BGPスキャナまたはルータプロセスによるCPU高使用率のトラブルシューティング

内容

概要

このドキュメントでは、BGPスキャナまたはルータを使用した場合のCPU高読み取りの原因をトラブルシューティングする方法について説明します。

前提条件

要件

このドキュメントを読むには、**show process cpu**コマンドの解釈方法に関する知識が必要です。

使用するコンポーネント

このドキュメントの情報は、Cisco IOS® ソフトウェア リリース 12.0 に基づいています。

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、初期（デフォルト）設定の状態から起動しています。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。

表記法

ドキュメント表記の詳細は、『シスコ テクニカル ティップスの表記法』を参照してください。

背景説明

このドキュメントでは、Border Gateway Protocol（BGP；ボーダーゲートウェイプロトコル）ルータプロセスまたはBGPスキャナプロセスが原因で、**show process cpu**コマンドの出力で示されるように、Cisco IOSルータでCPUの高使用率が発生する状況について説明します。CPU 高使用状態の持続期間は、インターネット ルーティング テーブルのサイズ、および特定のルータがルーティング テーブルおよび BGP テーブルに保持するルート数などの条件によって異なります。show process cpuを使用すると、直前の 5 秒間、1 分間、および 5 分間の平均 CPU 使用率が表示されます。CPU 使用率の数値は、負荷と正比例するわけではありません。

主な理由のいくつかを次に示します。

- 現実のネットワークでは、CPU は、ネットワーク管理などのさまざまなシステム メンテナンス機能を処理する必要がある。
- CPU は、定期的なルーティング更新およびイベントにより起動されるルーティング更新を処理する必要がある。
- リソースの可用性に対するポーリングなど、他の内部システムオーバーヘッド操作も存在します。これらはトラフィックの負荷に比例しません。

また、CPU の動作を示す指標を得るために、show processes cpu コマンドも使用できます。

> **注**：showコマンドの詳細については、『Cisco IOS設定の基本コマンドリファレンス』を参照してください

BGPプロセスについて

Cisco IOSプロセスは、一般に、システムメンテナンス、パケットのスイッチング、ルーティングプロトコルの実装などのタスクを実行する個々のスレッドと関連データで構成されます。ルータで実行されている複数の Cisco IOS プロセスによって、BGP が実行されます。show process cpu **| include BGP** コマンドを使用すると、BGPプロセスによるCPU使用率を確認できます。

次の表に、BGPプロセスの機能を示し、各プロセスが異なる時間に実行され、その時間が処理されるタスクによって異なることを示します。BGPスキャナプロセスとBGPルータプロセスは大量の計算を処理するため、これらのプロセスのいずれかが原因でCPUの使用率が高くなることがあります。次のセクションでは、これらのプロセスについて詳しく説明します。

プロセス名	説明	間隔
BGP オープン	BGP ピア確立を実行します。	初期化時に、BGPピアとのTC が確立されたとき。
BGP I/O	キュー内にあり、UPDATESやKEEPALIVESなどの処理されるBGPパケットを処理します。	BGP 制御パケットの受信時。
BGP スキャナ	BGP テーブルをスキャンしてネクストホップの到達可能性を確認します。また、BGPスキャナは条件付きアドバタイズメントをチェックして、BGPが条件プレフィクスをアドバタイズしてルート削減を実行するか	1 分に 1 回。

	どうかを判断します。また、MPLS VPN 環境では、特定の VPN Routing and Forwarding instance (VRF; VPN ルーティング/転送インスタンス) に対してルートをインポートおよびエクスポートします。	
BGP ルータ	最適なBGPパスを計算し、ルートのチャーンを処理します。また、ルートの送受信、ピアの確立、およびルーティング情報ベース(RIB)との対話も行います。	1秒に1回、およびBGPピアが追加、削除、またはソフト設定されるとき。

