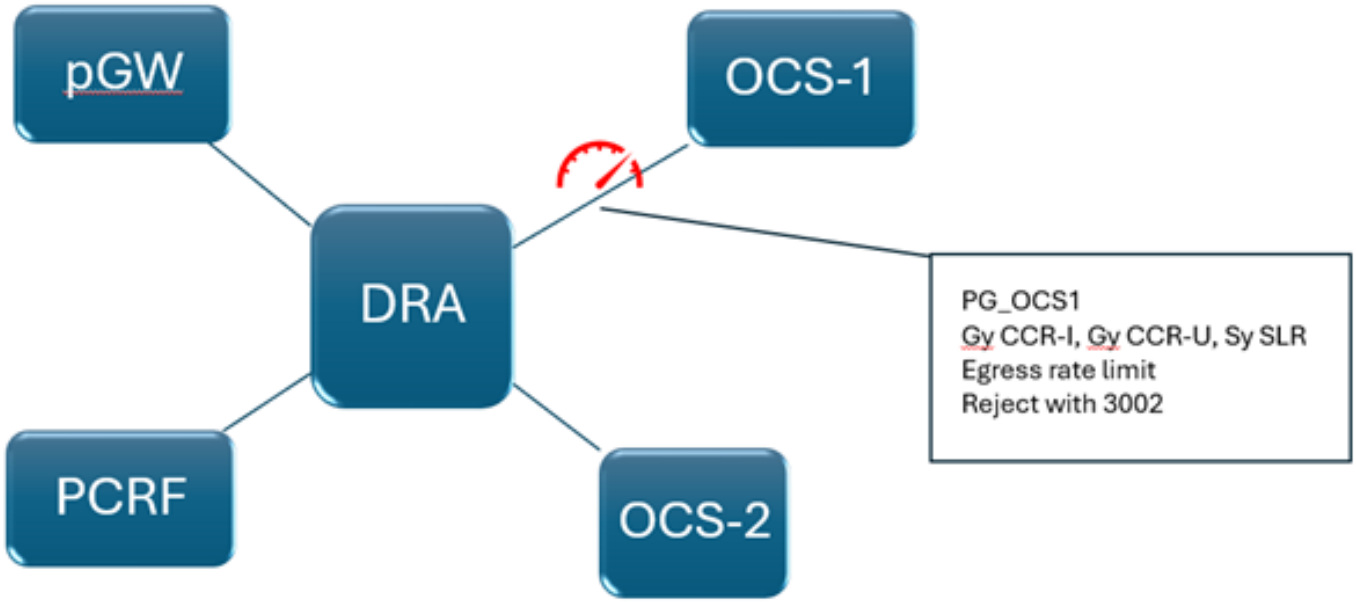
Considere un ejemplo de una transacción Gy que se intercambia entre pGW y Online Charging System (OCS). En caso de lentitud en la red en el canal DRA-OCS (debido al alto tráfico de pGW o a cualquier otro problema de red), la solicitud se agota debido a una infracción del SLA. Estos tiempos de espera no solo afectan a DRA, sino a toda la red.

Los recursos DRA se retrasan al intentar enviar la solicitud a OCS a través de la red lenta, lo que provoca que se agoten sus recursos. Esto da lugar a que DRA rechace varias solicitudes, aunque la capacidad asignada de DRA no se ha infringido.

Esto también afecta al tráfico que no se encuentra en el canal DRA-OCS. Estos rechazos/tiempos de espera y caídas activan la reconexión en varios elementos de la red.

En este caso, debe acelerar en el lado de salida: DRA acelera el mensaje justo antes de reenviarlo a OCS, en función del grupo de pares de OCS (que tiene limitaciones de capacidad o problemas de red).

