

Network Working Group
Internet Draft
Updates: 4271 (if approved)
Intended Status: Standards Track
Expiration Date: March 9, 2011

E. Chen
P. Mohapatra
K. Patel
Cisco Systems
September 8, 2010

Revised Error Handling for BGP Updates from External Neighbors
draft-chen-ebgp-error-handling-00.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on March 9, 2011.

Copyright Notice

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

This document is subject to BCP 78 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.

Abstract

In this document we partially revise the error handling of an UPDATE message from an external BGP neighbor. The essence of the revision is to avoid resetting an external BGP session by using the "treat-as-withdraw" approach when the whole NLRI field of a malformed UPDATE message can be parsed.

1. Introduction

The base BGP specification [RFC4271] requires that a BGP session be reset when an UPDATE message containing a malformed attribute is received. This behavior is undesirable in the case of optional transitive attributes as has been discussed and revised in [OPT-TRANS].

However, there are other situations where the behavior is also undesirable, but are outside the scope of [OPT-TRANS]. For example, there have been a few occurrences in the field where the AS-PATH attribute is malformed for a small number of routes. Resetting the BGP session would impact all the other valid routes in these cases.

Our goal is to minimize the scope of the network that is affected by a malformed UPDATE message, and also to limit the impact to only the routes involved. The constrain is that the protocol correctness must not be violated.

In this document we partially revise the error handling of an UPDATE message from an external BGP neighbor. The essence of the revision is to avoid resetting an external BGP session by using the "treat-as-withdraw" approach specified in [OPT-TRANS] when the whole NLRI field of a malformed UPDATE message can be parsed.

1.1. Specification of Requirements

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].

2. Revision to Base Specification

The revised error handling specified in this section is applicable only for processing an UPDATE message from an external BGP neighbor.

The error handling of the following case described in Section 6.3 of [RFC4271] remains unchanged:

If the Withdrawn Routes Length or Total Attribute Length is too large (i.e., if Withdrawn Routes Length + Total Attribute Length + 23 exceeds the message Length), then the Error Subcode MUST be set to Malformed Attribute List.

The error handling of all other cases described in Section 6.3 of [RFC4271] that specify a session reset is conditionally revised as follows.

If a path attribute in an UPDATE message from an external BGP neighbor is determined to be malformed, the message containing that attribute SHOULD be treated as though all contained routes had been withdrawn ("treat-as-withdraw") when the whole NLRI field in the message can be parsed.

One exception is that the "attribute discard" approach [OPT-TRANS] SHOULD be used to handle a malformed optional transitive attribute for which the "attribute discard" approach is specified.

A BGP speaker MUST provide debugging facilities to permit issues caused by malformed UPDATE messages to be diagnosed. At a minimum, such facilities SHOULD include logging an error when such an attribute is detected. The malformed UPDATE message SHOULD be analyzed, and the root cause SHOULD be investigated.

3. Parsing of NLRI Fields

As described in [OPT-TRANS], we observe that in order to use the "treat-as-withdraw" approach for a malformed UPDATE, the NLRI field and/or MP_REACH and MP_UNREACH [RFC4760] attributes need to be successfully parsed. If this were not possible, the UPDATE would necessarily be malformed in some other way beyond the scope of this document and therefore, the procedures of [RFC4271] would continue to apply.

To facilitate the determination of the NLRI field in an UPDATE with malformed attributes, we strongly RECOMMEND that the MP_REACH or MP_UNREACH attribute (if present) be encoded as the very first path attribute in an UPDATE.

Traditionally the NLRIs for the IPv4 unicast address family are carried immediately following all the attributes in an UPDATE [RFC4271]. When such an UPDATE is received, we observe that the NLRI field can be determined using the "Message Length" and the "Total Attribute Length" (when they are consistent) carried in the message instead of relying on the length of individual attributes in the message.

Furthermore, it is observed that the NLRIs for the IPv4 unicast address family can also be carried in the MP_REACH attribute of an UPDATE when the IPV4 unicast address family capability is shared (i.e., both advertised and received) over a BGP session. For the same sake of better debugging and fault handling, we also RECOMMEND that the MP_REACH attribute be used and be placed as the very first path attribute in an UPDATE in this case.

4. Discussion

As discussed in [OPT-TRANS], the approach of "treat-as-withdraw" is not always safe to use. In the case of internal BGP sessions, the resolution of recursive nexthops can result in forwarding loops and blakholes when the BGP speakers inside a network have inconsistent routing information.

Depending on the network topology, the routing table, routes involved, and whether "tunnels" are used inside a network, the approach of "treat-as-withdraw" may work for internal BGP sessions only in some specific cases. Thus it may be deployed for internal BGP sessions only as a temporary measure to stop continuous session flaps due to malformed UPDATE messages. Such deployment must be carefully evaluated on a case-by-case basis.

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

TBD

7. Acknowledgments

We would like to thank Robert Raszuk, Naiming Shen and Tony Li for their review and discussions.

8. References

8.1. Normative References

- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

8.2. Informative References

- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [OPT-TRANS] Scudder, J. and E. Chen, "Error Handling for Optional Transitive BGP Attributes", Work in Progress, March 2010.

9. Authors' Addresses

Enke Chen
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134

EMail: enkechen@cisco.com

Pradosh Mohapatra
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134

EMail: pmohapat@cisco.com

Keyur Patel
Cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134

EMail: keyupate@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 20, 2011

B. Decraene
France Telecom - Orange
P. Francois
UCL
October 17, 2010

Reserved BGP extended communities
draft-decraene-idr-reserved-extended-communities-00

Abstract

This document assigns two BGP extended community types, one transitive and one non-transitive. It also defines two IANA registries in order to allow the allocation of reserved transitive and non-transitive extended communities. These are similar to the existing reserved (formerly Well-known) BGP communities defined in RFC 1997 but provides an easier control of inter-AS community advertisement as a community could be chosen as transitive or non-transitive across ASes.

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].

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 April 20, 2011.

Copyright Notice

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

This document is subject to BCP 78 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.

1. Introduction

RFC1997 [RFC1997] defines the BGP community attribute and some BGP Well known communities whose meaning SHALL be understood by all implementations compliant with RFC1997 [RFC1997]). New reserved communities can be registered in the IANA "BGP Well-known Communities" registry but can't anymore be considered as well known. Implementations which do not recognize those new reserved communities will propagate them from BGP neighbour to BGP neighbour and from AS to AS with an unlimited scope.

RFC 4360 [RFC4360] defines the BGP extended community attribute with a structure including a type and a transitive bit "T". The transitive bit, when set, allows to restrict the scope of the community within an AS. Without structure, this can only be accomplished by explicitly enumerating all community values that will be denied or allowed and passed to BGP speakers in neighboring ASes. RFC 4360 [RFC4360] defines IANA registries to allocate BGP Extended Communities types. Each type is able to encode 2^{48} or 2^{56} values depending on the type being extended or regular. It does not define an IANA registry to allocate single reserved communities. Therefore, one needing to reserve a single non-transitive extended community would need to reserve an extended subtype which represents 2^{48} communities. This would both waste the resources and disable the ability to define global policies on reserved communities, such as to filter them out.

This document assigns two BGP extended community types, one transitive and one non-transitive. It also defines two IANA registries in order to allow the allocation of reserved transitive and non-transitive extended communities. These are similar to the existing reserved ("Well-known") BGP communities defined in RFC 1997 but provides an easier control of inter-AS community advertisement as a community could be chosen as transitive or non-transitive across ASes.

2. IANA Considerations

IANA is requested to assign, from the registry "BGP Extended Communities Type - extended, transitive type", a type value TBD for "BGP Reserved transitive extended communities":

Registry Name: BGP Extended Communities Type - extended, transitive

Name	Type Value
----	-----
BGP Reserved transitive extended communities	TBD (e.g. 0x9000)

IANA is requested to assign, from the registry "BGP Extended Communities Type - extended, non-transitive", a type value TBD for "BGP Reserved non-transitive extended communities":

Registry Name: BGP Extended Communities Type - extended, non-transitive

Name	Type Value
----	-----
BGP Reserved non-transitive extended communities	TBD (e.g. 0xd000)

Note to the IANA: suggested value for the two reserved BGP Extended Communities extended type are 0x9000 and 0xd000. Otherwise, both values should be identical, except for their T - Transitive bit (bit 1 as defined in RFC 4360 [RFC4360]).

The IANA is requested to create and maintain a registry entitled "BGP Reserved transitive extended communities".

Registry Name: BGP Reserved transitive extended communities

Range	Registration Procedures
-----	-----
0x000000000000-FFFFFFFFFFFF	Reserved
0xFFFFFFFF0000-00FFFFFF8000	First Come First Served
0x00FFFFFF8001-FFFFFFFFFFFF	Standards Action/Early IANA Allocation

The IANA is requested to create and maintain a registry entitled "BGP Reserved non-transitive extended communities".

Registry Name: BGP Reserved non-transitive extended communities

Range	Registration Procedures
-----	-----
0x000000000000-FFFFFFFFFFFF	Reserved
0xFFFFFFFF0000-00FFFFFF8000	First Come First Served
0x00FFFFFF8001-FFFFFFFFFFFF	Standards Action/Early IANA Allocation

An application may need both a transitive and non-transitive reserved community. It may be beneficial to have the same value for both communities. (Note that both extended community will still be different as they will differ from their T bit). The IANA SHOULD try to accomodate such request to have both a transitive and non-transitive reserved community with the same value for both.

3. Security Considerations

This document defines IANA actions. In itself, it has no impact on the security of the BGP protocol.

4. Normative References

- [RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

Authors' Addresses

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

Email: bruno.decraene@orange-ftgroup.com

Pierre Francois
UCL
Place Ste Barbe, 2
Louvain-la-Neuve 1348
BE

Email: francois@info.ucl.ac.be

GROW Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 29, 2011

E. Jasinska
Limelight Networks
N. Hilliard
INEX
R. Raszuk
Cisco Systems
N. Bakker
AMS-IX B.V.
October 26, 2010

Internet Exchange Route Server
draft-jasinska-ix-bgp-route-server-01

Abstract

The growing popularity of Internet exchange points (IXPs) brings a new set of requirements to interconnect participating networks. While bilateral exterior BGP sessions between exchange participants were previously the most common means of exchanging reachability information, the overhead associated with dense interconnection has caused substantial operational scaling problems for Internet exchange point participants.

This document outlines a specification for multilateral interconnections at IXPs. Multilateral interconnection is a method of exchanging routing information between three or more BGP speakers using a single intermediate broker system, referred to as a route server. Route servers are typically used on shared access media networks such as Internet exchange points (IXPs), to facilitate simplified interconnection between multiple Internet routers on such a network.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>.

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>.

This Internet-Draft will expire on April 29, 2011.

Copyright Notice

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

This document is subject to BCP 78 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 BSD License.

Table of Contents

1. Introduction to Multilateral Interconnection	4
1.1. Specification of Requirements	5
2. Bilateral Interconnection	5
3. Multilateral Interconnection	6
4. Technical Considerations for Route Server Implementations	7
4.1. Client UPDATE Messages	7
4.2. Attribute Transparency	7
4.2.1. NEXT_HOP Attribute	8
4.2.2. AS_PATH Attribute	8
4.2.3. MULTI_EXIT_DISC Attribute	8
4.2.4. Communities Attributes	8
4.3. Per-Client Prefix Filtering	9
4.3.1. Prefix Hiding on a Route Server	9
4.3.2. Mitigation Techniques	10
4.3.2.1. Multiple Route Server RIBs	10
4.3.2.2. Advertising Multiple Paths	10
5. Operational Considerations for Route Server Installations	12
5.1. Route Server Scaling	12
5.1.1. Tackling Scaling Issues	12
5.1.1.1. View Merging and Decomposition	12
5.1.1.2. Destination Splitting	13
5.1.1.3. NEXT_HOP Resolution	13
5.2. NLRI Leakage Mitigation	13
5.3. Route Server Redundancy	13
5.4. AS_PATH Consistency Check	14
5.5. Implementing Routing Policies	14
5.5.1. Communities	14
5.5.2. Internet Routing Registry	14
6. Security Considerations	15
7. IANA Considerations	15
8. Acknowledgments	15
9. References	15
9.1. Normative References	15
9.2. Informative References	16
Authors' Addresses	17

1. Introduction to Multilateral Interconnection

Internet exchange points (IXPs) provide IP data interconnection facilities for their participants, typically using shared Layer-2 networking media such as Ethernet. The Border Gateway Protocol (BGP) [RFC4271], an inter-Autonomous System routing protocol, is commonly used to facilitate exchange of network reachability information over such media.

In the case of bilateral interconnection between two exchange participant routers, each router must be configured with a BGP session to the other. At IXPs with many participants who wish to implement dense interconnection, this requirement can lead both to large router configurations and high administrative overhead. Given the growth in the number of participants at many IXPs, it has become operationally troublesome to implement densely meshed interconnections at these IXPs.

Multilateral interconnection is a method of interconnecting BGP speaking routers using a third party brokering system, commonly referred to as a route server and typically managed by the IXP operator. Each of the multilateral interconnection participants (usually referred to as route server clients) announces network reachability information to the route server using exterior BGP, and the route server in turn forwards this information to each other route server client connected to it, according to its configuration. Although a route server uses BGP to exchange reachability information with each of its clients, it does not forward traffic itself and is therefore not a router.

A route server can be viewed as similar in function to an [RFC4456] route reflector, except that it operates using EBGp instead of iBGP. Certain adaptations to [RFC4271] are required, to enable an EBGp router to operate as a route server, which are outlined in Section 4 of this document. Operational considerations to be taken into account in a route server deployment are subject of Section 5.

The term "route server" is often in a different context used to describe a BGP node whose purpose is to accept BGP feeds from multiple clients for the purpose of operational analysis and troubleshooting. A system of this form may alternatively be known as a "route collector" or a "route-views server". This document uses the term "route server" exclusively to describe multilateral peering brokerage systems.

1.1. Specification of Requirements

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 [RFC2119].

2. Bilateral Interconnection

Bilateral interconnection is a method of interconnecting routers using individual BGP sessions between each participant router on an IXP in order to exchange reachability information. While interconnection policies vary from participant to participant, most IXPs have significant numbers of participants who see value in interconnecting with as many other exchange participants as possible. In order for an IXP participant to implement a dense interconnection policy, it is necessary for the participant to liaise with each of their intended interconnection partners and if this partner agrees to interconnect, then both participants' routers must be configured with a BGP session to exchange network reachability information. If each exchange participant interconnects with each other participant, a full mesh of BGP sessions is needed, as detailed in Figure 1.

