IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 15, 2011

E. Jasinska
Limelight Networks
N. Hilliard
INEX
R. Raszuk
Cisco Systems
N. Bakker
AMS-IX B.V.

March 14, 2011

Internet Exchange Route Server
draft-jasinska-ix-bgp-route-server-02

Abstract

This document outlines a specification for multilateral interconnections at Internet exchange points (IXPs). Multilateral interconnection is a method of exchanging routing information between three or more exterior BGP speakers using a single intermediate broker system, referred to as a route server. Route servers are typically used on shared access media networks such as Internet exchange points (IXPs), to facilitate simplified interconnection between multiple Internet routers.

Table of Contents

1. Introduction to Multilateral Interconnection

Internet exchange points (IXPs) provide IP data interconnection facilities for their participants, typically using shared Layer-2 networking media such as Ethernet. The Border Gateway Protocol (BGP) [RFC4271], an inter-Autonomous System routing protocol, is commonly used to facilitate exchange of network reachability information over such media.

While bilateral exterior BGP sessions between exchange participants were previously the most common means of exchanging reachability information, the overhead associated with dense interconnection has caused substantial operational scaling problems for Internet exchange point participants.

Multilateral interconnection is a method of interconnecting BGP speaking routers using a third party brokering system, commonly referred to as a route server and typically managed by the IXP operator. Each of the multilateral interconnection participants (usually referred to as route server clients) announces network reachability information to the route server using exterior BGP, and the route server in turn forwards this information to each other route server client connected to it, according to its configuration. Although a route server uses BGP to exchange reachability information with each of its clients, it does not forward traffic itself and is therefore not a router.

A route server can be viewed as similar in function to an [RFC4456] route reflector, except that it operates using EBGP instead of iBGP. Certain adaptions to [RFC4271] are required, to enable an EBGP router to operate as a route server, which are outlined in Section 2 of this document.

The term "route server" is often in a different context used to describe a BGP node whose purpose is to accept BGP feeds from multiple clients for the purpose of operational analysis and troubleshooting. A system of this form may alternatively be known as a "route collector" or a "route-views server". This document uses the term "route server" exclusively to describe multilateral peering brokerage systems.

1.1. Specification of Requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Technical Considerations for Route Server Implementations

2.1. Client UPDATE Messages

A route server MUST accept all UPDATE messages received from each of its clients for inclusion in its Adj-RIB-In. These UPDATE messages MAY by omitted from the route server's Loc-RIB or Loc-RIBs, due to filters configured for the purposes of implementing routing policy. The route server SHOULD perform one or more BGP Decision Processes to select routes for subsequent advertisement to its clients, taking into account possible configuration to provide multiple NLRI paths to a particular client as described in Section 2.3.2.2 or multiple Loc-RIBs as described in Section 2.3.2.1. The route server SHOULD forward UPDATE messages where appropriate from its Loc-RIB or Loc-RIBs to its clients.

2.2. Attribute Transparency

As a route server primarily performs a brokering service, modification of attributes could cause route server clients to alter their BGP best-path selection process for received prefix reachability information, thereby changing the intended routing policies of exchange participants. Therefore, contrary to what is specified in section 5. of [RFC4271], route servers SHOULD NOT update well-known BGP attributes received from route server clients before redistributing them to their other route server clients. Optional recognized and unrecognized BGP attributes, whether transitive or non-transitive, SHOULD NOT be updated by the route server and SHOULD be passed on to other route server clients.

2.2.1. NEXT_HOP Attribute

The NEXT_HOP, a well-known mandatory BGP attribute, defines the IP address of the router used as the next hop to the destinations listed in the Network Layer Reachability Information field of the UPDATE message. As the route server does not participate in the actual routing of traffic, the NEXT_HOP attribute MUST be passed unmodified to the route server clients, similar to the "third party" next hop feature described in section 5.1.3. of [RFC4271].

2.2.2. AS_PATH Attribute

AS_PATH is a well-known mandatory attribute which identifies the autonomous systems through which routing information carried in the UPDATE message has passed.

