Problem Statement

• In Internet Exchange Point (IXP) environments, routers may peer with each other using BGP Route Servers (RFC 7947).
  • The IXP switching fabric connecting peers typically is supplied as a broadcast link, such as Ethernet.

• Routes learned from a Route Server are not learned directly from the peer supplying the destination or the BGP next-hop.

• In the event of data link issues, the switching fabric may become partitioned. Since the routes learned from the Route Server are not learned directly from a peer, the router may blackhole traffic to that destination.
When the data plane breaks, the control plane can detect this.

When the data plane breaks, the control plane doesn’t notice.
Solution

• Route Server Client routers verify connectivity to the next-hops of the learned routes using BFD (RFC 5880).
  • When a next-hop isn’t reachable via BFD, treat routes using that next-hop as unreachable.
  • This is still useful for non-Route Server scenarios.

• Route Servers tell their clients about available next-hops.
• Route Server clients use this knowledge to provision BFD sessions.
• Route Server clients tell their server about reachability of the monitored next-hops.
• The Route Server can use this next-hop reachability to influence the contents of a Client’s BGP routes in its view.
Next-hop tracking

• This document defines a “Next-Hop Information Base”:
  • Adj-NHIB-In: The nexthops you’ve learned from this mechanism from a peer.
  • Adj-NHIB-Out: The nexthops you’re telling a peer about.

• What you place in the NHIB will depend on the role – are you a route server or its client?
General Procedure

• The Route Server tells its clients about next-hops it knows about for a given client rib-out (view). I.e. puts the next-hops in its Adj-NHIB-Out.
  • It’ll include the next-hops it has in there. Basically, the next-hops in that view’s rib-in.
  • It’ll also include the BGP peering addresses if a next-hop isn’t available. You need this so the BFD session can be provisioned when you have asymmetric distribution of routes through the RS.

• Route Server clients set up BFD sessions to the received next-hops in its Adj-NHIB-In.

• The clients tell the Route Server about whether the next-hop is reachable or not. I.e puts the next-hop in its Adj-NHIB-Out.
What’s changed in this document?

• < -02, NHIB was distributed via nh-cost SAFI (draft-ietf-idr-bgp-nh-cost). This wasn’t quite a good fit for the feature, and overloaded the SAFI inappropriately.

• -02 used BGP-LS. Good fit! However, feedback was that the complexity was too high for this application.

• -03 Proposal to use new RS-Reachable SAFI. Very similar to Nexthop-SAFI but highly simplified for this use case. Expand detail on more of the procedure.
What needs improvement?

• Good discussion on the mailing list; suggestions queued for next rev.

• Discussion among the authors suggests that the NHIB model description is a bit convoluted since the behavior depends on the point of view of the BGP speaker. Is it the Route Server, or its client?
  • Will simplify next rev.

• Current proposal leaves “negative state” tracked by the Route Server that needs to be flushed in some circumstances.
  • Will move to a new mechanism that always sends current state.

• The document has been through three sets of editors and needs cleanup.
  • Will happen in -04.