CPS vDRAによるプリエンプティブ静的および動的レート制限

内容

はじめに

このドキュメントでは、Diameterメッセージをルーティングし、ネットワークトラフィックを管理する通信コンポーネントであるDRAのレート制限オプションについて説明します。

前提条件

要件

次の項目に関する知識があることが推奨されます。

- Cisco Policy Suite(CPS)Diameter Routing Agent(vDRA)
- Diameter Routing Agentの基本と仕様

使用するコンポーネント

このドキュメントの情報は、Cisco Policy Suite DRAに基づくものです。

このドキュメントの情報は、特定のラボ環境にあるデバイスに基づいて作成されました。このドキュメントで使用するすべてのデバイスは、クリアな（デフォルト）設定で作業を開始しています。本稼働中のネットワークでは、各コマンドによって起こる可能性がある影響を十分確認してください。

背景説明

DRAは、特にDiameterプロトコルベースのネットワークにおける通信ネットワークのコンポーネントです。DRAは、ポリシーサーバ、課金システム、その他のDiameter対応デバイスなどの異なるネットワーク要素間でDiameterメッセージを効率的にルーティングします。レート制限は、ネットワーク要素との間でやり取りされるトラフィックの量を制御するために使用されるネットワークトラフィック管理手法です。ネットワークリソースが枯渇しないようにし、QoSを維持し、ネットワークの誤使用や悪用を防止するのに役立ちます。

問題

ネットワーク内の各コンポーネントは、定格キャパシティに基づいてトラフィックの負荷を処理できますが、リアルタイムでは、生成されるトラフィックがシステムで処理できる量を超える場合があります。次のその一例を示します。

- ユーザの動作：短時間で大量のデータを生成するストリーミングイベントやソフトウェア更新などのアクティビティ。通常、ゲートウェイ(Gw)からDRAに送信されます。
- ネットワークの輻輳：ネットワークの使用率が高い時間帯には、輻輳が蓄積され、キューに入れられたデータが容量不足になったときにバースト送信されます。
- ネットワークの復元力メカニズム：サービスの停止やメンテナンス時にトラフィックを再ルーティングし、一時的にスパイクを引き起こします。これは、ネットワークに問題のないサイトのトラフィックフローに影響を与える可能性があります。
- ネットワーク要素の動作：過負荷と輻輳が発生した場合、1つ以上のネットワーク要素からの応答/タイムアウトが表示されなくなり、再接続が発生してシステムの過負荷がさらに進む可能性があります。
- ゲートウェイのフラッシュ：ゲートウェイは、ポリシー変更、トポロジ変更、メンテナンス、またはトラブルシューティングのアクティビティによって、既存のセッションをフラッシュできます。これらのシナリオ中にセッションはクリアされ、Gx Credit Control Request(CCR)-T要求のバーストを受信できます。

解決方法

DRAは、要求の効率的な処理を保証し、単一のサーバの過負荷を回避するために、複数のDiameterサーバ間で負荷を分散できます。サーバに障害が発生した場合、DRAは代替サーバにメッセージをリダイレクトし、ネットワークサービスの高可用性と信頼性を確保できます。

DRAのレート制限は、メッセージのフローを制御することで、DRAだけでなく他のエンティティも保護します。レート制限の主な利点は次のとおりです。

- サービスの継続性：重要なネットワークコンポーネントが過負荷になるのを防ぎ、サービスの停止を防ぐことで、継続的なサービスの可用性を維持します。
- 拡張性：パフォーマンスを低下させることなく、さまざまな負荷に対応できるネットワークを実現します。
- サービスレベル契約(SLA)へのコンプライアンス：パフォーマンスと信頼性のレベルを一貫して維持することで、ネットワークがSLAを確実に満たすようにします。

ロードバランサでの静的レート制限

これは、DRA/Packet Gateway(pGW)およびその他のネットワーク要素の定格容量に基づいて固定しきい値が設定され、ネットワーク条件またはシステムリソースに基づいて変更されない、単純なアプローチです。着信要求のレートを制限することにより、DRAが処理するトラフィックの量に関して予測可能な結果が得られます。

