

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 12, 2014

C. Camilo Cardona  
P. Pierre Francois  
IMDEA Networks  
S. Ray  
K. Patel  
P. Paolo Lucente  
Cisco Systems  
P. Mohapatra  
Cumulus Networks  
July 11, 2013

BGP Path Marking  
draft-bgp-path-marking-00

Abstract

The potential advertisement of non-best paths by a BGP speaker supporting the add-path or the best-external extensions makes it difficult for other BGP speakers to identify the paths that have been selected as best by those who advertise them. This information is required for proper operation of some applications. Towards that end, this document proposes marking the paths using extended communities that encode the path type.

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 12, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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 . . . . .	2
2. The BGP Path Type Community . . . . .	4
3. Rules . . . . .	5
4. Operational Considerations . . . . .	6
5. Applications . . . . .	7
5.1. Avoiding suboptimal routing in Inter-AS VPN . . . . .	7
5.2. Monitoring applications . . . . .	9
5.3. SDN applications . . . . .	9
5.4. Selective Best-path . . . . .	10
6. IANA Considerations . . . . .	10
7. Security Considerations . . . . .	10
8. Contributors . . . . .	10
9. Acknowledgments . . . . .	10
10. References . . . . .	11
10.1. Normative References . . . . .	11
10.2. Informative References . . . . .	11
Authors' Addresses . . . . .	11

## 1. Introduction

When there are multiple paths for a given address prefix, BGP chooses one of the paths as the "best-path" according to the best-path selection rules prescribed in [RFC4271] and installs the best-path in its forwarding table. Classically, each BGP speaker advertises only

the best-path to its peers. So when a BGP speaker receives a path from one of its peers, it is assured that the path is used by the peer for forwarding and all other peers have received the same path from this peer. This leads to consistent routing in a BGP network.

The classical advertisement rule of sending only the best-path does not convey the full routing state of a destination present on a BGP speaker to its peers.

- o In order to improve link bandwidth utilization, most BGP implementations choose additional paths, that satisfy certain conditions, as "multi-path", and install them in the forwarding table. Incoming packets for that destination are load-balanced across the best-path and the multi-path(s). I.e., there may be paths installed in the forwarding table that are not advertised to the peers.
- o When an Autonomous System (AS) deploys a route-reflector ([RFC4456]) instead of using full IBGP mesh, the BGP speakers receive only the route reflector's best-path and therefore lose information about the best-paths of other IBGP peers.
- o If an IBGP path is chosen as the best-path by a non-route-reflector BGP speaker, then the best-path is not sent to its IBGP peers. Thus the IBGP peers learn nothing from this BGP speaker even though it might have other EBGp paths for that destination.
- o Even when a BGP speaker selects an EBGp path as the best-path and advertises it to its peers, it may have additional EBGp paths for the destination. Should those paths be advertised a priori, they could be used by the peers in the event of loss of reachability of the best-path resulting in faster convergence.

There are extensions to the classical BGP advertisement rule to provide additional information about the routing state of a destination. A BGP speaker supporting the best-external [I-D.ietf-idr-best-external] extension sends its best external path to its IBGP peers when the best-path is an IBGP path. A BGP speaker supporting the add-path [I-D.ietf-idr-add-paths] extension advertises multiple paths for a given address prefix.

With best-external or add-path extensions in use, when a BGP speaker receives a path from a peer, that path may not be the best-path, or it may not be installed in the peer's forwarding table. In some scenarios, knowledge of the path type - i.e., whether the path is the best-path, or whether the path is installed in the forwarding table - is essential.

For instance, in a typical dual-homed VPN in primary-backup configuration, the backup path is created by advertising the best-external path from the backup PE with worse LOCAL\_PREF. However, when the customer adds a site in another AS, the LOCAL\_PREF information does not reach that site. As a result, data traffic coming from that site may incorrectly be forwarded over the backup link instead of the primary link.

Similarly when an add-path enabled peer receives multiple paths from a peer, it does not know which one among those paths is the best-path and which ones are installed in the forwarding table. An exogenous monitoring system, e.g., would require that information to properly tweak the policies on the router to effect desired forwarding optimization.

This draft proposes marking the advertised paths by an extended community, called Path Type community, that encodes the path type. The path type provides the necessary information to the BGP speakers about how the path is used by the sender when add-path or best-external extensions are in use.

## 2. The BGP Path Type Community

The BGP Path Type Community is an IPv4 Address Extended Community ([RFC4360]) defined as follows:

### Type Field:

The value of the high-order octet of the extended Type Field is 0x01, which indicates that it is transitive. The value of low-order octet of the extended type field for this community is TBD.

### Value Field:

The Value field contains two sub-fields, described below:

```
+-----+
| Router-ID (4 octet) |
+-----+
| Path type (2 octet) |
+-----+
```

The Router-ID field contains the BGP identifier of the BGP speaker that adds the Path Type community to a path.

The Path type field contains a bitfield where each bit encodes a specific role of the path. Multiple bits may be set when a path is used in multiple roles.

Value	Path type
0x0000	Unknown
0x0001	Best-path
0x0002	Best-external path
0x0004	Multi-path
0x0008	Backup path
0x0010	Uninstalled path
0x0020	Unreachable path

Table 1: Path Type Values

The best-path is defined in [RFC4271] and the best-external path is defined in [I-D.ietf-idr-best-external].

A multi-path is not the best-path but installed in the forwarding table and used for forwarding packets. We use the convention that the best-path is not considered a multi-path.

A backup path is installed in the forwarding table, but it is not used for forwarding until all multipath(s) and the best-path become unreachable. Backup paths are used for fast convergence in the event of failures.

All other reachable paths are marked as 'Uninstalled'.

Lastly, all paths that are considered unreachable are marked as 'Unreachable'. Unreachable paths may be sent only in special cases (such as to a monitoring application).

### 3. Rules

- o A BGP speaker MAY add the Path Type community to an originated path.
- o When a BGP speaker receives a path from a peer and propagates it without changing the NEXT\_HOP to self:
  - \* If the path contained a Path Type community, it MUST be retained in the propagated path.

