GROW Working Group
Internet-Draft
Intended status: Informational
Expires: September 4, 2014

N. Hilliard
INEX
E. Jasinska
Netflix, Inc
R. Raszuk
NTT I3
N. Bakker
Akamai Technologies B.V.
March 3, 2014

Internet Exchange Route Server Operations
draft-ietf-grow-ix-bgp-route-server-operations-02

Abstract

The popularity of Internet exchange points (IXPs) brings new challenges to interconnecting networks. While bilateral eBGP sessions between exchange participants were historically the most common means of exchanging reachability information over an IXP, the overhead associated with this interconnection method causes serious operational and administrative scaling problems for IXP participants.

Multilateral interconnection using Internet route servers can dramatically reduce the administrative and operational overhead of IXP participation and these systems used by many IXP participants as a preferred means of exchanging routing information.

This document describes operational considerations for multilateral interconnections at IXPs.

Table of Contents

1. Introduction

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] is normally used to facilitate exchange of network reachability information over these media.

As bilateral interconnection between IXP participants requires operational and administrative overhead, BGP route servers [I-D.ietf-idr-ix-bgp-route-server] are often deployed by IXP operators to provide a simple and convenient means of interconnecting IXP participants with each other. A route server redistributes prefixes received from its BGP clients to other clients according to a pre-specified policy, and it can be viewed as similar to an eBGP equivalent of an iBGP [RFC4456] route reflector.

Route servers at IXPs require careful management and it is important for route server operators to thoroughly understand both how they work and what their limitations are. In this document, we discuss several issues of operational relevance to route server operators and provide recommendations to help route server operators provision a reliable interconnection service.

1.1. Notational Conventions

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. Bilateral BGP Sessions

Bilateral interconnection is a method of interconnecting routers using individual BGP sessions between each participant router on an IXP, in order to exchange reachability information. If an IXP participant wishes to implement an open interconnection policy - i.e. a policy of interconnecting with as many other IXP participants as possible - it is necessary for the participant to liaise with each of their intended interconnection partners. Interconnection can then be implemented bilaterally by configuring a BGP session on both participants' routers to exchange network reachability information. If each exchange participant interconnects with each other participant, a full mesh of BGP sessions is needed, as shown in Figure 1.

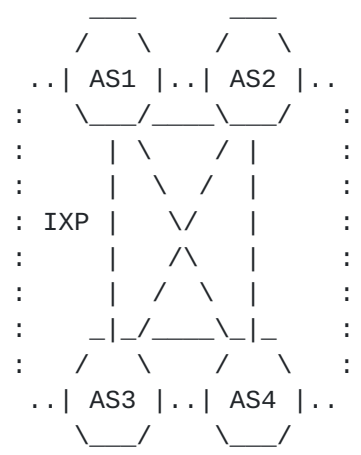

Figure 1: Full-Mesh Interconnection at an IXP

Figure 1 depicts an IXP platform with four connected routers, administered by four separate exchange participants, each of them with a locally unique autonomous system number: AS1, AS2, AS3 and AS4. Each of these four participants wishes to exchange traffic with all other participants; this is accomplished by configuring a full mesh of BGP sessions on each router connected to the exchange, resulting in 6 BGP sessions across the IXP fabric.

The number of BGP sessions at an exchange has an upper bound of n*(n-1)/2, where n is the number of routers at the exchange. As many exchanges have large numbers of participating networks, the amount of administrative and operation overhead required to implement an open interconnection scales quadratically. New participants to an IXP require significant initial resourcing in order to gain value from their IXP connection, while existing exchange participants need to commit ongoing resources in order to benefit from interconnecting with these new participants.

3. Multilateral Interconnection

Multilateral interconnection is implemented using a route server configured to use BGP to distribute network layer reachability information (NLRI) among all client routers. The route server preserves the BGP NEXT_HOP attribute from all received NLRI UPDATE messages, and passes these messages with unchanged NEXT_HOP to its route server clients, according to its configured routing policy, as described in [I-D.ietf-idr-ix-bgp-route-server]. Using this method of exchanging NLRI messages, an IXP participant router can receive an aggregated list of prefixes from all other route server clients using a single BGP session to the route server instead of depending on BGP sessions with each other router at the exchange. This reduces the

overall number of BGP sessions at an Internet exchange from n*(n-1)/2 to n, where n is the number of routers at the exchange.

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.

In practical terms, this allows dense interconnection between IXP participants with low administrative overhead and significantly simpler and smaller router configurations. In particular, new IXP participants benefit from immediate and extensive interconnection, while existing route server participants receive reachability information from these new participants without necessarily having to modify their configurations.