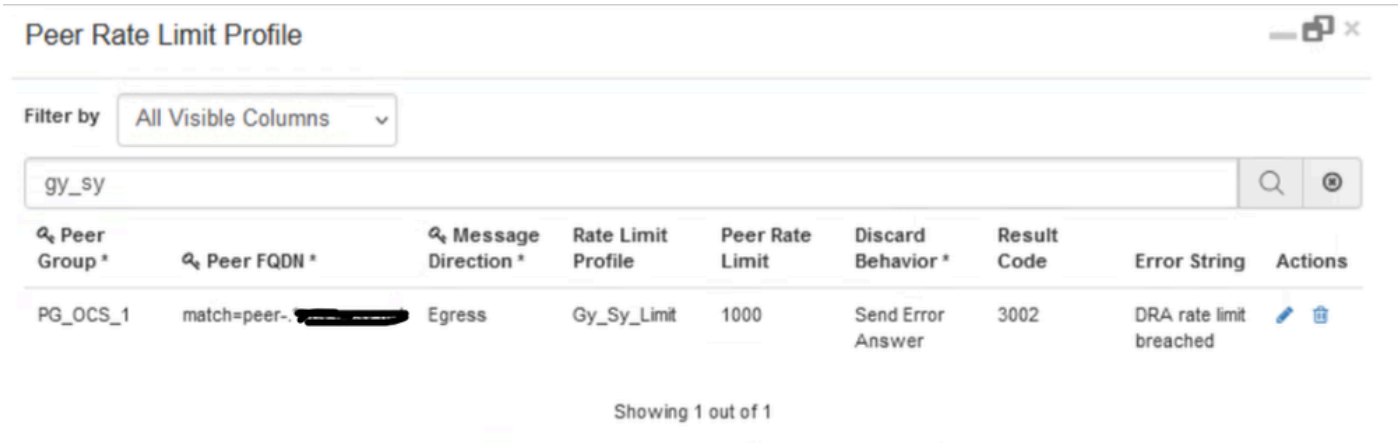
El comportamiento del acelerador puede ser rechazar el mensaje con un código de error y un mensaje de error concretos o descartarlo.



Este comportamiento se puede habilitar en CPS vDRA configurando las tablas CRD 'Peer Rate Limit Profile' y 'Message Rate Limit Profile'. En estas tablas de CRD, debe configurar estos valores:

Grupo de pares	Un grupo de peers es una agrupación lógica de nodos Diameter basada en su rango y host. Necesita configurar el grupo de peers que necesita ser acelerado.
FQDN de peer	FQDN (coincidencia exacta o regex) para los pares del grupo de pares que necesita para limitar la velocidad.
Dirección del mensaje	Dirección del límite: entrada o salida. En este caso, Egress.
Perfil de límite de velocidad	Nombre de perfil de límite de velocidad de mensajes que se utilizó para definir el tipo de mensaje que debe acelerarse.
Límite de velocidad de peer	Velocidad de solicitudes permitida para este grupo de pares. Esto incluye todos los tipos de mensajes de ese grupo de pares.
Descartar comportamiento	Puede optar por rechazar la solicitud o rechazarla con un código de error.

Código de resultado	Valor del código de resultado en caso de que rechace los mensajes. No se aplica en caso de que se descarten los mensajes.
Cadena de error	Cadena de error que se utiliza en el mensaje de respuesta de la solicitud que se rechaza. No se aplica en caso de que el mensaje se descarte.
Identificador de aplicación	ID de aplicación del mensaje que se va a limitar.
Código de comando	Código de comando del mensaje que se va a limitar.
Tipo de mensaje/solicitud	ID de aplicación y Tipo de solicitud de las solicitudes que deben acelerarse.
Límite de velocidad de mensajes	TPS de la solicitud de ese tipo de mensaje que procesa DRA. Las solicitudes más allá de este TPS deben acelerarse. Este valor es por peer en el grupo de peers.

