Limitação de taxa estática e dinâmica preventiva com CPS vDRA

Contents

Introdução

Este documento descreve as opções de limite de taxa no DRA, um componente de telecomunicações que roteia mensagens de Diâmetro e gerencia o tráfego de rede.

Pré-requisitos

Requisitos

A Cisco recomenda que você tenha conhecimento destes tópicos:

- Agente de roteamento de diâmetro (vDRA) do Cisco Policy Suite (CPS)
- Fundamentos e especificações do agente de roteamento de diâmetro

Componentes Utilizados

As informações neste documento são baseadas no DRA do Cisco Policy Suite.

As informações neste documento foram criadas a partir de dispositivos em um ambiente de laboratório específico. Todos os dispositivos utilizados neste documento foram iniciados com uma configuração (padrão) inicial. Se a rede estiver ativa, certifique-se de que você entenda o impacto potencial de qualquer comando.

Informações de Apoio

A DRA é um componente das redes de telecomunicações, particularmente dentro do contexto das redes baseadas em protocolos Diameter. O DRA roteia eficientemente mensagens de Diâmetro entre diferentes elementos de rede, como servidores de política, sistemas de carregamento e outros dispositivos habilitados para Diâmetro. A limitação de taxa é uma técnica de gerenciamento de tráfego de rede usada para controlar a quantidade de tráfego de ou para um elemento de rede. Ele ajuda a garantir que os recursos da rede não sejam esgotados, mantém a qualidade do serviço e evita o uso indevido ou abuso da rede.

Problema

Cada componente na rede pode lidar com a carga de tráfego com base em sua capacidade nominal, mas em tempo real pode haver cenários onde o tráfego gerado é mais do que o sistema pode lidar. Alguns deles são:

- Comportamento do usuário - Atividades como streaming de eventos ou atualizações de software que geram grandes quantidades de dados em um período curto. Geralmente enviado do Gateway (Gw) para o DRA.
- Congestionamento de rede - Em períodos de alto uso da rede, o congestionamento pode aumentar, levando a dados em fila que são enviados em surtos quando a capacidade se torna disponível.
- Mecanismos de resiliência de rede - redirecionamento de tráfego durante interrupções ou manutenção, causando picos temporários. Isso pode afetar o fluxo de tráfego em sites acasalados que não têm nenhum problema de rede.
- Comportamento do elemento de rede - Em caso de sobrecarga e congestionamento, você pode começar a ver nenhuma resposta/timeout de um ou mais elementos de rede que podem causar a reconexão, contribuindo para sobrecarga adicional no sistema.
- Liberação de gateway - o gateway pode liberar as sessões existentes devido a alterações de política, alteração de topologia ou qualquer atividade de manutenção ou solução de problemas. Durante esses cenários, as sessões são limpas e você pode receber uma rajada de solicitações de CCR (Credit Control Request, Solicitação de controle de crédito)-T Gx.

Solução

A DRA pode distribuir a carga entre vários servidores de Diâmetro para garantir um tratamento eficiente das solicitações e evitar sobrecarregar um único servidor. Em caso de falha do servidor, o DRA pode redirecionar mensagens para servidores alternativos, garantindo alta disponibilidade e confiabilidade dos serviços de rede.

Limitação de taxa no DRA, não apenas protege o DRA, mas também outras entidades, garantindo um fluxo controlado de mensagens. Os principais benefícios da limitação de taxa são:

- Continuidade do serviço - Manter a disponibilidade contínua do serviço, garantindo que os componentes críticos da rede não fiquem sobrecarregados e evitando interrupções.
- Escalabilidade - Permitindo que a rede lide com cargas variáveis sem degradação no desempenho.
- Conformidade com SLAs (Service Level Agreements, contratos de nível de serviço) -

garantir que a rede atenda aos SLAs mantendo níveis consistentes de desempenho e confiabilidade.

Limite de taxa estática no balanceador de carga

Essa é uma abordagem direta, em que um limite fixo é definido com base na capacidade nominal do DRA/Packet Gateway (pGW) e de outros elementos da rede e não é alterado com base nas condições da rede ou nos recursos do sistema. Ao limitar a taxa de solicitações de entrada, você tem um resultado previsível sobre a quantidade de tráfego que o DRA processa.

As configurações para limite de taxa estática dependem do caso de uso para o qual ele é aplicado.

