IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: July 19, 2021

R. Raszuk, Ed.
NTT Network Innovations
C. Cassar
Tesla
E. Aman

B. Decraene, Ed.
Orange
K. Wang
Juniper Networks
January 15, 2021

BGP Optimal Route Reflection (BGP-ORR)
draft-ietf-idr-bgp-optimal-route-reflection-22

Abstract

This document defines an extension to BGP route reflectors. On route reflectors, BGP route selection is modified in order to choose the best path from the standpoint of their clients, rather than from the standpoint of the route reflectors. Multiple types of granularity are proposed, from a per client BGP route selection or to a per peer group, depending on the scaling and precision requirements on route selection. This solution is particularly applicable in deployments using centralized route reflectors, where choosing the best route based on the route reflector IGP location is suboptimal. This facilitates, for example, best exit point policy (hot potato routing).

The solution relies upon all route reflectors learning all paths which are eligible for consideration. Best path selection is performed in each route reflector based on the IGP cost from a selected location in the link state IGP.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on July 19, 2021.

Table of Contents

1. Definitions of Terms Used in This Memo

NLRI - Network Layer Reachability Information

RIB - Routing Information Base

AS - Autonomous System number

VRF - Virtual Routing and Forwarding instance

PE - Provider Edge router

RR - Route Reflector

POP - Point Of Presence

L3VPN - Layer 3 Virtual Private Network [RFC4364]

6PE - IPv6 Provider Edge [RFC4798]

IGP - Interior Gateway Protocol

SPT - Shortest Path Tree

best path - the route chosen by the decision process detailed in [RFC4271] section 9.1.2 and its subsections

best path computation - the decision process detailed in [RFC4271] section 9.1.2 and its subsections

best path algorithm - the decision process detailed in [RFC4271] section 9.1.2 and its subsections

best path selection - the decision process detailed in [RFC4271] section 9.1.2 and its subsections

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 BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

2. Introduction

There are three types of BGP deployments within Autonomous Systems today: full mesh, confederations and route reflection. BGP route reflection [RFC4456] is the most popular way to distribute BGP routes between BGP speakers belonging to the same Autonomous System. However, in some situations, this method suffers from non-optimal path selection.

[RFC4456] asserts that, because the IGP cost to a given point in the network will vary across routers, "the route reflection approach may not yield the same route selection result as that of the full IBGP

mesh approach." One practical implication of this assertion is that the deployment of route reflection may thwart the ability to achieve hot potato routing. Hot potato routing attempts to direct traffic to the closest AS exit point in cases where no higher priority policy dictates otherwise. As a consequence of the route reflection method, the choice of exit point for a route reflector and its clients will be the exit point that is optimal for the route reflector - not necessarily the one that is optimal for its clients.

Section 11 of [RFC4456] describes a deployment approach and a set of constraints which, if satisfied, would result in the deployment of route reflection yielding the same results as the IBGP full mesh approach. This deployment approach makes route reflection compatible with the application of hot potato routing policy. In accordance with these design rules, route reflectors have traditionally often been deployed in the forwarding path and carefully placed on the POP to core boundaries.

The evolving model of intra-domain network design has enabled deployments of route reflectors outside of the forwarding path. Initially this model was only employed for new address families, e.g. L3VPNs and L2VPNs, however it has been gradually extended to other BGP address families including IPv4 and IPv6 Internet using either native routing or 6PE. In such environments, hot potato routing policy remains desirable.

Route reflectors outside of the forwarding path can be placed on the POP to core boundaries, but they are often placed in arbitrary locations in the core of large networks.

Such deployments suffer from a critical drawback in the context of best path selection: A route reflector with knowledge of multiple paths for a given prefix will typically pick its best path and only advertise that best path to its clients. If the best path for a prefix is selected on the basis of an IGP tie-break, the path advertised will be the exit point closest to the route reflector. However, the clients are in a different place in the network topology than the route reflector. In networks where the route reflectors are not in the forwarding path, this difference will be even more acute.

In addition, there are deployment scenarios where service providers want to have more control in choosing the exit points for clients based on other factors, such as traffic type, traffic load, etc. This further complicates the issue and makes it less likely for the route reflector to select the best path from the client's perspective. It follows that the best path chosen by the route reflector is not necessarily the same as the path which would have

been chosen by the client if the client had considered the same set of candidate paths as the route reflector.

