Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2017

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

March 11, 2017

Making Route Servers Aware of Data Link Failures at IXPs
draft-ietf-idr-rs-bfd-02

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 a newly defined BGP SAFI 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

In configurations (typically Internet Exchange Points (IXPs)) where EBGP routing information is exchanged between client routers through the agency of a route server [RFC7947], but traffic is exchanged

directly, operational issues can arise when partial data plane connectivity exists among the route server client routers. Since 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, resulting in significant data loss.

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 of the failure back to the route server.

The first can be solved by application of Bidirectional Forwarding Detection [RFC5880]. The second can be solved by exchanging BGP routes which use the RS-Reachable SAFI defined in this document.

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. [RFC7947]. 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 nexthops by sending it RS-Reachable routes, the client router verifies connectivity to those other client routers using BFD and communicates its findings back to the route server using RS-Reachable routes. The route server uses the received routes with RS-Reachable SAFI as input to the route selection process it performs on behalf of the client.

2.1. Mutual Discovery of Route Server Client Next-Hops

Strictly speaking, a route server client does not need to know of other control-plane clients. For validation purposes, it only needs to know the set of next hops the route server might choose to send to it; i.e., to know all potential forwarding plane relationships.

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. [RFC7947]). 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). The NHIB is communicated to the Route Server using RS-Reachable routes.

```
   +--------------------------------------------------------+
   |                     +------------+                     |
   |                     |    Per-    |                     |
   |          .----------->   Client   |----------.          |
   |          |          |    NHIB    |          |          |
   |          |          +------------+          |          |
   |   +------+-----+                      +-----v------+   |
   |   |Adj-NHIB-In |                      |Adj-NHIB-Out|   |
   |   +------^-----+     Route Server     +-----+------+   |
   +----------|----------------------------------|----------+
              |                                  |
              |                                  |
              |                                  |
              |                                  |
   +----------|----------------------------------|----------+
   |   +------+-----+       RS Client      +-----v------+   |
   |   |Adj-NHIB-Out|                      |Adj-NHIB-In |   |
   |   +------^-----+                      +-----+------+   |
   |          |          +------------+          |          |
   |          |          |            |          |          |
   |          `----------+    NHIB    <----------'          |
   |                     |            |                     |
   |                     +------------+                     |
   +--------------------------------------------------------+
```

Figure 1: Route Server, RS Client, and NHIBs with In/Out Queues

The NHIB is not large; the set of routers in the ASs the client has asked the RS to maintain in its view.

At the route server, the Adj-NHIB-Out 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.

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 in its NHIB as received from the RS's Adj-NHIB-Out.

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 exists to that next hop. Here we assume BFD.

The client router maintains its own NHIB in order to keep track of its (potential) next hops 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 attempt to establish a BFD session if one is not already established, and track the reachability of this next hop.

For each nexthop that is determined to be reachable, an entry should be added in the client router's Adj-NHIB-Out to be advertised to the route server. Similarly, when that nexthop is determined to no longer be reachable, the entry should be removed from the client router's Adj-NHIB-Out. This may also be done as a result of policy even if connectivity exists.

If the client can not establish a BFD session with an entry in its NHIB, the next hop is put it in the Adj-NHIB-Out for backward compatibility.

If the test of connectivity between one client router and another client router fails, the client router detecting this failure should perform the connectivity test for a configurable amount of time, preferably 24 hours. 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 client. This client-dependent Adj-RIB-Out is then advertised to this client. In particular, the route server MUST exclude any routes whose next hops the client has declared to be not reachable.

4. Advertising NHIB state in BGP

Two distinct pieces of per-peer state have been identified in the sections above:

- The set of next-hops for BGP routes received from the BGP speaker, the Adj-NHIB-In.
- The set of next-hops the BGP speaker is advertising as reachable, i.e., has potential connectivity to, the Adj-NHIB-Out.

4.1. Using the RS-Reachable SAFI to carry NHIB state

A new BGP SAFI, the RS-Reachable SAFI, is defined in this document. It has been assigned a value TBD. A route server or a route server client using the procedures in this document negotiate the RS-Reachable SAFI for the IPv4 and/or IPv6 AFIs to carry NHIB entries.

NHIB entries are exchanged as host routes using the NLRI format described in [RFC4271], section 4.3. If a NHIB entry for a given AFI is received with an inappropriate prefix length, that NLRI MUST BE ignored.

NHIB entries MUST NOT be propagated from one BGP peering session to another; the routes are not transitive. To help enforce this expected behavior, RS-Reachable routes MUST carry the NO_ADVERTISE community [RFC1997]. RS-Reachable routes not carrying this community MUST BE ignored.

If a NHIB entry is received from a BGP speaker and that entry is not part of the sub-network for that BGP session, that NLRI MUST BE ignored. This prevents erroneous BFD peering session being provisioned outside of the IXP network.