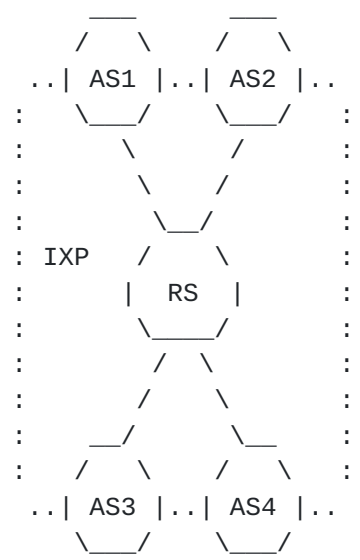

Figure 2: IXP-based Interconnection with Route Server

As illustrated in Figure 2, each router on the IXP fabric requires only a single BGP session to the route server, from which it can receive reachability information for all other routers on the IXP which also connect to the route server.

4. Operational Considerations for Route Server Installations

4.1. Path Hiding

"Path hiding" is a term used in [I-D.ietf-idr-ix-bgp-route-server] to describe the process whereby a route server may mask individual paths by applying conflicting routing policies to its Loc-RIB. When this happens, route server clients receive incomplete information from the route server about network reachability.

There are several approaches which may be used to mitigate against the effect of path hiding; these are described in [I-D.ietf-idr-ix-bgp-route-server]. However, the only method which does not require explicit support from the route server client is for the route server itself to maintain a individual Loc-RIB for each client which is the subject of conflicting routing policies.

4.2. Route Server Scaling

While deployment of multiple Loc-RIBs on the route server presents a simple way to avoid the path hiding problem noted in Section 4.1, this approach requires significantly more computing resources on the route server than where a single Loc-RIB is deployed for all clients. As the [RFC4271] BGP decision process must be applied to all Loc-RIBs deployed on the route server, both CPU and memory requirements on the host computer scale approximately according to O(P * N), where P is the total number of unique paths received by the route server and N is the number of route server clients which require a unique Loc-RIB. As this is a super-linear scaling relationship, large route servers may derive benefit from deploying per-client Loc-RIBs only where they are required.

Regardless of any Loc-RIB optimization technique is implemented, the route server's control plane bandwidth requirements will scale according to O(P * N), where P is the total number of unique paths received by the route server and N is the total number of route server clients. In the case where P_avg (the arithmetic mean number of unique paths received per route server client) remains roughly constant even as the number of connected clients increases, this relationship can be rewritten as O((P_avg * N) * N) or O(N^2). This quadratic upper bound on the network traffic requirements indicates that the route server model will not scale to arbitrarily large sizes.

This scaling analysis presents problems in three key areas: route processor CPU overhead associated with BGP decision process calculations, the memory requirements for handling many different BGP path entries, and the network traffic bandwidth required to distribute these prefixes from the route server to each route server client.

4.2.1. Tackling Scaling Issues

The network traffic scaling issue presents significant difficulties with no clear solution - ultimately, each client must receive a UPDATE for each unique prefix received by the route server. However, there are several potential methods for dealing with the CPU and memory resource requirements of route servers.

4.2.1.1. View Merging and Decomposition

View merging and decomposition, outlined in [RS-ARCH], describes a method of optimising memory and CPU requirements where multiple route server clients are subject to exactly the same routing policies. In this situation, the multiple Loc-RIB views required by each client are merged into a single view.

There are several variations of this approach. If the route server operator has prior knowledge of interconnection relationships between route server clients, then the operator may configure separate Loc-RIBs only for route server clients with unique outbound routing policies. As this approach requires prior knowledge of interconnection relationships, the route server operator must depend on each client sharing their interconnection policies, either in a internal provisioning database controlled by the operator, or else in an external data store such as an Internet Routing Registry Database.

Conversely, the route server implementation itself may implement internal view decomposition by creating virtual Loc-RIBs based on a single in-memory master Loc-RIB, with delta differences for each prefix subject to different routing policies. This allows a more granular and flexible approach to the problem of Loc-RIB scaling, at the expense of requiring a more complex in-memory Loc-RIB structure.

Whatever method of view merging and decomposition is chosen on a route server, pathological edge cases can be created whereby they will scale no better than fully non-optimised per-client Loc-RIBs. However, as most route server clients connect to a route server for the purposes of reducing overhead, rather than implementing complex per-client routing policies, edge cases tend not to arise in practice.

4.2.1.2. Destination Splitting

Destination splitting, also described in [RS-ARCH], describes a method for route server clients to connect to multiple route servers and to send non-overlapping sets of prefixes to each route server. As each route server computes the best path for its own set of prefixes, the quadratic scaling requirement operates on multiple smaller sets of prefixes. This reduces the overall computational and memory requirements for managing multiple Loc-RIBs and performing the best-path calculation on each. In order for this method to perform well, destination splitting would require significant co-ordination between the route server operator and each route server client. In practice, this level of close co-ordination between IXP operators and their participants tends not to occur, suggesting that the approach is unlikely to be of any real use on production IXPs.