BGP スキャナによる CPU の高使用

BGPスキャナプロセスによるCPUの高使用は、大規模なインターネットルーティングテーブルを伝送するルータでは短期間で発生する可能性があります。BGP スキャナは、1 分に 1 回 BGP RIB テーブルをスキャンし、重要なメンテナンス タスクを実行します。これらのタスクには、ルータのBGPテーブルで参照されるネクストホップの調査が含まれ、ネクストホップデバイスに到達できることを確認します。そのため、大きなBGPテーブルを取得するには、同等に長い時間がかかります。

BGP スキャナ プロセスは BGP テーブル全体に渡り実行されるため、CPU 使用率が高い状態になる時間の長さは、ネイバーの数やネイバーごとに学習されたルートの数によって変動します。この情報をキャプチャするには、show ip bgp summary コマンドおよび **show ip route summary コマンドを使用します。**BGP スキャナ プロセスでは、データ ストラクチャをアップデートするために BGP テーブルが走査され、経路再配布のためにルーティング テーブルが走査されます(この場合、ルーティングテーブルはRouting Information Base (RIB ; ルーティング情報ベース) とも呼ばれ、;show ip routeコマンドを実行するとルータから出力されます)。両方のテーブルはルータのメモリに別々に保存され、大きくなってCPUサイクルを消費することがあります。

次のdebug ip bgp updatesコマンドの出力例では、BGPスキャナの実行がキャプチャされています。

```
router#
2d17h: BGP: scanning routing tables
2d17h: BGP: 10.0.0.0 computing updates, neighbor version 8,
     table version 9, starting at 0.0.0.0
2d17h: BGP: 10.0.0.0 update run completed, ran for 0ms, neighbor
     version 8, start version 9, throttled to 9, check point net 0.0.0.0
2d17h: BGP: 10.1.0.0 computing updates, neighbor version 8,
     table version 9, starting at 0.0.0.0
2d17h: BGP: 10.1.0.0 update run completed, ran for 4ms, neighbor
     version 8, start version 9, throttled to 9, check point net 0.0.0.0
router#
```

BGP スキャナが実行されると、優先順位の低いプロセスは、CPU にアクセスするまで長時間待機する必要があります。優先順位が低いプロセスの 1 つは、PING などの Internet Control Message Protocol (ICMP; インターネット制御メッセージ プロトコル) パケットを制御します。ICMPプロセスはBGPスキャナの背後で待機する必要があるため、ルータを宛先とする、またはルータから発信されるパケットは、予想される遅延よりも大きくなる可能性があります。サイクルとしては、BGP スキャナが実行され、このスキャナが中断され、その後 ICMP が実行されます。これに対して、ルータを介して送信されるpingは、Cisco Express Forwarding(CEF)を介してス

イッチングされる必要があり、遅延が発生することはありません。遅延が周期的に急増する問題をトラブルシューティングする場合は、ルータを経由して転送されるパケットの転送時間を、ルータのCPUによって直接処理されるパケットと比較します。

> **注**：レコードルートなどのIPオプションを指定するpingコマンドでも、CPUによる直接処理が必要になるため、転送遅延が長くなる可能性があります。

show process | include bgp scannerコマンドを使用して、CPUの優先度を表示します。次の出力例の**Lsi**の値は、低優先度プロセスを示すために**L**を使用しています。

```
6513#show processes | include BGP Scanner
 172 Lsi 407A1BFC       29144     29130    1000 8384/9000  0 BGP Scanner
```

BGP ルータ プロセスによる CPU の高使用

BGP ルータ プロセスは、作業確認のために 1 秒に 1 回程度実行されます。BGP のコンバージェンスでは、最初の BGP ピアが確立された時点から BGP がコンバージされた時点までの時間長が定義されます。コンバージェンスのための時間をできるだけ最短にするために、BGP ルータでは空いているすべての CPU サイクルが使用されます。ただし、BGP ルータは、開始されると CPU を断続的に解放（または中断）します。

コンバージェンス時間は、BGP ルータが CPU を使用した時間を直接計測したものであり、合計時間ではありません。この手順では、BGP コンバージェンス時の高い CPU 使用状態を表示し、2 つの外部 BGP（eBGP）ピアと BGP プレフィックスを交換します。