4.2. Specific Procedures for Route Server Clients

A route server SHALL always create an entry in its Adj-NHIB-Out for its clients that are peering with each other through the route server, even if a next hop has not been received for this client. This self-originated entry permits BFD sessions at the clients to be provisioned even if the route exchange via the route server is asymmetric and one router sends routes to the second router in the route server view but not vice versa.

Route server clients are considered to be peering with each other if the configuration of the route server permits routes from a given pair of peers to be mutually exchanged through the route server.

4.3. The RS-Reachable Control Extended Community

```
    0                   1                   2                   3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |      0x43     | Sub-Type TBD1 |    Reserved (Must be Zero)    |
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
   |     Reserved (Must be Zero)   |             Flags           |F|
   +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The RS-Reachable Control Extended Community is used to signal additional information in RS-Reachable NLRI. Currently, a two-octet flag field is utilized for Flags. The remainder of the extended community is currently reserved and its contents MUST be set to zero when originated and SHOULD be ignored upon receipt.

A single flag is currently reserved in this proposal:

F: Flush received NHIB state.

5. Processing NHIB State Changes

5.1. Route Server Client Procedures for NHIB Changes

When entries are added to the a route server client's Adj-NHIB-In for a route server peering session, it will then attempt to verify connectivity to the BGP nexthop for that entry. The procedure described in this specification utilizes BFD; other mechanisms are permitted but are out of scope of this document.

If no existing BFD session exists to this nexthop, a BFD session is provisioned to that IP address and the Adj-NHIB-In (In?) Reachable state is set to Unknown. Since this session requires the remote BFD session to also be provisioned, it may stay in the Down/AdminDown state for a period of time.

If the client can not establish a BFD session with an entry in its NHIB, the next hop is put it in the Adj-NHIB-Out as Reachable for backward compatibility.

Once the BFD session moves to the Up state, the Adj-NHIB-In Reachable state is set to Up. This NHIB entry is now eligible to be placed in Adj-NHIB-Out table and distributed according to the procedures above. Additionally, local BGP route selection may be impacted by this state. See Section 6.

When the BFD session transitions out of the Up state to the Down state, the Adj-NHIB-In Reachable state is set to Down. The NHIB entry MUST be removed from the Adj-NHIB-Out table. This informs the route server that the next hop is no longer reachable.

If the BFD session transitions out of the Up state to the AdminDown state, the Adj-NHIB-In Reachable state is set to AdminDown. During this transition, the NHIB entry is not be removed from the Adj-NHIB-Out table. Instead, the RS-Reachable Extended Community is added to the route with the F (flush) bit set. This signals the route server should remove cached state for this entry.

The motivation for this behavior is that AdminDown could imply one of two possible circumstances:

- The local BFD session has been deconfigured and BFD validation is no longer possible. While the nexthop may still be usable, it is no longer able to be determined using BFD whether that can happen. Removing the entry from the Adj-NHIB-Out will inform the route server that the next hop is no longer reachable and may adversely impact the route server's view supplied to that route server client.
- The remote BFD session has been deconfigured with similar impact.

An implementation of these procedures MUST provide an administrative mechanism to clear such AdminDown entries from the Adj-NHIB-Out table.

When entries are removed from the route server client's Adj-NHIB-In for a route server peering session, the client MAY delay de-provisioning the BFD peering session. If the client delays de-provisioning the session, it should remove it if the BFD session transitions to the Down or AdminDown states. The client should remove the entry from its Adj-NHIB-Out table regardless of the state of the BFD session.

5.2. Route Server Procedures for NHIB Changes

A route server is tracking two distinct types of next hop state for its clients:

- The BGP next hops received from those clients' BGP routes.
- The Adj-NHIB-Out state from each client representing next hops to which the clients believe they have connectivity.

The route-server will place the collection of received BGP next hops from its clients into its per client Adj-NHIB-Out tables when at least one of the route server peers that supports this procedure has negotiated the RS-Reachable SAFI. It will then advertise them per the procedures above. This informs the route server clients of the available BGP nexthops visible to the route server supporting this feature.

In the event that a given client that supports this feature does not provide any routes containing BGP next hops that would be used to populate an Adj-NHIB-Out entry, the route server SHOULD advertise an entry for such a router using the provided self-originated entry. This permits the provisioning of BFD peering sessions for continuity check when route exchange via the route server is asymmetric and one client has routes from a second client, but not vice-versa.

A route server will not generally delete NHIB entries learned in its per client Adj-NHIB-In table when processing a withdraw from the route server client. It derives the following information from the presence and state, or absence, of an entry:

- When an NHIB entry is present, it means that the route server client has noted the BGP next hop from the route server and has validated connectivity to it. Such an entry has the Received state of Active.
- When an entry is withdrawn but was previously present, it means that the route server client previously had validated connectivity to that next hop and NO LONGER has connectivity to it. Such an entry has the Received state of Cached. The route server may choose to adjust what routes are present in that client's view (Adj-Rib-Out) based on that information according to local capability and configuration.
