

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 17, 2014

P. Fan
Z. Li
China Mobile
February 13, 2014

IPv6 BGP Identifier Capability for BGP-4
draft-fan-idr-ipv6-bgp-id-00

Abstract

To solve the problem of BGP Identifier in an IPv6-only network without special configuration and planning considerations, this document extends BGP to allow a BGP Identifier to be a valid IPv6 global unicast address assigned to the BGP speaker. Protocol extension includes the definition of a BGP capability code, "IPv6 BGP Identifier capability", to be used by a BGP speaker to indicate its support for IPv6 address as a BGP Identifier. This document updates RFC 4271.

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 August 17, 2014.

Copyright Notice

Copyright (c) 2014 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	2
2. Protocol Extension	3
3. Operations	3
3.1. Processing Received OPEN messages	3
3.2. Connection Collision Detection	3
3.3. Route Selection Decision	3
4. Security Considerations	4
5. IANA Considerations	4
6. Normative References	4
Authors' Addresses	4

1. Introduction

The BGP Identifier of a BGP speaker is specified as a valid IPv4 host address assigned to the BGP speaker [RFC4271]. In addition, the deployed BGP code requires that two BGP speakers be of distinct BGP Identifiers in order to establish a BGP connection.

In an IPv6-only network, the requirements for the BGP Identifier in [RFC4271] are not met as no IPv4 address is configured to the BGP speaker. To accommodate this situation, [RFC6286] relaxes the definition of the BGP Identifier to be a 4-octet, unsigned, non-zero integer and requires AS-wide uniqueness of the BGP Identifiers.

The proposal in [RFC6286] requires that a 4-octet integer be configured as BGP Identifier on the BGP speaker after basic IPv6 information (such as IPv6 addresses) configuration has been completed. The 4-octet integer to be configured as BGP Identifier has to be planned carefully in advance to guarantee uniqueness within the AS. In a large IPv6-only AS the extra configuration and planning work introduced by the special integers can be troublesome.

To solve the problem of BGP Identifier in an IPv6-only network without special configuration and planning considerations, this document extends BGP to allow a BGP Identifier to be a valid IPv6 global unicast address assigned to the BGP speaker. Protocol extension includes the definition of a BGP capability code, "IPv6 BGP Identifier capability", to be used by a BGP speaker to indicate its support for IPv6 address as a BGP Identifier.

2. Protocol Extension

A new BGP capability [RFC5492] is defined to convey the IPv6 global unicast address to be used as the BGP Identifier. A BGP speaker uses BGP Capabilities Advertisements in its OPEN message to advertise its neighbors this IPv6 BGP Identifier ability. The BGP Identifier field of the OPEN message is set to zero, indicating that actual BGP ID is in the Capability Optional Parameter.

The Capability Length field of the IPv6 BGP Identifier Capability is set to 16, and the Capability value field is set to one of the IPv6 global unicast addresses that have been assigned to the BGP speaker.

The BGP Identifier is also used in the AGGREGATOR attribute, so a BGP speaker that uses IPv6 BGP Identifier Capability sets the AGGREGATOR attribute accordingly. The BGP Identifier carried in the attribute is encoded as a 16-octet entity.

3. Operations

3.1. Processing Received OPEN messages

A BGP speaker checks the BGP Identifier field of the OPEN message received first. If the BGP Identifier field is not zero, then the OPEN message is processed in the way of the message that does not contain IPv6 BGP Identifier, and any IPv6 BGP Identifier Capability in the Capability Optional Parameter of the message is ignored. If the BGP Identifier field is zero, then the BGP speaker checks if any IPv6 BGP Identifier Capability is carried in the Capability Optional Parameter. If there is no IPv6 BGP ID Capability, or the capability value of the IPv6 BGP ID Capability is not a valid IPv6 global unicast address, then a Notification message is generated, with Error Code set to 2 (OPEN Message Error) and Error subcode set to 3 (Bad BGP Identifier).

3.2. Connection Collision Detection

In case of collision detection, the BGP Identifiers of the peers involved in the collision are compared and only the connection initiated by the BGP speaker with the higher-valued BGP Identifier is retained. Comparing BGP Identifiers is done by converting them to host byte order and treating them as 16-octet unsigned integers.

3.3. Route Selection Decision

If a route is advertised by an IPv4 BGP speaker and an IPv6 BGP speaker respectively, then the route advertised by the IPv6 BGP speaker is selected; if a route is advertised by two IPv6 BGP

speakers respectively, then their IPv6 BGP IDs are compared, and the route advertised by the BGP speaker with the lower-valued BGP Identifier is selected.

4. Security Considerations

TBD.

5. IANA Considerations

This document requests a new BGP Capability Code to be allocated by IANA.

6. Normative References

- [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.
- [RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, June 2011.

Authors' Addresses

Peng Fan
China Mobile
32 Xuanwumen West Street, Xicheng District
Beijing 100053
P.R. China

Email: fanpeng@chinamobile.com

Zhenqiang Li
China Mobile
32 Xuanwumen West Street, Xicheng District
Beijing 100053
P.R. China

Email: lizhenqiang@chinamobile.com

Network Working Group
Internet-Draft
Expires: August 11, 2014

Pierre Francois
Camilo Cardona
Institute IMDEA Networks
Adam Simpson
Alcatel-Lucent
Jeffrey Haas
Juniper Networks
February 7, 2014

ADD-PATH limit capability
draft-francois-idr-addpath-limit-00

Abstract

In this draft, we propose a new capability that allows BGP speakers supporting ADD-PATH to announce a limit on the number of paths they want to receive from their peers.

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 August 11, 2014.

Copyright Notice

Copyright (c) 2014 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. The ADD-PATH Limit capability	3
3. Operation	4
4. References	5
Authors' Addresses	5

1. Introduction

ADD-PATH is an extension to BGP that allows a BGP speaker to send and receive more than one path for the same NLRI [AddPath]. The ADD-PATH capability does not include any mechanism for restricting the number or type of paths that a peer can receive from others. Thus, a receiving BGP speaker has no control on the number of paths sent by its peer, which depends only on the mode the operator desires to implement in the transmitting side [AddPathGuidelines]. This restriction can make network operations more difficult in some situations:

- o In cases in which all network devices are managed by the same group, operators select the ADD-PATH mode that best fits the resources of each of their devices. Operators must configure manually each speaker with a mode that does not overload the resources of the other devices. The overhead of this procedure can be high in some networks, as this configuration must be performed at the session level. If a ADD-PATH router could signal a limit in the number of paths it wants to receive, operators could achieve the same resource control by performing a more simple configuration.
- o In cases in which devices are managed by different operators, such as in networks spanning large geographical regions or in Internet Exchange Points, operators must agree on the ADD-PATH mode to use between BGP speakers. If one of the devices has constraints on the number of paths it can receive, the other party must configure an ADD-PATH mode that does not overload the memory of other device. Under these circumstances, allowing the receiving side to limit the number of paths can ease the management process for all administration sides.

In this document, we propose a new capability that allows an ADD-PATH capable BGP speaker to set an explicit upperbound on the number of paths it wants to receive from its peer.

2. The ADD-PATH Limit capability

```
+-----+
| Address Family Identifier (2 octets) |
+-----+
| Subsequent Address Family Identifier (1 octet) |
+-----+
| Flags(1 octet) |
+-----+
| Receive bound (2 octet) |
+-----+
```

Figure 1: Capability

The meaning and use of the fields are as follows:

Address Family Identifier (AFI):

This field is the same as the one used in [RFC4760].

Subsequent Address Family Identifier (SAFI):

This field is the same as the one used in [RFC4760].

Flag Field

This field contains different bit flags related to the ADD-PATH limitation capability.

The most significant bit of the field (limit-capable bit) is used to communicate if the device supports the limitation of paths announced to the peer. If the router supports the ADD-PATH Limit capability, but it is not capable of limiting the number of announced paths, it should set the limit-capable bit to 0.

The remaining bits MUST be set to 0 and SHOULD be ignored upon receipt.

Receive Bound

This field indicates the maximum number of paths the device wants to receive from a peer for the <AFI, SAFI>. If the field is set to 0, the device has no restriction on the number of paths it can receive.

3. Operation

The ADD-PATH Limit capability SHOULD be announced by BGP speakers that require their neighbors to send a limited number of paths per prefix. A router that is capable of sending a limited number of paths to a BGP ADD-PATH neighbor SHOULD also announce the ADD-PATH Limit capability. For cases in which a router is not capable of setting a limit in the number of paths it sends to a peer, it should set the limit-capable bit in the add-path capability to 0.

The ADD-PATH capability is a prerequisite of the ADD-PATH Limit. A BGP peer SHOULD ignore the ADD-PATH Limit capability from a peer that did not also announce the ADD-PATH capability.

The ADD-PATH limit capability is used to set an upper bound on the

number of paths that the router wants to receives from a peer. A sender capable of limiting the number of paths per peer SHOULD NOT send more paths than requested by the receiver.

It might be impractical for a BGP speaker to strictly stick to each of the upper bounds specified by its peers. Thus, the sender MAY send a lower amount of paths than the upper bound indicated by its peer.

A router SHOULD include the best path in the subset of paths to send to a peer, except when the best path is received from that peer. The choice of the rest of paths to be sent is left free to the implementation.

A BGP speaker could receive more paths than the number defined in the ADD-PATH capability, even when the BGP peer supports the limitation of paths. This event should be logged, but the session with the peer should be preserved. The receiving speaker should implement the required mechanism to deal with more paths that it can support.

4. References

[AddPath] D. Walton, E. Chen, A. Retana, and J. Scudder,
"Advertisement of Multiple Paths in BGP",
draft-ietf-idr-add-paths-09.txt (work in progress).

[AddPathGuidelines]
J. Uttaro, P. Francois, R. Fragassi, A. Simpson, and K.
Patel, "Best Practices for Advertisement of Multiple Paths
in IBGP", draft-ietf-idr-add-paths-guidelines-05.txt (work
in progress).

Authors' Addresses

Pierre Francois
Institute IMDEA Networks
Avda. del Mar Mediterraneo, 22
Leganes 28918
ES

Email: pierre.francois@imdea.org

Camilo Cardona
Institute IMDEA Networks
Avda. del Mar Mediterraneo, 22
Leganes 28918
ES

Email: juancamilo.cardona@imdea.org

Adam Simpson
Alcatel-Lucent
600 March Road
Ontario K2K 2E6
CA

Email: adam.simpson@alcatel-lucent.com

Jeffrey Haas
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale 94089
USA

Email: jhaas@juniper.net

Network Working Group
Internet-Draft
Expires: August 11, 2014

Pierre Francois
Camilo Cardona
Institute IMDEA Networks
Adam Simpson
Alcatel-Lucent
Jeffrey Haas
Juniper Networks
February 7, 2014

ADD-PATH for Route Servers
draft-francois-idr-rs-addpaths-00

Abstract

BGP speakers in Internet Exchange Points exchange routes with a large number of peers. To reduce the burden of maintaining many sessions, IXPs implement and administrate BGP route servers. Route servers announce to their clients the paths of multiple peers by using a single eBGP session. Route servers, however, are restricted to propagating a single path per NLRI per eBGP session. This constraint affects the path diversity received by clients, which could use paths that they would not have chosen, had they known all possible paths. To overcome this limitation, we propose in this draft the extension of ADD-PATH to eBGP peers in the context of route servers.

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 August 11, 2014.

Copyright Notice

Copyright (c) 2014 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. Motivation	3
3. Operation of eBGP ADD-PATH capability for IXP route Server . .	4
3.1. Operation of eBGP ADD-PATH in route servers with the limited paths capability	4
4. Error conditions	4
5. IANA considerations	5
6. Security Considerations	5
7. References	5
Authors' Addresses	6

1. Introduction

IXP route servers were designed to help network operators reduce the difficulties associated with maintaining a large number of sessions [IXPRouteServer]. Every route server client can receive paths from multiple ASes using the same eBGP session with the route server. In some cases, usually when there are many members in the IXP, multiple clients might announce a path to the same NLRI. Path diversity is an advantage for IXPs, as members can choose the path that better suits their policy. However, as a normal eBGP speaker, route servers can only advertise a single path to its clients. This limitation causes the route server to potentially hide paths that would be useful for their clients.

ADD-PATH [AddPath] is a capability that allows BGP speakers to announce more than one path to their peers. Works related to ADD-PATH have focused on applications for iBGP deployments. We propose the use of ADD-PATH over eBGP sessions to overcome the problems associated to the limit in the number of paths that route servers can announce. In this document, we define the operation and error conditions of ADD-PATH for these scenarios and describe additional benefits for the route servers that implement it.

2. Motivation

By collecting paths from all their clients, route servers potentially accumulate various paths for some destination prefix. Multiple of these paths may be compliant with the policy of some client of the route server. However, route servers typically maintain a single session with their clients, and hence advertise at most a single path towards each of them. As a result, a route server client will typically know only one of these paths.

We believe that this aspect of route serving is an unfortunate limitation, as it artificially hides paths from clients that may have wanted to use them.

First, it prevents the member from performing a policy based decision that is finer than the one advertised to the route server platform. That is, the arbitrary best path picked among the policy-compliant ones by the route server may be actually different from the one that the client would have picked, had it known about all of them.

Second, it prevents the member from doing temporary preference tweaking among the set of available paths in order to perform traffic engineering. That is, a member may only receive a path for a destination through a peer that is saturated, while alternate paths

through non saturated nexthops are available and would have been used if the router (and the operator) were aware of their existence.

ADD-PATH was designed to advertise more than one path towards a given NLRI. Multiple paths installed in the forwarding planes, as well as alternate paths, can be advertised among speakers supporting ADD-PATH. ADD-PATH can be used by a route server to announce all paths available for the same NLRI that still fulfill the policy of the route server client.

3. Operation of eBGP ADD-PATH capability for IXP route Server

A route server that supports the advertisement of multiple paths toward the same NLRI SHOULD announce the ADD-PATH capability to its clients. Likewise, a client supporting the reception of multiple paths SHOULD announce the ADD-PATH capability to the route server.

In an IXP context, only the route server should propagate multiple paths to the route server clients. The advertisement of multiple paths in the other direction is currently out of the specification of this document. Therefore, a route Server client should set the Send/Receive field for the Add-Path capability with a value of 1. The route Server should set the same field in the capability with a 2.

A route server supporting ADD-PATH can announce to its clients all paths that comply with their policy. This operational mode is similar to the ADD-PATH ALL mode described in [AddPathGuidelines].

A route server could also support other type of ADD-PATH modes that restrict the paths announced to the client. In an Add-N mode, for instance, the route server would announce at most N paths to their clients.

3.1. Operation of eBGP ADD-PATH in route servers with the limited paths capability

The limited paths capability provides ADD-PATH speakers with a method to communicate the maximum number of paths towards the same NLRI that BGP speakers are willing to receive. The use of this capability in IXP environments is recommended, as it provides the clients with the ability to control the resources used in their devices by limiting the total amount of paths received from the route Server.

4. Error conditions

In the specific context of route servers, third party nexthops are

being used so as to have the client actually be able to select the appropriate nexthop. This is achieved by letting the route server leave the nexthop field of the propagated paths unchanged.

Similarly, the propagation of multiple paths by the route server to one of its clients must be made in a way that allows the receiver to actually select one among those paths. As a result, a route server advertising two different paths for the same destination, with equal nexthops, is out of specification. If this situation occurs, the client SHOULD log the event and let the normal decision process decide the best path.

A typical route server client will have only one usable path towards a given destination announced to the other clients of the route server. As a result, a route server client advertising more than one path towards a given destination, to its route server, is out of specification. If this situation occurs, the client SHOULD log the event and let the normal decision process decide the best path.

5. IANA considerations

None

6. Security Considerations

The use of eBGP ADD-PATH in the route server environment does not increase the number of destinations for which paths are being advertised. However, the potential number of paths per destination is now larger than one, potentially increasing the memory load of the Adj-Rib-In. Systems risking to be short on memory due to this increase should be configured to constrain the amount of paths being advertised to them by a value which ensures proper operations.

7. References

[AddPath] D. Walton, E. Chen, A. Retana, and J. Scudder,
"Advertisement of Multiple Paths in BGP",
draft-ietf-idr-add-paths-06.txt (work in progress).

[AddPathGuidelines]
J. Uttaro, P. Francois, R. Fragassi, A. Simpson, and K.
Patel, "Best Practices for Advertisement of Multiple Paths
in IBGP", draft-ietf-idr-add-paths-guidelines-05.txt (work
in progress).

[IXPRouteServer]

E. Jasinska, N. Hilliard, R. Raszuk, and N. Bakker, "Best Practices for Advertisement of Multiple Paths in IBGP", draft-ietf-idr-ix-bgp-route-server-03 (work in progress).