スタティックレート制限の設定は、それが適用されるユースケースによって異なります。

入力レート制限

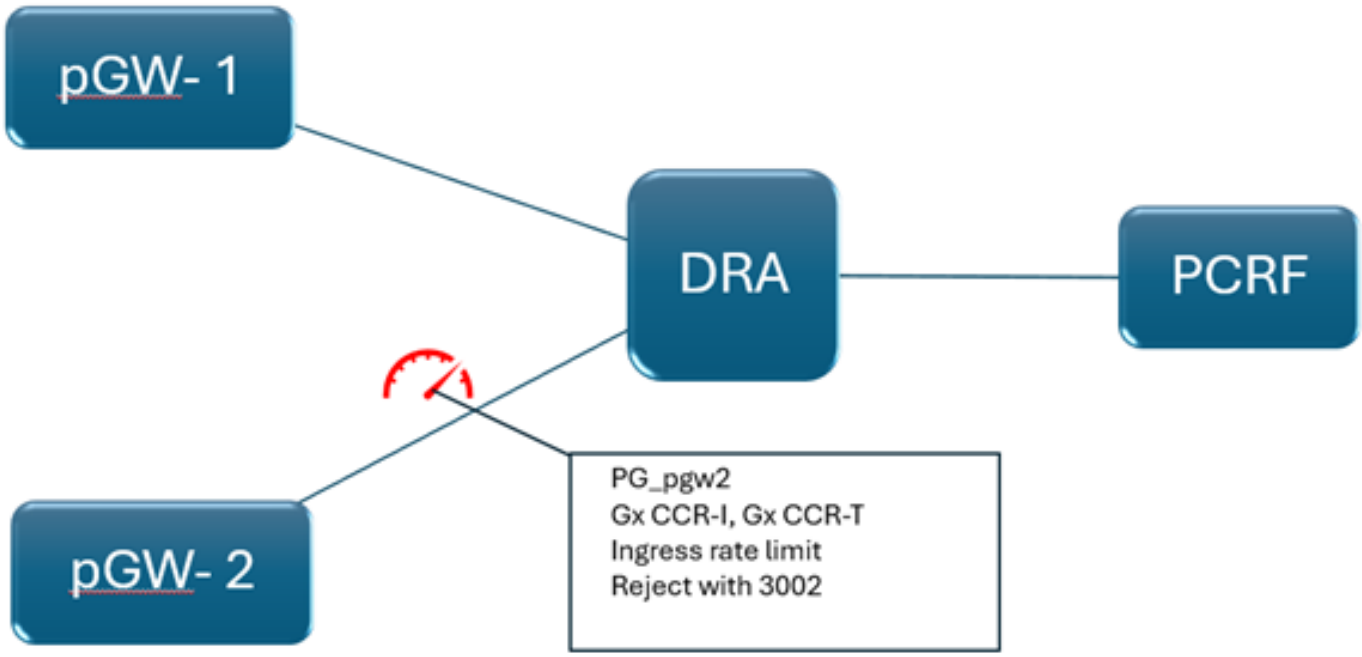
シナリオ：pGWからのバースト

これらのトラフィックバーストの影響を受けやすいpGWに固有のしきい値を設定します。この値は、これらのバースト時に見られる通常のトラフィック/ピークトラフィック数に基づいて算出する必要があります。

しきい値の数値は、バーストトラフィックだけを抑制するために各メッセージタイプに固有に定義できます。たとえば、ゲートウェイからのGx CCR-IおよびGx CCR-T要求だけを抑制する必要がありますが、Gx CCR-Uトラフィックまたはトラフィックは受信済みとして転送されます。

この場合、入力側でスロットルできます。つまり、DRAはメッセージを受信するとすぐにメッセージをスロットルします。これは、この目的が、要求を受信するネットワーク要素に基づいて拒否し、DRAが処理できる数よりも多い要求の処理を避けることであるためです。

スロットルの動作は、特定のエラーコードとエラーメッセージを含むメッセージを拒否するか、ドロップするかのいずれかです。



CPS vDRAでこの動作を有効にするには、カスタム参照データ(CRD)テーブルの「ピアレート制限プロファイル」および「メッセージレート制限プロファイル」を設定します。 これらのCRDテーブルでは、次の値を設定する必要があります。

ピアグループ	ピアグループは、レルムとホストに基づくDiameterノードの論理グループです。調整する必要があるピアグループを設定する必要があります。
ピアの完全修飾ドメイン名(FQDN)	レート制限が必要なピアグループ内のピアのFQDN（完全一致または正規表現一致）。
メッセージの方向	スロットリングの方向：入力または出力。この例では、入力です。
レートリミットプロファイル	スロットルが必要なメッセージタイプの定義に使用されるメッセージレート制限プロファイル名。
ピアレート制限	このピアグループに許可されている要求の割合です。これには、そのピアグループからのすべてのメッセージタイプが含まれます。
破棄動作	要求をドロップするか、エラーコードを使用して拒否するかを選択できます。
結果コード	メッセージを拒否する場合の結果コード値。メッセージがドロップされた場合は適用されません。
エラー文字列	拒否された要求の応答メッセージで使用されるエラー文字列。メッセージがドロップされた場合は適用されません。
アプリケーションID	調整するメッセージのアプリケーションID。
コマンドコード	調整するメッセージのコマンドコードです。
メッセージ/要求タイプ	調整する必要がある要求のアプリケーションIDおよび要求タイプ。
メッセージレート制限	DRAによって処理されるそのメッセージタイプの要求のTelePresence Server(TPS)。このTPSを超える要求は抑制されます。この値は、ピアグループ内のピア単位です。