- When an entry is missing, i.e. never has been seen, the route server can't derive any information about the reachability of a given next hop from the perspective of the route server client. The route server SHOULD NOT negatively bias the client's view according to this information.

However, if the route server receives an NHIB entry with the F (flush) bit set the RS-Reachable Control Extended Community, it will remove the entry from the Adj-NHIB-In table for that peer. Similarly, if the entry is being removed because the peering session with the client has closed, entries will also be removed.

6. 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 the above should apply mechanisms to avoid unnecessary route flapping. Such mechanisms exist in IGP implementations which should be applied to this scenario.

7. 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 shield BFD sessions against packets coming from outside the IXP.

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

- If BFD is administratively shut down by the administrator of a client router then the functions described in this document MUST also be administratively shut down.
- If the administrator enables the functions 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 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 excessive BFD packets, especially at larger IXPs with many hundreds of client routers.

The configuration values above were chosen to accept brief interruptions in the data plane. Otherwise, if a BFD session detects a brief data plane interruption to a particular client router, it will signal to the route server that it should remove routes from this client router and shortly thereafter to add the routes again. This is disruptive and computationally 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 interruptions faster than the BGP advertisement time, a data plane connectivity flap could be detected by BFD but the route server is not informed about it because BGP is not able to transport this information quickly enough.

As discussed, finding good configuration values is hard, so a client router administrator MAY select more appropriate values to meet the special needs of a particular deployment.

8. Bootstrapping

During route server start-up, 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.

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. IANA Considerations

IANA is requested to allocate a value from the Subsequent Address Family Identifiers (SAFI) Parameters registry for this proposal. Its Description in that registry shall bgp RS-Reachable with a Reference of this RFC.

IANA is request to allocate a value from the Non-Transitive Opaque Extended Community Sub-Types registry. Its Name will be "RS-Reachable Control Extended Community" with a Reference of this RFC.

11. Security Considerations

The mechanism in this document permits route server clients to influence the contents of the route server's Adj-Ribs-Out through its reports of NHIB state using the Rs-Reachable SAFI. Since this state is per-client, if a route server client is able to inject Rs-Reachable routes for another route server's BGP session to a client, it can cause the route server to select different forwarding than otherwise expected. This issue may be mitigated using transport security on its BGP session to route server clients. See [RFC4272].

Should route server clients provision the RS-Reachable SAFI amongst themselves, it would be an error but would have no undesired impact on forwarding. It is incorrect provisioning for an IXP client which is using a Route Server to have a BGP session with another IXP client. Should they negotiate the RS-Reachable SAFI and send RS-Reachable routes, this only serves to signal that BGP Speaker, when not operating as a route server, to attempt to set verify connectivity with the hosts in the received NLRI. While this may potentially request a large number of sessions, the default BFD timers prevent excess packets from being sent from inappropriately provisioned sessions.

The reachability tests between route server clients themselves may be a target for attack. Such attacks may include forcing a BFD session Down through injecting false BFD state. A less likely attack includes forcing a BFD session to stay Up when its real state is Down. These attacks may be mitigated using the BFD security mechanisms defined in [RFC5880].

12. References

12.1. Normative References

[RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <http://www.rfc-editor.org/info/rfc1997>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <http://www.rfc-editor.org/info/rfc5880>.

[RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, DOI 10.17487/RFC5881, June 2010, <http://www.rfc-editor.org/info/rfc5881>.

[RFC7947] Jasinska, E., Hilliard, N., Raszuk, R., and N. Bakker, "Internet Exchange BGP Route Server", RFC 7947, DOI 10.17487/RFC7947, September 2016, <http://www.rfc-editor.org/info/rfc7947>.

12.2. Informative References

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <http://www.rfc-editor.org/info/rfc4272>.

Appendix A. Summary of Adj-NHIB-In state

The Adj-NHIB-In state is maintained per BGP peering session. It consists of per-peer state and per-peer, per-nexthop state.

```
    +------------------------------------+---------------------------+
    | Client Role                        | (Route-Server |           |
    |                                    |  Route-Server-Client      |
    +------------------------------------+---------------------------+
                Fig. 1  Per-peer Adj-NHIB-In Table State


   +----------------------------+-------------------------------------+
   | NextHop                    | <IPv4 Address | IPv6 Address        |
   +----------------------------+-------------------------------------+
   | Reachable                  | (Unknown | Up | Down | AdminDown)   |
   +----------------------------+-------------------------------------+
                 Fig. 2  Per-peer, per-nexthop  Adj-NHIB-In State
```

Appendix B. Summary of Document Changes

idr-01 to idr-02: Move from BGP-LS to RS-Reachable SAFI. Lots of editorial changes.
idr-00 to idr-01: Add BGP Capability. Move from NH-Cost to BGP-LS.
ymbk-01 to idr-00: No technical changes; adopted by IDR.
ymbk-00 to ymbk-01: Clarifications to BFD procedures. Use BFD state as an input to BGP route selection.

Authors' Addresses

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

Email: randy@psg.com

Jeffrey Haas
Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, CA 94089
US

Email: jhaas@juniper.net

John G. Scudder
Juniper Networks, Inc.
1133 Innovation Way
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