Limite de taxa de ingresso

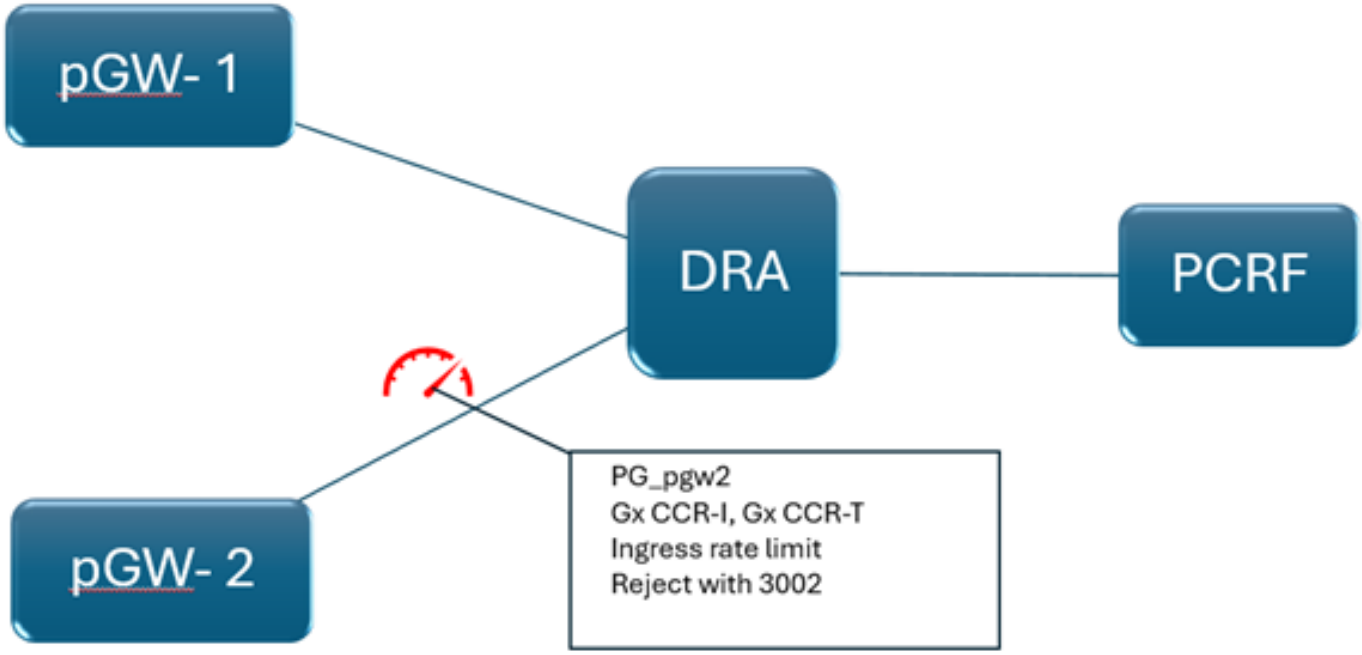
Cenário: Intermitências de pGW

São configurados limites específicos para pGWs que são susceptíveis a essas intermitências de tráfego. O valor deve ser obtido com base no tráfego regular/números de pico de tráfego que podem ser vistos durante esses bursts.

Os números de limiar podem ser definidos especificamente para cada tipo de mensagem para garantir que apenas o tráfego de intermitência seja limitado, como, somente as solicitações Gx CCR-I e Gx CCR-T de um GW devem ser limitadas, mas o tráfego Gx CCR-U ou Gy é encaminhado como recebido.

Nesse caso, você pode acelerar no lado do ingresso, ou seja, o DRA acelera a mensagem assim que ela é recebida, já que a finalidade aqui é rejeitar com base no elemento de rede do qual ela está recebendo a solicitação e evitar o processamento de um número maior de solicitações do que o DRA pode tratar.

O comportamento do acelerador pode ser rejeitar a mensagem com um determinado Error-Code e Error-Message ou descartá-la.



Este comportamento pode ser habilitado no CPS vDRA configurando as tabelas de dados de referência personalizados (CRD) 'Perfil de limite de taxa de mesmo nível' e 'Perfil de limite de taxa de mensagem'. Nessas tabelas de CRD, você precisa configurar estes valores:

Grupo de Pares	Um grupo de pares é um agrupamento lógico de nós de Diâmetro com base em seu território e host. Você precisa configurar o grupo de peer que precisa ser acelerado.
Nome de Domínio Totalmente Qualificado (FQDN) de Mesmo Nível	FQDN (correspondência exata ou regex) para os correspondentes no grupo de correspondentes que você precisa para o limite de taxa.
Direção da mensagem	Direção da limitação - entrada ou saída. Neste caso - Ingress.
Perfil de Limite de Taxa	Nome do perfil de limite de taxa de mensagens usado para definir o tipo de mensagem que precisa ser limitado.
Limite de taxa de mesmo nível	Taxa de solicitações permitidas para este Grupo de Pares. Isso inclui todos os tipos de mensagem desse grupo de pares.
Comportamento de descarte	Você pode optar por descartar a solicitação ou rejeitá-la com um Código de Erro.

Código do resultado	Valor do código de resultado caso você esteja rejeitando as mensagens. Não aplicável caso as mensagens sejam descartadas.
String de erro	A cadeia de caracteres de erro usada na mensagem de resposta da solicitação que foi rejeitada. Não aplicável caso a mensagem seja descartada.
Identificador do aplicativo	ID do aplicativo da mensagem a ser limitada.
Código de Comando	Código de Comando da mensagem a ser limitada.
Tipo de Mensagem/Solicitação	ID do Aplicativo e Tipo de Solicitação das solicitações que precisam ser limitadas.
Limite de taxa de mensagens	TelePresence Server (TPS) da solicitação desse tipo de mensagem que será processada pelo DRA. As solicitações além deste TPS são limitadas. Este valor é por peer no grupo de peers.

Aqui está um exemplo, onde você está configurando um limite de taxa geral de 1000 mensagens para pGW2 e um limite de taxa de 200 Gx CCR-I e 200 Gx CCR-T. Qualquer solicitação além dessa taxa é rejeitada com um 3002 e uma mensagem de erro indica que o limite de taxa foi violado.