Authors' Addresses

Pierre Francois
Institute IMDEA Networks
Avda. del Mar Mediterraneo, 22
Leganes 28918
ES

Email: pierre.francois@imdea.org

Camilo Cardona
Institute IMDEA Networks
Avda. del Mar Mediterraneo, 22
Leganes 28918
ES

Email: juancamilo.cardona@imdea.org

Adam Simpson
Alcatel-Lucent
600 March Road
Ontario K2K 2E6
CA

Email: adam.simpson@alcatel-lucent.com

Jeffrey Haas
Juniper Networks
1194 N. Mathilda Ave
Sunnyvale 94089
USA

Email: jhaas@juniper.net

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: August 18, 2014

H. Gredler, Ed.
Juniper Networks, Inc.
S. Ray, Ed.
S. Previdi
C. Filsfils
Cisco Systems, Inc.
M. Chen
Huawei Technologies
J. Tantsura
Ericsson
February 14, 2014

BGP Link-State extensions for Segment Routing
draft-gredler-idr-bgp-ls-segment-routing-extension-01

Abstract

Segment Routing (SR) allows for a flexible definition of end-to-end paths within link-state graphs by encoding paths as sequences of topological sub-paths, called "segments".

The link-state routing protocols (IS-IS, OSPF and OSPFv3) have been extended to advertise the segments. But flooding based propagation of path segments using IGPs is limited by the perimeter of the IGP domain. For building paths which span across IGP domains (e.g. multiple ASes), the Border Gateway Protocol (BGP) is better suited as its propagation perimeter is not limited like the IGPs.

This draft defines extensions to the BGP Link-state address-family to carry path segment information via BGP.

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 August 18, 2014.

Copyright Notice

Copyright (c) 2014 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. BGP-LS Extensions for Segment Routing	6
2.1. Node Attribute TLVs	7
2.2. Link Attribute TLVs	7
2.3. Prefix Attribute TLVs	8
3. IANA Considerations	8
4. Manageability Considerations	8
4.1. Operational Considerations	8
4.1.1. Operations	8
5. TLV/Sub-TLV Code Points Summary	8
6. Security Considerations	9
7. Acknowledgements	9
8. References	9
8.1. Normative References	9
8.2. Informative References	10
Authors' Addresses	11

1. Introduction

Segment Routing (SR) allows for a flexible definition of end-to-end paths within link-state topologies by encoding paths as sequences of topological sub-paths, called "segments". Segment routing is an amalgamation of source routing and MPLS. In Segment Routing, the ingress node prepends a sequence of instructions, called "segments", to the packet. The SR capable nodes sequentially execute the instructions on the packet to achieve packet forwarding via required topological paths as well as service paths.

The segments can be thought of, in a simple way, to represent instructions such as "go to node N using the shortest path", "follow the shortest path for prefix P", "use link/node/ERO L", etc. Each segment is identified by a 32 bit Segment Identifier (SID) (when unmodified MPLS data-plane is used, the SIDs are restricted to 20 bits). There are "global" segments that are known to all SR nodes in the local domain, and there are local segments whose semantics are known only to the nodes that originate them. The segment routing architecture is described in [I-D.filsfils-rtgwg-segment-routing] and segment routing use-cases in [I-D.filsfils-rtgwg-segment-routing-use-cases].

Segment routing is enabled in a network by advertising the segments (including the associated SIDs) to the nodes in the network. The IGP link-state routing protocols (IS-IS [I-D.previdi-isis-segment-routing-extensions], OSPFv2 [I-D.psenak-ospf-segment-routing-extensions] and OSPFv3 [I-D.psenak-ospf-segment-routing-ospfv3-extension]) have been extended to advertise the segments. Using these extensions, segment routing can be enabled within an IGP domain.

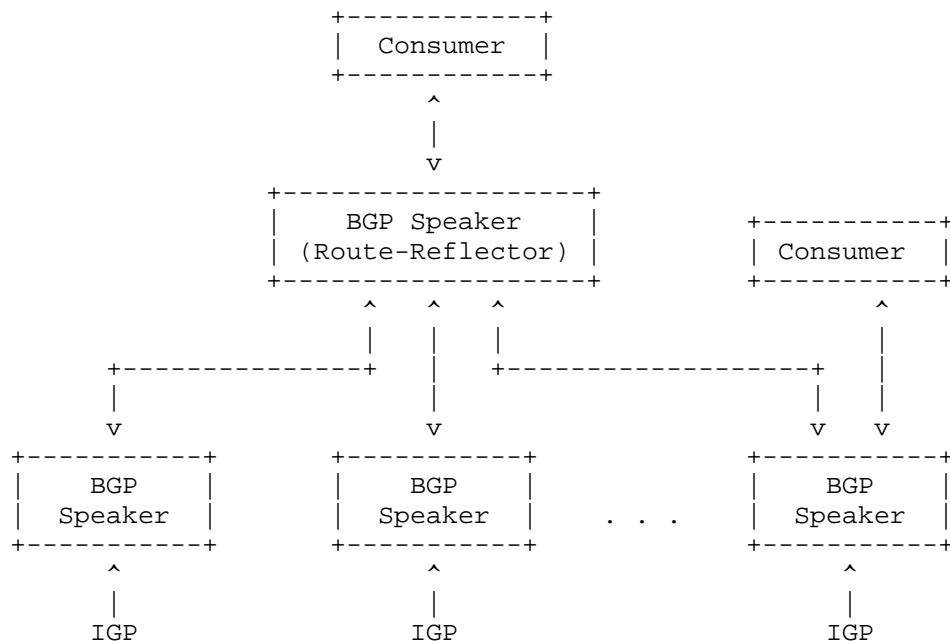


Figure 1: Link State info collection

Segment Routing (SR) allows advertisement of single or multi-hop paths. The flooding scope for the IGP extensions for Segment routing is IGP area-wide. Consequently, the contents of a Link State Database (LSDB) or a Traffic Engineering Database (TED) has the scope of an IGP area and therefore by using the IGP alone it is not possible to construct segments across an IGP Area or AS boundaries.

To address the need for applications that require visibility into LSDB across IGP areas, or even across ASes, the BGP-LS address-family/sub-address-family have been defined that allows BGP to carry LSDB information. The BGP Network Layer Reachability Information (NLRI) encoding format for BGP-LS and a new BGP Path Attribute called BGP-LS attribute are defined in [I-D.ietf-idr-ls-distribution]. The identifying key of each LSDB object, namely a node, a link or a prefix, is encoded in the NLRI and the properties of the object are encoded in the BGP-LS attribute. Figure Figure 1 describes a typical deployment scenario. In each IGP area, one or more nodes are configured with BGP-LS. These BGP speakers form an IBGP mesh by connecting to one or more route-reflectors. This way, all BGP speakers - specifically the route-reflectors - obtain LSDB information from all IGP areas (and from other ASes from EBGP peers). An external component connects to the route-reflector to obtain this information (perhaps moderated by a policy regarding what information

is sent to the external component, and what information isn't).

This document describes extensions to BGP-LS to carry the segments. An external component - a Controller - then can collect segment information in the "northbound direction" across IGP areas/autonomous systems and construct the segment stack that need to be added to an incoming packet to achieve the desired end-to-end forwarding.

2. BGP-LS Extensions for Segment Routing

The BGP-LS NLRI can be a node NLRI, a link NLRI or a prefix NLRI. The corresponding BGP-LS attribute is a node attribute, a link attribute or a prefix attribute. BGP-LS [I-D.ietf-idr-ls-distribution] defines the TLVs that map link-state information to BGP-LS NLRI and BGP-LS attribute. This document adds additional BGP-LS attribute TLVs to encode SR information.

[I-D.previdi-isis-segment-routing-extensions] defines the following TLVs to encode SR information.

- o TLV for Prefix-SID
- o TLV for Adjacency-SID between two nodes as well as between nodes in a LAN
- o TLV for SID/Label binding for advertising paths from other protocols (and their optional ERO)
- o TLV for SR Capabilities
- o TLV for SR Algorithm

These TLVs are mapped to BGP-LS attribute TLVs in the following way.

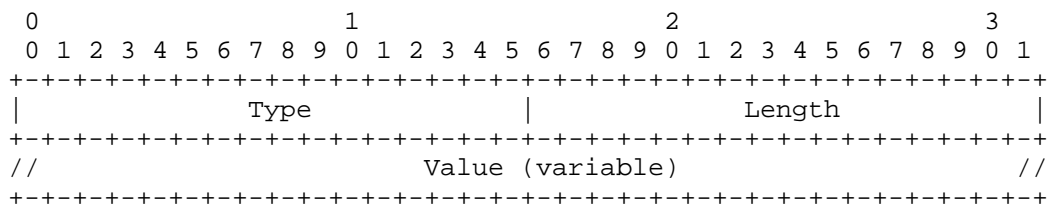


Figure 2: TLV format

The 2 octet Type field values are defined in Table 1, Table 2, and Table 3. The next 2 octet Length field encodes length of the rest of the TLV. The Value portion of the TLV is variable and is equal to

the corresponding Value portion of the TLV defined in [I-D.previdi-isis-segment-routing-extensions].

In each case, multiple TLVs for a given type are allowed to be added. The semantics of multiple such values are determined by [I-D.previdi-isis-segment-routing-extensions].

2.1. Node Attribute TLVs

The following 'Node Attribute' TLVs are defined:

TLV Code Point	Description	Length	IS-IS SR TLV/sub-TLV
1033	SID/Label Binding	variable	149
1034	SR Capabilities	variable	2
1035	SR Algorithm	variable	15

Table 1: Node Attribute TLVs

The sections refer to [I-D.previdi-isis-segment-routing-extensions].

These TLVs can ONLY be added to the Node Attribute associated with the Node NLRI that originates the corresponding SR TLV.

2.2. Link Attribute TLVs

The following 'Link Attribute' TLVs are defined:

TLV Code Point	Description	Length	IS-IS SR TLV/sub-TLV
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	31
1100	LAN Adjacency Segment Identifier (Adj-SID) TLV	variable	32

Table 2: Link Attribute TLVs

The sections refer to [I-D.previdi-isis-segment-routing-extensions].

These TLVs can ONLY be added to the Link Attribute associated with the link whose local node originates the corresponding SR TLV.

For a LAN, normally a node only announces its adjacency to the pseudo-node. [I-D.previdi-isis-segment-routing-extensions] allows a node to announce adjacency to all other nodes attached to the LAN. In such a case, the corresponding BGP-LS link NLRI must be originated for each additional link in order to add the SR TLVs to the Link Attribute.

2.3. Prefix Attribute TLVs

The following 'Prefix Attribute' TLVs are defined:

TLV Code Point	Description	Length	IS-IS SR TLV/sub-TLV
1158	Prefix SID	variable	3

Table 3: Prefix Attribute TLVs

The sections refer to [I-D.previdi-isis-segment-routing-extensions].

These TLVs can ONLY be added to the Prefix Attribute whose local node in the corresponding prefix NLRI is the node that originates the corresponding SR TLV.

3. IANA Considerations

This document requests assigning code-points from the registry for BGP-LS attribute TLVs based on table Table 4.

4. Manageability Considerations

This section is structured as recommended in [RFC5706].

4.1. Operational Considerations

4.1.1. Operations

Existing BGP and BGP-LS operational procedures apply. No new operation procedures are defined in this document.

5. TLV/Sub-TLV Code Points Summary

This section contains the global table of all TLVs/Sub-TLVs defined in this document.

TLV Code Point	Description	Length	IS-IS SR TLV/sub-TLV
1033	SID/Label Binding	variable	149
1034	SR Capabilities	variable	2
1035	SR Algorithm	variable	15
1099	Adjacency Segment Identifier (Adj-SID) TLV	variable	31
1100	LAN Adjacency Segment Identifier (Adj-SID) TLV	variable	32
1158	Prefix SID	variable	3

Table 4: Summary Table of TLV/Sub-TLV Codepoints

6. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See the 'Security Considerations' section of [RFC4271] for a discussion of BGP security. Also refer to [RFC4272] and [RFC6952] for analysis of security issues for BGP.

7. Acknowledgements

TBD.

8. References

8.1. Normative References

[I-D.ietf-idr-ls-distribution]

Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution-04 (work in progress), November 2013.

[I-D.previdi-isis-segment-routing-extensions]

Previdi, S., Filsfils, C., Bashandy, A., Gredler, H., and S. Litkowski, "IS-IS Extensions for Segment Routing", draft-previdi-isis-segment-routing-extensions-04 (work in progress), October 2013.

[I-D.psenak-ospf-segment-routing-extensions]

Psenak, P., Previdi, S., Filsfils, C., Gredler, H.,

Shakir, R., and W. Henderickx, "OSPF Extensions for Segment Routing",
draft-psenak-ospf-segment-routing-extensions-03 (work in progress), October 2013.

[I-D.psenak-ospf-segment-routing-ospfv3-extension]
Psenak, P., Previdi, S., Filsfils, C., Gredler, H.,
Shakir, R., and W. Henderickx, "OSPFv3 Extensions for Segment Routing",
draft-psenak-ospf-segment-routing-ospfv3-extension-00 (work in progress), October 2013.

[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.

8.2. Informative References

[I-D.filsfils-rtgwg-segment-routing]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture",
draft-filsfils-rtgwg-segment-routing-01 (work in progress), October 2013.

[I-D.filsfils-rtgwg-segment-routing-use-cases]
Filsfils, C., Francois, P., Previdi, S., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., Kini, S., and E. Crabbe, "Segment Routing Use Cases",
draft-filsfils-rtgwg-segment-routing-use-cases-02 (work in progress), October 2013.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, January 2006.

[RFC5706] Harrington, D., "Guidelines for Considering Operations and Management of New Protocols and Protocol Extensions", RFC 5706, November 2009.

[RFC6952] Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP, and MSDP Issues According to the Keying and Authentication for Routing Protocols (KARP) Design Guide", RFC 6952, May 2013.

Authors' Addresses

Hannes Gredler (editor)
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: hannes@juniper.net

Saikat Ray (editor)
Cisco Systems, Inc.
170, West Tasman Drive
San Jose, CA 95134
US

Email: sairay@cisco.com

Stefano Previdi
Cisco Systems, Inc.
Via Del Serafico, 200
Rome 00142
Italy

Email: sprividi@cisco.com

Clarence Filsfils
Cisco Systems, Inc.
Brussels
Belgium

Email: cfilsfil@cisco.com

Mach(Guoyi) Chen
Huawei Technologies
Huawei Building, No. 156 Beiqing Rd.
Beijing 100095
China

Email: mach.chen@huawei.com

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
US

Email: jeff.tantsura@ericsson.com

Inter-Domain Routing
Internet-Draft
Intended status: Informational
Expires: August 18, 2014

H. Gredler, Ed.
B. Rajagopalan
Juniper Networks, Inc.
S. Ray, Ed.
M. Bhardwaj
Cisco Systems, Inc.
February 14, 2014

BGP Link-State Information Distribution Implementation Report
draft-gredler-idr-ls-distribution-impl-00

Abstract

This document is an implementation report for the BGP Link-State Information Distribution protocol as defined in [I-D.ietf-idr-ls-distribution]. The editors did not verify the accuracy of the information provided by respondents. The respondents are experts with the implementations they reported on, and their responses are considered authoritative for the implementations for which their responses represent. Respondents were asked to only use the YES answer if the feature had at least been tested in the lab.

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 August 18, 2014.

Copyright Notice

Copyright (c) 2014 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. Implementation Forms	3
3. NLRI subtypes	3
4. Link NLRI TLV support	4
5. Node NLRI TLV support	7
6. Prefix NLRI TLV support	8
7. Interoperable Implementations	9
7.1. Cisco Implementation	9
7.2. Juniper Implementation	10
7.3. TBD Implementation	10
8. IANA Considerations	10
9. Security considerations	10
10. Acknowledgements	10
11. Informative References	10
Authors' Addresses	10

1. Introduction

In order to share network link-state and traffic engineering information collected with external components using the BGP routing protocol a new BGP Network Layer Reachability Information (NLRI) encoding format is required.

This document provides an implementation report for the BGP Link-State Information Distribution NLRI Format as defined in [I-D.ietf-idr-ls-distribution].

The editors did not verify the accuracy of the information provided by respondents or by any alternative means. The respondents are experts with the implementations they reported on, and their responses are considered authoritative for the implementations for which their responses represent. Respondents were asked to only use the YES answer if the feature had at least been tested in the lab.

2. Implementation Forms

Contact and implementation information for person filling out this form:

IOS-XR

Name: Manish Bhardwaj
Email: manbhard@cisco.com
Vendor: Cisco Systems, Inc.
Release: IOS-XR
Protocol Role: Sender, Receiver

JUNOS

Name: Balaji Rajagopalan
Email: balajir@juniper.net
Vendor: Juniper Networks, Inc.
Release: JUNOS
Protocol Role: Sender, Receiver

3. NLRI subtypes

Does the implementation support the Network Layer Reachability (NLRI) subtypes as described in Section 3.2 of [I-D.ietf-idr-ls-distribution] ?

