Network Working Group
Internet-Draft
Intended status: Informational
Expires: October 5, 2016

R. White
Linkedin
A. Retana
Cisco Systems, Inc.
S. Hares
Huawei

April 4, 2016

Filtering of Overlapping Routes
draft-white-grow-overlapping-routes-04

Abstract

This document proposes an optional mechanism to remove a prefix when it overlaps with a functionally equivalent shorter prefix. The proposed mechanism does not require any changes to the BGP protocol.

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 http://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, 2016.

Copyright Notice

Table of Contents

1. Introduction

One cause of the growth of the global Internet's default free zone table size is overlapping routes injected into the routing system to steer traffic among various entry points into a network. Because padding AS Path lengths can only steer inbound traffic in a very small set of cases, and other mechanisms used to steer traffic to a particular inbound point are ineffective when multiple upstream providers are in use, advertising longer prefixes is often the only possible way for an AS to steer traffic into specific entry points along its edge.

These longer prefix routes, called overlapping routes in this document, are often advertised along with a shorter prefix route, called a covering route, in order to ensure connectivity in the case

of link or device failures. Overlapping routes not only add to the load on routers in the Internet core by simply expanding the table size; these routes may be less stable than the covering routes they are paired with.

Given the importance of an autonomous system's ability to steer traffic into specific entry points, simply removing the longer prefixes in a longer prefix (overlapping)/shorter prefix (covering) pair of routes isn't a viable solution.

This document proposes an optional mechanism to remove overlapping routes that are no longer useful for steering traffic towards a specific entry point in a particular AS. Removing these routes would reduce the global table in size, and reduce its instability, while removing no capabilities, nor increasing the average path length.

The mechanism proposed is simple to implement, requiring no changes to BGP [RFC4271] either in packet format or in the decision process. The removal described in this document is akin to filtering, not to route aggregation.

The intent of the mechanism is for it to be used based on local decisions and policies, not on an Internet-wide fashion. It is assumed that network operators using this mechanism have an incentive to do so.

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. Overlapping Route Filtering Mechanism

The handling of overlapping prefixes received from an external peer can be broken down into four parts: marking overlapping routes, preferring marked routes, handling marked routes within the AS, and handling marked routes at the AS exit point.

The initial step in successfully filtering overlapping routes is to identify and mark them. This document proposes the use of a BGP community called BOUNDED for that purpose. Because the operation suggested takes place inside an Autonomous System (AS), then any locally assigned community can be used.

The term BOUNDED is used to refer to a locally assigned community used to mark overlapping routes, and to these marked routes as well.

3.1. Marking Overlapping Routes

As each prefix is received by a BGP speaker from an external peer, it is evaluated in the light of other prefixes already received. If two prefixes overlap in space (such as 192.0.2.0/24 and 192.0.2.128/25, or 2001:DB8::/32 and 2001:DB8:1:/48), the longer prefix SHOULD be BOUNDED if it fully overlaps the covering prefix and it is the best path to the destination.

An overlapping prefix is said to fully overlap the corresponding covering prefix if both have identical AS_PATH attributes (both in length and contents) and the same NEXT_HOP.

3.2. Preferring Marked Routes

Since the same overlapping route may be received at several peering points along the edge of the AS, and the covering route may not be present at each of these points, BOUNDED routes SHOULD be preferred over unmarked routes for overlapping routes to be properly handled. A router which marks an overlapping route should also use one of the two mechanisms described here to insure the marked route is preferred throughout the AS.

Only one method described in this section SHOULD be deployed in any given AS.

3.2.1. Using a Cost Community

The recommended method for preferring BOUNDED routes is to use a Cost Community [I-D.ietf-idr-custom-decision] with the Point of Insertion set to ABSOLUTE_VALUE. This mechanism leaves all existing local policy controls in place within the AS.

If this method is used, only the BOUNDED routes need to be tagged using a lower than default Cost, as routes without a Cost Community are considered to have the default value.