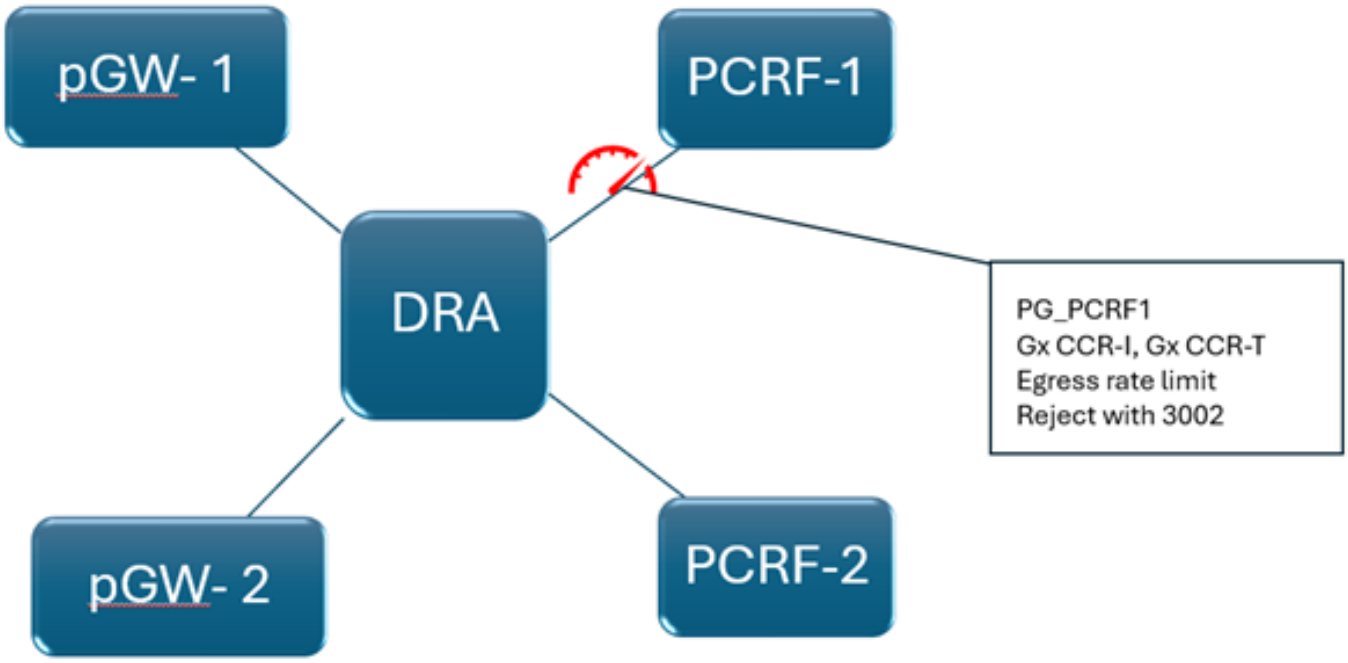
Limite de taxa de saída

Cenário: Proteção do elemento de rede que recebe a solicitação

Considere um exemplo de uma transação Gx em que a solicitação é recebida do pGW e encaminhada para a Policy and Charging Rules Function (PCRF). Se houver limitações para a quantidade de dados que o PCRF pode tratar, mesmo que o DRA possa tratar o tráfego de entrada, você pode usar o DRA para acelerar a mensagem no DRA em vez de encaminhar a solicitação ao PCRF e sobrecarregá-lo.

Aqui você precisa acelerar no lado de saída, isto é, o DRA acelera a mensagem antes de encaminhá-la para o PCRF, com base no grupo de peers do PCRF que é identificado com base na lógica de roteamento do DRA.

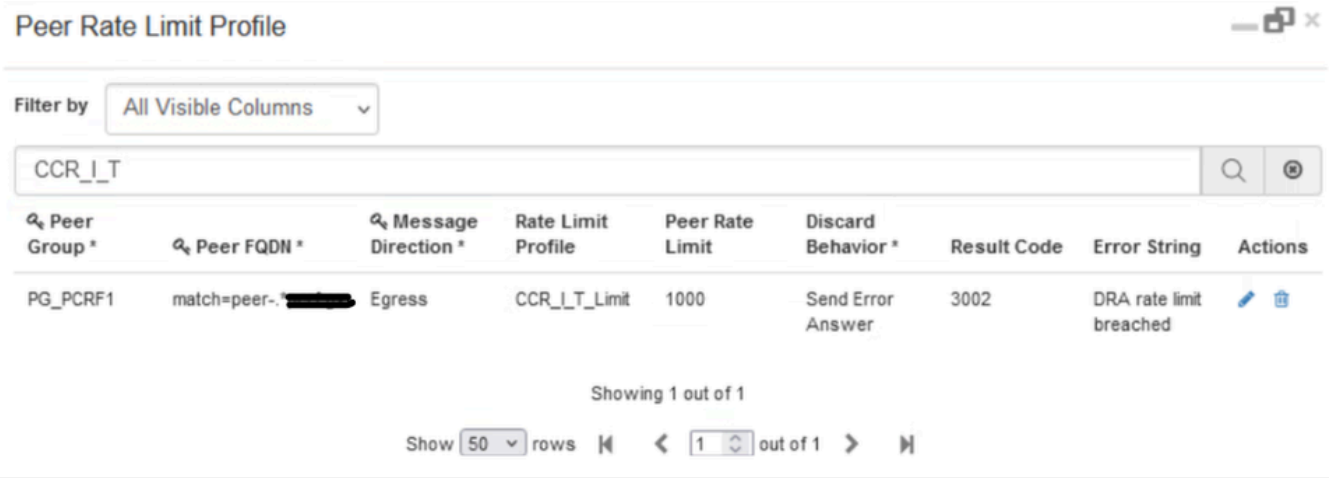
O comportamento do acelerador pode ser rejeitar a mensagem com um determinado Error-Code e Error-Message ou descartá-la.