N1: Node NLRI

N2: Link NLRI

N3: IPv4 Topology Prefix NLRI

N4: IPv6 Topology Prefix NLRI

	IOS-XR	JUNOS	TBD
Rcv.N1	YES	YES	---
Snd.N1	YES	YES	---
Rcv.N2	YES	YES	---
Snd.N2	YES	YES	---
Rcv.N3	YES	NO(1)	---
Snd.N3	YES	NO(1)	---
Rcv.N4	YES	NO(1)	---
Snd.N4	YES	NO(1)	---

Note 1: Topology Prefix NLRIs get transparently relayed.

4. Link NLRI TLV support

Does the implementation support the TLVs described in Section 7 of [I-D.ietf-idr-ls-distribution] ?

TLV 256: Local Node Descriptor

TLV 257: Remote Node Descriptor

TLV 258: Link Local/Remote Identifier

TLV 259: IPv4 Interface address

TLV 260: IPv4 Neighbor address

TLV 261: IPv6 Interface address

TLV 262: IPv6 Neighbor address

TLV 263: Multi-Topology IDs

TLV 512: Autonomous System
 TLV 513: BGP-LS Identifier
 TLV 514: Area ID
 TLV 515: IGP Router ID
 TLV 1028: IPv4 router-ID of Local Node
 TLV 1029: IPv6 router-ID of Local Node
 TLV 1030: IPv4 router-ID of Remote Node
 TLV 1031: IPv6 router-ID of Remote Node
 TLV 1088: Administrative group (color)
 TLV 1089: Maximum link bandwidth
 TLV 1090: Maximum reservable link bandwidth
 TLV 1091: Unreserved link bandwidth
 TLV 1092: TE default Metric
 TLV 1093: Link Protection Type
 TLV 1094: MPLS Protocol Mask
 TLV 1095: IGP Metric
 TLV 1096: Shared Risk Link Group
 TLV 1097: Opaque Link attribute
 TLV 1098: Link name attribute

	IOS-XR	JUNOS	TBD
Rcv.TLV 256	YES	YES	---
Snd.TLV 256	YES	YES	---
Rcv.TLV 257	YES	YES	---
Snd.TLV 257	YES	YES	---
Rcv.TLV 258	YES	YES	---
Snd.TLV 258	YES	YES	---
Rcv.TLV 259	YES	YES	---

Snd.TLV 259	YES	YES	---
Rcv.TLV 260	YES	YES	---
Snd.TLV 260	YES	YES	---
Rcv.TLV 261	YES	YES	---
Snd.TLV 261	YES	YES	---
Rcv.TLV 262	YES	YES	---
Snd.TLV 262	YES	YES	---
Rcv.TLV 263	---	NO	---
Snd.TLV 263	---	NO	---
Rcv.TLV 512	YES	YES	---
Snd.TLV 512	YES	YES	---
Rcv.TLV 513	---	YES	---
Snd.TLV 513	---	NO	---
Rcv.TLV 514	---	YES	---
Snd.TLV 514	---	NO	---
Rcv.TLV 515	YES	YES	---
Snd.TLV 515	YES	YES	---
Rcv.TLV 1028	YES	YES	---
Snd.TLV 1028	YES	YES	---
Rcv.TLV 1029	YES	YES	---
Snd.TLV 1029	YES	YES	---
Rcv.TLV 1030	YES	YES	---
Snd.TLV 1030	YES	YES	---
Rcv.TLV 1031	YES	YES	---
Snd.TLV 1031	YES	YES	---
Rcv.TLV 1088	YES	YES	---
Snd.TLV 1088	YES	YES	---
Rcv.TLV 1089	YES	YES	---
Snd.TLV 1089	YES	YES	---
Rcv.TLV 1090	---	YES	---
Snd.TLV 1090	---	YES	---
Rcv.TLV 1091	---	YES	---
Snd.TLV 1091	---	YES	---
Rcv.TLV 1092	---	YES	---
Snd.TLV 1092	---	YES	---
Rcv.TLV 1093	---	NO	---
Snd.TLV 1093	---	NO	---
Rcv.TLV 1094	NO	NO	---
Snd.TLV 1094	NO	NO	---
Rcv.TLV 1095	---	NO	---
Snd.TLV 1095	---	NO	---
Rcv.TLV 1096	YES	YES	---
Snd.TLV 1096	---	YES	---
Rcv.TLV 1097	---	YES	---
Snd.TLV 1097	---	NO	---
Rcv.TLV 1098	NO	NO	---
Snd.TLV 1098	NO	NO	---

5. Node NLRI TLV support

Does the implementation support the TLVs described in Section 7 of [I-D.ietf-idr-ls-distribution] ?

TLV 256: Local Node Descriptor

TLV 263: Multi-Topology IDs

TLV 512: Autonomous System

TLV 513: BGP-LS Identifier

TLV 514: Area ID

TLV 515: IGP Router ID

TLV 1024: Node flag bits

TLV 1025: Opaque Node properties

TLV 1026: Node name

TLV 1027: IS-IS Area Identifier

TLV 1028: IPv4 router-ID of Local Node

TLV 1029: IPv6 router-ID of Local Node

	IOS-XR	JUNOS	TBD
Rcv.TLV 256	YES	YES	---
Snd.TLV 256	YES	YES	---
Rcv.TLV 263	---	NO	---
Snd.TLV 263	---	NO	---
Rcv.TLV 512	YES	YES	---
Snd.TLV 512	YES	YES	---
Rcv.TLV 513	---	YES	---
Snd.TLV 513	---	NO	---
Rcv.TLV 514	---	YES	---
Snd.TLV 514	---	NO	---
Rcv.TLV 515	YES	YES	---
Snd.TLV 515	YES	YES	---
Rcv.TLV 1024	YES	NO	---
Snd.TLV 1024	YES	NO	---
Rcv.TLV 1025	---	NO	---
Snd.TLV 1025	---	NO	---
Rcv.TLV 1026	---	NO	---
Snd.TLV 1026	---	NO	---
Rcv.TLV 1027	---	NO	---
Snd.TLV 1027	---	NO	---
Rcv.TLV 1028	YES	YES	---
Snd.TLV 1028	YES	YES	---
Rcv.TLV 1029	YES	YES	---
Snd.TLV 1029	YES	YES	---

6. Prefix NLRI TLV support

Does the implementation support the TLVs described in Section 7 of [I-D.ietf-idr-ls-distribution] ?

TLV 256: Local Node Descriptor

TLV 263: Multi-Topology IDs

TLV 264: OSPF route type

TLV 265: IP Reachability information

TLV 1152: IGP Flags

TLV 1153: Route Tag

TLV 1154: Extended Tag

TLV 1155: Prefix Metric

TLV 1156: OSPF Forwarding Address

TLV 1157: Opaque Prefix Attribute

	IOS-XR	JUNOS	TBD
Rcv.TLV 256	YES	NO	---
Snd.TLV 256	YES	NO	---
Rcv.TLV 263	YES	NO	---
Snd.TLV 263	YES	NO	---
Rcv.TLV 264	YES	NO	---
Snd.TLV 264	YES	NO	---
Rcv.TLV 265	YES	NO	---
Snd.TLV 265	YES	NO	---
Rcv.TLV 1152	YES	NO	---
Snd.TLV 1152	YES	NO	---
Rcv.TLV 1153	YES	NO	---
Snd.TLV 1153	YES	NO	---
Rcv.TLV 1154	YES	NO	---
Snd.TLV 1154	YES	NO	---
Rcv.TLV 1155	YES	NO	---
Snd.TLV 1155	YES	NO	---
Rcv.TLV 1156	YES	NO	---
Snd.TLV 1156	YES	NO	---
Rcv.TLV 1157	---	NO	---
Snd.TLV 1157	---	NO	---

7. Interoperable Implementations

List other implementations that you have tested interoperability of BGP-LS Protocol Implementation.

7.1. Cisco Implementation

Cisco: The Cisco Systems, Inc. IOS-XR implementation should be interoperable with other vendor BGP-LS Protocol implementations. In particular we have tested our interoperability with Juniper's JUNOS and Telefonica's XXX implementation.

7.2. Juniper Implementation

Juniper: The Juniper Networks, Inc. JUNOS implementation should be interoperable with other vendor BGP-LS Protocol implementations. In particular we have tested our interoperability with Cisco Systems, Inc. IOS-XR implementation.

7.3. TBD Implementation

TBD: The TBD implementation has been tested by us with other implementations. It was so buggy that we were rolling on the floor laughing. We think this was either due to bad star alignment or perhaps increased solar flare activity.

8. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: The IANA has requested that this section remain in the document upon publication as an RFC. This note to the RFC Editor, however, may be removed.

9. Security considerations

No new security issues are introduced to the BGP Link-State Information Distribution Protocol defined in [I-D.ietf-idr-ls-distribution].

10. Acknowledgements

The authors would like to thank Stefano Previdi, Jan Medved for their contributions to this document.

11. Informative References

[I-D.ietf-idr-ls-distribution]
Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", draft-ietf-idr-ls-distribution-04 (work in progress), November 2013.

Authors' Addresses

Hannes Gredler (editor)
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: hannes@juniper.net

Balaji Rajagopalan
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: balajir@juniper.net

Saikat Ray (editor)
Cisco Systems, Inc.
170, West Tasman Drive
San Jose, CA 95134
US

Email: sairay@cisco.com

Manish Bhardwaj
Cisco Systems, Inc.
170, West Tasman Drive
San Jose, CA 95134
US

Email: manbhard@cisco.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: August 18, 2014

W. George
Time Warner Cable
S. Amante
Apple, Inc.
February 14, 2014

Autonomous System (AS) Migration Features and Their Effects on the BGP
AS_PATH Attribute
draft-ietf-idr-as-migration-00

Abstract

This draft discusses common methods of managing an ASN migration using some BGP features that while commonly-used are not formally part of the BGP4 protocol specification and may be vendor-specific in exact implementation. It is necessary to document these de facto standards to ensure that they are properly supported in future BGP protocol work such as BGPsec.

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 August 18, 2014.

Copyright Notice

Copyright (c) 2014 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	2
1.1. Documentation note	3
2. ASN Migration Scenario Overview	3
3. External BGP Autonomous System Migration Features	5
3.1. Local AS: Modify Inbound BGP AS_PATH Attribute	6
3.2. Replace AS: Modify Outbound BGP AS_PATH Attribute	7
4. Internal BGP Autonomous System Migration Features	8
4.1. Internal BGP Alias	9
5. Additional Operational Considerations	11
6. Conclusion	12
7. Acknowledgements	13
8. IANA Considerations	13
9. Security Considerations	13
10. References	13
10.1. Normative References	13
10.2. Informative References	13
Authors' Addresses	14

1. Introduction

This draft discusses common methods of managing an ASN migration using some BGP features that while commonly-used are not formally part of the BGP4 [RFC4271] protocol specification and may be vendor-specific in exact implementation. This draft does not attempt to standardize these features, because they are local to a given implementation and do not require negotiation with or cooperation of BGP neighbors. The deployment of these features do not need to interwork with one another to accomplish the desired results. However, it is necessary to document these de facto standards to ensure that any future protocol enhancements to BGP that propose to read, copy, manipulate or compare the AS_PATH attribute can do so without inhibiting the use of these very widely used ASN migration features.

It is important to understand the business need for these features and illustrate why they are critical, particularly for ISPs' operations. However, these features are not limited to ISPs and organizations of all sizes use these features for similar reasons to ISPs. During a merger, acquisition or divestiture involving two organizations it is necessary to seamlessly migrate BGP speakers from one ASN to a second ASN. The overall goal in doing so, particularly in the case of a merger or acquisition, is to achieve a uniform

operational model through consistent configurations across all BGP speakers in the combined network. In addition, and perhaps more importantly, it is common practice in the industry for ISPs to bill customers based on utilization. ISPs bill customers based on the 95th percentile of the greater of the traffic sent or received, over the course of a 1-month period, on the customer's PE-CE access circuit. Given that the BGP Path Selection algorithm selects routes with the shortest AS_PATH attribute, it is critical for the ISP to not increase AS_PATH length during or after ASN migration, toward both downstream transit customers as well as settlement-free peers, who are likely sending or receiving traffic from those transit customers. This would not only result in sudden changes in traffic patterns in the network, but also (substantially) decrease utilization driven revenue at the ISP.

Lastly, it is important to note that by default, the BGP protocol requires an operator to configure a single remote ASN for the eBGP neighbor inside a router, in order to successfully negotiate and establish an eBGP session. Prior to the existence of these features, it would have required an ISP to work with, in some cases, tens of thousands of customers. In particular, the ISP would have to encourage those customers to change their CE router configs to use the new ASN in a very short period of time, when the customer has no business incentive to do so. Thus, it becomes critical to allow the ISP to make this process a bit more asymmetric, so that it could seamlessly migrate the ASN within its network(s), but not disturb existing customers, and allow the customers to gradually migrate to the ISP's new ASN at their leisure.

1.1. Documentation note

This draft uses Autonomous System Numbers (ASNs) from the range reserved for documentation as described in RFC 5398 [RFC5398]. In the examples used here, they are intended to represent Globally Unique ASNs, not private use ASNs as documented in RFC 6996 [RFC6996] section 10.

2. ASN Migration Scenario Overview

The use case being discussed here is an ISP merging two or more ASNs, where eventually one ASN subsumes the other(s). In this use case, we will assume the most common case where there are two ISPs, A and B, that use AS 64500 and 64510, respectively, before the ASN migration is to occur. AS 64500 will be the permanently retained ASN used going forward across the consolidated set of both ISPs network equipment and AS 64510 will be retired. Thus, at the conclusion of the ASN migration, there will be a single ISP A' with all internal

BGP speakers configured to use AS 64500. To all external BGP speakers, the AS_PATH length will not be increased.

In this same scenario, AS 64496 and AS 64499 represent two, separate customer networks: C and D, respectively. Originally, customer C (AS 64496) is attached to ISP B, which will undergo ASN migration from AS 64510 to AS 64500. Furthermore, customer D (AS 64496) is attached to ISP A, which does not undergo ASN migration since ISP A's ASN will remain constant, (AS 64500). Although this example refers to AS 64496 and 64499 as customer networks, either or both may be settlement-free or other types of peers. In this use case they are referred to as "customers" merely for convenience.

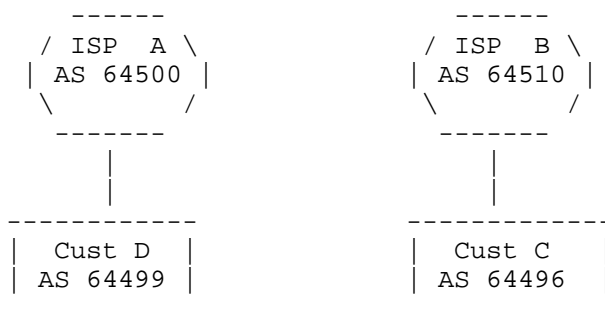


Figure 1: Before Migration

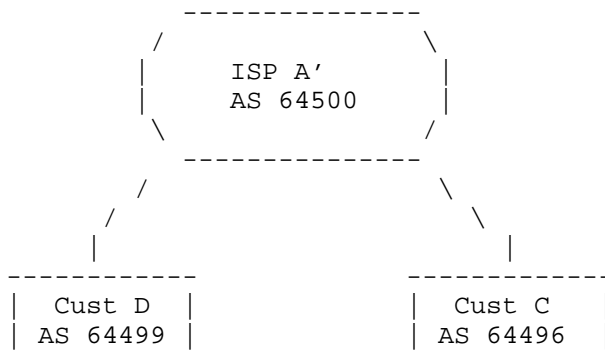


Figure 2: After Migration

The general order of operations, typically carried out in a single maintenance window by the network undergoing ASN migration, ISP B, are as follows. First, ISP B, will change the global BGP ASN used by a PE router, from ASN 64510 to 64500. At this point, the router will no longer be able to establish eBGP sessions toward the existing CE

devices that are attached to it and still using AS 64510. Second, ISP B will configure two separate, but related ASN migration features discussed in this document on all eBGP sessions toward all CE devices. These features modify the AS_PATH attribute received from and transmitted toward CE devices to achieve the desired effect of not increasing the length of the AS_PATH.

At the conclusion of the ASN migration, the CE devices at the edge of the network are not aware of and do not observe any change in the length of the AS_PATH attribute. However, after the changes discussed in this document are put in place by ISP A', there is a change to the contents of the AS_PATH attribute to ensure the AS_PATH is not artificially lengthened for the duration of time that these AS migration parameters are used.

In this use case, neither ISP is using BGP Confederations RFC 5065 [RFC5065] internally.

Additional information about this scenario, including vendor-specific implementation details can be found as follows:

- o Cisco [CISCO]
- o Juniper [JUNIPER]
- o Alcatel-Lucent [ALU]

Equivalent features do exist in other implementations, however the authors were unable to find publicly available documentation of the vendor-specific implementation to reference. Finally, the examples cited below use Cisco IOS CLI for ease of illustration purposes only.