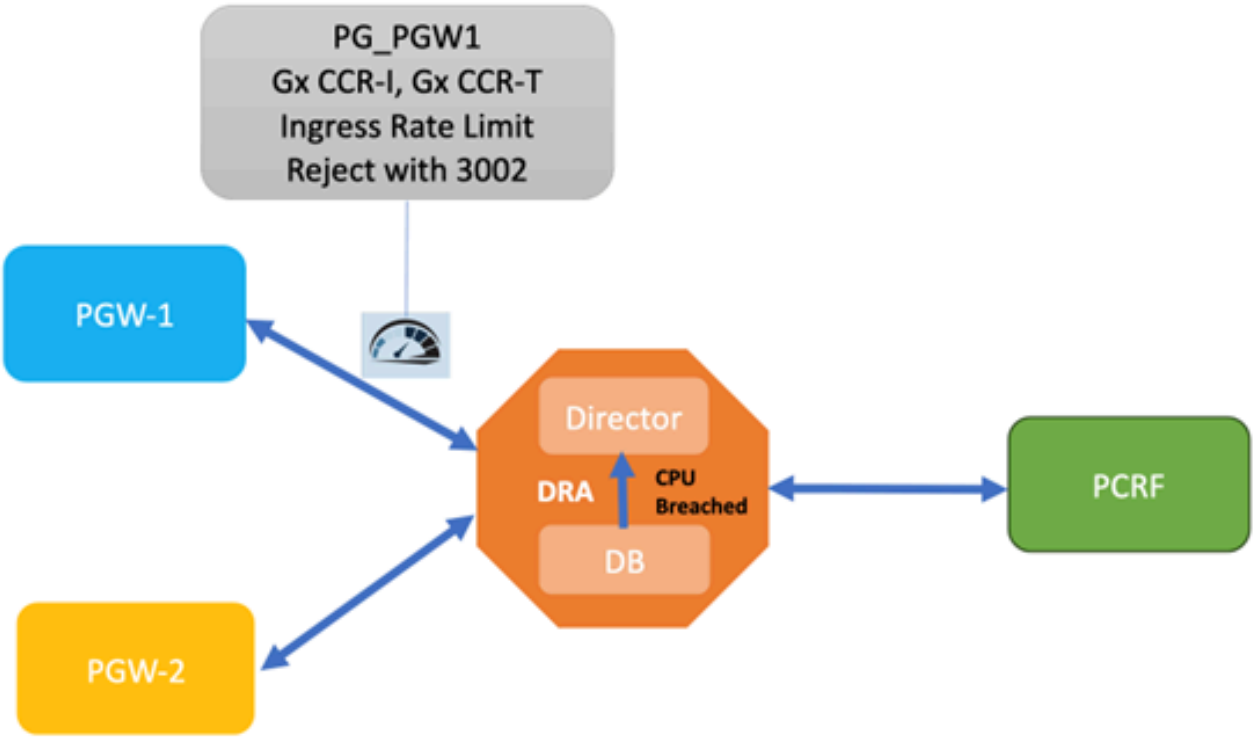


Límite de velocidad dinámica

Cuando se producen ráfagas CCR-I o CCR-T, puede producirse una sobrecarga en la base de datos (DB), que puede provocar la desestabilización del sistema. Para superar este problema, DRA admite la limitación de velocidad dinámica (sólo para Gx CCR-I y Gx CCR-T) en función de la capacidad de BD disponible.

DRA supervisa la utilización de la CPU de la base de datos y, siempre que se supera el umbral, regula las solicitudes entrantes. Los umbrales de CPU para la regulación y el tráfico entrante que se va a regular son configurables.

Se pueden configurar diferentes umbrales de CPU con los porcentajes de aceleración correspondientes. DRA ajusta el nivel de aceleración en función del uso actual de la CPU de la base de datos. Cuando el uso de la CPU se vuelve estable, la regulación se detiene gradualmente.



Este comportamiento se puede habilitar en CPS vDRA configurando las tablas de CRD 'Peer Rate Limit Profile', 'Message Rate Limit Profile', 'Dynamic Peer Rate Limit Profile' y 'Dynamic Throttling DB CPU Profile'. En estas tablas de CRD, debe configurar estos valores:

Grupo de pares	Un grupo de peers es una agrupación lógica de nodos Diameter basada en su rango y host. En este ejemplo usted configura el grupo de peers del pGW.
FQDN de peer	FQDN (coincidencia exacta o regex) para los pares del grupo de pares que necesita para limitar la velocidad.
Tipo de mensaje/solicitud	ID de aplicación y Tipo de solicitud de las solicitudes que deben acelerarse. En este ejemplo Gx CCR-I, Gx CCR-T.
Dirección del mensaje	Dirección del límite: entrada o salida. En este caso, Ingress.
Perfil de límite de velocidad	Nombre de perfil de límite de velocidad de mensajes que se utilizó para definir el tipo de mensaje que debe acelerarse.
Límite de velocidad de peer	Velocidad de solicitudes permitida para este grupo de pares.