次の例では、pGW2に対する全体的なレート制限である1000メッセージと、200 Gx CCR-Iおよび

200 Gx CCR-Tのレート制限を設定しています。このレートを超える要求は3002で拒否され、レート制限に違反したことを示すエラーメッセージが表示されます。

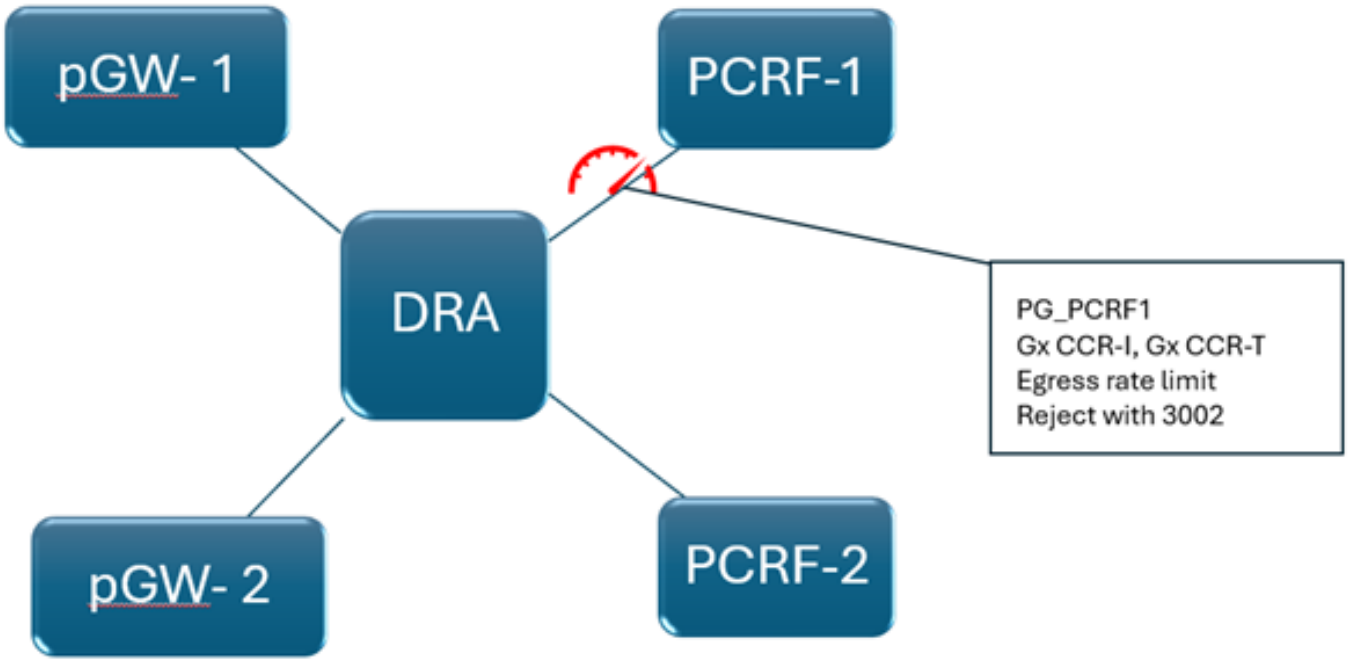
出力レート制限

シナリオ：要求を受信するネットワーク要素の保護

要求がpGWから受信され、ポリシー/課金ルール機能(PCRF)に転送されるGxトランザクションの例を考えてみます。 PCRFが処理できるデータ量に制限がある場合、DRAが着信トラフィックを処理できる場合でも、PCRFに要求を転送して過負荷にする代わりに、DRAを使用してDRAでメッセージを抑制することができます。

出力側でスロットルする必要があります。つまり、DRAは、DRAルーティングロジックに基づいて識別されるPCRFピアグループに基づいて、PCRFに転送される直前にメッセージをスロットルします。

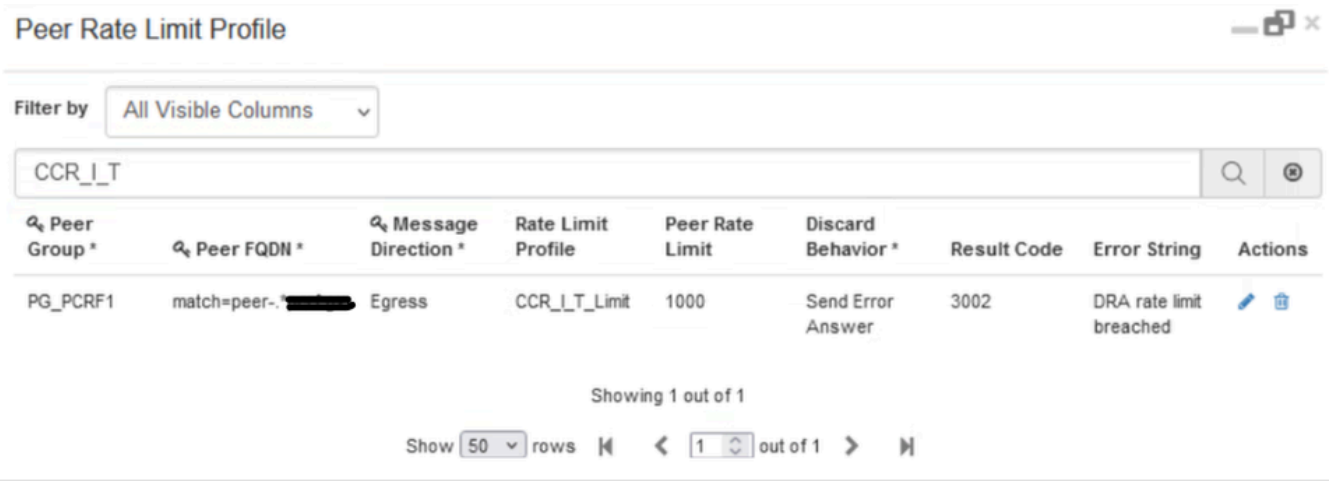
スロットルの動作は、特定のエラーコードとエラーメッセージを含むメッセージを拒否するか、ドロップするかのいずれかです。

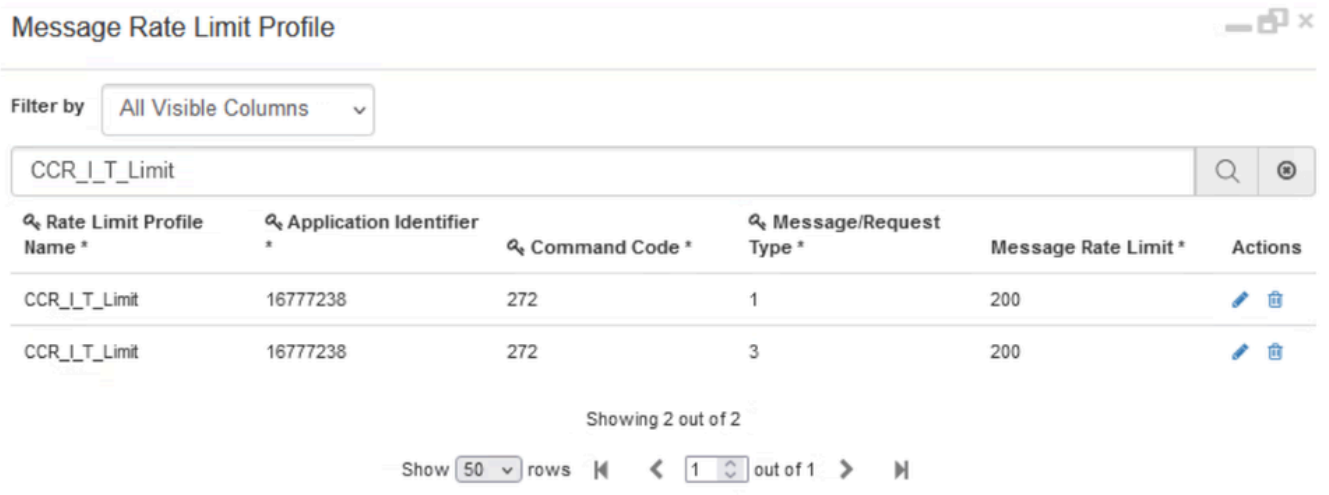


CPS vDRAでこの動作を有効にするには、CRDテーブルの「ピアレート制限プロファイル」および「メッセージレート制限プロファイル」を設定します。 これらのCRDテーブルでは、次の値を設定する必要があります。

