路由反射器

RFC 4456

陶志豪

zhihao.tao@outlook.com

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

Network Working Group
Request for Comments: 4456
Obsoletes: 2796, 1966
Category: Standards Track
Chandra

T. Bates
E. Chen
Cisco Systems
R.

Sonoa Systems
April 2006

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 路由反射：

全连接内部 BGP（IBGP）的替代方案

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

边界网关协议（BGP）是一种为 TCP/IP 互联网设计的自治系统间路由协议。通常，单个 AS 中的所有 BGP speaker 必须为全连接的，以便任何外部路由信息必须重新分配给自治系统（AS）内所有其他路由器。这是一个严重的规模问题，已经在几个替代方案很好地记录了。

本文档描述了一种称为“路由反射”方法的使用和设计来减轻对“全连接”内部 BGP（IBGP）的需求。

本文档废弃 RFC 2796 和 RFC 1966。

1. 介绍

通常，单个 AS 中的所有 BGP speaker 必须为全连接的，以便任何外部路由信息必须重新分配给自治系统（AS）内所有其他路由器。对于一个 AS 中的 n 个 BGP speaker 需要维护 n*（n-1）/2 个唯一的内部 BGP（IBGP）会话。这个“全连通”的要求显然没有缩小大量 IBGP speaker 每次交换大量路由信息，正如现在许多网络中常见的。

这个规模问题已经有很好的记录，还有一些已经提出的建议来减轻这一点的[2,3]。这个文件描绘着另一种替代方式来减轻对“全连接”的需要，被称为“路由反射”，这种方法允许 BGP speaker（称为“路由反射器”）来公告 IBGP 学习的路由到某些 IBGP 对等体。它描述了一般理解 IBGP 的概念的变化，并增加两个新的可选非传递 BGP 属性，以防止路由更新中的环路。

本文档删除了 RFC 2796 [6]和 RFC 1966 [4]。

2. 要求规格

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。

3. 设计标准

路线反射设计满足以下标准。

- o 简单

  任何替代方案必须配置简单和易于理解。

- o 轻松转换

  必须可以无需更改拓扑或 AS 从全连通配置转换。这是在[3]中提出令人遗憾的技术管理开销。

- o 兼容性

  固执的 IBGP peer 必须可以继续成为原始 AS 或域的一部分，不会丢失任何 BGP 路由信息。

这些标准是由操作经验所驱动的,其具有一个非常大型和拓扑丰富的存在许多外部连接的网络的操作经验。

4. 路由反射

路线反射的基本思路很简单。让我们考虑以下图 1 所示的简单示例。

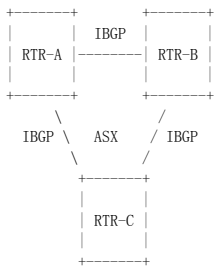

图 1 ： 全连通 IBGP

在 ASX 中，有三个 IBGP speaker（路由器 RTR-A，RTR-B 和 RTR-C）。使用现有的 BGP 模型，如果 RTR-A 收到外部的路由，它被选为最佳路径，它必须公告外部路由到 RTR-B 和 RTR-C。RTR-B 和 RTR-C（作为 IBGP speaker）不会将这些 IBGP 学习路由重新发布给其他 IBGP speaker。

如果放宽此规则，并允许 RTR-C 发布 IBGP 学习到的路由到 IBGP peer，则可以重新公告（或反射）从 RTR-A 学习到的 IBGP 路由到 RTR-B，反之亦然。这将消除 RTR-A 和 RTR-B 之间对 IBGP 会话的需要，如下图 2 所示。

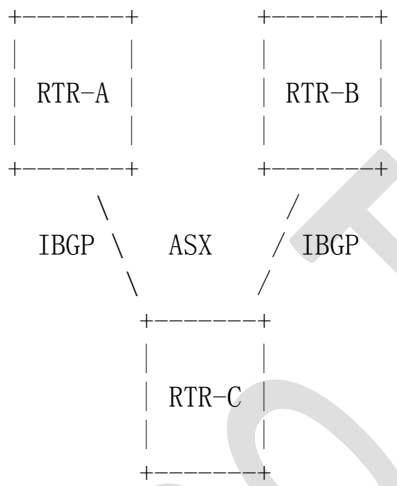

图 2 ： 路由反射 IBGP

路线反射方案是基于这一基本原则。

5. 术语和概念

我们使用术语“路由反射”来描述 BGP speaker 公告一个 IBGP 学习到的路由到另一个 IBGP peer 的操作。BGP 这样的 BGP speaker 被称为“路由反射器”（RR），这样路由被称为反射路由。