	Esto incluye todos los tipos de mensajes de ese grupo de pares.
Descartar comportamiento	Puede optar por rechazar la solicitud o rechazarla con un código de error.
Código de resultado	Valor del código de resultado en caso de que rechace los mensajes. No se aplica en caso de que se descarten los mensajes.
Cadena de error	Cadena de error que se utiliza en el mensaje de respuesta de la solicitud que se rechaza. No se aplica en caso de que el mensaje se descarte.
Identificador de aplicación	ID de aplicación del mensaje que se va a limitar.
Código de comando	Código de comando del mensaje que se va a limitar.
Límite de velocidad de mensajes	TPS de la solicitud de ese tipo de mensaje que procesa DRA. Las solicitudes más allá de este TPS son aceleradas. Este valor es por peer en el grupo de peers.
Perfil de CPU de BD de aceleración dinámica	Hace referencia al nombre del perfil de CPU, que se utiliza para definir el porcentaje de aceleración para un intervalo de CPU diferente.
Umbral de utilización de CPU DB	Puede elegir el valor correcto de los niveles de CPU configurados como límites de infracción según el patrón de tráfico de su implementación.
Porcentaje del acelerador	% de límite de velocidad que se aplica cuando se infringe el nivel de CPU correspondiente.

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

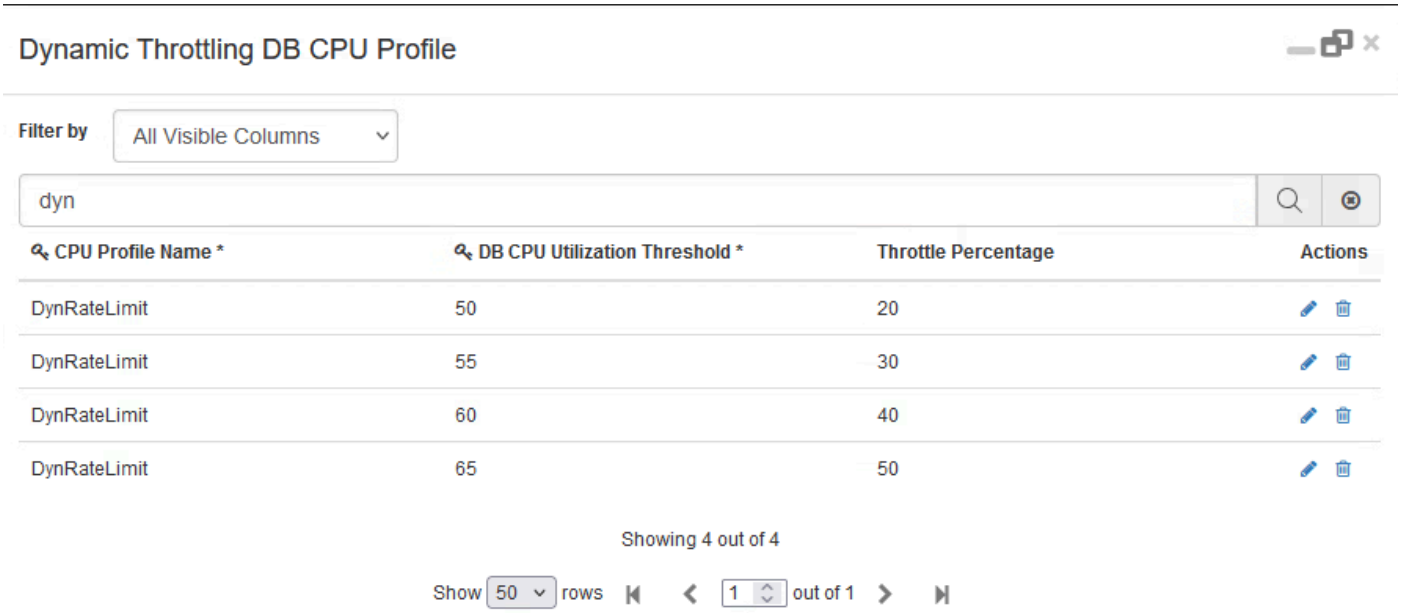
Filter by: All Visible Columns

DynRateLimit

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

Showing 1 out of 1

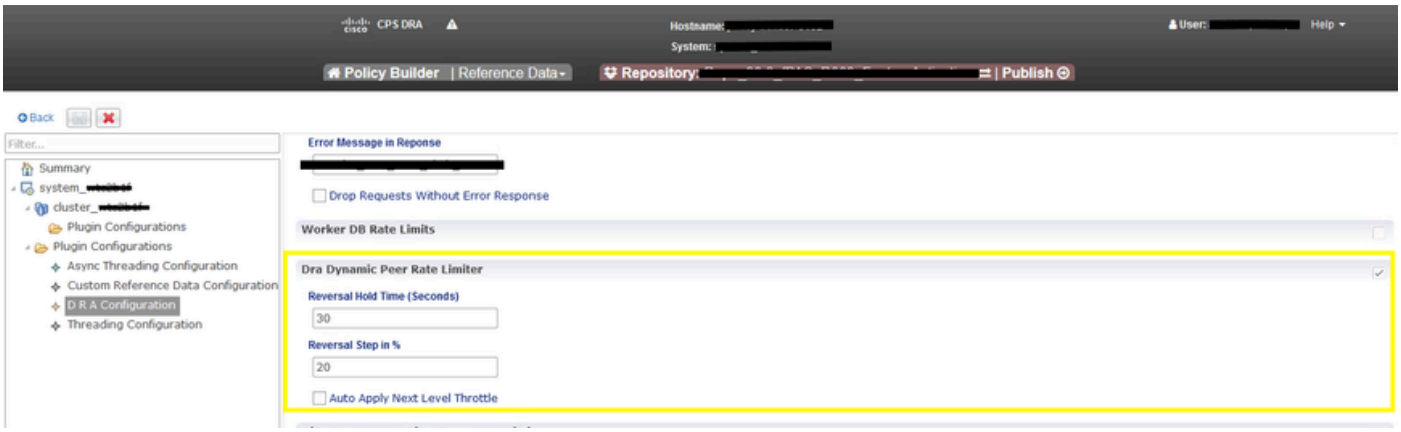
Show 50 rows 1 out of 1



Además, este comportamiento debe estar habilitado, activando la casilla de verificación en Policy Builder, en "Complemento de configuración DRA", en la sección "Limitador de velocidad de par dinámica DRA".

Tiempo de Retención de Reversión: Período de tiempo para el que se controla el uso de la CPU antes de aplicar la reversión.

Paso de Reversión en %: el porcentaje de aceleración que se invierte.



Situación: Limitación de velocidad dinámica basada en el uso de la CPU

Considere esta configuración en DRA:

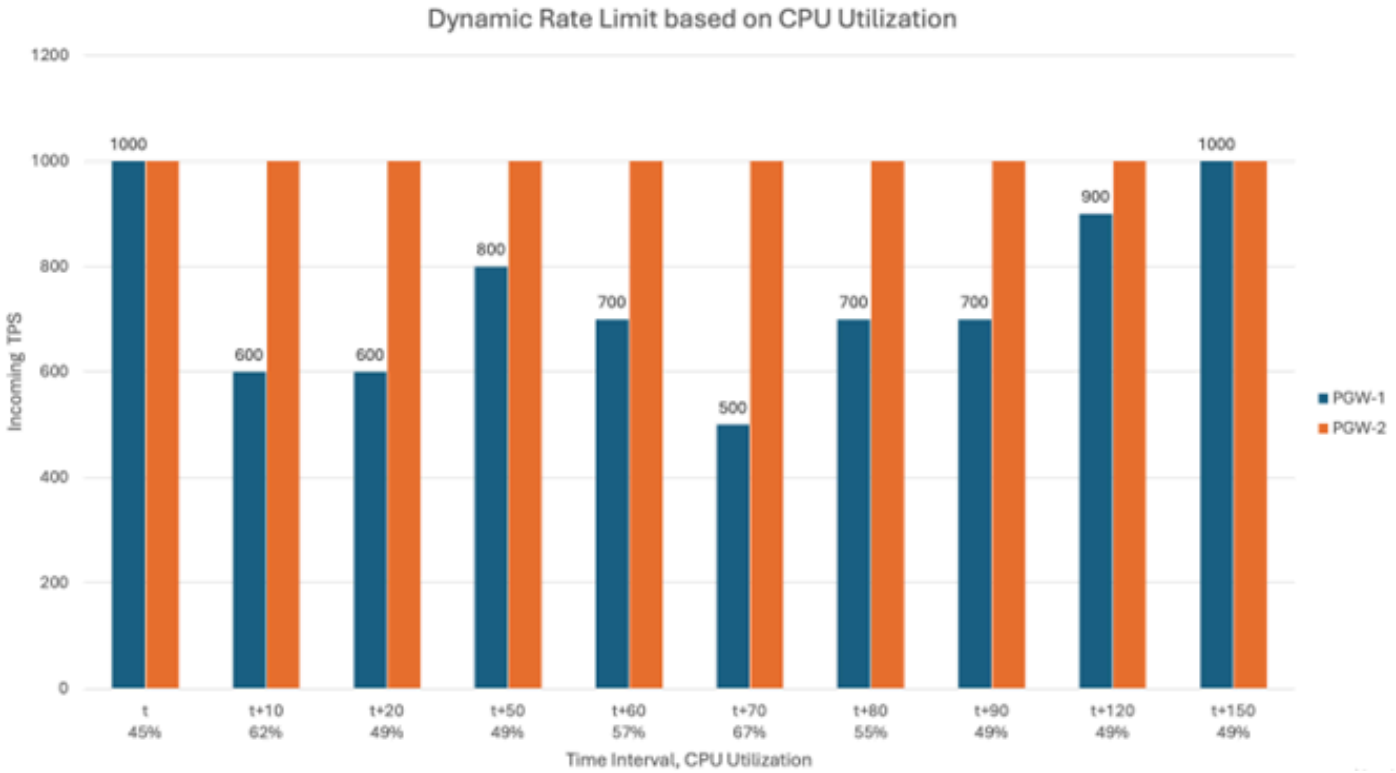
Límite de velocidad de mensajes estáticos: 1000 (es por lo tanto el valor del TPS entrante)

