排除由BGP扫描程序或路由器进程导致的高CPU故障

目录

简介

本文档介绍当使用BGP扫描程序或路由器时如何排除CPU读数高的原因。

先决条件

要求

本文档需要了解如何解释**show process cpu**命令。

使用的组件

本文档中的信息基于 Cisco IOS® 软件版本 12.0。

本文档中的信息都是基于特定实验室环境中的设备编写的。本文档中使用的所有设备最初均采用原始（默认）配置。如果您的网络处于活动状态，请确保您了解所有命令的潜在影响。

规则

有关文档规则的详细信息，请参阅 Cisco 技术提示规则。

背景信息

本文描述了Cisco IOS路由器由于边界网关协议(BGP)路由器进程或BGP扫描程序进程而可能遇到CPU使用率较高的情况，如**show process cpu命令的输**出所示。CPU 使用率过高这一情况的持续时间取决于多个条件，尤其是 Internet 路由表的大小和特定路由器在其路由和 BGP 表中保留的路由数。show process cpu 命令可显示过去五秒钟、一分钟和五分钟的 CPU 平均使用率。CPU 使用率数值显示出使用率与流入负载并不具有真实的线性关系。

以下是一些主要原因：

- 在实际的全球网络中，CPU 必须处理网络管理等多种系统维护功能。
- CPU 必须处理定期的和事件触发的路由更新。
- 还有其他内部系统开销操作，例如轮询资源可用性，这些操作与流量负载不成比例。

您还可以使用**show processes cpu命**令来获取CPU活动的一些指示。

> **注意**：有关show命令的详细信息，请参阅《Cisco IOS配置基础命令参考》

了解BGP进程

通常，Cisco IOS进程由执行任务（如系统维护、交换数据包和路由协议实施）的各个线程和关联数据组成。路由器上执行的多个Cisco IOS进程使BGP能够运行。请使用 show process cpu **| include BGP**命令，查看由BGP进程导致的CPU使用率。

此表列出了BGP进程的功能，并显示每个进程在不同的时间运行，这些时间取决于所处理的任务。由于BGP扫描程序和BGP路由器进程负责大量计算，因此，由于这些进程中的任一进程，您都会看到CPU使用率较高。接下来的部分将更详细地讨论这些流程。

进程名	描述	间隔
BGP Open	执行 BGP 对等体的建立。	初始化时，与BGP对等体建立T接时。
BGP I/O	处理队列中并被处理的BGP数据包，例如UPDATES和KEEPALIVES。	收到 BGP 控制数据包时。
BGP Scanner	扫描 BGP 表，并确认下一跳的可达性。BGP扫描程序还检查条件通告以确定BGP是否通告条件前缀并执行路由缩减。在 MPLS VPN 环境中，BGP 扫描程序将路由导入和导出到特定的 VPN 路由和转发实例 (VRF) 中。	每分钟执行一次。
BGP Router	计算最佳BGP路径并处理任何路**由流**动。它还发送和接收路由、建立对等体以及与路由信息库(RIB)交互。	每秒一次，当添加、删除或软配BGP对等体时。

BGP 扫描程序引起的高 CPU 使用率

由于BGP扫描程序进程导致的CPU使用率较高，在承载大型Internet路由表的路由器上可能会持续较短的时间。BGP 扫描程序每分钟扫描一次 BGP RIB 表并执行重要的维护任务。这些任务包括检查路由器BGP表中引用的下一跳，并验证是否可以到达下一跳设备。因此，大型BGP表需要相当长的时间才能进行浏览和验证。

由于 BGP 扫描程序进程将扫描整个 BGP 表，因此，CPU 使用率过高这一情况的持续时间根据邻居数量和每个邻居获知的路由数而有所不同。请使用 **show ip bgp summary 和 show ip route summary 命令获取此信息。**BGP 扫描程序进程扫描 BGP 表以更新所有数据结构，并扫描路由表以进行路由重分配。(在此情景中，路由表也称为路由信息库(RIB)，在执行；show ip route命令时，路由器会输出该路由信息库。) 两个表单独存储在路由器的内存中，可能较大并占用CPU周期。

debug ip bgp updates命令的下**一个示例输**出捕获BGP扫描程序的执行：

```
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#
```

