Abstract

The growing popularity of Internet exchange points (IXPs) brings a new set of requirements to interconnect participating networks. 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 IXP participants.

This document outlines a specification for multilateral interconnections at IXPs. Multilateral interconnection describes a method of exchanging routing information between three or more 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 Exchanges (IXPs), to facilitate simplified interconnection between multiple Internet routers on such a network.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.
Table of Contents

1. Introduction to Multilateral Interconnection .......................... 3
   1.1. Specification of Requirements ................................. 3
2. Bilateral Interconnection ........................................... 3
3. Multilateral Interconnection .......................................... 5
4. Technical Considerations for Route Server Implementations ....... 6
   4.1. Client UPDATE Messages ....................................... 6
   4.2. Attribute Transparency ......................................... 6
   4.2.1. NEXT_HOP Attribute ......................................... 6
   4.2.2. AS_PATH Attribute ........................................... 6
   4.2.3. MULTI_EXIT_DISC Attribute ................................ 7
   4.2.4. BGP Community Attributes ................................... 7
4.3. Per-Client Prefix Filtering ........................................ 7
   4.3.1. Prefix Hiding on a Route Server ............................ 7
   4.3.2. Mitigation Techniques ....................................... 9
   4.3.2.1. Multiple Route Server RIBs ............................... 9
   4.3.2.2. Advertising Multiple Paths ............................... 9
5. Operational Considerations for Route Server Installations ...... 10
   5.1. Route Server Scaling ............................................ 10
   5.2. NLRI Leakage Mitigation ........................................ 11
   5.3. Route Server Redundancy ....................................... 11
6. Security Considerations ................................................ 12
7. IANA Considerations .................................................. 12
8. Acknowledgments ....................................................... 12
9. References ............................................................. 12
   9.1. Normative References .......................................... 12
   9.2. Informative References ........................................ 13
Authors’ Addresses ...................................................... 13
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.

In the case of bilateral interconnection between two exchange participant routers, each router must be configured with a BGP session to the other. At IXPs with many participants who wish to implement dense interconnection, this requirement can lead both to large router configurations and high administrative overhead. Given the growth in the number of participants at many IXPs, it has become operationally troublesome to implement densely meshed interconnections at these IXPs.

Multilateral interconnection describes 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 exchange fabric 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.

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. Bilateral Interconnection

Bilateral interconnection describes a method of interconnecting routers using individual BGP sessions between each participant router on an IXP in order to exchange BGP reachability information. While interconnection policies vary from participant to participant, most IXPs have significant numbers of participants who see value in
interconnecting with as many other exchange participants as possible. In order for an IXP participant to implement a dense interconnection policy, it is necessary for the participant to liaise with each of their intended interconnection partners and if this partner agrees to interconnect, then both participants’ routers must be configured with a BGP session to exchange network reachability information. If each exchange participant interconnects with each other participant, a full mesh of BGP sessions is needed, as detailed in Figure 1.

![Diagram of Full-Mesh Interconnection at an IXP](image)

Figure 1 depicts an IXP platform with four connected routers, administered by four separate exchange participants, each 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.

It can be calculated that 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 relatively large numbers of participating networks, the super-linear scaling requirements of dense interconnection tend to cause operational and administrative overhead at large IXPs. In particular, new participants to an IXP require significant initial resourcing in order to gain value from their IXP connection.
3. Multilateral Interconnection

Multilateral interconnection is implemented using a route server configured to use BGP to exchange reachability information between each client router. 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. Using this method of exchanging NLRIs, 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 multiple BGP sessions with each other participant router. This reduces the overall number of BGP sessions over an Internet exchange from $n^2 - n$ to $n$, where $n$ is the number of routers at the exchange.

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 adapt their configurations.

![Diagram of IXP-based Interconnection with Route Server]

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. Technical Considerations for Route Server Implementations

4.1. Client UPDATE Messages

A route server MUST accept all UPDATE messages transmitted to it from each of its clients for inclusion in its Adj-RIB-In. These UPDATE messages may subsequently not be included in the route server’s Loc-RIB or Loc-RIBs, due to filters configured for the purposes of implementing policy routing. 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 4.3.2.2 or multiple Loc-RIBs as described in Section 4.3.2.1. The route server SHOULD forward UPDATE messages where appropriate from its Loc-RIB or Loc-RIBs to its clients.

4.2. 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 5.1 of [RFC4271], route servers should not generally modify BGP attributes received from route server clients before distributing them to their other route server clients.

4.2.1. NEXT_HOP Attribute

The BGP NEXT_HOP 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 [RFC2283].

4.2.2. AS_PATH Attribute

AS_PATH is a mandatory transitive 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 either prepend its own AS number to the AS_PATH segment or modify the AS_PATH segment in any other way.