3.2.2. Using the Local Preference

An alternate mechanism which may be used to prefer BOUNDED routes is to set their Local Preference to some number higher than the normal standard policy settings for a particular prefix. It's not important that any particular BOUNDED route win over any other one; so simply adding a small amount to the normal Local Preference, as dictated by local policy, will ensure a BOUNDED route will always win over an unmarked route, so only these routes reach the outbound edge of the AS.

3.3. Handling Marked Routes Within the AS

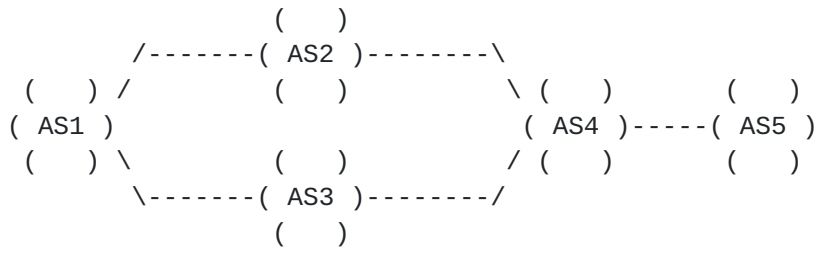
Routes marked with the BOUNDED community MAY not be installed in the local RIB of routers within the AS. This optional step will reduce local RIB and forwarding table usage and volatility within the AS.

3.4. Handling Marked Routes at the Outbound Edge

If local policy dictates, routes marked with the BOUNDED community SHOULD NOT be advertised to external peers. If they are advertised, they MAY then be marked with the NO_EXPORT community.

4. Examples of Filtering Overlapping Routes

Assume the following configuration of autonomous systems:



This network is used in both of the following examples.

4.1. IPv4 Example

- AS1 is advertising 192.0.2.128/25 to both AS2 and AS3.
- AS2 is advertising both 192.0.2.128/25 and 192.0.2.0/24 into AS4.
- AS3 is advertising 192.0.2.128/25 into AS4
- Each BGP connection (session) is handled by a separate router within each AS (for instance, AS4 peers with AS2 and AS3 on separate routers).

When the router in AS4 peering with AS2 receives both the 192.0.2.128/25 and the 192.0.2.0/24 prefixes, it will mark 192.0.2.128/25 as BOUNDED, and set a Cost Community (as described in Section 3.2.1) so the marked overlapping route is preferred over unmarked routes within AS4.

The border router between AS4 and AS3 will receive the longer prefix from AS3, and the preferred BOUNDED overlapping route through iBGP. It will prefer the marked route, so the unmarked route towards 192.0.2.128/25 will not be advertised throughout AS4.

If the link between AS1 and AS2 fails, the longer length prefix will be withdrawn from AS2, and thus the peering point between AS2 and AS4 will no longer have an overlapping set of prefixes. Within AS4, the border router which peers with AS2 will cease advertising the 192.0.2.128/25 prefix, which allows the AS3/AS4 border router to begin advertising it into AS4, and through AS4 into AS5, restoring connectivity to AS1.

4.2. IPv6 Example

- AS1 is advertising 2001:DB8:1:/48 to both AS2 and AS3.
- AS2 is advertising both 2001:DB8:1:/48 and 2001:DB8::/32 into AS4.
- AS3 is advertising 2001:DB8:1:/48 into AS4
- Each BGP connection (session) is handled by a separate router within each AS (for instance, AS4 peers with AS2 and AS3 on separate routers).

When the router in AS4 peering with AS2 receives both the 2001:DB8:1:/48 and 2001:DB8::/32 prefixes, it will mark 2001:DB8:1:/48 as BOUNDED, and set a Cost Community (as described in Section 3.2.1) so the marked overlapping route is preferred over unmarked routes within AS4.