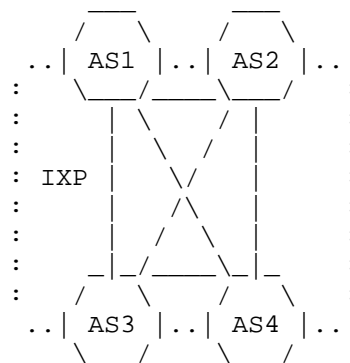


Figure 1: Full-Mesh Interconnection at an IXP

Figure 1 depicts an IXP platform with four connected routers, administered by four separate exchange participants, each of them with a locally unique autonomous system number: AS1, AS2, AS3 and AS4. Each of these four participants wishes to exchange traffic with all other participants; this is accomplished by configuring a full mesh of BGP sessions on each router connected to the exchange, resulting in 6 BGP sessions across the IXP fabric.

The number of BGP sessions at an exchange has an upper bound of $n*(n-1)/2$, where n is the number of routers at the exchange. As many exchanges have relatively large numbers of participating networks, the quadratic scaling requirements of dense interconnection tend to cause operational and administrative overhead at large IXPs. Consequently, new participants to an IXP require significant initial resourcing in order to gain value from their IXP connection, while existing exchange participants need to commit ongoing resources in order to benefit from interconnecting with these new participants.

3. Multilateral Interconnection

Multilateral interconnection is implemented using a route server configured to use BGP to distribute network layer reachability information (NLRI) among all client routers. The route server preserves the BGP NEXT_HOP attribute from all received NLRI UPDATE messages, and passes these messages with unchanged NEXT_HOP to its route server clients, according to its configured routing policy. Using this method of exchanging NLRI messages, an IXP participant router can receive an aggregated list of prefixes from all other route server clients using a single BGP session to the route server instead of depending on BGP sessions with each other router at the exchange. This reduces the overall number of BGP sessions at an Internet exchange from $n*(n-1)/2$ to n , where n is the number of routers at the exchange.

In practical terms, this allows dense interconnection between IXP participants with low administrative overhead and significantly simpler and smaller router configurations. In particular, new IXP participants benefit from immediate and extensive interconnection, while existing route server participants receive reachability information from these new participants without necessarily having to adapt their configurations.

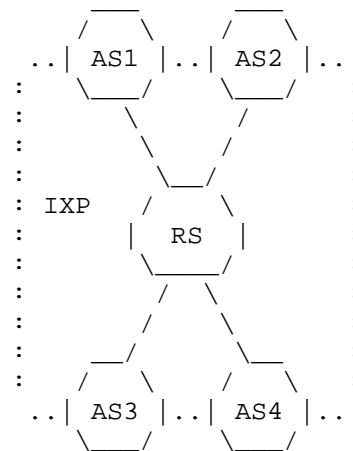


Figure 2: IXP-based Interconnection with Route Server

As illustrated in Figure 2, each router on the IXP fabric requires only a single BGP session to the route server, from which it can receive reachability information for all other routers on the IXP which also connect to the route server.

4. Technical Considerations for Route Server Implementations

4.1. Client UPDATE Messages

A route server **MUST** accept all UPDATE messages received from each of its clients for inclusion in its Adj-RIB-In. These UPDATE messages **MAY** be omitted from the route server's Loc-RIB or Loc-RIBs, due to filters configured for the purposes of implementing routing policy. The route server **SHOULD** perform one or more BGP Decision Processes to select routes for subsequent advertisement to its clients, taking into account possible configuration to provide multiple NLRI paths to a particular client as described in Section 4.3.2.2 or multiple Loc-RIBs as described in Section 4.3.2.1. The route server **SHOULD** forward UPDATE messages where appropriate from its Loc-RIB or Loc-RIBs to its clients.

4.2. Attribute Transparency

As a route server primarily performs a brokering service, modification of attributes could cause route server clients to alter their BGP best-path selection process for received prefix reachability information, thereby changing the intended routing policies of exchange participants. Therefore, contrary to what is

specified in section 5. of [RFC4271], route servers SHOULD NOT update well-known BGP attributes received from route server clients before redistributing them to their other route server clients. Optional recognized and unrecognized BGP attributes, whether transitive or non-transitive, SHOULD NOT be updated by the route server and SHOULD be passed on to other route server clients.

4.2.1. NEXT_HOP Attribute

The NEXT_HOP, a well-known mandatory BGP attribute, defines the IP address of the router used as the next hop to the destinations listed in the Network Layer Reachability Information field of the UPDATE message. As the route server does not participate in the actual routing of traffic, the NEXT_HOP attribute MUST be passed unmodified to the route server clients, similar to the "third party" next hop feature described in section 5.1.3. of [RFC4271].

4.2.2. AS_PATH Attribute

AS_PATH is a well-known mandatory attribute which identifies the autonomous systems through which routing information carried in the UPDATE message has passed.

As a route server does not participate in the process of forwarding data between client routers, and because modification of the AS_PATH attribute could affect route server client best-path calculations, the route server SHOULD NOT prepend its own AS number to the AS_PATH segment nor modify the AS_PATH segment in any other way.

4.2.3. MULTI_EXIT_DISC Attribute

MULTI_EXIT_DISC is an optional non-transitive attribute intended to be used on external (inter-AS) links to discriminate among multiple exit or entry points to the same neighboring AS. If applied to an NLRI UPDATE sent to a route server, the attribute (contrary to section 5.1.4 of [RFC4271]) SHOULD be propagated to other route server clients and the route server SHOULD NOT modify the value of this attribute.

4.2.4. Communities Attributes

The BGP COMMUNITIES ([RFC1997]) and Extended Communities ([RFC4360]) attributes are attributes intended for labeling information carried in BGP UPDATE messages. Transitive as well as non-transitive Communities attributes applied to an NLRI UPDATE sent to a route server SHOULD NOT be modified, processed or removed. However, if such an attribute is intended for processing by the route server itself, it MAY be modified or removed.

4.3. Per-Client Prefix Filtering

4.3.1. Prefix Hiding on a Route Server

While IXP participants often use route servers with the intention of interconnecting with as many other route server participants as possible, there are several circumstances where control of prefix distribution on a per-client basis is important for ensuring that desired interconnection policies are met.

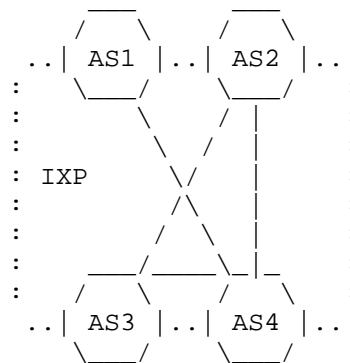


Figure 3: Filtered Interconnection at an IXP

Using the example in Figure 3, AS1 does not directly exchange prefix information with either AS2 or AS3 at the IXP, but only interconnects with AS4.

In the traditional bilateral interconnection model, prefix filtering to a third party exchange participant is accomplished either by not engaging in a bilateral interconnection with that participant or else by implementing outbound prefix filtering on the BGP session towards that participant. However, in a multilateral interconnection environment, only the route server can perform outbound prefix filtering in the direction of the route server client; route server clients depend on the route server to perform their filtering for them.

If the same prefix is sent to a route server from multiple route server clients with different BGP attributes, and traditional best-path route selection is performed on that list of prefixes, then the route server will select a single best-path prefix for propagation to all connected clients. If, however, the route server has been configured to filter the calculated best-path prefix from reaching a particular route server client, then that client will receive no

reachability information for that prefix from the route server, despite the fact that the route server has received alternative reachability information for that prefix from other route server clients. This phenomenon is referred to as "prefix hiding".

For example, in Figure 3, if the same prefix were sent to the route server via AS2 and AS4, and the route via AS2 was preferred according to BGP's traditional best-path selection, but AS2 was filtered by AS1, then AS1 would never receive this prefix, even though the route server had previously received a valid alternative path via AS4. This happens because the best-path selection is performed only once on the route server for all clients.

It should be noted that prefix hiding will only occur on route servers which employ per-client prefix filtering; if an IXP operator deploys a route server without prefix filtering, then prefix hiding does not occur, as all paths are considered equally valid from the point of view of the route server.

There are several techniques which may be employed to prevent the prefix hiding problem from occurring. Route server implementations SHOULD implement at least one method to prevent prefix hiding.

4.3.2. Mitigation Techniques

4.3.2.1. Multiple Route Server RIBs

The most portable means of preventing the route server prefix hiding problem is by using a route server BGP implementation which performs the per-client best-path calculation for each set of prefixes which results after the route server's client filtering policies have been taken into consideration. This can be implemented by using per-client Loc-RIBs, with prefix filtering implemented between the Adj-RIB-In and the per-client Loc-RIB. Implementations MAY optimize this by maintaining prefixes not subject to filtering policies in a global Loc-RIB, with per-client Loc-RIBs stored as deltas.

This problem mitigation technique is highly portable, as it makes no assumptions about the feature capabilities of the route server clients.

4.3.2.2. Advertising Multiple Paths

The prefix distribution model described above assumes standard BGP session encoding where the route server sends a single path to its client for any given prefix. This path is selected using the BGP path selection decision process described in [RFC4271]. If, however, it were possible for the route server to send more than a single path

to a route server client, then route server clients would no longer depend on receiving a single best path to a particular prefix; consequently, the prefix hiding problem described in Section 4.3.1 would disappear.

We present two methods which describe how such increased path diversity could be implemented.

4.3.2.2.1. Diverse BGP Path Approach

The Diverse BGP Path proposal as defined in [I-D.ietf-grow-diverse-bgp-path-dist] is a simple way to distribute multiple prefix paths from a route server to a route server client by using a separate BGP session from the route server to a client for each different path.

The number of paths which may be distributed to a client is constrained by the number of BGP sessions which the server and the client are willing to establish with each other. The distributed paths may be established from the global BGP Loc-RIB on the route server in addition to any per-client Loc-RIB. As there may be more potential paths to a given prefix than configured BGP sessions, this method is not guaranteed to eliminate the prefix hiding problem in all situations. Furthermore, this method may significantly increase the number of BGP sessions handled by the route server, which may negatively impact its performance.

4.3.2.2.2. BGP ADD-PATH Approach

The [I-D.ietf-idr-add-paths] Internet draft proposes a different approach to multiple path propagation, by allowing a BGP speaker to forward multiple paths for the same prefix on a single BGP session. As [RFC4271] specifies that a BGP listener must implement an implicit withdraw when it receives an UPDATE message for a prefix which already exists in its Adj-RIB-In, this approach requires explicit support for the feature both on the route server and on its clients.

If the ADD-PATH capability is negotiated bidirectionally between the route server and a route server client, and the route server client propagates multiple paths for the same prefix to the route server, then this could potentially cause the propagation of inactive, invalid or suboptimal paths to the route server, thereby causing loss of reachability to other route server clients. For this reason, ADD-PATH implementations on a route server SHOULD enforce send-only mode with the route server clients, which would result in negotiating receive-only mode from the client to the route server.

5. Operational Considerations for Route Server Installations

5.1. Route Server Scaling

While deployment of multiple Loc-RIBs on the route server presents a simple way to avoid the prefix hiding problem noted in Section 4.3.1, this approach requires significantly more computing resources on the route server than where a single Loc-RIB is deployed for all clients. As the [RFC4271] Decision Process must be applied to all Loc-RIBs deployed on the route server, both CPU and memory requirements on the host computer scale approximately according to $O(P * N)$, where P is the total number of unique prefixes received by the route server and N is the number of route server clients which require a unique Loc-RIB. As this is a super-linear scaling relationship, large route servers may derive benefit from deploying per-client Loc-RIBs only where they are required.

Regardless of any Loc-RIB optimization implemented, the route server's control plane bandwidth requirements will scale according to $O(P * N)$, where P is the total number of unique prefixes received by the route server and N is the total number of route server clients. In the case where P_{avg} (the arithmetic mean number of unique prefixes received per route server client) remains roughly constant even as the number of connected clients increases, this relationship can be rewritten as $O((P_{avg} * N) * N)$ or $O(N^2)$. This quadratic upper bound on the network traffic requirements indicates that the route server model will not scale to arbitrarily large sizes.

5.1.1. Tackling Scaling Issues

The network traffic scaling issue presents significant difficulties with no clear solution - ultimately, each client must receive a UPDATE for each unique prefix received by the route server. However, there are several potential methods for dealing with the CPU and memory resource requirements of route servers.

5.1.1.1. View Merging and Decomposition

View merging and decomposition, outlined in [RS-ARCH], describes a method of optimising memory and CPU requirements where multiple route server clients are subject to exactly the same routing policies. In this situation, the multiple Loc-RIB views required by each client are merged into a single view.

A variation of this approach may be implemented on route servers by ensuring that separate Loc-RIBs are only configured for route server clients with unique export peering policies.

5.1.1.2. Destination Splitting

Destination splitting, also described in [RS-ARCH], describes a method for route server clients to connect to multiple route servers and to send non-overlapping sets of prefixes to each route server. As each route server computes the best path for its own set of prefixes, the quadratic scaling requirement operates on multiple smaller sets of prefixes. This reduces the overall computational and memory requirements for managing multiple Loc-RIBs and performing the best-path calculation on each. In order for this method to perform well, destination splitting would require significant co-ordination between the route server operator and each route server client. In practice, such levels of co-ordination are unlikely to work successfully, thereby diminishing the value of this approach.

5.1.1.3. NEXT_HOP Resolution

As route servers are usually deployed at IXPs which use flat layer 2 networks, recursive resolution of the NEXT_HOP attribute is generally not required, and can be replaced by a simple check to ensure that the NEXT_HOP value for each prefix is a network address on the IXP LAN's IP address range.

5.2. NLRI Leakage Mitigation

NLRI leakage occurs when a BGP client unintentionally distributes NLRI UPDATE messages to one or more neighboring BGP routers. NLRI leakage of this form to a route server can cause connectivity problems at an IXP if each route server client is configured to accept all prefix UPDATE messages from the route server. It is therefore RECOMMENDED when deploying route servers that, due to the potential for collateral damage caused by NLRI leakage, route server operators deploy NLRI leakage mitigation measures in order to prevent unintentional prefix announcements or else limit the scale of any such leak. Although not foolproof, per-client inbound prefix limits can restrict the damage caused by prefix leakage in many cases. Per-client inbound prefix filtering on the route server is a more deterministic and usually more reliable means of preventing prefix leakage, but requires more administrative resources to maintain properly.