Esse comportamento pode ser habilitado no vDRA do CPS configurando as tabelas CRD 'Peer Rate Limit Profile' e 'Message Rate Limit Profile'. Nessas tabelas CRD, você precisa configurar os seguintes valores:

Grupo de Pares	Um grupo de pares é um agrupamento lógico de nós de Diâmetro com base em seu território e host. Você precisa configurar o grupo de peer que precisa ser acelerado.
FQDN de Par	FQDN (correspondência exata ou regex) para os correspondentes no grupo de correspondentes que você precisa para o limite de taxa.
Direção da mensagem	Direção da limitação - entrada ou saída. Neste caso - Saída.
Perfil de Limite de Taxa	Nome do perfil de limite de taxa de mensagens usado para definir o tipo de mensagem que precisa ser limitado.
Limite de taxa de mesmo nível	Taxa de solicitações que devem ser permitidas para este Grupo de Mesmo Nível. Isso inclui todos os tipos de mensagem desse grupo de pares.
Comportamento de descarte	Você pode optar por descartar a solicitação ou rejeitá-la com um Código de Erro.

Código do resultado	Valor do código de resultado caso você esteja rejeitando as mensagens. Não aplicável caso as mensagens sejam descartadas.
String de erro	A cadeia de caracteres de erro usada na mensagem de resposta da solicitação que foi rejeitada. Não aplicável caso a mensagem seja descartada.
Identificador do aplicativo	ID do aplicativo da mensagem a ser limitada.
Código de Comando	Código de Comando da mensagem a ser limitada.
Tipo de Mensagem/Solicitação	ID do Aplicativo e Tipo de Solicitação das solicitações que precisam ser limitadas.
Limite de taxa de mensagens	TPS da solicitação desse tipo de mensagem processada pelo DRA. As solicitações além deste TPS são limitadas. Este valor é por peer no grupo de peers.



Cenário: Lentidão na rede, resultando em congestionamento de tráfego, causando falha total/parcial do DRA

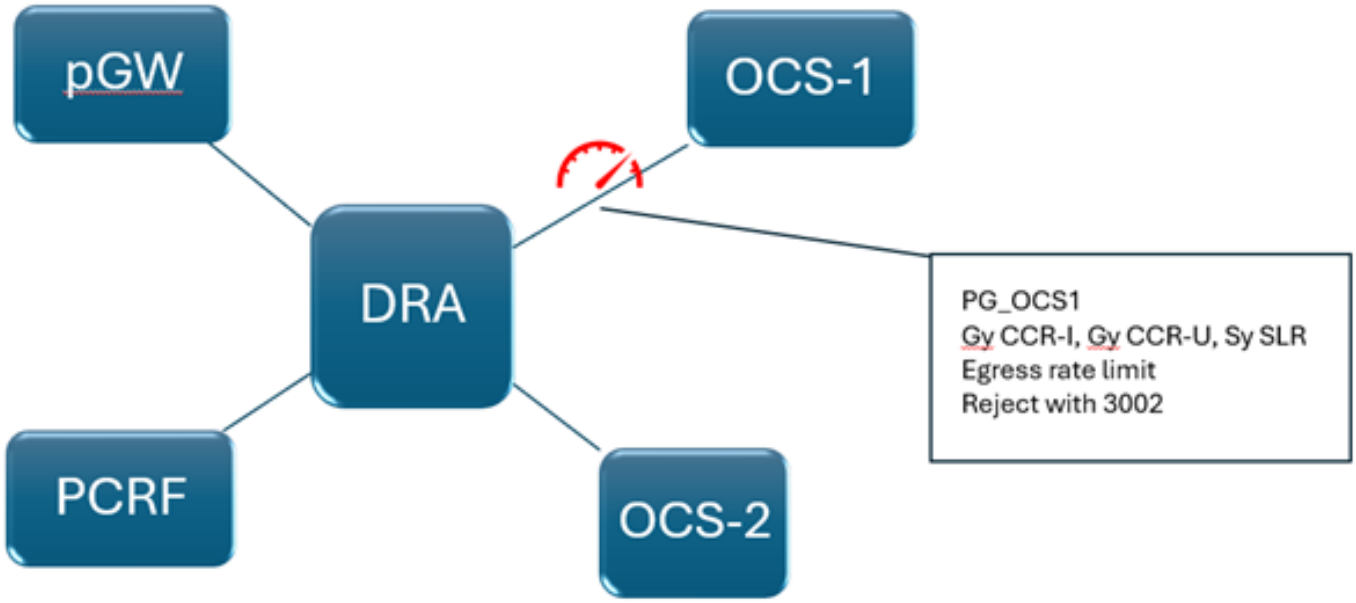
Considere um exemplo de uma transação Gy que é trocada entre o pGW e o Sistema de Cobrança On-line (OCS). Em caso de lentidão na rede no canal DRA-OCS (devido ao alto tráfego de pGW ou devido a qualquer outro problema de rede), a solicitação atinge o tempo limite devido à violação do SLA. Esses tempos limite afetam não apenas o DRA, mas toda a rede.

Os recursos de DRA são bloqueados ao tentar enviar a solicitação ao OCS pela rede lenta, fazendo com que seus recursos se esgotem. Isso faz com que várias solicitações sejam rejeitadas pela DRA, embora a capacidade nominal da DRA não seja violada.

Isso também afeta o tráfego que não está no canal DRA-OCS. Essas rejeições/intervalos e quedas acionam a reconexão em vários elementos da rede.

Nesse caso, você precisa acelerar no lado da saída - o DRA acelera a mensagem antes de encaminhá-la ao OCS, com base no grupo de pares do OCS que tem limitações de capacidade ou problemas de rede).