- \* If the path did not contain a Path Type community, the speaker MAY add a Path Type community with 'Unknown' value.
- o When a path received from a peer is propagated after changing the NEXT\_HOP to self:
  - \* If the path did not contain a Path Type community, the Path Type community indicating the path role MAY be added.
  - \* If the path contained a Path Type community:
    - + If data traffic entering the router for the given destination may be forwarded over other paths (e.g., for doing load balancing), then the existing Path Type community MUST be removed. The BGP speaker MAY add its own Path Type community.
    - + If data traffic entering the router for the given destination is forwarded only along the given path, then the existing Path Type community MAY be retained.

In all cases, when a BGP speaker adds its own Path Type community, it sets its own router-id in the community. Note that BGP router-id need not be unique across ASes.

The above rule-set prevents a route reflector from modifying the Path Type community set by its client (unless the route reflector is changing the NEXT\_HOP to self).

When a peer is capable of sending only one path for a given address prefix and it sends the path without any Path Type community, the path MAY be considered as the best-path of the peer. In all other cases, a path without any Path Type community SHOULD be considered to have an 'Unknown' Path type.

A local policy might modify the above rules. For instance, if a monitoring application peers with a BGP speaker with add-path capability for the sole purpose of learning its paths and their types, then the speaker may always add its own Path Type community when it advertises the paths to that peer even if it does not change the NEXT\_HOP to self. Such overriding policies should be used with caution if the advertised paths may impact forwarding decisions in the network.

#### 4. Operational Considerations

If a speaker receives a path with a Path Type community with an invalid combination of bits (e.g., both 'Multi-path' and 'Backup'

bits are set), the path MUST NOT be considered invalid. Such error cases SHOULD be logged through other means.

An implementation SHOULD provide a configurable option for the user to indicate whether a path should be readvertised when its type is changed. If the user does not configure the option, the BGP speaker MUST NOT readvertise a path just to update its Path Type community (e.g., when a path type changes from 'Multi-path' to 'Uninstalled' due to a change in IGP metric).

An implementation SHOULD provide a configurable option for removing Path Type communities from paths that are advertised to untrusted peers.

An implementation SHOULD mark all paths for a given address prefix consistently. If one of the paths is marked, then all other paths SHOULD be marked.

An implementation MAY modify its best-path selection algorithm to take path type information into account. For instance, paths with type 'Best-path' MAY be preferred over paths of other types. Similarly, paths of type 'Best-external' MAY be considered ineligible for being a multipath.

## 5. Applications

In this section, we illustrate some applications that benefit from the Path Type community proposed in this draft.

### 5.1. Avoiding suboptimal routing in Inter-AS VPN

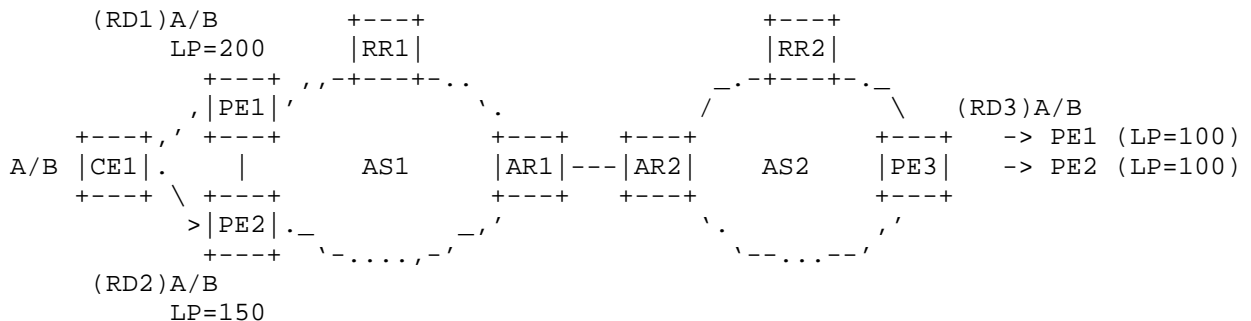


Figure 1: Inter-AS VPN scenario

Figure 1 depicts an L3VPN network that spans two ASes: AS1 and AS2. The ASes may be connected using either Option-B or Option-C

techniques [RFC4364]. A customer site with equipment CE1 is dual-homed in AS1, connected to PE1 and PE2. For prefix A/B, the customer prefers to use the link between CE1 and PE1. This routing preference is expressed by setting the LOCAL\_PREF of the prefix advertised by PE1 to a higher value than that of the prefix advertised by PE2. This causes PE2 to use PE1's route as the best-path and its own EBGp path becomes the best-external path. PE2 is configured to advertise its best-external path. Therefore, both PEs continue to advertise their own EBGp path. The provider uses unique route-distinguishers for its VPNs. So PE1 and PE2 advertises different VPN prefixes: (RD1)A/B and (RD2)A/B. Both these prefixes are advertised to PE3 in AS2. PE3 imports both paths to its own VPN with route-distinguisher RD3.

#### Existing behavior:

Since LOCAL\_PREF is not sent across AS boundary, both paths on PE3 have the default LOCAL\_PREF of 100. As a result the best-path selection on PE3 may boil down to tie breaking steps and the path towards PE2, which is the best-external path, may be chosen. Alternately, the path from PE2 may be chosen as the multipath and may be used for load-balancing. Therefore, some or all data traffic entering PE3 would reach CE1 via PE2, which is not what the customer desired.

#### Behavior with Path Type Community:

When PE2 advertises its path, it adds the best-external Path Type community. This community is preserved across AS boundary. If option C is used, then RR1 or RR2 does not change the NEXT\_HOP and hence the community is preserved according to the rule-set (Section 3). If option B is used, then the community reaches AR1 since RR1 does not change the NEXT\_HOP. At AR1, (RD2)A/B has only one path and forwarding traffic entering AR1 from AR2 for this destination (determined by the outer label) would use this path. Therefore, AR1 retains the Path Type community set by PE2. The same applies to AR2. So at PE3, the path to PE2 has the best-external Path Type community and therefore PE3 can choose to not use this path for forwarding.