1. テストを開始する前に、通常のCPU使用率のベースラインを取得します。
   ```
   router#show process cpu
     CPU utilization for five seconds: 0%/0%; one minute: 4%; five minutes: 5%
   ```
2. テストを開始すると、CPU の使用率は 100 % になります。**show process cpuコマ**ンドは、CPU高使用率の状態がBGPルータによって引き起こされ、次の出力で139（BGPルータのCisco IOSプロセスID）と示しています。
   ```
   router#show process cpu
     CPU utilization for five seconds: 100%/0%; one minute: 99%; five minutes: 81%

   !--- Output omitted. 139 6795740 1020252 6660 88.34% 91.63% 74.01% 0 BGP Router
   ```
3. この時点で、**show ip bgp summary**コマンドと**show process cpu**コマンドの複数の出力をモニタしてキャプチャできます。**show ip bgp summary コマンドでは、BGP ネイバーの状態をキャプチャできます。**
   ```
   router#show ip bgp summary
     Neighbor    V    AS  MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down State/PfxRcd
     10.0.0.0    4  64512  309453  157389    19981    0  253 22:06:44 111633
     10.1.0.0  4  65101  188934    1047    40081   41    0 00:07:51 58430
   ```
4. ルータがBGPピアとのプレフィックス交換を完了すると、CPU使用率は通常のレベルに戻ります。計算された1分と5分の平均値も同様に落ち着き、5秒のレートよりも長い期間、通常のレベルよりも高いレベルを示すことができます。
   ```
   router#show process cpu
     CPU utilization for five seconds: 3%/0%; one minute: 82%; five minutes: 91%
   ```
5. 前のshowコマンドのキャプチャ出力を使用して、BGPコンバージェンス時間を計算します。特に、**show ip bgp summary**コマンドの**Up/Down< /strong>列を使用して、CPU高使用状態の開始時刻と終了時刻を比較します。通常、BGPコンバージェンスは、大きなインターネットルーティングテーブルが作成されるまでに数分かかることがあります。スワップされる**

注：デバイス上の CPU は、BGP テーブルの不安定性が原因で使用率が高くなることがあります。これは、ルータがルーティングテーブルのコピーを2つ受信した場合に発生します。1つはISPとのEBGPピアリングから、もう1つはネットワークのIBGPピアリングから受信します。この動作の根本的な原因は、デバイスに搭載されているメモリの量にあります。インターネット ルーティング テーブルのコピー 1 つに対して、最低 1 ギガの RAM を用意することをお勧めします。この不安定性を回避するには、デバイスに装備する RAM の量を増やすか、プレフィックスをフィルタ処理して、使用される BGP テーブルとメモリが少なくなるようにします。

パフォーマンスの向上

インターネットルーティングテーブル内のルートの数が増加すると、BGPのコンバージェンスに要する時間も増加します。一般に、コンバージェンスは、すべてのルートテーブルが一貫性のある状態になるプロセスとして定義されます。BGPは、次の条件に該当する場合に収束したと見なされます。

- すべてのルートが受け入れられている。
- すべてのルートがルーティング テーブルにインストールされている。
- すべてのピアのテーブル バージョンが BGP テーブルのテーブル バージョンと同じである。
- すべてのピアの InQ および OutQ がゼロである。

このセクションでは、BGPコンバージェンス時間を短縮し、BGPプロセスによる高いCPU状態を減らすための、Cisco IOSのパフォーマンス向上について説明します。

TCPピア接続へのキュー

BGPは、OutQが完全に枯渇するまで、BGP OutQから各ピアのTCPソケットにデータを積極的にキューイングします。BGP の送信レートが高速になったため、BGP がコンバージする時間が短縮されました。

