Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 7, 2016

R. Bush
Internet Initiative Japan
J. Haas
J. Scudder
Juniper Networks, Inc.
A. Nipper
T. King, Ed.
DE-CIX Management GmbH
July 6, 2015

Making Route Servers Aware of Data Link Failures at IXPs
draft-ietf-idr-rs-bfd-01

Abstract

When route servers are used, the data plane is not congruent with the control plane. Therefore, the peers on the Internet exchange can lose data connectivity without the control plane being aware of it, and packets are dropped on the floor. This document proposes the use of BFD between the two peering routers to detect a data plane failure, and then uses BGP next hop cost to signal the state of the data link to the route server(s).

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without normative meaning.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction 2
2. Operation 3
  2.1. Mutual Discovery of Route Server Client Routers 3
  2.2. Tracking Connectivity 4
3. Advertising Client Router Connectivity to the Route Server . 5
4. Modelling the IXP Network using BGP Link-State 5
5. Utilizing Next Hop Unreachability Information at Client Routers 6
6. Recommendations for Using BFD 6
7. Bootstrapping 8
8. Capability Detection 8
9. Other Considerations 8
10. Acknowledgments 8
11. Normative References 8
Authors' Addresses 9

1. Introduction

In configurations (typically Internet Exchange Points (IXP)) where EBGP routing information is exchanged between client routers through the agency of a route server [I-D.ietf-idr-ix-bgp-route-server], but traffic is exchanged directly, operational issues can arise when partial data plane connectivity exists among the route server client routers. This is because, as the data plane is not congruent with the control plane, the client routers on the IXP can lose data connectivity without the control plane - the route server - being aware of it, and packets are dropped on the floor.

To remedy this, two basic problems need to be solved:

1. Client routers must have a means of verifying connectivity amongst themselves, and

2. Client routers must have a means of communicating the knowledge so gained back to the route server.

The first can be solved by application of Bidirectional Forwarding Detection [RFC5880]. The second can be solved by use of BGP Link-State [I-D.ietf-idr-ls-distribution]. There is a subsidiary problem that must also be solved. Since one of the key value propositions offered by a route server is that client routers need not be configured to peer with each other:

3. Client routers must have a means (other than configuration) to know of one another's existence.

This can also be solved by an application of BGP Link-State.

Throughout this document, we generally assume that the route server being discussed is able to represent different RIBs towards different clients, as discussed in section 2.3.2.1. [I-D.ietf-idr-ix-bgp-route-server]. These procedures (other than the use of BFD to track next hop reachability) have limited value if this is not the case.

2. Operation

Below, we detail procedures where a route server tells its client routers about other client routers (by sending it their next hops using BGP Link-State), the client router verifies connectivity to those other client routers (using BFD) and communicates its findings back to the route server (again using BGP Link-State). The route server uses the received BGP Link-State routes as input to the route selection process it performs on behalf of the client.

2.1. Mutual Discovery of Route Server Client Routers

Strictly speaking, what is needed is not for a route server client router to know of other (control-plane) client routers, but rather to know (so that it can validate) all the next hops the route server might choose to send the client router, i.e. to know of potential forwarding plane relationships.

In effect, this requirement amounts to knowing the BGP next hops the route server is aware of for the particular per-client Loc-RIB (see section 2.3.2.1. [I-D.ietf-idr-ix-bgp-route-server]). We introduce a new table for each client to store known next hops, their compatibility with this proposed solution and their learned

reachability. We call these tables per-client Next Hop Information Base (NHIB). BGP Link-State is used to transfer the NHIBs from the route server to route server clients.

At the route server, the NHIB for each client is populated with the next hops from its Loc-RIB. If the BGP capabilities learned during BGP session setup identify a next hop as compatible with this proposal, this is reflected in the NHIB. Initially, it is assumed that the client router is able to reach its next hops which is stored in the NHIB.

If a next hop is added to the NHIB for a particular client, a route SHOULD be added to the router server's Adj-NHIB-Out. This route contains a BGP Link-State SAFI and models the next hop as node (see section 3.2.1 [I-D.ietf-idr-ls-distribution]) and the connectivity between the route server and the next hop as link (see section 3.2.2 [I-D.ietf-idr-ls-distribution]). If a next hop is removed from a NHIB, the corresponding route in the Adj-NHIB-Out SHOULD be removed.