If the best-path algorithm takes the Path Type community values into account, it eliminates the need for setting LOCAL\_PREF to deprefer the best-external path even within a single AS. This simplifies the network design and management.

Instead of using Path Type communities, it is possible to use policies on the border routers (AR1 and AR2 for option B, or RR1 and



RR2 for option C) to recreate the LOCAL\_PREF in AS2 (e.g., by matching on the RD and the prefix). However, the recreated LOCAL\_PREF may interfere with the local policies set in AS2 (e.g., if there are other paths in AS2 for A/B that the customer wants to use as secondary paths). In addition, such policies are error-prone and complex to manage, especially when the customer is allowed to change the primary/backup relationships between PE1 and PE2 on its own. The standardized mechanism of Path Type community is free from such drawbacks.

## 5.2. Monitoring applications

A modern Service Provider (SP) network may contain thousands of BGP routers. For planning, proper engineering and operation of a backbone, it is a good practice to continuously monitor the routers' states and perhaps keep a history. Many Network Management Systems (NMS) establish IBGP sessions with BGP speakers to collect the paths the speaker has. When the speaker supports add-path (or best-external), the NMS receives non-best-paths. There are also monitoring protocols such as BMP [I-D.ietf-grow-bmp] that similarly receives all paths from a speaker.

When an NMS receives multiple paths for a destination, it is important for its operation to know which path is the best-path, which paths are installed in forwarding table, which path is used as a backup, etc. The NMS system may run the best-path algorithm on those paths on its own. However, its information, especially on IGP metric, local policies, etc., may be incomplete and hence its own calculations may not match that of the router's. It is also noted that even if the NMS system collected additional information to run the best-path algorithm from the point-of-view of the router, it would have to do so for every router in the network, which would impose a very high computational burden on the NMS.

When Path Type community is in use, the router provides the required information directly, thus avoiding computational load on the NMS as well as potential discrepancies between the point-of-view of the router and that of the NMS.

## 5.3. SDN applications

Similar to the monitoring applications, a "Software Defined Networking" application monitors the routing state and based on it, may change the policies on the router, or inject additional paths, to influence the forwarding. When a BGP speaker supports Path Type communities and add-path, an SDN application can simply peer with the router to receive its routing state in real-time even if the router does not provide vendor-specific APIs for doing the same.

#### 5.4. Selective Best-path

When the classical BGP advertisement rule is followed, all paths a BGP speaker considers for best-path are already installed in the forwarding table of the peer. However, when add-path, or best-external extensions are used, that no longer holds. If the BGP speakers support the Path Type communities, then the classical behavior can be reinstated by considering only those paths in the best-path algorithm that are marked as best-path or multi-path. Detailed discussions on the rules and benefits of such an approach are outside the scope of this draft.

#### 6. IANA Considerations

Section 2 defines an IPv4 Address specific transitive extended community called the Path Type extended community. IANA is requested to assign a sub-type value for the Path Type extended community. The last 2 bytes of the value field of the Path Type extended community contains a bitfield that encodes the type of the advertised path. IANA is expected to maintain a registry for these bits. Section 2 defines 6 of those bits. The rest of the bits are to be assigned by IANA using the "IETF Consensus" policy defined in [RFC2434].

#### 7. Security Considerations

This document introduces no new security concerns to BGP or other specifications referenced in this document.

#### 8. Contributors

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

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

#### 9. Acknowledgments

We would like to thank Bruno Decraene for his feedback on this work.

## 10. References

### 10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2434] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 2434, October 1998.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

### 10.2. Informative 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.
- [I-D.ietf-idr-add-paths] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-08 (work in progress), December 2012.
- [I-D.ietf-idr-best-external] Marques, P., Fernando, R., Chen, E., Mohapatra, P., and H. Gredler, "Advertisement of the best external route in BGP", draft-ietf-idr-best-external-05 (work in progress), January 2012.
- [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.

## Authors' Addresses

Camilo Cardona  
IMDEA Networks  
Avenida del Mar Mediterraneo  
Leganes 28919  
Spain

Email: [juancamilo.cardona@imdea.org](mailto:juancamilo.cardona@imdea.org)

Pierre Francois  
IMDEA Networks  
Avenida del Mar Mediterraneo  
Leganes 28919  
Spain

Email: [pierre.francois@imdea.org](mailto:pierre.francois@imdea.org)

Saikat Ray  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

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

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

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

Paolo Lucente  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

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

Pradosh Mohapatra  
Cumulus Networks  
140 C. Whisman Rd.  
Mountain View, CA 94041  
USA

Email: [pmohapat@cumulusnetworks.com](mailto:pmohapat@cumulusnetworks.com)

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 16, 2014

J. Dong  
M. Chen  
Huawei Technologies  
H. Gredler  
Juniper Networks, Inc.  
S. Previdi  
Cisco Systems, Inc.  
July 15, 2013

Distribution of MPLS Traffic Engineering (TE) LSP State using BGP  
draft-dong-idr-te-lsp-distribution-03

Abstract

This document describes a mechanism to collect the Traffic Engineering (TE) LSP information using BGP. Such information can be used by external components for path reoptimization, service placement and network visualization.

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 16, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1. Introduction . . . . .	2
2. Carrying LSP State Information in BGP . . . . .	4
2.1. LSP Identifier Information . . . . .	4
2.2. LSP State Information . . . . .	5
3. IANA Considerations . . . . .	6
4. Security Considerations . . . . .	7
5. References . . . . .	7
5.1. Normative References . . . . .	7
5.2. Informative References . . . . .	7
Authors' Addresses . . . . .	8

## 1. Introduction

In some network environments, the states of established Multi-Protocol Label Switching (MPLS) Traffic Engineering (TE) Label Switched Paths (LSPs) in the network are required by some components external to the network domain. Usually this information is directly maintained by the ingress Label Edge Routers (LERs) of the MPLS TE LSPs.