RR 的内部 peer 分为两组：

1） 客户端
2） 非客户端 peer

RR 反射这些组之间的路由，可能会在客户 peer 之间反射路由。RR 与其客户端形成一个簇。非客户端 peer 必须全连接，但客户端 peer 不需要全连接。图 3 描绘了一个简单的例子，使用上述术语概述了基本的 RR 组件。

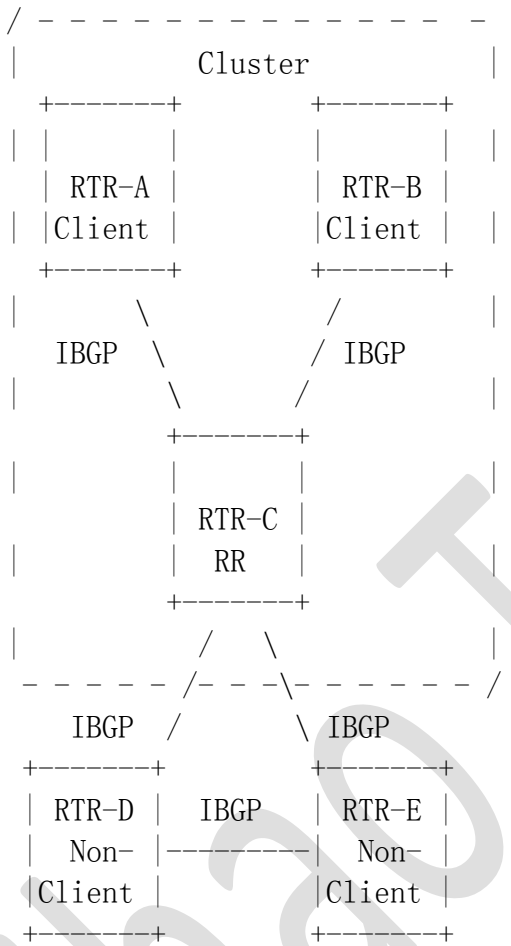

图 3： RR 组件

6. 操作

当 RR 接收到来自 IBGP peer 的路由时，它基于路径选择规则选择最佳的路径。最优路径选择后，它必须根据正在接收最优路径的 peer 的类型进行以下操作：

1） 来自非客户端 IBGP peer 的路由：
反射给所有客户。
2） 来自客户端 peer 的路由：
反射给所有非客户端 peer，也反映给客户端 peer。（因此客户端 peer 不需要全连接。）

一个自治系统可能有很多 RRs。RR 对待其他 RRs 就像任何其他 IBGP speaker 一样。可以配置 RR 的客户端组或非客户端组中具有其他 RR。

在简单的配置中，主干可以分为许多集群。每个 RR 将被配置为其他 RRs 作为非客户端 peers（因此所有 RR 是全连接）。客户端将会配置为仅维护与集群中 RR 的 IBGP 会话。由于路由反射，所有的 IBGP speaker 都会收到反射的路由信息。

在自治系统中可以存在 BGP，其不了解路由反射器的概念（让我们称他们为常规 BGP speaker）。路由反射器方案允许与这样传统的 BGP speaker 共存。传统 BGP speaker 可以是非客户端组或客户端组的成员。这使当前 IBGP 模型轻松和逐步迁移到路由反射模型。可以开始创建集群，通过将单个路由器配置为指定的 RR，且将其他 RRs 及其客户端配置为正常的 IBGP peer。可以逐渐创建额外的集群。

7. 冗余 RR

通常，一组客户端将具有一个单独的 RR。在这种情况下，该集群将由 RR 的 BGP 标识符来标识。然而，这体现单点故障，使其可能在同一个集群中有多个 RRs，同一个集群中所有 RRs 都可以配置一个 4 字节的 CLUSTER_ID，使 RR 可以丢弃同一集群中其他 RRs 的路由。