ピアグループ	ピアグループは、レルムとホストに基づくDiameterノードの論理グループです。調整する必要があるピアグループを設定する必要があります。
ピアFQDN	レート制限が必要なピアグループ内のピアのFQDN（完全一致または正規表現一致）。
メッセージの方向	スロットリングの方向：入力または出力。この場合は出力です。
レートリミットプロファイル	スロットルが必要なメッセージタイプの定義に使用されるメッセージレート制限プロファイル名。
ピアレート制限	このピアグループに許可される要求の割合です。これには、そのピアグループからのすべてのメッセージタイプが含まれます。
破棄動作	要求をドロップするか、エラーコードを使用して拒否するかを選択できます。

結果コード	メッセージを拒否する場合の結果コード値。メッセージがドロップされた場合は適用されません。
エラー文字列	拒否された要求の応答メッセージで使用されるエラー文字列。メッセージがドロップされた場合は適用されません。
アプリケーションID	調整するメッセージのアプリケーションID。
コマンドコード	調整するメッセージのコマンドコードです。
メッセージ/要求タイプ	調整する必要がある要求のアプリケーションIDおよび要求タイプ。
メッセージレート制限	DRAによって処理されるそのメッセージタイプの要求のTPS。このTPSを超える要求は抑制されます。この値は、ピアグループ内のピア単位です。





シナリオ：ネットワークの速度低下によりトラフィックが輻輳し、DRAの完全な障害または部分的な障害が発生する

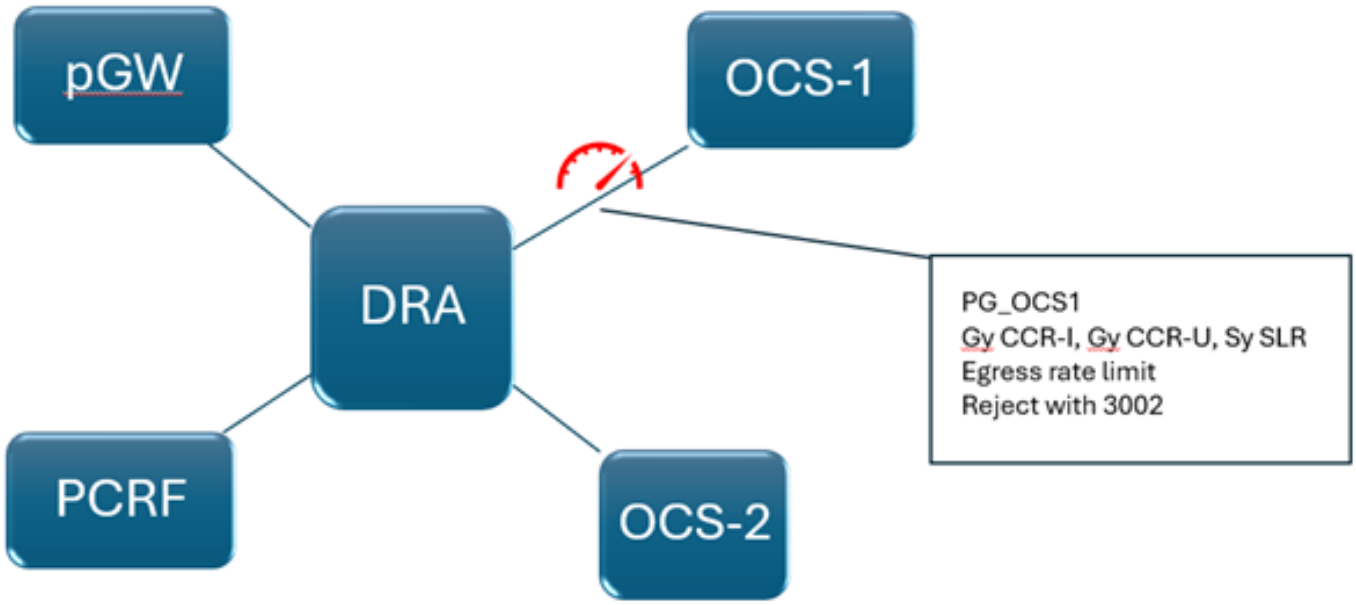
pGWとOnline Charging System(OCS)の間で交換されるGyトランザクションの例を考えてみましょう。 DRA-OCSチャネル上のネットワークの速度が遅い場合（pGWからのトラフィックが多いか、その他のネットワークの問題が原因）、SLA違反により要求がタイムアウトします。これらのタイムアウトは、DRAだけでなく、ネットワーク全体に影響を与えます。

低速ネットワークを介してOCSに要求を送信しようとすると、DRAリソースが停滞し、その結果、リソースが使い果たされます。この結果、DRAの定格容量は違反していませんが、複数の要求がDRAによって拒否されます。

これは、DRA-OCSチャネル上にないトラフィックにも影響します。これらの拒否/タイムアウトとドロップにより、複数のネットワーク要素で再接続がトリガーされます。

この場合、出力側でスロットルする必要があります。DRAは、OCSに転送される直前に、キャパシティ制限またはネットワークの問題があるOCSピアグループに基づいてメッセージをスロットルします。

