IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 13, 2021

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

B. Decraene, Ed.
Orange
K. Wang
Juniper Networks
May 12, 2021

BGP Optimal Route Reflection (BGP-ORR)
draft-ietf-idr-bgp-optimal-route-reflection-23

Abstract

This document defines an extension to BGP route reflectors. On route reflectors, BGP route selection is modified in order to choose the best route from the standpoint of their clients, rather than from the standpoint of the route reflectors. Depending on the scaling and precision requirements, route selection can be specific for one client, common for a set of clients or common for all clients of a route reflector. This solution is particularly applicable in deployments using centralized route reflectors, where choosing the best route based on the route reflector's 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. BGP Route Selection is performed in the route reflectors based on the IGP cost from configured locations in the link state IGP.

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 November 13, 2021.

Table of Contents

1. 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 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 services, e.g. IP VPNs [RFC4364], however it has been gradually extended to other BGP services including IPv4 and IPv6 Internet. 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 BGP Route 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.

2. Terminology

This memo makes use of the terms defined in [RFC4271] and [RFC4456].

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.

3. Modifications to BGP Route Selection

The core of this solution is the ability for an operator to specify the IGP location for which the route reflector calculates interior cost for the NEXT_HOP. The IGP location is defined as a node in the IGP topology and may be configured on a per route reflector basis, per set of clients, or per client 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, set of clients, or client) 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 has 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, for BGP Route Reflectors [RFC4456], two changes to the BGP Route 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 processes using different perspective or policies.

A significant advantage of these approaches is that the route reflector clients do not need to be modified.

3.1. Route Selection from a different IGP location

In this approach, optimal refers to the decision where the interior cost of a route is determined during step e) of [RFC4271] section 9.1.2.2 "Breaking Ties (Phase 2)". 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, [RFC4271] section 9.1.2.2 is 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 way the IGP location is configured is outside 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. One or more backup locations SHOULD be allowed to be specified for redundancy.

3.1.1. Restriction when BGP next hop is a 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 MUST treat such paths as least preferred during next hop metric comparison. However such paths MUST still be considered valid for BGP Phase 2 Route Selection.

3.2. Multiple Route Selections

BGP Route Reflector as per [RFC4456] runs a single BGP Decision Process. Optimal route reflection may require multiple BGP Decision Processes or subsets of the Decision Process 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) and below as defined [RFC4271] section 9.1.2.2, needs to be run multiple times.

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 run multiple times, 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. 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.

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.

4. Deployment Considerations

BGP Optimal Route Reflection provides a model for integrating the client perspective into the BGP Route Selection decision function for route reflectors. More specifically, the choice of BGP path factors in either the IGP cost between the client and the NEXT_HOP (rather than the IGP cost from the route reflector to the NEXT_HOP) or other user configured policies.

The achievement of optimal routing between clients of different clusters relies upon all route reflectors learning all paths that are eligible for consideration. In order to satisfy this requirement, BGP add-path [RFC7911] needs to be deployed between route reflectors.

This solution 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.

As discussed in section 11 of [RFC4456], the IGP locations of BGP route reflectors is important and has routing implications. This equally applies to the choice of the IGP locations configured on optimal route reflectors. After selecting suitable IGP locations, an operator may let one or multiple route reflectors handle route selection for all of them. The operator may alternatively deploy one or multiple route reflector for each IGP location or create any design in between. This choice may depend on operational model (centralized vs per region), acceptable blast radius in case of failure, acceptable number of IBGP sessions for the mesh between the route reflectors, performance and configuration granularity of the equipment.

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. Compared with a deployment of ADD-PATH on all routers, BGP ORR reduces the amount of state which needs to be pushed to the edge of the network in order to perform hot potato routing.

Modifying the IGP location of BGP ORR does not interfere with policies enforced before IGP tie-breaking (step e) in the BGP Decision Process Route.

Calculating routes for different IGP locations requires multiple SPF calculations and multiple (subsets of) BGP Decision Processes, which requires more computing resources. This document allows for different granularity such as one Decision Process per route reflector, per set of clients or per client. A more fine grained granularity may translate into more optimal hot potato routing at the cost of more computing power. The ability to run fine grained computations depends on the platform/hardware deployed, the number of clients, the number of BGP routes and the size of the IGP topology. In essence, sizing considerations are similar to the deployments of BGP Route Reflector.

5. Security Considerations

Similarly to [RFC4456], this extension to BGP does not change the underlying security issues inherent in the existing IBGP.

This document does not introduce requirements for any new protection measures.

6. IANA Considerations

This document does not request any IANA allocations.

7. 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.

8. 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

9. References

9.1. 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>.

9.2. 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>.

[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>.

[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>.

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