8. 避免路由信息环路

当路由被反射时，可能会通过配置错误形成路由重新分配环路。路线反射方法定义以下属性来检测和避免路由信息环路：

ORIGINATOR_ID

ORIGINATOR_ID 是一个新的可选非传递 BGP 属性，其类型代码为 9。这个属性是 4 个字节长，它将由反射路由中的一个 RR 创建。该属性将携带本地 AS 中路由的发起方的 BGP 标识符。一个 BGP speaker 不应该创建一个已经存在的 ORIGINATOR_ID 属性。识别 ORIGINATOR_ID 属性的路由器应该忽略使接收用其 BGP 标识符作为 ORIGINATOR_ID 的路由。

CLUSTER_LIST

CLUSTER_LIST 是一个新的可选非传递 BGP 属性，其类型代码为 10.它是一个 CLUSTER_ID 值的序列，表示已经经过路由的反射路径。

当 RR 反射路由时，它必须将本地 CLUSTER_ID 添加到 CLUSTER_LIST。如果 CLUSTER_LIST 是空的，它必须创建一个新的。使用此属性，RR 可以识别由于配置错误导致的路由信息循环回到同一个集群。如果在 CLUSTER_LIST 中找到本地 CLUSTER_ID，收到的公告应该被忽略。

9. 对路线选择的影响

BGP 决策过程 Tie Breaking 规则（Sect9.1.2.2，[1]）修改如下：

如果路由携带 ORIGINATOR_ID 属性，则在步骤 f）中 ORIGINATOR_ID 应该被视为已发布路由的 BGP speaker 的标识符。

另外，在步骤 f）和步骤 g）之间应该插入以下规则：BGP Speaker 应该更喜欢 CLUSTER_LIST 长度较短的路由。如果路由不携带 CLUSTER_LIST 属性，则 CLUSTER_LIST 长度为零。

10. 实现考虑

要注意确保，在 RRs 和客户端及非客户机之间交换的内部路由信息时，没有一条上面定义的 BGP 路径属性可以通过配置进行修改。他们的修改可能会导致路由环路。

另外，当 RR 反射路由时，不应该修改以下路径属性：NEXT_HOP，AS_PATH，LOCAL_PREF 和 MED。

它们的修改可能会导致路由环路。

11. 配置和部署注意事项

BGP 协议不能让客户端动态地识别自身作为 RR 的客户端。最简单的方法来实现这一点是通过手动配置。

路由反射是解决地址扩展问题的关键组成部分之一，RR 汇总路由信息，且只反射其最佳路径。

多出口识别器（MEDs）和内部网关协议（IGP）度量都可能影响 BGP 路由选择。因为 MEDs 是并不总是具有可比性，并且每个路由器的 IGP 度量可能彼此不同，具有某些路由反射拓扑的路由反射方式其可能不会产生与全连接 IBGP 方法相同的路由选择结果。一种使路由选择与全连接 IBGP 方法相同的方法，是确保路由反射器从不被强制执行，基于与客户的 IGP 度量显着不同的 IGP 度量，或基于不可比的 MEDs 的 BGP 路由选择。前者可以通过配置集群内 IGP 度量来实现优于群集间 IGP 度量，并保持在集群内全连通。后者可以实现

- 设置边界路由器上的路由的本地优先级来反映 MED 值，或
- 当 AS 路径长度用作路由选择标准时，确保来自不同 AS 的 AS 路径长度是不同的，或
- 配置基于联盟策略来影响路由选择。

人们可能争辩，在某些情况下后者的要求是过度的，也是不实际的。可能进一步认为，只要没有路由环路，就有没有强制的理由强制路由选择与路由反射器与使用全连接 IBGP 的方法相同。

为了防止路由环路和维护一致的路由视图，在设计路由反射拓扑时，它必须仔细考虑网络拓扑。一般来说，当前缀存在多条路径时，路由反射拓扑应与网络拓扑一致。一个常用的方法是基于存在点（POP）的反射器，其中每个 POP 都维护着自己的路由反射器，为 POP 客户提供服务,且所有路由反射器都是全连接的。另外每个 POP 中的反射器的客户为了最优的 POP 内路由通常是全连接的，以及 POP 内部 IGP 度量被配置为优于 POP 间的 IGP 度量。

12. 安全注意事项

对 BGP 的这种扩展不会改变现有 IBGP 固有的潜在安全问题[1,5]。