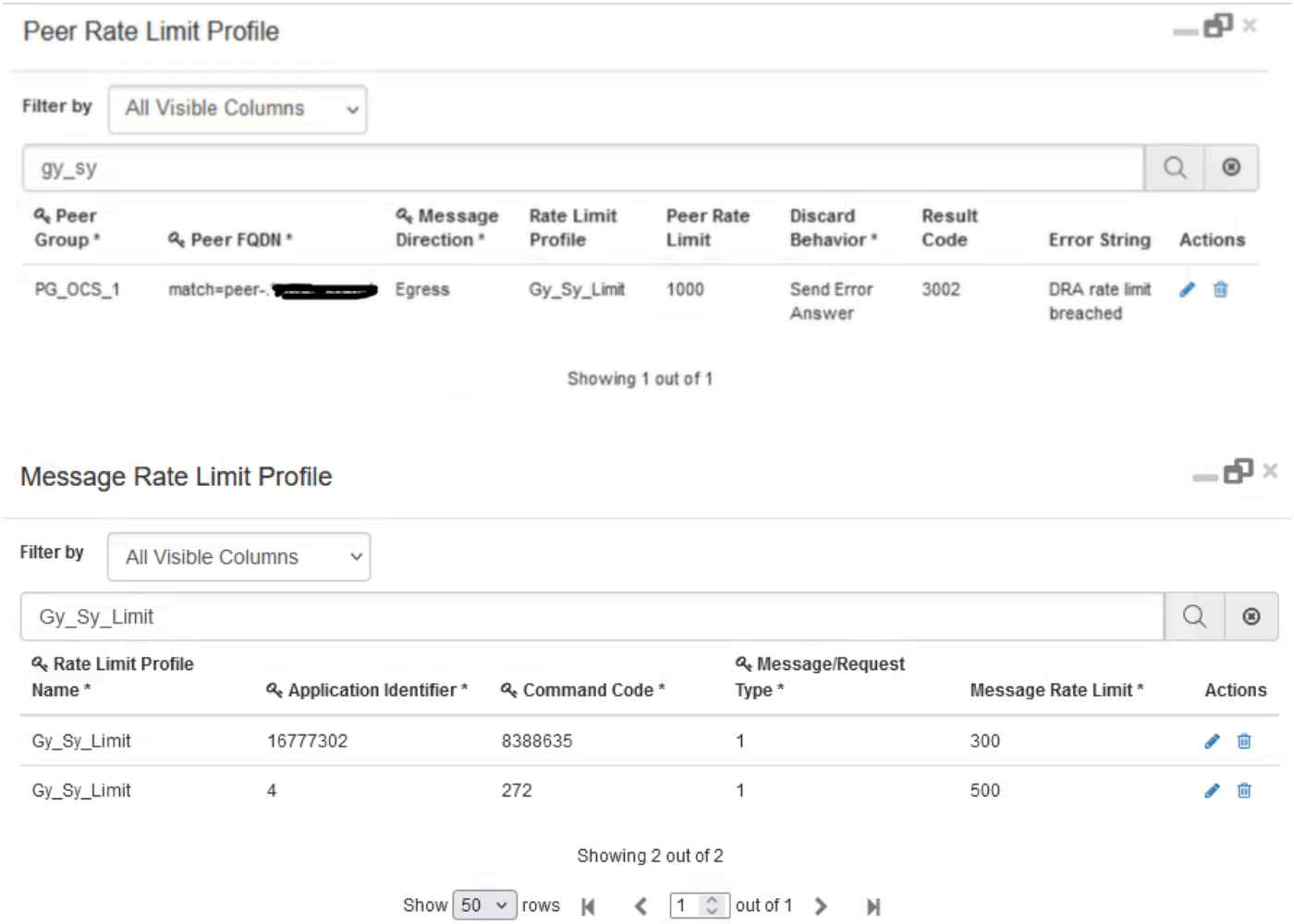
スロットルの動作は、特定のエラーコードとエラーメッセージを含むメッセージを拒否するか、ドロップするかのいずれかです。



CPS vDRAでこの動作を有効にするには、CRDテーブルの「ピアレート制限プロファイル」および「メッセージレート制限プロファイル」を設定します。 これらのCRDテーブルでは、次の値を設定する必要があります。

ピアグループ	ピアグループは、レルムとホストに基づくDiameterノードの論理グループです。調整する必要があるピアグループを

	設定する必要があります。
ピアFQDN	レート制限が必要なピアグループ内のピアのFQDN（完全一致または正規表現一致）。
メッセージの方向	スロットリングの方向：入力または出力。この場合は出力です。
レート制限プロファイル	スロットルが必要なメッセージタイプの定義に使用されるメッセージレート制限プロファイル名。
ピアレート制限	このピアグループに許可されている要求の割合です。これには、そのピアグループからのすべてのメッセージタイプが含まれます。
破棄動作	要求をドロップするか、エラーコードを使用して拒否するかを選択できます。
結果コード	メッセージを拒否する場合の結果コード値。メッセージがドロップされた場合は適用されません。
エラー文字列	拒否された要求の応答メッセージで使用されるエラー文字列。メッセージがドロップされた場合は適用されません。
アプリケーションID	調整するメッセージのアプリケーションID。
コマンドコード	調整するメッセージのコマンドコードです。
メッセージ/要求タイプ	調整する必要がある要求のアプリケーションIDおよび要求タイプ。
メッセージレート制限	DRAによって処理されるそのメッセージタイプの要求のTPS。このTPSを超える要求は抑制されます。この値は、ピアグループ内のピア単位です。