当 BGP 扫描程序运行时，优先级低的进程需要等待更长的时间才可访问 CPU。一个低优先级进程控制Internet控制消息协议(ICMP)数据包，例如ping。发往路由器或从路由器发起的数据包可能比预期延迟更高，因为ICMP进程必须在BGP扫描程序后等待。周期是 BGP 扫描程序运行一段时间并自行暂停，然后 ICMP 运行。相反，通过路由器发送的ping必须通过思科快速转发(CEF)交换，并且不能遇到任何额外延迟。排除延迟的周期性峰值故障时，将通过路由器转发的数据包的转发时间与路由器上CPU直接处理的数据包进行比较。

> **注意：**指定IP选项（如记录路由）的ping命令也要求CPU直接处理它们，这可能导致更长的转发延迟。

请使用 **show process | include bgp scanner命**令查看CPU优先级。下一个示例**输出**中的Lsi值使用**L**来指代低优先级进程。

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

BGP 路由器进程引起的高 CPU 使用率

BGP 路由器进程约每秒钟运行一次，以检查工作。BGP 收敛定义了首个 BGP 对等体的建立时间和 BGP 收敛时间之间的持续时间。为确保收敛时间尽可能最短，BGP 路由器将会占用所有空闲的 CPU 周期。但是在开始之后，它将会间歇性地释放（或暂停）CPU。

收敛时间是对 BGP 路由器在 CPU 上所花费时间（而非总时间）的直接测量。此过程显示BGP收敛期间的高CPU条件，并与两个外部BGP(eBGP)对等体交换BGP前缀。

1. 在开始测试之前，捕获正常CPU使用率的基线。
   ```
   router#show process cpu
     CPU utilization for five seconds: 0%/0%; one minute: 4%; five minutes: 5%
   ```
2. 测试开始之后，CPU 使用率达到 100%。**show process** cpu命令显示高CPU条件是由BGP路由器引起的，在下一输出中以139（BGP路由器的Cisco IOS进程ID）表示。
   ```
   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. 此时，您可以监控并捕获show ip bgp summary和show process cpu命令的多个输出结果。show ip bgp summary 命令可捕获 BGP 邻居的状态。

```
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. 当路由器完成与其BGP对等体的前缀交换时，CPU使用率将返回正常级别。计算出的一分钟和五分钟平均值也能稳定下来，并且在比五秒速率更长的时间内显示高于正常水平。

```
router#show process cpu
  CPU utilization for five seconds: 3%/0%; one minute: 82%; five minutes: 91%
```

5. 使用前面show命令捕获的输出计算BGP收敛时间。特别是，使用show ip bgp summary命令的Up/Down< /strong>列，并比较高CPU条件的启动和停止时间。通常，当大型Internet路由表时，BGP融合可能需要几分钟。交换

注意：设备上的CPU使用率较高也可能是由于BGP表不稳定。如果路由器收到两个路由表副本，一个来自与ISP的EBGP对等，另一个来自网络中的IBGP对等，则会发生这种情况。造成这种情况的根本原因是设备上的内存量。思科建议对互联网路由表的单个副本至少使用1 Gig的RAM。要避免这种不稳定，请增加设备上的RAM或过滤前缀，以便释放BGP表和它占用的内存。**性能改进**随着Internet路由表中路由数的增加，BGP收敛所花的时间也在增加。一般来说，收敛定义为所有路由表进入一致状态的过程。当以下条件为真时，BGP被视为已收敛：

