Internet Engineering Task Force (IETF)
Request for Comments: 7964
Category: Standards Track
ISSN: 2070-1721

D. Walton
Cumulus Networks
A. Retana
E. Chen
Cisco Systems, Inc.
J. Scudder
Juniper Networks

September 2016

Solutions for BGP Persistent Route Oscillation

Abstract

Routing information reduction by BGP Route Reflection or Confederation can result in persistent internal BGP route oscillations with certain routing setups and network topologies. This document specifies two sets of additional paths that can be used to eliminate these route oscillations in a network.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7964.

Table of Contents

1. Introduction

As documented in [RFC3345], routing information reduction by BGP Route Reflection [RFC4456] or BGP Confederation [RFC5065] can result in persistent Internal BGP (IBGP) route oscillations with certain routing setups and network topologies. Except for a couple of artificially engineered network topologies, the MULTI_EXIT_DISC (MED) attribute [RFC4271] has played a pivotal role in virtually all known persistent IBGP route oscillations. For the sake of brevity, we use the term "MED-induced route oscillation" hereafter to refer to a persistent IBGP route oscillation in which the MED plays a role.

In order to eliminate MED-induced route oscillations and to achieve consistent routing in a network, a route reflector or a confederation Autonomous System Border Router (ASBR) needs to advertise more than just the best path for an address prefix. Our goal is to identify the necessary set of paths for an address prefix that needs to be advertised by a route reflector or a confederation ASBR to prevent the condition.

In this document, we describe two sets of paths for an address prefix that can be advertised by a BGP route reflector or confederation ASBR to eliminate MED-induced route oscillations in a network. The first set involves all the available paths, and would achieve the same routing consistency as the full IBGP mesh. The second set, which is a subset of the first one, involves the neighbor-AS-based Group Best Paths, and would be sufficient to eliminate MED-induced route oscillations (subject to certain commonly adopted topological constraints).

These paths can be advertised using the mechanism described in ADD-PATH [RFC7911] for advertising multiple paths. No other assumptions in functionality beyond the base BGP specification [RFC4271] are made.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Advertise All the Available Paths

Observe that in a network that maintains a full IBGP mesh, all the BGP speakers have consistent and equivalent routing information. Such a network is thus free of MED-induced route oscillations and other routing inconsistencies such as forwarding loops.

Therefore, one approach is to allow a route reflector or a confederation ASBR to advertise all the available paths for an address prefix. Clearly this approach would yield the same amount of routing information and achieve the same routing consistency as the full IBGP mesh in a network. In this document, "Available Paths" refers to the advertisement of all the available paths.

This approach can be implemented using the mechanism described in ADD-PATH [RFC7911] for advertising multiple paths for certain prefixes.

For the sake of scalability, the advertisement of multiple paths should be limited to those prefixes that are affected by MED-induced route oscillation in a network carrying a large number of alternate paths. A detailed description of how these oscillations can occur can be found in [RFC3345]; the description of how a node would locally detect such conditions is outside the scope of this document.

4. Advertise the Group Best Paths

The term "neighbor-AS" for a route refers to the neighboring autonomous system (AS) from which the route was received. The calculation of the neighbor-AS is specified in Section 9.1.2.2 of [RFC4271], and Section 5.3 of [RFC5065]. By definition, the MED is comparable only among routes with the same neighbor-AS. Thus, the route selection procedures specified in [RFC4271] would conceptually involve two steps: first, organize the paths for an address prefix into groups according to their respective neighbor-ASes, and

calculate the most preferred one (termed "Group Best Path") for each of the groups; then, calculate the overall best path among all the Group Best Paths.

As a practice that is generally recommended (in [RFC4456] and [RFC5065]) and widely adopted, a route reflection cluster or a confederation sub-AS should be designed such that BGP routes from within the cluster (or confederation sub-AS) are preferred over routes from other clusters (or confederation sub-AS) when the decision is based on the IGP cost to the BGP NEXT_HOP. This is typically done by setting IGP metrics for links within a cluster (or confederation sub-AS) to be much smaller than the IGP metrics for the links between the clusters (or confederation sub-AS). This practice helps achieve consistent routing within a route reflection cluster or a confederation sub-AS.

When the aforementioned practice for devising a route reflection cluster or confederation sub-AS is followed in a network, we claim that the advertisement of all the Group Best Paths by a route reflector or a confederation ASBR is sufficient to eliminate MED-induced route oscillations in the network. This claim is validated in Appendix A.

Note that a Group Best Path for an address prefix can be identified by the combination of the address prefix and the neighbor-AS. Thus, this approach can be implemented using the mechanism described in ADD-PATH [RFC7911] for advertising multiple paths, and in this case, the neighbor-AS of a path may be used as the path identifier of the path.

It should be noted that the approach of advertising the Group Best Paths requires certain topological constraints to be satisfied in order to eliminate MED-induced route oscillation. Specific topological considerations are described in [RFC3345].

5. Route Reflection and Confederation

To allow a route reflector or a confederation ASBR to advertise either the Available Paths or Group Best Paths using the mechanism described in ADD-PATH [RFC7911], the following revisions are proposed for BGP Route Reflection and BGP Confederation.

5.1. Route Reflection

For a particular <Address Family Identifier (AFI), Subsequent Address Family (SAFI)>, a route reflector MUST include the <AFI, SAFI> with the "Send/Receive" field set to 2 (send multiple paths) or 3 (send/receive multiple paths) in the ADD-PATH Capability [RFC7911] advertised to an IBGP peer. When the ADD-PATH Capability is also received from the IBGP peer with the "Send/Receive" field set to 1 (receive multiple paths) or 3 (send/receive multiple paths) for the same <AFI, SAFI>, then the following procedures apply:

If the peer is a route reflection client, the route reflector MUST advertise to the peer the Group Best Paths (or the Available Paths) received from its non-client IBGP peers. The route reflector MAY also advertise to the peer the Group Best Paths (or the Available Paths) received from its clients.