3. External BGP Autonomous System Migration Features

The following section addresses features that are specific to modifying the AS_PATH attribute at the Autonomous System Border Routers (ASBRs) of an organization, (typically a single Service Provider). This ensures that external BGP customers/peers are not forced to make any configuration changes on their CE routers before or during the exact time the Service Provider wishes to migrate to a new, permanently retained ASN. Furthermore, these features eliminate the artificial lengthening of the AS_PATH both transmitted from and received by the Service Provider that is undergoing AS Migration, which would have negative implications on path selection by external networks.

3.1. Local AS: Modify Inbound BGP AS_PATH Attribute

ISP B needs to reconfigure its router(s) to participate as an internal BGP speaker in AS 64500, to realize the business goal of becoming a single Service Provider: ISP A'. ISP B needs to do this without coordinating the change of its ASN with all of its eBGP peers, simultaneously. The first step is for ISP B to change the global AS in its router configuration, used by the local BGP process as the system-wide Autonomous System ID, from AS 64510 to AS 64500. The next step is for ISP B to establish iBGP sessions with ISP A's existing routers, thus consolidating ISP B into ISP A resulting in operating under a single AS: ISP A', (AS 64500).

The next step is for ISP B to reconfigure its PE router(s) so that each of its eBGP sessions toward all eBGP speakers with a feature called "Local AS". This feature allows ISP B's PE router to re-establish a eBGP session toward the existing CE devices using the legacy AS, AS 64510, in the eBGP session establishment. Ultimately, the CE devices, (i.e.: customer C), are completely unaware that ISP B has reconfigured its router to participate as a member of a new AS. Within the context of ISP B's PE router, the second effect this feature has is that, by default, it prepends all received BGP UPDATE's with the legacy AS of ISP B: AS 64510. Thus, within ISP A' the AS_PATH toward customer C would appear as: 64510 64496, which is an increase in AS_PATH length from previously. Therefore, a secondary feature "No Prepend" is required to be added to the "Local AS" configuration toward every eBGP neighbor on ISP B's PE router. The "No Prepend" feature causes ISP B's PE router to not prepend the legacy AS, AS 64510, on all received eBGP UPDATE's from customer C. This restores the AS_PATH within ISP A' toward customer C so that it is just one ASN in length: 64496.

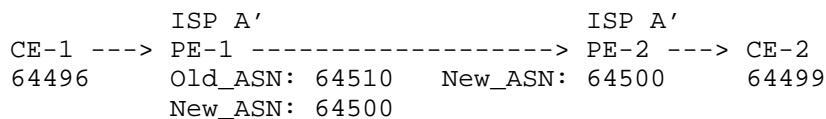
In the direction of CE -> PE (inbound):

1. 'local-as <old_ASN>': appends the <old_ASN> value to the AS_PATH of routes received from the CE
2. 'local-as <old_ASN> no-prepend': does not prepend <old_ASN> value to the AS_PATH of routes received from the CE

As stated previously, local-as <old_ASN> no-prepend, (configuration #2), is critical because it does not increase the AS_PATH length. Ultimately, this ensures that routes learned from ISP B's legacy customers will be transmitted through legacy eBGP sessions of ISP A, toward both customers and peers, will contain only two AS'es in the AS_PATH: 64500 64496. Thus, the legacy customers and peers of ISP A will not see an increase in the AS_PATH length to reach ISP B's legacy customers. Ultimately, it is considered mandatory by

operators that both the "Local AS" and "No Prepend" configuration parameters always be used in conjunction with each other in order to ensure the AS_PATH length is not increased.

PE-1 is a PE that was originally in ISP B. PE-1 has had its global configuration ASN changed from AS 64510 to AS 64500 to make it part of the permanently retained ASN. This now makes PE-1 a member of ISP A'. PE-2 is a PE that was originally in ISP A. Although its global configuration ASN remains AS 64500, throughout this exercise we also consider PE-2 a member of ISP A'.



Note: Direction of BGP UPDATE as per the arrows.

Figure 3: Local AS BGP UPDATE Diagram

The final configuration on PE-1 after completing the "Local AS" portion of the AS migration is as follows:

```

router bgp 64500
  neighbor <CE-1_IP> remote-as 64496
  neighbor <CE-1_IP> local-as 64510 no-prepend
  
```

As a result of the "Local AS No Prepend" configuration, on PE-1, CE-2 will see an AS_PATH of: 64500 64496. CE-2 will not receive a BGP UPDATE containing AS 64510 in the AS_PATH. (If only the "local-as 64510" feature was configured without the keyword "no-prepend" on PE-1, then CE-2 would see an AS_PATH of: 64496 64510 64500, which is unacceptable).

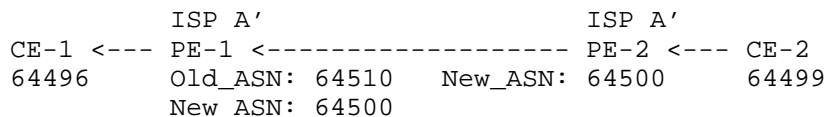
3.2. Replace AS: Modify Outbound BGP AS_PATH Attribute

The previous feature, "Local AS No Prepend", was only designed to modify the AS_PATH Attribute received from CE devices by the ISP, when CE devices still have an eBGP session established with the ISPs legacy AS, (AS64510). Use of "Local AS No Prepend" has an unfortunate side effect where its use does not concurrently modify the AS_PATH Attribute for BGP UPDATES that are transmitted by the ISP to CE devices. Specifically, with "Local AS No Prepend" enabled on ISP A's PE-1, it automatically causes a lengthening of the AS_PATH in outbound BGP UPDATES from ISP A' toward directly attached eBGP speakers, (Customer C in AS 64496). This is the result of the "Local AS No Prepend" feature automatically appending the new global configuration ASN, AS64500, after the legacy ASN, AS64510, on ISP A'

PE-1 in BGP UPDATES that are transmitted by PE-1 to CE-1. The end result is that customer C, in AS 64496, will receive the following AS_PATH: 64510 64500 64499. Therefore, if ISP A' takes no further action, it will cause an increase in AS_PATH length within customer's networks directly attached to ISP A', which is unacceptable.

A second feature, called "Replace AS", was designed to resolve this problem. This feature allows ISP A' to not append the global configured AS in outbound BGP UPDATES toward its customer's networks configured with the "Local AS" feature. Instead, only the historical (or, legacy) AS will be prepended in the outbound BGP UPDATE toward customer's network, restoring the AS_PATH length to what it was before AS Migration occurred.

To re-use the above diagram, but in the opposite direction, we have:



Note: Direction of BGP UPDATE as per the arrows.

Figure 4: Replace AS BGP UPDATE Diagram

The final configuration on PE-1 after completing the "Replace AS" portion of the AS migration is as follows:

```

router bgp 64500
  neighbor <CE-1_IP> remote-as 64496
  neighbor <CE-1_IP> local-as 64510 no-prepend replace-as
  
```

By default, without "Replace AS" enabled, CE-1 would see an AS_PATH of: 64510 64500 64499, which is artificially lengthened by the ASN Migration. After ISP A' changes PE-1 to include the "Replace AS" feature, CE-1 would receive an AS_PATH of: 64510 64499, which is the same AS_PATH length pre-AS migration.

4. Internal BGP Autonomous System Migration Features

The following section describes features that are specific to performing an ASN migration within medium to large networks in order to realize the business and operational benefits of a single network using one, globally unique Autonomous System. These features assist with a gradual and least service impacting migration of Internal BGP sessions from a legacy ASN to the permanently retained ASN. It should be noted that the following feature is very valuable to

networks undergoing AS migration, but its use does not cause changes to the AS_PATH attribute.

4.1. Internal BGP Alias

In this case, all of the routers to be consolidated into a single, permanently retained ASN are under the administrative control of a single entity. Unfortunately, though, the traditional method of migrating all Internal BGP speakers, particularly within larger networks, is both time consuming and widely service impacting.

The traditional method to migrate Internal BGP sessions was strictly limited to reconfiguration of the global configuration ASN and, concurrently, changing of iBGP neighbor's remote ASN from the legacy ASN to the new, permanently retained ASN on each router within the legacy AS. These changes can be challenging to swiftly execute in networks with more than a few dozen internal BGP speakers. There is also the concomitant service interruptions as these changes are made to routers within the network, resulting in a reset of iBGP sessions and subsequent reconvergence times to reestablish optimal routing paths. Operators do not and, in some cases, cannot make such changes given the associated risks and highly visible service interruption; rather, they require a more gradual method to migrate Internal BGP sessions, from one ASN to a second, permanently retained ASN, that is not visibly service-impacting to its customers.

With the "Internal BGP Alias" [JUNIPER] feature, it allows an Internal BGP speaker to form a single iBGP session using either the old, legacy ASN or the new, permanently retained ASN. The benefits of using this feature are several fold. First, it allows for a more gradual and less service-impacting migration away from the legacy ASN to the permanently retained ASN. Second, it (temporarily) permits the coexistence of the legacy and permanently retained ASN within a single network, allowing for uniform BGP path selection among all routers within the consolidated network.

When the "Internal BGP Alias" feature is enabled, typically just on one side of a iBGP session, it allows that iBGP speaker to establish a single iBGP session with either the legacy ASN or the new, permanently retained ASN, depending on which one it receives in the "My Autonomous System" field of the BGP OPEN message from its iBGP session neighbor. It is important to recognize that enablement of the "Internal BGP Alias" feature preserves the semantics of a regular iBGP session, (using identical ASNs). Thus, the BGP attributes transmitted by and the acceptable methods of operation on BGP attributes received from iBGP sessions configured with "Internal BGP Alias" are no different than those exchanged across an iBGP session

without "Internal BGP Alias" configured, as defined by [RFC4271] and [RFC4456].

Typically, in medium to large networks, BGP Route Reflectors [RFC4456] (RRs) are used to aid in reduction of configuration of iBGP sessions and scalability with respect to overall TCP (and, BGP) session maintenance between adjacent iBGP speakers. Furthermore, BGP Route Reflectors are typically deployed in pairs within a single Route Reflection cluster to ensure high reliability of the BGP Control Plane. As such, the following example will use Route Reflectors to aid in understanding the use of the "Internal BGP Alias" feature. It should be noted that Route Reflectors are not a prerequisite to enable "Internal BGP Alias" and this feature can be enabled independent of the use of Route Reflectors.

The general order of operations is as follows:

1. Within the legacy network, (the routers comprising the set of devices that still have a globally configured legacy ASN), take one member of a redundant pair of RRs and change its global configuration ASN to the permanently retained ASN. Concurrently, enable use of "Internal BGP Alias" on all iBGP sessions. This will comprise Non-Client iBGP sessions to other RRs as well as Client iBGP sessions, typically to PE devices, both still utilizing the legacy ASN. Note that during this step there will be a reset and reconvergence event on all iBGP sessions on the RRs whose configuration was modified; however, this should not be service impacting due to the use of redundant RRs in each RR Cluster.
2. Repeat the above step for the other side of the redundant pair of RRs. The one alteration to the above procedure is to disable use of "Internal BGP Alias" on the Non-Client iBGP sessions toward the other (previously reconfigured) RRs, since it is no longer needed. "Internal BGP Alias" is still required on all RRs for all RR Client iBGP sessions. Also during this step, there will be a reset and reconvergence event on all iBGP sessions whose configuration was modified, but this should not be service impacting. At the conclusion of this step, all RRs should now have their globally configured ASN set to the permanently retained ASN and "Internal BGP Alias" enabled and in use toward RR Clients.
3. At this point, the network administrators would then be able to establish iBGP sessions between all Route Reflectors in both the legacy and permanently retained networks. This would allow the network to appear to function, both internally and externally, as

a single, consolidated network using the permanently retained network.

4. The next steps to complete the AS migration are to gradually modify each RR Client, (PE), in the legacy network still utilizing the legacy ASN. Specifically, each legacy PE would have its globally configured ASN changed to use the permanently retained ASN. The ASN used by the PE for the iBGP sessions, toward each RR, would be changed to use the permanently retained ASN. (It is unnecessary to enable "Internal BGP Alias" on the migrated iBGP sessions). During the same maintenance window, External BGP sessions would be modified to include the above "Local AS No Prepend" and "Replace-AS" features, since all of the changes are service interrupting to the eBGP sessions of the PE. At this point, all PE's will have been migrated to the permanently retained ASN.
5. The final step is to excise the "Internal BGP Alias" configuration from the first half of the legacy RR Client pair -- this will expunge "Internal BGP Alias" configuration from all devices in the network. After this is complete, all routers in the network will be using the new, permanently retained ASN for all iBGP sessions with no vestiges of the legacy ASN on any iBGP sessions.

The benefit of using "Internal BGP Alias" is a more gradual and less externally visible, service-impacting change to accomplish an AS migration. Previously, without "Internal BGP Alias", such an AS migration change would carry a high risk and need to be successfully accomplished in a very short timeframe, (e.g.: at most several hours). In addition, it would cause substantial routing churn and, likely, rapid fluctuations in traffic carried -- potentially causing periods of congestion and resultant packet loss -- during the period the configuration changes are underway to complete the AS Migration. On the other hand, with "Internal BGP Alias", the migration from the legacy ASN to the permanently retained ASN can occur over a period of days or weeks with little disruption experienced by customers of the network undergoing AS migration. (The only observable service disruption should be when each PE undergoes the changes discussed in step 4 above.)

5. Additional Operational Considerations

This document describes several implementation-specific features to support ISPs and other organizations that need to perform ASN migrations. Other variations of these features may exist, for example, in legacy router software that has not been upgraded or

reached End of Life, but continues to operate in the network. Such variations are beyond the scope of this document.

Companies routinely go through periods of mergers, acquisitions and divestitures, which in the case of the former cause them to accumulate several legacy ASNs over time. ISPs often do not have control over the configuration of customer's devices, (i.e.: the ISPs are often not providing a managed CE router service, particularly to medium and large customers that require eBGP). Furthermore, ISPs are using methods to perform ASN migration that do not require coordination with customers. Ultimately, this means there is not a finite period of time after which legacy ASNs will be completely expunged from the ISP's network. In fact, it is common that legacy ASNs and the associated External BGP AS Migration features discussed in this document can and do persist for several years, if not longer. Thus, it is prudent to plan that legacy ASNs and associated External BGP AS Migration features will persist in a operational network indefinitely.

With respect to the Internal BGP AS Migration Features, all of the routers to be consolidated into a single, permanently retained ASN are under the administrative control of a single entity. Thus, completing the migration from iBGP sessions using the legacy ASN to the permanently retained ASN is more straightforward and could be accomplished in a matter of days to months. Finally, good operational hygiene would dictate that it is good practice to avoid using "Internal BGP Alias" over a long period of time for reasons of not only operational simplicity of the network, but also reduced reliance on that feature during the ongoing lifecycle management of software, features and configurations that are maintained on the network.

6. Conclusion

Although the features discussed in this document are not formally recognized as part of the BGP4 specification, they have been in existence in commercial implementations for well over a decade. These features are widely known by the operational community and will continue to be a critical necessity in the support of network integration activities going forward. Therefore, these features are extremely unlikely to be deprecated by vendors. As a result, these features must be acknowledged by protocol designers, particularly when there are proposals to modify BGP's behavior with respect to handling or manipulation of the AS_PATH Attribute. More specifically, assumptions should not be made with respect to the preservation or consistency of the AS_PATH Attribute as it is transmitted along a sequence of ASN's. In addition, proposals to manipulate the AS_PATH that would gratuitously increase AS_PATH

length or remove the capability to use these features described in this document will not be accepted by the operational community.

7. Acknowledgements

Thanks to Kotikalapudi Sriram, Stephane Litkowski, Terry Manderson, and David Farmer for their comments.

8. IANA Considerations

This memo includes no request to IANA.

9. Security Considerations

This draft discusses a process by which one ASN is migrated into and subsumed by another. This involves manipulating the AS_PATH Attribute with the intent of not increasing the AS_PATH length, which would typically cause the BGP route to no longer be selected by BGP's Path Selection Algorithm in other's networks. This could result in a loss of revenue if the ISP is billing based on measured utilization of traffic sent to/from entities attached to its network. This could also result in sudden, and unexpected shifts in traffic patterns in the network, potentially resulting in congestion, in the most extreme cases.

Given that these features can only be enabled through configuration of router's within a single network, standard security measures should be taken to restrict access to the management interface(s) of routers that implement these features.

10. References

10.1. Normative References

[RFC5398] Huston, G., "Autonomous System (AS) Number Reservation for Documentation Use", RFC 5398, December 2008.

10.2. Informative References

[ALU] Alcatel-Lucent, "BGP Local AS attribute", 2006-2012, <https://infoproducts.alcatel-lucent.com/html/0_add-h-f/93-0074-10-01/7750_SR_OS_Routing_Protocols_Guide/BGP-CLI.html#709567>.

[CISCO] Cisco Systems, Inc., "BGP Support for Dual AS Configuration for Network AS Migrations", 2003, <http://www.cisco.com/en/US/docs/ios/12_3t/12_3t11/feature/guide/gtbgpdas.html>.