5.3. Route Server Redundancy

As the purpose of an IXP route server implementation is to provide a reliable reachability brokerage service, it is RECOMMENDED that exchange operators who implement route server systems provision multiple route servers on each shared Layer-2 domain. There is no requirement to use the same BGP implementation or operating system

for each route server on the IXP fabric; however, it is RECOMMENDED that where an operator provisions more than a single server on the same shared Layer-2 domain, each route server implementation be configured equivalently and in such a manner that the path reachability information from each system is identical.

5.4. AS_PATH Consistency Check

As per [RFC4271] every BGP speaker who advertises a route to another external BGP speaker prepends its own AS number as the last element of the AS_PATH sequence. Therefore the leftmost AS in an AS_PATH attribute is equal to the autonomous system number of the BGP speaker that sent an UPDATE message.

[RFC4271] suggests in section 6.3 that a BGP speaker MAY check the AS_PATH attribute of each UPDATE message received for consistency, if the leftmost AS in the AS_PATH is in fact the one of the sender.

Route servers do not modify the AS_PATH attribute (as described in Section 4.2.2), since they do not participate in the traffic exchange. Therefore a consistency check on the AS_PATH of an UPDATE received by a route server client would fail. It is therefore RECOMMENDED that route server clients disable the AS_PATH consistency check towards the route server.

5.5. Implementing Routing Policies

Prefix filtering is commonly implemented on route servers to provide prefix distribution control mechanisms for route server clients. There are a few commonly used strategies available.

5.5.1. Communities

Prefixes sent to the route server are tagged with certain COMMUNITIES attributes agreed upon beforehand between the operator and all participants. Based on the values, routes are propagated to all other participants, a subset of participants, or none. This allows for one-way filtering policies to be implemented on the route server; if a participant chooses not to exchange routes with a certain other participant, he will have to instruct the route server to not announce his own routes and filter incoming routes on his own router.

5.5.2. Internet Routing Registry

Filters configured on the route server can be constructed by querying an Internet Routing Registry database for RPSL [RFC2622] objects placed there by participating operators. Import and export statements for the route server's ASN in an aut-num object define

their desired policy, from which the configured filters are derived.

6. Security Considerations

On route server installations which do not employ prefix-hiding mitigation techniques, the prefix hiding problem outlined in section Section 4.3.1 can be used in certain circumstances to proactively block third party prefix announcements from other route server clients.

7. IANA Considerations

The new set of mechanism for route servers does not require any new allocations from IANA.

8. Acknowledgments

The authors would like to thank Chris Hall, Ryan Bickhart and Steven Bakker for their valuable input.

In addition, the authors would like to acknowledge the developers of BIRD, OpenBGPD and Quagga, whose open source BGP implementations include route server capabilities which are compliant with this document.

9. References

9.1. Normative References

- [I-D.ietf-grow-diverse-bgp-path-dist]
Raszuk, R., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of diverse BGP paths.",
draft-ietf-grow-diverse-bgp-path-dist-02 (work in progress), July 2010.
- [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-04 (work in progress),
August 2010.
- [RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2622] Alaettinoglu, C., Villamizar, C., Gerich, E., Kessens, D., Meyer, D., Bates, T., Karrenberg, D., and M. Terpstra, "Routing Policy Specification Language (RPSL)", RFC 2622, June 1999.
- [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.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.
- [RS-ARCH] Govindan, R., Alaettinoglu, C., Varadhan, K., and D. Estrin, "A Route Server Architecture for Inter-Domain Routing", 1995, <<http://www.cs.usc.edu/research/95-603.ps.Z>>.

9.2. Informative References

- [RFC1863] Haskin, D., "A BGP/IDRP Route Server alternative to a full mesh routing", RFC 1863, October 1995.
- [RFC3418] Presuhn, R., "Management Information Base (MIB) for the Simple Network Management Protocol (SNMP)", STD 62, RFC 3418, December 2002.
- [RFC4223] Savola, P., "Reclassification of RFC 1863 to Historic", RFC 4223, October 2005.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, August 2007.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

Authors' Addresses

Elisa Jasinska
Limelight Networks
2220 W 14th St
Tempe, AZ 85281
US

Email: elisa@llnw.com

Nick Hilliard
INEX
4027 Kingswood Road
Dublin 24
IE

Email: nick@inex.ie

Robert Raszuk
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: raszuk@cisco.com

Niels Bakker
AMS-IX B.V.
Westeinde 12
Amsterdam, NH 1017 ZN
NL

Email: niels.bakker@ams-ix.net

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 20, 2011

R. Raszuk
E. Chen
Cisco Systems
B. Decraene
France Telecom
October 17, 2010

BGP Diagnostic Message
draft-raszuk-bgp-diagnostic-message-00

Abstract

BGP protocol lacks self diagnostic tools which would allow for monitoring and detection of any possible bgp state database differences between BGP_RIB_Out of the sender and BGP_RIB_In of the receiver over BGP peering session. It also lacks of build in mechanism to inform peer about subset of prefixes received over session which experienced some errors and which per protocol specification either resulted in attribute drop or "treat-as-withdraw" action.

The intention of this document is to start a new class of work which will make BGP protocol and therefor assuring services constructed with the help of BGP protocol to become much more reliable and robust.

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 April 20, 2011.

Copyright Notice

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

This document is subject to BCP 78 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.

Table of Contents

1. Introduction	3
2. Applications	3
3. BGP diagnostic message	4
3.1. BGP DIAGNOSTIC Message Encoding	4
3.2. BGP DIAGNOSTIC Message TLVs	5
3.2.1. Operational TLVs	6
3.2.2. BGP database counters exchange	9
3.2.3. Diagnostics for encoding errors in BGP messages	10
3.2.4. AFI/SAFI signaling when malformed update	12
3.2.5. Prefix specific BGP debugging	12
3.2.6. Intra-domain bgp decision monitoring	13
3.2.7. Exchange of installed Route Target filters	14
4. Operation	14
5. Capability negotiation	15
6. Security considerations	16
7. IANA Considerations	16
8. Acknowledgments	17
9. Normative References	17
Authors' Addresses	18

1. Introduction

In this document we will first define a new diagnostic communication channel in the form of new BGP message then construct the set of basic message encoding to be used for simple diagnostic self test routines periodically exchanged between BGP speakers. We will also define set of other TLVs which can be very useful in precise description of prefixes affected by various cases of BGP session malfunctions.

The goal of this document is to provide the background which will in turn allow for very easy extensibility once new needs and new BGP diagnostic ideas surface.

2. Applications

Authors would like to propose four main applications which BGP Diagnostic TLVs are designed to address. New TLVs can be easily added to enhance further current applications or to propose new applications.

The set of TLVs is organized in the following application groups:

General TLVs used for operational purposes of the described mechanism.

Set of TLVs designed to carry information about BGP state across BGP peers that include per neighbor counters and global counters. There are two modes this functionality can be used - on demand by explicit query as well as periodic in an automated mode. The scope of messages is to be able to operate both on the iBGP as well as eBGP boundaries. It is in the control of the operator to decide which set of information would be send to a given set of peers.

Messages which operate in an automated push mode (as long as peer negotiated listen capability for them) and are designed to inform BGP peer on the list of impacted NLRIs which were received along with malformed attribute or within malformed update message.

Following recommendation from MP-BGP4 RFC4760 next group of messages are used to indicate which AFI/SAFIs were disabled for any further processing by BGP peer due to detection of an incorrect attribute present in the BGP Update message.

In number of troubleshooting efforts in real networks it is often very helpful to verify state of a given prefix in the neighboring

router's BGP database. This is particularly useful on the EBGp boundaries where there is no CLI/SNMP access to the router. Authors define a new way of query peer's BGP for the state of particular prefix.

Last set of messages is an attempt to allow for intra-domain better analysis of the BGP best path selection tie break decisions.

3. BGP diagnostic message

When defining any self test tool the critical element is to find a right separation balance between the test object and testing instruments.

For the vast majority of real BGP issues found in the life production networks authors believe that the right balance is the definition of new BGP message which could be exchanged along with any negotiated AFI/SAFI between those BGP speakers which will during initial OPEN message exchange new BGP diagnostic message capability.

The two extreme alternatives which were considered were the definition of new BGP attribute which may inherit and share potential issues of given BGP address family it is designed to diagnose and on the other extreme to build a separate and independent network diagnostic protocol. The use of BGP message seems to provide sufficient isolation from any service address family and is much easier to deploy then enabling an entire new intra and inter-domain protocol. Another very important issue with using any other protocol for detection of potential differences of BGP databases state is lack of synchronization with BGP UPDATE messages. This alone in the continuously churning BGP environment would not allow for any benefit.

3.1. BGP DIAGNOSTIC Message Encoding

BGP message as defined in RFC 4271 consists of a fixed-size header followed by two octet length field and one octet of type value. RFC 4271 limits maximum message size to 4096 octets. As one of the applications of BGP Diagnostic message is to be able to carry entire potentially malformed BGP message this specification extends the maximum size of BGP Diagnostic message to be always 128 octets bigger then any other BGP Message. Considering the current RFC 4271 maximum BGP message size to be 4096 octets maximum size of BGP diagnostic message would be 4224 octets.

For the purpose of diagnostic message information encoding we will

use one or more Type-Length-Value containers where each TLV will have the following format:

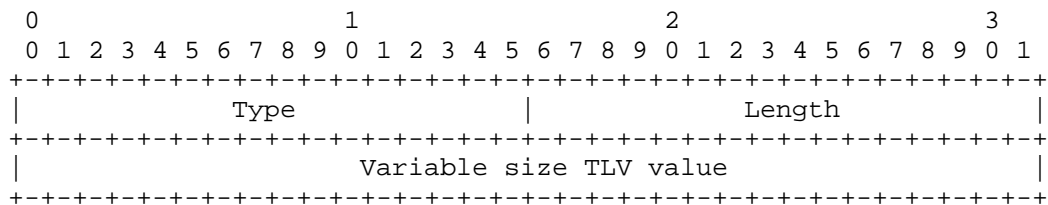


Figure 1: DIAGNOSTIC message TLV Format

Type - 2 octet value indicating the TLV type
 Length - 2 octet value indicating the TLV length in octets
 Value - Variable length value field depending on the type of the TLVs carried.

To work around continued BGP churn issue some types of TLVs will need to contain a sequence number to correlate request with associated to it replies. The sequence number will consist of 8 octets and will be of form: 4 octet `bgp_router_id` + local 4 octet number. When local 4 octet number reaches 0xFFFF it should restart from 0x0000.

Typical application scenario for use of sequence number is to include it in the diagnostic request message and during reply to copy it into reply messages triggered by such request message.

3.2. BGP DIAGNOSTIC Message TLVs

This document defines the following diagnostic TLV types:

- * Operational TLVs
- * BGP database counters exchange
- * Diagnostics for encoding errors in BGP messages
- * AFI/SAFI signaling when malformed update
- * Prefix specific BGP debugging
- * Intra-domain bgp decision monitoring

* Exchange of Route Target filters

3.2.1. Operational TLVs

Type 1 - Diagnostic Message Periodic Request

Length - 2 octets - variable value

Value (N x 2 octets):

TLV type - 2 octets

Use: To indicate the request to periodically receive listed TLV information. TLV type of 0xFFFF indicates request to receive all available diagnostic TLVs from the peer.

Type 2 - Max frequency permitted

Length - 2 octets - variable value

Value (N x 4 octets):

TLV type - 2 octets

Frequency value in seconds two octets 0..65535

Special values:

0 - never send given diagnostic TLV

65535 - no TLV inter-gap minimum set

Use: To indicate in seconds the maximum frequency given TLV may be periodically sent to the bgp speaker

Type 3 - Diagnostic Message Query
Length - 2 octets - variable value
Sequence number - 8 octets

Value (N x 2 octets):
TLV type - 2 octets

Use: To interactively (during debugging/troubleshooting) request to receive listed TLV information. TLV type of 0xFFFF indicates request to receive all available diagnostic TLVs from the peer. TLV of type 0x0000 indicates request to receive a list of all enabled and available diagnostic TLV types from the peer towards querying BGP speaker. The support of this TLV type is mandatory.

Type 4 - Counter's reset request
Length - 2 octets - variable value

Value (N x 2 octets):
TLV type - 2 octets - List of TLVs subject to counter's reset.

Use: To request rest of per neighbor counters of a given TLV type. TLV type of 0xFFFF indicates request to zero all per neighbor counters.

Type 5 - Not supported TLV reply
Length - 2 octets - variable value

Value (N x 3 octets):
 TLV type - 2 octets - TLV that is not supported by the peer
 but where part of TLV Request or TLV Query message
 Error Code - 1 octet - Error code

 Error codes:

 0x01 - Wrong TLV value
 0x02 - TLV not supported for this peer
 0x03 - Max query frequency exceeded
 0x04 - Administratively disabled

Use: To indicate to the peer that the TLV he has requested
 either in TLV Request or in TLV Query message is not
 supported. The support of this TLV type is mandatory.

Type 6 - Enabled and supported TLV types
Length - 2 octets - variable value

Value (N x 2 octets):
 TLV type - 2 octets - TLV that is enabled and supported
 by the peer

Use: To indicate to the peer that the enclosed list of TLVs
 can be requested either in TLV Request or in TLV Query
 messages. The support of this TLV type is mandatory.

3.2.2. BGP database counters exchange

Type 7 - Number of Reachable Prefixes Transmitted/Received
Length - 2 octets - variable value
Sequence number - 8 octets

Value (N x 11 octets):
 AFI/SAFI - 3 octets
 Number of prefixes transmitted - 4 octets
 Number of prefixes received - 4 octets

Use: To indicate number of reachable prefixes exchanged for a given AFI/SAFI between two bgp speakers. This message can be sent only based on the remote query Type 3 which contains the query sequence number to be placed in the reply.

Type 8 - Number of prefixes in BGP_RIB_Out
Length - 2 octets - variable value

Value (N x 7 octets):
 AFI/SAFI - 3 octets
 Number of prefixes 4 octets

Use: To indicate number of prefixes kept in BGP_RIB_Out between bgp speakers for a given AFI/SAFI between two bgp speakers.