If the peer is a non-client, the route reflector MUST advertise to the peer the Group Best Paths (or the Available Paths) received from its clients.

5.2. Confederation

For a particular <AFI, SAFI>, a confederation ASBR MUST include the <AFI, SAFI> with the "Send/Receive" field set to 2 (send multiple paths) or 3 (send/receive multiple paths) in the ADD-PATH Capability [RFC7911] advertised to an IBGP peer, and to a confederation external peer. When the ADD-PATH Capability is also received from the IBGP peer or the confederation-external peer with the "Send/Receive" field set to 1 (receive multiple paths) or 3 (send/receive multiple paths) for the same <AFI, SAFI>, then the following procedures apply:

If the peer is internal, the confederation ASBR MUST advertise to the peer the Group Best Paths (or the Available Paths) received from its confederation-external peers.

If the peer is confederation-external, the confederation ASBR MUST advertise to the peer the Group Best Paths (or the Available Paths) received from its IBGP peers.

6. Deployment Considerations

Some route oscillations, once detected, can be eliminated by simple configuration workarounds. As carrying additional paths impacts the memory usage and routing convergence in a network, it is recommended that the impact be evaluated and the approach of using a configuration workaround be considered in deciding whether to deploy the proposed mechanism in a network. In addition, the advertisement of multiple paths should be limited to those prefixes that are affected by MED-induced route oscillation.

While the route reflectors or confederation ASBRs in a network need to advertise the Group Best Paths or Available Paths, the vast majority of the BGP speakers in the network only need to receive the Group Best Paths or Available Paths, which would involve only minor software changes.

It should be emphasized that, in order to eliminate MED-induced route oscillations in a network using the approach of advertising the Group Best Paths, the recommended practice for devising a route reflection cluster or confederation sub-AS with respect to the IGP metrics ([RFC4456] [RFC5065]) should be followed.

It is expected that the approach of advertising the Group Best Paths would be adequate to achieve consistent routing for the vast majority of the networks. For a network that has a large number of alternate paths, the approach should be a good choice as the number of paths advertised by a reflector or a confederation ASBR is bounded by the number of the neighbor-ASes for a particular address prefix. The additional states for an address prefix would also be per neighbor-AS rather than per path. The number of neighbor-ASes for a particular address prefix is typically small because of the limited number of upstream providers for a customer and the nature of advertising only customer routes at the inter-exchange points.

The approach of advertising the Group Best Paths, however, may still be inadequate for certain networks to avoid other routing inconsistencies such as forwarding loops. The required topological constraints could also be operationally challenging. In these cases the approach of advertising the Available Paths may be used, but should be limited to those prefixes that are affected by MED-induced route oscillation in a network carrying a large number of alternate paths.

7. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing BGP [RFC4271].

8. References

8.1. 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, <http://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, <http://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, <http://www.rfc-editor.org/info/rfc4456>.

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, DOI 10.17487/RFC5065, August 2007, <http://www.rfc-editor.org/info/rfc5065>.

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

8.2. Informative References

[RFC3345] McPherson, D., Gill, V., Walton, D., and A. Retana, "Border Gateway Protocol (BGP) Persistent Route Oscillation Condition", RFC 3345, DOI 10.17487/RFC3345, August 2002, <http://www.rfc-editor.org/info/rfc3345>.

Appendix A. Why the Group Best Paths Are Adequate

It is assumed that the following common practice is followed. A route reflection cluster or a confederation sub-AS should be designed such that the IGP metrics for links within a cluster (or confederation sub-AS) are much smaller than the IGP metrics for the links between the clusters (or confederation sub-AS). This practice helps achieve consistent routing within a route reflection cluster or a confederation sub-AS.

Observe that in a network that maintains full IBGP mesh only, the paths that survive the (Local_Pref, AS-PATH Length, Origin, and MED) comparisons [RFC4271] would contribute to route selection in the network.

Consider a route reflection cluster that sources one or more paths that would survive the (Local_Pref, AS-PATH Length, Origin, and MED) comparisons among all the paths in the network. One of these surviving paths would be selected as the Group Best Path by the route reflector in the cluster. Due to the constraint on the IGP metrics as described previously, this path would remain as the Group Best Path and would be advertised to all other clusters even after a path is received from another cluster.

On the other hand, when no path in a route reflection cluster would survive the (Local_Pref, AS-PATH Length, Origin, and MED) comparisons among all the paths in the network, the Group Best Path (when it exists) for a route reflector would be from another cluster. Clearly, the advertisement of the Group Best Path by the route reflector to the clients only depends on the paths received from other clusters.

Therefore, there is no MED-induced route oscillation in the network as the advertisement of a Group Best Path to a peer does not depend on the paths received from that peer.

The claim for the confederation can be validated similarly.

Acknowledgements

We would like to thank David Cook and Naiming Shen for their contributions to the design and development of the solutions.

Many thanks to Tony Przygienda, Sue Hares, Jon Mitchell, and Paul Kyzivat for their helpful suggestions.

Authors' Addresses

Daniel Walton
Cumulus Networks
140C S. Whisman Rd.
Mountain View, CA 94041
United States of America

Email: dwalton@cumulusnetworks.com

Alvaro Retana
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709
United States of America

Email: aretana@cisco.com

Enke Chen
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134
United States of America

Email: enkechen@cisco.com

John Scudder
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
United States of America

Email: jgs@juniper.net