

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 1, 2015

J. Dong  
M. Chen  
Huawei Technologies  
June 30, 2014

BGP Extensions for Inter-AS Traffic Engineering (TE) Link Distribution  
draft-dong-idr-inter-as-te-link-distribution-00

Abstract

Protocol extensions to Interior Gateway Protocols (IGPs) have been specified for the flooding of Traffic Engineering (TE) information of the Inter-Autonomous System (AS) links into the local AS (RFC 5392 and RFC 5316), in which some information of the inter-AS links needs to be manually configured. This document proposes BGP extensions for dynamic advertisement of TE information of Inter-AS links between adjacent ASes. Such mechanism may also be used for the distribution of Inter-AS TE link information to some external entities, such as Path Computation Element (PCE).

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 January 1, 2015.

## 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. Carrying Inter-AS Link Information in BGP . . . . .	3
3. Operational Considerations . . . . .	4
4. IANA Considerations . . . . .	4
5. Security Considerations . . . . .	5
6. References . . . . .	5
6.1. Normative References . . . . .	5
6.2. Informative References . . . . .	5
Authors' Addresses . . . . .	6

## 1. Introduction

Protocol extensions to Interior Gateway Protocols (IGPs) have been specified for the flooding of Traffic Engineering (TE) information of the Inter-Autonomous System (AS) links in local AS [RFC5392] [RFC5316]. With those IGP extension mechanisms, some of the TE information of the inter-AS links, such as remote AS number and remote AS Border Router (ASBR) IDs are manually configured on the ASBRs of local AS. This requires additional human intervention and may be error-prone. Besides, an ASBR of local AS needs to generate a local link-state information for the inter-AS TE link, and also needs to 'proxy' for the remote ASBR to generate an additional link-state information, in order for the two-way check of the Inter-AS link during the path calculation. This introduces additional processing on the ASBR of local AS and the 'proxy' information may be not quite accurate. As bandwidth and other TE information of the Inter-AS links are useful for establishing TE Label Switched Paths (LSPs) across multiple ASes, such information needs to be dynamically exchanged between the peering ASes.

This document specifies BGP extensions for dynamic advertisement of Inter-AS TE link information between the adjacent ASes. This mechanism may also be used for the distribution of Inter-AS TE link information to some external entities, such as Path Computation Element (PCE).

## 2. Carrying Inter-AS Link Information in BGP

The Inter-AS link information is advertised in BGP UPDATE messages using the MP\_REACH\_NLRI and MP\_UNREACH\_NLRI attributes [RFC4760]. The Link-State NLRI defined in [I-D.ietf-idr-ls-distribution] is extended to carry the Inter-AS link information.

A new Protocol-ID is defined in the Link-State NLRI:

- o Protocol-ID = 7: Inter-AS, The NLRI information has been sourced from an Inter-AS connection

And a new Sub-TLV is defined in the Node Descriptor Sub-TLVs:

Sub-TLV Code Point	Description	Length
TBD	BGP Identifier	4

BGP Identifier is the 4-octet unsigned integer that indicates a BGP speaker, as defined in [RFC4271] [RFC6286].

The format of the link NLRI with Protocol-ID 7 is shown in the figure below:

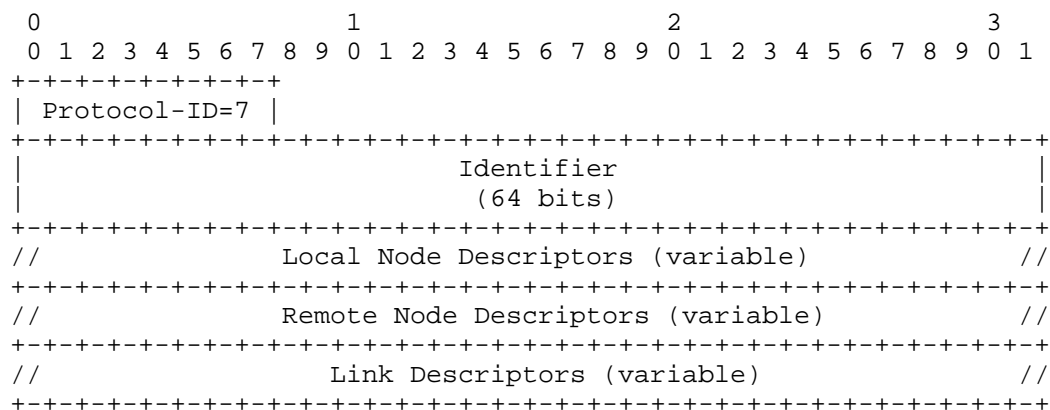


Figure 1. Inter-AS Link NLRI

The "Local Node Descriptors" field MUST contain the "Autonomous System" Sub-TLV defined in [I-D.ietf-idr-ls-distribution] to identify

the local AS number, and the "BGP Identifier" Sub-TLV defined in this document to identify the local ASBR.

The "Remote Node Descriptors" field MUST contain the "Autonomous System" Sub-TLV defined in [I-D.ietf-idr-ls-distribution] to identify the remote AS number, and the "BGP Identifier" Sub-TLV defined in this document to identify the remote ASBR.

For IPv4 Inter-AS link, the "Link Descriptors" field MUST use "IPv4 interface address" Sub-TLV to specify the local IPv4 address, and use "IPv4 neighbor address" Sub-TLV to specify the peering IPv4 address on the remote ASBR. The local and peering addresses are the IPv4 addresses used for the specific EBGP session between the local and remote ASBRs.

For IPv6 Inter-AS link, the "Link Descriptors" field MUST use "IPv6 interface address" Sub-TLV to specify the local IPv6 address, and use "IPv6 neighbor address" Sub-TLV to specify the peering IPv6 address on the remote ASBR. The local and peering addresses are the IPv6 addresses used for the specific EBGP session between the local and remote ASBRs.

The TE characteristics of the Inter-AS link, such as bandwidth, Shared Risk Link Group (SRLG), IPv4/IPv6 TE Router ID, etc., SHOULD be carried in the Link attribute TLVs of the BGP-LS attribute as specified in [I-D.ietf-idr-ls-distribution]. No further extension to the BGP-LS attribute is defined in this document.

### 3. Operational Considerations

The advertisement of Inter-AS TE link information SHOULD be constrained to only between the adjacent ASes connected by the Inter-AS link. BGP speakers SHOULD NOT advertise the Inter-AS TE link information received from the peering AS to any other ASes. The ASBR receiving the Inter-AS TE link information SHOULD redistribute such information into the IGP of the local AS, using mechanisms defined in [RFC5392] and [RFC5316].

The Inter-AS TE link information may optionally be advertised to an external entity, for example PCE. Such advertisement SHOULD be performed under agreement and policy control of the involved administrative domains.

### 4. IANA Considerations

IANA needs to assign one new Protocol-ID for "Inter-AS" from the BGP-TE/LS registry of Protocol-IDs.

IANA needs to assign one new Sub-TLV for "BGP Identifier" from the "node anchor, link descriptor and link attribute TLVs" registry.

## 5. 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.

## 6. References

### 6.1. 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-05 (work in progress), May 2014.
- [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.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, June 2011.

### 6.2. Informative References

- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, January 2006.
- [RFC5316] Chen, M., Zhang, R., and X. Duan, "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, December 2008.
- [RFC5392] Chen, M., Zhang, R., and X. Duan, "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, January 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

Jie Dong  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: jie.dong@huawei.com

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

Email: mach.chen@huawei.com

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: December 21, 2014

J. Dong  
M. Chen  
Huawei Technologies  
June 19, 2014

Extensions to RT-Constrain in Hierarchical Route Reflection Scenarios  
draft-dong-idr-rtc-hierarchical-rr-00

Abstract

The Route Target (RT) Constrain mechanism specified in RFC 4684 is used to build a route distribution graph in order to restrict the propagation of Virtual Private Network (VPN) routes. In network scenarios where hierarchical route reflection (RR) is used, the existing RT-Constrain mechanism cannot build a correct route distribution graph. This document refines the route distribution rules of RT-Constrain to address the hierarchical RR scenarios.

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 December 21, 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. Proposed Solution . . . . .	3
3. IANA Considerations . . . . .	4
4. Security Considerations . . . . .	4
5. Acknowledgements . . . . .	4
6. Normative References . . . . .	4
Authors' Addresses . . . . .	4

## 1. Introduction

The Route Target (RT) Constrain mechanism specified in RFC 4684 is used to build a route distribution graph in order to restrict the propagation of Virtual Private Network (VPN) routes. In network scenarios where hierarchical route reflection (RR) is used, the existing RT-Constrain mechanism cannot build a correct route distribution graph.

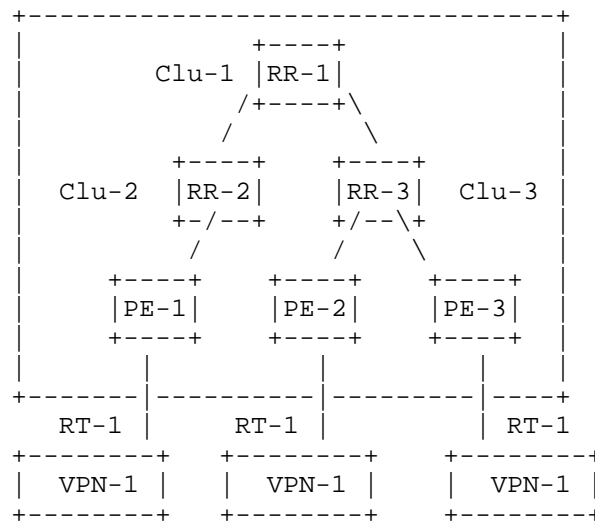


Figure 1. RT-Constrain with Hierarchical RR



As shown in Figure 1, hierarchical RRs are deployed in the network, RR-2 and RR-3 are route-reflectors of their connecting PEs, and are also the clients of RR-1. If each PE advertises RT membership information of RT-1 to the upstream RR, after the best path selection, both RR-2 and RR-3 would create the CLUSTER\_LIST attribute, prepend their local CLUSTER\_ID and then advertise the best path to RR-1 and their clients respectively.

On receipt of the RT-Constrain routes from RR-2 and RR-3, RR-1 will select one of the received routes as the best route, assume the route received from RR-2 is selected by RR-1 as the best path. Then RR-1 needs to advertise the best path to both RR-2 and RR-3 to create the route distribution graph of VPN-1. RR-1 would prepend its CLUSTER\_ID to the CLUSTER\_LIST of the path, and according to the rules in Section 3.2 of [RFC4684], it sets the ORIGINATOR\_ID to its own router-id, and the NEXT\_HOP to the local address for the session. Then RR-1 would advertise this route to both RR-2 and RR-3. On receipt of the RT-Constrain route from RR-1, RR-2 checks the CLUSTER\_LIST and find its own CLUSTER\_ID in the list, so this route will be ignored by RR-2. As a result, RR-2 will not form the outbound filter of RT-1 towards RR-1, hence will not advertise VPN routes with RT-1 to RR-1.

## 2. Proposed Solution

The problem described in the above section is that the best path is sent back to the BGP speaker which advertised the path and get discarded due to the BGP loop detection mechanisms. Since the advertisement of RT-Constrain route is to set up a route distribution graph and not to guide the data packet forwarding, all the available paths can be considered in setting up the route distribution graph, not just the best path. Thus in addition to the rules specified in section 3.2 of [RFC4684], the following rule applies in the advertisement of RT-Constrain routes:

- o When advertising an RT membership NLRI to a route-reflector client, if the best route as selected by the path selection procedure described in Section 9.1 of [RFC4271] is the path received from this client, and there are alternative paths received from other peers, the most disjoint alternative route SHOULD be advertised to that client; The most disjoint alternative path is the path whose CLUSTER\_LIST and ORIGINATOR\_ID attributes are different from the attributes of the best path.

With this additional rule, RR-1 in Figure 1 would advertise to RR-2 the RT-Constrain route received from RR-3, although the best route is received from RR-2. Thus RR-2 will not discard the RT-constrain

route received from RR-1, and the route distribution graph can be set up completely.

### 3. IANA Considerations

This document makes no request of IANA.

### 4. Security Considerations

This document does not change the security properties of BGP based VPNs and [RFC4684].

### 5. Acknowledgements

The authors would like to thank Yaqun Xiao for the discussion about RT-Constrain in hierarchical RR scenario.

### 6. Normative References

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

### Authors' Addresses

Jie Dong  
Huawei Technologies  
Huawei Campus, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: jie.dong@huawei.com

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

Email: mach.chen@huawei.com

Internet Engineering Task Force  
Internet-Draft  
Intended status: Informational  
Expires: December 15, 2014

W. George  
Time Warner Cable  
S. Amante  
Apple, Inc.  
June 13, 2014

Autonomous System (AS) Migration Features and Their Effects on the BGP  
AS\_PATH Attribute  
draft-ietf-idr-as-migration-01

## 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 December 15, 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. Requirements Language . . . . .	3
1.2. Documentation note . . . . .	3
2. ASN Migration Scenario Overview . . . . .	4
3. External BGP Autonomous System Migration Features . . . . .	6
3.1. Modify Inbound BGP AS_PATH Attribute . . . . .	6
3.2. Modify Outbound BGP AS_PATH Attribute . . . . .	8
3.3. Implementation . . . . .	9
4. Internal BGP Autonomous System Migration Features . . . . .	10
4.1. Internal BGP Alias . . . . .	10
4.2. Implementation . . . . .	13
5. Additional Operational Considerations . . . . .	14
6. Conclusion . . . . .	14
7. Acknowledgements . . . . .	15
8. IANA Considerations . . . . .	15
9. Security Considerations . . . . .	15
10. Appendix: Implementation report . . . . .	15
11. References . . . . .	16
11.1. Normative References . . . . .	16
11.2. Informative References . . . . .	16
Authors' Addresses . . . . .	17

## 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. These features are local to a given BGP Speaker 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, so slight variations between existing vendor implementations exist. 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.

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. 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].

#### 1.2. 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.

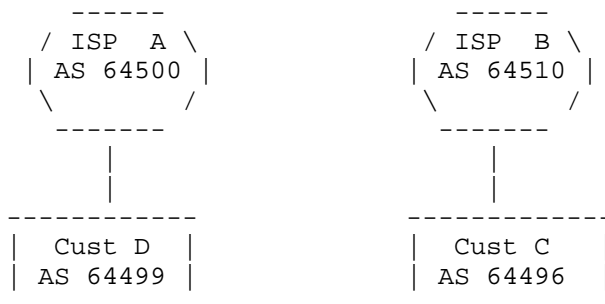


Figure 1: Before Migration

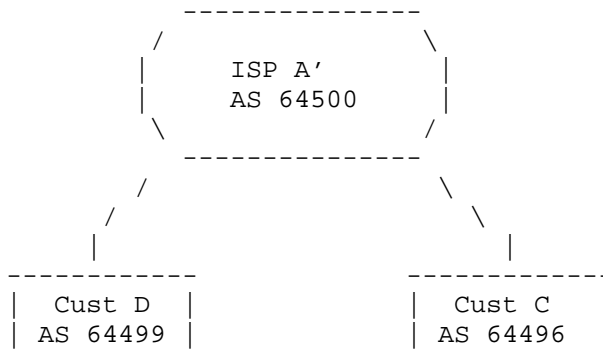


Figure 2: After Migration

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.

There are multiple implementations with equivalent features deployed and in use. Some documentation pointers to these implementations, as well as additional documentation on migration scenarios can be found in the appendix. 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. 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).