BGP ピア グループ

BGP ピア グループにより、BGP の設定が簡略化されるだけでなく、スケーラビリティも向上します。すべてのピア グループ メンバは、共通の発信ポリシーを共有する必要があります。したがって、同じアップデートパケットを各グループメンバーに送信できるため、BGPがルートをピアにアドバタイズするために必要なCPUサイクルの数が減少します。つまり、ピア グループにより、BGP はピア グループ リーダーの BGP テーブルだけをスキャンし、発信ポリシーを使用してプレフィクスをフィルタし、アップデートを生成してピア グループ リーダーに送信します。次に、リーダーはアップデートを複製し、同期対象としてこのアップデートをグループ メンバに配布します。ピア グループがないと、BGP は各ピアについてテーブルをスキャンして、発信ポリシーを使用してプレフィクスをフィルタし、1 つのピアだけに送信するアップデートを作成する必要があります。

パス MTU と ip tcp path-mtu-discovery コマンド

すべての TCP セッションは、1 つのパケットで転送できるバイト数の上限によって制限を受けます。この上限は Maximum Segment Size（MSS; 最大セグメント サイズ）と呼ばれており、デフォルトでは 536 バイトです。つまり、TCPは送信キュー内のパケットを536バイトのチャンクに分割してからIPレイヤに渡します。show ip bgp neighbors **| include max data**コマンドを使用して、BGPピアのMSSを表示します。

```
Router#show ip bgp neighbors | include max data
Datagrams (max data segment is 536 bytes):
Datagrams (max data segment is 536 bytes):
Datagrams (max data segment is 536 bytes):
Datagrams (max data segment is 536 bytes):
```

ほとんどのリンクでは少なくとも 1500 バイトの MTU が使用されているため、MSS が 536 バイトの場合、パケットは宛先へのパスの途中にある IP デバイスで断片化されにくくなります。ただし、パケット サイズが小さくなると、転送オーバーヘッドに使用される帯域幅が増加することになります。BGP はすべてのピアに対して TCP 接続を構築するため、536 バイトの MSS は BGP コンバージェンス時間に影響を与えます。

解決策は、ip tcp path-mtu-discoveryコマンドでパスMTU(PMTU)機能を有効にすることです。この機能を使用すると、MSS値の大きさを動的に決定でき、その間はフラグメント化が必要なパケットを作成しません。PMTU を使用すると、TCP は TCP セッションにあるすべてのリンクから最小の MTU を決定できます。次に、TCP はこの MTU 値から IP ヘッダーおよび TCP ヘッダー用の領域を差し引き、TCP セッションの MSS とします。TCPセッションがイーサネットセグメントのみを通過する場合、MSSは1460バイトです。Packet over SONET(POS)セグメントのみを通過する場合、MSSは4430バイトです。MSS が 536 バイトから 1460 または 4430 バイトに増加すると、TCP/IP オーバーヘッドが低減され、BGP のコンバージェンスが高速化されます。

PMTUを有効にした後、再度show ip bgp neighbors | include max dataコマンドを使用して、ピアごとのMSS値を表示します。

```
Router#show ip bgp neighbors | include max data
Datagrams (max data segment is 1460 bytes):
Datagrams (max data segment is 1460 bytes):
Datagrams (max data segment is 1460 bytes):
Datagrams (max data segment is 1460 bytes):
```

インターフェイス入力キューの拡大

BGPが数千のルートを多数のピアにアドバタイズする場合、TCPは短時間で数千のパケットを送信する必要があります。BGPピアはこれらのパケットを受信し、アドバタイズするBGPスピーカにTCP確認応答を送信します。これにより、BGPスピーカは短時間で大量のTCP ACKを受信します。ACK がルータ プロセッサの能力よりも高いレートで到達すると、パケットは着信インターフェイス キューに戻ります。デフォルトでは、ルータ インターフェイスで使用される入力キューサイズは 75 パケットです。また、BGP UPDATES などの特別な制御パケットは、Selective Packet Discard（SPD; 選択パケット廃棄）とともに特別なキューを使用します。 この特別なキューでは、100 パケット保持されます。BGPコンバージェンスが発生すると、TCP ACKによって175箇所の入力バッファリングがすぐにいっぱいになり、到着した新しいパケットを廃棄する必要があります。15以上のBGPピアを持ち、完全なインターネットルーティングテーブルを交換しているルータでは、1分あたりインターフェイスあたり10,000を超えるドロップが発生する可能性があります。再起動後 15 分経過したルータの出力例を次に示します。

```
Router#show interface pos 8/0 | include input queue
Output queue 0/40, 0 drops; input queue 0/75, 278637 drops
Router#
```

