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

February 5, 2015

Making Route-Servers Aware of Data Link Failures at IXPs
draft-ymbk-idr-rs-bfd-00

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

Table of Contents

1. Introduction

In configurations (typically Internet exchanges) 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 Internet exchange 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 NH-SAFI [I-D.ietf-idr-bgp-nh-cost]. 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 NH-SAFI.

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 NH-SAFI), the client router verifies connectivity to those other client routers (using BFD) and communicates its findings back to the route server (again using NH-SAFI). The route server uses the received NH-SAFI 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 in its Adj-RIBs-In. Fortunately, [I-D.ietf-idr-bgp-nh-cost] defines a construct that contains exactly this data, the "Next-Hop Information Base", or NHIB, as well as procedures for a BGP speaker to communicate its NHIB to its peer. Thus, the problem can be solved by the route server advertising its NHIB to its client router, following those procedures.

We observe that (as per NH-SAFI) the cost advertised in the route server's Adj-NHIB-Out need not reflect a "real" IGP cost, the only requirement being that the advertised costs are commensurate. A route server MAY choose to advertise any fixed cost other than all-ones (which is a reserved value in NH-SAFI). This specification does not suggest semantics be imputed to the NH-SAFI advertised by the route server and received by the client, other than "this next hop is present in the control plane, you might like to track it". The route server is not allowed to advertise a next hop as NH_UNREACHABLE.

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

2.2. Tracking Connectivity

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

One way that is highly recommended is the use of BFD in echo mode. Authentication may or may not be used. For each next hop in the Adj-NHIB-In received from the route server, the client router SHOULD setup a BFD session to it if not one is already available and track the reachability of this next hop.

For each next hop being tracked, a corresponding NH-SAFI route should be placed in the client router's own Adj-NHIB-Out to be advertised to the route server. Any next hop for which connectivity has failed should have its cost advertised as NH_UNREACHABLE. (This may also be done as a result of policy even if connectivity exists.) Any other next hop should have some feasible cost advertised. The values advertised may be all equal, or may be set according to policy or other implementation-specific means.

If the test of connectivity between on 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 and this client router is advertised as NH_UNREACHABLE until manually changed or the client router is rebooted.

A client router tracking next hop reachability should also use that determination as input to its own bestpath determination, as per section 9.1 [RFC4271].

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 use this information as input to its own decision process when computing the Adj-RIB-Out for this peer. This peer-dependent Adj-RIB-Out is then be advertised to this peer. In particular, the route server MUST exclude any routes whose next hops the client has declared to be NH_UNREACHABLE. The route server MAY also consider the advertised cost to be the "IGP cost" section 9.1 [RFC4271] when doing this computation.

4. Bootstraping

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.

5. 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.

6. Normative References

[I-D.ietf-idr-bgp-nh-cost]
Varlashkin, I. and R. Raszuk, "Carrying next-hop cost information in BGP", draft-ietf-idr-bgp-nh-cost-01 (work in progress), March 2012.

[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-06 (work in progress), December 2014.

[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.

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