- [JUNIPER] Juniper Networks, Inc., "Configuring the BGP Local Autonomous System Attribute", 2012, <https://www.juniper.net/techpubs/en_US/junos12.3/topics/reference/configuration-statement/local-as-edit-protocols-bgp.html>.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.
- [RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, August 2007.
- [RFC6996] Mitchell, J., "Autonomous System (AS) Reservation for Private Use", BCP 6, RFC 6996, July 2013.

Authors' Addresses

Wesley George
Time Warner Cable
13820 Sunrise Valley Drive
Herndon, VA 20171
US

Phone: +1 703-561-2540
Email: wesley.george@twcable.com

Shane Amante
Apple, Inc.
1 Infinite Loop
Cupertino, CA 95014
US

Email: samante@apple.com

Internet Engineering Task Force
Internet-Draft
Updates: 1997, 4271, 4360, 4456, 4760, 5701 (if approved)
Intended status: Standards Track
Expires: August 18, 2014

E. Chen, Ed.
Cisco Systems, Inc.
J. Scudder, Ed.
Juniper Networks
P. Mohapatra
Cumulus Networks, Inc.
K. Patel
Cisco Systems, Inc.
February 14, 2014

Revised Error Handling for BGP UPDATE Messages
draft-ietf-idr-error-handling-06

Abstract

According to the base BGP specification, a BGP speaker that receives an UPDATE message containing a malformed attribute is required to reset the session over which the offending attribute was received. This behavior is undesirable as a session reset would impact not only routes with the offending attribute, but also other valid routes exchanged over the session. This document partially revises the error handling for UPDATE messages, and provides guidelines for the authors of documents defining new attributes. Finally, it revises the error handling procedures for a number of existing attributes.

This document updates error handling for RFCs 1997, 4271, 4360, 4456, 4760, and 5701.

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 August 18, 2014.

Copyright Notice

Copyright (c) 2014 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.

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

Table of Contents

1. Introduction	3
1.1. Requirements Language	3
2. Revision to Base Specification	4
3. Parsing of NLRI Fields	6
3.1. Inconsistency of Attribute Length Fields	6
3.2. Syntactic Correctness of NLRI Fields	7
4. Operational Considerations	7
5. Error Handling Procedures for Existing Attributes	8
5.1. ORIGIN	8
5.2. AS_PATH	8
5.3. NEXT_HOP	9
5.4. MULTI_EXIT_DESC	9
5.5. LOCAL_PREF	9
5.6. ATOMIC_AGGREGATE	9
5.7. AGGREGATOR	9
5.8. Community	10
5.9. Extended Community	10
5.10. IPv6 Address Specific BGP Extended Community Attribute	10
5.11. ORIGINATOR_ID	11

5.12. CLUSTER_LIST	11
6. IANA Considerations	11
7. Security Considerations	11
8. Acknowledgements	11
9. Normative References	12
Appendix A. Why not discard UPDATE messages?	12
Authors' Addresses	13

1. Introduction

According to the base BGP specification [RFC4271], a BGP speaker that receives an UPDATE message containing a malformed attribute is required to reset the session over which the offending attribute was received. This behavior is undesirable as a session reset would impact not only routes with the offending attribute, but also other valid routes exchanged over the session. In the case of optional transitive attributes, the behavior is especially troublesome and may present a potential security vulnerability. The reason is that such attributes may have been propagated without being checked by intermediate routers that do not recognize the attributes -- in effect the attribute may have been tunneled, and when they do reach a router that recognizes and checks them, the session that is reset may not be associated with the router that is at fault.

The goal for revising the error handling for UPDATE messages is to minimize the impact on routing by a malformed UPDATE message, while maintaining protocol correctness to the extent possible. This can be achieved largely by maintaining the established session and keeping the valid routes exchanged, but removing the routes carried in the malformed UPDATE from the routing system.

This document partially revises the error handling for UPDATE messages, and provides guidelines for the authors of documents defining new attributes. Finally, it revises the error handling procedures for a number of existing attributes. Specifically, the error handling procedures of, [RFC1997], [RFC4271], [RFC4360], [RFC4456], [RFC4760] and [RFC5701] are revised.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Revision to Base Specification

The first paragraph of Section 6.3 of [RFC4271] is revised as follows:

Old Text:

All errors detected while processing the UPDATE message MUST be indicated by sending the NOTIFICATION message with the Error Code UPDATE Message Error. The error subcode elaborates on the specific nature of the error.

New text:

An error detected while processing the UPDATE message for which a session reset is specified MUST be indicated by sending the NOTIFICATION message with the Error Code UPDATE Message Error. The error subcode elaborates on the specific nature of the error.

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 the following case described in Section 6.3 of [RFC4271] is revised

If any recognized attribute has Attribute Flags that conflict with the Attribute Type Code, then the Error Subcode MUST be set to Attribute Flags Error. The Data field MUST contain the erroneous attribute (type, length, and value).

as follows:

If any recognized attribute has Attribute Flags that conflict with the Attribute Type Code, then the attribute MUST be treated as malformed and the treat-as-withdraw approach (see below) used, unless the specification for the attribute mandates different handling for incorrect Attribute Flags.

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

When a path attribute (other than the MP_REACH_NLRI attribute [RFC4760] or the MP_UNREACH_NLRI attribute [RFC4760]) in an UPDATE message is determined to be malformed, the UPDATE message containing that attribute MUST be treated as though all contained routes had been withdrawn just as if they had been listed in the WITHDRAWN ROUTES field (or in the MP_UNREACH_NLRI attribute if appropriate) of the UPDATE message, thus causing them to be removed from the Adj-RIB-In according to the procedures of [RFC4271]. In the case of an attribute which has no effect on route selection or installation, the malformed attribute MAY instead be discarded and the UPDATE message continue to be processed. For the sake of brevity, the former approach is termed "treat-as-withdraw", and the latter as "attribute discard".

If any of the well-known mandatory attributes are not present in an UPDATE message, then the approach of "treat-as-withdraw" MUST be used for the error handling.

The approach of "treat-as-withdraw" MUST be used for the error handling of the cases described in Section 6.3 of [RFC4271] that specify a session reset and involve any of the following attributes: ORIGIN, AS_PATH, NEXT_HOP, MULTI_EXIT_DISC, and LOCAL_PREF.

The approach of "attribute discard" MUST be used for the error handling of the cases described in Section 6.3 of [RFC4271] that specify a session reset and involve any of the following attributes: ATOMIC_AGGREGATE and AGGREGATOR.

If the MP_REACH_NLRI attribute or the MP_UNREACH_NLRI attribute appears more than once in the UPDATE message, then a NOTIFICATION message MUST be sent with the Error Subcode "Malformed Attribute List". If any other attribute appears more than once in an UPDATE message, then all the occurrences of the attribute other than the first one SHALL be discarded and the UPDATE message continue to be processed.

When multiple attribute errors exist in an UPDATE message, if the same approach (either "session reset", or "treat-as-withdraw" or "attribute discard") is specified for the handling of these malformed attributes, then the specified approach MUST be used. Otherwise the approach with the strongest action MUST be used following the order of "session reset", "treat-as-withdraw" and "attribute discard" from the strongest to the weakest.

A document which specifies a new attribute MUST provide specifics regarding what constitutes an error for that attribute and how that error is to be handled.

Finally, we observe that in order to use the approach of "treat-as-withdraw", the entire NLRI field and/or the MP_REACH_NLRI and MP_UNREACH_NLRI attributes need to be successfully parsed. If this is not possible, the procedures of [RFC4271] continue to apply. Alternatively the error handling procedures specified in [RFC4760] for disabling a particular AFI/SAFI MAY be followed. One notable case where it would be not possible to successfully parse the NLRI is if the NLRI field is found to be "syntactically incorrect" (see Section 3.2). It can be seen that therefore, this part of [RFC4271] Section 6.3 necessarily continues to apply:

The NLRI field in the UPDATE message is checked for syntactic validity. If the field is syntactically incorrect, then the Error Subcode MUST be set to Invalid Network Field.

Furthermore, this document extends RFC 4271 by mandating that the Withdrawn Routes field SHALL be checked for syntactic correctness in the same manner as the NLRI field.

3. Parsing of NLRI Fields

To facilitate the determination of the NLRI field in an UPDATE with a malformed attribute, the MP_REACH_NLRI or MP_UNREACH_NLRI attribute (if present) SHALL be encoded as the very first path attribute in an UPDATE. An implementation, however, MUST still be prepared to receive these fields in any position.

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

3.1. Inconsistency of Attribute Length Fields

There are two error cases in which the Total Attribute Length value can be in conflict with the enclosed path attributes, which themselves carry length values. In the "overflow" case, as the enclosed path attributes are parsed, the length of the last encountered path attribute would cause the Total Attribute Length to be exceeded. In the "underflow" case, as the enclosed path attributes are parsed, after the last successfully-parsed attribute, fewer than three bytes remain, or fewer than four bytes, if the Attribute Flags field has the Extended Length bit set -- that is, there remains unconsumed data in the path attributes but yet insufficient data to encode a single minimum-sized path attribute. In either of these

cases an error condition exists and the treat-as-withdraw approach MUST be used (unless some other, more severe error is encountered dictating a stronger approach), and the Total Attribute Length MUST be relied upon to enable the beginning of the NLRI field to be located.

3.2. Syntactic Correctness of NLRI Fields

The NLRI field or Withdrawn Routes field SHALL be considered "syntactically incorrect" if either of the following are true:

- o The length of any of the included NLRI is greater than 32,
- o When parsing NLRI contained in the field, the length of the last NLRI found exceeds the amount of unconsumed data remaining in the field.

Similarly, the MP_REACH or MP_UNREACH attribute of an update SHALL be considered to be incorrect if any of the following are true:

- o The length of any of the included NLRI is inconsistent with the given AFI/SAFI (for example, if an IPv4 NLRI has a length greater than 32 or an IPv6 NLRI has a length greater than 128),
- o When parsing NLRI contained in the attribute, the length of the last NLRI found exceeds the amount of unconsumed data remaining in the attribute.

4. Operational Considerations

Although the "treat-as-withdraw" error-handling behavior defined in Section 2 makes every effort to preserve BGP's correctness, we note that if an UPDATE received on an IBGP session is subjected to this treatment, inconsistent routing within the affected Autonomous System may result. The consequences of inconsistent routing can include long-lived forwarding loops and black holes. While lamentable, this issue is expected to be rare in practice, and more importantly is seen as less problematic than the session-reset behavior it replaces.

When a malformed attribute is indeed detected over an IBGP session, we RECOMMEND that routes with the malformed attribute be identified and traced back to the ingress router in the network where the routes were sourced or received externally, and then a filter be applied on the ingress router to prevent the routes from being sourced or received. This will help maintain routing consistency in the network.

Even if inconsistent routing does not arise, the "treat-as-withdraw" behavior can cause either complete unreachability or sub-optimal routing for the destinations whose routes are carried in the affected UPDATE message.

Note that "treat-as-withdraw" is different from discarding an UPDATE message. The latter violates the basic BGP principle of incremental update, and could cause invalid routes to be kept. (See also Appendix A.)

For any malformed attribute which is handled by the "attribute discard" instead of the "treat-as-withdraw" approach, it is critical to consider the potential impact of doing so. In particular, if the attribute in question has or may have an effect on route selection or installation, the presumption is that discarding it is unsafe, unless careful analysis proves otherwise. The analysis should take into account the tradeoff between preserving connectivity and potential side effects.

Because of these potential issues, a BGP speaker MUST provide debugging facilities to permit issues caused by a malformed attribute to be diagnosed. At a minimum, such facilities MUST include logging an error listing the NLRI involved, and containing the entire malformed UPDATE message when such an attribute is detected. The malformed UPDATE message SHOULD be analyzed, and the root cause SHOULD be investigated.

5. Error Handling Procedures for Existing Attributes

5.1. ORIGIN

The attribute is considered malformed if its length is not 1, or it has an undefined value [RFC4271].

An UPDATE message with a malformed ORIGIN attribute SHALL be handled using the approach of "treat-as-withdraw".

5.2. AS_PATH

The error conditions for the attribute have been defined in [RFC4271].

An UPDATE message with a malformed AS_PATH attribute SHALL be handled using the approach of "treat-as-withdraw".

5.3. NEXT_HOP

The error conditions for the NEXT_HOP attribute have been defined in [RFC4271].

An UPDATE message with a malformed NEXT_HOP attribute SHALL be handled using the approach of "treat-as-withdraw".

5.4. MULTI_EXIT_DESC

The attribute is considered malformed if its length is not 4 [RFC4271].

An UPDATE message with a malformed MULTI_EXIT_DESC attribute SHALL be handled using the approach of "treat-as-withdraw".

5.5. LOCAL_PREF

The attribute is considered malformed if its length is not 4 [RFC4271].

An UPDATE message with a malformed LOCAL_PREF attribute SHALL be handled as follows:

- o using the approach of "attribute discard" if the UPDATE message is received from an external neighbor, or
- o using the approach of "treat-as-withdraw" if the UPDATE message is received from an internal neighbor.

In addition, if the attribute is present in an UPDATE message from an external neighbor, the approach of "attribute discard" SHALL be used to handle the unexpected attribute in the message.

5.6. ATOMIC_AGGREGATE

The attribute SHALL be considered malformed if its length is not 0 [RFC4271].

An UPDATE message with a malformed ATOMIC_AGGREGATE attribute SHALL be handled using the approach of "attribute discard".

5.7. AGGREGATOR

The error conditions specified in [RFC4271] for the attribute are revised as follows:

The AGGREGATOR attribute SHALL be considered malformed if any of the following applies:

- o Its length is not 6 (when the "4-octet AS number capability" is not advertised to, or not received from the peer [RFC6793]).
- o Its length is not 8 (when the "4-octet AS number capability" is both advertised to, and received from the peer).

An UPDATE message with a malformed AGGREGATOR attribute SHALL be handled using the approach of "attribute discard".

5.8. Community

The error handling of [RFC1997] is revised as follows:

The Community attribute SHALL be considered malformed if its length is nonzero and is not a multiple of 4.

An UPDATE message with a malformed Community attribute SHALL be handled using the approach of "treat-as-withdraw".

5.9. Extended Community

The error handling of [RFC4360] is revised as follows:

The Extended Community attribute SHALL be considered malformed if its length is nonzero and is not a multiple of 8.

An UPDATE message with a malformed Extended Community attribute SHALL be handled using the approach of "treat-as-withdraw".

Note that a BGP speaker MUST NOT treat an unrecognized Extended Community Type or Sub-Type as an error.

5.10. IPv6 Address Specific BGP Extended Community Attribute

The error handling of [RFC5701] is revised as follows:

The IPv6 Address Specific Extended Community attribute SHALL be considered malformed if its length is nonzero and is not a multiple of 20.

An UPDATE message with a malformed IPv6 Address Specific Extended Community attribute SHALL be handled using the approach of "treat-as-withdraw".

Note that a BGP speaker MUST NOT treat an unrecognized IPv6 Address Specific Extended Community Type or Sub-Type as an error.

5.11. ORIGINATOR_ID

The error handling of [RFC4456] is revised as follows.

- o If the ORIGINATOR_ID attribute is received from an external neighbor, it SHALL be discarded using the approach of "attribute discard", or
- o if received from an internal neighbor, it SHALL be considered malformed if its length is not equal to 4. If malformed, the UPDATE SHALL be handled using the approach of "treat-as-withdraw".

5.12. CLUSTER_LIST

The error handling of [RFC4456] is revised as follows.

- o If the CLUSTER_LIST attribute is received from an external neighbor, it SHALL be discarded using the approach of "attribute discard", or
- o if received from an internal neighbor, it SHALL be considered malformed if its length is not a multiple 4. If malformed, the UPDATE SHALL be handled using the approach of "treat-as-withdraw".

6. IANA Considerations

This document makes no request of IANA.

7. Security Considerations

This specification addresses the vulnerability of a BGP speaker to a potential attack whereby a distant attacker can generate a malformed optional transitive attribute that is not recognized by intervening routers (which thus propagate the attribute unchecked) but that causes session resets when it reaches routers that do recognize the given attribute type.

In other respects, this specification does not change BGP's security characteristics.

8. Acknowledgements

The authors wish to thank Juan Alcaide, Ron Bonica, Mach Chen, Andy Davidson, Bruno Decraene, Dong Jie, Rex Fernando, Joel Halpern, Akira Kato, Miya Kohno, Tony Li, Alton Lo, Shin Miyakawa, Tamas Mondal,

Jonathan Oddy, Robert Raszuk, Yakov Rekhter, Eric Rosen, Rob Shakir, Naiming Shen, Shyam Sethuram, Ananth Suryanarayana, Kaliraj Vairavakkalai and Lili Wang for their observations and discussion of this topic, and review of this document.

9. 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.
- [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.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, November 2009.
- [RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, December 2012.

Appendix A. Why not discard UPDATE messages?