インターフェイスの入力キューの深さを(hold-queue inコマンドを使用して)増やすと、廃棄されるTCP ACKの数が減り、BGPがコンバージするために実行する必要がある作業量が減ります。通常、入力キューの廃棄が原因で発生する問題は、値に 1000 を指定すると解消されます。

注：入力キューの増加によって遅延が生じる可能性があるため、この方法は慎重に使用して

ください。

Cisco IOSのその他の改善点

Cisco IOSには、アップデートのパッキングと複製を改善するために、BGPピアグループコードに対するいくつかの最適化が含まれています。これらの改善点を確認する前に、更新プログラムのパッキングとレプリケーションの詳細を確認してください。

BGP のアップデートは、アトリビュートの組み合せ（MED = 50 と LOCAL_PREF = 120 など）、およびこのアトリビュートの組み合せを共有する Network Layer Reachability Information（NLRI; ネットワーク レイヤ到着可能性情報）プレフィクスのリストで構成されています。BGP が 1 つのアップデート内でリストする NLRI プレフィクスが多いほど、オーバーヘッド（IP、TCP、および BGP の各ヘッダー）が低下するため、BGP のコンバージェンスが高速になります。 **アップデートパッキング**とは、NLRIをBGPアップデートにパッキングすることです。たとえば、BGP テーブルにおいて固有なアトリビュートの組み合せが 15,000 含まれているルートが 100,000 保持されており、NLRI が 100 % の効率でパックされている場合、BGP は 15,000 のアップデートを送信するだけで済みます。

> **注：**パッキング効率 0 % とは、この環境で BGP が 100,000 のアップデートを送信する必要があることを意味します。

BGP アップデートの効率を表示するには、show ip bgp peer-group コマンドを使用します。

ピアグループメンバーが同期している場合、BGPルータはピアグループリーダー用にフォーマットされた更新メッセージを取得し、それをメンバー用に複製します。アップデートをピア グループ メンバ用に複製することは、アップデートを再フォーマットするよりも効率的です。たとえば、ピア グループに 20 のメンバが存在し、すべてのメンバが 100 BGP メッセージを受信する必要があると想定します。100 % の複製とは、1 つの BGP ルータがピア グループ リーダー用に 100 のメッセージをフォーマットし、これらのメッセージを他の 19 のピア グループ メンバに複製することです。複製の改善点を確認するには、**show ip bgp peer-group コマンドで表示される複製されたメッセージの数とフォーマットされたメッセージの数を比較します。**これにより、コンバージェンス時間が大幅に改善され、BGP で数多くのピアがサポートされるようになります。

たとえば、アップデートのパッキングとアップデートの複製の効率を確認するには、**show ip bgp peer-group**コマンドを使用します。次の出力は、6つのピアグループ(内部BGP(iBGP)ピア)、最初の5つのピアグループ（eBGPピア）それぞれに20のピア、および6番目のピアグループ(内部BGP(iBGP)ピア)に100のピアを持つコンバージェンステストからのものです。 また、使用されたBGP テーブルには、アトリビュートの組み合せが 36,250 あります。

show ip bgp peer-group | include replicatedコマンドをCisco IOS 12.0(18)Sが稼働するルータで実行すると、次の情報が表示されます。

```
Update messages formatted 836500, replicated 1668500
Update messages formatted 1050000, replicated 1455000
Update messages formatted 660500, replicated 1844500
Update messages formatted 656000, replicated 1849000
Update messages formatted 501250, replicated 2003750

!-- The first five lines are for eBGP peer groups. Update messages formatted 2476715, replicated
12114785 !-- The last line is for an iBGP peer group.
```