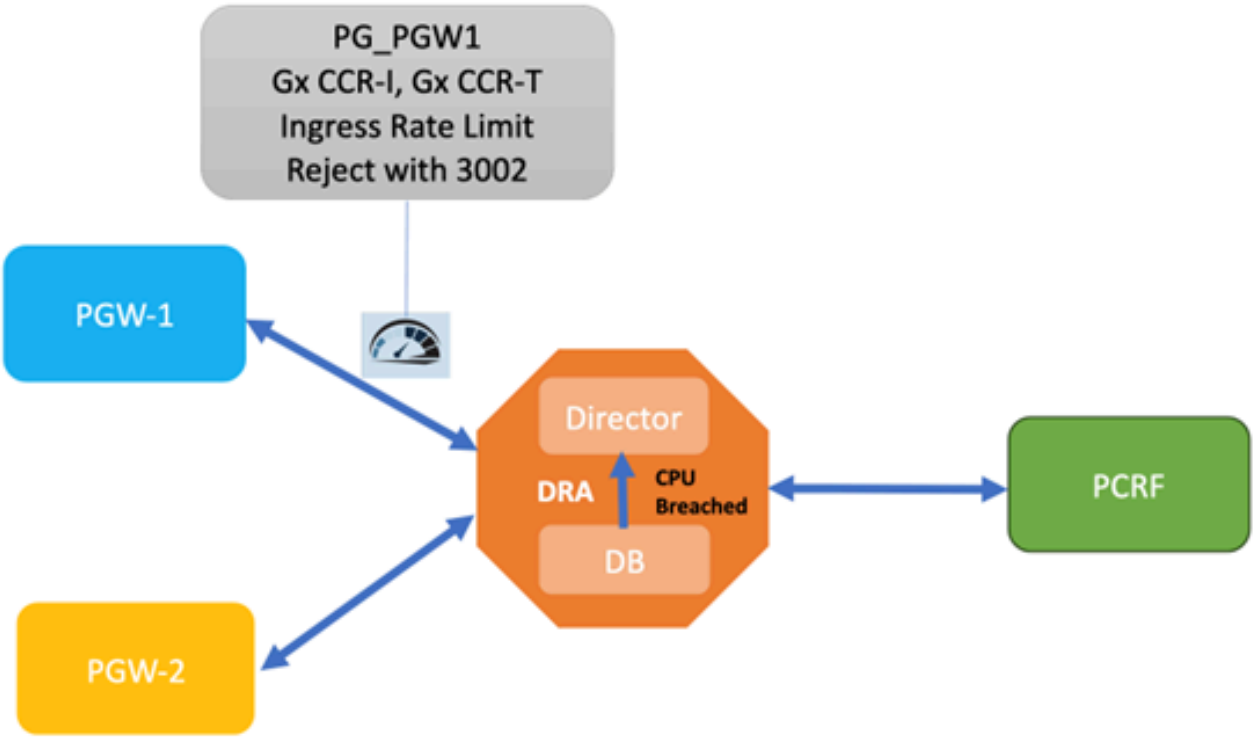


ダイナミックレート制限

CCR-IまたはCCR-Tのバーストが発生すると、データベース(DB)に過負荷が発生し、システムが不安定になる可能性があります。これを解決するために、DRAは使用可能なDB容量に基づくダイナミックレート制限（Gx CCR-IおよびGx CCR-Tのみ）をサポートしています。

DRAはDBのCPU使用率を監視し、しきい値を超えると必ず着信要求を抑制します。スロットリングのCPUしきい値、およびスロットリングされる着信トラフィックは設定可能です。

対応するスロットル率を使用して、異なるCPUしきい値を設定できます。DRAは、現在のDB CPU使用率に基づいてスロットリングレベルを調整します。CPU使用率が安定した場合、スロットリングは徐々に停止します。



この動作は、CPS vDRAでCRDテーブル「ピアレート制限プロファイル」、「メッセージレート制限プロファイル」、「動的ピアレート制限プロファイル」、および「動的調整DB CPUプロファイル」を設定することで有効にできます。 これらのCRDテーブルでは、次の値を設定する必要があります。

ピアグループ	ピアグループは、レルムとホストに基づくDiameterノードの論理グループです。この例では、pGWのピアグループを設定します。
ピアFQDN	レート制限が必要なピアグループ内のピアのFQDN（完全一致または正規表現一致）。
メッセージ/要求タイプ	調整する必要がある要求のアプリケーションIDおよび要求タイプ。この例では、Gx CCR-I、Gx CCR-Tです。
メッセージの方向	スロットリングの方向：入力または出力。この例では、入力です。
レート制限プロファイル	スロットルが必要なメッセージタイプの定義に使用されるメッセージレート制限プロファイル名。

ピアレート制限	このピアグループに許可されている要求の割合です。これには、そのピアグループからのすべてのメッセージタイプが含まれます。
破棄動作	要求をドロップするか、エラーコードを使用して拒否するかを選択できます。
結果コード	メッセージを拒否する場合の結果コード値。メッセージがドロップされた場合は適用されません。
エラー文字列	拒否された要求の応答メッセージで使用されるエラー文字列。メッセージがドロップされた場合は適用されません。
アプリケーションID	調整するメッセージのアプリケーションID。
コマンドコード	調整するメッセージのコマンドコードです。
メッセージレート制限	DRAによって処理されるそのメッセージタイプの要求のTPS。このTPSを超える要求は抑制されます。この値は、ピアグループ内のピア単位です。
動的スロットリングDB CPUプロファイル	これはCPUプロファイル名を参照します。これは、異なるCPU範囲のスロットルパーセンテージを定義するために使用されます。
DB CPU使用率のしきい値	導入環境のトラフィックパターンに従って、違反制限として設定されたCPUレベルの正しい値を選択できます。
スロットル率	対応するCPUレベルを超えたときに適用されるレート制限の%。

