BGP 的 GR 机制

RFC 4724

陶志豪

zhihao.tao@outlook.com

https://github.com/netwiki/share-doc

Network Working Group
Request for Comments: 4724
Category: Standards Track

S. Sangli
E. Chen
Cisco Systems
R. Fernando
J. Scudder
Y. Rekhter
Juniper Networks
January 2007

Thank Zhihao Tao for your hard work in Translation. The translator spent countless nights and weekends, using his hard work to make it convenient for everyone.
If you have any questions, please send a email to zhihao.tao@outlook.com

BGP 的 Graceful Restart 机制

备忘录状态

本文档为互联网社区规定的互联网标准化协议，并请讨论和建议来改进。

请参考当前版本的“互联网官方协议标准”（STD 1）和该协议的状态。此备忘录的传播不受限制。

版权声明

概述

本文档描述了一种 BGP 机制，其有助于最小化 BGP 重启导致对路由的负面影响。End-of-RIB 标记被详细描述并且可以用来传送路由收敛信息。一种新的 BGP 能力，称为“GR 能力”定义为允许 BGP speaker 表达其具有 BGP 重启期间保持转发状态的能力。最后，概述了 TCP 会话终止/重新建立期间暂时保留路由信息的处理。

本文档中描述的机制适用于所有路由器，即包括在 BGP 重启期间能够保持转发状态和那些没有（尽管后者只需要实现此文件描述的机制的一个子集）。

1. 介绍

通常情况下，当一台路由器重新启动 BGP 时，所有的 BGP 对等体都会检测到会话 DOWN，然后 UP。这种“DOWN/UP”的过渡导致了“路由振荡”并导致 BGP 路由重新计算，产生 BGP 路由更新，以及不必要的扰动转发表。它可能跨越多个路由域。这样的路由振荡可能会产生暂时的转发黑洞和/或瞬态转发环路。他们也消耗受振荡影响的路由器的控制平面的资源。就这样，他们不利于整体网络性能。

本文档描述了一种 BGP 机制，其有助于最小化 BGP 重启导致对路由的负面影响。End-of-RIB 标记被详细描述并且可以用来传送路由收敛信息。一种新的 BGP 能力，称为“GR 能力”定义为允许 BGP speaker 表达其具有 BGP 重启期间保持转发状态的能力。最后，概述了 TCP 会话终止/重新建立期间暂时保留路由信息的处理。

1.1 要求规范

关键词“必须”，“不得”，“所需”，“已”，“不”，“应该”，“不应该”，“推荐”，“可能”和“可选”在文档[RFC2119]中所述。

2. End-of-RIB 标记

没有可达网络层可达性信息（NLRI）和空的撤销 NLRI 的 UPDATE 消息被描述为 End-of-RIB 标记，其可以被 BGP speaker 用来指示其 peer 建立会话后完成了初始路由的更新。对于 IPv4 单播地址族，End-of-RIB 标记具有最小长度的 UPDATE 消息[BGP-4]。对于任何其他地址族，这是一个只包含 MP_UNREACH_NLRI 属性[BGP-MP]的 UPDATE 消息，没有该<AFI，SAFI>的撤销路由。

尽管 End-of-RIB 标记被用于 BGP graceful restart 的目的，注意到这样一个标志的产生为了初始更新完成后通常将对有用于路由聚合，因此推荐实际使用。

另外，如果一个 BGP speaker 可以预先向其 peer 表明它会产生 End-of-RIB 标记有益于路由收敛，不管其有什么保护在 BGP 重启期间转发状态的能力。这可以使用在下一节中描述的 GR 能力。

3. GR 能力

GR 能力是一种新的 BGP 能力[BGP-CAP]，BGP speaker 可以使用它来表示其在 BGP 重启期间保留转发状态的能力。它也可以用来向其 peer 传达其完成初始路由更新后，会产生 End-of-RIB 标记的意图。

这个能力被定义如下：

能力代码： 64
能力长度： 可变
能力值： 由“重启标志/Restart Flags”字段，“重启时间/Restart Time”字段，0-63 个元组<AFI，SAFI，Flags 地址族>组成，如下：

```
+-------------------------------------------------+
| Restart Flags (4 bits)                          |
+-------------------------------------------------+
| Restart Time in seconds (12 bits)               |
+-------------------------------------------------+
| Address Family Identifier (16 bits)             |
+-------------------------------------------------+
| Subsequent Address Family Identifier (8 bits)   |
+-------------------------------------------------+
| Flags for Address Family (8 bits)               |
+-------------------------------------------------+
| ...                                             |
+-------------------------------------------------+
| Address Family Identifier (16 bits)             |
+-------------------------------------------------+
| Subsequent Address Family Identifier (8 bits)   |
+-------------------------------------------------+
| Flags for Address Family (8 bits)               |
+-------------------------------------------------+
```