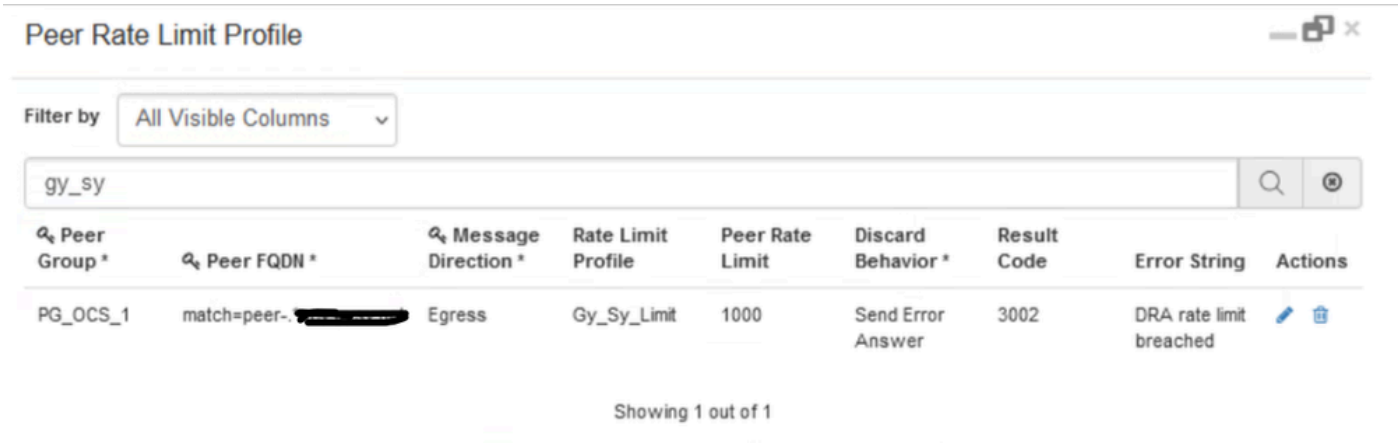
O comportamento do acelerador pode ser rejeitar a mensagem com um determinado Error-Code e Error-Message ou descartá-la.



Esse comportamento pode ser habilitado no vDRA do CPS configurando as tabelas CRD 'Peer Rate Limit Profile' e 'Message Rate Limit Profile'. Nestas tabelas CRD, você precisa configurar estes valores:

Grupo de Pares	Um grupo de pares é um agrupamento lógico de nós de Diâmetro com base em seu território e host. Você precisa configurar o grupo de peer que precisa ser acelerado.
FQDN de Par	FQDN (correspondência exata ou regex) para os correspondentes no grupo de correspondentes que você precisa para o limite de taxa.
Direção da mensagem	Direção da limitação - entrada ou saída. Neste caso - Saída.
Perfil de limite de taxa	Nome do perfil de limite de taxa de mensagens usado para definir o tipo de mensagem que precisa ser limitado.
Limite de taxa de mesmo nível	Taxa de solicitações permitidas para este Grupo de Pares. Isso inclui todos os tipos de mensagem desse grupo de pares.
Comportamento de descarte	Você pode optar por descartar a solicitação ou rejeitá-la com um Código de Erro.

Código do resultado	Valor do código de resultado caso você esteja rejeitando as mensagens. Não aplicável caso as mensagens sejam descartadas.
String de erro	A cadeia de caracteres de erro usada na mensagem de resposta da solicitação que foi rejeitada. Não aplicável caso a mensagem seja descartada.
Identificador do aplicativo	ID do aplicativo da mensagem a ser limitada.
Código de Comando	Código de Comando da mensagem a ser limitada.
Tipo de Mensagem/Solicitação	ID do Aplicativo e Tipo de Solicitação das solicitações que precisam ser limitadas.
Limite de taxa de mensagens	TPS da solicitação desse tipo de mensagem processada pelo DRA. As solicitações além deste TPS serão limitadas. Este valor é por peer no grupo de peers.