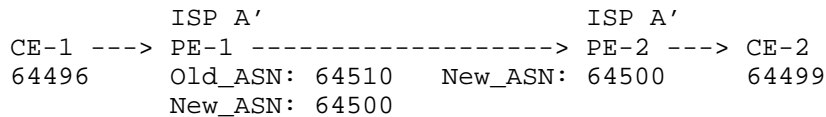
The next step is for ISP B to reconfigure its PE router(s) so that each of its eBGP sessions toward all eBGP speakers with a feature called "Local AS". This feature allows ISP B's PE router to re-establish a eBGP session toward the existing CE devices using the legacy AS, AS 64510, in the eBGP session 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
  
```

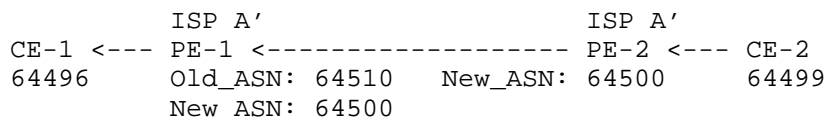
As a result of the "Local AS No Prepend" configuration, on PE-1, CE-2 will see an AS\_PATH of: 64500 64496. CE-2 will not receive a BGP UPDATE containing AS 64510 in the AS\_PATH. (If only the "local-as 64510" feature was configured without the keyword "no-prepend" on PE-1, then CE-2 would see an AS\_PATH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

### 3.2. Modify Outbound BGP AS\_PATH Attribute

The previous feature, "Local AS No Prepend", was only designed to modify the AS\_PATH Attribute received by the ISP in updates from CE devices, when CE devices still have an eBGP session established with the ISPs legacy AS, (AS64510). In some existing implementations, "Local AS No Prepend" 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 was designed to resolve this problem (continuing the use of Cisco CLI in the examples, it is called "Replace AS" in the examples below). This feature allows ISP A' to prevent routers configured with this feature from appending 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.

### 3.3. Implementation

While multiple implementations already exist, the following should document the expected behavior such that a new implementation of this feature could be done on other platforms.

These features MUST be configurable on a per-neighbor or per peer-group basis to allow for maximum flexibility. When this feature set is invoked, an ASN that is different from the globally-configured ASN is provided as a part of the command as exemplified above. To implement this feature, a BGP speaker MUST send BGP OPEN messages to the configured eBGP peer using the ASN configured for this session as the value sent in MY ASN. The speaker MUST NOT use the ASN configured globally within the BGP process as the value sent in MY ASN in the OPEN message. This will avoid the BGP OPEN Error message BAD PEER AS, and is typically used to re-establish eBGP sessions with peers expecting the legacy ASN after a router has been moved to a new ASN. Additionally, when the BGP speaker configured with this feature receives updates from its neighbor, it MUST append the configured ASN in the AS\_PATH attribute before processing the update as normal. Note that processing the update as normal will include appending the globally configured ASN to the AS\_PATH, thus processing this update will result in the addition of two ASNs to the AS\_PATH attribute. Similarly, for outbound updates sent by the configured BGP speaker to its neighbor, the speaker MUST append the configured ASN to the AS\_PATH attribute, adding to the existing global ASN in the AS\_PATH, for a total of two ASNs added to the AS\_PATH.

Two options exist to manipulate the behavior of this feature. They modify the behavior as described below:

No prepend inbound - When the BGP speaker configured with this option receives inbound updates from its neighbor, it MUST NOT append the configured ASN in the AS\_PATH attribute and instead MUST append only the globally configured ASN.

No prepend outbound - When the BGP speaker configured with this option generates outbound BGP updates to the configured peer, the BGP speaker MUST remove the globally configured ASN from the AS\_PATH attribute, and MUST append the locally configured ASN to the AS\_PATH attribute before sending outbound BGP updates to the configured peer.

While the exact command syntax is an implementation detail beyond the scope of this document, the following consideration may be helpful for implementers: Implementations MAY integrate the behavior of the options described above into a single command that addresses both inbound and outbound updates, but if this is done, implementations MUST provide a method to select its applicability to inbound updates, outbound updates, or updates in both directions. Several existing implementations use separate commands (e.g. local-as no-prepend vs local-as replace-as) for maximum flexibility in controlling the behavior on the session to address the widest range of possible migration scenarios.

#### 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, 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. NB: Cisco doesn't have an exact equivalent to "Internal BGP Alias", but the combination of the Cisco features iBGP local-AS and dual-as provides similar functionality.

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. Note 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 that it 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.)

#### 4.2. Implementation

When configured with this feature, a BGP speaker MUST accept BGP OPEN and establish an iBGP session from configured iBGP peers if the ASN value in MY ASN is either the globally configured ASN or the locally configured ASN provided in this command. Additionally, a BGP speaker configured with this feature MUST send its own BGP OPEN using both the globally configured and the locally configured ASN in MY ASN. To avoid potential deadlocks when two BGP speakers are attempting to establish a BGP peering session and are both configured with this feature, the speaker SHOULD send BGP OPEN using the globally configured ASN first, and only send a BGP OPEN using the locally configured ASN as a fallback if the remote neighbor responds with the BGP error BAD PEER ASN. In each case, the BGP speaker MUST treat updates sent and received to this peer as if this was a natively configured iBGP session, as defined by [RFC4271] and [RFC4456].

Implementations of this feature MAY integrate the functionality from the eBGP features (Section 3) section as a part of this command in order to simplify support for eBGP migrations as well as iBGP migrations, such that an eBGP session to a configured neighbor could be established via either the global ASN or the locally configured ASN. If the eBGP session is established with the global ASN, no modifications to AS\_PATH are required, but if the eBGP session is established with the locally configured ASN, the modifications discussed in eBGP features (Section 3) MUST be implemented to properly manipulate the AS\_PATH.



## 5. Additional Operational Considerations

This document describes several 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, David Farmer, Jaroslaw Adam Gralak, Gunter Van de Velde, and Juan Alcaide 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. Appendix: Implementation report

As noted elsewhere in this document, this set of migration features has multiple existing implementations in wide use.

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

This is not intended to be an exhaustive list, as equivalent features do exist in other implementations, however the authors were unable to find publicly available documentation of the vendor-specific implementation to reference.

## 11. References

### 11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5398] Huston, G., "Autonomous System (AS) Number Reservation for Documentation Use", RFC 5398, December 2008.

### 11.2. 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](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/c/en/us/td/docs/ios-xml/ios/iproute\\_bgp/configuration/xe-3s/asr1000/irg-xe-3s-asr1000-book/irg-dual-as.html](http://www.cisco.com/c/en/us/td/docs/ios-xml/ios/iproute_bgp/configuration/xe-3s/asr1000/irg-xe-3s-asr1000-book/irg-dual-as.html)>.
- [JUNIPER] Juniper Networks, Inc., "Configuring the BGP Local Autonomous System Attribute", 2012, <[http://www.juniper.net/techpubs/en\\_US/junos13.3/topics/concept/bgp-local-as-introduction.html](http://www.juniper.net/techpubs/en_US/junos13.3/topics/concept/bgp-local-as-introduction.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,  
5543, 5701, 6368, 6790 (if  
approved)  
Intended status: Standards Track  
Expires: December 15, 2014

E. Chen, Ed.  
Cisco Systems, Inc.  
J. Scudder, Ed.  
Juniper Networks  
P. Mohapatra  
Sproute Networks  
K. Patel  
Cisco Systems, Inc.  
June 13, 2014

Revised Error Handling for BGP UPDATE Messages  
draft-ietf-idr-error-handling-13

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, 5543, 5701, 6368 and 6790.

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 December 15, 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 . . . . .	4
2. Error-Handling Approaches . . . . .	4
3. Revision to BGP UPDATE Message Error Handling . . . . .	4
4. Attribute Length Fields . . . . .	6
5. Parsing of NLRI Fields . . . . .	7
5.1. Encoding NLRI . . . . .	7
5.2. Missing NLRI . . . . .	7
5.3. Syntactic Correctness of NLRI Fields . . . . .	8
5.4. Typed NLRI . . . . .	8
6. Operational Considerations . . . . .	9
7. Error Handling Procedures for Existing Attributes . . . . .	10
7.1. ORIGIN . . . . .	10
7.2. AS_PATH . . . . .	10
7.3. NEXT_HOP . . . . .	11
7.4. MULTI_EXIT_DISC . . . . .	11
7.5. LOCAL_PREF . . . . .	11
7.6. ATOMIC_AGGREGATE . . . . .	12
7.7. AGGREGATOR . . . . .	12

7.8. Community . . . . .	12
7.9. ORIGINATOR_ID . . . . .	12
7.10. CLUSTER_LIST . . . . .	13
7.11. MP_REACH_NLRI . . . . .	13
7.12. MP_UNREACH_NLRI . . . . .	14
7.13. Traffic Engineering path attribute . . . . .	14
7.14. Extended Community . . . . .	14
7.15. IPv6 Address Specific BGP Extended Community Attribute . . . . .	14
7.16. BGP Entropy Label Capability Attribute . . . . .	15
7.17. ATTR_SET . . . . .	15
8. Guidance for Authors of BGP Specifications . . . . .	15
9. IANA Considerations . . . . .	16
10. Security Considerations . . . . .	16
11. Acknowledgements . . . . .	16
12. References . . . . .	16
12.1. Normative References . . . . .	16
12.2. Informative References . . . . .	18
Authors' Addresses . . . . .	18

## 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. To make matters worse, in such cases although the problematic attributes may have originated with a single update transmitted by a single BGP speaker, by the time they encounter a router that checks them they may have been replicated many times, and thus may cause the reset of many peering sessions. Thus the damage inflicted may be multiplied manyfold.

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], [RFC5543], [RFC5701], [RFC6368] and [RFC6790] 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. Error-Handling Approaches

In this document we refer to four different approaches to handling errors found in BGP path attributes. They are as follows (listed in order, from the one with the "strongest" action to the one with the "weakest" action):

- o Session reset: This is the approach used throughout the base BGP specification [RFC4271], where a NOTIFICATION is sent and the session terminated.
- o AFI/SAFI disable: [RFC4760] specifies a procedure for disabling a particular AFI/SAFI.
- o Treat-as-withdraw: In this approach, the UPDATE message containing the path attribute in question 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].
- o Attribute discard: In this approach the malformed attribute MUST be discarded and the UPDATE message continues to be processed. This approach must not be used except in the case of an attribute that has no effect on route selection or installation.

## 3. Revision to BGP UPDATE Message Error Handling

This specification amends [RFC4271] Section 6.3 in a number of ways. See also Section 7 for treatment of specific path attributes.

- a. The first paragraph 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.

- b. Error handling for the following case 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.

- c. Attribute Flag error handling is revised as follows:

## Old Text:

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).

## New Text:

If the value of either the Optional or Transitive bits in the Attribute Flags is in conflict with their specified values, then the attribute MUST be treated as malformed and the treat-as-withdraw approach used, unless the specification for the attribute mandates different handling for incorrect Attribute Flags.

- d. If any of the well-known mandatory attributes are not present in an UPDATE message, then "treat-as-withdraw" MUST be used. (Note that [RFC4760] reclassifies NEXT\_HOP as what is effectively discretionary.)

- e. "Treat-as-withdraw" MUST be used for the cases that specify a session reset and involve any of the attributes ORIGIN, AS\_PATH, NEXT\_HOP, MULTI\_EXIT\_DISC, or LOCAL\_PREF.
- f. "Attribute discard" MUST be used for any of the cases that specify a session reset and involve ATOMIC\_AGGREGATE or AGGREGATOR.
- g. If the MP\_REACH\_NLRI attribute or the MP\_UNREACH\_NLRI [RFC4760] 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 (whether recognized or unrecognized) 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.
- h. When multiple attribute errors exist in an UPDATE message, if the same approach (either "session reset", "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.
- i. The Withdrawn Routes field MUST be checked for syntactic correctness in the same manner as the NLRI field. This is discussed further below, and in Section 5.3.
- j. 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 -- what this entails is discussed in more detail in Section 5. If this is not possible, the procedures of [RFC4271] and/or [RFC4760] continue to apply, meaning that the "session reset" approach (or the "AFI/SAFI disable" approach) MUST be followed.

#### 4. 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 octets remain, or fewer than four octets, 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.

For all path attributes other than those specified as having an attribute length that may be zero it SHALL be considered a syntax error for the attribute to have a length of zero. (Of the path attributes considered in this specification, only AS\_PATH and ATOMIC\_AGGREGATE may validly have an attribute length of zero.)

## 5. Parsing of NLRI Fields

### 5.1. Encoding NLRI

To facilitate the determination of the NLRI field in an UPDATE with a malformed attribute:

- o The MP\_REACH\_NLRI or MP\_UNREACH\_NLRI attribute (if present) SHALL be encoded as the very first path attribute in an UPDATE.
- o An UPDATE message MUST NOT contain more than one of the following: non-empty Withdrawn Routes field, non-empty Network Layer Reachability Information field, MP\_REACH\_NLRI attribute, and MP\_UNREACH\_NLRI attribute.

Since older BGP speakers may not implement these restrictions, an implementation MUST still be prepared to receive these fields in any position or combination.

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.

### 5.2. Missing NLRI

[RFC4724] specifies an End-of-RIB message ("EoR") that can be encoded as an UPDATE message that contains only a MP\_UNREACH\_NLRI attribute that encodes no NLRI (it can also be a completely empty UPDATE message in the case of the "legacy" encoding). In all other well-specified cases, an UPDATE either carries only withdrawn routes (either in the Withdrawn Routes field, or the MP\_UNREACH\_NLRI attribute), or it advertises reachable routes (either in the Network

Layer Reachability Information field, or the MP\_REACH\_NLRI attribute).

Thus, if an UPDATE message is encountered that does contain path attributes other than MP\_UNREACH\_NLRI and doesn't encode any reachable NLRI, we cannot be confident that the NLRI have been successfully parsed as Section 3 (j) requires. For this reason, if any path attribute errors are encountered in such an UPDATE message, and if any encountered error specifies an error-handling approach other than "attribute discard", then the "session reset" approach MUST be used.

### 5.3. 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\_NLRI or MP\_UNREACH\_NLRI 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.
- o The attribute flags of the attribute are inconsistent with those specified in [RFC4760].
- o The length of the MP\_UNREACH\_NLRI attribute is less than 3, or the length of the MP\_REACH\_NLRI attribute is less than 5.

### 5.4. Typed NLRI

Certain address families, for example MCAST-VPN [RFC6514], MCAST-VPLS [RFC7117] and EVPN [I-D.ietf-l2vpn-evpn] have NLRI that are typed. Since supported type values within the address family are not expressed in the MP-BGP capability [RFC4760], it is possible for a BGP speaker to advertise support for the given address family and

sub-address family while still not supporting a particular type of NLRI within that AFI/SAFI.

A BGP speaker advertising support for such a typed address family MUST handle routes with unrecognized NLRI types within that address family by discarding them, unless the relevant specification for that address family specifies otherwise.

## 6. 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.

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.

## 7. Error Handling Procedures for Existing Attributes

In the following subsections, we elaborate on the conditions for error-checking various path attributes, and specify what approach(es) should be used to handle malformations. It is possible that implementations may apply other error checks not contemplated here. If so, the error handling approach given here should generally be applied.

This section addresses all path attributes that are defined at the time of this writing, that were not defined with error-handling consistent with Section 8, and that are not marked as "deprecated" in [IANA-BGP-ATTRS]. Attributes 17 (AS4\_PATH), 18 (AS4\_AGGREGATOR), 22 (PMSI\_TUNNEL), 23 (Tunnel Encapsulation Attribute), 26 (AIGP), 27 (PE Distinguisher Labels) and 29 (BGP-LS Attribute) do have error-handling consistent with Section 8 and thus are not further discussed herein. Attributes 11 (DPA), 12 (ADVERTISER), 13 (RCID\_PATH / CLUSTER\_ID), 19 (SAFI Specific Attribute), 20 (Connector Attribute) and 21 (AS\_PATHLIMIT) are deprecated and thus are not further discussed herein.

### 7.1. 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".

### 7.2. AS\_PATH

An AS\_PATH is considered malformed if an unrecognized segment type is encountered, or if it contains a malformed segment. A segment is considered malformed if any of the following obtains:

- o There is an overrun, where the path segment length field of the last segment encountered would cause the Attribute Length to be exceeded.
- o There is an underrun, where after the last successfully-parsed segment, there is only a single octet remaining (that is, there is not enough unconsumed data to provide even an empty segment header).
- o It has a path segment length field of zero.

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

[RFC4271] also says that an implementation optionally "MAY check whether the leftmost ... AS in the AS\_PATH attribute is equal to the autonomous system number of the peer that sent the message". A BGP implementation SHOULD also handle routes that violate this check using "treat-as-withdraw", but MAY follow the session reset behavior if configured to do so.

### 7.3. NEXT\_HOP

According to [RFC4271] the attribute is considered malformed if it is syntactically incorrect. To quote from that document, "Syntactic correctness means that the NEXT\_HOP attribute represents a valid IP host address", but it does not go on to define what it means to be a "valid IP host address". Therefore:

An IP host address SHOULD be considered invalid if it appears in the "IANA IPv4 Special-Purpose Address Registry" [IANA-IPV4] and either the "destination" or the "forwardable" boolean in that registry is given as "false". An implementation SHOULD provide a means to modify the list of invalid host addresses by configuration -- these are sometimes referred to as "Martians".

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

### 7.4. MULTI\_EXIT\_DISC

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

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

### 7.5. LOCAL\_PREF

The error handling of [RFC4271] is revised as follows.

- o If the LOCAL\_PREF 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".

#### 7.6. 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".

#### 7.7. 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".

#### 7.8. Community

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

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

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

#### 7.9. 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".



## 7.10. 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 nonzero multiple of 4. If malformed, the UPDATE SHALL be handled using the approach of "treat-as-withdraw".

## 7.11. MP\_REACH\_NLRI

[RFC4760] references the error-handling of the base BGP specification for validation of the next hop. ("The rules for the next hop information are the same as the rules for the information carried in the NEXT\_HOP BGP attribute".) Thus just as in Section 7.3 we must consider what it means for the Next Hop field of the MP\_REACH attribute to be a "valid host address":

- o If the Next Hop field contains an IPv4 address (possibly as a sub-field), the field SHOULD be considered invalid if the IPv4 address appears in the "IANA IPv4 Special-Purpose Address Registry" [IANA-IPV4] and either the "destination" or the "forwardable" boolean in that registry is given as "false".
- o If the Next Hop field contains an IPv6 address (possibly as a sub-field), the field SHOULD be considered invalid if the IPv6 address appears in the "IANA IPv6 Special-Purpose Address Registry" [IANA-IPV6], the address is not an IPv4-mapped IPv6 address, and either the "destination" or the "forwardable" boolean in that registry is given as "false".
- o If the Next Hop field contains an IPv4-mapped IPv6 address (possibly as a sub-field), the field SHOULD be considered invalid unless the use of such addresses has been explicitly allowed for the particular AFI/SAFI that occurs in this MP\_REACH\_NLRI attribute. (E.g., see [RFC4659] and [RFC4798].)
- o If the Next Hop field is some other form of address, it should be considered invalid in circumstances analogous to the above -- if it is found in the relevant IANA special-purpose address registry (if any) and its "destination" or "forwardable" boolean is given as "false".

- o An implementation SHOULD provide a means to modify the list of invalid host addresses by configuration -- these are sometimes referred to as "Martians".

Section 3 and Section 5 provide further discussion of the handling of this attribute.

#### 7.12. MP\_UNREACH\_NLRI

Section 3 and Section 5 discuss the handling of this attribute.

#### 7.13. Traffic Engineering path attribute

We note that [RFC5543] does not detail what constitutes "malformation" for the Traffic Engineering path attribute. A future update to that specification may provide more guidance. In the interim, an implementation that determines (for whatever reason) that an UPDATE message contains a malformed Traffic Engineering path attribute MUST handle it using the approach of "treat-as-withdraw".

#### 7.14. Extended Community

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

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

#### 7.15. 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 not a nonzero 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.

#### 7.16. BGP Entropy Label Capability Attribute

The error handling of [RFC6790] is revised as follows.

No syntax errors are defined for the Entropy Label Capability attribute (ELCA). However, if any implementation does for some local reason determine that a syntax error exists with the ELCA, the error SHALL be handled using the approach of "attribute discard".

#### 7.17. ATTR\_SET

The final paragraph of Section 5 of [RFC6368] is revised as follows:

Old Text:

An UPDATE message with a malformed ATTR\_SET attribute SHALL be handled as follows. If its Partial flag is set and its Neighbor-Complete flag is clear, the UPDATE is treated as a route withdraw as discussed in [OPT-TRANS-BGP]. Otherwise (i.e., Partial flag is clear or Neighbor-Complete is set), the procedures of the BGP-4 base specification [RFC4271] MUST be followed with respect to an Optional Attribute Error.

New Text:

An UPDATE message with a malformed ATTR\_SET attribute SHALL be handled using the approach of "treat as withdraw".

Furthermore, the normative reference to [OPT-TRANS-BGP] in [RFC6368] is removed.

### 8. Guidance for Authors of BGP Specifications

A document that specifies a new BGP attribute MUST provide specifics regarding what constitutes an error for that attribute and how that error is to be handled. Allowable error-handling approaches are detailed in Section 2. The treat-as-withdraw approach is generally preferred. The document SHOULD also provide consideration of what debugging facilities may be required to permit issues caused by a malformed attribute to be diagnosed.

For any malformed attribute that 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.

Authors can refer to Section 7 for examples.

## 9. IANA Considerations

This document makes no request of IANA.

## 10. 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.

## 11. Acknowledgements

The authors wish to thank Juan Alcaide, Deniz Bahadir, Ron Bonica, Mach Chen, Andy Davidson, Bruno Decraene, Rex Fernando, Jeff Haas, Chris Hall, Joel Halpern, Dong Jie, Akira Kato, Miya Kohno, Tony Li, Alton Lo, Shin Miyakawa, Tamas Mondal, Jonathan Oddy, Tony Przygienda, Robert Raszuk, Yakov Rekhter, Eric Rosen, Shyam Sethuram, Rob Shakir, Naiming Shen, Adam Simpson, Ananth Suryanarayana, Kaliraj Vairavakkalai, Lili Wang and Ondrej Zajicek for their observations and discussion of this topic, and review of this document.

## 12. References

### 12.1. Normative References

- [IANA-BGP-ATTRS]  
"BGP Path Attributes", <<http://www.iana.org/assignments/bgp-parameters/bgp-parameters.xhtml#bgp-parameters-2>>.
- [IANA-IPV4]  
"IANA IPv4 Special-Purpose Address Registry",  
<<http://www.iana.org/assignments/iana-ipv4-special-registry>>.

- [IANA-IPV6] "IANA IPv4 Special-Purpose Address Registry",  
<<http://www.iana.org/assignments/iana-ipv6-special-registry>>.
- [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.
- [RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5543] Ould-Brahim, H., Fedyk, D., and Y. Rekhter, "BGP Traffic Engineering Attribute", RFC 5543, May 2009.
- [RFC5701] Rekhter, Y., "IPv6 Address Specific BGP Extended Community Attribute", RFC 5701, November 2009.
- [RFC6368] Marques, P., Raszuk, R., Patel, K., Kumaki, K., and T. Yamagata, "Internal BGP as the Provider/Customer Edge Protocol for BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 6368, September 2011.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.
- [RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, December 2012.

## 12.2. Informative References

- [I-D.ietf-l2vpn-evpn]  
Sajassi, A., Aggarwal, R., Bitar, N., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-07 (work in progress), May 2014.
- [RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, September 2006.
- [RFC4798] De Clercq, J., Ooms, D., Prevost, S., and F. Le Faucheur, "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, February 2007.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.
- [RFC7117] Aggarwal, R., Kamite, Y., Fang, L., Rekhter, Y., and C. Kodeboniya, "Multicast in Virtual Private LAN Service (VPLS)", RFC 7117, February 2014.

## Authors' Addresses

Enke Chen (editor)  
Cisco Systems, Inc.

Email: enkechen@cisco.com

John G. Scudder (editor)  
Juniper Networks

Email: jgs@juniper.net

Pradosh Mohapatra  
Sproute Networks

Email: mpradosh@yahoo.com

Keyur Patel  
Cisco Systems, Inc.

Email: keyupate@cisco.com

Network Working Group  
Internet-Draft  
Intended status: Informational  
Expires: January 2, 2015

Z. Li  
Z. Zhuang  
Huawei Technologies  
July 1, 2014

BGP Extensions for Service-Oriented MPLS Path Programming (MPP)  
draft-li-idr-mpls-path-programming-00

## Abstract

Service-oriented MPLS programming is to provide customized service process based on flexible label combinations. BGP will play an important role for MPLS path programming to allocate MPLS segment, download programmed MPLS path and the mapping of the service path to the transport path. This document defines BGP extensions to support service-oriented MPLS path programming.

## 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 January 2, 2015.

## 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. Terminology . . . . .	2
3. MPLS Segment Allocations . . . . .	3
4. Download of MPLS Path . . . . .	3
5. Download of Mapping of Service Path to Transport Path . . . . .	5
5.1. Extended Unicast Tunnel Attributes . . . . .	5
5.2. Extended PMSI Tunnel Attribute . . . . .	7
6. Capability Negotiation . . . . .	9
7. IANA Considerations . . . . .	9
8. Security Considerations . . . . .	9
9. References . . . . .	9
9.1. Normative References . . . . .	9
9.2. Informative References . . . . .	10
Authors' Addresses . . . . .	10

## 1. Introduction

Service-oriented MPLS programming proposed by [I-D.li-spring-mpls-path-programming] is to provide customized service process based on flexible label combinations. BGP will play an important role for MPLS path programming to allocate MPLS segment, download programmed MPLS path and the mapping of the service path to the transport path. This document defines BGP extensions to support service-oriented MPLS path programming.

## 2. Terminology

BGP: Border Gateway Protocol

EVPN: Ethernet VPN

L2VPN: Layer 2 VPN

L3VPN: Layer 3 VPN

MPP: MPLS Path Programming

MVPN: Multicast VPN



RR: Route Reflector

SDN: Software-Defined Network

S-EVPN: Segment-based EVPN

SR-Path: Segment Routing Path

NLRI: Network Layer Reachability Information

### 3. MPLS Segment Allocations

MPLS Segment is the component to compose the MPLS path. [I-D.li-spring-mpls-path-programming] proposes the use cases for service-oriented MPLS path programming which needs following MPLS segments:

1. MPLS path programming for unicast service
  - o MPLS Segment for VPN identification
  - o MPLS Segment for ECMP
  - o MPLS Segment for OAM (Source identification)
  - o MPLS Segment for Traffic Steering
2. MPLS path programming for multicast service
  - o MPLS Segment for MVPN identification
  - o MPLS Segment for Source identification
3. MPLS virtual network for services
  - o MPLS Segment for MPLS virtual network

These MPLS Segments are defined in individual drafts. It is out of the scope of this document.

### 4. Download of MPLS Path

According to the service requirements, the central controller can combine MPLS segments flexibly. Then it can download the service label combination for specific prefix related with the service. BGP extensions are necessary to advertise label stacks for prefix in NLRI field.

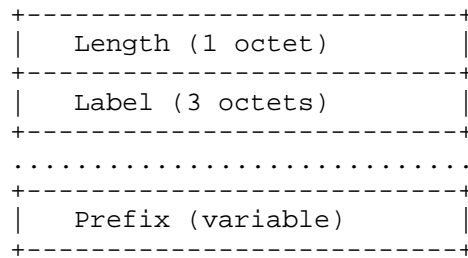


Figure 1: NLRI Definition in RFC3107

[RFC3107] defines above NLRI to advertise label binding for specific prefix. The label field can carry one or more labels. Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack". But for other AFI/SAFIs using label binding such as VPNv4, VPNv6, EVPN, MVPN, etc., it does not support the capability to carry more labels for the specific prefix. Moreover for the AFI/SAFIs which do not support label binding capability originally, but may possibly adopt MPLS path programming now, there is no label field in the NLRI. In order to support flexible MPLS path programming, this document defines and uses a new BGP attribute called the "Extended Label attribute". This is an optional transitive BGP attribute. The format of this attribute is defined as follows:

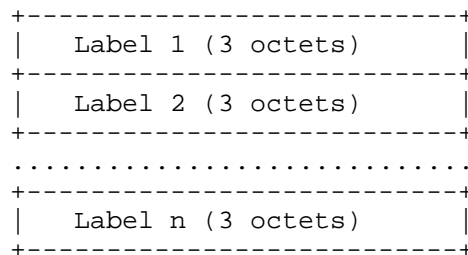


Figure 2: Extended Label Attribute

The Label field carries one or more labels (that corresponds to the stack of labels [[RFC3032]]). Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack" (as defined in [[RFC3032]]).

The Central Controller for MPLS path programming could build a route with Extended Label attribute and send it to the ingress routers.

Upon receiving such a route from the MPP Controller, the ingress router SHOULD select such a route as the best path. If a packet

comes into the ingress router and uses such a path, the ingress router will encapsulate the stack of labels which gets from the Extended Label Attribute of the route into the packet and forward the packet along the path.

The "Extended Label attribute" can be used for various BGP address families. Before using this attribute, firstly, it is necessary to negotiate the capability between two nodes to support MPLS path programming for a specific BGP address family. If negotiation fails, a node MUST NOT send this attribute and MUST discard this attribute when it receives.

## 5. Download of Mapping of Service Path to Transport Path

Since the transport path is also to satisfy the service bearing the requirement, it can also be programmed according to traffic engineering requirements of service. Or the transport path can be set up according to general traffic engineering requirements. Then there needs to be implements the mapping of the service path to the transport path. BGP Extensions is necessary that through the community attribute of BGP, the identifier of the transport path can be carried when distribute label stack for a specific prefix.

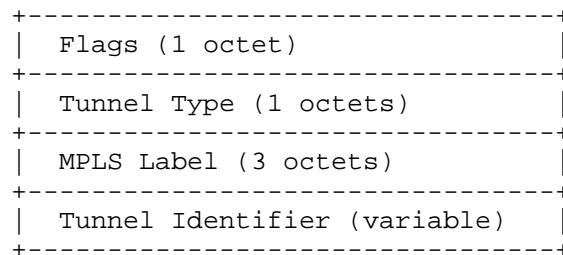


Figure 3: PMSI Tunnel Attribute in RFC6514

[RFC6514] defines the "P-Multicast Service Interface Tunnel (PMSI Tunnel) attribute". It is shown in the above figure. Since the attribute is always for the specific usage in BGP-based MVPN, it can not be applied to all possible use cases of service-oriented MPLS path programming. This document accordingly defines two new types of BGP attribute for both usage of unicast service path and the multicast service path: Extended Unicast Tunnel Attribute and Extended PMSI Tunnel Attribute.

### 5.1. Extended Unicast Tunnel Attributes

This document defines and uses a new BGP attribute called the "Extended Unicast Tunnel attribute". This is an optional transitive BGP attribute. The format of this attribute is defined as follows:

```

+-----+
| Flags (1 octet) |
+-----+
| Tunnel Type (1 octets) |
+-----+
| Tunnel Identifier (variable) |
+-----+
| Tunnel Specific Attributes (Variable)(Optional) |
+-----+

```

This document defines the following flags:

```

0 1 2 3 4 5 6 7
+-----+
| reserved |S|
+-----+

```

+ Unicast Tunnel Setup Required (S)

If the S flag is not set, the client node is just to map the service path to the corresponding tunnel. If the S flag is set, the client node MUST set up the tunnel according to the tunnel identifier and the tunnel specific attribute firstly. Then it maps the service path to the corresponding tunnel.

The Tunnel Type identifies the type of the tunneling technology used for the unicast service path. The type determines the syntax and semantics of the Tunnel Identifier field. This document defines the following Tunnel Types:

- + 0 - No tunnel information present
- + 1 - RSVP-TE LSP
- + 2 - LDP LSP
- + 3 - GRE Tunnel
- + 4 - MPLS-based Segment Routing Best-effort Path
- + 5 - MPLS-based Segment Routing Traffic Engineering Path

Tunnel Specific Attributes contains the attributes of the tunnel. The field is optional. The value depends on the tunnel type. It will be defined in the future versions.

When the Tunnel Type is set to "No tunnel information present", the Tunnel attribute carries no tunnel information (no Tunnel Identifier). when the type is used, the tunnel used for the service path is determined by the ingress router.

When the Tunnel Type is set to RSVP - Traffic Engineering (RSVP-TE) Label Switched Path (LSP), the Tunnel Identifier is <C-Type, Tunnel

Sender Address, Tunnel ID, Tunnel End-point Address> as specified in [RFC3209]. If C-Type = 7, Tunnel Sender Address and Tunnel End-point Address are IPv4 address in 4 octets. If C-Type = 8, Tunnel Sender Address and Tunnel End-point Address are IPv6 address in 16 octets. The other fields in the RSVP-TE LSP Identifier are the same as specified in [RFC3209].

When the Tunnel Type is set to LDP LSP, the Tunnel Identifier is <Ingress Router's IP Address, Address Family, Prefix Length, Prefix> as specified in [RFC5036].

When the Tunnel Type is set to GRE Tunnel, the Tunnel Identifier is <Ingress Router's IP Address, Address Family, Source IP Address, Destination IP Address>.

When the Tunnel Type is set to MPLS-based Segment Routing Best-effort Path, the Tunnel Identifier is <Ingress Router's IP Address, Address Family, Destination Address>. When the ingress router receives a BGP route with MPLS-based Segment Routing Path Tunnel Identifier in the Extended Unicast Tunnel attribute, it will find the best-effort SR-path based on the destination address.

When the Tunnel Type is set to MPLS-based Segment Routing Traffic Engineering Path, the Tunnel Identifier is <C-Type, Tunnel Sender Address, Tunnel ID, Tunnel End-point Address>. If C-Type = 7, Tunnel Sender Address and Tunnel End-point Address are IPv4 address in 4 octets. If C-Type = 8, Tunnel Sender Address and Tunnel End-point Address are IPv6 address in 16 octets. The tunnel identifier is similar as that of RSVP-TE LSP.

## 5.2. Extended PMSI Tunnel Attribute

This document defines and uses a new BGP attribute called the "Extended PMSI Tunnel attribute". This is an optional transitive BGP attribute. The format of this attribute is defined as follows:

```

+-----+
| Flags (1 octet) |
+-----+
| Tunnel Type (1 octets) |
+-----+
| Tunnel Identifier (variable) |
+-----+
| Tunnel Specific Attributes (Variable)(Optional) |
+-----+

```

This document defines the following flags:

```

0 1 2 3 4 5 6 7
+-----+
| reserved |S|
+-----+

```

+ PMSI Tunnel Setup Required (S)

If the S flag is not set, the client node is just to map the service path to the corresponding tunnel. If the S flag is set, the client node MUST set up the tunnel according to the tunnel identifier and the tunnel specific attribute firstly. Then it maps the service path to the corresponding tunnel.

The Tunnel Type identifies the type of the tunneling technology used for the multicast service path. The type determines the syntax and semantics of the Tunnel Identifier field. This document defines the following Tunnel Types:

```

+ 0 - No tunnel information present
+ 1 - RSVP-TE P2MP LSP
+ 2 - mLDP P2MP LSP
+ 3 - PIM-SSM Tree
+ 4 - PIM-SM Tree
+ 5 - BIDIR-PIM Tree
+ 6 - Ingress Replication
+ 7 - mLDP MP2MP LSP

```

Tunnel Identifier: The definition of Tunnel Identifier is the same as those specified in section 5 of [RFC6514].

Tunnel Specific Attributes contains the attributes of the PMSI tunnel. The field is optional. The value depends on the PMSI tunnel type. It will be defined in the future versions.

## 6. Capability Negotiation

It is necessary to negotiate the capability to support MPLS path programming. The MPLS-Path-Programming Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is to be specified by the IANA. The Capability Length field of this capability is variable. The Capability Value field consists of one or more of the following tuples:

```
+-----+
| Address Family Identifier (2 octets) |
+-----+
| Subsequent Address Family Identifier (1 octet) |
+-----+
| Send/Receive (1 octet) |
+-----+
```

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].

Send/Receive: This field indicates whether the sender is (a) willing to receive programming MPLS paths from its peer (value 1), (b) would like to send programming MPLS paths to its peer (value 2), or (c) both (value 3) for the <AFI, SAFI>.

## 7. IANA Considerations

TBD.

## 8. Security Considerations

TBD.

## 9. References

### 9.1. Normative References

[I-D.li-spring-mpls-path-programming]  
Li, Z., "Use Cases and Framework of Service-Oriented MPLS Path Programming (MPP)", draft-li-spring-mpls-path-programming-00 (work in progress), July 2014.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

## 9.2. Informative References

- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

## Authors' Addresses

Zhenbin Li  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: lizhenbin@huawei.com



Shunwan Zhuang  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: zhuangshunwan@huawei.com

Interdomain Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 4, 2015

S. Litkowski  
Orange Business Service  
K. Patel  
Cisco Systems  
J. Haas  
Juniper Networks  
July 3, 2014

Timestamp support for BGP paths  
draft-litkowski-idr-bgp-timestamp-00

## Abstract

BGP is more and more used to transport routing information for critical services. Some BGP updates may be critical to be received as fast as possible : for example, in a layer 3 VPN scenario where a dual-attached site is loosing primary connection, the BGP withdraw message should be propagated as fast as possible to restore the service. The same criticity exists for other address-families like multicast VPNs where "join" messages should also be propagated very fast.

Experience of service providers shows that BGP path propagation time may vary depending on network conditions (especially load of BGP speaker on the path) and too long propagation time are affecting customer service.

It is important for service providers to keep track of BGP updates propagation time to monitor quality of service for the customers. It is also important to be able to identify BGP Speakers that are slowing down the propagation.

This document presents a solution to transport timestamps of a BGP path.

## Requirements Language

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

## 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 January 4, 2015.

#### 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. Problem statement . . . . .	3
2. Proposal . . . . .	4
3. BGP timestamp attribute . . . . .	4
4. Processing the BGP timestamp attribute . . . . .	6
4.1. Inspection list . . . . .	6
4.2. Originating a timestamped route in BGP . . . . .	6
4.3. Receiving a timestamped route in BGP . . . . .	6
4.4. Sending a timestamped route in BGP . . . . .	7
4.5. Inter-AS considerations . . . . .	7
4.5.1. Drop option . . . . .	8
4.5.2. Summary option . . . . .	9
4.5.3. Propagate option . . . . .	9
4.6. Handling malformed attribute . . . . .	10
5. Monitoring BGP Update propagation time . . . . .	10
5.1. An architecture to measure BGP Update propagation time . . . . .	10
5.2. Measurement accuracy . . . . .	11
5.3. Dealing with stale information . . . . .	12
6. Compared to BMP . . . . .	13
7. Deployment considerations . . . . .	14

8. Security considerations . . . . .	14
9. Acknowledgements . . . . .	15
10. IANA Considerations . . . . .	15
11. Normative References . . . . .	15
Authors' Addresses . . . . .	15

## 1. Problem statement

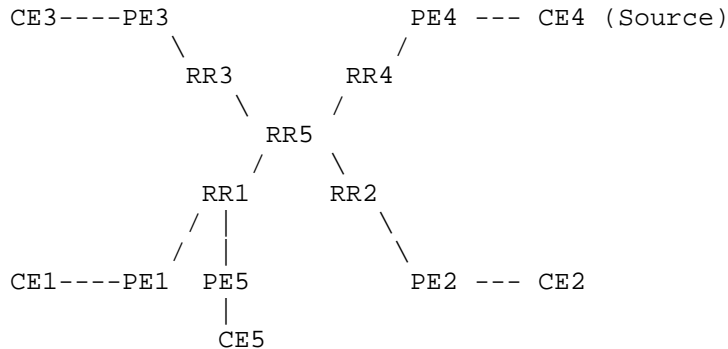


Figure 1

The figure 1 describes a typical hierarchical RR design where PEs are meshed to local RRs and local RRs are meshed to more centric RRs. We consider a single multicast VPN between all CEs. CE4 is the source, all others may be receivers. The BGP controlplane also supports some other BGP service like L3VPN service.

We consider an event in L3VPN service leading to RR1 being temporarily overloaded (for example, RR1 is processing massive updates due to a router failure or formatting updates for a route-refresh). In the same timeframe, CE1 wants to join the multicast flow from CE4. PE1 propagates the C-multicast route to RR1, but RR1 fails to propagate the route to RR5 because it is busy processing L3VPN. When RR1 finishes the L3VPN job, it would send the C-multicast route to RR5 and updates would be imported by PE4. The long time to join the flow may cause CE4 to miss part of the multicast flow.

All BGP implementations are different in term of internal processing within an address family or between address family. The issue described above is just given as an example, and the document does not presume that all implementations are suffering from this exact issue. But whatever the implementation, their always be cases where BGP update processing could be delayed.

Service providers currently lack of performant solution to keep track of BGP update propagation time as well as solution to identify the BGP speakers causing issues.

BMP (BGP Monitoring Protocol) may be a solution but as several drawbacks (see Section 6).

## 2. Proposal

Our proposal is based on the path vector property of BGP. Each hop within the path would add a tuple (ID,timestamp) information in the BGP path. An ordered list of timestamps would so be built along the path.

BGP Update	BGP Update	BGP Update	BGP Update
10.0.0.0/8	10.0.0.0/8	10.0.0.0/8	10.0.0.0/8
Timestamp:	Timestamp:	Timestamp:	Timestamp:
R1:T1	R1:T1	R1:T1	R1:T1
	R2:T2	R2:T2	R2:T2
		R3:T3	R3:T3
			R4:T4

R1 -----> R2 -----> R3 -----> R4 -----> R5

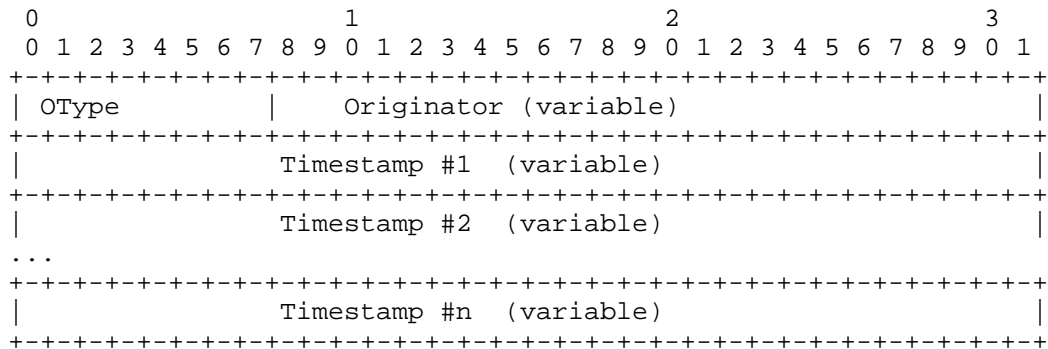
Using this mechanism, we can easily identify if a hop within a path is slowing down the propagation.

We propose to use a new BGP attribute, BGP timestamp attribute to encode timestamps information.

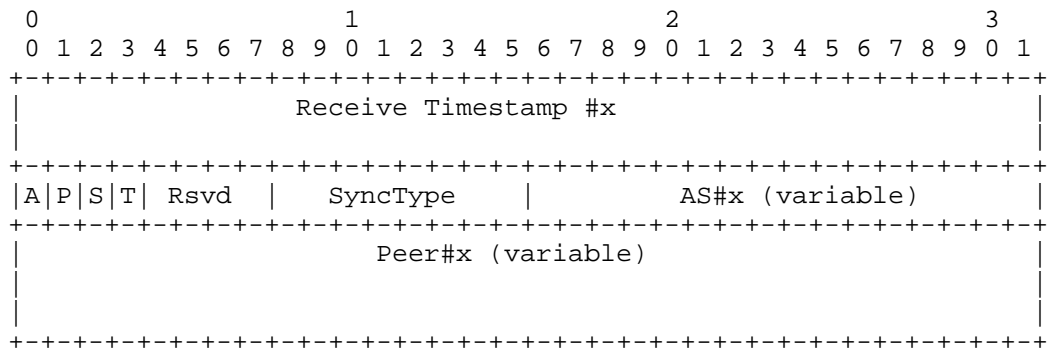
## 3. BGP timestamp attribute

The BGP timestamp (BGP-TS) Attribute is an optional transitive BGP Path Attribute. The attribute type code is TBD.

The value field of the BGP timestamp attribute is defined here :



- o OType : A single octet encoding the originator type
  - \* Type 1 : 2 bytes ASN.
  - \* Type 2 : 4 bytes ASN.
  - \* Type 3 : IPv4 address.
  - \* Type 4 : IPv6 address.
- o Originator : IP address or AS number identifying the first router or AS that added the BGP timestamp attribute.
- o Timestamp : ordered list of timestamps, the first timestamp is the first router information. The timestamps are encoded as follows :



- \* Receive timestamp : the time at which the BGP path was received. When originating a path in BGP, the timestamp is the originating time. The format of the timestamp is the same as in [RFC5905] and is as follows: the first 32 bits represent the unsigned integer number of seconds elapsed since 0h on 1

January 1900; the next 32 bits represent the fractional part of a second that has elapsed since then.

- \* Flags :
  - + A : AS type, if unset the AS field is a 2 bytes ASN, otherwise the AS field is a 4 bytes ASN.
  - + P : Peer type, if unset the peer field is an IPv4 address, otherwise the peer field is an IPv6 address.
  - + S : Summary, if set, the timestamp is a summary entry and does not contain a peer field. If set, the P bit MUST be set to zero.
  - + T : Synchronized, if set, the BGP speaker clock is synchronized to an external system.
- \* SyncType : defines the stratum as defined in [RFC5905].
- \* AS : the local AS of the BGP Speaker.
- \* Peer : the routerID of the BGP Speaker.

#### 4. Processing the BGP timestamp attribute

##### 4.1. Inspection list

A BGP Speaker supporting the BGP-TS can decide to timestamp only some specific BGP paths. An inspection list may be configured by the user (filter) to apply timestamping on a specific set of BGP prefixes or paths. By default, we suggest that a BGP Speaker supporting BGP-TS SHOULD NOT timestamp any BGP paths.

##### 4.2. Originating a timestamped route in BGP

When a BGP Speaker supporting BGP-TS originates a new path in BGP that matches the inspection list, it MUST add the BGP-TS attribute to the BGP path and MUST set the receive timestamp field to the time the path was originated in BGP. If the BGP Speaker is synchronized to an external system when originating the route, the S-bit MUST be set in the attribute and the SyncType MUST be set to the current stratum.

##### 4.3. Receiving a timestamped route in BGP

When a BGP Speaker supporting BGP-TS receives a BGP path that matches the inspection list and does not contain a BGP-TS attribute, it MUST add a BGP-TS attribute containing :

- o The originator type and originator field are set according to local BGP Speaker informations.
- o The timestamp entry contains information related to the local BGP Speaker.
- o If the BGP Speaker is synchronized to an external system when receiving the route, the S-bit MUST be set in the attribute and the SyncType MUST be set to the current stratum.

When a BGP Speaker supporting BGP-TS receives a BGP path that matches the inspection list and contains a BGP-TS attribute, it MUST append its own timestamp entry in the existing attribute. If the BGP Speaker is synchronized to an external system when receiving the route, the S-bit MUST be set in the attribute and the SyncType MUST be set to the current stratum.

When a BGP Speaker supporting BGP-TS receives a BGP path that does not the inspection list and contains a BGP-TS attribute, it MUST NOT change the existing attribute.

When a BGP Speaker not supporting BGP-TS receives a BGP path that contains a BGP-TS attribute, it MUST follow the standard BGP procedures described in [RFC4271].

#### 4.4. Sending a timestamped route in BGP

For a manageability/security purpose, the authors suggest that BGP timestamp attribute MAY NOT be sent to a peer unless it was explicitly configured for. This would prevent timestamp and internal address informations to be propagated to some external peers for example. See Section 4.5 for more information.

If a BGP path containing a BGP-TS attribute must be sent to be peer not configured with BGP timestamp option, the BGP-TS attribute should be dropped when the update message is sent to the peer.

#### 4.5. Inter-AS considerations