The border router between AS4 and AS3 will receive the longer prefix from AS3, and the preferred BOUNDED overlapping route through iBGP. It will prefer the marked route, so the unmarked route towards 2001:DB8:1:/48 will not be advertised throughout AS4.

If the link between AS1 and AS2 fails, the longer length prefix will be withdrawn from AS2, and thus the peering point between AS2 and AS4 will no longer have an overlapping set of prefixes. Within AS4, the border router which peers with AS2 will cease advertising the 2001:DB8:1:/48 prefix, which allows the AS3/AS4 border router to begin advertising it into AS4, and through AS4 into AS5, restoring connectivity to AS1.

5. Operational Considerations

The intent of the mechanism described in this document is for it to be used based on local policies, not on an Internet-wide fashion. It is assumed that network operators using this mechanism have an incentive to do so.

The practice of filtering exists today on the Internet. While there may be local benefits to applying manual filters and/or the mechanism

specified in this document, the operator should be aware of the impact it may have on neighboring autonomous systems' policies [I-D.cardona-filtering-threats].

The benefits and implications associated with this proposal are discussed in the sections below. The text references the sample network in Section 4.

5.1. Advantages to the Service Provider

AS4, in each of the situations, reduces the number of prefixes advertised to transit peering autonomous systems by the number of longer prefixes that overlap with aggregates of those prefixes, so that AS5 receives fewer total routes, and a more stable routing table. While one copy of the prefix continues to be carried through the autonomous system, this entry can be removed from the local forwarding table.

5.2. Implications for Router processing

This proposal requires a BGP speaker to perform an additional check on receiving a route, checking the route against existing routes for overlapping coverage of a set of reachable destinations. This additional work, in terms of processing requirements, should be easily offset by the overall savings in processing through the reduction of the forwarding table size, and the additional stability in the routing table due to the removal of longer length prefixes.

5.3. Implications for Convergence Time

If the route to the AS providing the route to the covering route should be lost, the overlapping route must now propagate into the autonomous systems which had formerly received only the covering route. This behavior increases convergence time and may create situations in which reachability is temporarily compromised. Unlike the case where manual filters are used, normal BGP behavior should restore reachability without changes to the router configuration.

6. Security Considerations

This document presents a mechanism for an autonomous system to mark and filter overlapping prefixes. Note that the result of this operation is akin to the implementation of local route filtering at an AS boundary. As such, this document doesn't introduce any new security risks.

7. IANA Considerations

This document has no IANA actions.

8. Acknowledgements

Cengiz Alaentinoglu, Daniel Walton, David Ball, Ted Hardie, Jeff Hass, Barry Greene, Bill Herrin and Robert Raszuk gave valuable comments on this document.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

9.2. Informative References

[I-D.cardona-filtering-threats]
Cardona, C. and P. Francois, "Making BGP filtering a habit: Impact on policies", draft-cardona-filtering-threats-02 (work in progress), July 2013.

[I-D.ietf-idr-custom-decision]
Retana, A. and R. White, "BGP Custom Decision Process", draft-ietf-idr-custom-decision-04 (work in progress), November 2013.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

Appendix A. Change Log

A.1. Changes between the -00 and -01 versions.

- Updated authors' contact information.
- Changed intended status to Informational.
- General editorial changes.
- Clarified the intent of the draft in several places.
- Clarified when a route should be marked (3.1).
- Edited the operational considerations section.

- Updated ACKs.

A.2. Changes between the -01 and -02 versions

- Updated authors' contact information.
- General editorial changes.
- Refined the text about marking routes.

A.3. Changes between the -02 and -03 versions

- Updated authors' contact information.
- Added IPv6 examples.
- Minor editorial changes.

A.4. Changes between the -03 and -04 versions

- Updated authors' contact information.

Authors' Addresses

Russ White
Linkedin

Email: russ@riw.us

Alvaro Retana
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709
USA

Email: aretana@cisco.com

Susan Hares
Huawei

Email: shares@ndzh.com