IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 5, 2019

R. Raszuk, Ed.
Bloomberg LP
C. Cassar
Tesla
E. Aman
Telia Company
B. Decraene
Orange
K. Wang
Juniper Networks
April 3, 2019

BGP Optimal Route Reflection (BGP-ORR)
draft-ietf-idr-bgp-optimal-route-reflection-18

Abstract

This document proposes a solution for BGP route reflectors to allow them to choose the best path for their clients that the clients themselves would have chosen under the same conditions, without requiring further state or any new features to be placed on the clients. This facilitates, for example, best exit point policy (hot potato routing). This solution is primarily applicable in deployments using centralized route reflectors.

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 a configured virtual location in the IGP. The location can be the same for all clients or different per peer/update group or per peer. Best path selection can also be performed based on user configured policies in each route reflector.

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 October 5, 2019.

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 Networks RFC4364

6PE - IPv6 Provider Edge Router

IGP - Interior Gateway Protocol

SPT - Shortest Path Tree

best path - the route chosen by the decision process detailed in [RFC 4271] section 9.1.2 and its subsections

best path computation - the decision process detailed in [RFC 4271] section 9.1.2 and its subsections

best path algorithm - the decision process detailed in [RFC 4271] section 9.1.2 and its subsections

best path selection - the decision process detailed in [RFC 4271] 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. Authors

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

Stephane Litkowski
Orange
9 rue du chene germain
Cesson Sevigne, 35512
France

stephane.litkowski@orange.com

Adam Chappell
Interoute Communications
31st Floor
25 Canada Square
London, E14 5LQ
United Kingdom

adam.chappell@interoute.com

3. 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.

3.1. Problem Statement

[RFC4456] asserts that, because the Interior Gateway Protocol (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 best 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 best for the route reflector - not necessarily the one best for the route reflector 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.

3.2. Existing/Alternative Solutions

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 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)

4. Proposed Solutions

The goal of this document is to propose a solution to allow a route reflector to choose the best path for its clients that the clients themselves would have chosen had they considered the same set of candidate paths. 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. Accordingly, we propose two distinct modifications to the best path

algorithm, to address these two distinct factors. A route reflector can implement either or both of the modifications, as needed, 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.

Both modifications rely upon all route reflectors learning all paths that are eligible for consideration. In order to satisfy this requirement, path diversity enhancing mechanisms such as add-path/ diverse paths may need to be deployed between route reflectors.

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

4.1. Client's Perspective IGP Based Best Path Selection

The core of this solution is the ability for an operator to specify on a per route reflector basis or per peer/update group basis or per peer basis the virtual IGP location placement of the route reflector. This enables having a given group of clients receive routes with shortest distance to the next hops from the position of the configured virtual IGP location. This provides for freedom of route reflector location, and allows transient or permanent migration of this network control plane function to an arbitrary location.

The choice of specific granularity left as an implementation decision. An implementation is considered compliant with the document if it supports at least one listed grouping of virtual 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. This approach does not apply to path selection preference based on other policy steps and provisions.

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

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.

This solution does not require any BGP or IGP protocol changes, as all required changes are contained within the route reflector implementation.

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

4.2. Client's Perspective Policy Based Best Path Selection

Optimal route reflection based on virtual IGP location could reflect the best path to the client from IGP cost perspective. However, there are also cases where the client might want the best path based on factors beyond IGP cost. Examples include, but not limited to:

- Selecting the best path for the clients from a traffic engineering perspective.

- Dedicating certain exit points for certain ingress points.

The solution proposed here allows the user to apply a general policy on the route reflector to select a subset of exit points as the candidate exit points for its clients. For a given client, the policy SHOULD also allow the operator to select different candidate exit points for different address families. Regular path selection, including client's perspective IGP based best path selection stated above, will be applied to the candidate paths to select the final paths to advertise to the clients.

Since the policy is applied on the route reflector on behalf of its clients, the route reflector will be able to reflect only the optimal paths to its clients. An additional advantage of this approach is that configuration need only be done on a small number of route reflectors, rather than on a significantly larger number of clients.

4.3. Solution Interactions

Depending on the actual deployment scenarios, service providers may configure IGP based optimal route reflection or policy based optimal route reflection. It is also possible to configure both approaches together. In cases where both are configured together, policy based optimal route reflection will be applied first to select the candidate paths, then IGP based optimal route reflection will be applied on top of the candidate paths to select the final path to advertise to the client.

The expected use case for optimal route reflection is to avoid reflecting all paths to the client because the client either does not support add-paths or does not have the capacity to process all of the

paths. Typically the route reflector would just reflect a single optimal route to the client. However, the solutions MUST NOT prevent reflecting more than one optimal path to the client as path diversity may be desirable for load balancing or fast restoration. In cases where add-path and optimal route reflection are configured together, the route reflector MUST reflect n optimal paths to a client, where n is the add-path count.

The most complicated scenario is where add-path is configured together with both IGP based and policy based optimal route reflection. In this scenario, the policy based optimal route reflection will be applied first to select the candidate paths. Subsequently, IGP based optimal route reflection will be applied on top of the candidate paths to select the best n paths to advertise to the client.

With IGP based optimal route reflection, even though the virtual 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. A virtual IGP location is then specified for the peer/update group. The virtual 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 virtual 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. Advanced optimizations like partial and incremental SPF may also be exploited. The number of SPTs computed 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.

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 the client's IGP best-path selection, 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

No new security issues are introduced to the BGP protocol by this specification.

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. References

10.1. 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>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <https://www.rfc-editor.org/info/rfc5492>.

[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>.

10.2. Informative References

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

[RFC1998] Chen, E. and T. Bates, "An Application of the BGP Community Attribute in Multi-home Routing", RFC 1998, DOI 10.17487/RFC1998, August 1996, <https://www.rfc-editor.org/info/rfc1998>.

[RFC4384] Meyer, D., "BGP Communities for Data Collection", BCP 114, RFC 4384, DOI 10.17487/RFC4384, February 2006, <https://www.rfc-editor.org/info/rfc4384>.

[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>.

[RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, DOI 10.17487/RFC4893, May 2007, <https://www.rfc-editor.org/info/rfc4893>.

[RFC5283] Decraene, B., Le Roux, JL., and I. Minei, "LDP Extension for Inter-Area Label Switched Paths (LSPs)", RFC 5283, DOI 10.17487/RFC5283, July 2008, <https://www.rfc-editor.org/info/rfc5283>.

[RFC5668] Rekhter, Y., Sangli, S., and D. Tappan, "4-Octet AS Specific BGP Extended Community", RFC 5668, DOI 10.17487/RFC5668, October 2009, <https://www.rfc-editor.org/info/rfc5668>.

[RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, DOI 10.17487/RFC5714, January 2010, <https://www.rfc-editor.org/info/rfc5714>.

[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>.

[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>.

Authors' Addresses

Robert Raszuk (editor)
Bloomberg LP
731 Lexington Ave
New York City, NY 10022
USA

Email: robert@raszuk.net

Christian Cassar
Tesla
43 Avro Way
Weybridge KT13 0XY
UK

Email: ccassar@tesla.com

Erik Aman
Telia Company
Solna SE-169 94
Sweden

Email: erik.aman@teliacompany.com

Bruno Decraene
Orange
38-40 rue du General Leclerc
Issy les Moulineaux cedex 9 92794
France

Email: bruno.decraene@orange.com

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

Email: kfwang@juniper.net