As a route server does not participate in the process of forwarding data between client routers, and because modification of the AS_PATH

attribute could affect route server client best-path calculations, the route server SHOULD NOT prepend its own AS number to the AS_PATH segment nor modify the AS_PATH segment in any other way.

2.2.3. MULTI_EXIT_DISC Attribute

MULTI_EXIT_DISC is an optional non-transitive attribute intended to be used on external (inter-AS) links to discriminate among multiple exit or entry points to the same neighboring AS. If applied to an NLRI UPDATE sent to a route server, the attribute (contrary to section 5.1.4 of [RFC4271]) SHOULD be propagated to other route server clients and the route server SHOULD NOT modify the value of this attribute.

2.2.4. Communities Attributes

The BGP COMMUNITIES ([RFC1997]) and Extended Communities ([RFC4360]) attributes are attributes intended for labeling information carried in BGP UPDATE messages. Transitive as well as non-transitive Communities attributes applied to an NLRI UPDATE sent to a route server SHOULD NOT be modified, processed or removed. However, if such an attribute is intended for processing by the route server itself, it MAY be modified or removed.

2.3. Policy in Multilateral Interconnection

While IXP participants often use route servers with the intention of interconnecting with as many other route server participants as possible, there are circumstances where control of path distribution on a per-client basis is important for ensuring that desired interconnection policies are met.

The control of path distribution on a per-client basis can lead to a path being hidden from the route server client, we refer to this as "path hiding" described in Section 2.3.1, evn though it is not a path controlled by the policy.

Route server implementations SHOULD implement one of the methods described in Section 2.3.2, for the operator to be able to allow the control of path distribution on a per-client basis without the occurance of "path hiding".

2.3.1. Path Hiding on a Route Server

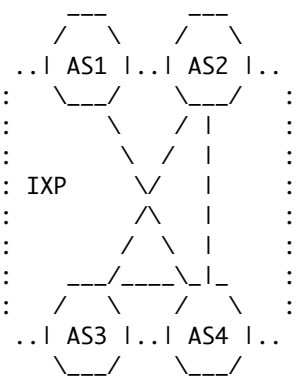

Figure 1: Controlled Interconnection at an IXP

Using the example in Figure 1, AS1 does not directly exchange prefix information with either AS2 or AS3 at the IXP, but only interconnects with AS4.

In the traditional bilateral interconnection model, per-client policy control to a third party exchange participant is accomplished either by not engaging in a bilateral interconnection with that participant or else by implementing outbound filtering on the BGP session towards that participant. However, in a multilateral interconnection environment, only the route server can perform outbound filtering in the direction of the route server client; route server clients depend on the route server to perform their filtering for them.

If the same prefix is sent to a route server from multiple route server clients with different BGP attributes, and traditional best-path route selection is performed on that list of prefixes, then the route server will select a single best-path prefix for propagation to all connected clients. If, however, the route server has been configured to filter the calculated best-path prefix from reaching a particular route server client, then that client will receive no reachability information for that prefix from the route server, despite the fact that the route server has received alternative reachability information for that prefix from other route server clients. This phenomenon is referred to as "path hiding".

For example, in Figure 1, if the same prefix were sent to the route server via AS2 and AS4, and the route via AS2 was preferred according to BGP's traditional best-path selection, but AS2 was filtered by AS1, then AS1 would never receive a path to this prefix, even though

the route server had previously received a valid alternative path via AS4. This happens because the best-path selection is performed only once on the route server for all clients.

Path hiding will only occur on route servers which employ per-client policy control; if an IXP operator deploys a route server without the possibility for policy control, then path hiding does not occur, as all paths are considered equally valid from the point of view of the route server.

2.3.2. Implementing Per-Client Policy Control

For the IXP operator to be able to employ per-client policy control, the route server implementation SHOULD implement a method to allow per-client policy control without the occurance of path hiding.

2.3.2.1. Multiple Route Server RIBs

The most portable means to allow for per-client policy control without the occurance of path hiding, is by using a route server BGP implementation which performs the per-client best-path calculation for each set of paths to a prefix, which results after the route server's client policies have been taken into consideration. This can be implemented by using per-client Loc-RIBs, with path filtering implemented between the Adj-RIB-In and the per-client Loc-RIB. Implementations MAY optimize this by maintaining paths not subject to filtering policies in a global Loc-RIB, with per-client Loc-RIBs stored as deltas.