A commonly asked question is "why not simply discard the UPDATE message instead of treating it like a withdraw? Isn't that safer and easier?" The answer is that it might be easier, but it would compromise BGP's correctness so is unsafe. Consider the following example of what might happen if UPDATE messages carrying bad attributes were simply discarded:

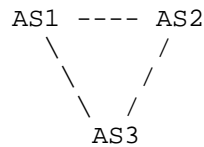


Figure 1

- o AS1 prefers to reach AS3 directly, and advertises its route to AS2.
- o AS2 prefers to reach AS3 directly, and advertises its route to AS1.
- o Connections AS3-AS1 and AS3-AS2 fail simultaneously.
- o AS1 switches to prefer AS2's route, and sends an update message which includes a withdraw of its previous announcement. The withdraw is bundled with some advertisements. It includes a bad attribute. As a result, AS2 ignores the message.
- o AS2 switches to prefer AS1's route, and sends an update message which includes a withdraw of its previous announcement. The withdraw is bundled with some advertisements. It includes a bad attribute. As a result, AS1 ignores the message.

The end result is that AS1 forwards traffic for AS3 towards AS2, and AS2 forwards traffic for AS3 towards AS1. This is a permanent (until corrected) forwarding loop.

Although the example above discusses route withdraws, we observe that in BGP the announcement of a route also withdraws the route previously advertised. The implicit withdraw can be converted into a real withdraw in a number of ways; for example, the previously-announced route might have been accepted by policy, but the new announcement might be rejected by policy. For this reason, the same concerns apply even if explicit withdraws are removed from consideration.

Authors' Addresses

Enke Chen (editor)
Cisco Systems, Inc.

Email: enkechen@cisco.com

John G. Scudder (editor)
Juniper Networks

Email: jgs@juniper.net

Pradosh Mohapatra
Cumulus Networks, Inc.

Email: pmohapat@cumulusnetworks.com

Keyur Patel
Cisco Systems, Inc.

Email: keyupate@cisco.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 24, 2014

V. Lopez
O. Gonzalez de Dios
Telefonica I+D
D. King
Old Dog Consulting
S. Previdi
Cisco Systems, Inc.
October 21, 2013

Traffic Engineering Database dissemination for Hierarchical PCE
scenarios
draft-lopez-pce-hpce-ted-00

Abstract

The PCE architecture is well-defined and may be used to compute the optimal path for LSPs across domains in MPLS-TE and GMPLS networks. The Hierarchical Path Computation Element (H-PCE) [RFC6805] was developed to provide an optimal path when the sequence of domains is not known in advance. The procedure and mechanism for populating the Traffic Engineering Database (TED) with domain topology and link information used in H-PCE-based path computations is open to interpretation. This informational document describes how topology dissemination mechanisms may be used to provide TE information between Parent and Child PCEs (within the H-PCE context). In particular, it describes how BGP-LS might be used to provide inter-domain connectivity. This document is not intended to define new extensions, it demonstrates how existing procedures and mechanisms may be used.

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 24, 2014.

Copyright Notice

Copyright (c) 2013 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	2
1.1. Parent PCE Domain Topology	3
1.2. Parent PCE TED requirements	3
2. H-PCE Domain Topology Dissemination and Construction Methods	4
3. H-PCE architecture using BGP-LS	5
4. Including Inter-domain connectivity in BGP-LS	8
4.1. Mapping from OSPF-TE	8
4.2. Mapping from ISIS-TE	8
5. BGP considerations	8
6. Manageability Considerations	8
7. Security Considerations	8
8. Acknowledgements	9
9. References	9
9.1. Normative References	9
9.2. Informative References	9
Authors' Addresses	10

1. Introduction

In scenarios with multiple domains in both MPLS-TE and GMPLS networks, the hierarchical Path Computation Element (H-PCE) Architecture, defined in [RFC6805], allows to obtain the optimum end-to-end path. The architecture exploits a hierarchical relation among domains.

[RFC6805] defines the architecture and requirements for the end-to-end path computation across domains. The solution draft for the H-PCE [I-D.draft-ietf-pce-hierarchy-extensions-00] is focused on the PCEP protocol extensions to support such H-PCE procedures, including negotiation of capabilities and errors. However, neither the architecture nor the solution draft specify which mechanism must to

be used to build and populate the parent PCE (pPCE) Traffic Engineering Database (TED).

The H-PCE architecture documents define the minimum content needed in the traffic engineering database required to compute paths. The information required by parent TEDB are identified in [RFC6805] and further elaborated in [I-D.draft-ietf-pce-inter-area-as-applicability-03]. For instance, [RFC6805] and [I-D.draft-ietf-pce-inter-area-as-applicability-03] suggest that BGP-LS could be used as a "northbound" TE advertisement. This means that a PCE does not need to listen IGP in its domain, but its TED is populated by messages received (for example) from a Route Reflector.

This document highlights the applicability of BGP-LS to the dissemination of domain topology within the H-PCE architecture. In particular, it describes how can BGP-LS be used to send the inter-domain connectivity. It also shows how can OSPF-TE and ISIS-TE updates be mapped into BGP-LS.

Note that this document is not intended to define new protocol extensions, it is an informational document and where required it highlights where existing mechanisms and protocols may be applied.

1.1. Parent PCE Domain Topology

The pPCE maintains a domain topology map of the child domains and their interconnectivity. This map does not include any visibility into the child domains. Where inter-domain connectivity is provided by TE links, the capabilities of those links may also be known to the pPCE. The pPCE maintains a TED for the parent domain, the nodes in the parent domain are abstractions of the cPCE domains (connected by real or virtual TE links), but the pPCE domain may also include real nodes and links.

The procedure and protocol mechanism for disseminating and construction of the pPCE TED may be provided using a number of mechanisms, including manually configuring the necessary information or automated using a separate instance of a routing protocol to advertise the domain interconnectivity. Since inter-domain TE links can be advertised by the IGPs operating in the child domains, this information could then be exported to the parent PCE either by the child PCEs or using north-bound export mechanisms.

1.2. Parent PCE TED requirements

The information that would be exchanged includes:

- o Identifier of advertising child PCE.
- o Identifier of PCE's domain.
- o Identifier of the link.
- o TE properties of the link (metrics, bandwidth).
- o Other properties of the link (technology-specific).
- o Identifier of link endpoints.
- o Identifier of adjacent domain.

2. H-PCE Domain Topology Dissemination and Construction Methods

A variety of methods exist to provide are different alternatives so the parent PCE can get the topological information from the child PCEs (cPCEs):

- o Statically configure all inter-domain link and topology information.
- o Membership of an IGP instance. The necessary topological information could be disseminated by joining the IGP instance of each child PCE domain. However, by doing so, it would break the domain confidentiality principles and is subject to scalability issues.
- o PCEP Notification Messages. Another solution is to send the interconnection information between domains using PCEP Notifications (see section 4.8.4 of [RFC6805]). One approach, followed in research work, is embedding in PCEP Notifications the Inter-AS OSPF-TE Link State Advertisements (LSA) to send the Inter-Domain Link information from child PCEs to the parent PCE and to send reachability information (list of end-points in each domain). However, it is argued that the utilization of PCEP to disseminate topology is beyond scope of the protocol.
- o Separate IGP instance. [RFC6805] points out that in models such as ASON it is possible to consider a separate instance of an IGP running within the parent domain where the participating protocol speakers are the nodes directly present in that domain and the PCEs (parent and child PCEs).
- o Use north-bound distribution of TE information. The North-Bound Distribution of Link-State and TE Information using BGP has been recently propose in the IEFT

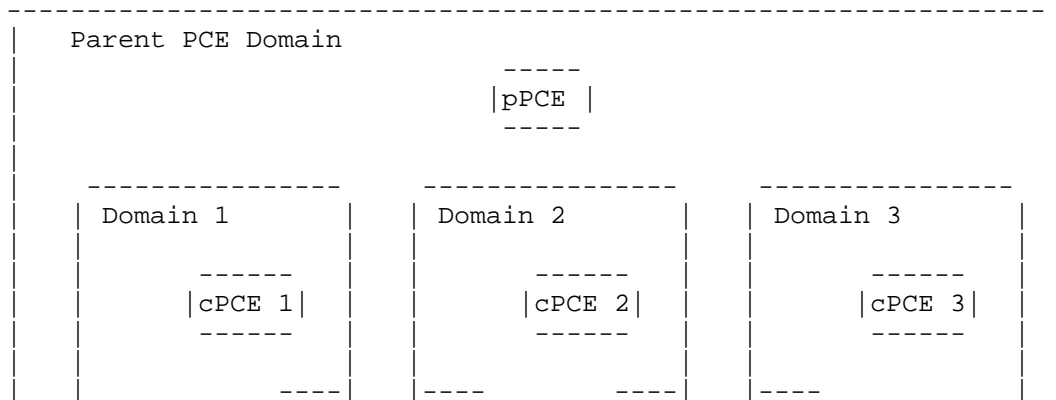
[I-D.draft-ietf-idr-ls-distribution-03]. This approach is known as BGP-LS and defines a mechanism by which links state and traffic engineering information can be collected from networks and exported to external elements using the BGP routing protocol. By using BGP-LS as northbound distribution mechanism, there would be a BGP speaker in each domains that sends the necessary information to a BGP speaker in the parent domain. This architecture is further elaborated in this document.

3. H-PCE architecture using BGP-LS

As mentioned in [I-D.draft-dugeon-pce-ted-reqs-01] PCE has to retrieve Traffic Engineering (TE) information to carry out its path computation. This is required not only for intra-domain information, which can be got using IGP (like OSPF-TE or ISIS-TE), but also for inter-domain information in the Hierarchical PCE (H-PCE) architecture.

Figure 1 shows an example of a H-PCE architecture. In this example, there is a parent PCE and three child PCEs, and they are organized in multiple domains. The parent PCE does not have information of the whole network, but is only aware of the connectivity among the domains and provides coordination to the child PCEs. Figure 2 shows which is the visibility that parent PCE has from the network according to the definition in [RFC6805].

Thanks to this topological information, when there is a request to a child PCE with the destination in another domain, this path request is sent to the parent PCE, which selects a set of candidate domain paths and sends requests to the child PCEs responsible for these domains. Then, the parent PCE selects the best solution and it is transmitted to the source PCE.



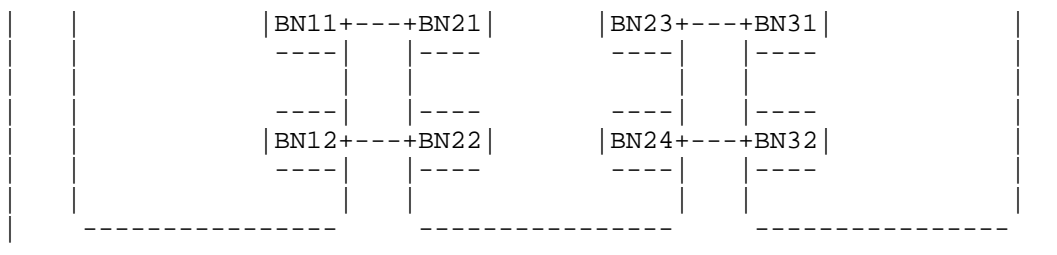


Figure 1: Example of Hierarchical PCE architecture

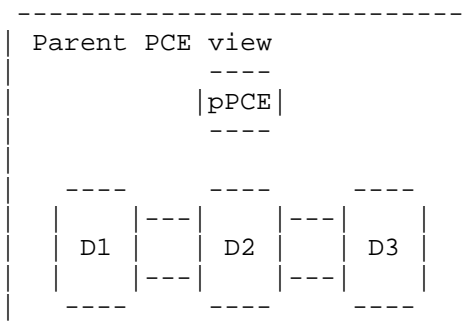
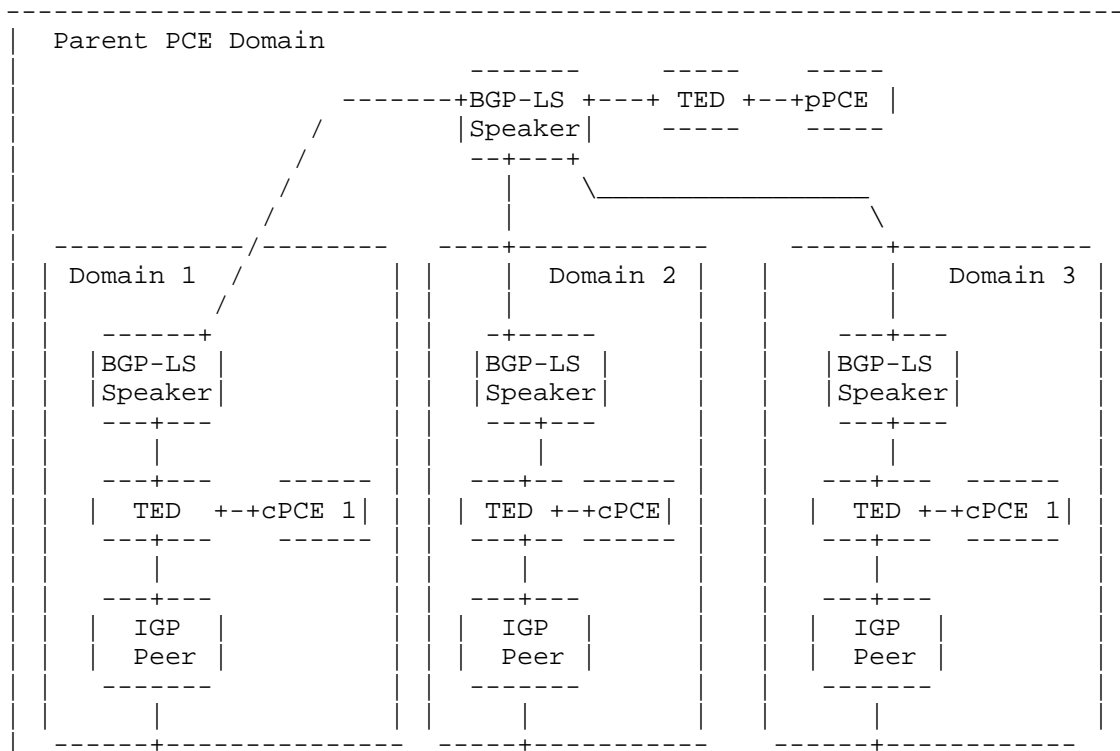


Figure 2: Parent PCE topology information

Thanks to the dissemination of inter-domain adjacency information from each cPCE to the pPCE, the pPCE can have a view of reachability between the domains. The H-PCE architecture with BGP-LS is shown in Figure 3. Each domain has a cPCE that is able to compute paths in the domain. This child PCE has access to a domain TED, which is built using IGP information. In each domain, a BGP speaker has access to such domain TED and acts as BGP-LS Route Reflector to provide network topology to the pPCE. Next to the pPCE, there is a BGP speaker that maintains a BGP session with each of the BGP speakers in the domains to receive the topology and build the parent TED. A policy can be applied to the BGP-LS speakers to decide which information is sent to its peer speaker. The minimum amount of information that needs to be exchanged is the inter-domain connectivity, including the details of the Traffic Engineering Inter-domain Links [RFC6805]. With this information, the parent PCE is able to have access to a domain topology map and its connectivity. Additionally, the BGP-LS speaker can be configured to send the

complete list of TE Links, including its details. In this case, the parent PCE has access to an extended database, with visibility of both intra-domain and inter-domain information and can compute the sequence of domains with better accuracy. Even, the pPCE could have enough information to compute the whole end-to-end path by itself.

BGP-LS [I-D.draft-ietf-idr-ls-distribution-03] extends the BGP Update messages to advertise link-state topology thanks to new BGP Network Layer Reachability Information (NLRI). The Link State information is sent in two BGP attributes, the MP_REACH (defined in [RFC4670]) and a LINK_STATE attribute (defined in [I-D.draft-ietf-idr-ls-distribution-03]). To describe the inter domain links, in the MP_REACH attribute, a Link NLRI can be used with the local node descriptors the address of the source, and in the remote descriptors, the address of the destination of the link. The Link Descriptors field has a TLV (Link Local/Remote Identifiers), which carries the prefix of the Unnumbered or Numbered Interface. In case of the message informs about an intra-domain link, the standard traffic engineering information is included in the LINK_STATE attribute.



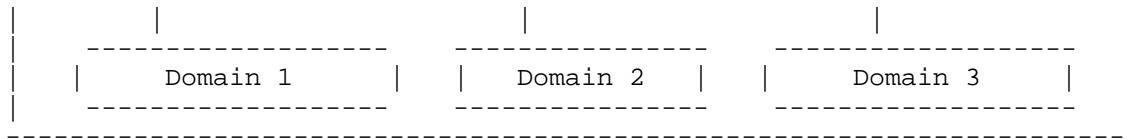


Figure 3: Example of Hierarchical PCE architecture with BGP-LS

4. Including Inter-domain connectivity in BGP-LS

TBD

4.1. Mapping from OSPF-TE

TBD

4.2. Mapping from ISIS-TE

TBD

5. BGP considerations

TBD

- o Supporting BGP-4
- o BGP Speakers
- o Graceful Restart
- o SRLGs
- o Multiprotocol extensions

6. Manageability Considerations

TBD

7. Security Considerations

TBD

8. Acknowledgements

Authors would like to thank Stefano Previdi for his comments.