各ピア グループについて複製率を計算するには、次のように複製されたアップデートの数を、フ

ォーマットされたアップデートの数で割ります。

1668500/836500 = 1.99 1455000/1050000 = 1.38 1844500/660500 = 2.79 1849000/656000 = 2.81 2003750/501250 = 3.99 12114785/2476715 = 4.89

- BGPが完全に複製された場合、ピアグループには20のピアが存在するため、eBGPピアグループの複製率はそれぞれ19になります。アップデートはピアグループリーダー用にフォーマットされ、他の19のピアに複製されます。これにより、最適な複製レートは19になります。ピア数が100であるため、iBGPピアグループの理想的な複製レートは99になります。
- BGPがアップデートを完全にパックした場合、フォーマットされたアップデートは36,250だけになります。BGPテーブル内の属性の組み合わせの数がピアグループごとに36,250のアップデートを生成するだけで済みます。iBGP ピア グループだけで約 2,500,000 のアップデートがフォーマットされますが、eBGP ピア グループでは、平均 500,000 ~ 1,000,000 のアップデートが生成されます。

Cisco IOS 12.0(19)Sが稼働するルータでは、**show ip bgp peer-group | include replicated**commandは、次の情報を提供します。

```
Update messages formatted 36250, replicated 688750
Update messages formatted 36250, replicated 688750
Update messages formatted 36250, replicated 688750
Update messages formatted 36250, replicated 688750
Update messages formatted 36250, replicated 688750
Update messages formatted 36250, replicated 3588750
```

> **注**：アップデート パッキングは最適です。正確に36,250のアップデートがピアグループごとにフォーマットされます。688750/36250 = 19 688750/36250 = 19 688750/36250 = 19 688750/36250 = 19 688750/36250 = 19 3588750/36250 = 99

> **注**：アップデートの複製も完全です。

トラブルシューティング手順

BGP スキャナまたは BGP ルータが原因で CPU の使用率が高くなる問題をトラブルシューティングするには、次の手順を実行します。

- BGP トポロジに関する情報を収集します。BGPピアの数と、各ピアによってアドバタイズされるルートの数を決定します。CPU の高使用状態の持続時間は、ご使用の環境に対して妥当ですか。
- CPU の高使用が発生する時期を特定します。この時期は、通常スケジュールされている BGP テーブルのスキャニングと一致しますか。
- CPU の高使用は、インターフェイスのフラップの後に発生しましたか。ダンプニングが有効な場合は、コマンド**show ip bgp dampening flap-statistics**コマンドを使用できます。
- ルータを経由する PING を実行し、ルータから PING を実行します。ICMP エコーは、優先順位が低いプロセスとして処理されます。詳細については、『pingおよびtracerouteコマンドについて』を参照してください。通常の転送が影響を受けていないかどうかを確認します。
- 着信インターフェイスと発信インターフェイスでファーストスイッチングやCEFが有効になっているかどうかを確認する際には、パケットが高速転送パスを通過できることを確認する必要があります。インターフェイスに関する **ip route-cache cef、またはグローバル コンフィギュレーションに関する no ip cefが表示されないことを確認します。**グローバル コンフィギ

ュレーション モードで CEF を有効にするには、ip cef コマンドを使用します。
- ほとんどの場合、このような状況を引き起こすデバイスの過負荷が原因であるため、プラットフォームの規模を確認します。また、ルータ上に適切なTernary Content Addressable Memory(TCAM)スペースが使用可能であることを確認します。

- ルータに十分なメモリがあることを確認します。推奨事項に従って、完全なインターネットルーティングテーブルを送信するBGPピアごとに、Cisco IOSスペースに最低1 GBのDRAMを割り当てます。ここに示した DRAM 空間は、BGP に必要なメモリだけを表してます。ルータで実行されるその他の機能には、追加のスペースが必要な場合があります。

関連情報

- IP ルーティングに関するサポート ページ
- テクニカルサポート - Cisco Systems