

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: February 9, 2011

K. Patel
C. Appanna
P. Mohapatra
Cisco Systems
J. Scudder
Juniper Networks
J. Uttaro
AT&T
August 8, 2010

Root cause notification to solve BGP path hunting
draft-keyupate-bgp-rcn-00.txt

Abstract

Whenever a prefix is withdrawn using BGP withdrawal mechanism, it triggers a number of updates in certain scenarios before the prefix is completely withdrawn from the entire BGP network. This phenomenon is popularly known as `_path exploration_` or `_path hunting_` and occurs because of path vector property of BGP. It results in a series of unwanted or redundant transitions that overloads the BGP network.

This document describes a mechanism to help limit the amount of such path exploration by defining two optional transitive path attributes for BGP: `SPEAKERID_PATH` and `ROOT_CAUSE`.

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 February 9, 2011.

Copyright Notice

Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.

Internet-Draft

Root Cause Notification

August 2010

This document is subject to [BCP 78](http://trustee.ietf.org/license-info) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Internet-Draft

Root Cause Notification

August 2010

Table of Contents

1.	Introduction	4
1.1.	Requirements Language	5
2.	Reference Diagram	5
3.	SPEAKERID_PATH attribute	6
4.	ROOT_CAUSE attribute	8
5.	Operation	9
5.1.	Sending SPEAKERID_PATH attribute	9
5.2.	Sending ROOT_CAUSE attribute	9
5.2.1.	At the point of occurrence	9
5.2.2.	At an intermediate point	9
5.3.	Receiving ROOT_CAUSE Attribute	10
5.4.	Usage of BGP Aggregates	10
5.5.	BGP Confederation	10
5.6.	BGP Inactive Timer	10
6.	Acknowledgements	11
7.	IANA Considerations	11
8.	Security Considerations	11
9.	References	11
9.1.	Normative References	11
9.2.	Informative References	11
	Authors' Addresses	11

1. Introduction

Whenever a prefix is withdrawn using BGP withdrawal mechanism, it triggers a number of updates in certain scenarios before the prefix is completely withdrawn from the entire BGP network. This phenomenon is popularly known as `_path exploration_` or `_path hunting_` and occurs because of path vector property of BGP. It results in a series of unwanted or redundant transitions that overloads the BGP network ([\[I-D.li-bgp-stability\]](#)).

It is interesting to note that these redundant transitions can end up triggering route dampening ([\[RFC2439\]](#), if deployed in the network. Additionally, route dampening itself is known to cause path exploration in the network due to the delay it introduces ([\[I-D.li-bgp-stability\]](#)). This effectively creates a spiral effect on BGP instability. Both the generation of unwanted update messages and the triggering of route dampening can adversely affect the BGP convergence time.

The problem lies in the way BGP path vector is defined. With a link state protocol, each router stores a complete view of the entire network and derives reachability information from that view. In the event of a flap, each router can correctly determine all paths that suffer from the same root cause. This is not scalable in large networks in which BGP operates. By design, BGP advertises only the path it is using in terms of ASes to its neighbors with each prefix. Unfortunately, this information is coarse even in a simple topology as the number of possible paths through the routers is quite large. When a route is not reachable, because the detail route information

is not included, BGP selection process may end up choosing an alternative path that is actually not available. After sets of such transitions, BGP speaker will resolve this abnormality and decide on correct available path based on receiver side loop detection.

This document proposes a mechanism to identify unreachable paths for which BGP withdrawals are not received and prevent them from being selected as preferred paths. This helps avoid unnecessary route flapping within the network. A new optional transitive path attribute, `SPEAKERID_PATH` is tagged in BGP announcements as the prefix travels through the network, essentially creating more granular information about routers in the path. When a prefix is withdrawn, another optional transitive attribute, `ROOT_CAUSE` is attached to the implicit or explicit withdrawals that are generated at different points in the network. This attribute is created once at the point of occurrence of the fault and gets attached to the resulting UPDATE message throughout the network unchanged. At a receiving speaker, the `ROOT_CAUSE` attribute is matched against the `SPEAKERID_PATH` attributes of available paths to help identify and

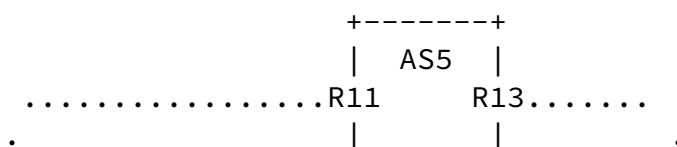
avoid those that are unreachable since they are affected by the same root cause.