One example of using the LSP information is stateful Path Computation Element (PCE) [I-D.ietf-pce-stateful-pce], which could provide benefits in path reoptimization. While some extensions are proposed in Path Computation Element Communication Protocol (PCEP) for the Path Computation Clients (PCCs) to report the LSP states to the PCE, this mechanism may not be applicable in a management-based PCE architecture as specified in section 5.5 of [RFC4655]. As illustrated in the figure below, the PCC is not an LSR in the routing domain, thus the head-end nodes of the TE-LSP may not implement the PCEP protocol. In this case some general mechanism to collect the TE-LSP states from the ingress LERs is needed. This document

proposes an LSP state collection mechanism complementary to the mechanism defined in [I-D.ietf-pce-stateful-pce].

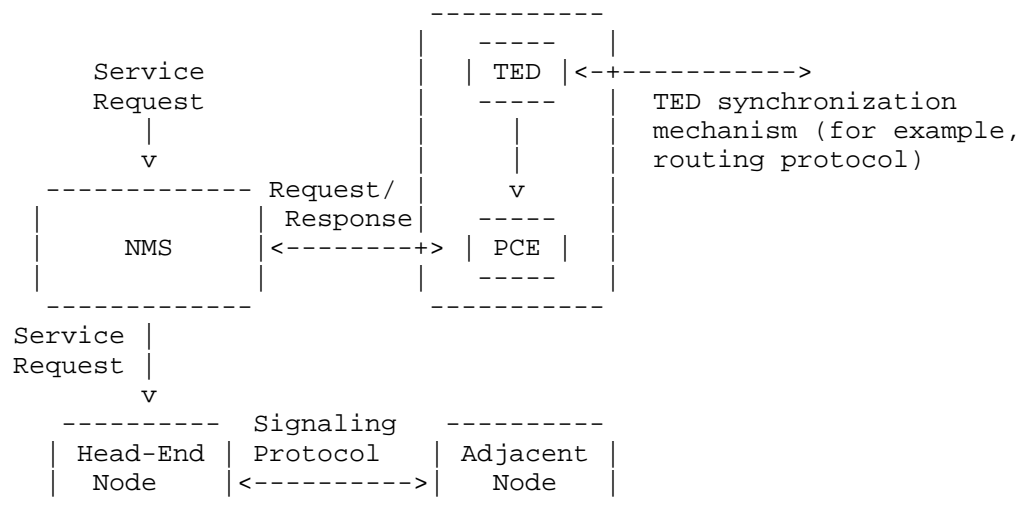


Figure 1. Management-Based PCE Usage

In networks with composite PCE nodes as specified in section 5.1 of [RFC4655], the PCE is implemented on several routers in the network, and the PCCs in the network can use the mechanism described in [I-D.ietf-pce-stateful-pce] to report the LSP information to the PCE nodes. An external component may further need to collect the LSP information from all the PCEs in the network to get a global view of the LSP states in the network.

In some networks, a centralized controller is used for service placement. Obtaining the TE LSP state information is quite important for making appropriate service placement decisions with the purpose of both meeting the application's requirements and utilizing the network resource efficiently.

The Network Management System (NMS) may need to provide global visibility of the TE LSPs in the network as part of the network visualization function.

BGP has been extended to distribute link-state and traffic engineering information and share with some external components [I-D.ietf-idr-ls-distribution]. Using the same protocol to collect other network layer information would be desired by the external components, which avoids introducing multiple protocols for network



information collection. This document describes a mechanism to distribute the TE LSP information to external components using BGP.

## 2. Carrying LSP State Information in BGP

### 2.1. LSP Identifier Information

The TE LSP Identifier 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 TE LSP Identifier information. BGP speakers that wish to exchange TE LSP information MUST use the BGP Multiprotocol Extensions Capability Code (1) to advertise the corresponding (AFI, SAFI) pair, as specified in [RFC4760].

The format of "Link State NLRI" is defined in [I-D.ietf-idr-ls-distribution]. Two new "NLRI Type" are defined for TE LSP Identifier Information as following:

- o NLRI Type = 5: IPv4 TE LSP NLRI
- o NLRI-Type = 6: IPv6 TE LSP NLRI

The IPv4 TE LSP NLRI (NLRI Type = 5) is shown in the following figure:

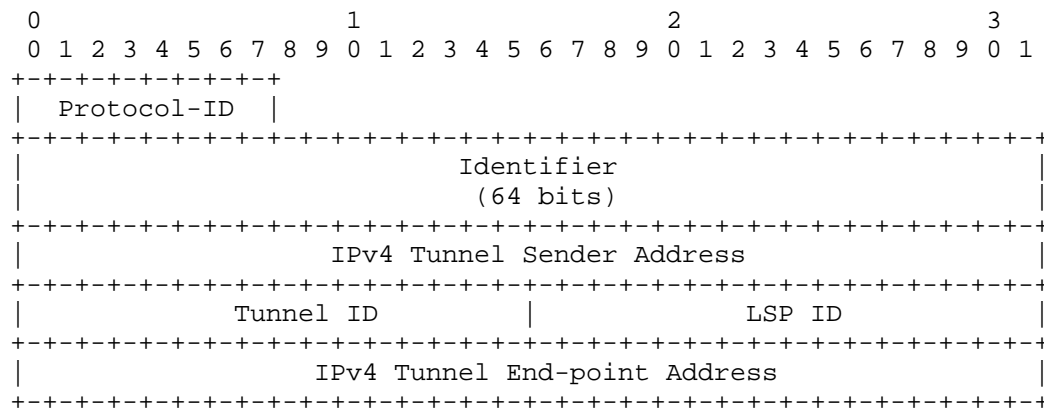


Figure 2. IPv4 TE LSP NLRI

The IPv6 TE LSP NLRI (NLRI Type = 6) is shown in the following figure:

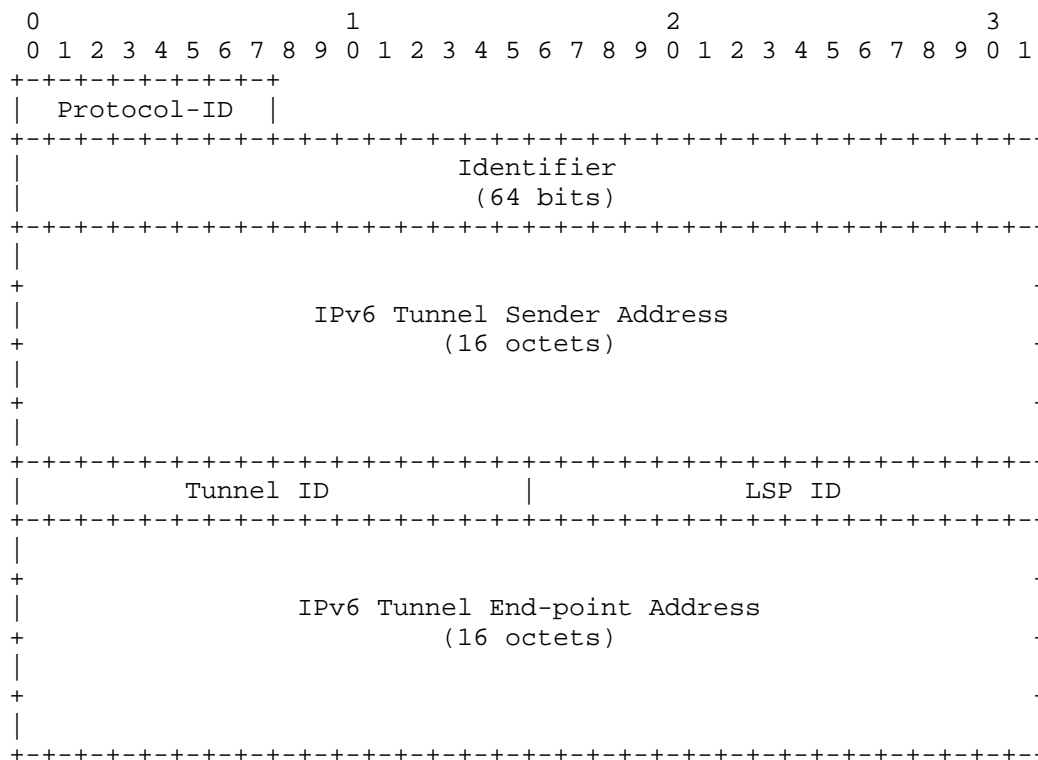


Figure 3. IPv6 TE LSP NLRI

For IPv4 TE LSP NLRI and IPv6 TE LSP NLRI, the Protocol-ID field is set to 6, which indicates that the NLRI information has been sourced by RSVP-TE.

The Identifier field is used to discriminate between instances with different LSP technology - e.g. one identifier can identify the instance for packet path, and another one is to identify the instance of optical path.

The other fields in the IPv4 TE LSP NLRI and IPv6 TE LSP NLRI are the same as specified in [RFC3209].

## 2.2. LSP State Information

The LSP State TLV is used to describe the characteristics of the TE LSPs, which is carried in the optional non-transitive BGP Attribute "LINK\_STATE Attribute" defined in [I-D.ietf-idr-ls-distribution].

The "Value" field of the LSP State TLV corresponds to the format and semantics of a set of objects defined in [RFC3209], [RFC3473] and [RFC5440] for TE LSPs. Rather than replicating all RSVP-TE related objects in this document the semantics and encodings of existing RSVP-TE objects are re-used. Hence all RSVP-TE LSP objects are regarded as sub-TLVs. The LSP State TLV SHOULD only be used with IPv4/IPv6 TE LSP NLRI.

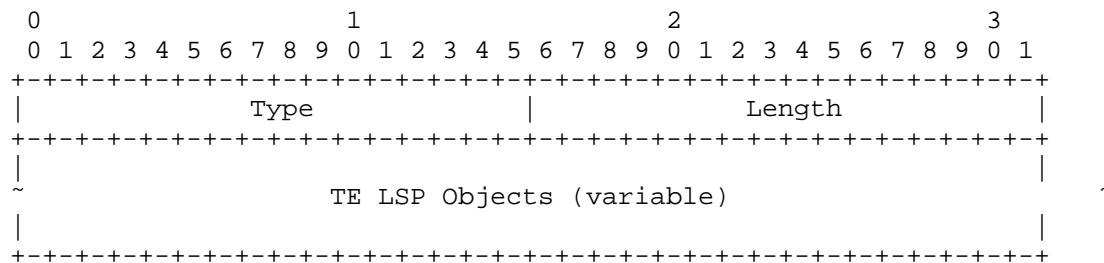


Figure 4. LSP State TLV

Currently the TE LSP Objects that can be carried in the LSP State TLV include:

- o LSP Attributes (LSPA) Object [RFC5440]
- o Explicit Route Object (ERO) [RFC3209]
- o Record Route Object (RRO) [RFC3209]
- o BANDWIDTH Object [RFC5440]
- o METRIC Object [RFC5440]
- o Protection Object [RFC3473]
- o Admin\_Status Object [RFC3473]

Other TE LSP objects may also be carried in LSP state TLV, which is for further study.

### 3. IANA Considerations

IANA needs to assign one new TLV type for "LSP State TLV" from the TLV registry of Link\_State Attribute.

IANA needs to assign one Protocol-ID for 'RSVP-TE' from the BGP-TE/LS registry of Protocol-IDs.

#### 4. Security Considerations

TBD

#### 5. References

##### 5.1. 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-03 (work in progress), May 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [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.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

##### 5.2. Informative References

- [I-D.ietf-pce-stateful-pce]  
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-05 (work in progress), July 2013.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.

Authors' Addresses

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

Email: jie.dong@huawei.com

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

Email: mach.chen@huawei.com

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

Email: hannes@juniper.net

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

Email: sprevidi@cisco.com

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

J. Dong  
M. Chen  
Huawei Technologies  
H. Gredler  
Juniper Networks, Inc.  
S. Previdi  
Cisco Systems, Inc.  
October 21, 2013

Distribution of MPLS Traffic Engineering (TE) LSP State using BGP  
draft-dong-idr-te-lsp-distribution-04

Abstract

This document describes a mechanism to collect the Traffic Engineering (TE) LSP information using BGP. Such information can be used by external components for path reoptimization, service placement and network visualization.