```

    BGP update
    CE2 add timestamp
    10.0.0.0/8
when receiving path
    TS:
    CE1:T1

```

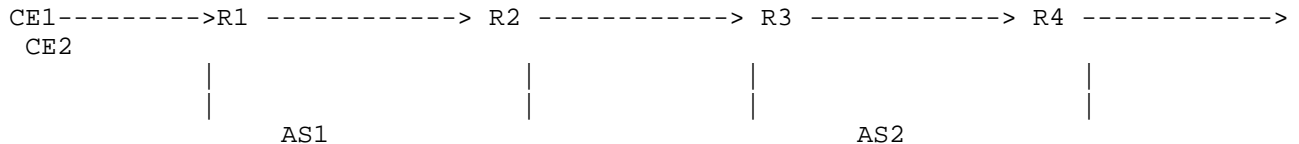


Figure 2

In the figure above, we consider that customer wants to monitor BGP updates propagation time between its two sites.

If AS1 and AS2 BGP Speakers does not support BGP-TS, the attribute will be transported transparently accross AS1 without any processing. CE2 will so receive the BGP path with only a single timestamp entry from CE1.

If AS1 and AS2 BGP Speakers does support BGP-TS, three different options are offered : drop, summarize, propagate.

#### 4.5.1. Drop option

If AS1 and/or AS2 BGP Speakers support BGP-TS, they may not want to expose their timestamps or internal BGP topology to other ASes. If a service does not want to propagate timestamp information to external peers, it can decide to not activate the "timestamp" option on the peer configuration , as explained in Section 4.4.

BGP update	BGP update	BGP update	BGP update	BGP update
10.0.0.0/8	10.0.0.0/8	10.0.0.0/8	10.0.0.0/8	10.0.0.0/8
TS:	TS:			
CE1:T1	CE1:T1			

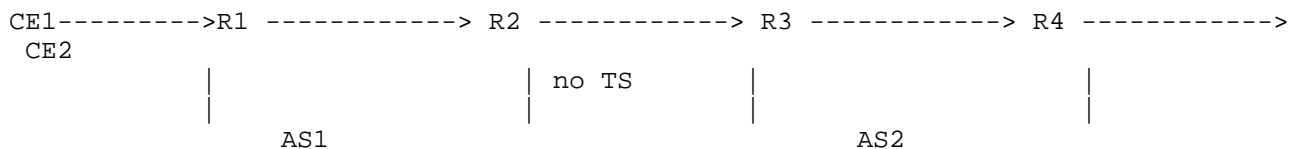


Figure 3

#### 4.5.2. Summary option

If AS1 and/or AS2 BGP Speakers support BGP-TS, they may want to offer timestamp service to their customers but they want to hide their internal topology. In order to achieve the expected behavior, AS1/AS2 can activate a timestamp summary option on the external peer.

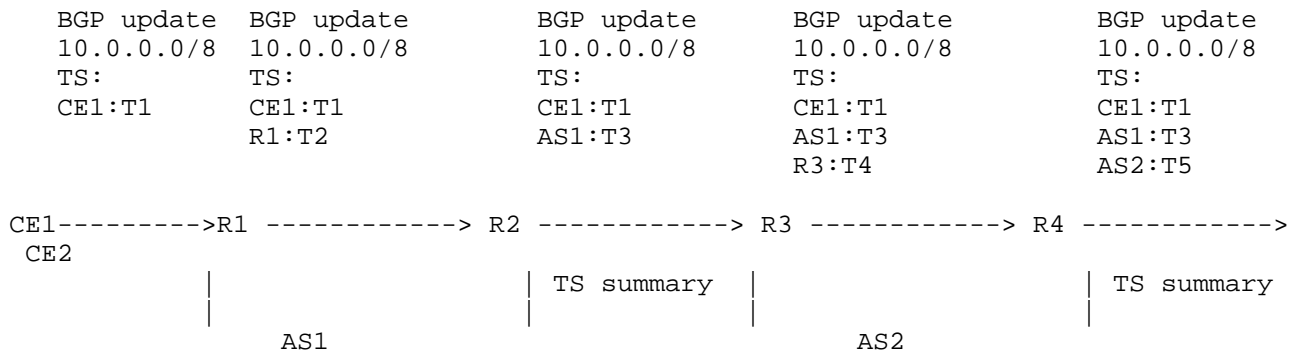


Figure 4

When using summary option, the BGP-TS attribute is modified as follows when exporting the route :

- o All timestamp entries containing the local AS in AS field are removed.
- o A new timestamp entry is created and inserted in place of removed entries (n entries replaced by 1).
- o The new timestamp entry MUST have the S bit set.
- o The new timestamp entry MUST NOT have a peer field.
- o The timestamp of the new timestamp entry is the receiving timestamp of the last timestamp entry that has been removed.

#### 4.5.3. Propagate option

If AS1 and/or AS2 BGP Speakers support BGP-TS, they may want to offer timestamp service to their customers with a full view. The behavior is the default intraAS behavior.

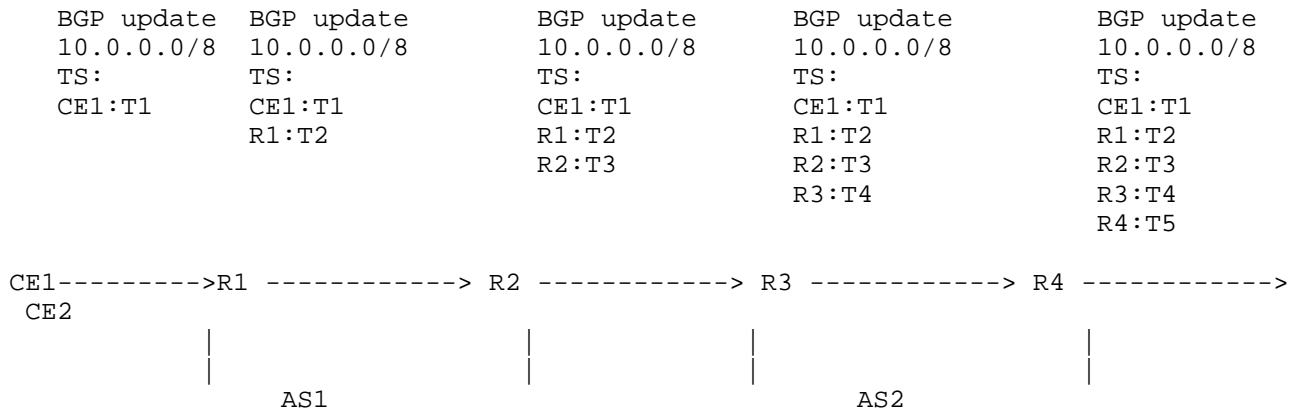


Figure 5

#### 4.6. Handling malformed attribute

When receiving a BGP Update message containing a malformed BGP-TS attribute, an "attribute-discard" action **MUST** be applied as defined in .

### 5. Monitoring BGP Update propagation time

#### 5.1. An architecture to measure BGP Update propagation time

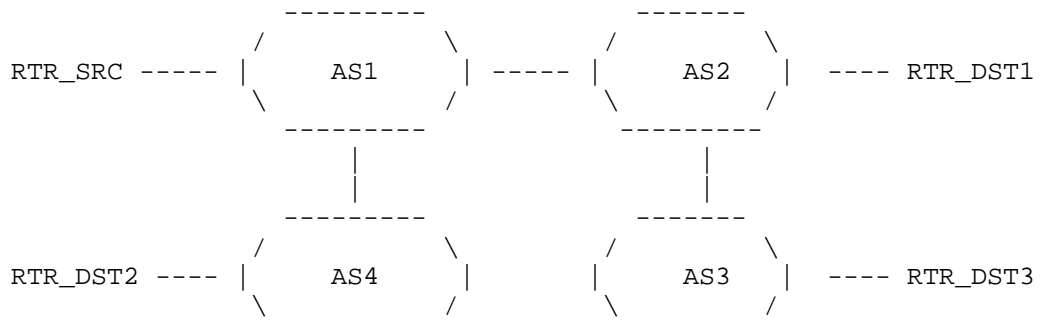


Figure 6

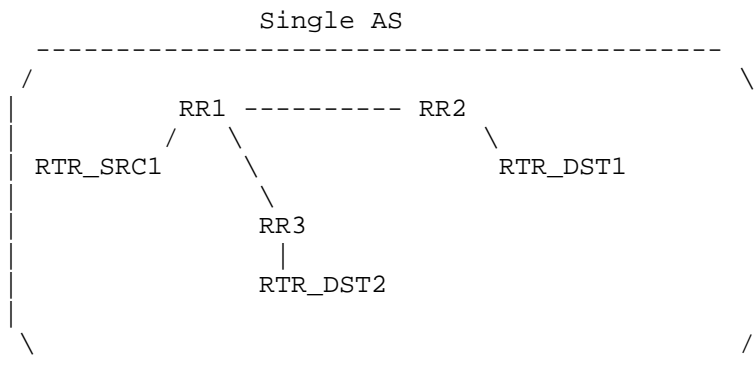


Figure 7

Figure 6 and Figure 7 describes an interAS and a single AS scenario where a service provider wants to monitor BGP Update propagation time from a router to multiple routers. In Figure 6, multiple probing routers are attached to multiple ASes. In Figure 7, all probing routers are in the same AS.

An external tool should command RTR\_SRC to originate a probing BGP path. Each probing router is configured to match the path in its inspection list. The BGP path would propagate across ASes whatever they are supporting BGP TS or not. Each probing router would receive the BGP path and add timestamp information. Authors suggest to implementors to use a local wrapping buffer on each node and record entries in the buffer each time a BGP path is timestamped. An external tool should then retrieve timestamps information from RTR\_DSTx. How the information is retrieved is out of scope of the document but we can imagine using :

- o BMP from the external tool to each RTR\_DSTx.
- o NetConf call to retrieve wrapping buffer information.
- o SNMP call to retrieve wrapping buffer information.
- o CLI command to retrieve wrapping buffer information.

## 5.2. Measurement accuracy

For the solution to be accurate, it is mandatory for BGP Speaker to be synchronized. This could be achieved easily within a single AS but in a inter domain scenario, it is hard to ensure that all Speakers are synchronized to a good clock source.

The S bit and SyncType fields are set to help operators to understand the accuracy of the timestamp measurements and being able to compare timestamps between them.

### 5.3. Dealing with stale information

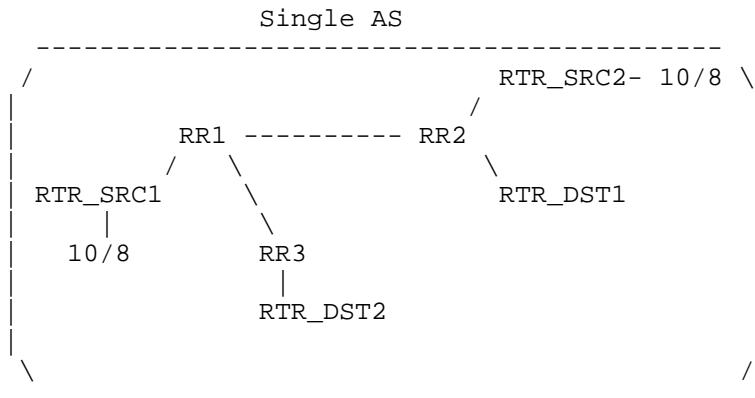


Figure 8

In the figure above, consider that the service provider is keep tracking of propagation time for real NLRIs (corresponding to customer routes). All the BGP Speakers in our figure are configured to inspect the NLRI 10/8 which is multihomed. We consider that the network is starting and the NLRI has not been propagated yet.

RTR\_SRC1 starts to propagate 10/8 within the BGP controlplane. All BGP Speakers considers the path as best and this path will be propagated within the whole controlplane. Each BGP Speaker would add its timestamp information and RTR\_DST1 and RTR\_DST2 would be able to record the timestamp vector. In this case, the timestamp vector is quite accurate because it represents an end to end propagation.

Now RTR\_SRC2 starts to propagate its own path. RR2 has two paths for 10/8 and will choose the best one, let's consider that RTR\_SRC2 path is the best one, RTR\_SRC2 path will so be propagated and timestamp vector will be updated. RR1 will also have two paths, and we consider that RR1 prefers RTR\_SRC1 path, so RTR\_SRC2 path will not be propagated by RR1. In this situation, RTR\_DST1 will receive the path from RR2 with accurate timestamp (end to end propagation) but RTR\_DST2 will never receive it.

We could also consider a stable network situation, where both paths have been advertised for a long time. A network event may occur (e.g. IGP metric change) that would cause a BGP Speaker within a path vector to change its best path. In Figure 8, an IGP event, may

cause RR1 to change its decision and prefers the path originated by RTR\_SRC2 as best, the path will be propagated with previous received timestamp information that are no more accurate. RTR\_DST2 will receive a BGP timestamp vector containing stale timestamp informations as well as new ones.

The case of sending stale timestamp information can also appear with a single originator as soon as some redundancy in the BGP design is involved (multiple RRs, multiple ASBRs ...).

An external tool that monitors BGP timestamp should take care about analysing only end to end propagation scenarios.

## 6. Compared to BMP

BMP (BGP Monitoring Protocol) [I-D.ietf-grow-bmp] is a solution to monitor BGP sessions and provides a convenient interface for obtaining route views. BMP is a complete suite of messages to exchange informations regarding a BGP session.

We can imagine to use BMP as a solution to monitor BGP update propagation time but there is multiple drawbacks associated with such solution :

- o BMP provides dump of all received BGP update (per peer). If we are interested only in probing BGP routes, a strong filtering of information may be needed in BMP messages.
- o BMP does not mandate timestamping of messages (as per [I-D.ietf-grow-bmp] Section 5) : "If the implementation is able to provide information about when routes were received, it MAY provide such information in the BMP timestamp field. Otherwise, the BMP timestamp field MUST be set to zero, indicating that time is not available."
- o BMP may provide (if implementation available) timestamps information only for a single router point of view. If we want to retrieve timestamps of all BGP Speakers on a path, a BMP session is required to all BGP speakers. Correlation (based on known design) is also required at the external tool to order timestamps from each BMP session.
- o If BMP provides timestamp information, it does not provide information on how the router clock is synchronized (free run, NTP, GPS ...).

Using BMP to monitor BGP update propagation may complexify the design of the monitor solution.

## 7. Deployment considerations

This solution is not intended to perform timestamp imposition on all BGP updates.

Service provider implementing the BGP timestamp attribute must be aware of the propagation rules of the NLRIs to be inspected. If we consider an implementation scenario, where a path for NLRI is already propagated, a new path may appear and starts to be propagated, propagation of this new path may stop at a certain point because a BGP Speaker may consider the old path as the best one. Another scenario, could be that the two paths are installed, and for a BGP Speaker within the path vector, the best path is changing because of an IGP metric change, this BGP Speaker will send a new BGP update and timestamp information of the path will be updated but will have no more sense : origin timestamp will be quite old, but timestamps recorded after this BGP Speaker will be recent. This kind of scenario is complex to understand.

The deployment scenario we are targeting is really to inspect some specific NLRIs identified by the service provider where the propagation rules are well known (see Section 5 as an example). Service provider may rely on existing NLRIs (real routes), or ephemeral NLRIs (dedicated NLRIs for beaconing). Whatever the NLRI used, the tool used by the service provider to collect and interpret the timestamp must be aware of the propagation rules and must record events only if propagation is end to end (from originator to listener).

The inspection list should be kept as small as possible in order to not introduce processing overhead and as a consequence slow down propagation. Implementors should take care about reducing as much as possible the processing overhead introduced by the inspection list and timestamp imposition.

## 8. Security considerations

Depending of the implementation and router capacity, adding timestamps to BGP path may consume some router resources. As proposed in Section 4.1, by default a BGP Speaker will not timestamp any path and inspection list should be configured to activate timestamping on a subset of paths. Using this approach, we consider that overhead that may be introduced by timestamping BGP paths is well controlled by operators. An external router cannot force an internal router to timestamp.

Providing detailed timestamps information to other ASes may introduce security issues by exposing internal datas (part of BGP

topology, IP addresses, internal performance) to external entities. The proposal we make in Section 4.5 solves this security issue by giving flexibility to operators on the level of information he wants to expose to external peers.

## 9. Acknowledgements

## 10. IANA Considerations

IANA shall assign a codepoint for the BGP Timestamp attribute. This codepoint will come from the "BGP Path Attributes" registry.

## 11. Normative References

- [I-D.ietf-grow-bmp]  
Scudder, J., Fernando, R., and S. Stuart, "BGP Monitoring Protocol", draft-ietf-grow-bmp-07 (work in progress), October 2012.
- [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.
- [RFC5905] Mills, D., Martin, J., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.

## Authors' Addresses

Stephane Litkowski  
Orange Business Service  
  
Email: stephane.litkowski@orange.com

Keyur Patel  
Cisco Systems  
  
Email: keyupate@cisco.com

Jeff Haas  
Juniper Networks  
  
Email: jhaas@juniper.net



Routing Area Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 1, 2015

S. Litkowski  
Orange  
A. Simpson  
Alcatel Lucent  
K. Patel  
Cisco  
J. Haas  
Juniper Networks  
June 30, 2014

Applying BGP flowspec rules on a specific interface set  
draft-litkowski-idr-flowspec-interfaceset-00

Abstract

BGP Flow-spec is an extension to BGP that allows for the dissemination of traffic flow specification rules. The primary application of this extension is DDoS mitigation where the flowspec rules are applied in most cases to all peering routers of the network.

This document will present another use case of BGP Flow-spec where flow specifications are used to maintain some access control lists at network boundary. BGP Flowspec is a very efficient distributing machinery that can help in saving OPEX while deploying/updating ACLs. This new application requires flow specification rules to be applied only on a specific subset of interfaces and in a specific direction.

The current specification of BGP Flow-spec does not detail where the flow specification rules need to be applied.

This document presents a new interface-set flowspec action that will be used in complement of other actions (marking, rate-limiting ...). The purpose of this extension is to inform remote routers on where to apply the flow specification.

This extension can also be used in a DDoS mitigation context where a provider wants to apply the filtering only on specific peers.

Requirements Language

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

## 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 January 1, 2015.

## 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. Use case . . . . .	3
1.1. Specific filtering for DDoS . . . . .	3
1.2. ACL maintenance . . . . .	4
2. Interface specific filtering using BGP flowspec . . . . .	5
3. Interface-set extended community . . . . .	5
4. Security Considerations . . . . .	6
5. Acknowledgements . . . . .	7
6. IANA Considerations . . . . .	7
7. Normative References . . . . .	7
Authors' Addresses . . . . .	7

## 1. Use case

### 1.1. Specific filtering for DDoS

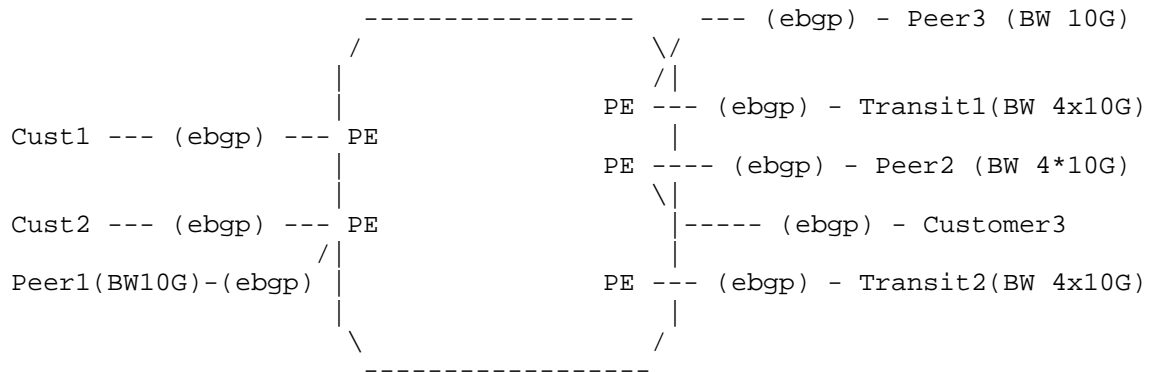


Figure 1

The figure 1 above displays a typical service provider Internet network owing Customers, Peers and Transit. To protect proactively against some attacks (e.g. DNS, NTP ...), the service provider may want to deploy some rate-limiting of some flows on peers and transit links. But depending on link bandwidth, the provider may want to apply different rate-limiting values.

For 4\*10G links peer/transit, it may want to apply a rate-limiting of DNS flows of 1G, while on 10G links, the rate-limiting would be set to 250Mbps. Customer interfaces must not be rate-limited.

BGP Flow-spec infrastructure may already be present on the network, and all PEs may have a BGP session running flowspec address family. The Flowspec infrastructure may be reused by the service provider to implement such rate-limiting in a very quick manner and being able to adjust values in future quickly without having to configure each node one by one. Using the current BGP flowspec specification, it would not be possible to implement different rate limiter on different interfaces of a same router. The flowspec rule is applied to all interfaces in all directions or on some interfaces where flowspec is activated but flowspec rule set would be the same among all interfaces.

Section Section 2 will detail a solution to address this use case using BGP Flowspec.

## 1.2. ACL maintenance

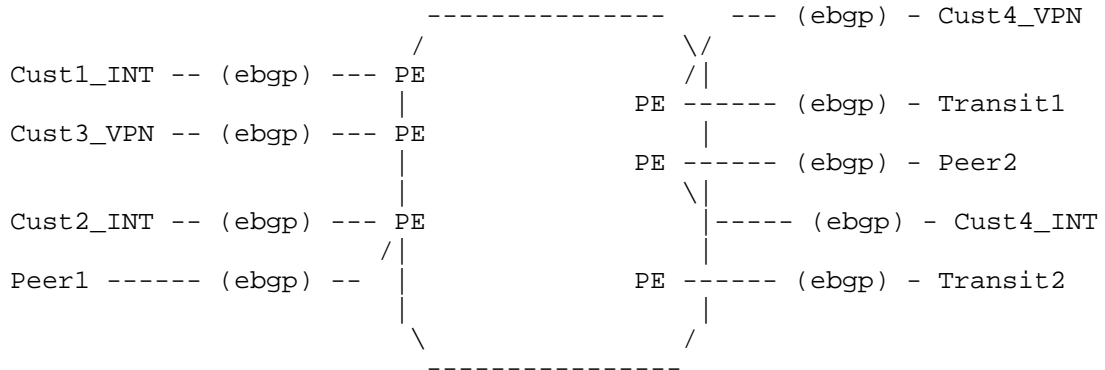


Figure 2

The figure 1 above displays a typical service provider multiservice network owing Customers, Peers and Transit for Internet, as well as VPN services. The service provider requires to ensure security of its infrastructure by applying ACLs at network boundary. Maintaining and deploying ACLs on hundreds/thousands of routers is really painful and time consuming and a service provider would be interested to deploy/updates ACLs using BGP Flowspec. In this scenario, depending on the interface type (Internet customer, VPN customer, Peer, Transit ...) the content of the ACL may be different.

We can imagine two cases :

- o Maintaining complete ACLs using flowspec : in this case all the ingress ACL are maintained and deployed using BGPFlowspec. See section Section 4 for more details on security aspects.
- o Requirement of a quick deployment of a new filtering term due to a security alert : new security alerts often requires a fast deployment of new ACL terms. Using traditional CLI and hop by hop provisionning, such deployment takes time and network is unprotected during this time window. Using BGP flowspec to deploy such rule, a service provider can protect its network in few seconds. Then the SP can decide to keep the rule permanently in BGP Flowspec or update its ACL or remove the entry (in case equipments are not vulnerable anymore).

Section Section 2 will detail a solution to address this use case using BGP Flowspec.

## 2. Interface specific filtering using BGP flowspec

The use case detailed above requires application of different BGP Flowspec rules on different set of interfaces. The basic specification detailed in [RFC5575] does not address this and does not give any detail on where the FlowSpec filter need to be applied.

We propose to introduce an identification of interfaces within BGP Flowspec. All interfaces may be associated to one or more group-identifiers and a BGP Flowspec rule may also be associated with one or more group-identifiers including a filtering direction (input/output/both) , so the FlowSpec rule will be applied only on interfaces belonging the the group identifier included in the BGP FlowSpec update.

Considering figure 2, we can imagine the following design :

- o Internet customer interfaces are associated with group-identifier 1.
- o VPN customer interfaces are associated with group-identifier 2.
- o All customer interfaces are associated with group-identifier 3.
- o Peer interfaces are associated with group-identifier 4.
- o Transit interfaces are associated with group-identifier 5.
- o All external provider interfaces are associated with group-identifier 6.
- o All interfaces are associated with group-identifier 7.

If the service provider wants to deploy a specific inbound filtering on external provider interfaces only, the provider can send the BGP flow specification using group-identifier 6 and including inbound direction.

## 3. Interface-set extended community

This document proposes a new BGP extended community called "flow spec interface-set". This new BGP extended community is part of TRANSITIVE FOUR-OCTET AS-SPECIFIC EXTENDED COMMUNITY and has subtype TBD.

The Global Administrator field of this community MUST be set to the ASN of the originating router. The Local Administrator field is encoded as follows :