- 已接受所有路由。
- 所有路由都已安装在路由表中。
- 所有对等体的表版本与 BGP 表的表版本相同。
- 所有对等体的 InQ 和 OutQ 都为零。

本节介绍一些Cisco IOS性能改进，以缩短BGP收敛时间，这会减少BGP进程导致的高CPU条件。**TCP对等连接队列**BGP现在将数据从BGP OutQ主动排队到每个对等体的TCP套接字，直到OutQ完全耗尽。由于 BGP 现在的发送速率更快，因此 BGP 的收敛速度也更快。**BGP 对等体组**BGP 对等体组不但有助于简化 BGP 配置，同时还可以提高可扩展性。所有对等体组成员必须共享一个公用出站策略。因此，可以向每个组成员发送相同的更新数据包，这减少了BGP向对等体通告路由所需的CPU周期数。换言之，使用对等体组时，BGP 仅扫描对等体组引导路由器上的 BGP 表，通过出站策略过滤前缀，并生成更新，然后向对等体组引导路由器发送更新。接着，引导路由器将更新复制到与其同步的组成员中。不使用对等体组时，BGP 必须扫描每个对等体的表，通过出站策略过滤前缀，并生成更新，然后仅向一个对等体发送更新。**路径 MTU 和 ip tcp path-mtu-discovery 命令**单个数据包中可传输的字节数存在一个限值，所有 TCP 会话都受此限值的限定。默认情况下，此限制称为最大分段大小(MSS)为536字节。换句话说，TCP将传输队列中的数据包分成536字节块，然后再将数据包向下传送到IP层。请使用 show ip bgp neighbors | include max data命令以显示BGP对等体的MSS:

```
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):
```

536 个字节的 MSS 所具有的优点是：由于大多数链路使用的 MTU 至少为 1500 个字节，因此，在通往目的地的路径上 IP 设备不可能将数据包分段。缺点是较小的数据包将会增加用于传输的带宽开销。由于 BGP 建立了到所有对等体的 TCP 连接，因此，536 个字节的 MSS 将会影响 BGP 收敛时间。解决方案是使用ip tcp path-mtu-discovery命令启用路径MTU(PMTU)功能。您可以使用此功能动态确定MSS值的大小，同时，不创建需要分段的数据包。通过 PMTU，TCP 可以确定 TCP 会话所有链路中的最小 MTU 大小。然后，TCP 将使用此 MTU 值（减去 IP 报头和 TCP 报头占用的空间）作为会话的 MSS。如果TCP会话仅通过以太网段，则MSS为1460字节。如果它仅通过SONET上的数据包(POS)段，则MSS为4430字节。MSS 从 536 个字节增加到 1460 或 4430 个字节可减少 TCP/IP 开销，这有助于 BGP 更快收敛。启用PMTU后，再次使用show ip bgp neighbors | include max data命令，查看每个对等体的MSS值：

```
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):
```

增加接口输入队列如果BGP向许多对等体通告数千条路由，则TCP必须在短时间内传输数千个数据包。BGP对等体接收这些数据包并向通告的BGP扬声器发送TCP确认消息，这会导致BGP扬声器在短时间内接收大量TCP ACK。如果 ACK 到达路由处理器的速率过高，则数据包会在入站接口队列中备份。默认情况下，路由器接口使用的输入队列大小为 75 个数据包。此外，特殊控制数据包（如BGP更新）使用具有选择性数据包丢弃(SPD)的特殊队列。 此特殊队列可存放 100 个数据包。当BGP收敛时，TCP ACK可以快速填充175个输入缓冲点，并且必须丢弃到达的新数据包。在具有15个或更多BGP对等体和交换完整Internet路由表的路由器上，每分钟每个接口可看到超过10,000个丢包。以下为重新启动 15 分钟之后一个路由器的示例输出：

