IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 28, 2015

R. Raszuk
Mirantis Inc.
C. Cassar
Cisco Systems
E. Aman
TeliaSonera
B. Decraene
S. Litkowski
Orange

April 26, 2015

BGP Optimal Route Reflection (BGP-ORR)
draft-ietf-idr-bgp-optimal-route-reflection-09

Abstract

[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 closest AS egress 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 egress point closest to the route reflector - and not necessarily closest to the RR clients.

Section 11 of [RFC4456] describes a deployment approach and a set of constraints which, if satsified, would result in the deployment of route reflection yielding the same results as the iBGP full mesh approach. Such a deployment approach would make route reflection compatible with the application of hot potato routing policy.

As networks evolved to accommodate architectural requirements of new services, tunneled (LSP/IP tunneling) networks with centralized route reflectors became commonplace. This is one type of common deployment where it would be impractical to satisfy the constraints described in Section 11 of [RFC4456]. Yet, in such an environment, hot potato routing policy remains desirable.

This document proposes a new solution which can be deployed to facilitate the application of closest exit point policy in centralized route reflection deployments.

Status of This Memo

Copyright Notice

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 is the most popular way to distribute BGP routes between BGP speakers belonging to the same administrative domain. Traditionally route reflectors have been deployed in the forwarding path and carefully placed on the POP to core boundaries. That model of BGP route reflector placement has started to evolve. The placement of route reflectors outside the forwarding path was triggered by applications which required traffic to be tunneled from AS ingress PE to egress PE: for example L3VPN.

This evolving model of intra-domain network design has enabled deployments of centralized route reflectors. Initially this model was only employed for new address families e.g. L3VPNs, L2VPNs etc

With edge to edge MPLS or IP encapsulation also being used to carry internet traffic, this model has been gradually extended to other BGP address families including IPv4 and IPv6 Internet routing. This is also applicable to new services achieved with BGP as control plane for example 6PE.

Such centralized route reflectors 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 (unless other techniques like add paths are in use) pick the best path and only advertise that best path to the the route reflector clients. If the best path for a prefix is selected on the basis of an IGP tie break, the best path advertised from the route reflector to its clients will be the exit point closest to the route reflector. But route reflector clients will be in a place in the network topology which is different from the route reflector. In networks with centralized route reflectors, this difference will be even more acute. 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 considered the same set of candidate paths as the route reflector. Furthermore, the path chosen by the client might have been a better path from that chosen by the route reflector for traffic entering the network at the client. The path chosen by the client would have guaranteed the lowest cost and delay trajectory through the network.

Route reflector clients switch packets using routing information learnt from route reflectors which are not on the forwarding path of the packet through the network even in the absence of end-to-end encapsulation. In those cases the path chosen as best and propagated to the clients will often not be the optimal path chosen by the client given all available paths.

Eliminating the IGP distance to the BGP nexthop as a tie breaker on centralized route reflectors does not address the issue. Ignoring IGP distance to the BGP next hop results in the tie breaking procedure contributing the best path by differentiating between paths using attributes otherwise considered less important than IGP cost to the BGP nexthop.

One possible valid solution or workaround to this problem requires sending all domain external paths from the RR to all its clients. This approach suffers the significant drawback of pushing a large amount of BGP state to all the edge routers. In many networks, the number of EBGP peers over which full Internet routing information is received would correlate directly to the number of paths present in each ASBR. This could easily result in tens of paths for each prefix.

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 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. While the route reflector chooses one set of n paths and distributes those same n paths 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 egress 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.

2. Proposed solution

This document proposes a simple solution to the problem described above - overwrite of the default IGP location placement of the route reflector - which is used for determining cost to the next hop contained in BGP paths.

The presented solution makes it possible for route reflector clients to direct traffic to their closest exit point in hot potato routing deployments, without requiring further state to be pushed out to the edge. This solution is primarily applicable in deployments using centralized route reflectors, which are typically implemented in devices without a capable forwarding plane or which are being moved to the NFV enabled cloud.

The solution rely upon all route reflectors learning all paths which are eligible for consideration for hot potato routing. In order to satisfy this requirement, path diversity enhancing mechanisms such as add paths/diverse paths may need to be deployed between route reflectors.