Type 9 - Number of paths in BGP_RIB_Out
Length - 2 octets - variable value

Value (N x 6 octets):
 AFI/SAFI - 3 octets
 Number of paths 4 octets

Use: To indicate number of paths kept in BGP_RIB_Out between bgp speakers for a given AFI/SAFI between two bgp speakers.

Type 10 - Number of prefixes present in BGP_RIB
Length - 2 octets - variable value

Value (N x 6 octets):
 AFI/SAFI - 3 octets
 Number of prefixes 4 octets

Use: To indicate number of prefixes kept in BGP RIB for a given
 AFI/SAFI.

Type 11 - Number of paths present in BGP_RIB
Length - 2 octets - variable value

Value (N x 7 octets):
 AFI/SAFI - 3 octets
 Number of prefixes 4 octets

Use: To indicate number of paths kept in BGP RIB for a given
 AFI/SAFI.

3.2.3. Diagnostics for encoding errors in BGP messages

Type 12 - Reachable prefixes present in dropped attribute UPDATE msg
Length - 2 octets - variable value

Value (N octets):
 AFI/SAFI - 3 octets
 1 .. M - List of prefixes

Use: To list reachable prefixes present in the update message
 where optional transitive attribute with partial bit set
 was malformed and has been removed from the update message.
 Prefix encoding should follow given AFI/SAFI definition.

Type 13 - Unreachable prefixes present in dropped attribute UPDATE msg
Length - 2 octets - variable value

Value (N octets):
 AFI/SAFI - 3 octets
 1 .. M - List of prefixes

Use: To list unreachable prefixes present in the update message where optional transitive attribute with partial bit set was malformed and has been removed from the update message. Prefix encoding should follow given AFI/SAFI definition.

Type 14 - Reachable prefixes present in malformed UPDATE msg
Length - 2 octets - variable value

Value (N octets):
 AFI/SAFI - 3 octets
 1 .. M - List of prefixes

Use: To list reachable prefixes present in the malformed update message which were subject to "treat-as-withdraw" behaviour. Prefix encoding should follow given AFI/SAFI definition.

Type 15 - Entire malformed update message enclosure
Length - 2 octets - variable value
Sequence number - 8 octets

Value:
 Malformed message

Use: Propagate the malformed message to the peer upon it's request or at the event of error detection. That includes propagation of messages which had malformed attribute, unparsable content or any other abnormal encoding. If more than a single message has been determined as malformed the subsequent replies will contain the same sequence number and should not be treated as an override.

3.2.4. AFI/SAFI signaling when malformed update

Type 16 - List of ignored AFI/SAFIs by the peer over given session
Length - 2 octets - variable value

Value (N octets):

1..M AFI/SAFI - 3 octets each

Use: To list those AFI/SAFIs which were detected to be malformed by the peer and while session is up were transitioned to IGNORE state.

Such case is inline with Multiprotocol Extensions RFC 4760 as per it's section 7 Error Handling:

"For the duration of the BGP session over which the UPDATE message was received, the speaker then SHOULD ignore all the subsequent routes with that AFI/SAFI received over that session".

3.2.5. Prefix specific BGP debugging

Type 17 - Prefix specific BGP query
Length - 2 octets - variable value

Value (N octets):

AFI/SAFI - 3 octets

Prefix under query

Prefix mask (optional)

Use: To query peer for the status of prefix under examination. When prefix mask is present the request is for exact match. When prefix mask is not present the request is for the longest match. Prefix encoding should follow given AFI/SAFI definition.

Type 18 - Prefix specific BGP response
Length - 2 octets - variable value

Value (N octets):

AFI/SAFI - 3 octets

Prefix under query

Prefix mask (optional)

Prefix status (1 octet)

Status:

0x01 - prefix not found in BGP table

0x02 - prefix in BGP table and active (in FIB)

0x03 - prefix in BGP table and not-active (not in FIB)

0x04 - administratively disabled

Use: To inform peer querying about the status of particular prefix status. Prefix encoding should follow given AFI/SAFI definition.

3.2.6. Intra-domain bgp decision monitoring

Type 19 - Number of IGP metric best path tie breaks executed
Length - 2 octets - variable value

Value (N x 7 octets):

AFI/SAFI - 3 octets

Number of tie breaks 4 octets

Use: To indicate number of prefixes with their best path selected by tie break of IGP metric to their BGP next hop distance step of BGP best path selection algorithm.

Type 20 - Number of BGP best path tie breaks in each selection step
Length - 2 octets - variable value

Value (N x 7 octets):

AFI/SAFI - 3 octets

Best path selection step N - Number of tie breaks 4 octets

Use: To indicate number of cases where in BGP best path selection algorithm given step has been used as a tie break during overall best path selection process for a given prefix.

3.2.7. Exchange of installed Route Target filters

Type 21 - Request for reception of route target filters
 installed towards given peer by RFC4684

Length - 2 octets - variable value

Sequence number - 8 octets

Value (N x 7 octets):

 AFI/SAFI - 3 octets

 BGP Router ID of the peer - 4 octets

Use: To request reception of full table of route target
 filters installed towards listed BGP peer for a requested
 AFI/SAFI. Single request may contain multiple pairs of
 AFI/SAFIs and/or BGP Router IDs.

Type 22 - Reply containing all route target filters installed
 towards given peer

Length - 2 octets - variable value

Sequence number - 8 octets

Value (7 + N * 12 or 24 octets):

 AFI/SAFI - 3 octets

 BGP Router ID of the peer - 4 octets

 List of route targets - each 12 or 24 octets

Use: Allows for troubleshooting purposes to share list of
 route targets installed for a given AFI/SAFI towards
 indicated BGP peer. In the event that RT filtering
 table size will not fit in single BGP Diagnostic
 Message reply the subsequent reply should include
 the same sequence number.

4. Operation

BGP implementation which supports DIAGNOSTIC message can support all
or subset of defined diagnostic types. The range of supported TLV
types will be signaled in the new BGP capability message during BGP
connection establishment phase.

The operation of this extension can be realized on a pool/query based
or push based principles. An implementation may provide, a timer to
periodically send selected Diagnostic types TLVs to the peer or to
the management station.

Similarly BGP peer may periodically or by manual cli request the reception of selected or all of the defined diagnostic TLV types.

The received values are then compared against local counters. When discrepancy is found operator is alarmed and further analysis should follow. The repair actions is out of scope of this document.

Example:

Under some situations when determined that the discrepancy is detected an automated or manual Route Refresh message can be triggered with it's extension for Start_of_Refresh and End_of_Refresh markers . That would allow for purge of any stalled data across two BGP databases.

An important point which needs to be discussed is the exchange of counter's values in light of continued BGP churn presence. As BGP is never stable it is expected that any sort of described counters will also be subject to continues value change making any comparison of their values questionable.

There are three classes of counters defined in this document: sent counters, received counters and current table state counters.

Only "sent" counters can be used for not correlated comparison and problem detection between any two BGP speakers. They are not subject to BGP churn issue due to the fact that DIAGNOSTIC messages would be exchanged inline with BGP UPDATE messages on a given session. An implementation must be able to freeze the received counters when comparing or displaying the received "sent" counters from BGP peer.

Received counters send in the Diagnostic messages are only meaningful in the context of explicit request trigger situation generated by the BGP speaker. BGP speaker should stop transmitting any BGP message of a given AFI/SAFI or freeze corresponding counter after sending diagnostic message request to the peer and before reception of actual diagnostic message reply. In order to correlate diagnostic message requests with associated replies use of build in sequence numbers is provided.

Table state counters (for example number of BGP RIB entries) are exchanged only for informational reasons and they should not be subject to comparison with any local counter values.

5. Capability negotiation

A BGP speaker that is willing to send or receive the BGP DIAGNOSTIC

Messages from its peer should advertise the new DIAGNOSTIC Messages Capability to the peer using BGP Capabilities advertisement [BGP-CAP]. A BGP speaker may send a DIAGNOSTIC message to its peer only if it has received the DIAGNOSTIC message capability from its peer.

The Capability Code for this capability is specified in the IANA Considerations section of this document.

The Capability Length field of this capability is 2 octets. The Capability Value field consists of reserved flags field.

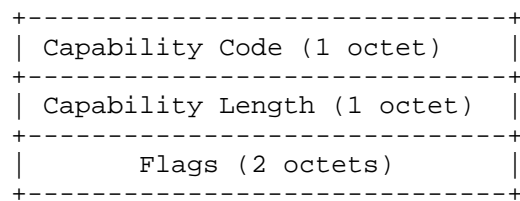


Figure 2: DIAGNOSTIC message BGP Capability Format

6. Security considerations

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

7. IANA Considerations

IANA is requested to allocate a type code for the DIAGNOSTIC message from the BGP Message Types registry, as well as requesting a type code for the new Diagnostic Message Capability negotiation from BGP Capability Codes registry.

This document requests IANA to define and maintain a new registry named: "DIAGNOSTIC Message Type Values". The reserved types are: 0x0000 0xFFFF. The allocation policy is on a first come first served basis.

This document makes the following assignments for the DIAGNOSTIC Message Type Values:

- Type 1 - Diagnostic Message TLV(s) Request
- Type 2 - Max frequency permitted
- Type 3 - Diagnostic Message TLV(s) Query
- Type 4 - Counter's reset request
- Type 5 - Not supported TLV
- Type 6 - Enabled and supported TLV types

- Type 7 - Number of Reachable Prefixes Transmitted/Received
- Type 8 - Number of prefixes in BGP_RIB_Out
- Type 9 - Number of paths in BGP_RIB_Out
- Type 10 - Number of prefixes present in BGP_RIB
- Type 11 - Number of paths present in BGP_RIB

- Type 12 - Reachable prefixes present in dropped attribute message
- Type 13 - Unreachable prefixes present in dropped attribute message
- Type 14 - Reachable prefixes present in malformed UPDATE message
- Type 15 - Entire malformed update message enclosure

- Type 16 - List of ignored AFI/SAFIs by the peer over given session

- Type 17 - Prefix specific BGP query
- Type 18 - Prefix specific BGP response

- Type 19 - Number of IGP metric best path tie breaks executed
- Type 20 - Number of BGP best path tie breaks in each selection step

- Type 21 - Request for reception of route target filters
- Type 22 - Reply containing all route target filters installed

- Type 23 - 65534 Free for future allocation.
- Type 65535 - Reserved

8. Acknowledgments

Authors would like to thank Alton Lo for his valuable input.

9. 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.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement

with BGP-4", RFC 5492, February 2009.

Authors' Addresses

Robert Raszuk
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: raszuk@cisco.com

Enke Chen
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: enkechen@cisco.com

Bruno Decraene
France Telecom
38-40 rue du General Leclerc
Issi Moulineaux cedex 9 92794
France

Email: bruno.decraene@orange-ftgroup.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 19, 2011

R. Raszuk
C. Cassar
Cisco Systems
E. Aman
TeliaSonera
B. Decraene
France Telecom
October 16, 2010

BGP Optimal Route Reflection (BGP-ORR)
draft-raszuk-bgp-optimal-route-reflection-00

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 satisfied, 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 two new solutions which can be deployed to facilitate the application of closest exit point policy centralized route reflection deployments.

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 April 19, 2011.

Copyright Notice

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

This document is subject to BCP 78 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.

Table of Contents

1. Introduction	4
2. Proposed solutions	5
3. Best path selection for BGP hot potato routing from customized IGP network position	6
3.1. Client's perspective best path selection algorithm	7
3.1.1. Flat IGP network	7
3.1.2. Hierarchical IGP network	8
3.2. Aside: Configuration-based flexible route reflector placement	9
3.3. Discussion	10
3.4. Advantages	11
4. Angular distance approximation for BGP warm potato routing	11
4.1. Problem statement	12
4.2. Proposed solution	13
4.3. Centralized vs distributed route reflectors	14
5. Deployment considerations	15
6. Security considerations	15
7. IANA Considerations	16
8. Acknowledgments	16
9. References	16
9.1. Normative References	16
9.2. Informative References	16
Authors' Addresses	17

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 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

learned 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. Protocol extensions like add-paths [I-D.ietf-idr-add-paths] or diverse-path [I-D.ietf-grow-diverse-bgp-path-dist] allow for such improved path diversity and can be used to address the same problems addressed by the mechanisms proposed in this draft. 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 solutions

This document proposes two simple solutions to the problem described above. Both of these solutions make 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. These solutions are primarily applicable in deployments using centralized route reflectors, which are typically implemented in devices without a capable forwarding plane.

The two alternatives are:

"Best path selection for BGP hot potato routing from client's IGP network position"

"Angular distance approximation for BGP warm potato routing"

Both solutions 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.

In both of these solutions the route reflector selects and distributes a route to each client based on what would be optimal from the client's perspective. In the respective solutions the choice is made either factoring in IGP costs or the configured angular distance to the next hop. The route reflector makes different decisions for different clients only in the case where the tie breaker for path selection would have been the IGP distance to the BGP nexthop (as in hot potato routing).

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

3. Best path selection for BGP hot potato routing from customized IGP network position

This section describes a method for calculating the order of preference of BGP paths from the point of view of each separate route reflector client. More specifically, the route reflector will compute the IGP metric to the BGP nexthop from the position of the client to which the resulting path will be distributed, if the IGP metric is the tie breaker applied to a set of possible paths. In the subsequent model authors will propose virtual reflector placement at operator's selected IGP location.

In the case of a hierarchical IGP deployment where the client is in a different level in the hierarchy to the route reflector, the route reflector will compute IGP distance to the BGP nexthop from the Area Border Routers (ABR) leading to the client in lieu of the route reflector client itself, and use the shortest distance from these

ABRs to the nexthop. This provides an approximation to the desired functionality. Rather than a client picking the closest path, the client would be picking the exit point closest to the client region as defined by area or level. In cases where one or more nexthops are in the same region as the client, one of those nexthops would be preferred, with tie breaking within those nexthops performed from the route reflector's position in the network.

It is assumed that reachability through a set of ABRs is always advertised through identical prefixes from those ABRs. If a nexthop is reachable through multiple ABRs but the ABRs advertise reachability through prefixes of different length, then only the ABR advertising the longest prefix will be considered as a viable path to the nexthop.

BGP best path selection and its distribution has a natural consequence of limiting the amount of state in the network. That is not in itself a drawback. BGP speakers will rarely need to receive all available BGP paths. In network deployments with multiple upstream peerings or with very dense peering schemes, the number of available BGP paths for a given BGP prefix can be high. Real network deployments with the number of paths for a prefix ranging from 10s to 100s have been observed. It would be wasteful to propagate all of those paths to all clients, such that each client can select paths according to the position of the nexthop relative to the client.