```

      0   1   2   3   4   5   6   7
+---+---+---+---+---+---+---+---+
| 0 | I |   Group Identifier   :
+---+---+---+---+---+---+---+---+
: Group Identifier (cont.)    |
+---+---+---+---+---+---+---+---+

```

The flags are :

- o 0 : if set, the flow specification rule MUST be applied in outbound direction to the interface set referenced by the following group-identifier.
- o I : if set, the flow specification rule MUST be applied in input direction to the interface set referenced by the following group-identifier.

Both flags can be set at the same time in the interface-set extended community leading to flow rule to be applied in both directions. An interface-set extended community with both flags set to zero MUST be treated as an error and as consequence, the FlowSpec update MUST be discarded.

The Group Identifier is coded as a 14-bit number (values goes from 0 to 16383).

Multiple instances of the interface-set community may be present in a BGP update. This may appear if the flow rule need to be applied to multiple set of interfaces.

Multiple instances of the community in a BGP update MUST be interpreted as a "OR" operation : if a BGP update contains two interface-set communities with group ID 1 and group ID 2, the filter would need to be installed on interfaces belonging to Group ID 1 or Group ID 2.

#### 4. Security Considerations

Managing permanent Access Control List by using BGP Flowspec as described in Section 1.2 helps in saving roll out time of such ACL. However some ACL especially at network boundary are critical for the network security and loosing the ACL configuration may lead to network open for attackers.

By design, BGP flowspec rules are ephemeral : the flow rule exists in the router while the BGP session is UP and the BGP path for the rule is valid. We can imagine a scenario where a Service Provider is

managing the network boundary ACLs by using only FlowSpec. In this scenario, if , for example, an attacker succeed to make the internal BGP session of a router to be down , it can open all boundary ACLs on the node, as flowspec rules will disappear due to the BGP session down.

In reality, the chance for such attack to occur is low, as boundary ACLs should protect the BGP session from being attacked.

In order to complement the BGP flowspec solution is such deployment scenario and provides security against such attack, a service provider may activate Long lived Graceful Restart [I-D.uttaro-idr-bgp-persistence] on the BGP session owning Flowspec address family. So in case of BGP session to be down, the BGP paths of Flowspec rules would be retained and the flowspec action will be retained.

## 5. Acknowledgements

Authors would like to thanks Wim Hendrickx for his valuable comments.

## 6. IANA Considerations

This document requests a new sub-type from the "TRANSITIVE FOUR-OCTET AS-SPECIFIC EXTENDED COMMUNITY SUB-TYPES" extended community registry. The sub-type name shall be 'Flow spec interface-set'.

## 7. Normative References

- [I-D.uttaro-idr-bgp-persistence]  
Uttaro, J., Chen, E., Decraene, B., and J. Scudder,  
"Support for Long-lived BGP Graceful Restart", draft-uttaro-idr-bgp-persistence-03 (work in progress), November 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, August 2009.

## Authors' Addresses

Stephane Litkowski  
Orange

Email: stephane.litkowski@orange.com

Adam Simpson  
Alcatel Lucent

Email: adam.simpson@alcatel-lucent.com

Keyur Patel  
Cisco

Email: keyupate@cisco.com

Jeff Haas  
Juniper Networks

Email: jhaas@juniper.net



Interdomain Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: December 20, 2014

S. Litkowski  
Orange Business Service  
J. Haas  
Juniper Networks  
June 18, 2014

Inter Domain considerations for Constrained Route distribution  
draft-litkowski-idr-rtc-interas-00

Abstract

[RFC4684] defines Multi-Protocol BGP (MP-BGP) procedures that allow BGP speakers to exchange Route Target reachability information in order to limit the propagation of Virtual Private Networks (VPN) Network Layer Reachability Information (NLRI).

[RFC4684] addresses both intra domain and inter domain distributions. Based on operational deployments, the current distribution model defined in [RFC4684] may cause some issue in specific scenarios.

This document refines the route distribution rules for inter domain NLRIs in order to address these specific scenarios.

Requirements Language

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

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 December 20, 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. Problem statement . . . . .	2
2. Proposal . . . . .	4
3. Security considerations . . . . .	5
4. Acknowledgements . . . . .	5
5. IANA Considerations . . . . .	5
6. Normative References . . . . .	5
Authors' Addresses . . . . .	6

## 1. Problem statement

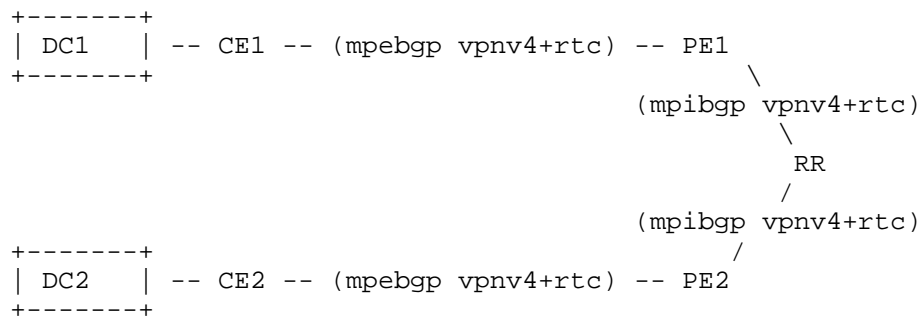


Figure 1

The figure above describes a typical service provider scenario where datacenters are connected through MPLS VPN interas option B with the Service Provider network. Route Target Constraint (RTC) is deployed on MPeBGP sessions as well as internally in the service provider network to ensure optimal distribution of VPN routes (required for scaling reason). In this scenario, both Datacenters are using the same AS number, generally a private ASN (65000) like a typical PE-CE connection. As we expect DCs to communicate between each other, some

features like "as-override" are deployed on PEs to overcome ASPATH loop issue.

[RFC4684] Section 3.1 and 3.2 describes propagation of Route Target NLRI between ASes and inside an AS and distinguish two types of NLRIs :

- o Locally originated NLRI where origin-as field of the NLRI is equal to the local AS number.
- o External NLRI where origin-as field of the NLRI is different from the local AS number.

Regarding External NLRI, the idea of Section 3.1 and 3.2 is to establish the route distribution tree over the shortest path considering that BGP routing is internally consistent for a given AS.

Extract from [RFC4684] Section 3.2 :

"As indicated above, the inter-AS VPN route distribution graph, for a given route-target, is constructed by creating a directed arc on the inverse direction of received Route Target membership UPDATES containing an NLRI of the form {origin-as#, route-target}.

Inside the BGP topology of a given autonomous-system, as far as external RT membership information is concerned (route-targets where the as# is not the local as), it is easy to see that standard BGP route selection and advertisement rules [4] will allow a transit AS to create the necessary flooding state."

In the Figure 1, CE1 and CE2 are advertising the RT 1:1 respectively to PE1 and PE2, the generated NLRI would be 65000:2:1:1/96. According to procedures defined in [RFC4684] Section 3.2, both PEs are using the standard BGP route selection and advertising rules. So both PEs are advertising their path for 65000:2:1:1/96 to the route-reflector. The route-reflector would also use the standard BGP route selection to create the RT flooding state. Considering that path from PE1 is the best one, a flooding tree branch for RT 1:1 is created only towards PE1.

Due to this behavior, VPN routes from DC1 would never to send to DC2 because PE2 is not part of the flooding tree and as DC1 and DC2 are disjoint, even if they are using the same ASN, there is no communication possible between them.

The same issue may appear if two MPeBGP sites using the same ASN are connected on the same PE like in figure 2.

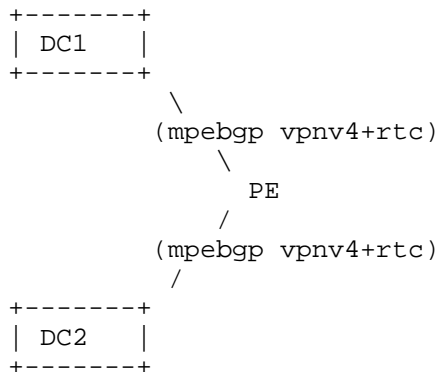


Figure 2

## 2. Proposal

This document proposes to modify the following procedures defined in [RFC4684] :

### 1. [RFC4684] Section 3.1 :

"Using RT membership information that includes both route-target and originator AS number allows BGP speakers to use standard path selection rules concerning as-path length (and other policy mechanisms) to prune duplicate paths in the RT membership information flooding graph, while maintaining the information required to reach all autonomous systems advertising the Route Target."

### 2. [RFC4684] Section 3.2 :

"As indicated above, the inter-AS VPN route distribution graph, for a given route-target, is constructed by creating a directed arc on the inverse direction of received Route Target membership UPDATES containing an NLRI of the form {origin-as#, route-target}.

Inside the BGP topology of a given autonomous-system, as far as external RT membership information is concerned (route-targets where the as# is not the local as), it is easy to see that standard BGP route selection and advertisement rules [4] will allow a transit AS to create the necessary flooding state."

In order to support our scenario, path pruning may be disabled by configuration for a given origin AS (different from the local AS). Implementations may also permit path pruning to be disabled for private AS numbers by default, but must make provision for it to be selectively enabled if such a feature is present.

This modification in establishing route distribution tree may create unnecessary flooding states in the situations where a real AS is multihomed to a service provider network (as displayed in Figure 3).

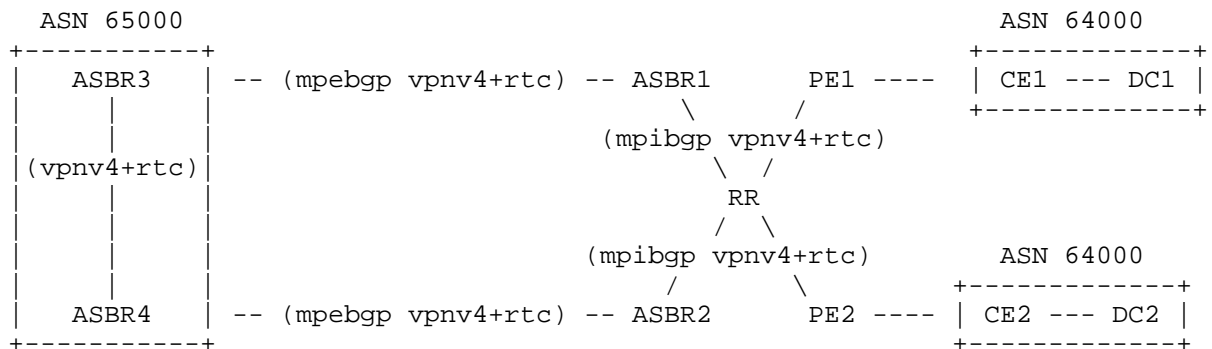


Figure 3

In the figure above, disabling pruning is required for AS64000 but it may be interesting to keep it enabled for AS65000. Implementations may require support for such granularity as proposed previously.

### 3. Security considerations

This document does not introduce any new security issue compared to [RFC4684].

### 4. Acknowledgements

### 5. IANA Considerations

There is no IANA consideration.

### 6. Normative References

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

Authors' Addresses

Stephane Litkowski  
Orange Business Service

Email: [stephane.litkowski@orange.com](mailto:stephane.litkowski@orange.com)

Jeff Haas  
Juniper Networks

Email: [jhaas@juniper.net](mailto:jhaas@juniper.net)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: November 27, 2014

S. Previdi, Ed.  
C. Filsfils  
S. Ray  
K. Patel  
Cisco Systems, Inc.  
May 26, 2014

Segment Routing Egress Peer Engineering BGPLS Extensions  
draft-previdi-idr-bgpls-segment-routing-epe-00

## Abstract

Segment Routing (SR) leverages source routing. A node steers a packet through a controlled set of instructions, called segments, by prepending the packet with an SR header. A segment can represent any instruction, topological or service-based. SR allows to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node of the SR domain.

The Segment Routing architecture can be directly applied to the MPLS dataplane with no change on the forwarding plane. It requires minor extension to the existing link-state routing protocols.

This document outline a BGPLS extension for exporting BGP egress point topology information (including its peers, interfaces and peering ASs) in a way that is exploitable in order to compute efficient Egress Point Engineering policies and strategies.

## 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 November 27, 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. Segment Routing Documents . . . . .	3
3. BGP Peering Segments . . . . .	3
4. Peering Segment NLRI-Type . . . . .	4
4.1. Peer Descriptors . . . . .	4
4.2. Peer Attributes . . . . .	5
5. Definition of PeerNode and PeerAdj . . . . .	5
5.1. PeerNode Segment (PeerNodeSID) . . . . .	5
5.2. PeerAdj Segment (PeerAdjSID) . . . . .	6
5.3. PeerSet Segment (PeerSetSID) . . . . .	7
6. Illustration . . . . .	7
6.1. Reference Diagram . . . . .	7
6.1.1. PeerNode for Node D . . . . .	9
6.1.2. PeerNode for Node H . . . . .	9
6.1.3. PeerNode for Node E . . . . .	9
6.1.4. PeerAdj for Node E, Link 1 . . . . .	10
6.1.5. PeerAdj for Node E, Link 2 . . . . .	10
7. IANA Considerations . . . . .	11
8. Manageability Considerations . . . . .	11
9. Security Considerations . . . . .	11
10. Acknowledgements . . . . .	11
11. References . . . . .	11
11.1. Normative References . . . . .	11
11.2. Informative References . . . . .	11
Authors' Addresses . . . . .	12



## 1. Introduction

Segment Routing (SR) leverages source routing. A node steers a packet through a controlled set of instructions, called segments, by prepending the packet with an SR header. A segment can represent any instruction, topological or service-based. SR allows to enforce a flow through any topological path and service chain while maintaining per-flow state only at the ingress node of the SR domain.

The Segment Routing architecture can be directly applied to the MPLS dataplane with no change on the forwarding plane. It requires minor extension to the existing link-state routing protocols.

This document outline a BGPLS extension for exporting BGP egress point topology information (including its peers, interfaces and peering ASs) in a way that is exploitable in order to compute efficient Egress Point Engineering policies and strategies.

## 2. Segment Routing Documents

The main reference for this document is the SR architecture defined in [I-D.filsfils-spring-segment-routing].

The Segment Routing Egress Peer Engineering architecture is described in [I-D.filsfils-spring-segment-routing-central-epe].

## 3. BGP Peering Segments

As defined in [draft-filsfils-spring-segment-routing-epe], an EPE enabled Egress PE node MAY advertise segments corresponding to its attached peers. These segments are called BGP peering segments or BGP Peering SIDs. They enable the expression of source-routed inter-domain paths.

An ingress border router of an AS may compose a list of segments to steer a flow along a selected path within the AS, towards a selected egress border router C of the AS and through a specific peer. At minimum, a BGP Peering Engineering policy applied at an ingress PE involves two segments: the Node SID of the chosen egress PE and then the BGP Peering Segment for the chosen egress PE peer or peering interface.

Hereafter, we will define three types of BGP peering segments/SID's: PeerNodeSID, PeerAdjSID and PeerGroupSID.

#### 4. Peering Segment NLRI-Type

This section described a new NLRI-Type in the BGP-LS specification ([I-D.ietf-idr-ls-distribution]). The new NLRI-Type (5) is called the Peer NLRI-Type and describes the connectivity of a BGP Egress router.

The format of the Peer NLRI Type is as follows:

```

      0                               1                               2                               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Protocol-ID |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Identifier                                     |
|                                     (64 bits)                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
//                               Local Node Descriptors (variable)                               //
+-----+-----+-----+-----+-----+-----+-----+-----+
//                               Peer Descriptors (variable)                               //
+-----+-----+-----+-----+-----+-----+-----+-----+
//                               Link Descriptors (variable)                               //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Where:

Local Node Descriptors: as defined in  
[I-D.ietf-idr-ls-distribution] Section 3.2.1.2.

Link Descriptors: as defined in [I-D.ietf-idr-ls-distribution]  
Section 3.2.2.

##### 4.1. Peer Descriptors

The following Sub-TLVs are allowed to be used as Peer Descriptors:

Sub-TLV Code Point	Description	Length
512	Peer Autonomous System	4
513	BGP-LS Identifier	4

Peer Descriptors Sub-TLVs are defined in  
[I-D.ietf-idr-ls-distribution].

#### 4.2. Peer Attributes

The Peer Attributes Sub-TLVs codepoints and lengths are listed in the following table:

TLV Code Point	Description	Length	IS-IS SR TLV /sub-TLV
1099	Adjacency Segment Identifier (Adj-SID)	variable	31 (section 2.3.1)
1100	LAN Adjacency Segment Identifier (Adj-LAN SID)	variable	32 (section 2.3.2)
TBA	Peer Set SID	variable	31 (section 2.3.1)

Sections refer to [I-D.ietf-idr-ls-distribution].

The value of the Adj-SID, Adj-LAN-SID and Peer Set SID Sub-TLV SHOULD be persistent across router restart.

#### 5. Definition of PeerNode and PeerAdj

In this section the following Peer Segments are defined:

PeerNode Segment (PeerNodeSID)

PeerAdj Segment (PeerAdjSID)

PeerSet Segment (PeerSetSID)

##### 5.1. PeerNode Segment (PeerNodeSID)

A BGP PeerNode segment/SID is a local segment. At the BGP node advertising it, its semantics is:

- o SR header operation: NEXT (as defined in [I-D.filsfils-spring-segment-routing]).
- o Next-Hop: the connected peering node to which the segment is related.

The PeerNode is advertised with a Peering Segment NLRI, where:

- o Local Node Descriptor is the IGP node describing the EPE enabled egress PE.

- o Peer Descriptor is the ASN of the peer.
- o Link Descriptors, as defined in [I-D.ietf-idr-ls-distribution] contain the addresses used by the BGP session:
  - \* IPv4 Interface Address (Sub-TLV 259) contains the BGP session IPv4 local address.
  - \* IPv4 Neighbor Address (Sub-TLV 260) contains the BGP session IPv4 peer address.
  - \* IPv6 Interface Address (Sub-TLV 261) contains the BGP session IPv6 local address.
  - \* IPv6 Neighbor Address (Sub-TLV 262) contains the BGP session IPv6 peer address.
- o Peer Attribute contains the Adj-SID TLV

## 5.2. PeerAdj Segment (PeerAdjSID)

A BGP PeerAdj segment/SID is a local segment. At the BGP node advertising it, its semantics is:

- o SR header operation: NEXT (as defined in [I-D.filsfils-spring-segment-routing]).
- o Next-Hop: the peer connected through the interface to which the segment is related.

The PeerAdj is advertised with a Peering Segment NLRI, where:

- o Local Node Descriptor is the IGP node describing the EPE enabled egress PE.
- o Peer Descriptor is the ip address and ASN of the peer.
- o Link Descriptors, as defined in [I-D.ietf-idr-ls-distribution] contain the addresses used by the BGP session:
  - \* IPv4 Interface Address (Sub-TLV 259) contains the BGP session IPv4 local address.
  - \* IPv4 Neighbor Address (Sub-TLV 260) contains the BGP session IPv4 peer address.

- \* IPv6 Interface Address (Sub-TLV 261) contains the BGP session IPv6 local address.
- \* IPv6 Neighbor Address (Sub-TLV 262) contains the BGP session IPv6 peer address.
- o Peer Attribute contains the Adj-SID TLV

In addition, BGPLS Link Attributes, as defined in [I-D.ietf-idr-ls-distribution] MAY be inserted in order to advertise the characteristics of the link.

### 5.3. PeerSet Segment (PeerSetSID)

A PeerSet segment/SID is a local segment. At the BGP node advertising it, its semantics is:

- o SR header operation: NEXT (as defined in [I-D.filsfils-spring-segment-routing]).
- o Next-Hop: loadbalance across any connected interface to any peer in the related set.

The PeerSet is advertised in a Peering Segment NLRI (PeerNode or PeerAdj) as a BGPLS attribute.

The PeerSet Attribute contains an Adj-SID TLV, defined in Section 4.2 identifying the Set the PeerNode or PeerAdj is part of.

## 6. Illustration

### 6.1. Reference Diagram

The following reference diagram is used throughout this document. The solution is described for IPv4 with MPLS-based segments.

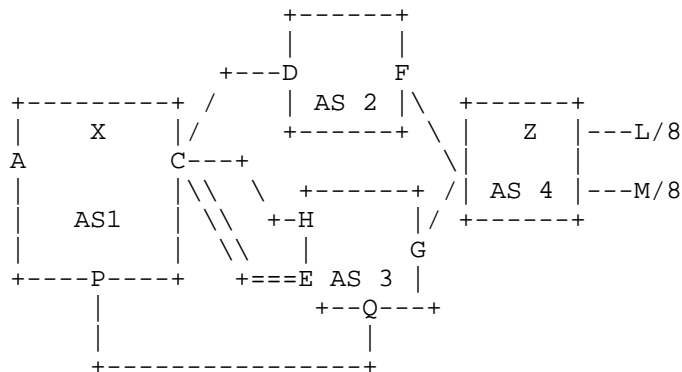


Figure 1: Reference Diagram

IPv4 addressing:

- o C's interface to D: 1.0.1.1/24, D's interface: 1.0.1.2/24
- o C's interface to H: 1.0.2.1/24, H's interface: 1.0.2.2/24
- o C's upper interface to E: 1.0.3.1/24, E's interface: 1.0.3.2/24
- o C's lower interface to E: 1.0.4.1/24, E's interface: 1.0.4.2/24
- o Loopback of E used for eBGP multi-hop peering to C: 1.0.5.2/32
- o C's loopback is 3.3.3.3/32 with SID 64

C's BGP peering:

- o Single-hop eBGP peering with neighbor 1.0.1.2 (D)
- o Single-hop eBGP peering with neighbor 1.0.2.2 (H)
- o Multi-hop eBGP peering with E on ip address 1.0.5.2 (E)

C's resolution of the multi-hop eBGP session to E:

- o Static route 1.0.5.2/32 via 1.0.3.2
- o Static route 1.0.5.2/32 via 1.0.4.2

Node C configuration is such that:

- o A PeerNode segment is allocated to each peer (D, H and E).

- o A PeerAdj segment is defined for each recursing interface to a multi-hop peer (CE upper and lower interfaces).
- o A PeerSet is defined to include all peers in AS3 (peers H and E).

#### 6.1.1.1. PeerNode for Node D

##### Descriptors:

- o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1
- o Peer Descriptors (peer ASN): AS2
- o Link Descriptors (IPv4 interface address, neighbor IPv4 address): 1.0.1.1, 1.0.1.2

##### Attributes:

- o Adj-SID: 1012

#### 6.1.1.2. PeerNode for Node H

##### Descriptors:

- o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1
- o Peer Descriptors (peer ASN): AS3
- o Link Descriptors (IPv4 interface address, neighbor IPv4 address): 1.0.2.1, 1.0.2.2

##### Attributes:

- o Adj-SID: 1022
- o PeerSetSID: 1060
- o Link Attributes: see section 3.3.2 of [I-D.ietf-idr-ls-distribution]

#### 6.1.1.3. PeerNode for Node E

##### Descriptors:

- o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1
- o Peer Descriptors (peer ASN): AS3

- o Link Descriptors (IPv4 interface address, neighbor IPv4 address):  
3.3.3.3, 1.0.5.2

Attributes:

- o Adj-SID: 1052
- o PeerSetSID: 1060

#### 6.1.4. PeerAdj for Node E, Link 1

Descriptors:

- o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1
- o Peer Descriptors (peer ASN): AS3
- o Link Descriptors (IPv4 interface address, neighbor IPv4 address):  
1.0.3.1 , 1.0.3.2

Attributes:

- o Adj-SID: 1032
- o LinkAttributes: see section 3.3.2 of  
[I-D.ietf-idr-ls-distribution]

#### 6.1.5. PeerAdj for Node E, Link 2

Descriptors:

- o Node Descriptors (router-ID, ASN): 3.3.3.3 , AS1
- o Peer Descriptors (peer ASN): AS3
- o Link Descriptors (IPv4 interface address, neighbor IPv4 address):  
1.0.4.1 , 1.0.4.2

Attributes:

- o Adj-SID: 1042
- o LinkAttributes: see section 3.3.2 of  
[I-D.ietf-idr-ls-distribution]



## 7. IANA Considerations

This document defines a new BGPLS NLRI TYPE known as the Peer NLRI Type and a new BGP attribute known as the Peer Set SID TLV. The code points are to be assigned by IANA.

## 8. Manageability Considerations

TBD

## 9. Security Considerations

TBD

## 10. Acknowledgements

TBD

## 11. References

### 11.1. Normative References

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

### 11.2. Informative References

[I-D.filsfils-spring-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-spring-segment-routing-01 (work in progress), May 2014.

[I-D.filsfils-spring-segment-routing-central-epe]

Filsfils, C., Previdi, S., Patel, K., Aries, E., shaw@fb.com, s., Ginsburg, D., and D. Afanasiev, "Segment Routing Centralized Egress Peer Engineering", draft-filsfils-spring-segment-routing-central-epe-01 (work in progress), May 2014.

[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-05 (work in progress), May 2014.

Authors' Addresses

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

Email: [sprevidi@cisco.com](mailto:sprevidi@cisco.com)

Clarence Filsfils  
Cisco Systems, Inc.  
Brussels  
BE

Email: [cfilsfil@cisco.com](mailto:cfilsfil@cisco.com)

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

Email: [sairay@cisco.com](mailto:sairay@cisco.com)

Keyur Patel  
Cisco Systems, Inc.  
170, West Tasman Drive  
San Jose, CA 95134  
US

Email: [keyupate@cisco.com](mailto:keyupate@cisco.com)

IDR Working Group  
Internet Draft  
Intended Status: Standards Track  
Updates: 4684  
Expires: December 16, 2014

Eric C. Rosen  
Keyur Patel  
Cisco Systems, Inc.

Jeffrey Haas  
Juniper Networks, Inc.

Robert Raszuk

June 16, 2014

## Route Target Constrained Distribution of Routes with no Route Targets

draft-rosen-idr-rtc-no-rt-00.txt

### Abstract

BGP routes sometimes carry an "Extended Communities" path attribute. An Extended Communities path attribute can contain one or more "Route Targets" (RTs). By means of a procedure known as "RT Constrained Distribution" (RTC), a BGP speaker can send BGP UPDATE messages that express its interest in a particular set of RTs. Generally, RTC has been applied only to address families whose routes always carry RTs. When RTC is applied to such an address family, a BGP speaker expressing its interest in a particular set of RTs is indicating that it wants to receive all and only the routes of that address family that have at least one of the RTs of interest. However, there are scenarios in which the originator of a route chooses not to include any RTs at all, assuming that the distribution of a route with no RTs at all will be unaffected by RTC. This has led to interoperability problems in the field, where the originator of a route assumes that RTC will not affect the distribution of the route, but intermediate BGP speakers refuse to distribute that route because it does not carry any RT of interest. The purpose of this document is to clarify the effect of the RTC mechanism on routes that do not have any RTs.