Requirements Language

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

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on April 24, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1. Introduction . . . . .	2
2. Carrying LSP State Information in BGP . . . . .	4
2.1. LSP Identifier Information . . . . .	4
2.2. LSP State Information . . . . .	5
3. IANA Considerations . . . . .	6
4. Security Considerations . . . . .	7
5. References . . . . .	7
5.1. Normative References . . . . .	7
5.2. Informative References . . . . .	7
Authors' Addresses . . . . .	8

## 1. Introduction

In some network environments, the states of established Multi-Protocol Label Switching (MPLS) Traffic Engineering (TE) Label Switched Paths (LSPs) in the network are required by some components external to the network domain. Usually this information is directly maintained by the ingress Label Edge Routers (LERs) of the MPLS TE LSPs.

One example of using the LSP information is stateful Path Computation Element (PCE) [I-D.ietf-pce-stateful-pce], which could provide benefits in path reoptimization. While some extensions are proposed in Path Computation Element Communication Protocol (PCEP) for the Path Computation Clients (PCCs) to report the LSP states to the PCE, this mechanism may not be applicable in a management-based PCE architecture as specified in section 5.5 of [RFC4655]. As illustrated in the figure below, the PCC is not an LSR in the routing domain, thus the head-end nodes of the TE-LSP may not implement the PCEP protocol. In this case some general mechanism to collect the TE-LSP states from the ingress LERs is needed. This document

proposes an LSP state collection mechanism complementary to the mechanism defined in [I-D.ietf-pce-stateful-pce].

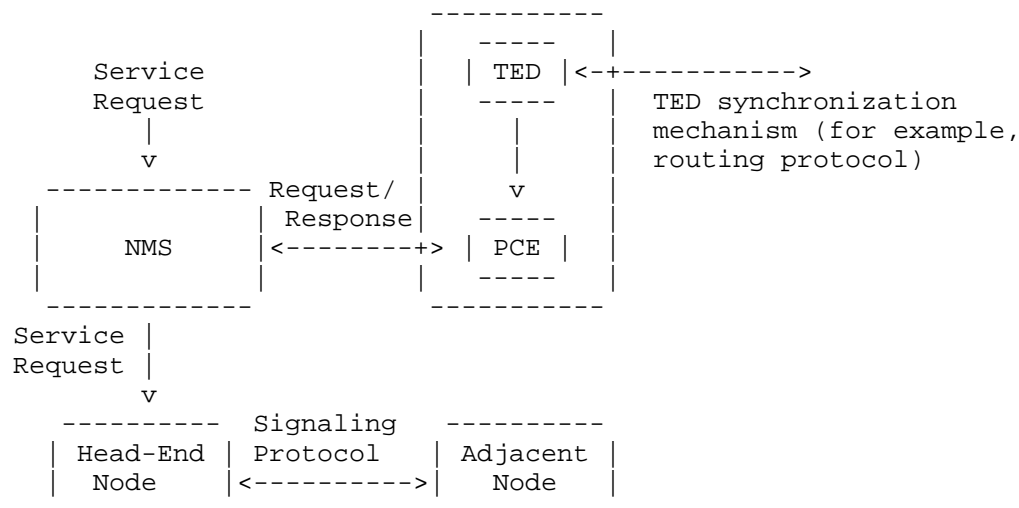


Figure 1. Management-Based PCE Usage

In networks with composite PCE nodes as specified in section 5.1 of [RFC4655], the PCE is implemented on several routers in the network, and the PCCs in the network can use the mechanism described in [I-D.ietf-pce-stateful-pce] to report the LSP information to the PCE nodes. An external component may further need to collect the LSP information from all the PCEs in the network to get a global view of the LSP states in the network.

In some networks, a centralized controller is used for service placement. Obtaining the TE LSP state information is quite important for making appropriate service placement decisions with the purpose of both meeting the application's requirements and utilizing the network resource efficiently.

The Network Management System (NMS) may need to provide global visibility of the TE LSPs in the network as part of the network visualization function.

BGP has been extended to distribute link-state and traffic engineering information and share with some external components [I-D.ietf-idr-ls-distribution]. Using the same protocol to collect other network layer information would be desired by the external components, which avoids introducing multiple protocols for network



information collection. This document describes a mechanism to distribute the TE LSP information to external components using BGP.

## 2. Carrying LSP State Information in BGP

### 2.1. LSP Identifier Information

The TE LSP Identifier 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 TE LSP Identifier information. BGP speakers that wish to exchange TE LSP information MUST use the BGP Multiprotocol Extensions Capability Code (1) to advertise the corresponding (AFI, SAFI) pair, as specified in [RFC4760].

The format of "Link State NLRI" is defined in [I-D.ietf-idr-ls-distribution]. Two new "NLRI Type" are defined for TE LSP Identifier Information as following:

- o NLRI Type = 5: IPv4 TE LSP NLRI
- o NLRI-Type = 6: IPv6 TE LSP NLRI

The IPv4 TE LSP NLRI (NLRI Type = 5) is shown in the following figure:

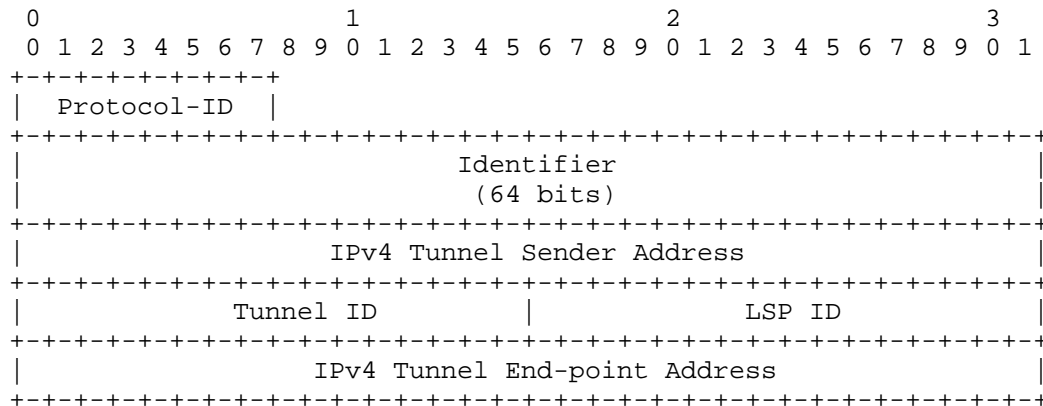


Figure 2. IPv4 TE LSP NLRI

The IPv6 TE LSP NLRI (NLRI Type = 6) is shown in the following figure:

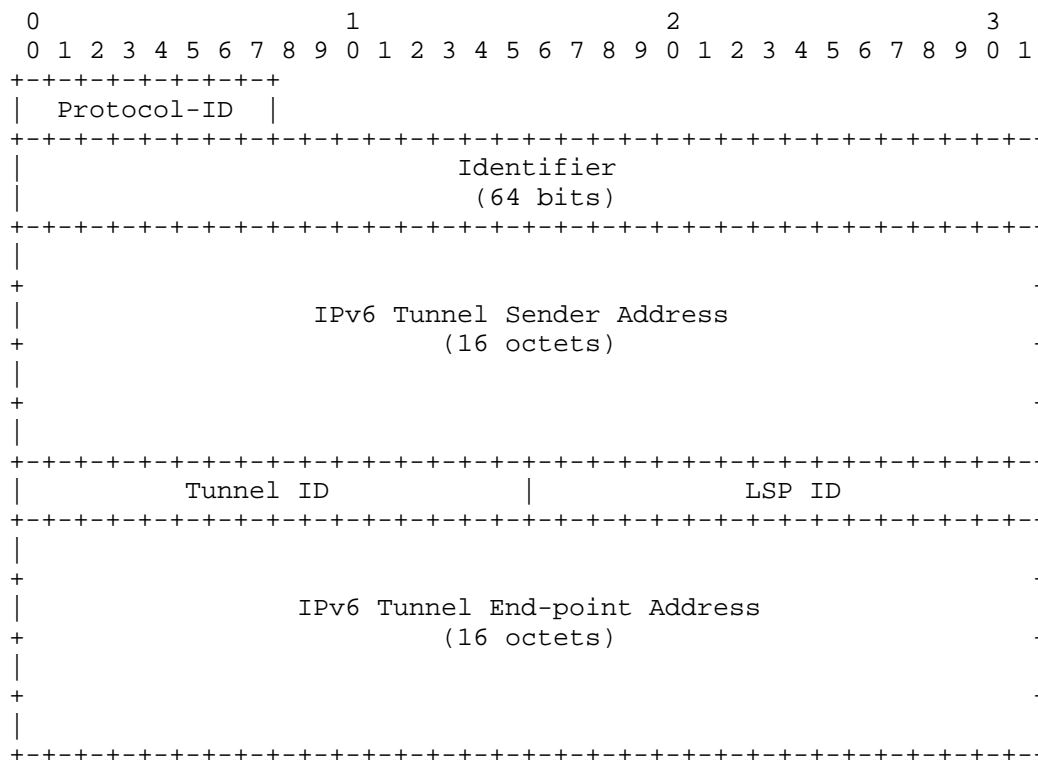


Figure 3. IPv6 TE LSP NLRI

For IPv4 TE LSP NLRI and IPv6 TE LSP NLRI, the Protocol-ID field is set to 6, which indicates that the NLRI information has been sourced by RSVP-TE.

The Identifier field is used to discriminate between instances with different LSP technology - e.g. one identifier can identify the instance for packet path, and another one is to identify the instance of optical path.

The other fields in the IPv4 TE LSP NLRI and IPv6 TE LSP NLRI are the same as specified in [RFC3209].

## 2.2. LSP State Information

The LSP State TLV is used to describe the characteristics of the TE LSPs, which is carried in the optional non-transitive BGP Attribute "LINK\_STATE Attribute" defined in [I-D.ietf-idr-ls-distribution].

The "Value" field of the LSP State TLV corresponds to the format and semantics of a set of objects defined in [RFC3209], [RFC3473] and [RFC5440] for TE LSPs. Rather than replicating all RSVP-TE related objects in this document the semantics and encodings of existing RSVP-TE objects are re-used. Hence all RSVP-TE LSP objects are regarded as sub-TLVs. The LSP State TLV SHOULD only be used with IPv4/IPv6 TE LSP NLRI.

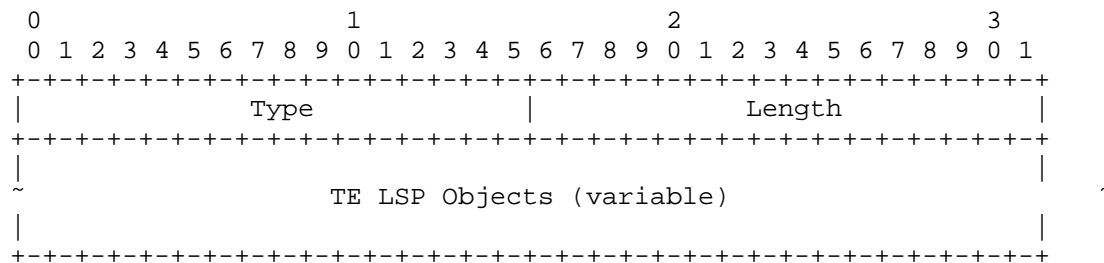


Figure 4. LSP State TLV

Currently the TE LSP Objects that can be carried in the LSP State TLV include:

- o LSP Attributes (LSPA) Object [RFC5440]
- o Explicit Route Object (ERO) [RFC3209]
- o Record Route Object (RRO) [RFC3209]
- o BANDWIDTH Object [RFC5440]
- o METRIC Object [RFC5440]
- o Protection Object [RFC3473]
- o Admin\_Status Object [RFC3473]

Other TE LSP objects may also be carried in LSP state TLV, which is for further study.