Whenever a BGP route reflector would need to decide what path or paths need to be selected for advertisement to one of its clients, the route reflector would need to virtually position itself in its client IGP network location in order to choose the right set of paths based on the IGP metric to the next hops from the client's perspective.

This technique applies in deployments with or without diverse paths or the various path selection modes contemplated in add-paths.

3.1. Client's perspective best path selection algorithm

For each centralized route reflector the proposal assumes that the route reflector participates in a common IGP with its clients. There are two scenarios to consider - flat versus hierarchical IGP network.

3.1.1. Flat IGP network

Reflectors run SPF from the client IGP node point of view such that the cost of BGP nexthops from the client can be determined if necessary. For the purpose of BGP path selection the interesting product of this calculation is the ability to determine the IGP

distance from a client to a BGP next hop. This distance to a nexthop would be interesting in cases where that next hop is for a path which is contending with otherwise equally preferred paths. This approach works in tunneled as well as conventional hop-by-hop IP forwarding cores.

When the path selection tie breaker for a prefix is the IGP metric to the BGP nexthops of the contending paths, then the route reflector will determine the order of preference of the contending paths by considering the distance from the client to the path nexthops in order to decide what path/s to advertise to a client (or group of clients where feasible). It should be noted that an operator may wish to provide a distance tolerance value, such that beyond a certain granularity, differences between IGP metric are invisible to the path selection algorithm. This will allow a route reflector some leeway in selecting between paths such that rather than pick one path over another on the basis of a difference in distance which is operationally irrelevant, the route reflector can choose to optimise for update generation grouping. Furthermore, this tolerance will reduce the likelihood of generation of BGP updates when the IGP topology changes in a way which is not operationally relevant. In the case that a path is selected from a set for a given prefix while ignoring differences in distance within the tolerance figure, then that same path must always be preferred for all clients where the paths are within the tolerance figure

3.1.2. Hierarchical IGP network

Hierarchy introduces two challenges:

The first challenge is that the RR IGP view may differ from a client IGP view by virtue of one or the other having a summarised view versus the other. Summarisation, by its nature, loses information. Consider the example where a client within a PoP sees two prefixes with two metrics for two egress points within the PoP, but where the RR only sees a single summary covering reachability to both nexthops as injected by the ABR. However it needs to be observed that inter area networks running LDP are required to disable summarization of all FEC advertised in LDP (typically all loopbacks) unless [RFC5283] is deployed. Such deployments are not likely to suffer summarisation difficulties.

The second challenge is that in cases where the client is in a different level of hierarchy from the RR, the RR can not build a Shortest Path First (SPF) tree with the client node as root, simply because the topology derived by the IGP will not include the client node. It will instead only include reachability to the

client from one or more ABRs. In order to overcome this problem, the RR could compute an SPF tree from the ABRs in the area. The RR would then determine the shortest distance from a client which lives behind the ABRs, to a nexthop, by adding the advertised distances from an ABR to the client and the distance from the ABR to a nexthop, for each ABR, and picking the minimum. This assumes that IGP metrics on links are symmetric; i.e. that the distance from the ABR to the client or nexthop is equal to the distance from the client or nexthop to the ABR.

There are cases where the above approach does not help. If RR is trying to arbitrate amongst a set of paths for a client which is in the same hierarchy as some of those paths, and in a different hierarchy to the RR, the opaqueness of the region containing the client at the RR defeats the selection process. It is impossible to determine the relative position of the RR client and the paths within the client region.

The solution for hierarchical IGP networks also assumes that if RRs are present and are responsible for calculation of BGP best path to clients they are either placed in each local area coinciding with area containing clients or they are placed in the core (area 0/level 2) of the network.

3.2. Aside: Configuration-based flexible route reflector placement

The ability to exploit topology information available in the IGP in ways described above can also be used to virtually place the RR at different points in the network for purposes other than hot potato routing.

A route reflector can be globally configured to "pretend" its logical location is one of any of the other nodes within a given IGP area/level flooding scope regardless of its physical connectivity.

Such flexibility provides a useful tool for reflector virtualization, and supports moving or replacing physical route reflectors without any effect on routing. Such a change can be permanent or it could be performed during network maintenance in order to minimize network impact.

A possible variation would allow the virtual placement of RR to be effected on a per-AF or AF plus update/peer group granularity. It should be noted that this approach provides for splitting one centralized route reflector such that it is virtually positioned at various network locations, with the network location depending upon of address family or address family plus update/peer group.

Virtual slicing of a centralized route reflector relaxes the need to propagate all BGP paths between RRs in a alternative conventional distributed RR deployment. It is expected that such RRs would be deployed in redundant sets, and that those RRs would not need to be physically colocated, while still benefiting from the possibility of being logically colocated, and therefore not compromising any of the best path selection symmetry.

3.3. Discussion

This is not the first instance where a router participating in an IGP is required to build the SPF tree using a root other than itself. Determination of loop free alternate paths as described in [RFC5714] is one such example.

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 optimisations 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.

It may be appropriate to allow the operator, or the route reflector itself, to group clients together using IGP distance between clients to determine grouping. All the operation discussed above which relied upon computing best path for each client, and measuring distances from each client to different nexthops, would instead be performed for each group of clients. A configurable thresholds can be used to determine which IGP metric changes should be visible to BGP, and trigger best paths recomputation. The latter would be beneficial in existng BGP RR code too.

3.4. Advantages

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 unencapsulated hop-by-hop forwarding, such optimisations should be enabled 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 to the 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.

4. Angular distance approximation for BGP warm potato routing

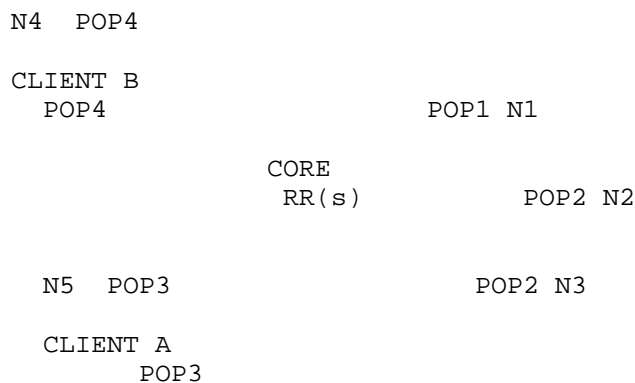
This section describes an alternative solution to the use of IGP topology information to virtually position the RR at the client location in the network. This solution involves modelling the network topology as a set of elements (regions, PoPs or routers) arranged in a circle. Route reflector clients and inter-domain exit points would then be statically assigned to those elements such that one can compute the angular distance between route-reflector clients and the various exit points in order to infer the distance between

any two elements. This measure of distance can be used as an effective alternative to the IGP distance as a tie breaker in the path selection algorithm if necessary.

4.1. Problem statement

This solution addresses the problem described in earlier sections, while attempting to minimise computational overhead. The aim of the proposed solution is to enable a route reflector to provide a route reflector client with an exit point for a prefix which is 'closest' to the client rather than the route-reflector, without having to distribute all paths to that client, or having to derive each client's view of the network topology. The measure of closest is based on a simplistic description of network topology provided by the operator.

Consider the following example of an ISP network topology drawn to reflect the location of the nodes and POPs:



N - represents the different exit points for a given prefix. POP2 is a geographically large PoP with two paths; N2 and N3.

In a deployment where the centralized RRs tie break on the basis of their IGP-based view of the network, N1 above would be advertised to all clients on the basis that it is closest to the RR. Path N4 would be a more appropriate choice for client B. Similarly, N5 would be more appropriate for client A since path N5 is closer to client A than path N1.

4.2. Proposed solution

The proposed solution revolves around the operator establishing the angular position of the route-reflector clients and inter-domain exit points in the network. The route reflector then picks the path to advertise to a client based on the client's angular position versus the angular position of the inter-domain exit points originating the paths. The operator can choose the granularity of angular position appropriate to the desired goals. On one hand, the coarseness of the angular position will effect the operator overhead; versus the optimality of routing on the other. The finest granularity possible will be the relative position of originating clients.

Note that this solution has nothing to do with actual IGP link metrics and resulting topology in the network.

It can be shown that for each network topology, elements such as AS exit points can be mapped on to a circle. By putting POPs, Regions or individual clients onto the hypothetical circle we can identify an angular location for each element relative to some fixed direction; for example defining the angular north of the circle at 0 degrees.

The angular position of elements in the network can be conveyed to a route reflector in a number of ways:

- Assignment of angular position of each RR client through configuration on the route reflector itself; per client configuration on RR

- Assignment of angular position of an RR client at each client, then propagating it to RRs.

The proposed angular distance approximation is compatible with both flat and hierarchical IGP deployments.

In the example illustrated above the route reflector might learn or be configured with the following set of paths and corresponding angular positions:

Prefix X/Y	N1	N2	N3	N4	N5
Location in degrees	60	85	120	290	260

If the absolute angular position of clients A and B were as follows:

Client A: 260 degrees

Client B: 290 degrees

Then the corresponding angular distances for those clients versus the exit points can be calculated as follows:

Prefix X/Y	N1	N2	N3	N4	N5
Client A	200	175	140	30	0
Client B	230	205	170	0	30

With an RR running the BGP best path algorithm modified to use the angular distance from the client to the nexthops, rather than its IGP distance to the nexthops as tie breaker, each client is provided with its closest path with the measure of closeness reflecting the angular position as configured by the operator.

The model used by the operator in order to determine the angular position of a client or exit point, might involve grouping elements together by region or PoP, or might involve no grouping at all. Implementations should allow the operator to pick the appropriate granularity.

4.3. Centralized vs distributed route reflectors

In an environment where the RR clusters are distributed (yet centralized enough to make hot potato routing hard), and each RR cluster serves a subset of clients, it becomes necessary to propagate the angular position of the clients between route reflectors. This can be achieved as follows:

Deploy add-paths between route reflectors in order to maximise path diversity within the cluster.

A non AS transitive BGP community of type (TBA by IANA) can be used to encode and propagate angular position between 0 and 359 of a client. This community is only relevant to the route reflectors of a given BGP domain and should be stripped either at the ASBR boundary or when propagating updates to BGP peers which are not route reflectors.

The angular position marking could also be added by clients and advertised to the route reflector. This would require some configuration effort.

5. Deployment considerations

The solutions are primarily intended for end-to-end tunneled environments, i.e. where traffic is label switched or IP tunneled across the core. If unencapsulated hop-by-hop forwarding is used, either misconfigurations or conflicts between these optimizations and classical BGP path selection rules could lead to intra-domain forwarding loops. Under certain circumstances the solutions can also be deployable without end-to-end tunneling. In particular the best path selection based on the client's IGP best-path selection is guaranteed not to cause any forwarding loops (other than micro loops associated with reconvergence) when deployed in a flat IGP area provided that no distance tolerance value is used so that the path choice is truly made on a per-client basis.

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 and which can be tie broken by IGP distance to the nexthop.

It should be noted that customized per-client or group of clients best path selection is already in use today in the context of Internet Exchange Point (IXP) route servers. In an IXP route server the client best path is selected as a result of different policies rather than IGP metric distance to BGP next hop.

A possible scalability impact of optimising path selection to take account of the RR client position is that different RR clients receive different paths, and therefore update/peer group efficiency diminishes. This cost is imposed by the requirement given the requirement is to optimise the egress path from the client's perspective. It is also not unlikely that groups of clients will end up receiving the same best path/s, in which case, inefficiency of update generation will be minimised. It should be noted that in the cases described under flexible router placement where placement is determined on a per update/peer group basis or per route reflector, the scale benefits of peer groupings are retained.

6. Security considerations

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

7. 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.

8. Acknowledgments

Authors would like to thank Clarence Filsfils and Mike Shand for their valuable input.

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.
- [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.

9.2. Informative References

- [I-D.ietf-grow-diverse-bgp-path-dist]
Raszuk, R., Fernando, R., Patel, K., McPherson, D., and K. Kumaki, "Distribution of diverse BGP paths.", draft-ietf-grow-diverse-bgp-path-dist-02 (work in progress), July 2010.
- [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-04 (work in progress), August 2010.
- [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.

Authors' Addresses

Robert Raszuk
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: raszuk@cisco.com

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
France Telecom
38-40 rue du General Leclerc
Issi Moulineaux cedex 9, 92794
France

Email: bruno.decraene@orange-ftgroup.com

IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 19, 2011

R. Raszuk
Cisco Systems
J. Haas
Juniper Networks
R. Steenbergen
nLayer Communications, Inc.
B. Decraene
France Telecom
P. Jakma
DCS, Uni. of Glasgow
October 16, 2010

Wide BGP Communities Attribute
draft-raszuk-wide-bgp-communities-01

Abstract

Communicating various routing policies via route tagging plays an important role in external BGP peering relations. It is also a very common best practice among operators to propagate various additional information about routes intra domain. The most common tool used today to attach various information about routes is realized with the use of BGP communities.

Such information is important for the BGP speakers to perform some mutually agreed actions without the need to maintain a separate offline database for each pair of prefix and an associated with it requested set of action entries.

This document defines a new encoding which will enhance and simplify what can be accomplished today with the use of BGP communities. The most important addition this specification brings over currently defined BGP communities is the ability to specify, carry as well as use for execution operator's defined set of parameters. Specification also provides an extensible platform for any new community encoding needs in the future.

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 April 19, 2011.

Copyright Notice

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

This document is subject to BCP 78 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.

Table of Contents

1. Introduction	4
2. Wide BGP Community Attribute	4
3. Wide BGP Community Attribute Containers	5
3.1. Fixed size container template	6
3.2. Variable size container template	6
4. Container Type 1: Wide Community	7
4.1. Container Type 1 - TTL	7
4.2. Container Type 1 - Length	8
4.3. Container Type 1 - Community Value	8
4.4. Container Type 1 - Source AS number	8
4.5. Container Type 1 - Community Parameters	8
5. Well Known Standard BGP Communities	9
6. Operational considerations	9
7. Example	10
8. Security considerations	11
9. IANA Considerations	11
10. Contributors	12
11. Acknowledgments	13
12. References	13
12.1. Normative References	13
12.2. Informative References	13
Authors' Addresses	14

1. Introduction

RFC 1997 [RFC1997] defines a BGP Community Attribute to be used as a tool to contain in BGP update message various additional information about routes which may help to automate peering administration. As defined in RFC 1997 [RFC1997] BGP Communities Attribute consists of one or more sets of four octet values, where each one of them specifies a different community. Except two reserved ranges the encoding of community values mandates that first two octets are to contain the Autonomous System number followed by next two octets containing locally defined value.