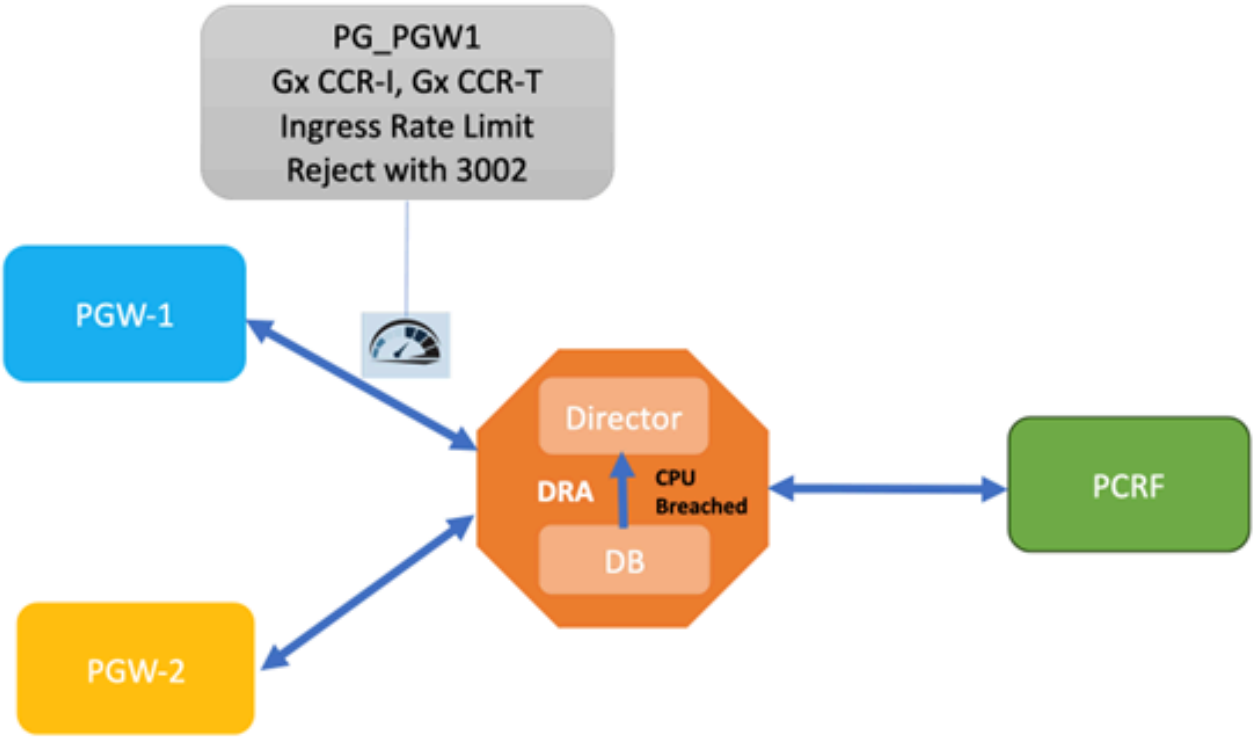


Limite de taxa dinâmica

Quando ocorrem intermitências de CCR-I ou CCR-T, pode haver uma sobrecarga no banco de dados (DB), que pode causar a desestabilização do sistema. Para superar isso, a DRA suporta limitação de taxa dinâmica (somente para Gx CCR-I e Gx CCR-T) com base na capacidade disponível do DB.

O DRA monitora a utilização da CPU do BD e, sempre que o limite é ultrapassado, ele controla o fluxo das solicitações recebidas. Os limites de CPU para controle e o tráfego de entrada a ser controlado são configuráveis.

Diferentes limites de CPU com percentuais de aceleração correspondentes podem ser configurados. O DRA ajusta o nível de otimização com base no uso atual da CPU do BD. Quando o uso da CPU se torna estável, o controle pára gradualmente.



Esse comportamento pode ser habilitado no CPS vDRA configurando as tabelas CRD 'Peer Rate Limit Profile', 'Message Rate Limit Profile', 'Dynamic Peer Rate Limit Profile' e 'Dynamic Throttling DB CPU Profile'. Nestas tabelas CRD, você precisa configurar estes valores:

Grupo de Pares	Um grupo de pares é um agrupamento lógico de nós de Diâmetro com base em seu território e host. Neste exemplo, você configura o grupo de peer do pGW.
FQDN de Par	FQDN (correspondência exata ou regex) para os correspondentes no grupo de correspondentes que você precisa para o limite de taxa.
Tipo de Mensagem/Solicitação	ID do Aplicativo e Tipo de Solicitação das solicitações que precisam ser limitadas. Neste exemplo, Gx CCR-I, Gx CCR-T.
Direção da mensagem	Direção da limitação - entrada ou saída. Neste caso - Ingress.
Perfil de limite de taxa	Nome do perfil de limite de taxa de mensagens usado para definir o tipo de mensagem que precisa ser limitado.

Limite de taxa de mesmo nível	Taxa de solicitações permitidas para este Grupo de Pares. Isso inclui todos os tipos de mensagem desse grupo de pares.
Comportamento de descarte	Você pode optar por descartar a solicitação ou rejeitá-la com um Código de Erro.
Código do resultado	Valor do código de resultado caso você esteja rejeitando as mensagens. Não aplicável caso as mensagens sejam descartadas.
String de erro	A cadeia de caracteres de erro usada na mensagem de resposta da solicitação que foi rejeitada. Não aplicável caso a mensagem seja descartada.
Identificador do aplicativo	ID do aplicativo da mensagem a ser limitada.
Código de Comando	Código de Comando da mensagem a ser limitada.
Limite de taxa de mensagens	TPS da solicitação desse tipo de mensagem processada pelo DRA. As solicitações além deste TPS são limitadas. Este valor é por peer no grupo de peers.
Perfil de CPU do Banco de Dados de Limitação Dinâmica	Refere-se ao nome do perfil de CPU, que é usado para definir a porcentagem de aceleração para diferentes intervalos de CPU.
Limite de Utilização da CPU do BD	Você pode escolher o valor correto dos níveis de CPU configurados como limites de violação de acordo com o padrão de tráfego em sua implantação.
Porcentagem de Aceleração	O % do limite de taxa aplicado quando o nível de CPU correspondente é violado.