3. Modifications to BGP Best Path selection

The core of this solution is the ability for an operator to specify the IGP location for which the route reflector should calculate routes. This can be done on a per route reflector basis, per peer/ update group basis, or per peer basis. This ability enables the route reflector to send to a given set of clients routes with shortest distance to the next hops from the position of the selected IGP location. This provides for freedom of route reflector physical location, and allows transient or permanent migration of this network control plane function to an arbitrary location.

The choice of specific granularity (route reflector, peer/update group, or peer) is configured by the network operator. An implementation is considered compliant with this document if it supports at least one listed grouping of IGP location.

For purposes of route selection, the perspective of a client can differ from that of a route reflector or another client in two

distinct ways:

- it can, and usually will, have a different position in the IGP topology, and
- it can have a different routing policy.

These factors correspond to the issues described earlier.

This document defines, on BGP Route Reflectors [RFC4456], two changes to the BGP Best Path selection algorithm:

- The first change, introduced in Section 3.1, is related to the IGP cost to the BGP Next Hop in the BGP decision process. The change consists in using the IGP cost from a different IGP location than the route reflector itself.
- The second change, introduced in Section 3.2, is to extend the granularity of the BGP decision process, to allow for running multiple decisions process using different perspective or policies.

A route reflector can implement either or both of the modifications in order to allow it to choose the best path for its clients that the clients themselves would have chosen given the same set of candidate paths.

A significant advantage of these approaches is that the route reflector clients do not need to run new software or hardware.

3.1. Best Path Selection from a different IGP location

In this approach, optimal refers to the decision made during best path selection at the IGP metric to BGP next hop comparison step. It does not apply to path selection preference based on other policy steps and provisions.

In addition to the change specified in [RFC4456] section 9, the BGP Decision Process tie-breaking rules ([RFC4271] section 9.1.2.2) are modified as follows.

The below text in step e)

e) Remove from consideration any routes with less-preferred interior cost. The interior cost of a route is determined by calculating the metric to the NEXT_HOP for the route using the Routing Table.

...is replaced by this new text:

e) Remove from consideration any routes with less-preferred interior cost. The interior cost of a route is determined by calculating the metric from the selected IGP location to the NEXT_HOP for the route using the shortest IGP path tree rooted at the selected IGP location.

In order to be able to compute the shortest path tree rooted at the selected IGP locations, knowledge of the IGP topology for the area/level that includes each of those locations is needed. This knowledge can be gained with the use of the link state IGP such as IS-IS [ISO10589] or OSPF [RFC2328] [RFC5340] or via BGP-LS [RFC7752].

The configuration of the IGP location is outside of the scope of this document. The operator may configure it manually, an implementation may automate it based on heuristics, or it can be computed centrally and configured by an external system.

This solution does not require any change (BGP or IGP) on the clients, as all required changes are limited to the route reflector.

This solution applies to NLRIs of all address families that can be route reflected.

3.1.1. Restriction when BGP next hop is BGP prefix

In situations where the BGP next hop is a BGP prefix itself, the IGP metric of a route used for its resolution SHOULD be the final IGP cost to reach such next hop. Implementations which can not inform BGP of the final IGP metric to a recursive next hop SHOULD treat such paths as least preferred during next hop metric comparison. However such paths SHOULD still be considered valid for best path selection.

3.2. Multiple Best Path Selections

BGP Route Reflector as per [RFC4456] runs a single best path selection. Optimal route reflection may require calculation of multiple best path selections or subsets of best path selection in order to consider different IGP locations or BGP policies for different sets of clients.

If the required routing optimization is limited to the IGP cost to the BGP Next-Hop, only step e) as defined [RFC4271] section 9.1.2.2, needs to be duplicated.

If the routing optimization requires the use of different BGP policies for different sets of clients, a larger part of the decision process needs to be duplicated, up to the whole decision process as defined in section 9.1 of [RFC4271]. This is for example the case when there is a need to use different policies to compute different degree of preference during Phase 1. This is needed for use cases involving traffic engineering or dedicating certain exit points for certain clients. In the latter case, the user MAY specify and apply a general policy on the route reflector for a set of clients. For a given set of clients, the policy SHOULD in that case allow the operator to select different candidate exit points for different address families. Regular path selection, including IGP perspective for a set of clients as per Section 3.1, is then applied to the candidate paths to select the final paths to advertise to the clients.