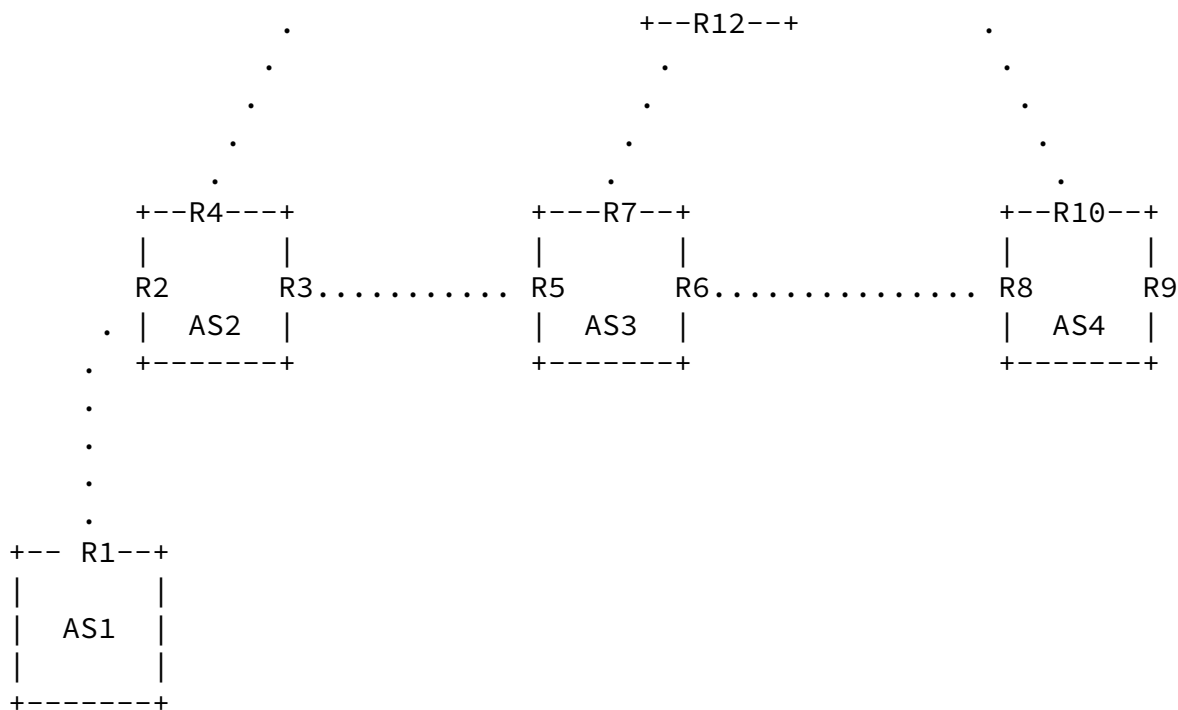
Path exploration caused by new prefix advertisements is not discussed in this document.

[1.1](#). 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 [RFC 2119](#) [[RFC2119](#)].

[2](#). Reference Diagram





The figure above describes a topology that leads to classic path hunting problem. In steady state, AS5 has 3 paths for prefixes received from AS1:

Path	AS_PATH
p1(best)	2 1
p2	3 2 1
p3	4 3 2 1

When the link between AS1 and AS2 goes down, it leads to a series of events and actions at AS5 as follows:

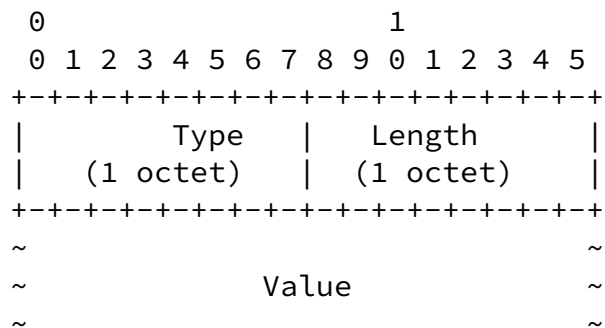
Step	Event	Action
1	Recv withdraw of p1	Select p2 as best

		Send AS_PATH (5 3 2 1) upstream
--	--	--
2	Recv withdraw of p2	Select p3 as best
		Send AS_PATH (5 4 3 2 1) upstream
--	--	--
3	Recv withdraw of p3	Prefixes have no path
		Send withdraw for the prefixes upstream

This trivial example creates unnecessary churn in the network till the end state is reached.