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.
- [RFC4670] Nelson, D., "RADIUS Accounting Client MIB for IPv6", RFC 4670, August 2006.
- [RFC6805] King, D. and A. Farrel, "The Application of the Path Computation Element Architecture to the Determination of a Sequence of Domains in MPLS and GMPLS", RFC 6805, November 2012.

9.2. Informative References

- [I-D.draft-dugeon-pce-ted-reqs-01]
Dugeon, O., Meuric, J., Douville, R., Casellas, R., and O. Gonzalez de Dios, "Path Computation Element (PCE) Traffic Engineering Database (TED) Requirements", March 2012.
- [I-D.draft-ietf-idr-ls-distribution-03]
Gredler, H., Medved, J., Previdi, S., Farrel, A., and S. Ray, "North-Bound Distribution of Link-State and TE Information using BGP", May 2013.
- [I-D.draft-ietf-pce-hierarchy-extensions-00]
Zhang, F., Zhao, Q., Gonzalez de Dios, O., Casellas, R., and D. King, "Extensions to Path Computation Element Communication Protocol (PCEP) for Hierarchical Path Computation Elements (PCE)", August 2013.
- [I-D.draft-ietf-pce-inter-area-as-applicability-03]
King, D., Meuric, J., Dugeon, O., Zhao, Q., and O. Gonzalez de Dios, "Applicability of the Path Computation Element to Inter-Area and Inter-AS MPLS and GMPLS Traffic Engineering", March 2012.

Authors' Addresses

Victor Lopez
Telefonica I+D
Don Ramon de la Cruz 82-84
Madrid 28045
Spain

Phone: +34913128872
Email: vlopez@tid.es

Oscar Gonzalez de Dios
Telefonica I+D
Don Ramon de la Cruz 82-84
Madrid 28045
Spain

Phone: +34913128832
Email: ogondio@tid.es

Daniel King
Old Dog Consulting
UK

Email: daniel@olddog.co.uk

Stefano Previdi
Cisco Systems, Inc.
Via Del Serafico 200
Rome 00144
IT

Email: sprevidi@cisco.com

Inter-Domain Routing
Internet-Draft
Intended status: Standards Track
Expires: August 18, 2014

S. Ray
A. Sreekantiah
Cisco Systems, Inc.
February 14, 2014

Signaling AFI-SAFI scope for Constrained Route Distribution
draft-ray-idr-route-constrain-scope-00

Abstract

The Route Constrain address family can be used by a BGP speaker to signal a neighbor its interest in receiving only the routes with a matching route target (RT) extended community. This signaling is afi-safi agnostic; the sender of a route constrain NLRI with an RT expresses its interest in receiving routes with that RT for all afi-safi. The ability to further scope a given RT to a list of afi-safi would simplify network operations; then optimal route filtering would no longer require that the set of RTs used for different afi-safi be disjoint.

This document proposes a simple extended community based backward compatible method to associate the list of afi-safi of interest to a route constrain NLRI and discusses the operational procedure.

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 August 18, 2014.

Copyright Notice

Copyright (c) 2014 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	2
1.1. Sub-optimality of route constrain in shared RT deployments	3
1.1.1. Unused route advertisement and retention	3
1.1.2. Need for route-refresh on some configuration changes	4
1.2. Solutions	4
1.2.1. Use of disjoint set of RTs	5
1.2.2. New route constrain NLRI	5
1.2.3. Using BGP Path Attribute	5
2. AFI-SAFI Extended Community	6
3. Operation	6
4. IANA Considerations	10
5. Manageability Considerations	10
5.1. Configuration Management	10
5.2. Operational Considerations	10
6. Security Considerations	10
7. Acknowledgements	10
8. References	10
8.1. Normative References	10
8.2. Informative References	11
Authors' Addresses	12

1. Introduction

A VPN route (such as VPNv4/VPNv6 unicast [RFC4364], Multicast VPN [RFC6513], Layer 2 VPN [RFC6624], Flow-spec [RFC5575], etc.) is retained by a BGP speaker only if the route is of interest to itself - e.g., if the route belongs to a local VPN, or if it needs to be sent to one of its neighbors. The VPN membership of a route is determined by the set of route target (RT) extended communities attached to the route. Therefore the speaker's neighbors need to

send only those routes to the speaker that carry one or more RTs that are of interest to the speaker.

[RFC4684] defines a new address family called Route Target Constrain (RTC) using which a BGP speaker can signal its neighbors the set of RTs it is interested in. For instance, if one of speaker A's locally configured VPNs is a member of RT:1:1 (i.e., if a local VRF imports RT:1:1), then speaker A advertises an RTC NLRI with RT:1:1 to speaker B. In turn speaker B sends to speaker A only the VPN routes with RT:1:1. [I-D.ietf-idr-bgp-ipv6-rt-constrain] extends the NLRI definition to accommodate longer IPv6 Specific Route Target extended communities [RFC5701].

The scope of an RTC NLRI RT:X:Y spans all address family routes with RT:X:Y. This design choice makes RTC a standard route filtering mechanism for all BGP based VPN solutions. On the flip side, RTC can deliver optimal filtering only when the set of RTs used by an address family afi=x/safi=y is disjoint from the set of RTs used by any other address family afi=z/safi=w. The following section illustrates some of the problems when shared RTs are used.

1.1. Sub-optimality of route constrain in shared RT deployments

1.1.1. Unused route advertisement and retention

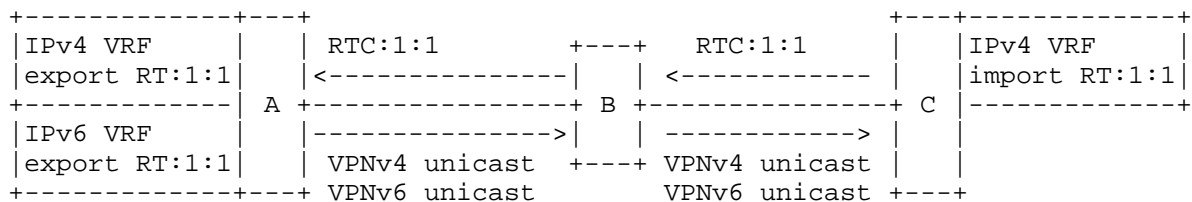


Figure 1: Sub-optimality: Unused route advertisement and retention

Suppose speaker A has local VPNv4 unicast and VPNv6 unicast routes, both with membership in RT:1:1. Speaker B is a route-reflector with client speaker C. Speaker C only needs VPNv4 unicast routes with RT:1:1 for its VPN with membership in RT:1:1. Speaker C sends an RTC NLRI RT:1:1 to speaker B and speaker B in turn sends an RTC NLRI RT:1:1 to speaker A. Speaker A sends its VPNv4 unicast and VPNv6 unicast routes with RT:1:1 to speaker B. Speaker B retains routes from both address families and sends all of them to speaker C. Speaker C retains only the VPNv4 unicast routes and drops VPNv6 unicast routes.

In this deployment, advertisement of VPNv6 unicast routes are not needed, but they are still advertised. Second, speaker B (the route-reflector) does not need to retain any VPNv6 unicast route since its only client, speaker C, does not need them. But speaker B still retains the VPNv6 unicast routes. Therefore, route constrain does not lead to optimal route filtering.

1.1.2. Need for route-refresh on some configuration changes

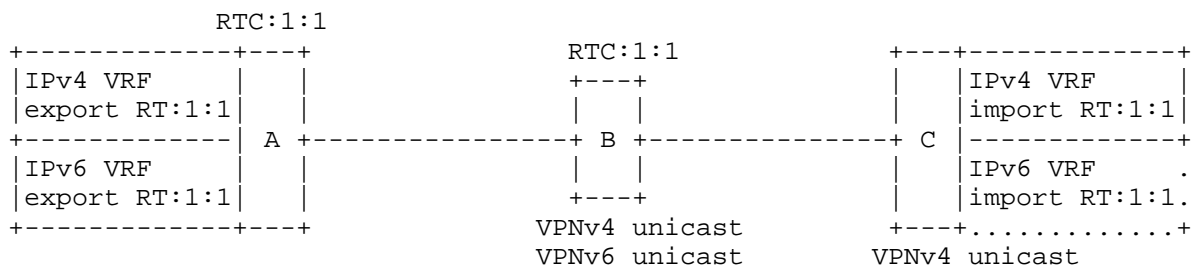


Figure 2: Sub-optimality: Route refresh

In the deployment shown in Figure 1, in steady state speaker C retains the VPNv4 unicast routes for RT:1:1 and speaker B has an RTC NLRI path for RT:1:1 from speaker C. Now suppose the operator adds a new IPv6 VRF on speaker C that imports RT:1:1 depicted in Figure 2. Now speaker C needs to get the VPNv6 unicast routes from speaker B with RT:1:1. At this point, if speaker C advertises the RTC NLRI RT:1:1 to speaker B again, speaker B would receive an identical path from speaker C. Standard BGP implementation practice is to ignore identical updates (i.e., not mark the local prefix for further processing), in which case speaker B will not send the VPN routes to speaker C again. Thus, speaker C would need to send a route refresh message to speaker B and receive all VPNv6 routes (this set of routes is larger than necessary if RT:1:1 is not the first RT imported by an IPv6 VRF on speaker C).

There are possible implementation tricks to get around this issue (e.g., readvertise the RTC NLRI RT:1:1 from speaker C after changing an attribute such as local-preference). However, a standardized solution without possible side-effects is much more preferable.

1.2. Solutions

1.2.1. Use of disjoint set of RTs

Not using shared RTs by ensuring that the set of RTs used by one address family is disjoint from the set of RTs used by any other address family avoids the problems. However, this constraint poses a burden on the network operation, especially in large networks that are run by multiple loosely coupled departments, where configurations change frequently. Therefore, a protocol level solution is more preferable.

1.2.2. New route constrain NLRI

[I-D.dong-idr-vpn-route-constrain] proposes a new NLRI format that includes the safi value (among other fields) and use different afi values during capability negotiations. This is not a backward compatible solution. Given many existing (large) deployments of [RFC4684] based multi-vendor networks, backward compatibility is necessary. In addition, [I-D.ietf-idr-bgp-ipv6-rt-constrain] solves the IPv6 specific RT issue in a backward compatible manner that [I-D.dong-idr-vpn-route-constrain] addresses.

1.2.3. Using BGP Path Attribute

The list of afi-safi that a speaker needs in the scope for a given RT can be carried in the path attribute of the RTC NLRI. This approach does not require any change in the NLRI format as the new information is not carried in the NLRI making the approach backward compatible. In addition, RTC being a hop-by-hop technique by nature, best path selection done on, say, a route-reflector, does not lead to missing information. In this document, we adopt this approach which leads to a light-weight, backward compatible solution. In addition, if the policy language supported on the BGP speakers allow attaching arbitrary extended communities to a route, then the proposed solution can be deployed on edge routers (i.e., on leafs of the VPN distribution graph) even without any software upgrade.

While one could define a new BGP path attribute to carry the list of afi-safi as the scope, use of communities suffices for the present purpose. Specifically, we propose using a new type of opaque extended community called AFI-SAFI extended community to encode each afi-safi pair that is of interest to the speaker and use multiple extended communities to form the list. Extended communities instead of standard communities are used since the latter are used widely by the providers for communicating internal information.

2. AFI-SAFI Extended Community

The AFI-SAFI Community is a Non-Transitive Opaque Extended Community ([RFC4360], [I-D.ietf-idr-extcomm-iana]) defined as follows:

Type Field:

The value of the high-order octet of the extended Type Field is 0x43, which indicates that it is non-transitive. The value of low-order octet of the extended type field for this community is TBD.

Value Field:

The first 3 octets of the Value field contains two sub-fields, described below. The last 3 octets of the Value field are reserved. The originator of an AFI-SAFI community must set the reserved octets to 0, and a receiver of an AFI-SAFI community must ignore the reserved octets.

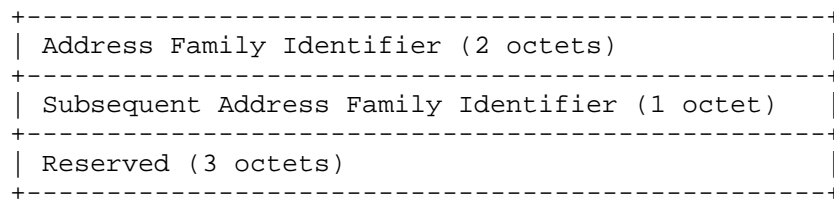


Figure 3: AFI-SAFI Extended Community Format

We denote an AFI-SAFI extended community with Address Family Identifier field x and Subsequent Address Family Identifier field y as AFI-SAFI-EC:(x/y).

A route carrying AFI-SAFI-EC:(x/y) implies that the route is correlated to the address family with index x/y . The route itself may belong to a different address family. The semantics of the correlation is context dependent. This document defines the correlation semantics for route constrain routes carrying AFI-SAFI-EC.

3. Operation

Suppose BGP speakers A and B have negotiated route constrain capability. Speaker A receives an RTC NLRI $RT:X:Y$ from speaker B with set S of AFI-SAFI-EC attached. We define the following semantics:

- o If S is empty, then speaker A SHOULD send all otherwise eligible routes from all address families to speaker B.
- o If S is nonempty, then speaker A SHOULD send an otherwise eligible AFI=x/SAFI=y route with RT:X:Y to speaker B ONLY if S contains AFI-SAFI-EC:(x/y).
- o Suppose speaker A has RTC route RT:X:Y with path i with AFI-SAFI-EC set S_i, for i=1..n.
 - * If all S_i are nonempty, then speaker A attaches AFI-SAFI-EC set S to RTC NLRI RT:X:Y that it advertises to its neighbors, where S is the union of S_i, i=1..n.
 - * If there is an empty S_i, then speaker A does not attach any AFI-SAFI-EC to RTC NLRI RT:X:Y that it advertises to its neighbors.

The essential idea is to treat an RTC NLRI path with no AFI-SAFI-EC attached as a path with AFI-SAFI-EC for all address families attached. We illustrate the rules with a couple of examples shown in Figure 4 and Figure 5.

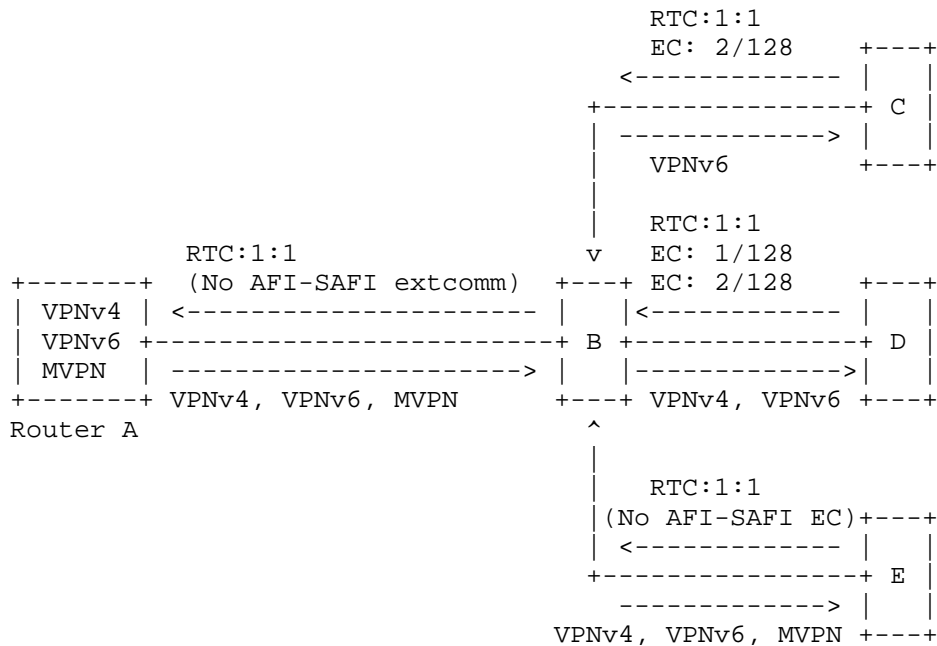


Figure 4: Operational rule in asymmetric cases

In Figure 4, speaker B receives RTC NLRI RT:1:1 from neighbors C, D and E. Neighbor C attaches AFI-SAFI-EC:(2/128) and neighbor D attaches AFI-SAFI-EC:(1/128) and AFI-SAFI-EC:(2/128). Neighbor E does not attach any AFI-SAFI-EC to its route.

From the received set of AFI-SAFI-EC, speaker B knows that among all the routes with RT:1:1, speaker C needs only VPNv6 unicast routes and speaker D needs only VPNv4 unicast and VPNv6 unicast routes. However, speaker B does not know which address family routes node E needs. Therefore, speaker B must request routes with RT:1:1 for all address families (whose capabilities have been negotiated) from its neighbors. So speaker B advertises RTC NLRI RT:1:1 to speaker A with no AFI-SAFI-EC.