With the introduction of 4-octet Autonomous System numbers by RFC 4893 [RFC4893] it became obvious that BGP Communities as specified in RFC 1997 will not be able to accommodate new AS encoding. In fact RFC 4893 explicitly recommends use of four octets AS specific extended communities as a way to encode new 4 octet AS numbers.

While encoding of 4 octet AS numbers are being addressed by [draft-ietf-idr-as4octet-extcomm-generic-subtype] neither the base BGP communities (both standard or extended) nor as4octet-extcomm-generic document define sufficient level of encoding freedom which could be of practical use. Authors believe that defining a new BGP Path Attribute which will provide ability to contain locally defined parameters will enhance current level of network policies as well as simplify BGP policy management. Proposed simple encoding will also enable to deliver a set of new network services without a need to define a new BGP extension each time.

While defining a new type of any tool there is always a unique opportunity to specify a subset of well recognized behaviors. List of the most commonly used today BGP communities as well as provision for a new registry for future definitions will be contained in a separate document.

2. Wide BGP Community Attribute

For the purposes of encoding for Wide BGP Communities a new BGP Path Attribute has been defined. The attribute type code is of the value (TBC by IANA).

Wide BGP Community Attribute is an optional, transitive BGP attribute, and may be present only once in the update message.

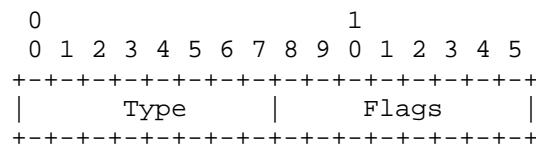
The attribute contains a number of typed containers, which are either fixed or variable in size. Any given container type may appear multiple times, unless that container type's definition says

otherwise.

3. Wide BGP Community Attribute Containers

Two container templates are defined for carrying BGP community information, to hold fixed or variably sized data. All container definitions MUST conform with one of these two templates.

Containers always start with the following header:



Container header

Flags are defined globally, to apply to all community container types.

- Bit 0: 0 => local community value
 - 1 => registered community value
- 1: 0 => do not decrement TTL field across confederation boundaries
 - 1 => decrement TTL across confederation boundaries
- 2...7: => ignored, preserve or set to zero.

Bit 0 set (value 1) indicates that the given container carries a Wide BGP Community which is registered with IANA. When not set (value 0) it indicates that community value which follows is locally assigned with a local meaning. Ignored bits SHOULD be preserved in any received containers, or set to 0 otherwise. Bit 1 is used to manage propagation scope of given community across confederation boundaries. When not set (value of 0) TTL field is not consider at the sub-AS boundaries. When set (value of 1) sub-AS border router follows the same procedure reg handling TTL field as applicable to ASBR at the domain boundary.

3.1. Fixed size container template

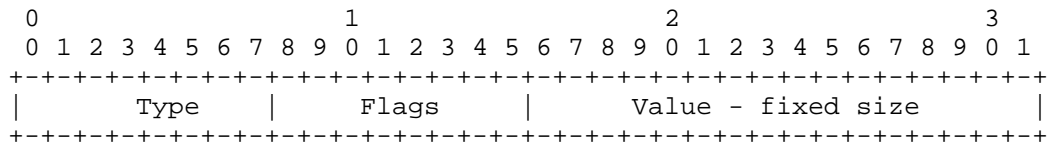


Figure 3: Fixed size type container

3.2. Variable size container template

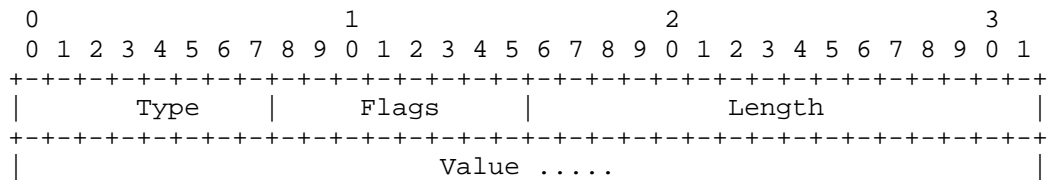
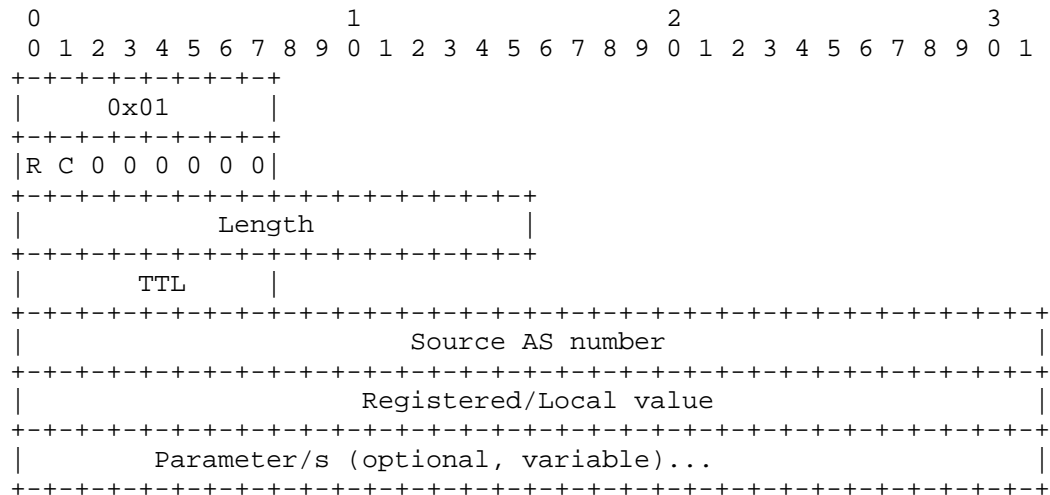


Figure 4: Variable size type container (TLV Format)

4. Container Type 1: Wide Community

Wide BGP Community Type 1 container is of variable size and is encoded as follows:



R is the value of the registered/local bit. C is the value indicating how to treat TTL field across confederation boundaries.

Figure 4: Wide BGP Community Type 1

4.1. Container Type 1 - TTL

TTL: 1 octet

This field represents the forwarding radius in the unit of AS hops for given Wide BGP Community. At each AS boundary when propagating given community over an EBGp session the TTL field must be decremented by value of 1 by the sending EBGp speaker. TTL with value of zero received to the ASBR over IBGP session indicates that this community must not cross an AS boundary.

The special value of 0xFF indicates that the enclosed community may be always propagated over EBGp boundary. Value of 0xFF must not be decremented during propagation.

The exact same procedures as described above applies also to sub-confederation boundaries when the global C flag is set to 1.

4.2. Container Type 1 - Length

The length represents the total lengths of a given container in octets. The minimum length when no optional parameters are attached is 13 octets.

4.3. Container Type 1 - Community Value

Community Value: 2 octets

The Wide BGP Community value encoded in this field indicates private/local or registered Wide BGP Community type which defines what set of actions a router is requested or recommended to take upon reception of routes with such BGP communities.

4.4. Container Type 1 - Source AS number

Source Autonomous System number: 4 octets

The Autonomous System number which indicates the originator of given Wide BGP Community.

When Autonomous System is a two octet number the first two octets of this 4 octet value are to be filled with zeros.

4.5. Container Type 1 - Community Parameters

Parameters: variable size

Community parameter are defined to contain additional data for execution of given BGP community.

Community parameter field could consist of an autonomous system number(s) which should be conditionally compared when executing given community, AS PATH prepend count to be added, local preference value to be inserted under some conditions, markers indicating number of BGP speakers traversed, cumulative IGP metrics to be used for transparent redistribution, etc...

For consistent Autonomous System treatment all encoded AS numbers SHOULD be encoded as 4 octet values. When such AS is a two octet number the first two octets of this 4 octet value are to be filled with zeros.

Two special values are reserved in the Parameter Autonomous System number field: 0x00000000 - to indicate "None of Autonomous Systems" and value of 0xFFFFFFFF - to indicate "All of Autonomous Systems".

The detailed interpretation of each set of parameters will be provided when describing given community type in a separate document or when locally defined by an operator.

5. Well Known Standard BGP Communities

According to RFC 1997 as well as to IANA's Well-Known BGP Communities registry today the following BGP communities are defined to have global significance:

0xFFFF0000	planned-shut	[draft-francois-bgp-gshut]
0xFFFFFFFF01	NO_EXPORT	[RFC1997]
0xFFFFFFFF02	NO_ADVERTISE	[RFC1997]
0xFFFFFFFF03	NO_EXPORT_SUBCONFED	[RFC1997]
0xFFFFFFFF04	NOPEER	[RFC3765]

This document recommends for simplicity as well as for avoidance of backward compatibility issues the continued use of BGP Standard Community Attribute type 8 as defined in RFC 1997 to distribute non Autonomous System specific Well-Known BGP Communities.

For the same reason the described registry does not intended to obsolete BGP Extended Community Attribute and any already defined and already deployed extended communities.

6. Operational considerations

Having two different ways to propagate locally assigned BGP communities, one via use of Standard BGP Community attribute and the other one via use of Wide BGP Community may seem to potentially cause problems when considering propagation of conflicting actions.

However it needs to be noticed and pointed out that today even within Standard BGP Communities operator or operators may append similar conflicting information to already existing community propagation tool set.

It is therefor recommended that any implementation when supporting both standard and wide BGP communities will allow for their easy inbound and outbound policing. For the actual execution all communities should be treated as union and if supported by an implementation their execution permission are to be a local configuration matter.

When advertising as well as during insertion of Wide BGP Communities

which are predefined as range of values - only use of one value of selected range is allowed.

7. Example

An operator wishes to tag incoming routes with a policy specifying that during their advertisement to two peering ASes 2424 and 8888 or during their advertisement to peers marked as RED (0xFF0000) the routes carrying such community will be advertised with AS_PREPEND equal to 4.

That can be easily accomplished by locally defining by an operator a new wide community value using type 1 proposed encoding as below:

PREPEND 4 TIMES TO AS 2424 or 8888 or to peers marked as RED

TTL - 0x00
LENGTH - 26 octets
VALUE - 01 / 0x12
PARAMETERS - 2 x 4 octets AS number
 1 x class of peers
 1 octet prepend's number