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

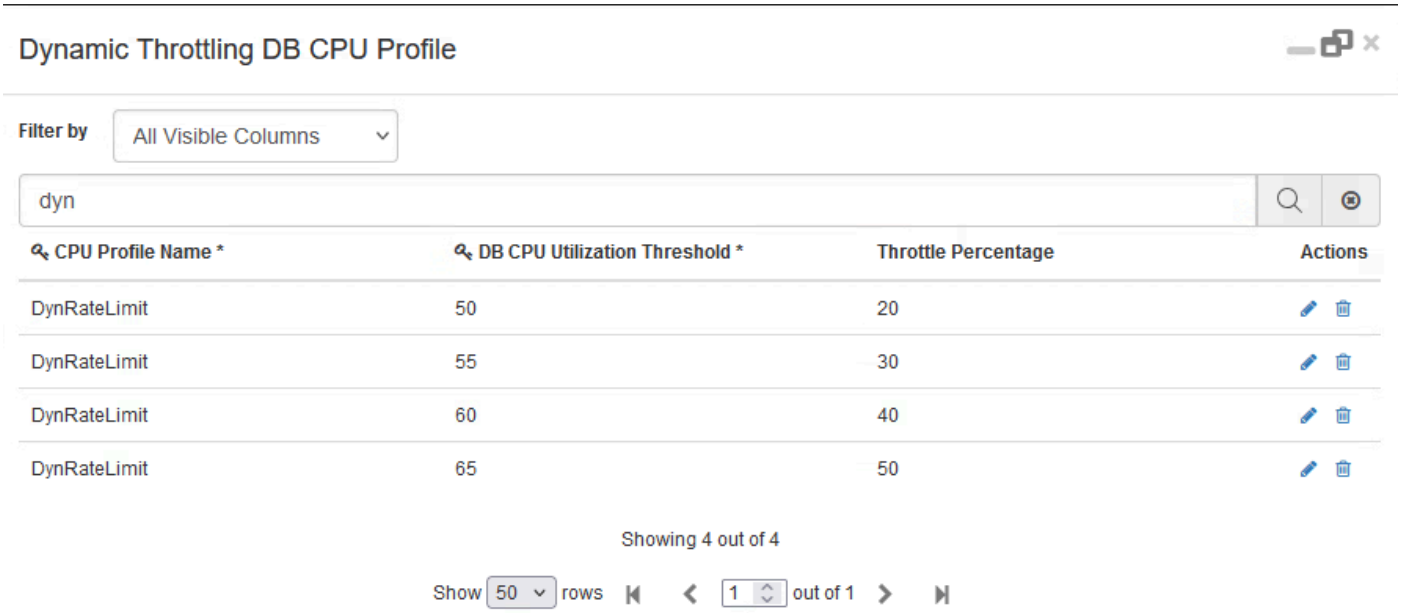
Filter by: All Visible Columns

DynRateLimit

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

Showing 1 out of 1

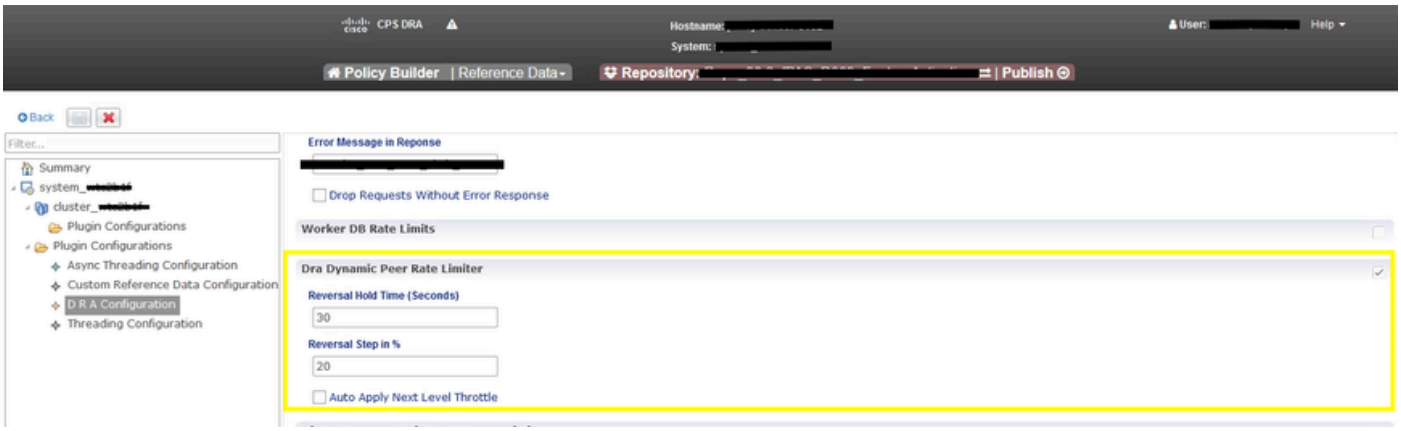
Show 50 rows 1 out of 1



Além disso, esse comportamento deve ser ativado, marcando a caixa de seleção no Policy Builder em 'DRA Configuration Plugin', na seção 'DRA Dynamic Peer Rate Limiter'.

Tempo de Retenção de Estorno - O período para o qual a utilização da CPU é monitorada antes da aplicação do estorno.

Etapa de Estorno em % - A porcentagem de limitação revertida.



Cenário: Limitação dinâmica de taxa com base na utilização da CPU

Considere esta configuração no DRA:

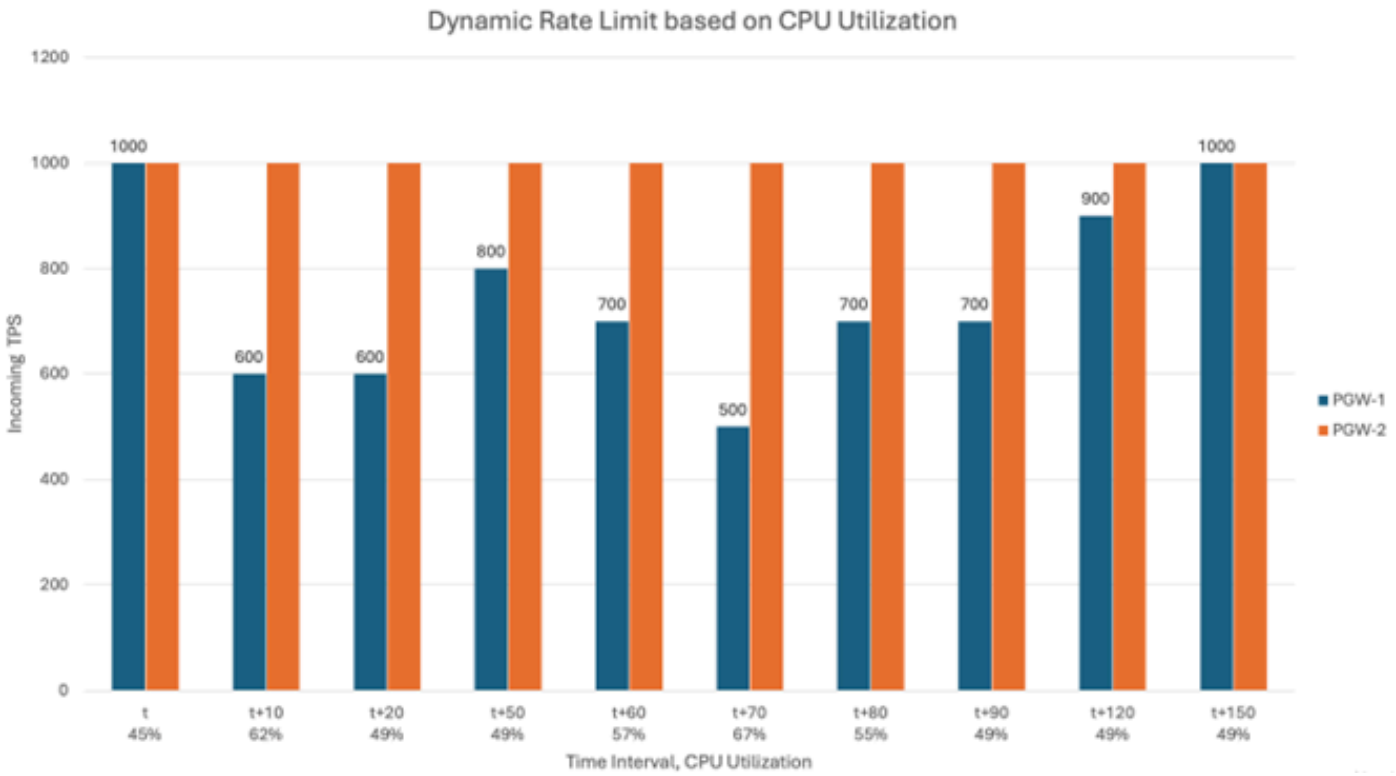
Limite de taxa de mensagens estáticas: 1000 (é, portanto, o valor de TPS de entrada)

Tempo de Retenção de Estorno: 30 s

Etapa de reversão em %: 20%

Sempre que a Utilização da CPU do BD ultrapassa o limite, ela se refere à configuração 'Dynamic Throttling DB CPU Profile' e acelera o TPS de entrada de acordo, notificando o direcionador. Como a limitação se baseia em valores de utilização da CPU que são sempre alterados, você

pode dizer que sua taxa dinâmica limita o tráfego.



- Inicialmente, a utilização da CPU do BD está abaixo do limite, portanto, não há limitação. Além disso, o PGW-2 não tem a configuração de Limitação de Taxa Dinâmica e, portanto, nenhum controle de fluxo acontece lá, independentemente da utilização da CPU.
- Quando a utilização da CPU do BD é de 62%, o tráfego é limitado em 40% e o limite de taxa efetivo é de 600 (TPS de entrada é 1000, DRA permite apenas 600).
- Se a utilização da CPU permanecer entre 60-65%, a limitação de 40% continuará a ser aplicada no limite de taxa configurado de 1000 e o limite de taxa efetivo será 600 (TPS de entrada é 1000, DRA permite apenas 600).
- A utilização da CPU é reduzida para 49%, a reversão da limitação começa em pGW-1.
- Se a utilização da CPU permanecer em 49% ou menos por 30 segundos, a limitação será reduzida em 20% a 20%. Agora o limite de taxa efetivo é 800 (TPS de entrada é 1000, DRA permite apenas 800.) Enquanto a reversão, de acordo com a configuração, é feita nas etapas de 20%.
- Quando a utilização da CPU do BD aumenta para 57%, o tráfego é limitado em 30% e o limite de taxa efetivo é 700 (o TPS de entrada é 1000, o DRA permite apenas 700).
- Quando a utilização da CPU do BD aumenta para 67%, o tráfego é limitado em 50% e o limite de taxa efetivo é 500 (o TPS de entrada é 1000, o DRA permite apenas 500).
- Quando a utilização da CPU do BD diminui para 55%, o tráfego é limitado em 30% e o limite de taxa efetivo é 700 (o TPS de entrada é 1000, o DRA permite apenas 700).
- Se a CPU cair para 49% ou menos nos próximos 30 segundos, a limitação será reduzida em 20% a 10% e o limite de taxa efetiva será 900 (TPS de entrada é 1000, DRA permite apenas 900).
- Se a CPU ficar além disso em 49% ou menos durante os próximos 30 segundos, a aceleração será reduzida em 20% a 0 e não haverá limite de taxa aplicado quando a reversão for concluída (TPS de entrada é 1000, DRA permite 1000).

Sobre esta tradução
A Cisco traduziu este documento com a ajuda de tecnologias de tradução automática e humana para oferecer conteúdo de suporte aos seus usuários no seu próprio idioma, independentemente da localização.

Observe que mesmo a melhor tradução automática não será tão precisa quanto as realizadas por um tradutor profissional.

A Cisco Systems, Inc. não se responsabiliza pela precisão destas traduções e recomenda que o documento original em inglês (link fornecido) seja sempre consultado.