The core of the proposed solution is the ability for operator to specify on a per route reflector basis or per peer/update group basis or per neighbour basis the virtual IGP location placement allowing to have given group of clients to consider optimal distance to the next hops from the position of the configured virtual IGP location. The choice of specific granularity is left to the implementation decision. Implementation is considered as compliant with the document if it supports at least one listed grouping of virtual IGP placement.

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

By optimal we refer in this document to the decision made during best path selection at the IGP metric to BGP next hop comparison step. Clearly the overall path selection preference may be chosen based other policy step and provisions as defined in this document would not apply.

A significant advantage of this approach is that the RR clients do not need to run new software or hardware.

3. Discussion

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 resource. However SPF tree generation code is now implemented efficiently in a number of implementations, and therefor this is not expected to be a major drawback. The number of SPTs computed in the general non-hierarchical case is expected to be of the order of the number of

clients of an RR whenever a topology change is detected. Advanced optimizations like partial and incremental SPF may also be exploited. 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 would be 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, which, in turn, will consume CPU and memory resources on all BGP speakers in the network. Building this client perspective into the route reflectors seems appropriate.

4. Advantages and deployment considerations

The solution described provides a model for integrating the client perspective into the best path computation for RRs. More specifically, the choice of BGP path factors in the IGP metric between the client and the nexthop, rather than the distance from the RR to the nexthop. The documented method does not require any BGP or IGP protocol changes as required changes are contained within the RR implementation.

This solution can be deployed in traditional hop-by-hop forwarding networks as well as in end-to-end tunneled environments. In the 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 and in turn lead to intra-domain forwarding loops.

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

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

The proposal allows for a fast and safe transition to BGP control plane route reflection without compromising an operator's closest exit operational principle. Hot potato routing is important to most

ISPs. The inability to perform hot potato routing effectively stops migrations to centralized route reflection and edge-to-edge LSP/IP encapsulation for traffic to IPv4 and IPv6 prefixes.

Regarding potential for intra-domain forwarding loops at ASBR level, this could be solved by enforcing external route preference or by performing tunnel to external interface switching action on ASBRs.

Regarding 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.

The solution applies to NLRIs of all address families which can be route reflected.

5. Security considerations

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

6. IANA Considerations

IANA is requested to allocate a type code for the Standard BGP Community to be used for inter cluster propagation of angular position of the clients.

IANA is requested to allocate a new type code from BGP OPEN Optional Parameter Types registry to be used for Group_ID propagation.

7. Acknowledgments

Authors would like to thank Keyur Patel, Eric Rosen, Clarence Filsfils, Uli Bornhauser, Russ White, Jakob Heitz, Mike Shand and Jon Mitchell for their valuable input.

8. References

8.1. Normative References

[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.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

8.2. Informative References

[I-D.ietf-idr-add-paths]
Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-10 (work in progress), October 2014.

[RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

[RFC1998] Chen, E. and T. Bates, "An Application of the BGP Community Attribute in Multi-home Routing", RFC 1998, August 1996.

[RFC4384] Meyer, D., "BGP Communities for Data Collection", BCP 114, RFC 4384, February 2006.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.

[RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", RFC 4893, May 2007.

[RFC5283] Decraene, B., Le Roux, JL., and I. Minei, "LDP Extension for Inter-Area Label Switched Paths (LSPs)", RFC 5283, July 2008.

[RFC5668] Rekhter, Y., Sangli, S., and D. Tappan, "4-Octet AS Specific BGP Extended Community", RFC 5668, October 2009.

[RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.

[RFC6774] Raszuk, R., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of Diverse BGP Paths", RFC 6774, November 2012.

Authors' Addresses

Robert Raszuk
Mirantis Inc.
615 National Ave. #100
Mt View, CA 94043
USA

Email: robert@raszuk.net

Christian Cassar
Cisco Systems
10 New Square Park
Bedfont Lakes, FELTHAM TW14 8HA
UK

Email: ccassar@cisco.com

Erik Aman
TeliaSonera
Marbackagatan 11
Farsta SE-123 86
Sweden

Email: erik.aman@teliasonera.com

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

Email: bruno.decraene@orange.com

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

Email: stephane.litkowski@orange.com