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+
|           0x1           |
+-----+-----+-----+-----+
| 0 0 0 0 0 0 0 0 0 0 |
+-----+-----+-----+-----+
| Length:      26         |
+-----+-----+-----+-----+
|      TTL: 0           |
+-----+-----+-----+-----+
|                               Own ASN                               |
+-----+-----+-----+-----+
|      Community: LOCAL PREPEND ACTION CATEGORY I                    |
+-----+-----+-----+-----+
|      Target ASN# 2424  (0x00000978)                                |
+-----+-----+-----+-----+
|      Target ASN# 8888  (0x000022B8)                                |
+-----+-----+-----+-----+
|      Peer color RED 0x00FF0000                                     |
+-----+-----+-----+-----+
|      Prepend #: 4 |
+-----+-----+-----+-----+

```

8. Security considerations

All the security considerations for BGP Communities as well as for BGP RFCs apply here.

9. IANA Considerations

This document defines a new BGP Path Attribute called Wide BGP Communities Attribute. For this new type IANA is to allocate new type value in the corresponding registry:

Registry Name: BGP Path Attributes

This document makes the following assignments for the optional, transitive Wide BGP Communities Attribute:

Name	Type Value
----	-----
Wide BGP Community Attribute	27

This document requests IANA to define and maintain a new registry named: "Wide BGP Communities Attribute Container Types".

The pool of: 0x00-0xFF has been defined for its allocations. The allocation policy is on a first come first served basis.

This document makes the following assignments for the Wide BGP Communities Attribute Types values:

Name	Type Value
----	-----
Reserved	0x00
Type 1	0x01
Types 2-254 to be allocated on FCFS basis	
Reserved	0xFF

10. Contributors

The following people contributed significantly to the content of the document:

Shintaro Kojima
OTEMACHI 1st. SQUARE EAST TOWER, 3F
1-5-1, Otemachi,
Chiyoda-ku, Tokyo 100-0004
Japan
Email: koji@mfeed.ad.jp

Juan Alcaide
Cisco Systems
Research Triangle Park, NC
United States
Email: jalcaide@cisco.com

Burjiz Pithawala
Cisco Systems
170 West Tasman Dr
San Jose, CA
United States
Email: bpithaw@cisco.com

Saku Ytti
TDC Oy
Mechelininkatu 1a
00094 TDC
Finland
Email: ytti@tdc.net

11. Acknowledgments

Authors would like to thank Enke Chen, Pedro Marques and Alton Lo for their valuable input.

12. References

12.1. 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.

12.2. Informative References

- [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.
- [RFC4893] Vohra, Q. and E. Chen, "BGP Support for Four-octet AS

Number Space", RFC 4893, May 2007.

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

Authors' Addresses

Robert Raszuk
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: raszuk@cisco.com

Jeffrey Haas
Juniper Networks
1194 N.Mathilda Ave
Sunnyvale, CA 94089
US

Email: jhaas@pfrc.org

Richard A Steenbergen
nLayer Communications, Inc.
209 W Jackson Blvd
Chicago, IL 60606
US

Email: ras@nlayer.net

Bruno Decraene
France Telecom
38-40 rue du General Leclerc
Issi Moulineaux cedex 9 92794
France

Email: bruno.decraene@orange-ftgroup.com

Paul Jakma
School of Computing Science, Uni. of Glasgow
Sir Alwyn Williams Building
University of Glasgow
Glasgow G1 5AE
UK

Email: paulj@dcs.gla.ac.uk

Network Working Group
Internet Draft
Intended status: Standards Track
Oct 22, 2010
Expires: Apr 22, 2011

J. Uttaro
AT&T
V. Van den Schrieck
P. Francois
UCLouvain
R. Fragassi
A. Simpson
Alcatel-Lucent
P. Mohapatra
Cisco Systems

Best Practices for Advertisement of Multiple Paths in BGP
draft-uttaro-idr-add-paths-guidelines-03.txt

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

This Internet-Draft will expire on April 22, 2011.

Copyright Notice

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

This document is subject to BCP 78 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 BSD License.

Abstract

Add-Paths is a BGP enhancement that allows a BGP router to advertise multiple distinct paths for the same prefix/NLRI. This provides a number of potential benefits, including reduced routing churn, faster convergence and better loadsharing.

This document provides recommendations to implementers of Add-Paths so that network operators have the tools needed to address their specific applications and to manage the scalability impact of Add-Paths. A router implementing Add-Paths may learn many paths for a prefix and must decide which of these to advertise to peers. This document analyses different algorithms for making this selection and provides recommendations based on the target application.

Table of Contents

1. Introduction.....	4
2. Terminology.....	4
3. Add-Paths Applications.....	5
3.1. Fast Connectivity Restoration.....	5
3.2. Load Balancing.....	7
3.3. Churn Reduction.....	7
3.4. Suppression of MED-Related Persistent Route Oscillation...	7
4. Implementation Guidelines.....	8
4.1. Capability Negotiation.....	8
4.2. Receiving Multiple Paths.....	9
4.3. Advertising Multiple Paths.....	9
4.3.1. Path Selection Modes.....	11
4.3.1.1. Advertise All Paths.....	11
4.3.1.2. Advertise N Paths.....	11
4.3.1.3. Advertise All AS-Wide Best Paths.....	12
4.3.1.4. Advertise ALL AS-Wide Best and Next-Best Paths (Double AS Wide).....	13
4.3.2. Derived Modes from Bounding the Number of Advertised Paths.....	14
5. Scalability and Routing Consistency Considerations.....	14
5.1. Scalability Considerations.....	14
5.2. Routing Consistency Considerations.....	14
5.3. Consistency between Advertised Paths and Forwarding Paths	15
6. Security Considerations.....	16

7. IANA Considerations.....	16
8. Conclusions.....	16
9. References.....	16
9.1. Normative References.....	16
9.2. Informative References.....	16
10. Acknowledgments.....	17
Appendix A. Other Path Selection Modes.....	18
A.1. Advertise Neighbor-AS Group Best Path.....	18
A.2. Best LocPref/Second LocPref.....	18
A.3. Advertise Paths at decisive step -1.....	19

1. Introduction

The BGP Add-Paths capability enhances current BGP implementations by allowing a BGP router to exchange with its BGP peers more than one path for the same destination/NLRI. The base BGP standard [RFC 4271] does not provide for such a capability. If a BGP router learns multiple paths for the same NLRI (from multiple peers), it selects only one as its best path and advertises the best path to its peers. The primary goal of Add-Paths is to increase the visibility of paths within an iBGP system. This has the effect of improving robustness in case of failure, reducing the number of BGP messages exchanged during such an event, and offering the potential for faster re-convergence. Through careful selection of the paths to be advertised, Add-Paths can also prevent routing oscillations.

The purpose of this document is to provide the necessary recommendations to the implementers of Add-Paths so that network operators have the tools needed to address their specific applications and to manage the scalability impact of Add-Paths while maintaining routing consistency. A router implementing Add-Paths may learn many paths for a prefix and must decide which of these to advertise to peers. This document analyses different algorithms for making this selection and provides recommendations based on the target application.

2. Terminology

In this document the following terms are used:

Add-Paths peer: refers a peer with which the local system has agreed to receive and/or send NLRI with path identifiers

Primary path: A path toward a prefix that is considered a best path by the BGP decision process [RFC 4271] and actively used for forwarding traffic to that prefix. A router may have multiple primary paths for a prefix if it implements multipath.

Backup path: One of the non-best paths toward a prefix.

Optimal backup path: the backup path that will be selected as the new best path for a prefix when all primary paths are removed/withdrawn.

AS-Wide preferred paths: All paths that are considered as best when applying rules of the BGP decision process up to the IGP tie-break.

Path diversity: The property that a router has several paths for a given prefix and each one is associated with a unique BGP next-hop (and BGP router).

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].

3. Add-Paths Applications

[draft-pmohapat] presents the applications that would benefit from multiple paths advertisement in iBGP. They are summarized in the following subsections.

3.1. Fast Connectivity Restoration

With the dissemination of backup paths, fast connectivity restoration and convergence can be achieved. If a router has a backup path, it can directly select that path as best upon failure of the primary path. This minimizes packet loss in the dataplane. Sending multiple paths in iBGP allows routers to receive backup paths when path visibility is not sufficient with classical BGP. This is especially useful when Route Reflection is used.

Consider a network such as the one depicted in Figure 1 and suppose that none of the routers support Add-Paths. From AS1 there are 3 paths (A, B and C) to a particular destination XYZ: two of the paths are via AS3 and one of the paths is via AS2. In this example, Path A is preferred over Path B due to Path A having a lower MED (multi-exit discriminator) (MED for Path A is lower than MED for path B).

AS1 uses a route reflector RR1 to reduce the scale of its iBGP mesh. During steady state, RR1 knows about (has in its RIB-IN) only 2 of the 3 paths. Router B suppresses the advertisement of its best external path (B) to RR, an iBGP peer, because its best overall path is A, learnt from router A (via the RR). RR1 chooses path A as the overall best since its IGP cost to router A is the lowest among path A and C. During normal conditions, router D has even less knowledge of the available paths to destination XYZ; it knows only about path (A), the best path from RR1's perspective.

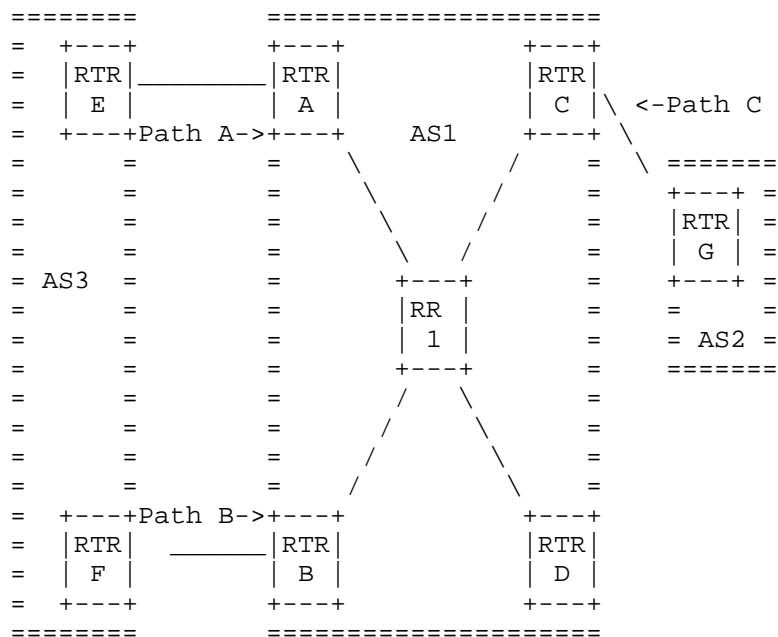


Figure 1: Example Topology

Consider now the steps required to restore traffic from router D to destination XYZ when the link between Router A and Router E fails.

1. Router A sends a BGP UPDATE message withdrawing its advertisement of path (A).
2. RR receives the withdrawal, and propagates it to its other client peers, routers B, C and D.
3. When router B receives the withdrawal of path (A) it reruns its decision process and selects path (B) as its new best path. Router B advertises path (B) to RR.
4. RR reruns its decision process and selects path (B) as its new best path. RR advertises path (B) to client peers A, C and D.
5. Router D reruns its decisions process, determines path (B) to be the best path, and updates its forwarding table. After this step traffic from router D to destination XYZ is restored (the traffic path has changed from A to B).

With the use of Add-Paths, the convergence time for the above path failure example can be reduced considerably. The main reason for the improvement is that Add-Paths allows router D to be aware of more than one path to destination XYZ prior to the failure of the best path (A). In steady-state (with no failures) router B decides, as before, that path (A) is its best path but it also advertises path (B) - which happens to be its next-best overall path and its best "external" path - to RR. With Add-Paths RR1 now has knowledge of all 3 paths to destination XYZ and it can advertise more than just the best path (A) to its peers. Suppose RR1 is allowed to advertise up to 3 paths for destination XYZ. In this case, with the appropriate path selection algorithm, it will advertise paths (A), (B) and (C) to router D. Now consider again the scenario where the link between Router A and Router E fails. In this case, with Add-Paths, fewer steps are required to achieve re-convergence:

1. Router A sends a BGP UPDATE message withdrawing its advertisement of path (A).
2. RR1 receives the withdrawal, and propagates it to its other client peers, routers B, C and D.
3. Router D receives the withdrawal, reruns the decision process and updates the forwarding entry for destination XYZ.

3.2. Load Balancing

Increased path diversity allows routers to install several paths in their forwarding tables in order to load balance traffic across those paths.

3.3. Churn Reduction

When Add-Paths is used in an AS, the availability of additional backup paths means failures can be recovered locally with much less path exploration in iBGP and therefore less Updates disseminated in eBGP. When the preferred backup path is the post-convergence path, churn is minimized.

3.4. Suppression of MED-Related Persistent Route Oscillation

As described in [oscillation], Add-Paths is a valuable tool in helping to stop persistent route oscillations caused by comparison of paths based on MED in topologies where route reflectors or the confederation structure hide some paths. With the appropriate path selection algorithm Add-Paths stops these route oscillations because the same set of paths are consistently advertised by the route

reflector or the confederation border router and the routers receiving this set of paths make stable routing decisions about the best path.

4. Implementation Guidelines

In this section, we discuss recommendations for the implementation of add-paths. We first discuss the BGP capability negotiations related to the use of Add-paths among iBGP peers, as well as their configuration aspects. Next, we provide an overview of RIB-IN management issues for the support of Add-paths. Finally, we discuss the properties of various algorithms for the selection of the paths to be advertised by a BGP speaker supporting Add-paths. The goal of this last section is to recommend, in future revisions of the draft, a default paths selection mode, as well as the minimal set of modes to be supported by a BGP speaker supporting Add-paths.

4.1. Capability Negotiation

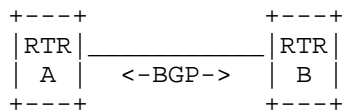


Figure 2: BGP Peering Example

In Figure 2, in order for a router A to receive multiple paths per NLRI from peer B, for a particular address family (AFI=x, SAFI=y), the BGP capabilities advertisements during session setup must indicate that peer B wants to send multiple paths for AFI=x, SAFI=y and that router A is willing to receive multiple paths for AFI=x, SAFI=y. Similarly, in order for router A to send multiple paths per NLRI to peer B, for a particular address family (AFI=x, SAFI=y), the BGP capabilities advertisements must indicate that router A wants to send multiple paths for AFI=x, SAFI=y and peer B is willing to receive multiple paths for AFI=x, SAFI=y. Refer to [Add-Paths] for details of the Add-Paths capabilities advertisement.

The capabilities of the local router shall be configurable per peer and per address family, with the ability to configure send-only operation or receive-only operation. The default mode of operation shall be to both send and receive.

4.2. Receiving Multiple Paths

Currently, per standard BGP behavior, if a BGP router receives an advertisement of an NLRI and path from a specific peer and that peer subsequently advertises the same NLRI with different path information (e.g. a different NEXT_HOP and/or different path attributes) the new path effectively overwrites the existing path.

When Add-Paths has been negotiated with the peer, the newly advertised path should be stored in the RIB-IN along with all of the paths previously advertised (and not withdrawn) by the peer.

When the Add-Paths receive capability for (AFIx, SAFIy) has been negotiated with a peer all advertisements and withdrawals of NLRI within that address family by that peer shall include a path identifier, as described in [Add-Paths]. The path identifiers have no significance to the receiving peer. If the combination of NLRI and path identifier in an advertisement from a peer is unique (does not match an existing route in the RIB-IN from that peer) then the route is added to the RIB-IN. If the combination of NLRI and path identifier in a received advertisement is the same as an existing route in the RIB-IN from the peer then the new route replaces the existing one. If the combination of NLRI and path identifier in a received withdrawal matches an existing route in the RIB-IN from the peer then that route shall be removed from the RIB-IN.

A BGP UPDATE message from a peer sending NLRI with the path identifier may advertise and withdraw more than one NLRI belonging to one or more address families. In this case Add-Paths may be supported for some of the address families and not others. In this situation the receiving BGP router should not expect that all of the path identifiers in the UPDATE message will be the same.

4.3. Advertising Multiple Paths

[Add-Paths] specifies how to encode the advertisement of multiple paths towards the same NLRI over an iBGP session, but provides no details about which set of multiple paths should be advertised. In this section, four path selection algorithms are described and compared with each other. These 4 algorithms are considered to be the most useful across the widest range of deployment scenarios. Of course the list of possible path selection algorithms is much larger and for the interested reader Appendix A provides information about other path selection modes that were considered in historical versions of this document.

In comparing any two path selection algorithms the following factors should be taken into account:

Control Plane Load: When a router receives multiples paths for a prefix from an iBGP client it has to store more paths in its Adj-Rib-Ins.

Control Plane Stress: Coping with multiple iBGP paths has two implications on the computation that a router has to handle. First, it has to compute the paths to send to its peers, i.e. more than the best path. Second, it also has to handle the potential churn related to the exchange of those multiple paths.

MED/IGP oscillations: BGP sometimes suffers from routing oscillations when the physical topology differs from the logical topology, or when the MED attribute is used. This is due to the limited path visibility when a single path is advertised and Route Reflection is used. Increasing the path visibility by advertising multiple paths can help solve this issue.

Path optimality: When a single path is advertised, border routers do not always receive the optimal path. As an example, Route Reflectors send a single path chosen based on their own IGP tie-break. Increasing path visibility would also help routers to learn the path that is best suited for them w.r.t. the IGP tie-break.