3. SPEAKERID_PATH attribute

SPEAKERID_PATH is an optional transitive attribute that is very similar in encoding and operation to the AS_PATH attribute. It is composed of a sequence of SPEAKERID path segments. Each segment is represented by a triple (type, length, value). Following is the format:



```
+---+---+---+---+---+---+---+---+
```

The type is a 1-octet field with the following value defined:

Value Type definition

- 1 AS_ID_SEQUENCE: ordered set of AS and Speaker ID pair
a route in the UPDATE message has traversed.

The length is a 1-octet field, containing the number of such pairs. Thus when the type is 1, the value contains one or more entries of the following:

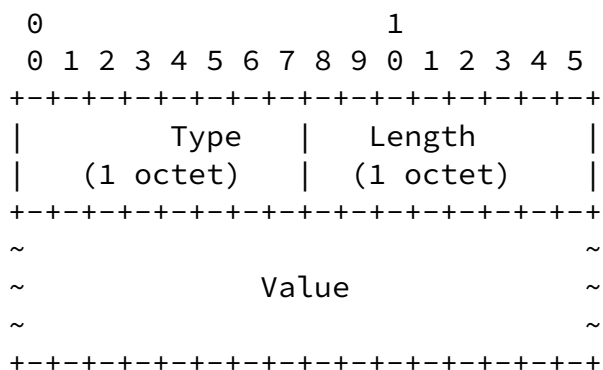
```
+-----+
|  AS      (4 bytes)  |
+-----+
|  SPEAKER-ID (4 bytes)  |
+-----+
```

The use and meaning of these fields are as follows:

AS: The AS is a four-octet field that indicates the AS number of the BGP speaker. If this ASN is from the public ASN space, it must have been assigned by the appropriate authority (use of ASN values from the private ASN space is strongly discouraged). Note that when a four-octet AS supporting speaker (NEW) announces an UPDATE to a two-octet AS supporting speaker (OLD), it encodes AS_TRANS as a two-octet AS in the AS_PATH attribute instead of its own AS ([[I-D.ietf-idr-rfc4893bis](#)]). But while encoding the SPEAKERID_PATH attribute, it MUST put its own four-octet AS in this field regardless of whether the neighbor to whom the UPDATE message is being sent is an OLD or NEW speaker.

SPEAKER-ID: The SPEAKER-ID is a four-octet field that indicates the router-id of the BGP speaker. If the router-id is from the public address space, it must have been assigned by the appropriate authority. (use of the private ip address as a router-id is strongly discouraged).

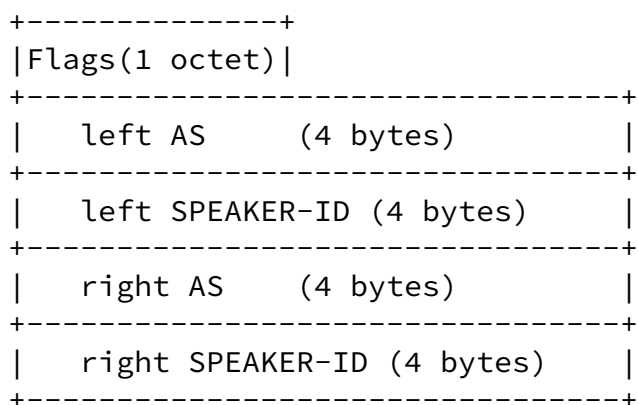
ROOT_CAUSE is an optional transitive attribute that is composed of one or more triple (type, length, value). Following is the format:



The type is a 1-octet field with the following value defined:

Value	Type definition
1	AS_ID_CONN: AS and router-ID pairs from both sides of the connection that is the point of occurrence for the withdraw.

The length is a 1-octet field, containing the length in octets of the value field. When the type is 1, the value contains the following:



[5. Operation](#)

[5.1. Sending SPEAKERID_PATH attribute](#)

When a BGP speaker supporting the mechanism described in this document propagates a route it learned from another BGP speaker's UPDATE message, it modifies the route's SPEAKERID_PATH attribute by prepending its own router-ID and AS number as the last pair of the sequence. If there is no such attribute, the local system creates the attribute, creates a new segment in the attribute of type AS_ID_SEQUENCE and places its own pair into that segment. If the act of prepending will cause an overflow in the existing segment (i.e. more than 255 pairs), it MUST prepend a new segment of type AS_ID_SEQUENCE and prepend its own pair to this new segment. This operation should be performed regardless of whether the peer is IBGP or EBGP.