4. Implementation considerations

4.1. Likely Deployments and need for backup

With IGP based optimal route reflection, even though the IGP location could be specified on a per route reflector basis or per peer/update group basis or per peer basis, in reality, it's most likely to be specified per peer/update group basis. All clients with the same or similar IGP location can be grouped into the same peer/update group. An IGP location is then specified for the peer/update group. The location is usually specified as the location of one of the clients

from the peer group or an ABR to the area where clients are located.

Also, one or more backup locations SHOULD be allowed to be specified for redundancy. Implementations may wish to take advantage of peer group mechanisms in order to provide for better scalability of optimal route reflector client groups with similar properties.

5. CPU and Memory Scalability

For IGP based optimal route reflection, determining the shortest path and associated cost between any two arbitrary points in a network based on the IGP topology learned by a router is expected to add some extra cost in terms of CPU resources. However, current SPF tree generation code is implemented efficiently in a number of implementations, and therefore this is not expected to be a major drawback. The number of SPTs computed is expected to be of the order of the number of clients of a route reflector whenever a topology change is detected. It is expected to be higher but comparable to some existing deployed features such as (Remote) Loop Free Alternate which computes a (r)SPT per IGP neighbor.

For policy based optimal route reflection, there will be some overhead to apply the policy to select the candidate paths. This overhead is comparable to existing BGP export policies and therefore should be manageable.

By the nature of route reflection, the number of clients can be split arbitrarily by the deployment of more route reflectors for a given number of clients. While this is not expected to be necessary in existing networks with best in class route reflectors available today, this avenue to scaling up the route reflection infrastructure is available.

If we consider the overall network wide cost/benefit factor, the only alternative to achieve the same level of optimality would require significantly increasing state on the edges of the network. This will consume CPU and memory resources on all BGP speakers in the network. Building this client perspective into the route reflectors seems appropriate.

6. Advantages and Deployment Considerations

The solutions described provide a model for integrating the client perspective into the best path computation for route reflectors. More specifically, the choice of BGP path factors in either the IGP cost between the client and the nexthop (rather than the IGP cost from the route reflector to the nexthop) or other user configured policies.

The achievement of optimal routing relies upon all route reflectors learning all paths that are eligible for consideration. In order to satisfy this requirement, path diversity enhancing mechanisms such as BGP add-path [RFC7911] may need to be deployed between route reflectors.

Implementations considered compliant with this document allow the configuration of a logical location from which the best path will be computed, on the basis of either a peer, a peer group, or an entire routing instance.

These solutions can be deployed in traditional hop-by-hop forwarding networks as well as in end-to-end tunneled environments. In networks where there are multiple route reflectors and hop-by-hop forwarding without encapsulation, such optimizations SHOULD be enabled in a consistent way on all route reflectors. Otherwise, clients may receive an inconsistent view of the network, in turn leading to intra-domain forwarding loops.

With this approach, an ISP can effect a hot potato routing policy even if route reflection has been moved out of the forwarding plane, and hop-by-hop switching has been replaced by end-to-end MPLS or IP encapsulation.

As per above, these approaches reduce the amount of state which needs to be pushed to the edge of the network in order to perform hot potato routing. The memory and CPU resources required at the edge of the network to provide hot potato routing using these approaches is lower than what would be required to achieve the same level of optimality by pushing and retaining all available paths (potentially 10s) per each prefix at the edge.

The solutions above allow for a fast and safe transition to a BGP control plane using centralized route reflection, without compromising an operator's closest exit operational principle. This enables edge-to-edge LSP/IP encapsulation for traffic to IPv4 and IPv6 prefixes.

Regarding Best Path Selection from a different IGP location, it should be self evident that this solution does not interfere with policies enforced above IGP tie-breaking in the BGP best path algorithm.

7. Security Considerations

Similarly to [RFC4456], this extension to BGP does not change the

underlying security issues inherent in the existing IBGP [RFC4456].

It however enables the deployment of base BGP Route Reflection as described in [RFC4456] to be possible using virtual compute environments without any negative consequence on the BGP routing path optimality.

This document does not introduce requirements for any new protection measures, but it also does not relax best operational practices for keeping the IGP network stable or to pace rate of policy based IGP cost to next hops such that it does not have any substantial effect on BGP path changes and their propagation to route reflection clients.