This implementation is highly portable, as it makes no assumptions about the feature capabilities of the route server clients.

2.3.2.2. Advertising Multiple Paths

The path distribution model described above assumes standard BGP session encoding where the route server sends a single path to its client for any given prefix. This path is selected using the BGP path selection decision process described in [RFC4271]. If, however, it were possible for the route server to send more than a single path to a route server client, then route server clients would no longer depend on receiving a single best path to a particular prefix; consequently, the path hiding problem described in Section 2.3.1 would disappear.

We present two methods which describe how such increased path diversity could be implemented.

2.3.2.2.1. Diverse BGP Path Approach

The Diverse BGP Path proposal as defined in [I-D.ietf-grow-diverse-bgp-path-dist] is a simple way to distribute multiple prefix paths from a route server to a route server client by using a separate BGP session from the route server to a client for each different path.

The number of paths which may be distributed to a client is constrained by the number of BGP sessions which the server and the client are willing to establish with each other. The distributed paths may be established from the global BGP Loc-RIB on the route server in addition to any per-client Loc-RIB. As there may be more potential paths to a given prefix than configured BGP sessions, this method is not guaranteed to eliminate the path hiding problem in all situations. Furthermore, this method may significantly increase the number of BGP sessions handled by the route server, which may negatively impact its performance.

2.3.2.2.2. BGP ADD-PATH Approach

The [I-D.ietf-idr-add-paths] Internet draft proposes a different approach to multiple path propagation, by allowing a BGP speaker to forward multiple paths for the same prefix on a single BGP session. As [RFC4271] specifies that a BGP listener must implement an implicit withdraw when it receives an UPDATE message for a prefix which already exists in its Adj-RIB-In, this approach requires explicit support for the feature both on the route server and on its clients.

If the ADD-PATH capability is negotiated bidirectionally between the route server and a route server client, and the route server client propagates multiple paths for the same prefix to the route server, then this could potentially cause the propagation of inactive, invalid or suboptimal paths to the route server, thereby causing loss of reachability to other route server clients. For this reason, ADD-PATH implementations on a route server SHOULD enforce send-only mode with the route server clients, which would result in negotiating receive-only mode from the client to the route server.

3. Security Considerations

The path hiding problem outlined in section Section 2.3.1 can be used in certain circumstances to proactively block third party path announcements from other route server clients.

4. IANA Considerations

The new set of mechanism for route servers does not require any new allocations from IANA.

5. Acknowledgments

The authors would like to thank Chris Hall, Ryan Bickhart and Steven Bakker for their valuable input.

In addition, the authors would like to acknowledge the developers of BIRD, OpenBGPD and Quagga, whose open source BGP implementations include route server capabilities which are compliant with this document.

6. References

6.1. Normative References

[I-D.ietf-grow-diverse-bgp-path-dist]
Raszuk, R., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of diverse BGP paths.", draft-ietf-grow-diverse-bgp-path-dist-03 (work in progress), January 2011.

[I-D.ietf-idr-add-paths]
Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-04 (work in progress), August 2010.

[RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

6.2. Informative References

[RFC1863] Haskin, D., "A BGP/IDRP Route Server alternative to a full mesh routing", RFC 1863, October 1995.

[RFC3418] Presuhn, R., "Management Information Base (MIB) for the Simple Network Management Protocol (SNMP)", STD 62, RFC 3418, December 2002.

[RFC4223] Savola, P., "Reclassification of RFC 1863 to Historic", RFC 4223, October 2005.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, August 2007.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

Authors' Addresses

Elisa Jasinska
Limelight Networks
2220 W 14th St
Tempe, AZ 85281
US

Email: elisa@llnw.com

Nick Hilliard
INEX
4027 Kingswood Road
Dublin 24
IE

Email: nick@inex.ie

Robert Raszuk
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: raszuk@cisco.com

Niels Bakker
AMS-IX B.V.
Westeinde 12
Amsterdam, NH 1017 ZN
NL

Email: niels.bakker@ams-ix.net