这些字段的用法和含义如下：

重启标志：

该字段包含与重启有关的位标志。

```
 0 1 2 3
+-+-+-+-+
|R|Resv.|
+-+-+-+-+
```

最重要的位被定义为重启状态（R）位，这可以用来避免多个 BGP speakers 互相窥视重新启动等待 End-of-RIB 标记可能造成的死锁。当设置（值 1），这一位表示 BGP speaker 已经重新启动，并且它的 peer 向 speaker 通告路由信息前必须禁止等待来自 speaker 的 End-of-RIB 标记。

剩下的位是保留的，必须被发送者置为 0，并被接收者忽略。

重启时间：

这是重新启动后，BGP 会话重新建立的估计时间（以秒为单位）。万一出现 BGP speaker 重启后不会恢复回来的情况，这可以用来加速路由收敛。

地址族标识符（AFI），子地址族标识符（SAFI）：

AFI 和 SAFI 结合起来表示，GR 支持具有相同 AFI 和 SAFI 的路由。路由可以明确地与特定的 AFI 和 SAFI 相关联，AFI/SAFI 使用[BGP-MP]的编码或隐含地与<AFI=IPv4, SAFI=Unicast>相关联（如果使用）的[BGP-4]的编码。

地址族的标志：

该字段包含与公告的给定 AFI 和 SAFI 的路由相关的位标志。

```
0 1 2 3 4 5 6 7
+-+-+-+-+-+-+-+-+
|F|   Reserved   |
+-+-+-+-+-+-+-+-+
```

最重要的位被定义为转发状态（F）位，可以用来指示在之前的 BGP 重启过程中，对于使用给定的 AFI 和 SAFI 进行公告的路由，转发状态是否确实被保留。当设置（值 1），该位指示转发状态具有被保存了。

剩下的位是预留的，必须被发送者置为 0，并被接收者忽略。

当此能力的发送者在其能力中不包含任何<AFI，SAFI>时，这意味着发送者在 BGP 重启期间不能保存它的转发状态，但接收 speaker 支持处理（如本文件第 4.2 节所定义）。在这种情况下，通过发送者公告的“重启时间”字段的值是无关紧要的。

BGP speaker 不能在能力通告[BGP-CAP]中包含多个实例的 GR 能力。如果在能力公告中携带多个 GR 能力，公告的接收者必须忽略除了最后一个实例的所有 GR 能力。

GR 能力中包含<AFI=IPv4，SAFI=unicast>并不意味着应该通过使用 BGP 多协议扩展[BGP-MP]携带 IPv4 单播路由信息--可以携带在 BGP UPDATE 消息的 NLRI 字段中。

4. 操作

当 BGP 重新启动时，如果一个 BGP speaker 有能力维护其自己的地址族转发状态，其可以公告一个地址族的 GR 能力到他的 peer。另外，即使在 BGP 重启期间 speaker 没有保存它的任何地址族的转发状态的能力，仍然建议发送者对 peer 公告 GR 能力（正如之前所说的那样，是通过在公告的能力中不包括任何的<AFI，SAFI>）。有这样做的两个原因。首先是表明它的意图，在其初始路由更新完成之后，会产生 End-of-RIB 标识，一般来说，这样做有利于路由聚合。二是表示支持希望执行 GR 的 peer。

在 BGP 会话建立之后，End-of-RIB 标识标记必须在 BGP speaker 完成了对某一地址族初始路由更新后（包括当时没有更新发送的情况），由其一次性发送给它的对等体。

需要注意的是，必须遵循正常的 BGP 过程，终止 TCP 会话，由于发送或接收 BGP NOTIFICATION 消息。

重启时间的建议默认值是小于或等于 OPEN 中携带的 HOLDTIME。

在下面的章节中，“重启 speaker”是指其 BGP 已重新启动的路由器，“接收 speaker”是指重新启动的 speaker 的 peer 路由器。

考虑一个地址族的 GR 能力由重启 speaker 公告，并被接收 speaker 悉知，并在它们之间的建立 BGP 会话。以下部分详细介绍了，一旦重启 speaker 重新启动，重启 speaker 以及接收 speaker 必须遵循的规程。

4.1 重启 speaker 的规程

当重启 speaker 重新启动时，如果可能的话，必须保留 Loc-RIB 中的 BGP 路由的转发状态，且必须标记他们为陈旧。转发期间它不能区分陈旧和其他的信息。

为了重新建立与 peer 会话，重启 speaker 必须在 OPEN 消息 GR 能力中设置“重启”位。除非配置允许，否则能力中为某一地址族的“转发状态”位，只有在重新启动时，该地址族转发状态确实已经被保存的情况下才被设置。