4.2.3. MULTI_EXIT_DISC Attribute

The multi-exit discriminator 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 SHOULD be treated as a transitive attribute (contrary to section 5.1.4 of [RFC4271]) and the route server SHOULD NOT modify the value of this attribute.

4.2.4. BGP Community Attributes

The BGP COMMUNITIES and Extended Communities attributes are optional transitive attributes intended for labeling information carried in BGP UPDATE messages. If applied to an NLRI UPDATE sent to a route server, these attributes SHOULD NOT generally be modified or removed, except in the case where the attributes are intended for processing by the route server itself.

4.3. Per-Client Prefix Filtering

4.3.1. Prefix Hiding on a Route Server

While IXP participants often use route servers with the intention of interconnecting with as many other route server participants as possible, there are several circumstances where control of prefix distribution on a per-client basis is important for ensuring that the desired interconnection policy is met.
Using the example in Figure 3, 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, prefix filtering to a third party exchange participant is accomplished either by not engaging in a bilateral interconnection with that participant or else by implementing outbound prefix filtering on the BGP session towards that participant. However, in a multilateral interconnection environment, only the route server can perform outbound prefix filtering in the direction of the route server client; other route server clients do not have direct control over what prefix filters the route server employs towards any particular client.

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 "prefix hiding".

For example, in Figure 3, 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 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.

It is noted that prefix hiding will only occur on route servers which employ per-client prefix filtering; if an IXP operator deploys a route server without prefix filtering, then prefix hiding does not occur, as all paths are considered equally valid from the point of view of the route server.

There are several techniques which may be employed to prevent the prefix hiding problem from occurring. Route server implementations SHOULD implement at least one method to prevent prefix hiding.

4.3.2. Mitigation Techniques

4.3.2.1. Multiple Route Server RIBs

The most portable means of preventing the route server prefix hiding problem is by using a route server BGP implementation which performs the per-client best-path calculation for each set of prefixes which results after the route server’s client filtering policies have been taken into consideration. This can be implemented by using per-client Loc-RIBs, with prefix filtering implemented between the Adj-RIB-In and the per-client Loc-RIB. Live implementations will usually optimise this by maintaining prefixes not subject to filtering policies in a global Loc-RIB, with per-client Loc-RIBs stored as deltas.

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

4.3.2.2. Advertising Multiple Paths

The prefix 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 this would alleviate the requirement for route server clients to depend on receiving a single best path to a particular prefix; consequently the prefix hiding problem described in Section 4.3.1 would disappear.

This document discusses two methods which describe how such increased path diversity could be implemented.
4.3.2.2.1. 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 between the route server and the route server 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 route server and the route server client are willing to establish with each other. The distributed paths may be established both 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 prefix hiding problem in all situations.

4.3.2.2.2. Add-Paths 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. Furthermore, if the add-paths 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.

5. Operational Considerations for Route Server Installations

5.1. Route Server Scaling

While deployment of multiple Loc-RIBs on the route server presents a simple way of avoid the prefix hiding problem noted in Section 4.3.1, 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] 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 prefixes 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 server deployments may derive benefit from only deploying per-client Loc-RIBs where they are required.

Regardless of any Loc-RIB optimization implemented, the route server’s network bandwidth requirements will continue to bounded above by a relationship of order $O(P \times N)$, where $P$ is the total number of unique prefixes received by the route server and $N$ is the total number of route server clients. In the case where $P_{avg}$, the mean number of unique prefixes received per route server client, remains roughly constant according as the number of connected clients, this relationship can be rewritten as $O((P_{avg} \times N) \times N)$ or $O(N^2)$. This polynomial upper bound on the network traffic requirements indicates that the route server model will not functionally scale to arbitrarily large sizes.

5.2. NLRI Leakage Mitigation

NLRI leakage occurs when a BGP client unintentionally distributes NLRI UPDATE messages to one or more neighboring BGP routers. NLRI leakage of this form to a route server can cause 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 NLRI 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.

5.3. 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 for each route server on the IXP fabric; however, it is RECOMMENDED that where an operators 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.
6. Security Considerations

On route server installations which do not employ prefix-hiding mitigation techniques, the prefix hiding problem outlined in section Section 4.3.1 can be used in certain circumstances to proactively block third party prefix announcements from other route server clients.

7. IANA Considerations

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

8. Acknowledgments

The authors would like to thank Niels Bakker and Ryan Bickhart 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.

9. References

9.1. 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.",

[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-03 (work in progress), February 2010.


9.2. Informative References


Authors’ Addresses

Elisa Jasinska
AMS-IX B.V.
Westeinde 12
Amsterdam, NH 1017 ZN
NL

Email: elisa.jasinska@ams-ix.net

Nick Hilliard
INEX
4027 Kingswood Road
Dublin 24
IE

Email: nick@inex.ie