A route server client SHOULD use BFD [RFC5880] (or other means beyond the scope of this document) to track forwarding plane connectivity to each next hop depicted in the received BGP Link-State information.

2.2. Tracking Connectivity

For each next hop in the NHIB received from the route server (called Adj-NHIB-In), the client router SHOULD use some means to confirm that data plane connectivity does exist to that next hop.

The client router maintains its own NHIB in order to keep track of its (potential) next hops, their capabilities as learned from the route server, and their reachability. The NHIB is updated according to the Adj-NHIB-In and client routers own tests to verify connectivity to next hops.

For each next hop in the Adj-NHIB-In received from the route server, the client router SHOULD evaluate the next hop's compatibility with this proposal. If the next hop supports this proposed mechanism the client router SHOULD setup a BFD session to it if one is not already available and track the reachability of this next hop.

For each next hop in the Adj-NHIB-In, a corresponding BGP Link-State SAFI containing a node NLRI route SHOULD be placed in the client router's own Adj-NHIB-Out to be advertised to the route server. If the next hop is not compatible with this proposal a route containing a BGP Link-State SAFI and a link NLRI SHOULD be placed in the client router's own Adj-NHIB-Out. The link NLRI is configured as follows: the local node is set to the client router, the remote node if set to

the particular next hop. Any next hop that is compatible with this proposal and for which connectivity is in the process of verification (in other words a BFD session is initiated) or is already verified a route containing a BGP Link-State SAFI and a link NLRI as described above SHOULD be placed to the client router's own Adj-NHIB-Out. For any next hop for which connectivity has failed a route SHOULD be placed in the client router's own Adj-NHIB-Out to withdraw the previously advertised link from the route server. (This may also be done as a result of policy even if connectivity exists.)

If the test of connectivity between one client router and another client router has failed the client router that detected this failure should perform connectivity test for a configurable amount of time (preferable 24 hours) on a regular basis (e.g. every 5 minutes). If during this time no connectivity can be restored no more testing is performed until manually changed or the client router is rebooted.

3. Advertising Client Router Connectivity to the Route Server

As discussed above, a client router will advertise its Adj-NHIB-Out to the route server. The route server SHOULD update the reachability information of next hops in the client's NHIB table accordingly. Furthermore, the route server SHOULD use reachability information from the NHIB as input to its own decision process when computing the Adj-RIB-Out for this peer. This peer-dependent Adj-RIB-Out is then advertised to this peer. In particular, the route server MUST exclude any routes whose next hops the client has declared to be not reachable.

4. Modelling the IXP Network using BGP Link-State

This section describes how BGP Link-State is used to a) transfer the per-client NHIB form the route server to the route server clients and b) transfer the reachability information about next hops from the route server client to the route server.

Each route server client and the route server are modeled as nodes (see section 3.2.1 [I-D.ietf-idr-ls-distribution]). As node ID the BGP identifier (see section 1.1 [RFC4271]) is used.

BGP Link-State defines as link a so-called half-way link (see section 3.2.2 [I-D.ietf-idr-ls-distribution]). To cover the bidirectional connectivity between two nodes two link definitions are required. In order to model the connectivity between two route server clients a link is used.

For both nodes and links the Protocol-ID is set to 5 to reflect the virtual modeling. The instance identifier for nodes and links is set to 0 as the default layer 3 routing topology is utilized.

The link descriptor TLV code points 259-262 are applied depending on the IP protocol version used. Prefix descriptors are not applied.

A way is needed to model whether a client router is compatible the mechanisms described in this document or not. For this, a new node descriptor Sub-TVLs (see section 3.2.1.4 [I-D.ietf-idr-ls-distribution]) is introduced.

Sub-TLV Code Point	Description	Length
516	Compatible to this document	1

Table 1: Node Descriptor Sub-TLV

The value of this Sub-TVL is set to 0 if a client router does not support the mechanisms described in this document (of if the support is administratively disabled). Otherwise the value is set to 1.

5. Utilizing Next Hop Unreachability Information at Client Routers

A client router detecting an unreachable next hop signals this information to the route server as described above. Also, it treats the routes as unresolvable as per section 9.1.2.1 [RFC4271] and proceeds with route selection as normal.

Changes in nexthop reachability via these mechanisms should receive some amount of consideration toward avoiding unnecessary route flapping. Similar mechanisms exist in IGP implementations and should be applied to this scenario.