4.2.1.3. NEXT_HOP Resolution

As route servers are usually deployed at IXPs which use flat layer 2 networks, recursive resolution of the NEXT_HOP attribute is generally not required, and can be replaced by a simple check to ensure that the NEXT_HOP value for each prefix is a network address on the IXP LAN's IP address range.

4.3. Prefix Leakage Mitigation

Prefix leakage occurs when a BGP client unintentionally distributes NLRI UPDATE messages to one or more neighboring BGP routers. Prefix leakage of this form to a route server can cause serious connectivity problems at an IXP if each route server client is configured to accept all prefix UPDATE messages from the route server. It is therefore RECOMMENDED when deploying route servers that, due to the potential for collateral damage caused by NLRI leakage, route server operators deploy prefix leakage mitigation measures in order to prevent unintentional prefix announcements or else limit the scale of any such leak. Although not foolproof, per-client inbound prefix limits can restrict the damage caused by prefix leakage in many cases. Per-client inbound prefix filtering on the route server is a more deterministic and usually more reliable means of preventing prefix leakage, but requires more administrative resources to maintain properly.

If a route server operator implements per-client inbound prefix filtering, then it is RECOMMENDED that the operator also builds in mechanisms to automatically compare the Adj-RIB-In received from each client with the inbound prefix lists configured for those clients. Naturally, it is the responsibility of the route server client to ensure that their stated prefix list is compatible with what they announce to an IXP route server. However, many network operators do not carefully manage their published routing policies and it is not uncommon to see significant variation between the two sets of prefixes. Route server operator visibility into this discrepancy can provide significant advantages to both operator and client.

4.4. Route Server Redundancy

As the purpose of an IXP route server implementation is to provide a reliable reachability brokerage service, it is RECOMMENDED that exchange operators who implement route server systems provision multiple route servers on each shared Layer-2 domain. There is no requirement to use the same BGP implementation or operating system for each route server on the IXP fabric; however, it is RECOMMENDED that where an operator provisions more than a single server on the same shared Layer-2 domain, each route server implementation be

configured equivalently and in such a manner that the path reachability information from each system is identical.

4.5. AS_PATH Consistency Check

[RFC4271] requires that every BGP speaker which advertises a route to another external BGP speaker prepends its own AS number as the last element of the AS_PATH sequence. Therefore the leftmost AS in an AS_PATH attribute should be equal to the autonomous system number of the BGP speaker which sent the UPDATE message.

As [I-D.ietf-idr-ix-bgp-route-server] suggests that route servers should not modify the AS_PATH attribute, a consistency check on the AS_PATH of an UPDATE received by a route server client would normally fail. It is therefore RECOMMENDED that route server clients disable the AS_PATH consistency check towards the route server.

4.6. Export Routing Policies

Policy filtering is commonly implemented on route servers to provide prefix distribution control mechanisms for route server clients. A route server "export" policy is a policy which affects prefixes sent from the route server to a route server client. Several different strategies are commonly used for implementing route server export policies.

4.6.1. BGP Communities

Prefixes sent to the route server are tagged with specific [RFC1997] or [RFC4360] BGP community attributes, based on pre-defined values agreed between the operator and all client. Based on these community tags, prefixes may be propagated to all other clients, a subset of clients, or none. This mechanism allows route server clients to instruct the route server to implement per-client export routing policies.

As both standard and extended BGP communities values are restricted to 6 octets, the route server operator should take care to ensure that the predefined BGP community values mechanism used on their route server is compatible with [RFC4893] 4-octet autonomous system numbers.

4.6.2. Internet Routing Registry

Internet Routing Registry databases (IRRDBs) may be used by route server operators to implement construct per-client routing policies. [RFC2622] Routing Policy Specification Language (RPSL) provides an comprehensive grammar for describing interconnection relationships,

and several toolsets exist which can be used to translate RPSL policy description into route server configurations.

4.6.3. Client-accessible Databases

Should the route server operator not wish to use either BGP community tags or the public IRRDBs for implementing client export policies, they may implement their own routing policy database system for managing their clients' requirements. A database of this form SHOULD allow a route server client operator to update their routing policy and provide a mechanism for allowing the client to specify whether they wish to exchange all their prefixes with any other route server client. Optionally, the implementation may allow a client to specify unique routing policies for individual prefixes over which they have routing policy control.

4.7. Layer 2 Reachability Problems

Layer 2 reachability problems on an IXP can cause serious operational problems for IXP participants which depend on route servers for interconnection. Ethernet switch forwarding bugs have occasionally been observed to cause non-commutative reachability. For example, given a route server and two IXP participants, A and B, if the two participants can reach the route server but cannot reach each other, then traffic between the participants may be dropped until such time as the layer 2 forwarding problem is resolved. This situation does not tend to occur in bilateral interconnection arrangements, as the routing control path between the two hosts is usually (but not always, due to IXP inter-switch connectivity load balancing algorithms) the same as the data path between them.