### Status of this Memo

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

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

Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

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

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

#### Copyright and License 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	Some Deployment Scenarios .....	4
3	Default Behavior .....	5
4	IANA Considerations .....	5
5	Security Considerations .....	5
6	Authors' Addresses .....	6
7	Normative References .....	6
8	Informational References .....	7

## 1. Introduction

A BGP route can carry a particular type of BGP path attribute known as an "Extended Communities Attribute" [RFC4360]. Each such attribute can contain a variable number of typed communities. Certain typed communities are known as "Route Targets" (RTs) ([RFC4360], [RFC4364]).

[RFC4684] defines a procedure, known as "RT Constrained Distribution" (RTC) that allows a BGP speaker to advertise its interest in a particular set of RTs. It does so by advertising "RT membership information". (See [RFC4684] for details.) It may advertise RT membership for any number of RTs. By advertising membership for a particular RT, a BGP speaker declares that it is interested in receiving BGP routes that carry that RT.

If RTC is enabled on a particular BGP session, the session must be provisioned with the set of "address family" and "subsequent address family" (AFI/SAFIs) values to which RTC is to be applied. In [RFC4684] it is implicitly assumed that RTC will only be applied to AFI/SAFIs where all the routes carry RTs. When this assumption is true, the RTC semantics are clear. A BGP speaker advertising its interest in RT1, RT2, ..., RTk is saying that, for the AFI/SAFIs to which RTC is being applied, it is interested in any route that carries at least one of those RTs, and it is not interested in any route that does not carry at least one of those RTs.

However, [RFC4684] does not specify how the RTC procedures are to be applied to address families whose routes sometimes carry RTs and sometimes do not. Consider a BGP session between routers R1 and R2, where R1 has advertised its interest in RT1, RT2, ..., RTk, and RTC is being applied to a particular AFI/SAFI. Suppose R2 has a route of that AFI/SAFI, and that route carries no RTs. Should R2 advertise this route to R1 or not?

There are two different answers to this question, each of which seems *prima facie* reasonable:

- No, R2 should not advertise the route, because it belongs to an AFI/SAFI to which RTC is being applied, and the route does carry any of the RTs in which R1 is interested.
- Yes, R2 should advertise the route; since the route carries no RTs, the intention of the route's originator is that the distribution of the route not be constrained by the RTC mechanism.

As might be expected, "one size does not fit all", and the best

answer depends upon the particular deployment scenario, and upon the particular AFI/SAFI to which RTC is being applied.

Section 3 defines a default behavior for each existing AFI/SAFI. This default behavior will ensure proper operation of that AFI/SAFI when RTC is applied. The default behavior may of course be overridden by a local policy.

Section 3 also defines a default "default behavior" for new AFI/SAFIs. When a new AFI/SAFI is defined, the specification defining it may specify a different default behavior; otherwise the default default behavior will apply.

## 2. Some Deployment Scenarios

There are at least three deployment scenarios where lack of a clearly defined default behavior for RTC is problematic.

- [RFC6037] describes a deployed Multicast VPN (MVPN) solution. It defines a BGP address family known as "MDT-SAFI". Routes of this address family may carry RTs, but are not required to do so. In order for the RFC6037 procedures to work properly, if an MDT-SAFI route does not carry any RTs, the distribution of that route must not be constrained by RTC. However, if an MDT-SAFI route does carry one or more RTs, its distribution may be constrained by RTC.
- [GTM] specifies a way to provide "global table" (as opposed to VPN) multicast, using procedures that are very similar to those described in [RFC6513] and [RFC6514] for MVPN. In particular, it uses routes of the MCAST-VPN address family that is defined in [RFC6514]. When used for MVPN, each MCAST-VPN route carries at least one RT. However, when used for global table multicast, it is optional for certain MCAST-VPN route types to carry RTs. In order for the procedures of [GTM] to work properly, if an MCAST-VPN route does not carry any RTs, the distribution of that route must not be constrained by RTC.
- Typically, Route Targets have been carried only by routes that are distributed as part of a VPN service. However, it may be desirable to be able to place RTs on non-VPN routes (e.g., on unicast IPv4 or IPv6 routes) and then to use RTC to constrain the delivery of the non-VPN routes. For example, if a BGP speaker desires to receive only a small set of IPv4 unicast routes, and the desired routes carry one or more RTs, the BGP speaker could use RTC to advertise its interest in one or more of those RTs. In this application, the intention would be that any IPv4 unicast

route not carrying an RT would be filtered. Note that this is the opposite of the behavior needed for the other use cases discussed in this section.

### 3. Default Behavior

In order to handle the use cases discussed in Section 3, this document specifies a default behavior for the case where RTC is applied to a particular address family (AFI/SAFI), and some (or all) routes of that address family do not carry any RTs.

When RTC is applied, on a particular BGP session, to routes of the MDT-SAFI address family (SAFI=66), the default behavior is that routes that do not carry any RTs are distributed on that session.

When RTC is applied, on a particular BGP session, to routes of the MCAST-VPN address family (SAFI=5), the default behavior is that routes that do not carry any RTs are distributed on that session.

When RTC is applied, on a particular BGP session, to routes of other address families, the default behavior is that routes without any RTs are not distributed on that session. This default "default behavior" applies to all AFI/SAFIs for which a different default behavior has not been defined.

A BGP speaker may be provisioned to apply a non-default behavior to a given AFI/SAFI. This is a matter of local policy.

### 4. IANA Considerations

This document contains no actions for IANA.

### 5. Security Considerations

No security considerations are raised by this document beyond those already discussed in [RFC4684].

## 6. Authors' Addresses

Eric C. Rosen  
Cisco Systems, Inc.  
1414 Massachusetts Avenue  
Boxborough, MA, 01719  
Email: [erosen@cisco.com](mailto:erosen@cisco.com)

Keyur Patel  
Cisco Systems, Inc.  
170 Tasman Drive  
San Jose, CA, 95134  
Email: [keyupate@cisco.com](mailto:keyupate@cisco.com)

Jeffrey Haas  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
USA  
Email: [jhaas@juniper.net](mailto:jhaas@juniper.net)

Robert Raszuk  
Email: [robert@raszuk.net](mailto:robert@raszuk.net)

## 7. Normative References

[RFC4360] "BGP Extended Communities Attribute", Sangli, Tappan, Rekhter, RFC 4360, February 2006

[RFC4684] "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", Marques, Bonica, Fang, Martini, Raszuk, Patel, Guichard, RFC 4684, November 2006



## 8. Informational References

[RFC6513] "Multicast in MPLS/BGP IP VPNs", E. Rosen and R. Aggarwal, editors, RFC 6513, February 2012

[RFC6514] "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", R. Aggarwal, E. Rosen, T. Morin, Y. Rekhter, RFC 6514, February 2012

[GTM] "Global Table Multicast with BGP-MVPN Procedures", Zhang, Giuliano, Rosen, Subramanian, Pacella, Schiller, draft-zzhang-l3vpn-mvpn-global-table-mcast-04.txt, May 2014

[RFC4364] "BGP/MPLS IP Virtual Private networks", Rosen, Rekhter, RFC 4364, February 2006

[RFC6037], "Cisco Systems' Solution for Multicast in BGP/MPLS IP VPNs", Rosen, Cai, Wijnands, RFC 6037, October 2010

IDR  
Internet-Draft  
Intended status: Standards Track  
Expires: January 1, 2015

G. Van de Velde  
K. Patel  
D. Rao  
Cisco Systems  
R. Raszuk  
NTT MCL Inc.  
R. Bush  
Internet Initiative Japan  
June 30, 2014

BGP Remote-Next-Hop  
draft-vandevelde-idr-remote-next-hop-07

Abstract

The BGP Remote-Next-Hop is an optional transitive attribute intended to facilitate automatic tunneling across an AS on a per address family basis. The attribute carries one or more tunnel end-points for a NLRI. Additionally, tunnel encapsulation information is communicated to successfully setup these tunnels.

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 January 1, 2015.

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. Requirements Language . . . . .	3
3. Tunnel Encapsulation attribute versus BGP Remote-Next-Hop attribute . . . . .	3
4. BGP Remote-Next-Hop attribute TLV Format . . . . .	4
4.1. Encapsulation sub-TLVs for virtual network overlays . . . . .	5
4.1.1. Encapsulation sub-TLV for VXLAN . . . . .	6
4.1.2. Encapsulation sub-TLV for NVGRE . . . . .	7
4.1.3. Encapsulation sub-TLV for GTP . . . . .	8
5. Use Case scenarios . . . . .	8
5.1. Stateless user-plane architecture for virtualized EPC (vEPC) . . . . .	8
5.2. Stateless User-plane Architecture for virtual Packet Edge . . . . .	9
5.3. Multi-homing for IPv6 . . . . .	9
5.4. Dynamic Network Overlay Infrastructure . . . . .	10
5.5. The Tunnel end-point is NOT the originating BGP speaker . . . . .	10
5.6. Networks that do not support BGP Remote-Next-Hop attribute . . . . .	10
5.7. Networks that do support BGP Remote-Next-Hop attribute . . . . .	10
6. BGP Remote-Next-Hop Community . . . . .	10
7. IANA Considerations . . . . .	10
8. Security Considerations . . . . .	11
8.1. Protecting the validity of the BGP Remote-Next-Hop attribute . . . . .	11
9. Privacy Considerations . . . . .	11
10. Acknowledgements . . . . .	11
11. Change Log . . . . .	11
12. References . . . . .	12
12.1. Normative References . . . . .	12
12.2. Informative References . . . . .	12
Authors' Addresses . . . . .	13

## 1. Introduction

[RFC5512] defines an attribute attached to an NLRI to signal tunnel end-point encapsulation information between two BGP speakers for a single tunnel. It assumes that the exchanged tunnel endpoint is the NLRI.

This document defines a new BGP transitive attribute known as a Remote-Next-Hop BGP attribute for Intra-AS and Inter-AS usage which removes the assumption of both a single tunnel and that the exchanged NLRI is the tunnel endpoint.

The tunnel endpoint information and the tunnel encapsulation information is carried within a Remote-Next-Hop BGP attribute. This attribute can be added to any BGP NLRI. This way the Address Family (AF) of the NLRI exchanged is decoupled from the tunnel SAFI address-family defined in [RFC5512].

## 2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in [RFC2119] only when they appear in all upper case. They may also appear in lower or mixed case as English words, without any normative meaning.

## 3. Tunnel Encapsulation attribute versus BGP Remote-Next-Hop attribute

The Tunnel Encapsulation attribute [RFC5512] is based on the principle that the tunnel end-point is the BGP speaker originating the update and is inserted as the NLRI in the exchange, with the consequence that it is impossible to set the endpoint to an arbitrary address. It is also assumed that there is only a single tunnel between endpoints.

There are use cases where it is desired that the tunnel end-point address should be a different address, or set of addresses, than the originating BGP speaker. It is also useful to be able to signal different encapsulation parameters for different prefixes with the same remote tunnel end-point. The BGP Remote-Next-Hop attribute provides the ability to have one or more different tunnel end-point addresses from IPv4, IPv6 and/or other address-families, and be able to signal next-hop encapsulation parameters along with any prefix.

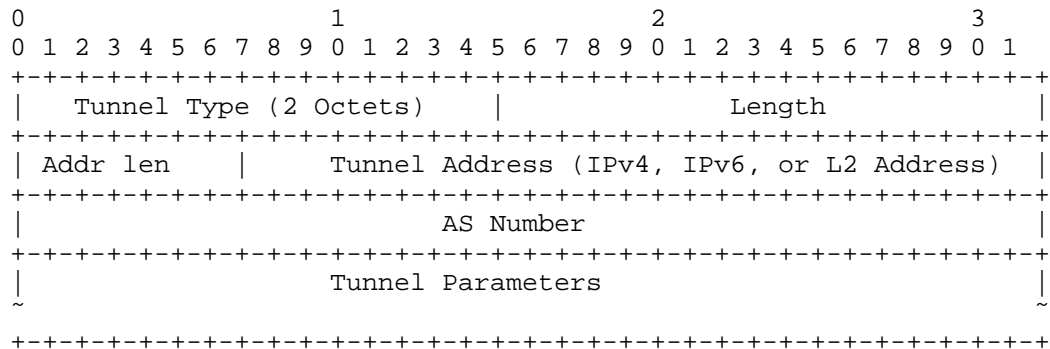
The sub-TLVs from the Tunnel Encapsulation Attribute [RFC5512] are reused for the BGP Next-Hop-Attribute.

Due to the intrinsic nature of both attributes, the tunnel encapsulation end-point assumes that the tunnel end-point is both the NLRI exchanged and the originating router, while the BGP Remote-Next-Hop attribute is inserted for an exchanged NLRI by adding a set of tunnel end-points and hence these two attributes are mutually exclusive.

#### 4. BGP Remote-Next-Hop attribute TLV Format

This attribute is an optional transitive attribute [RFC1771].

The BGP Remote-Next-Hop attribute is composed of a set of Type-Length-Value (TLV) encodings. The type code of the attribute is (IANA to assign). Each TLV contains information corresponding to a particular tunnel technology and tunnel end-point address. The TLV is structured as follows:



Tunnel Type (2 octets): identifies the type of tunneling technology being signaled. This document specifies the following types:

- L2TPv3 over IP [RFC3931]: Tunnel Type = 1
- GRE [RFC2784]: Tunnel Type = 2
- IP in IP [RFC2003] [RFC4213]: Tunnel Type = 7

This document also defines the following types:

- VXLAN: Tunnel Type = 8
- NVGRE: Tunnel Type = 9
- GTP: Tunnel Type = 10
- MPLS-in-GRE: Tunnel Type = 11

Unknown types MUST be ignored and skipped upon receipt.

Length (2 octets): the total number of octets of the value field.

Tunnel Address Length - Addr len (1 octet): Length of Tunnel Address. Set to 4 bytes for an IPv4 address, 16 bytes for an IPv6 address or 8 bytes for a MAC address.

AS Number - The AS number originating the BGP Remote-Next-Hop attribute and is either a 2-byte AS or 4-Byte AS number

Tunnel Parameter - (variable): comprised of multiple sub-TLVs. Each sub-TLV consists of three fields: a 1-octet type, 1-octet length, and zero or more octets of value. The sub-TLV definitions and the sub-TLV data are described in depth in [RFC5512].

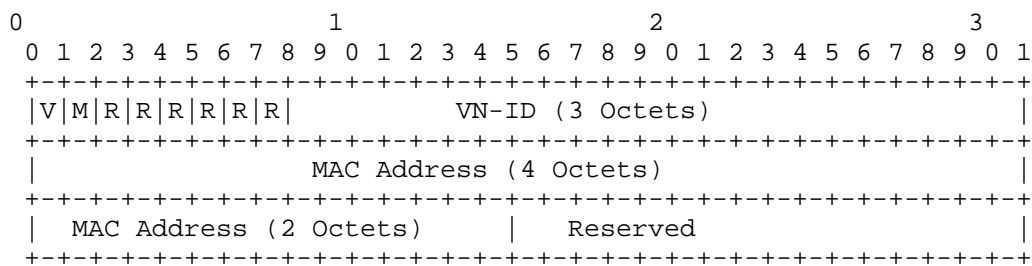
#### 4.1. Encapsulation sub-TLVs for virtual network overlays

A VN-ID may need to be signaled along with the encapsulation types for DC overlay encapsulations such as [VXLAN] and [NVGRE]. The VN-ID when present in the encapsulation sub-TLV for an overlay

encapsulation, MUST be processed by a receiving device if it is capable of understanding it. The details regarding how such a signaled VN-ID is processed and used is defined in specifications such as [IPVPN-overlay] and [EVPN-overlay].

#### 4.1.1.1. Encapsulation sub-TLV for VXLAN

This document defines a new encapsulation sub-TLV format, defined in [RFC5512], for VXLAN tunnels. When the tunnel type is VXLAN, the following is the structure of the value field in the encapsulation sub-TLV:



V: When set to 1, it indicates that a valid VN-ID is present in the encapsulation sub-TLV.

M: When set to 1, it indicates that a valid MAC Address is present in the encapsulation sub-TLV.

R: The remaining bits in the 8-bit flags field are reserved for further use. They MUST be set to 0 on transmit and MUST be ignored on receipt.

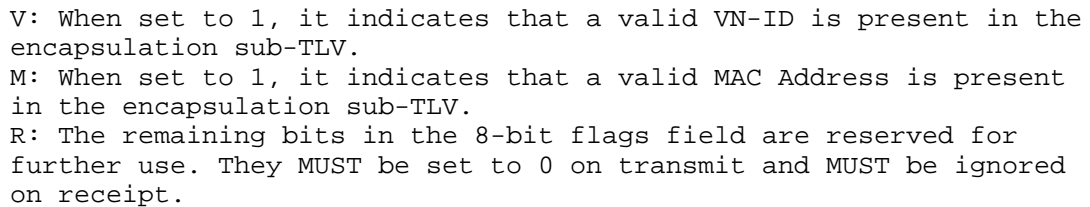
VN-ID: Contains a 24-bit VN-ID value, if the 'V' flag bit is set. If the 'V' flag is not set, it SHOULD be set to zero and MUST be ignored on receipt.

The VN-ID value is filled in the VNI field in the VXLAN packet header as defined in [VXLAN].

MAC Address: Contains an Ethernet MAC address if the 'M' flag bit is set. If the 'M' flag is not set, it SHOULD set to all zeroes and MUST be ignored on receipt.

The MAC address is local to the device advertising the route, and should be included as the destination MAC address in the inner Ethernet header immediately following the outer VXLAN header, in the packets destined to the advertiser.

This document defines a new encapsulation sub-TLV format, defined in [RFC5512], for NVGRE tunnels. When the tunnel type is NVGRE, the following is the structure of the value field in the encapsulation sub-TLV:



The MAC address is local to the device advertising the route, and should be included as the destination MAC address in the inner Ethernet header immediately following the outer NVGRE header, in the packets destined to the advertiser.





is offloaded at entity that is closer to the mobile node (ex. eNodeB or closer anchor).

### 5.2. Stateless User-plane Architecture for virtual Packet Edge

With the emergence of the NFV technologies, different architectures are proposed for virtualised Services. These functions will normally run in the datacenter. BGP remote-next-hop can be used to inject traffic into the virtualised services running in the datacenter for a optimized, simple and clean routing architecture. BGP Remote Next Hop can simplify the orchestration or provisioning layer by signalling the tunnel endpoint (virtual provider edge router) and the encapsulation protocol.

If this is used together with orchestrated traffic steering mechanisms (i.e. BGP Flowspec) , it is possible to differentiate at application level, and forward each different traffic types towards the desired destination.

### 5.3. Multi-homing for IPv6

When an end-user IPv6 network is multi-homed to the Internet, it may be assigned more than a single prefix originated by various upstream ASs. Each AS prefers to only announce a supernet of all its assigned IPv6 prefixes, unlike IPv4 where the AS announced the end-users assigned prefix. The goal of this BGP policy behaviour is to keep the number of entries in the IPv6 global BGP table to a minimum, it also it also results in well known resiliency improvements.

For example, if an end-user IPv6 is peering with 2 different Service providers AS1 and AS2. In this case the IPv6 end-user will have at least one prefix assigned from each of these service providers. The devices at the IPv6 end-user will each receive an address from these prefixes. The devices will in most cases, when building IPv6 sessions (TCP, etc...), do so with only a single IPv6 address. The decision which IPv6 address the device will use is documented in [RFC3484].

If one of the links between the end-user and one the neighboring AS's fails, a consequence will be that a set of sessions need to be reset, or that a section of the end-user network becomes unreachable.