Tiempo de espera de reversión: 30 s

Paso de reversión en %: 20%

Cada vez que la utilización de la CPU de la base de datos cruza el umbral, se refiere a la configuración del "perfil de CPU de la base de datos de aceleración dinámica" y regula el TPS

entrante en consecuencia notificándolo al director. Dado que se está limitando en función de los siempre cambiantes valores de utilización de la CPU, puede decir que su velocidad dinámica limita el tráfico.



- Inicialmente, la utilización de la CPU de la base de datos está por debajo del límite, por lo que no se produce aceleración. Además, PGW-2 no tiene la configuración de limitación de velocidad dinámica y, por lo tanto, no se produce ninguna limitación independientemente de la utilización de la CPU.
- Cuando la utilización de la CPU de la base de datos es del 62%, el tráfico se limita en un 40% y el límite de velocidad efectivo es de 600 (el TPS entrante es de 1000, el DRA sólo permite 600).
- Si la utilización de la CPU permanece entre el 60 y el 65%, se sigue aplicando la aceleración del 40% en el límite de velocidad configurado de 1000 y el límite de velocidad efectivo es 600 (el TPS entrante es 1000, el DRA permite solo 600).
- La utilización de la CPU se reduce al 49%, la inversión de la aceleración comienza en pGW-1.
- Si la utilización de la CPU se mantiene en el 49% o menos durante 30 segundos, la aceleración se reduce entre un 20% y un 20%. Ahora el límite de velocidad efectivo es 800 (TPS entrante es 1000, DRA permite solo 800). Mientras que la inversión, según la configuración, se realiza en los pasos del 20%.
- Cuando la utilización de la CPU de la base de datos aumenta hasta el 57%, el tráfico se limita en un 30% y el límite de velocidad efectivo es de 700 (el TPS entrante es de 1000, el DRA sólo permite 700).
- Cuando la utilización de la CPU de la base de datos aumenta hasta el 67%, el tráfico se limita en un 50% y el límite de velocidad efectivo es 500 (el TPS entrante es 1000, el DRA solo permite 500).
- Cuando la utilización de la CPU de la base de datos disminuye al 55%, el tráfico se limita en

un 30% y el límite de velocidad efectivo es de 700 (el TPS entrante es de 1000, el DRA sólo permite 700).

- Si la CPU desciende al 49% o menos durante el siguiente intervalo de 30 segundos, la aceleración se reduce entre un 20% y un 10% y el límite de velocidad efectivo es de 900 (el TPS entrante es de 1000, el DRA solo permite 900).
- Si la CPU permanece además en un 49% o menos durante el siguiente intervalo de 30 segundos, la aceleración se reduce en un 20% a 0 y no se aplica ningún límite de velocidad cuando se completa la inversión (el TPS entrante es 1000, el DRA permite 1000).

Acerca de esta traducción
Cisco ha traducido este documento combinando la traducción automática y los recursos humanos a fin de ofrecer a nuestros usuarios en todo el mundo contenido en su propio idioma.

Tenga en cuenta que incluso la mejor traducción automática podría no ser tan precisa como la proporcionada por un traductor profesional.

Cisco Systems, Inc. no asume ninguna responsabilidad por la precisión de estas traducciones y recomienda remitirse siempre al documento original escrito en inglés (insertar vínculo URL).