6. Recommendations for Using BFD

The RECOMMENDED way a client router can confirm the data plane connectivity to its next hops is available, is the use of BFD in asynchronous mode. Echo mode MAY be used if both client routers running a BFD session support this. The use of authentication in BFD is OPTIONAL as there is a certain level of trust between the operators of the client routers at a particular IXP. If trust cannot be assumed, it is recommended to use pair-wise keys (how this can be achieved is outside the scope of this document). The ttl/hop limit values as described in section 5 [RFC5881] MUST be obeyed in order to secure BFD sessions from packets coming from outside the IXP.

There is interdependence between the functionality described in this document and BFD from an administrative point of view. To streamline behaviour of different implementations the following is RECOMMENDED:

- If BFD is administratively shut down by the administrator of a client router then the functionality described in this document MUST also be administratively shut down.
- If the administrator enables the functionality described in this document on a client router then BFD MUST be automatically enabled.

The following values of the BFD configuration of client routers (see section 6.8.1 [RFC5880]) are RECOMMENDED in order to allow a fast detection of lost data plane connectivity:

- DesiredMinTxInterval: 1,000,000 (microseconds)
- RequiredMinRxInterval: 1,000,000 (microseconds)
- DetectMult: 3

The configuration values above are a trade-off between fast detection of data plane connectivity and the load client routers must handle keeping up the BFD communication. Selecting smaller DesiredMinTxInterval and RequiredMinRxInterval values generates lots of BFD packets, especially at larger IXPs with many hundreds of client routers.

The configuration values above are selected in order to handle brief interrupts on the data plane. Otherwise, if a BFD session detects a brief data plane interrupt to a particular client router, it will cause to signal the route server that it should remove routes from this client router and tell it shortly afterwards to add the routes again. This is disruptive and computational expensive on the route server.

The configuration values above are also partially impacted by BGP advertisement time in reaction to events from BFD. If the configuration values are selected so that BFD detects data plane interrupts a lot faster than the BGP advertisement time, a data plane connectivity flapping could be detected by BFD but the route server is not informed about them because BGP is not able to transport this information fast enough.

As discussed, finding good configuration values is hard so a client router administrator MAY select better suited values depending on the special needs of the particular deployment.

7. Bootstrapping

If the route server starts it does not know anything about connectivity states between client routers. So, the route server assumes optimistically that all client routers are able to reach each other unless told otherwise.

8. Capability Detection

In order for two BGP speakers to follow the mechanism defined in this document, they MUST use BGP Capabilities Advertisements [RFC5492]. This is done as specified in [RFC4760], by using capability code 1 (multiprotocol BGP), with an AFI XXX and SAFI XXX.

9. Other Considerations

For purposes of routing stability, implementations may wish to apply hysteresis ("holddown") to next hops that have transitioned from reachable to unreachable and back.

10. Acknowledgments

The authors would like to thank the authors of [I-D.ietf-idr-bgp-nh-cost] for their work as it was a basis for this proposal.

11. Normative References

[I-D.ietf-idr-bgp-nh-cost]
Varlashkin, I., Raszuk, R., Patel, K., Bhardwaj, M., and S. Bayraktar, "Carrying next-hop cost information in BGP", draft-ietf-idr-bgp-nh-cost-02 (work in progress), May 2015.

[I-D.ietf-idr-ix-bgp-route-server]
Jasinska, E., Hilliard, N., Raszuk, R., and N. Bakker, "Internet Exchange BGP Route Server", draft-ietf-idr-ix-bgp-route-server-07 (work in progress), June 2015.

[I-D.ietf-idr-ls-distribution]
Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution-11 (work in progress), June 2015.

[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.

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

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

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

Authors' Addresses

Randy Bush
Internet Initiative Japan
5147 Crystal Springs
Bainbridge Island, Washington 98110
US

Email: randy@psg.com

Jeffrey Haas
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: jhaas@juniper.net

John G. Scudder
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: jgs@juniper.net

Arnold Nipper
DE-CIX Management GmbH
Lichtstrasse 43i
Cologne 50825
Germany

Email: arnold.nipper@de-cix.net

Thomas King (editor)
DE-CIX Management GmbH
Lichtstrasse 43i
Cologne 50825
Germany

Email: thomas.king@de-cix.net