Speaker A has VPNv4 unicast, VPNv6 unicast and MVPN routes with RT:1:1 that are eligible for sending to speaker B. Speaker A sends all those routes to speaker B.

Speaker B now has VPNv4 unicast, VPNv6 unicast and MVPN routes with RT:1:1 that are eligible for sending to neighbors C, D and E. Among those routes, speaker B sends only the VPNv6 unicast routes to neighbor C, VPNv4 unicast and VPNv6 unicast routes to neighbor D and

VPNv4 unicast, VPNv6 unicast and MVPN (i.e., all) routes to neighbor E.

If speaker E actually only needs VPNv4 unicast routes, it would drop the VPNv6 unicast and MVPN routes it receives from speaker B for RT 1:1. This sub-optimal behavior improves when speaker E also uses AFI-SAFI-EC to signal the scope for RT:1:1 and the "island" of contiguous AFI-SAFI-EC users expands.

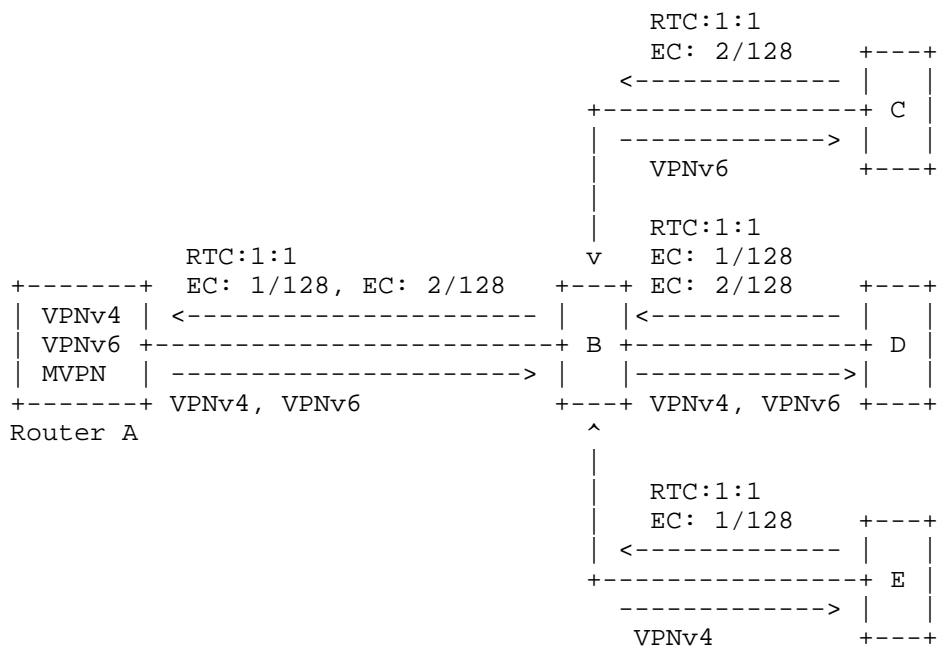


Figure 5: Operational rule in symmetric cases

In Figure 5, speaker B receives RTC NLRI RT:1:1 from neighbors C, D and E. Neighbor C attaches AFI-SAFI-EC:(2/128), neighbor D attaches AFI-SAFI-EC:(1/128) and AFI-SAFI-EC:(2/128), and neighbor E attaches AFI-SAFI-EC:(1/128).

From the received set of AFI-SAFI-EC, speaker B knows that among all routes with RT:1:1, speaker C needs only VPNv6 unicast routes, speaker D needs only VPNv4 unicast and VPNv6 unicast routes, and speaker E needs only VPNv4 unicast routes. Speaker B advertises RTC NLRI RT:1:1 to speaker A with AFI-SAFI-EC:(1/128) and AFI-SAFI-EC:(2/128), which is the union of all AFI-SAFI-EC from all paths of the RTC NLRI.

Speaker A therefore sends only VPNv4 unicast and VPNv6 unicast routes to speaker B.

Among those routes with RT:1:1 that speaker B has, speaker B sends only the VPNv6 unicast routes to neighbor C, VPNv4 unicast and VPNv6 unicast routes to neighbor D and only VPNv4 unicast routes to neighbor E. Therefore, when all speakers use AFI-SAFI-EC, optimal route filtering is restored even in shared RT deployments.

4. IANA Considerations

This document requests assignment of a codepoint from the Non-Transitive Opaque Extended Community Sub-Types registry for AFI-SAFI extended community.

5. Manageability Considerations

This section is structured as recommended in [RFC5706].

5.1. Configuration Management

TBD.

5.2. Operational Considerations

TBD.

6. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See the 'Security Considerations' section of [RFC4271] for a discussion of BGP security. Also refer to [RFC4272] and [I-D.ietf-karp-routing-tcp-analysis] for analysis of security issues for BGP.

7. Acknowledgements

TBD.

8. References

8.1. Normative References

[I-D.ietf-idr-extcomm-iana]
Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", draft-ietf-idr-extcomm-iana-02 (work in progress), December 2013.

- [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.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, January 2006.
- [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.

8.2. Informative References

- [I-D.dong-idr-vpn-route-constrain]
Li, Z., Dong, J., Ni, H., Chen, M., and G. Liu,
"Constrained Route Distribution for BGP based Virtual Private Networks(VPNs)", draft-dong-idr-vpn-route-constrain-02 (work in progress), September 2010.
- [I-D.ietf-idr-bgp-ipv6-rt-constrain]
Patel, K., Raszuk, R., Djernaes, M., Dong, J., and M. Chen, "IPv6 Extensions for Route Target Distribution", draft-ietf-idr-bgp-ipv6-rt-constrain-04 (work in progress), August 2013.
- [I-D.ietf-karp-routing-tcp-analysis]
Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP and MSDP Issues According to KARP Design Guide", draft-ietf-karp-routing-tcp-analysis-07 (work in progress), April 2013.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.
- [RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, August 2009.

- [RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, November 2009.
- [RFC5706] Harrington, D., "Guidelines for Considering Operations and Management of New Protocols and Protocol Extensions", RFC 5706, November 2009.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, May 2012.

Authors' Addresses

Saikat Ray
Cisco Systems, Inc.
170, West Tasman Drive
San Jose, CA 95134
US

Email: sairay@cisco.com

Arjun Sreekantiah
Cisco Systems, Inc.
170, West Tasman Drive
San Jose, CA 95134
US

Email: asreekan@cisco.com

Network Working Group
Internet Draft
Category: Standard Track

X. Xu
H. Ni
Huawei

M. Boucadair
C. Jacquenet
France Telecom

N. So
Tata Communications

Y. Fan
China Telecom

Expires: July 2014

January 16, 2014

Performance-based BGP Routing Mechanism

draft-xu-idr-performance-routing-00

Abstract

The current BGP specification doesn't use network performance metrics (e.g., network latency) in the route selection decision process. This document describes a performance-based BGP routing mechanism in which network latency metric is taken as one of the route selection criteria. This routing mechanism is useful for those server providers with global reach to deliver low-latency network connectivity services to their customers.

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 July 16, 2014.

Copyright Notice

Copyright (c) 2013 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.

Conventions used in this document

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].

Table of Contents

1. Introduction	3
2. Terminology	3
3. Performance Route Advertisement	4
4. Capability Advertisement	5
5. Performance Route Selection	6
6. Deployment Considerations	6
7. Security Considerations	6
8. IANA Considerations	7
9. Acknowledgements	7
10. References	7
10.1. Normative References	7
10.2. Informative References	7
Authors' Addresses	8

1. Introduction

Network performance, especially network latency is widely recognized as one of major obstacles in migrating business applications to the cloud, especially in the case where the network paths between cloud users and cloud data centers traverse more than one Autonomous System (AS), and would therefore stretch the forwarding path. However, the current Border Gateway Protocol (BGP) specification [RFC4271] which is used for path selection across ASes (Autonomous Systems) doesn't use network performance metrics (e.g., network latency) in the route selection process. As such, the best route selected based upon the existing BGP route selection criteria may not be the best from the customer experience perspective.

This document describes a performance-based BGP routing mechanism in which network performance metrics are conveyed as additional path attributes of the Network Layer Reachability Information (NLRI) and used in the route selection decisions. So far it's only the network latency metric that would be used in the performance-based route selection decisions. This mechanism is useful for those server providers with global reach, which usually own more than one AS, to deliver low-latency network connectivity services to their customers.

For the sake of simplicity, this document considers only one performance metric that's the network latency metric. The support of multiple attributes is out of scope of this document.

To make the performance routing paradigm and the vanilla routing paradigm coexist, performance routes should be exchanged as labeled routes as per [RFC3107] while using a specified Subsequent Address Family Identifier (SAFI). As such, network providers deploying such mechanism in their networks may provide the performance routing service as a value-added service to those customers with low latency need, while continually offering the vanilla routing service to the remaining customers as before.

A variant of this performance-based BGP routing is implemented [URL: <http://www.ist-mescal.org/roadmap/qbgp-demo.avi>].

2. Terminology

This memo makes use of the terms defined in [RFC4271].

Network latency indicates the amount of time it takes for a packet to traverse a given network path [RFC2679]. Provided a packet was forwarded along a path which contains multiple links and routers,

the network latency would be the sum of the transmission latency of each link (i.e., link latency), plus the sum of the internal delay occurred within each router (i.e., router latency) which includes queuing latency and processing latency. The sum of the link latency is also known as the cumulative link latency. In today's service provider networks which usually span across a wide geographical area, the cumulative link latency becomes the major part of the network latency since the total of the internal latency happened within each high-capacity router seems trivial compared to the cumulative link latency. In other words, the cumulative link latency could approximately represent the network latency in the above networks.

Furthermore, since the link latency is more stable than the router latency, such approximate network latency represented by the cumulative link latency is more stable. Therefore, if there was a way to calculate the cumulative link latency of a given network path, it is strongly recommended to use such cumulative link latency to approximately represent the network latency. Otherwise, the network latency would have to be measured frequently by some means (e.g., PING or other measurement tools).

3. Performance Route Advertisement

Performance routes SHOULD be exchanged between BGP peers by using a specified Subsequent Address Family Identifier (SAFI) of TBD (see IANA Section). Meanwhile, these routes SHOULD be carried as labeled routes as per [RFC3107].

A BGP speaker SHOULD NOT advertise performance routes to a particular BGP peer unless that peer indicates, through BGP capability advertisement (see Section 4), that it can process update messages with the specified SAFI field.

Network latency metric is attached to the performance routes as one additional path attribute, referred to as NETWORK_LATENCY path attribute, which is a well-known mandatory attribute. This attribute indicates the network latency in microseconds from the BGP speaker depicted by the NEXT_HOP path attribute to the address depicted by the NLRI prefix. The type code of this attribute is TBD (see IANA Section), and the value field is 4 octets in length. In some abnormal cases, if the cumulative link latency exceeds the maximum value of 0xFFFFFFFF, the value field SHOULD be set to 0xFFFFFFFF.

A BGP speaker SHOULD be configurable to enable or disable the origination/creation of performance routes. If enabled, a local latency value for a given to-be-originated performance route MUST be

configured to the BGP speaker so that it can be filled to the NETWORK_LATENCY attribute of that performance route.

When distributing a selected performance route learnt from one BGP peer to another, unless this BGP speaker has set itself as the NEXT_HOP of such route, the NETWORK_LATENCY path attribute of such route MUST NOT be modified. Otherwise when setting itself as the NEXT_HOP of such route, this BGP speaker SHOULD increase the value of the NETWORK_LATENCY path attribute by adding the network latency value from itself to the previous NEXT_HOP of such route. It is RECOMMENDED to use the cumulative link latency from this BGP speaker to the NEXT_HOP to represent the network latency between them if possible. Otherwise, the measured network latency between them can be used instead. It is RECOMMENDED that the type of network latency SHOULD be kept consistent across all these AS's (i.e., either cumulative link latency or measured network latency, choose one).

As for how to obtain the network latency to a given BGP NEXT_HOP is outside the scope of this document. However, note that the path latency to the NEXT_HOP SHOULD approximately represent the network latency of the exact forwarding path towards the NEXT_HOP. For example, if a BGP speaker uses a Traffic Engineering (TE) Label Switching Path (LSP) from itself to the NEXT_HOP, rather than the shortest path calculated by Interior Gateway Protocol (IGP), the latency to the NEXT_HOP SHOULD reflect the network latency of that TE LSP path, rather than the IGP shortest path.

To keep performance routes stable enough, a BGP speaker SHOULD use a configurable threshold of network latency fluctuation to suppress any update which would otherwise be triggered just by a minor network latency fluctuation below that threshold.

4. Capability Advertisement

A BGP speaker that uses multiprotocol extensions to advertise performance routes SHOULD use the Capabilities Optional Parameter, as defined in [RFC5492], to inform its peers about this capability.

The MP_EXT Capability Code, as defined in [RFC4760], is used to advertise the (AFI, SAFI) pairs available on a particular connection.

A BGP speaker that implements the Performance Routing Capability MUST support the BGP Labeled Route Capability, as defined in [RFC3107]. A BGP speaker that advertises the Performance Routing Capability to a peer using BGP Capabilities advertisement [RFC5492] does not have to advertise the BGP Labeled Route Capability to that peer.

5. Performance Route Selection

Performance route selection only requires the following modification to the tie-breaking procedures of the BGP route selection decision (phase 2) described in [RFC4271]: network latency metric comparison SHOULD be executed just ahead of the AS-Path Length comparison step.

Prior to executing the network latency metric comparison, the value of the NETWORK_LATENCY path attribute SHOULD be increased by adding the network latency from the BGP speaker to the NEXT_HOP of that route. In the case where a router reflector is deployed without next-hop-self enabled when reflecting received routes from one IBGP peer to other IBGP peer, it is RECOMMENDED to enable such route reflector to reflect all received performance routes by using some mechanisms such as [ADD-PATH], rather than reflecting only the performance route which is the best from its own perspective. Otherwise, it may result in a non-optimal choice by its clients and/or its IBGP peers.

The Loc-RIB of performance routing paradigm is independent from that of vanilla routing paradigm. Accordingly, the routing table of performance routing paradigm is independent from that of the vanilla routing paradigm. Whether performance routing paradigm or vanilla routing paradigm would be used for a given packet is a local policy issue which is outside the scope of this document.

6. Deployment Considerations

It is RECOMMENDED to deploy this performance-based BGP routing mechanism across multiple ASes which are within a single administrative domain. Within each AS, it is RECOMMENDED to deliver a packet from a BGP speaker to the BGP NEXT_HOP via tunnels, especially TE LSP tunnels. Furthermore, it is RECOMMENDED to use the latency metric carried in Unidirectional Link Delay Sub-TLV [OSPF-TE-EXT] [ISIS-TE-EXT] if possible, rather than the TE metric [RFC3630] [RFC5305] to perform the C-SPF calculation, unless the TE metric has already been set to the link latency metric. In this way, it could avoid the need for timely measurement of network latency between IBGP peers.

7. Security Considerations

In addition to the considerations discussed in [RFC4271], the following items should be considered:

Tweaking the value of the NETWORK_LATENCY by an illegitimate party may influence the route selection process. Means to check the integrity of BGP messages are RECOMMENDED.

Frequent updates of the NETWORK_LATENCY attribute may have a severe impact on the stability of the routing system. Such practice SHOULD be avoided.

8. IANA Considerations

A new BGP Capability Code for the Performance Routing Capability, a new SAFI specific for performance routing and a new path attribute for NETWORK_LATENCY are required to be allocated by IANA.

9. Acknowledgements

Thanks to Joel Halpern, Alvaro Retana, Jim Uttaro, Robert Raszuk, Eric Rosen, Qing Zeng, Jie Dong and Mach Chen for their valuable comments on the initial idea of this document.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.

10.2. Informative References

- [RFC5492] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.
- [RFC4760] Bates, T., Rekhter, Y., Chandra, R. and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

- [RFC2679] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999.
- [OSPF-TE-EXT] Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions-02 (work in progress), December 2012.
- [ISIS-TE-EXT] Previdi, S., Giacalone, S., Ward, D., Drake, J., Atlas, A., and C. Filsfils, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-previdi-isis-te-metric-extensions-02 (work in progress), October 2012.
- [RFC3630] Katz, D., Kompella, K., Yeung, D., "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [ADD-PATH] D. Walton, A. Retana, E. Chen, J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-09 (work in progress), October 2013.

Authors' Addresses

Xiaohu Xu
Huawei Technologies,
Beijing, China
Phone: +86-10-60610041
Email: xuxiaohu@huawei.com

Hui Ni
Huawei Technologies,
Beijing, China
Phone: +86-10-606100212
Email: nihui@huawei.com

Mohamed Boucadair
France Telecom
Rennes, France
EMail: mohamed.boucadair@orange.com

Christian Jacquenet

Orange
Rennes France
Email: christian.jacquetnet@orange.com

Ning So
Tata Communications
Plano, TX 75082, USA
Email: ning.so@tatacommunications.com

Yongbing Fan
China Telecom
Guangzhou, China.
Phone: +86 20 38639121
Email: fanyb@gsta.com