8. IANA Considerations

This document does not request any IANA allocations.

9. Acknowledgments

Authors would like to thank Keyur Patel, Eric Rosen, Clarence Filsfils, Uli Bornhauser, Russ White, Jakob Heitz, Mike Shand, Jon Mitchell, John Scudder, Jeff Haas, Martin Djernaes, Daniele Ceccarelli, Kieran Milne, Job Snijders and Randy Bush for their valuable input.

10. Contributors

Following persons substantially contributed to the current format of the document:

Stephane Litkowski
Cisco System

slitkows.ietf@gmail.com

Adam Chappell
GTT Communications, Inc.
Aspira Business Centre
Bucharova 2928/14a

158 00 Prague 13 Stodulky
Czech Republic

adam.chappell@gtt.net

11. References

11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://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, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

11.2. Informative References

[ISO10589] International Organization for Standardization, "Intermediate system to Intermediate system intra-domain routeing information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473)", ISO/IEC 10589:2002, Second Edition, Nov 2002.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4798] De Clercq, J., Ooms, D., Prevost, S., and F. Le Faucheur, "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, DOI 10.17487/RFC4798, February 2007, <https://www.rfc-editor.org/info/rfc4798>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC6774] Raszuk, R., Ed., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of Diverse BGP Paths", RFC 6774, DOI 10.17487/RFC6774, November 2012, <https://www.rfc-editor.org/info/rfc6774>.

[RFC7752] Gredler, H., Ed., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and Traffic Engineering (TE) Information Using BGP", RFC 7752, DOI 10.17487/RFC7752, March 2016, <https://www.rfc-editor.org/info/rfc7752>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

Appendix A. Appendix: alternative solutions with limited applicability

One possible valid solution or workaround to the best path selection problem requires sending all domain external paths from the route reflector to all its clients. This approach suffers the significant

drawback of pushing a large amount of BGP state and churn to all edge routers. Many networks receive full Internet routing information in a large number of locations. This could easily result in tens of paths for each prefix that would need to be distributed to clients.

Notwithstanding this drawback, there are a number of reasons for sending more than just the single best path to the clients. Improved path diversity at the edge is a requirement for fast connectivity restoration, and a requirement for effective BGP level load balancing.

In practical terms, add/diverse path deployments [RFC7911] [RFC6774] are expected to result in the distribution of 2, 3, or n (where n is a small number) good paths rather than all domain external paths. When the route reflector chooses one set of n paths and distributes them to all its route reflector clients, those n paths may not be the right n paths for all clients. In the context of the problem described above, those n paths will not necessarily include the closest exit point out of the network for each route reflector client. The mechanisms proposed in this document are likely to be complementary to mechanisms aimed at improving path diversity.

Another possibility to optimize exit point selection is the implementation of distributed route reflector functionality at key IGP locations in order to ensure that these locations see their viewpoints respected in exit selection. Typically, however, this requires the installation of physical nodes to implement the reflection, and if exit policy subsequently changes, the reflector placement and position can become inappropriate.

To counter the burden of physical installation, it is possible to build a logical overlay of tunnels with appropriate IGP metrics in order to simulate closeness to key locations required to implement exit policy. There is significant complexity overhead in this approach, however, enough so to typically make it undesirable.

Trends in control plane decoupling are causing a shift from traditional routers to compute virtualization platforms, or even third-party cloud platforms. As a result, without this proposal, operators are left with a difficult choice for the distribution and

reflection of address families with significant exit diversity:

- centralized path selection, and tolerate the associated suboptimal paths, or
- defer selection to end clients, but lose potential route scale capacity

The latter can be a viable option, but it is clearly a decision that needs to be made on an application and address family basis, with strong consideration for the number of available paths per prefix (which may even vary per prefix range, depending on peering policy, e.g. consider bilateral peerings versus onward transit arrangements)

Authors' Addresses

Robert Raszuk (editor)
NTT Network Innovations

Email: robert@raszuk.net

Christian Cassar
Tesla
43 Avro Way
Weybridge KT13 0XY
UK

Email: ccassar@tesla.com

Erik Aman

Email: erik.aman@aman.se

Bruno Decraene (editor)
Orange

Email: bruno.decraene@orange.com

Kevin Wang
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: kfwang@juniper.net