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

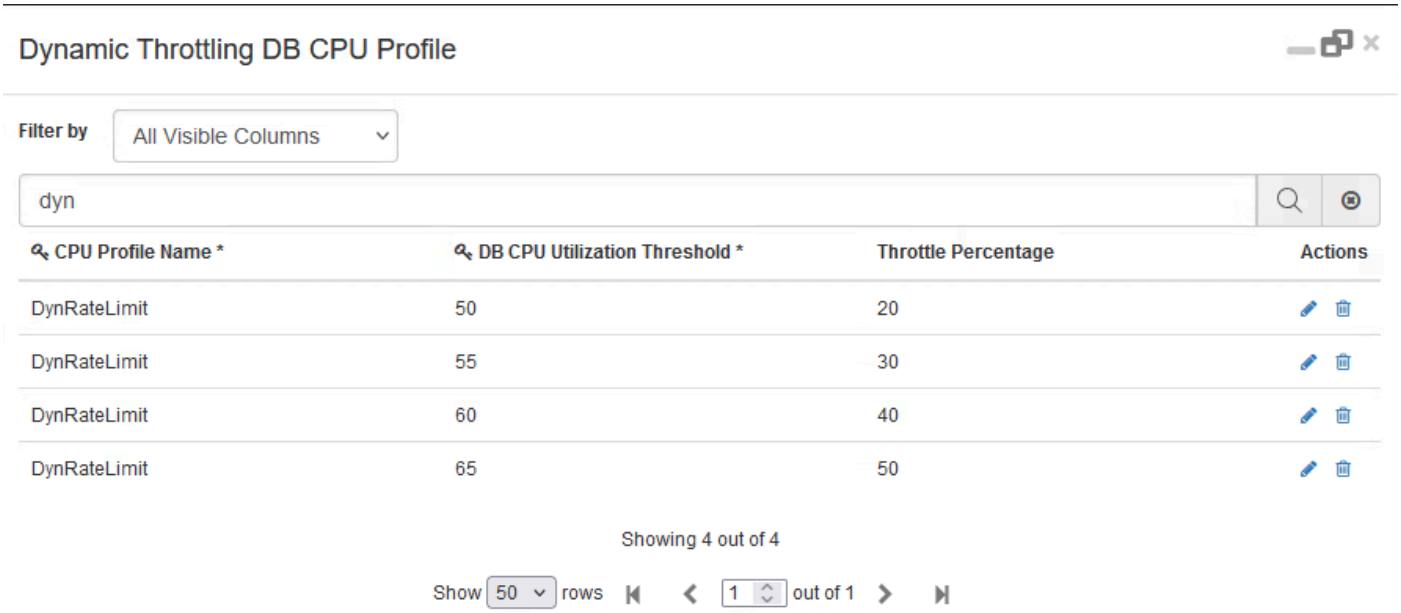
Filter by: All Visible Columns

DynRateLimit

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

Showing 1 out of 1

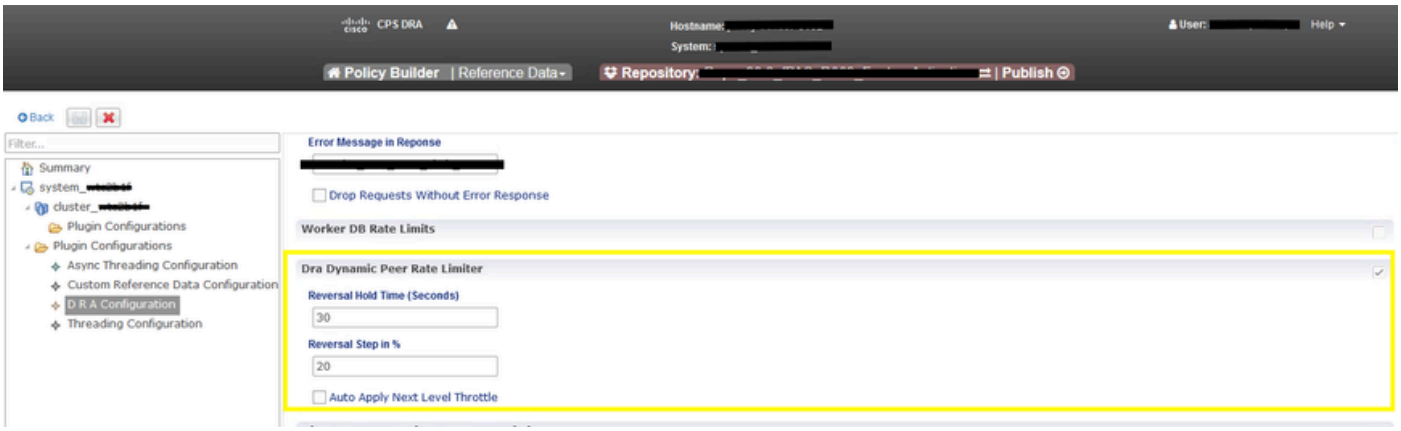
Show 50 rows 1 out of 1



さらに、「DRA Dynamic Peer Rate Limiter」セクションの「DRA Configuration Plugin」の下にあるポリシービルダーのチェックボックスを選択して、この動作を有効にする必要があります。