[5.2. Sending ROOT_CAUSE attribute](#)

[5.2.1. At the point of occurrence](#)

A BGP speaker originates the ROOT_CAUSE attribute into an UPDATE message in one of the following scenarios:

- o A session with a peer AS goes down or the associated link goes down and the received prefixes need to be withdrawn or their bestpath changes.
- o it receives withdraws for some prefixes without the ROOT_CAUSE attribute and they in turn need to be either withdrawn from the ASes upstream or re-advertised with new paths.

While originating the attribute, the speaker encodes the router-ID and AS of each side of the session.

[5.2.2. At an intermediate point](#)

Any speaker receiving a withdrawal UPDATE message with ROOT_CAUSE attribute should preserve and announce the resulting UPDATE message with the same attribute value. This can be an explicit withdraw for a prefix or an implicit withdraw.

Any speaker receiving a reachable UPDATE message with ROOT_CAUSE attribute should preserve the attribute and not announce the attribute in resulting UPDATE message unless the resulting UPDATE message is an explicit withdrawal message.

[5.3.](#) Receiving ROOT_CAUSE Attribute

Whenever a BGP speaker receives an update message to process withdrawn prefixes, it does the following:

- o Remove the BGP path of the prefix withdrawn.
- o Find all the other paths that have matching ROOT_CAUSE information to the one present in path that is removed. Place these paths on an Inactive timer for an Inactive time interval. Do not select these paths for the BGP bespath selection.

[5.4.](#) Usage of BGP Aggregates

Whenever a BGP speaker creates an aggregate route from more specific routes, it will not inherit any BGP SPEAKERID_PATH information from its more specific routes used for aggregation. Instead, it will create its own SPEAKERID_PATH attribute when it announces the aggregate route to its BGP peers, i.e. the attribute will contain one segment with only its own (AS, router-id) pair when it announces the aggregate.

[5.5.](#) BGP Confederation

BGP Confederation Speaker peering with EBGP peers and receiving routes from them will exchange BGP Route Originator attributes as well. Whenever a Special Withdrawal message is received, following is done:

- o Remove the path announced by peer (sending a Special Withdrawal message).
- o Not select any other BGP Paths with matching Route Originator Attribute (as one received in the Special Withdrawal).
- o If there arent any alternate paths available, forward the Special Withdrawal message (with originate Route Originator Attribute).

[5.6.](#) BGP Inactive Timer

BGP inactive timer is used for suppressing path information from being used in BGP bestpath selection. This prevents BGP from selecting such alternate paths for which withdrawals are not received yet. A BGP speaker should remove suppress paths whenever withdrawn. A BGP speaker must subject all the suppress paths for BGP bestpath selection if they are not withdrawn even after inactive timer expires. The timeout for an Inactive Timer should be kept big enough to allow the withdrawal information to propagate across the AS.

Patel, et al.

Expires February 9, 2011

[Page 10]

Internet-Draft

Root Cause Notification

August 2010

[6.](#) Acknowledgements

Authors would like to thank Robert Raszuk and Pedro Marques for their input.

[7.](#) IANA Considerations

IANA shall assign codepoints for the SPEAKERID_PATH and ROOT_CAUSE attributes. These codepoints will come from the "BGP Path Attributes" registry.

[8.](#) Security Considerations

This extension to BGP does not change the underlying security issues.

[9.](#) References

[9.1.](#) Normative References

[I-D.ietf-idr-rfc4893bis]

Vohra, Q. and E. Chen, "BGP Support for Four-octet AS Number Space", [draft-ietf-idr-rfc4893bis-01](#) (work in progress), October 2009.

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

[RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping", [RFC 2439](#), November 1998.

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

[9.2](#). Informative References

[I-D.li-bgp-stability]
Huston, G. and T. Li, "BGP Stability Improvements",
[draft-li-bgp-stability-01](#) (work in progress), June 2007.

Patel, et al.	Expires February 9, 2011	[Page 11]
---------------	--------------------------	-----------

Internet-Draft	Root Cause Notification	August 2010
----------------	-------------------------	-------------

Authors' Addresses

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: keyupate@cisco.com

Chandra Appanna
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: chandra@cisco.com

Pradosh Mohapatra
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: pmohapat@cisco.com

John Scudder
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale, CA 94089
USA

Email: jgs@juniper.net

James Uttaro
AT&T
200 S. Laurel Ave
Middletown, NJ 07748
USA

Email: uttaro@att.com