Backup path optimality: Multiple paths advertisement gives routers the opportunity to have a backup path. However, some backup paths are better than others. Indeed, when a link failure occurs, if a router already knows its post-convergence path, the BGP re-convergence is straightforward and traffic is less impacted by the transient use of non-best forwarding paths.

Convergence time: Advertising multiple paths in iBGP has an impact on the convergence time of the BGP system. More paths need to be exchanged, but on the other hand, the routing information is propagated faster. With an increased path visibility, there is less path exploration during the convergence. Also, with the availability of backup paths, convergence time in case of failure is also reduced.

Target application: Depending on the application type, the number of paths to advertise for a prefix will vary. For example, for fast connectivity restoration, it may be sufficient to advertise only 2 paths to a peer so that it will have the best path and the optimal backup path. For load balancing purposes, it may be desirable to advertise more paths, but inclusion of the optimal backup path in the

set may be less critical. For route oscillation elimination, it is required to advertise all group-best paths for a prefix.

4.3.1. Path Selection Modes

The following subsections describe the 4 main path selection modes considered in this draft. Each mode is considered either MANDATORY or OPTIONAL. A MANDATORY mode should be present in any implementation that claims compliance with [Add-Paths]. An OPTIONAL mode may be supported by some but not all implementations.

The path selection mode and any parameters applicable to the mode MUST be configurable per AFI/SAFI and per peer and SHOULD be configurable per prefix.

4.3.1.1. Advertise All Paths

A simple rule for advertising multiple paths in iBGP is to simply advertise to iBGP peers all received paths, provided they pass export filters. This solution is easy to implement, but the counterpart is that all those paths need to be stored by all routers that receive them, which can be quite expensive. If a path to a prefix P is advertised to N border routers, with a Full Mesh of iBGP sessions, all routers have N paths in their Adj-RIB-Ins. If Route Reflection is used and each client is connected to 2 Route Reflectors, it may learn up to 2*N paths.

This solution gives a perfect path visibility to all routers, thus limiting churn and losses of connectivity in case of failure. Indeed, this allows routers to select their optimal primary path, and to switch on their optimal backup path in case of failure.

However, as more paths are exchanged, the number of BGP messages disseminated during the initial iBGP convergence can be high, and convergence may be slower.

Routing oscillations are prevented with this rule, because a router won't need to withdraw a previously advertised path when its best path changes.

Routers that support Add-Path MAY support this path selection mode. It is an OPTIONAL mode.

4.3.1.2. Advertise N Paths

Another solution is for a router to advertise a maximum of N paths to iBGP peers. Here, the computational cost is the selection of the N

paths. Indeed, there must be a ranking of the paths in order to advertise the most interesting ones. A way for a router to select N paths is to run N times its decision process. At each iteration of the process only those paths not selected during a previous iteration and not having a NEXT_HOP or BGP Identifier (or Originator ID) in common with the previously-selected paths are eligible for consideration. The memory cost is bounded: a router receives a maximum of N paths for each prefix from each peer. With N equal to 2, all routers know at least two paths and can provide local recovery in case of failure. If multipath routing is to be deployed in the AS, N can be increased to provide more alternate paths to the routers.

Path optimality and backup path optimality are not guaranteed, but as path diversity is better, the nexthops of the chosen primary and backup path are more likely to be closer to the router than with classical BGP.

This solution helps to reduce routing oscillations, but not in all cases. Indeed, path visibility is still constrained by the maximum number of paths, and configurations with routing oscillations still exist.

Routers that support Add-Path MUST support this path selection mode. The default value of N must be 2. The value of N MUST be configurable and MAY be upper bounded by an implementation.

The default value of 2 ensures the availability of a backup path (if 2 or more paths have been received) while maintaining minimum impact to memory and churn. If Add-N with N equal to 2 is insufficient to meet another objective (e.g. loadsharing or MED/IGP oscillation) there is always a large enough value of N that can be selected, if N is configurable, to meet that objective.

4.3.1.3. Advertise All AS-Wide Best Paths

Another choice is to advertise all paths with the same AS-wide preference [Basu-ibgp-osc], i.e. the paths that all routers would select based on the rules of the decision process that are not router-dependent (i.e. Local-preference, ASPath length and MED rules). Thus, for a given router, those paths only differ by the IGP cost to the nexthop or by the tie-breaking rules.

The computational cost is reduced, as a router only has to send the paths remaining before applying the IGP tie-breaking rule. However, it is difficult to predict how many paths will be stored, as it depends on the number of eBGP sessions on which this prefix is advertised with the best AS-wide preference.

With this rule, the routing system is optimal: all routers can choose their best path (or best paths if multipath is used) based on their router-specific preferences, i.e. the IGP cost to the nexthop. Hot potato routing is respected. Also, MED oscillations are prevented, because the path visibility among the AS-wide preferred paths is total.

The existence of a backup path is not guaranteed. If only one path with the AS-wide best attributes exists, there is no backup path disseminated. However, if such a path exists, it is optimal as it has the same AS-wide preference as the primary

Routers that support Add-Path MAY support this path selection mode. It is an OPTIONAL mode.

4.3.1.4. Advertise ALL AS-Wide Best and Next-Best Paths (Double AS Wide)

This variant of "Advertise All AS Wide Best Paths" trades-off the number of paths being propagated within the iBGP system for post-convergence alternate paths availability and routing stability. A BGP speaker running this mode will select for advertisement its AS Wide Best paths, plus all the AS Wide Best paths obtained when removing the first ones from consideration.

Under this mode, a BGP speaker knows multiple AS-Wide best paths or the AS-Wide best path and all the second AS-Wide best paths, so that routing optimality and backup path availability are ensured. Note that the post-convergence paths will be known by each BGP node in an AS supporting this mode.

The computation complexity of this mode is relatively low as it requires to run the usual BGP Decision Process up to and including the MED rule. The set of paths remaining after that step form the AS-Wide best paths. Next, a best path selection algorithm is run up to and including the MED rule, based on the paths that are not in the set of AS-Wide best paths.

The number of paths for a prefix p, known by a given router of the AS, is the number of AS-Wide best and second AS-Wide best paths found at the Borders of the AS.

MED Oscillations are avoided by this mode, both for the primary and alternate paths being picked under this mode.

Routers that support Add-Path MAY support this path selection mode. It is an OPTIONAL mode.

4.3.2. Derived Modes from Bounding the Number of Advertised Paths

For some of the modes discussed in section 4.3.1 the number of paths selected by the algorithm (M) is not predictable in advance, and depends on factors such as network topology. For such modes, implementations MAY support the ability to limit the number of advertised paths to some value N that is less than M.

It must be noted that the resulting derivative mode may no longer meet the properties stated in section 4.3.1 (which assumes $N=M$). This is particularly true for the MED oscillation avoidance property. The use of such bounds thus needs to be considered carefully in deployments where MED oscillation avoidance is a key goal of deploying Add-path. If fast recovery is the main objective then it is reasonable and sufficient to set N to 2. If the main goal is improved load-balancing then limiting N to number of ECMP paths supported by the forwarding planes of the receiving routers is also a reasonable practice.

5. Scalability and Routing Consistency Considerations

When Add-Paths is introduced into a network it can have important implications on nodal and network scalability and routing consistency and correctness.

5.1. Scalability Considerations

In terms of scalability, we note that advertising multiple paths per prefix requires more memory and state than the current behavior of advertising the best path only. A BGP speaker that does not implement Add-Paths maintains send state information in its prefix data structure per neighbor as a way to determine that the prefix has been advertised to the neighbor. With Add-Paths, this information has to be replicated on a per path basis that needs to be advertised. Mathematically, if "send state" size per prefix is 's' bytes, number of neighbors is 'n', and number of paths being advertised is 'p', then the current memory requirement for BGP "send state" = $n * s$ bytes; with Add-Paths, it becomes $n * s * p$ bytes. In practice, this value may be reduced with implementation optimizations similar to attribute sharing. Receiving multiple paths per prefix also requires more memory and state since each path is a separate entry in the Adj-RIB-Ins.

5.2. Routing Consistency Considerations

As discussed in previous sections Add-Paths can help routers select more optimal paths and it can help deal with certain route

oscillation conditions arising from incomplete knowledge of the available paths. But depending on the path selection algorithm and how it is used Add-Paths is not immune to its own cases of routing inconsistencies. If the BGP routers within an AS do not make consistent routing decisions about how to reach a particular destination, route oscillations may occur and these route oscillations may result in traffic loss.

Optimizing an Add-Paths deployment for scalability may run counter to routing consistency goals, and in these circumstances operators have to decide the correct tradeoff for their particular deployment. For example the Advertise All Paths mode, if applied to many prefixes, is far from ideal from a scalability perspective but it does guarantee routing consistency and correctness. A path selection mode that allows better control over scalability is the Advertise N paths mode, but this is susceptible to routing inconsistency. First, if the N paths do not include the best path from each neighbor AS group then route oscillation cannot be precluded. Second, if the advertising router (e.g. an RR) advertises N paths to peer_n and M paths to peer_m, and $N < M$, care must be exercised to ensure that all paths advertised to peer_n are included in the paths advertised to peer_m. This can be assured as long as the advertising router has strictly ordered all of its paths

5.3. Consistency between Advertised Paths and Forwarding Paths

When using Add-Paths, routers may advertise paths that they have not selected as best, and that they are thus not using for traffic forwarding. If two levels of encapsulation are used in the network as described in [RFC4364], this is not an issue, as only the ingress router performs a lookup in its BGP-fed FIB. The traffic is encapsulated to the egress link, and no other router on the forwarding path needs to perform a BGP lookup. The dataplane path followed by the packets is the one intended by the ingress router, and corresponds to the control plane path it advertises.

However, in some networks using Add-Paths without double encapsulation, some scenarios can result in forwarding deflection or loops. Such forwarding anomalies already occur without Add-Paths, when the routers on the forwarding path do not use the same nexthop as the ingress router. They will deflect the traffic to their own nexthop, and, when multiple deflections occur, forwarding loops can appear. With Add-Paths, the issue can be exacerbated due to routers advertising non-best paths, even when one level of encapsulation is used. Indeed, both the ingress and the egress routers perform a BGP lookup, and traffic can be deflected by the egress router.

A first example of such issue is when the Local-Pref of paths received over iBGP sessions is modified. The ingress router may thus select as best a path non-preferred by the egress, and the egress router will thus deflect the traffic.

Another example is when the best path is selected based on tie-breaking rule. When the ingress and the egress base their path selection on the router-id of the neighbor that advertised the path to them, the result may be different for each of them. This specific issue is described and solved in [draft-pmohapat].

6. Security Considerations

TBD

7. IANA Considerations

TBD

8. Conclusions

TBD

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

[Add-Paths] Walton, D., Retana, A., Chen E., Scudder J., "Advertisement of Multiple Paths in BGP", February 6, 2010.

[draft-pmohapat] Mohapatra, P., Fernando, R., Filsfils, C., and R. Raszuk, "Fast Connectivity Restoration Using BGP Add-path", draft-pmohapat-idr-fast-conn-restore-00.txt (work in progress), September 2008.

[oscillation] Walton, D., Retana, A., Chen, E., Scudder, J., "BGP Persistent Route Oscillation Solutions", draft-walton-bgp-route-oscillation-stop-03.txt, May 10, 2010.

[Basu-ibgp-osc] Basu, A., Ong, C., Rasala, A., Sheperd, B., and G.

Wilfong, "Route oscillations in iBGP with Route Reflection", Sigcomm 2002.

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

10. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

Appendix A. Other Path Selection Modes

A.1. Advertise Neighbor-AS Group Best Path

[walton-osc] proposes that a router groups its paths based on the neighbor AS from which it was learned, and to advertise the best path in each of those groups.

The control plane stress induced by this solution is the computation of the per-neighbor path group, and the application of the decision process to each of them. The Control-Plane load is bounded by the number of neighboring ASes advertising a prefix, which cannot be known a-priori.

Path optimality and backup path optimality are not guaranteed, as the paths advertised are not all the AS-wide preferred paths. Backup path availability is not guaranteed. Indeed, if only one AS advertises this prefix, even on multiple eBGP sessions, only one of the paths may be selected and advertised.

A.2. Best LocPref/Second LocPref

This selection method consists in grouping the paths by Local Preference. A router sends to its peers all paths with the highest Local Preference. If there is only a single path with the highest Local Preference, it also sends all paths with the second best Local Preference.

This method ensures that all routers know all paths with the best local preference. As local preference are often related to the type of peering of the peer the path comes from, this ensures that in case of failure, routers have a backup path of equivalent quality. This prevents for example that a router switches temporarily on a peer path while an alternate path from a customer is available but hidden at the border of the AS. Such a situation could result in a temporary withdrawal of the prefix on some eBGP sessions when the router selects the path via the peer.

The advertisement of the Second Local Preference occurs when there is no alternate path with the same quality as the best path. This way, fast convergence is still ensured. Backup path is optimal, as it has the second AS-Wide preference, which becomes the AS-wide best preference upon failure of the primary one.

Sending all the paths with a given Local Preference also has a positive impact on routing optimality. Indeed, this allows border

routers to have an increased path visibility and to choose their best path based on their own criteria.

The computational cost of this solution is reduced when there are several paths with the best local preference. In this case, it is sufficient to stop the decision process after the first rule to have the set of paths to be advertised. When it is necessary to advertise the paths with second local-preference, the additional cost is to apply a second time the first rule of the decision process, which is still reasonable. The memory cost depends on the number of paths with the best local preference.

A.3. Advertise Paths at decisive step -1

When the goal is to provide fast recovery by advertising candidate post-reconvergence paths, one can choose to stop the decision process just before the step where only one path remains. If the decision process comes to IGP tie-break, all remaining paths are advertised. This way, routers advertise as many paths as possible with a quality as similar as possible.

This path selection is an intermediary solution between the two preceding ones. Here, instead of stopping the decision process at the local preference step or the IGP step, we stop it before the rule that removes the best potential backup paths. This way, we minimize the number of paths to advertise while guaranteeing the presence of a backup path. Primary and backup path optimality is ensured, as all paths with the same AS-wide preference as the best paths are included in the set of paths advertised.

Authors' Addresses

Jim Uttaro
AT&T
200 S. Laurel Avenue
Middletown, NJ 07748 USA
Email: uttaro@att.com

Virginie Van den Schrieck
UCLouvain
Place Ste Barbe, 2
Louvain-la-Neuve 1348 BE
Email: virginie.vandenschrieck@uclouvain.be
URI: <http://inl.info.ucl.ac.be/vvandens>

Pierre Francois
UCLouvain
Place Ste Barbe, 2
Louvain-la-Neuve 1348 BE
Email: pierre.francois@uclouvain.be
URI: <http://inl.info.ucl.ac.be/pfr>

Roberto Fragassi
Alcatel-Lucent
600 Mountain Avenue
Murray Hill, New Jersey
Email: roberto.fragassi@alcatel-lucent.com

Adam Simpson
Alcatel-Lucent
600 March Road
Ottawa, Ontario K2K 2E6
Canada
Email: adam.simpson@alcatel-lucent.com

Pradosh Mohapatra
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134 USA
Email: pmohapat@cisco.com