With usage of the BGP-remote-Next-Hop attribute the service provider can tunnel that packet towards an alternate BGP Remote-Next-Hop at the end-users alternate provider and restore the network connectivity even though the local link towards the end-user is broken.

#### 5.4. Dynamic Network Overlay Infrastructure

The BGP Remote-Next-Hop extension allows signaling tunnel encapsulations needed to build and dynamically create an overlay tunneled network with traffic isolation and virtual private networks.

#### 5.5. The Tunnel end-point is NOT the originating BGP speaker

Note that, in each network environment, the originating router is the preferred tunnel end-point server. It may be that the network administrator has deployed an independent set of tunnel end-point servers across their network, which may or may not speak BGP. The BGP Remote-Next-Hop attribute provides the ability to signal this via BGP.

#### 5.6. Networks that do not support BGP Remote-Next-Hop attribute

If a device does not support this attribute, and receives this attribute, then normal NLRI BGP forwarding is used as the attribute is optional and transitive.

#### 5.7. Networks that do support BGP Remote-Next-Hop attribute

If a BGP speaker does understand this attribute, and receives this attribute, then the BGP speaker MAY, by configuration, skip use or not use the information within this attribute.

#### 6. BGP Remote-Next-Hop Community

place-holder for an BGP extension to signal valid prefixes allowed to be considered as tunnel end-points. To be completed.

#### 7. IANA Considerations

This document defines a new BGP attribute known as a BGP Remote-Next-Hop attribute. We request IANA to allocate a new attribute code from the "BGP Path Attributes" registry with a symbolic name "Remote-Next-Hop" attribute.

We also request IANA to allocate four new BGP Tunnel Types from the "BGP Tunnel Encapsulation Attribute Tunnel Types" registry with the following symbolic names: "VXLAN" with Tunnel type 8, "NVGRE" with Tunnel type 9, "GTP" with Tunnel type 10, and "MPLS-in-GRE" with Tunnel type 11.

## 8. Security Considerations

This technology could be used as technology as man in the middle attack, however with existing RPKI validation for BGP that risk is reduced.

The distribution of Tunnel end-point address information can result in potential DoS attacks if the information is sent by malicious organisations. Therefore is it strongly recommended to install traffic filters, IDSs and IPSs at the perimeter of the tunneled network infrastructure.

### 8.1. Protecting the validity of the BGP Remote-Next-Hop attribute

It is possible to inject a rogue BGP Remote-Next-Hop attribute to an NLRI resulting in Monkey-In-The-Middle attack (MITM). To avoid this type of MITM attack, it is strongly recommended to use a technology a mechanism to verify that for NLRI it is the expected BGP Remote-Next-Hop. We anticipate that this can be done with an expansion of RPKI-Based origin validation, see [I-D.ietf-sidr-pfx-validate].

This does not avoid the fact that rogue AS numbers may be inserted or injected into the AS-Path. To achieve protection against that threat BGP Path Validation should be used, see [I-D.ietf-sidr-bgpsec-overview].

## 9. Privacy Considerations

This proposal may introduce privacy issues, however with BGP security mechanisms in place they should be prevented.

## 10. Acknowledgements

The authors would like to thanks Satoru Matsushima, Ryuji Wakikawa and Miya Kohno for their usefull vEPC discussions. Istvan Kakonyi provided insight in the vPE use case scenario.

## 11. Change Log

Initial Version: 16 May 2012

Hacked for -01: 17 July 2012

Hacked for -05: 07 January 2014

Hacked for -07: 30 June 2014

## 12. References

### 12.1. Normative References

- [RFC1771] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2784] Farinacci, D., Li, T., Hanks, S., Meyer, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 2784, March 2000.
- [RFC3484] Draves, R., "Default Address Selection for Internet Protocol version 6 (IPv6)", RFC 3484, February 2003.
- [RFC3931] Lau, J., Townsley, M., and I. Goyret, "Layer Two Tunneling Protocol - Version 3 (L2TPv3)", RFC 3931, March 2005.
- [RFC4213] Nordmark, E. and R. Gilligan, "Basic Transition Mechanisms for IPv6 Hosts and Routers", RFC 4213, October 2005.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.
- [RFC6459] Korhonen, J., Soininen, J., Patil, B., Savolainen, T., Bajko, G., and K. Iisakkila, "IPv6 in 3rd Generation Partnership Project (3GPP) Evolved Packet System (EPS)", RFC 6459, January 2012.

### 12.2. Informative References

- [I-D.ietf-sidr-bgpsec-overview]  
Lepinski, M. and S. Turner, "An Overview of BGPSEC", draft-ietf-sidr-bgpsec-overview-04 (work in progress), December 2013.
- [I-D.ietf-sidr-pfx-validate]  
Mohapatra, P., Scudder, J., Ward, D., Bush, R., and R. Austein, "BGP Prefix Origin Validation", draft-ietf-sidr-pfx-validate-10 (work in progress), October 2012.

[I-D.mahalingam-dutt-dcops-vxlan]

Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-02 (work in progress), August 2012.

[I-D.matsushima-stateless-uplane-vepc]

Matsushima, S. and R. Wakikawa, "Stateless user-plane architecture for virtualized EPC (vEPC)", draft-matsushima-stateless-uplane-vepc-01 (work in progress), July 2013.

[I-D.sridharan-virtualization-nvgre]

Sridharan, M., Greenberg, A., Venkataramaiah, N., Wang, Y., Duda, K., Ganga, I., Lin, G., Pearson, M., Thaler, P., and C. Tumuluri, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-02 (work in progress), February 2013.

#### Authors' Addresses

Gunter Van de Velde  
Cisco Systems  
De Kleetlaan 6a  
Diegem 1831  
Belgium

Phone: +32 2704 5473  
Email: gvandeve@cisco.com

Keyur Patel  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95124 95134  
USA

Email: keyupate@cisco.com

Dhananjaya Rao  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95124 95134  
USA

Email: dhrao@cisco.com

Robert Raszuk  
NTT MCL Inc.  
101 S Ellsworth Avenue Suite 350  
San Mateo, CA 94401  
US

Email: robert@raszuk.net

Randy Bush  
Internet Initiative Japan  
5147 Crystal Springs  
Bainbridge Island, Washington 98110  
US

Email: randy@psg.com

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 5, 2015

Z. Li  
L. Zhang  
S. Hares  
Huawei Technologies  
July 4, 2014

NEXTHOP\_PATH\_RECORD ATTRIBUTE for BGP  
draft-zhang-idr-nexthop-path-record-00

Abstract

As the BGP is deployed in a single Autonomous System using converged networks such as Seamless MPLS, it is desirable for BGP to carry more IGP nexthop pathway information to help select routing more intelligently. One example of a Seamless MPLS deployment is the Mobile BackHaul (MBH) deployment with multiple IGPs Areas per ASN. This document describes a new optional transitive path attribute, NEXTHOP\_PATH\_RECORD ATTRIBUTE for BGP that records the next hop path which can be used by BGP network management to monitor and manage the BGP infrastructure via management interfaces (such as I2RS).

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 January 5, 2015.



## 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. Definition of NEXTHOP_PATH_RECORD ATTRIBUTE . . . . .	3
3. Process of NEXTHOP_PATH_RECORD ATTRIBUTE . . . . .	4
3.1. Creating and Modifying the NEXTHOP_PATH_RECORD Attribute . . . . .	4
4. Deployment considerations . . . . .	5
4.1. Customized Best Path Selection . . . . .	5
4.2. Use in Seamless MPLS case with PE-RR . . . . .	5
5. IANA Considerations . . . . .	6
6. Security Considerations . . . . .	7
7. References . . . . .	7
7.1. Normative References . . . . .	7
7.2. Informative References . . . . .	7
Authors' Addresses . . . . .	8

## 1. Introduction

Network topologies have become more densely interconnected at the network level. Examples of this mesh topologies can be a data center or the converged carrier network that [I-D.ietf-mpls-seamless-mpls] architecture supports. In these mesh topologies, there may be multiple highly meshed IGPs connected by IBGP into a single AS. Scalability and redundancy may require exterior processes to calculate a better way through the array of next-hop pathways attached to BGP Routes of any AFI/SAFI pairing.

This document proposes a new path attribute that can record the next hop path of the route to help BGP route election and network management. In the deployment considerations section, this draft provides a deployment scenario linked to the use of I2RS and BGP Cost Community attribute.

## 2. Definition of NEXTHOP\_PATH\_RECORD ATTRIBUTE

The NEXTHOP\_PATH\_RECORD ATTRIBUTE is an optional transitive BGP Path Attribute. The NEXTHOP\_PATH\_RECORD ATTRIBUTE type is defined as below (refer to [RFC4271] ):

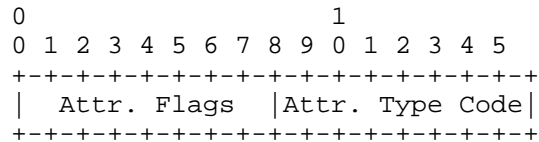


Figure 2 NEXTHOP\_PATH\_RECORD Type definition

Attr. Flags

SHOULD be optional transitive

Attr. Type Code

SHOULD be allocated by IANA

NEXTHOP\_PATH\_RECORD is composed of a sequence of next hop path segments. Each next hop path segment is represented by a triple <path segment type, path segment length, path segment value>. The format of the next hop path segment is shown in the figure 3.

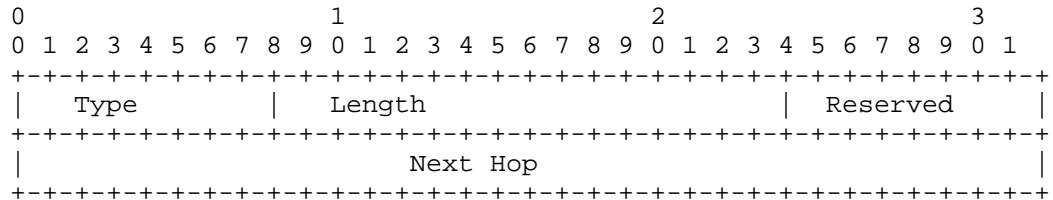


Figure 3 NH\_SEQUENCE\_V4 TLV

- Type: A single octet encoding the TLV Type. The Type of "NH\_SEQUENCE\_V4" is defined in this document, which needs to be allocated by IANA. The procedure for next hop path segment usage for IPv6 or other extensions will be described in the future revisions of this document.

Length: Two octets encoding the length in octets of the TLV, including the type and length fields. The length is encoded as an unsigned binary integer.

Reserved: A single octet that must be zero now.

NextHop: four octets encoding for the route next hop address.

### 3. Process of NEXTHOP\_PATH\_RECORD ATTRIBUTE

The NEXTHOP\_PATH\_RECORD attribute defined in this section is an optional transitive BGP Path Attribute as described in [RFC4271].

#### 3.1. Creating and Modifying the NEXTHOP\_PATH\_RECORD Attribute

When a BGP speaker distributes a route to its BGP peer within UPDATE message, the NEXTHOP\_PATH\_RECORD ATTRIBUTE should be processed based on different route states:

1. If the route is originated in this BGP speaker
  - \* If the NEXTHOP\_PATH\_RECORD ATTRIBUTE is supported, the NEXTHOP\_PATH\_RECORD ATTRIBUTE SHOULD be originated including the BGP speaker's own next hop address in a next hop path segment. In this case, the next hop address of the originating BGP speaker will be the only entry of the next hop path segment, and this path segment will be the only segment in NEXTHOP\_PATH\_RECORD ATTRIBUTE.
  - \* If the NEXTHOP\_PATH\_RECORD ATTRIBUTE is not supported, the route will be distributed without NEXTHOP\_PATH\_RECORD ATTRIBUTE.
2. if the route is received from one BGP speaker's UPDATE message
  - \* If the NEXTHOP\_PATH\_RECORD ATTRIBUTE is NULL and the local BGP speaker supports NEXTHOP\_PATH\_RECORD ATTRIBUTE, when the route is propagated to another IBGP speaker with next hop self (NHS ), the NEXTHOP\_PATH\_RECORD ATTRIBUTE SHOULD be originated including the BGP speaker's own next hop address in a next hop path segment. In this case, the next hop address of this BGP speaker will be the only entry to the next hop path segment, and this path segment will be the only segment in NEXTHOP\_PATH\_RECORD ATTRIBUTE
  - \* If the NEXTHOP\_PATH\_RECORD ATTRIBUTE is non-NULL and the local BGP speaker support NEXTHOP\_PATH\_RECORD ATTRIBUTE, when the route is propagated to another IBGP speaker with next hop self (NHS ), the BGP speaker MUST appends its own next hop address as the last one of the next hop path segments.
  - \* If the NEXTHOP\_PATH\_RECORD ATTRIBUTE is NULL and the local BGP speaker support NEXTHOP\_PATH\_RECORD ATTRIBUTE, when the route is propagated to another BGP speaker without changing the next

hop by the BGP speaker, the BGP speaker MUST NOT originate the NEXTHOP\_PATH\_RECORD ATTRIBUTE.

- \* If the NEXTHOP\_PATH\_RECORD ATTRIBUTE is non-NULL and the local BGP speaker support NEXTHOP\_PATH\_RECORD ATTRIBUTE, when the route is propagated to another BGP speaker without changing the next hop by the BGP speaker, the BGP speaker MUST NOT change the next hop path sequence.
- \* If the BGP speaker does not support NEXTHOP\_PATH\_RECORD ATTRIBUTE, it SHOULD keep the NEXTHOP\_PATH\_RECORD ATTRIBUTE unchanged whether the route is distribute with next hop self or not.

#### 4. Deployment considerations

Two deployment examples are given to demonstrate how the nexthop information can be deployed in existing BGP technologies to monitor and tune the IBGP cloud to provide better operation. maintenance. This attribute has records information.

##### 4.1. Customized Best Path Selection

The next\_hop information gathered on an IBGP or EBP route could be used by off-line decision processing to select paths, and re-inserted as policy to affect the decision making via I2RS. The I2RS BGP use case draft [I-D.keyupate-i2rs-bgp-usecases] describes this its section on customized best path selection (section 4.1) which uses the BGP feature [I-D.ietf-idr-custom-decision] to set a custom decision community.

##### 4.2. Use in Seamless MPLS case with PE-RR

In a Seamless MPLS network [I-D.ietf-mpls-seamless-mpls], the Area Border Routers (ABRs) which run IBGP may act RR-clients or be part of RR mesh as described in section 5.1.7. Seamless MPLS places restrictions on the BGP NEXT\_HOP to make Seamless MPLS work in the general case. With the transmittal of the next-hop-path attribute offline calculation can insert a better pathway decision using the BGP customer. A sample description of in a seamless MPLS is included below.

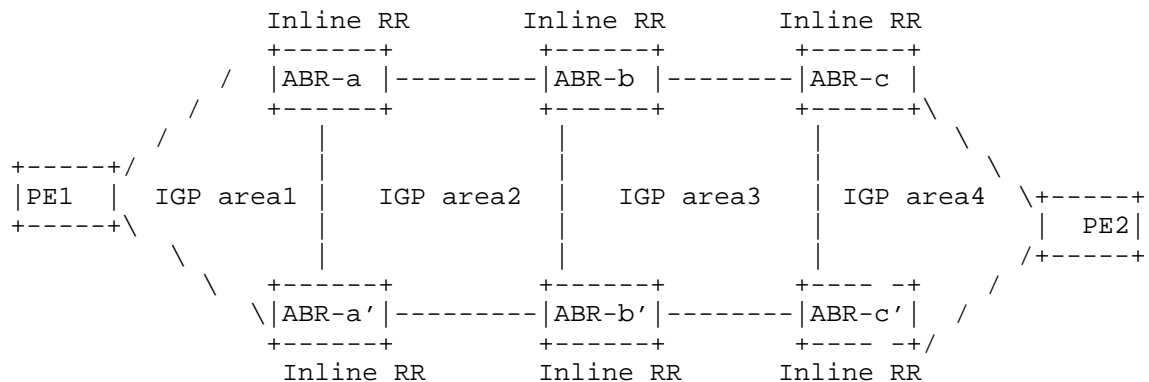


Figure 1 Seamless MPLS Network with Multiple IGP Areas

Just like Figure 1 shown, PE1 and PE2 are BGP VPN service end-point. IBGP peers runs contiguously between ABRs in different IGP areas, and each ABR works as inline RR. When labeled BGP routes or BGP VPN routes originated from PE1 is distributed to the other service end-point PE2, the route can be reflected by the ABRs one by one with next hop self (NHS).

The inline RR will distribute the route to all of the IBGP peers except the IBGP peer from which the route was received. As a result, an ABR may receive routes of the same prefix from different IBGP peers with different next hop. Traditionally the BGP RR should select the best route to reflect to other IBGP peers. But in this network the route selection process will be more complex which needs to introduce complex route policy.

The NEXTHOP\_PATH ATTRIBUTE can optionally collect information on the pathway the routes are taking through the IBGP mesh. This information may aid in monitoring paths in this complex path or in offline processing that reduces complex policy to simple instantiation of community policy or the BGP Custom Cost community to allow specialized pathways through the MPLS mesh.

## 5. IANA Considerations

IANA need to assign the codepoint in the "BGP Path Attributes" registry to the NEXTHOP\_PATH\_RECORD ATTRIBUTE.

IANA shall create a registry for "next hop path segment". The type field consists of a single octet, with possible values from 0 to 255. The allocation policy for this field is to be "Standards Action with Early Allocation". A new Type should be defined as "NH\_SEQUENCE\_V4".

## 6. Security Considerations

Note that, the NEXTHOP\_PATH\_RECORD ATTRIBUTE is defined as a optional transitive BGP Path attribute. Both the IBGP and EBGp speaker can use this attribute. When an ASBR propagates the route receive from a IBGP peer to an EBGp peer, the NEXTHOP\_PATH\_RECORD ATTRIBUTE will be distribute to the EBGp Speaker which may be controlled by other Service Provider. If the EBGp speaker can support the NEXTHOP\_PATH\_RECORD ATTRIBUTE, it can parse the NEXTHOP\_PATH\_RECORD ATTRIBUTE to get the inner network architecture of the other network.

BGP requires the use of TCP-MD5 [RFC2385] or TCP-AO [RFC5925]). Use of encryption will prevent unauthorized view of the NEXTHOP\_PATH\_RECORD attribute. For those not supporting the required TCP-MD5 or TCP-AO, the NEXTHOP\_PATH\_RECORD ATTRIBUTE capability SHOULD disabled for specific BGP speaker to prevent this attack.

## 7. References

### 7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.
- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, June 2010.

### 7.2. Informative References

- [I-D.ietf-idr-custom-decision]  
Retana, A. and R. White, "BGP Custom Decision Process", draft-ietf-idr-custom-decision-04 (work in progress), November 2013.
- [I-D.ietf-mpls-seamless-mpls]  
Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07 (work in progress), June 2014.

[I-D.keyupate-i2rs-bgp-usecases]

Patel, K., Fernando, R., Gredler, H., Amante, S., White, R., and S. Hares, "Use Cases for an Interface to BGP Protocol", draft-keyupate-i2rs-bgp-usecases-03 (work in progress), June 2014.

#### Authors' Addresses

Zhenbin Li  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: lizhenbin@huawei.com

Li Zhang  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China

Email: monica.zhangli@huawei.com

Susan Hares  
Huawei Technologies  
7453 Hickory Hill  
Saline, MI 48176  
USA

Email: shares@ndzh.com