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

如果增加接口输入队列深度(使用hold-queue in命令)，它有助于减少丢弃的TCP ACK数量，从而减少BGP必须完成的收敛工作量。通常情况下，如果值为 1000，则可以解决由输入队列丢包导致的问题。注意：有意识地使用此功能，因为输入队列增量会增加一些延迟。**Cisco IOS的其他改进**Cisco IOS对BGP对等体组代码进行了多次优化，以改进更新打包和复制。在查看这些改进之前，请更详细地查看更新打包和复制。BGP更新由属性（如MED = 50和LOCAL_PREF = 120）和共享该属性组合的网络层可达性信息(NLRI)前缀列表组成。BGP 可在单个更新中列出的 NLRI 前缀越多，由于减少了开销（如 IP、TCP 和 BGP 报头），BGP 收敛的速度就越快。 更新包装是指将NLRI打包到BGP更新中。例如，一个 BGP 表存放了包含 15,000 个唯一属性组合的 100,000 个路由，则如果以 100% 的效率打包 NLRI，BGP 将仅需要发送 15,000 个更新。注意：打包效率为 0% 意味着 BGP 需要在此环境中发送 100,000 个更新。请使用 show ip bgp peer-group 命令查看 BGP 更新的效率。如果对等组成员处于同步状态，BGP路由器会获取为对等组领导者格式化的更新消息，并为该成员复制该消息。复制对等体组成员的更新比重新格式化更新的效率更高。例如，假设对等体组包含 20 个成员，并且所有成员都需要接收 100 个 BGP 消息。百分之百复制意味着 BGP 路由器将为对等体组引导路由器格式化 100 个消息，然后将这些消息复制到其他 19 个对等体组成员中。要确认复制改进，请将复制的消息数与格式化的消息数（如 show ip bgp peer-group 命令所示）进行比较。改进可使收敛时间产生显著变化，并允许 BGP 支持更多的对等体。例如，使用show ip bgp peer-group命令检查更新打包和更新复制的效率。下一个输出来自一个融合测试，该测试包含6个对等组、前5个对等组（eBGP对等体）中每个组有20个对等体、第6个对等组(内部BGP(iBGP)对等体)中100个对等体。 此外，所使用的 BGP 表包含 36,250 个属性组合。show ip bgp peer-group的下一个示例输出 |在运行Cisco IOS 12.0(18)S的路由器上包含复制命令显示以下信息：

```
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.
```

要计算每个对等体组的复制率，请将复制的更新数除以格式化的更新数：1668500/836500 = 1.99 1455000/1050000 = 1.38 1844500/660500 = 2.79 1849000/656000 = 2.81 2003750/501250 = 3.99 12114785/2476715 = 4.89

- 如果BGP复制完美，则eBGP对等组的复制速率均为19，因为对等组中有20个对等体。更新为对等组领导格式，然后复制到其他19个对等体。这提供了19的最佳复制速率。iBGP对等体组的理想复制速率为99，因为有100个对等体。
- 如果BGP封装更新完美，则只有36,250个格式化更新。您只需为每个对等组生成36,250个更新，因为这是BGP表中的属性组合数。单个 iBGP 对等体组格式化约 2,500,000 个更新，而每个 eBGP 对等体组生成的更新数则在 500,000 至 1,000,000 范围内不等。

在运行Cisco IOS 12.0(19)S的路由器上，show ip bgp peer-group | include replicated命令提供以下信息：

```
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
```

注意：更新打包最佳。每个对等组的更新格式为36,250次。 688750/36250 = 19 688750/36250 = 19 688750/36250 = 19 688750/36250 = 19 688750/36250 = 19 36250/99注意：更新复制同样具有很好的效果。**故障排除步骤**要排除由于BGP扫描程序或BGP路由器而导致的高CPU故障，请使用以下步骤：

- 收集 BGP 拓扑相关信息。确定BGP对等体的数量和每个对等体通告的路由数量。根据您的环境，CPU 使用率过高这一情况的持续时间是否合理？
- 确定 CPU 使用率过高所发生的时间。它是否与 BGP 表的定期扫描时间一致？
- CPU 使用率过高是否发生在接口抖动之后？如果启用了阻尼，可以使用命令show ip bgp dampening flap-statistics命令。
- 通过路由器执行 Ping 操作，然后从路由器发出 ping。将 ICMP Echo 作为低优先级进程进行处理。本文档“了解Ping和Traceroute命令”详细解释了这一点。确保常规转发不受影响。
- 在检查入站和出站接口上是否启用了快速交换和/或CEF时，需要确保数据包可以遵循快速转发路径。确保在接口上未看到no ip route-cache cef命令，或在全局配置上未看到no ip cef命令。要在全局配置模式下启用CEF，请使用ip cef命令。
- 检查平台扩展，因为在大多数情况下，这是由于设备过载导致出现这种情况。此外，确保路由器上有适当的三态内容可寻址存储器(TCAM)空间。
- 检验路由器上的内存是否足够。根据建议，每个发送完整Internet路由表的BGP对等体至少为Cisco IOS空间分配1 GB DRAM。此处提及的DRAM空间仅是BGP所需的内存。路由器上运行的其它功能可能需要额外的空间。

相关信息

- IP 路由 支持页
- 技术支持 - Cisco Systems