Reversal Hold Time（リバーサル保留時間）：リバーサルが適用される前にCPU使用率が監視される時間。

Reversal Step in %：取り消される調整の割合。



シナリオ：CPU使用率に基づくダイナミックレート制限

DRAでの次の設定を検討します。

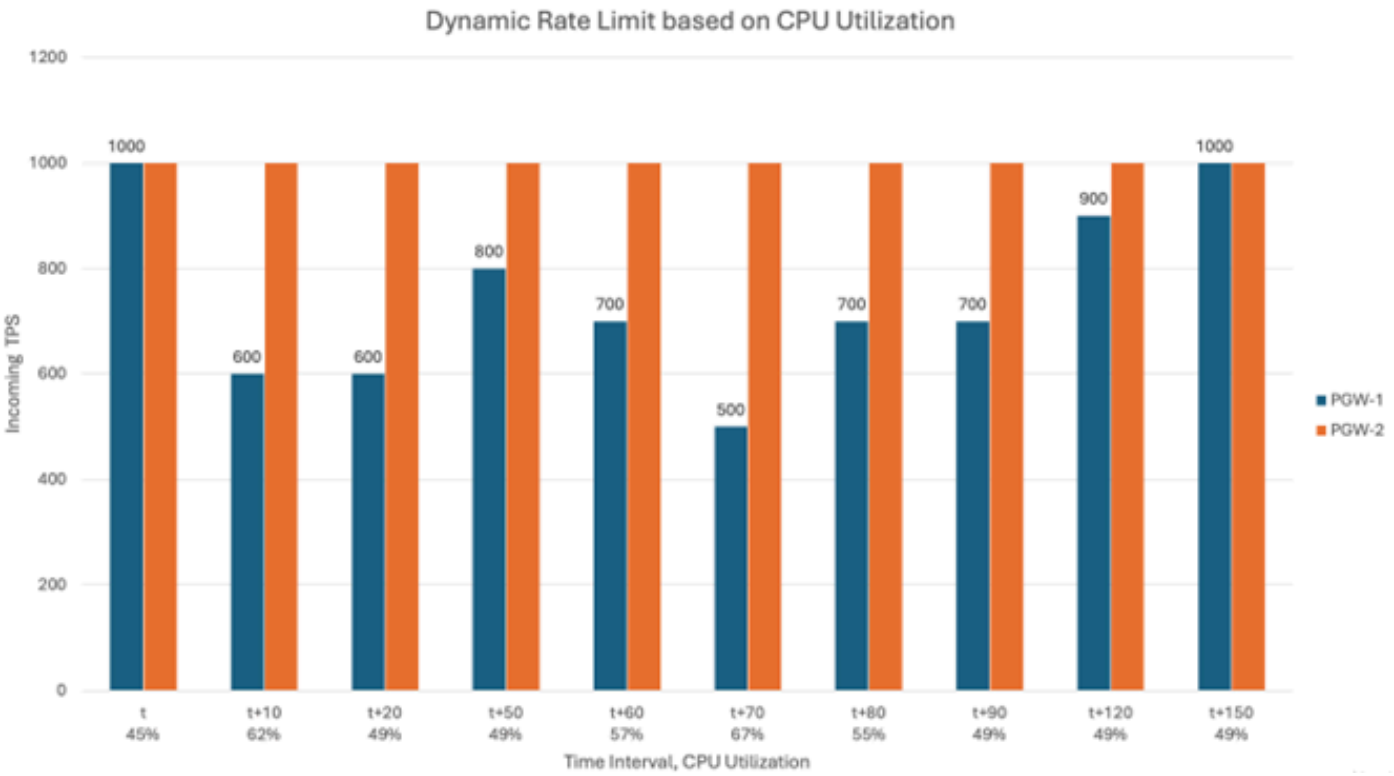
Static Message Rate Limit:1000（これは着信TPSの値です）

リバーサル保留時間：30秒

%の取消ステップ： 20%

DBのCPU使用率がしきい値を超えると、「Dynamic Throttling DB CPU Profile」設定が参照され、ダイレクタに通知することによって、それに応じて着信TPSが調整されます。常に変化するCPU使用率の値に基づいてスロットリングされるため、トラフィックを動的にレート制限してい

ると言えます。



- 最初は、DBのCPU使用率が制限を下回っているため、スロットリングは行われません。また、PGW-2にはDynamic Rate Limiting(DRL)の設定がないため、CPU使用率に関係なくスロットリングは行われません。
- DBのCPU使用率が62 %の場合、トラフィックは40 %抑制され、実効レート制限は600です（着信TPSは1000で、DRAは600のみを許可します）。
- CPU使用率が60 ~ 65 %の間に留まる場合、40 %のスロットリングが、設定されたレート制限1000に引き続き適用され、実効レート制限は600になります（着信TPSが1000の場合、DRAで許可されるのは600だけです）。
- CPU使用率が49 %に低下し、スロットリングの反転がpGW-1から始まります。
- CPU使用率が30秒間49 %以下の状態が続くと、スロットリングは20 ~ 20 %低下します。現在、実効レート制限は800です（着信TPSは1000で、DRAは800のみを許可します）。 リバーサルは、設定に従って、20%のステップで実行されます。
- DBのCPU使用率が57 %に増加すると、トラフィックは30 %抑制され、実効レート制限は700になります（着信TPSが1000の場合、DRAで許可されるのは700だけです）。
- DBのCPU使用率が67 %に増加すると、トラフィックは50 %抑制され、実効レート制限は500になります（着信TPSが1000の場合、DRAで許可されるのは500だけです）。
- DBのCPU使用率が55 %まで低下した場合、トラフィックは30 %抑制され、実効レート制限は700になります（着信TPSが1000の場合、DRAで許可されるのは700だけです）。
- 次の30秒間にCPUが49 %以下に低下した場合、スロットリングは20 %減らして10 %になり、実効レート制限は900になります（着信TPSが1000の場合、DRAでは900しか許可されません）。
- 次の30秒間にCPUがさらに49 %以下にとどまる場合、スロットリングは20 %減少して0になり、反転が完了したときに適用されるレート制限はありません（着信TPSは1000で、DRAは1000を許可します）。

翻訳について
シスコは世界中のユーザにそれぞれの言語でサポート コンテンツを提供するために、機械と人による翻訳を組み合わせて、本ドキュメントを翻訳しています。ただし、最高度の機械翻訳であっても、専門家による翻訳のような正確性は確保されません。シスコは、これら翻訳の正確性について法的責任を負いません。原典である英語版（リンクからアクセス可能）もあわせて参照することを推奨します。