Problems of this form can be dealt with using [RFC5881] bidirectional forwarding detection. However, as this is a bilateral protocol configured between routers, and as there is currently no means for automatic configuration of BFD between route server clients, BFD does not currently provide an optimal means of handling the problem.

4.8. BGP NEXT_HOP Hijacking

Section 5.1.3(2) of [RFC4271] allows eBGP speakers to change the NEXT_HOP address of an NLRI update to be a different internet address on the same subnet. This is the mechanism which allows route servers to operate on a shared layer 2 IXP network. However, the mechanism can be abused by route server clients to redirect traffic for their prefixes to other IXP participant routers.

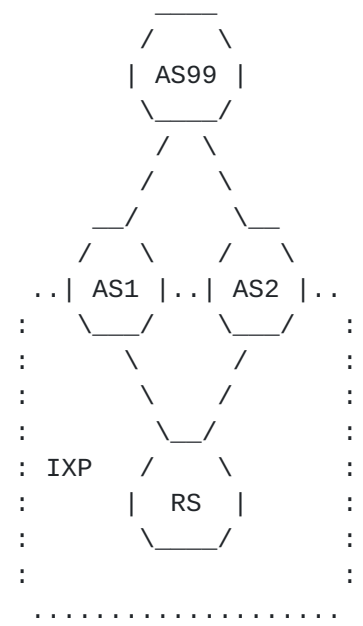

Figure 3: BGP NEXT_HOP Hijacking using a Route Server

For example in Figure 3, if AS1 and AS2 both announce prefixes for AS99 to the route server, AS1 could set the NEXT_HOP address for AS99's prefixes to be the address of AS2's router, thereby diverting traffic for AS99 via AS2. This may override the routing policies of AS99 and AS2.

Worse still, if the route server operator does not use inbound prefix filtering, AS1 could announce any arbitrary prefix to the route server with a NEXT_HOP address of any other IXP participant. This could be used as a denial of service mechanism against either the users of the address space being announced by illicitly diverting their traffic, or the other IXP participant by overloading their network with traffic which would not normally be sent there.

This problem is not specific to route servers and it can also be implemented using bilateral peering sessions. However, the potential damage is amplified by route servers because a single BGP session can be used to affect many networks simultaneously.

Route server operators SHOULD check that the BGP NEXT_HOP attribute for NLRIs received from a route server client matches the interface address of the client. If the route server receives an NLRI where these addresses are different and where the announcing route server client is in a different autonomous system to the route server client which uses the next hop address, the NLRI SHOULD be dropped.

5. Security Considerations

On route server installations which do not employ path hiding mitigation techniques, the path hiding problem outlined in section Section 4.1 can be used in certain circumstances to proactively block third party prefix announcements from other route server clients.

If the route server operator does not implement prefix leakage mitigation as described in section Section 4.3, it is trivial for route server clients to implement denial of service attacks against arbitrary Internet networks using a route server.

Route server installations SHOULD be secured against BGP NEXT_HOP hijacking, as described in section Section 4.8.

6. IANA Considerations

There are no IANA considerations.

7. Acknowledgments

The authors would like to thank Chris Hall, Ryan Bickhart, Steven Bakker and Eduardo Ascenco Reis 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.

8. References

8.1. Normative References

[I-D.ietf-idr-ix-bgp-route-server] Jasinska, E., Hilliard, N., Raszuk, R., and N. Bakker, "Internet Exchange Route Server", draft-ietf-idr-ix-bgp-route-server-03 (work in progress), August 2013.

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

8.2. Informative References

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

[RFC2622] Alaettinoglu, C., Villamizar, C., Gerich, E., Kessens, D., Meyer, D., Bates, T., Karrenberg, D., and M. Terpstra, "Routing Policy Specification Language (RPSL)", RFC 2622, June 1999.

[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.

[RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.

[RS-ARCH] Govindan, R., Alaettinoglu, C., Varadhan, K., and D. Estrin, "A Route Server Architecture for Inter-Domain Routing", 1995, <http://www.cs.usc.edu/research/95-603.ps.Z>.

Authors' Addresses

Nick Hilliard
INEX
4027 Kingswood Road
Dublin 24
IE

Email: nick@inex.ie

Elisa Jasinska
Netflix, Inc
100 Winchester Circle
Los Gatos, CA 95032
USA

Email: elisa@netflix.com

Robert Raszuk
NTT I3
101 S Ellsworth Avenue Suite 350
San Mateo, CA 94401
US

Email: robert@raszuk.net

Niels Bakker
Akamai Technologies B.V.
Kingsfordweg 151
Amsterdam 1043 GR
NL

Email: nbakker@akamai.com