一旦重启 speaker 和接收 speaker 之间的会话被重新建立，重启 speaker 将收到和处理来自其 peer 的 BGP 消息。但是，它必须推迟一个地址族的路由选择，直到（a）它收到来自所有 peers 的 End-of-RIB 标志（不包括接收到的能力中设置的“重启状态”位的，不包括那些不公告 GR 能力的），或者（b）下面提到的 Selection_Deferral_Timer 已经过期。注意，在路由选择之前，speaker 没有路由向其 peer 公告，且没有路由来更新转发状态。

在内部网关协议（IGP）和 BGP 都有重新启动的情况下，在 BGP speaker 执行路由选择之前，等待 IGP 收敛可能是有利的。

BGP speaker 执行路由选择后，speaker 的转发状态必须被更新，以及以前任何标记为陈旧的信息必须被删除。Adj-RIB-Out 可以被公告到其 peer。一旦一个地址族的初始更新被完成（包括那种没有路由更新被发送的情况），必须发送 End-of-RIB 标记。

为了设置路由器延迟其路由选择的时间上限，实现必须支持一个（可配置的）定时器强加这个上限。这个定时器被称为“Selection_Deferral_Timer”。这个计时器的值应该很大，以便为重启 speaker 的所有 peer 提供足够的时间，将所有路由发送到重启 speaker。

如果只想在计划的重新启动时应用 GR（而不是计划的和计划外的重新启动），然而完成这些的一种方式，在计划重新启动之后，将转发状态位设置为 1，在所有其他情况下为 0。其他完成这个的方法，超出了本文档的范围。

4.2 接收 speaker 的规程

当重启 speaker 重新启动时，接收 speaker 可能会或可能不会检测到与重启 speaker 的 TCP 会话终止，根据基础的 TCP 实现，不管[BGP-AUTH]是否正在使用，及重启的具体情况。如果它没有检测到旧 TCP 会话终止，并且仍然认为 BGP 会话正在处于建立状态，其必须把 peer 的后续 OPEN 连接视为一个旧的 TCP 会话终止和据此行动（当收到据此 peer 的 GR 能力时）。有关此行为的说明，请参见第 8 节 BGP 有限状态机的条款。

在这篇文章中“相应地行事”意味着以前的 TCP 会话必须被关闭，保留新的会话。注意这个行为与默认行为不同，如[BGP-4]中第 6.8 节所述。由于以前的连接被认为终止，不应该发送 NOTIFICATION 消息 - 前一个 TCP 会话简单地关闭。

当接收 speaker 检测到与发布了 GR 能力的 peer 之间 BGP 会话的 TCP 会话终止时，它必须保留来自 peer 的之前收到的 GR 能力中所有地址族的所有路由，并将其标记为陈旧的路由信息。为了处理可能的连续重启，以前标记为陈旧的路由（来自 peer）必须删除。路由器不能区分转发中的陈旧和其他路由信息。

在重新建立会话时，在接收方发送的 OPEN 消息中 GR 能力中“重启状态”位不得设置，除非接收 speaker 已重新启动。一个地址族的“转发状态”位的存在和设置取决于实际的转发状态和配置。

如果会话在 peer 之前公告的“重启时间”内没有重新建立，接收 speaker 必须删除所保留的 peer 的所有陈旧路由。

一个 BGP speaker 可以通过某种方式来确定是否其 peer 的转发状态仍然是可行的，例如通过双向转发检测[BFD]或通过监控二层信息。这种机制的细节超出了这个文件范围。如果它确定了它的 peer 的转发状态在会话重新建立之前是不可行的，speaker 可以删除所有保留的来自 peer 的陈旧路由。

一旦会话重新建立，如果在新收到的 GR 能力中，特定地址族的“转发状态”位没有设置，或者如果在新收到的 GR 能力中，不包括一特定地址族，或者在重新建立的会话中根本没有收到 GR 能力，那么接收 speaker 必须马上删除来自 peer 的，保留给该地址族的，所有的陈旧的路由。

一旦接收 speaker 完成地址族的初始更新（包括它没有路由发送），其必须发送 End-of-RIB 标记到 peer。

接收 speaker 必须通过从 peer 收到更新路由来替换陈旧的路由。一旦从 peer 收到的一个地址族的 End-of-RIB 标记，它必须马上删除来自 peer 的该地址族的任何仍然被标记为陈旧的路由。

为了设置路由器保留其陈旧路由的时间上限，实现必须支持一个（可配置的）定时器强加这个上限。

5. 更改 BGP 有限状态机

正如上面“接收 speaker 的规程”中所提到的，这个规范修改了 BGP 有限状态机。

对[BGP-4]的具体状态机修改，第 8.2.2 节，如下面所述。

在 IDLE 状态下，进行以下更改。

替换此文字：

- 为 peer 连接初始化所有 BGP 资源，

为

- 为 peer 连接初始化所有 BGP 资源，除了根据这个（GR）规范的“接收 speaker 的规程”一节，保留路由所需的资源，

在 Established 状态下，进行以下更改。

替换此文字：

TCP 连接成功建立（事件 16 或事件 17）的响应指示，第二个连接将被跟踪，直到它发送一个 OPEN 消息。

为

如果会话没有收到具有一个或多个 AFI/SAFI 的 GR 能力，则回应一个 TCP 连接已成功建立的指示（事件 16 或事件 17），第二个连接将被跟踪直到它发送 OPEN 消息。

但是，如果会话收到具有一个或多个 AFI/SAFI 的 GR 能力，则本地系统的事件 16 或事件 17 作为回应：

- 根据这个（GR）规范的“接收 speaker 的规程”一节，保留与此连接相关的所有路由，
- 释放其他 BGP 的所有资源，
- 丢弃与 ESTABLISHED 会话关联的 TCP 连接，
- 为 peer 连接初始化所有 BGP 资源，除了根据这个（GR）规范的“接收 speaker 的规程”一节，保留路由所需的资源，
- 将 ConnectRetryCounter 设置为零，
- 用初始值启动 ConnectRetryTimer，并且
- 将其状态更改为 Connect。

替换此文字：

如果本地系统收到一个 NOTIFICATION 消息（事件 24 或事件 25）或 TcpConnectionFails（事件 18）来自底层 TCP，本地系统：

- 将 ConnectRetryTimer 设置为零，
- 删除与此连接关联的所有路由，
- 释放所有的 BGP 资源，
- 丢弃 TCP 连接，
- 将 ConnectRetryCounter 递增 1，
- 将其状态更改为 IDLE。

为

如果本地系统收到一个 NOTIFICATION 消息（事件 24 或事件 25），或者如果本地系统收到来自底层的 TCP 的 TcpConnectionFails（事件 18）和该会话没有收到具有一个或多个 AFI/SAFI 的 GR 能力，本地系统：

- 将 ConnectRetryTimer 设置为零，
- 删除与此连接关联的所有路由，
- 释放所有的 BGP 资源，
- 丢弃 TCP 连接，
- 将 ConnectRetryCounter 增加 1
- 将其状态更改为闲置。

但是，如果本地系统从底层的 TCP 收到一个 TcpConnectionFails（Event18），以及该会话已经收到具有一个或多个 AFI/SAFI 的 GR 能力，本地系统：

- 将 ConnectRetryTimer 设置为零，
- 根据这个（GR）规范的“接收 speaker 的规程”一节，保留与此连接相关的所有路由，
- 释放所有其他 BGP 资源，
- 丢弃 TCP 连接，
- 将 ConnectRetryCounter 增加 1

- 将其状态更改为闲置。

6. 部署注意事项

虽然本文档中描述的规程将帮助最大限度地减少路由振荡的影响，注意当一个支持 BGP GR 能力的路由器重启，或者没有保留其转发状态的重启（例如，由于电力故障），如果在涉及的路由器完成路由更新和收敛之前，路由信息发生变化，则网络中存在一个潜在的瞬态路由环路或黑洞。另外，取决于网络拓扑，如果不是所有 IBGP speaker 都有 GR 能力的话，当 GR 过程执行时，可能会增加对瞬态路由环路或黑洞的暴露。

重新启动时间，保留路由的上限，以及延迟路由选择的上限，可能需要获得部署经验时调整。

最后，注意到，在 IGP 与 BGP 紧密耦合（即，BGP 和 IGPs 都会重新启动）和 IGP 没有相似的 GR 能力的自治系统（AS）中部署 BGP GR 的好处，相对于 IGP 具有类似的 GR 能力的场景有所降低。

7. 安全考虑

由于有了这个建议，一个新的连接可以导致一个旧的连接终止，似乎打开了拒绝服务攻击的大门。但是，注意到未经身份验证的 BGP 被熟知，在 TCP 传输中很容易通过拒绝服务攻击。TCP 传输通常通过使用[BGP-AUTH]来保护。这样的认证将同样防止通过虚假的新连接的拒绝服务。

如果攻击者能够通过冒充合法的 peer 成功打开 TCP 连接，攻击者的连接将会替换合法的连接，可能使攻击者能够做假路由。然而，我们注意到这样一个路由插入攻击的窗口非常小，因为通过正常的协议操作，合法 peer 将打开一个新的连接，反过来导致攻击者的连接被终止。因此，攻击发展到拒绝服务的形式。

因此得出这样的结论，这个建议并没有改变 BGP-4 的基本的安全模型（和问题）。

我们还注意到，实现可能允许由配置来控制 GR 的使用。如果 GR 不使能，理所当然地，BGP-4 的基本安全模型是不变。