### 3. IANA Considerations

IANA needs to assign one new TLV type for "LSP State TLV" from the TLV registry of Link\_State Attribute.

IANA needs to assign one Protocol-ID for 'RSVP-TE' from the BGP-TE/LS registry of Protocol-IDs.

#### 4. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See [RFC6952] for details.

#### 5. References

##### 5.1. 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-03 (work in progress), May 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [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.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

##### 5.2. Informative References

- [I-D.ietf-pce-stateful-pce]  
Crabbe, E., Medved, J., Minei, I., and R. Varga, "PCEP Extensions for Stateful PCE", draft-ietf-pce-stateful-pce-07 (work in progress), October 2013.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [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 Building, No. 156 Beiqing Rd.  
Beijing 100095  
China

Email: jie.dong@huawei.com

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

Email: mach.chen@huawei.com

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

Email: hannes@juniper.net

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

Email: sprevidi@cisco.com

Inter-Domain Routing  
Internet-Draft  
Intended status: Standards Track  
Expires: November 22, 2013

H. Gredler  
Juniper Networks, Inc.  
J. Medved  
S. Previdi  
Cisco Systems, Inc.  
A. Farrel  
Juniper Networks, Inc.  
S. Ray  
Cisco Systems, Inc.  
May 21, 2013

North-Bound Distribution of Link-State and TE Information using BGP  
draft-ietf-idr-ls-distribution-03

Abstract

In a number of environments, a component external to a network is called upon to perform computations based on the network topology and current state of the connections within the network, including traffic engineering information. This is information typically distributed by IGP routing protocols within the network

This document describes a mechanism by which links state and traffic engineering information can be collected from networks and shared with external components using the BGP routing protocol. This is achieved using a new BGP Network Layer Reachability Information (NLRI) encoding format. The mechanism is applicable to physical and virtual IGP links. The mechanism described is subject to policy control.

Applications of this technique include Application Layer Traffic Optimization (ALTO) servers, and Path Computation Elements (PCEs).

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 22, 2013.

#### Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1. Introduction . . . . .	5
2. Motivation and Applicability . . . . .	6
2.1. MPLS-TE with PCE . . . . .	6
2.2. ALTO Server Network API . . . . .	8
3. Carrying Link State Information in BGP . . . . .	9
3.1. TLV Format . . . . .	9
3.2. The Link State NLRI . . . . .	10
3.2.1. Node Descriptors . . . . .	13
3.2.2. Link Descriptors . . . . .	17
3.2.3. Prefix Descriptors . . . . .	18
3.3. The LINK_STATE Attribute . . . . .	20
3.3.1. Node Attribute TLVs . . . . .	20
3.3.2. Link Attribute TLVs . . . . .	23
3.3.3. Prefix Attribute TLVs . . . . .	27
3.4. BGP Next Hop Information . . . . .	30
3.5. Inter-AS Links . . . . .	31
3.6. Router-ID Anchoring Example: ISO Pseudonode . . . . .	31
3.7. Router-ID Anchoring Example: OSPFv2 to IS-IS Migration . . . . .	32
4. Link to Path Aggregation . . . . .	32
4.1. Example: No Link Aggregation . . . . .	33
4.2. Example: ASBR to ASBR Path Aggregation . . . . .	33
4.3. Example: Multi-AS Path Aggregation . . . . .	34
5. IANA Considerations . . . . .	34
6. Manageability Considerations . . . . .	34
6.1. Operational Considerations . . . . .	35
6.1.1. Operations . . . . .	35
6.1.2. Installation and Initial Setup . . . . .	35
6.1.3. Migration Path . . . . .	35
6.1.4. Requirements on Other Protocols and Functional Components . . . . .	35
6.1.5. Impact on Network Operation . . . . .	35
6.1.6. Verifying Correct Operation . . . . .	36
6.2. Management Considerations . . . . .	36
6.2.1. Management Information . . . . .	36
6.2.2. Fault Management . . . . .	36
6.2.3. Configuration Management . . . . .	36
6.2.4. Accounting Management . . . . .	36
6.2.5. Performance Management . . . . .	36
6.2.6. Security Management . . . . .	37
7. TLV/Sub-TLV Code Points Summary . . . . .	37
8. Security Considerations . . . . .	39
9. Contributors . . . . .	39
10. Acknowledgements . . . . .	39
11. References . . . . .	40
11.1. Normative References . . . . .	40
11.2. Informative References . . . . .	41



Authors' Addresses . . . . .	42
------------------------------	----

## 1. Introduction

The contents of a Link State Database (LSDB) or a Traffic Engineering Database (TED) has the scope of an IGP area. Some applications, such as end-to-end Traffic Engineering (TE), would benefit from visibility outside one area or Autonomous System (AS) in order to make better decisions.

The IETF has defined the Path Computation Element (PCE) [RFC4655] as a mechanism for achieving the computation of end-to-end TE paths that cross the visibility of more than one TED or which require CPU-intensive or coordinated computations. The IETF has also defined the ALTO Server [RFC5693] as an entity that generates an abstracted network topology and provides it to network-aware applications.

Both a PCE and an ALTO Server need to gather information about the topologies and capabilities of the network in order to be able to fulfill their function.

This document describes a mechanism by which Link State and TE information can be collected from networks and shared with external components using the BGP routing protocol [RFC4271]. This is achieved using a new BGP Network Layer Reachability Information (NLRI) encoding format. The mechanism is applicable to physical and virtual links. The mechanism described is subject to policy control.

A router maintains one or more databases for storing link-state information about nodes and links in any given area. Link attributes stored in these databases include: local/remote IP addresses, local/remote interface identifiers, link metric and TE metric, link bandwidth, reservable bandwidth, per CoS class reservation state, preemption and Shared Risk Link Groups (SRLG). The router's BGP process can retrieve topology from these LSDBs and distribute it to a consumer, either directly or via a peer BGP Speaker (typically a dedicated Route Reflector), using the encoding specified in this document.

The collection of Link State and TE link state information and its distribution to consumers is shown in the following figure.

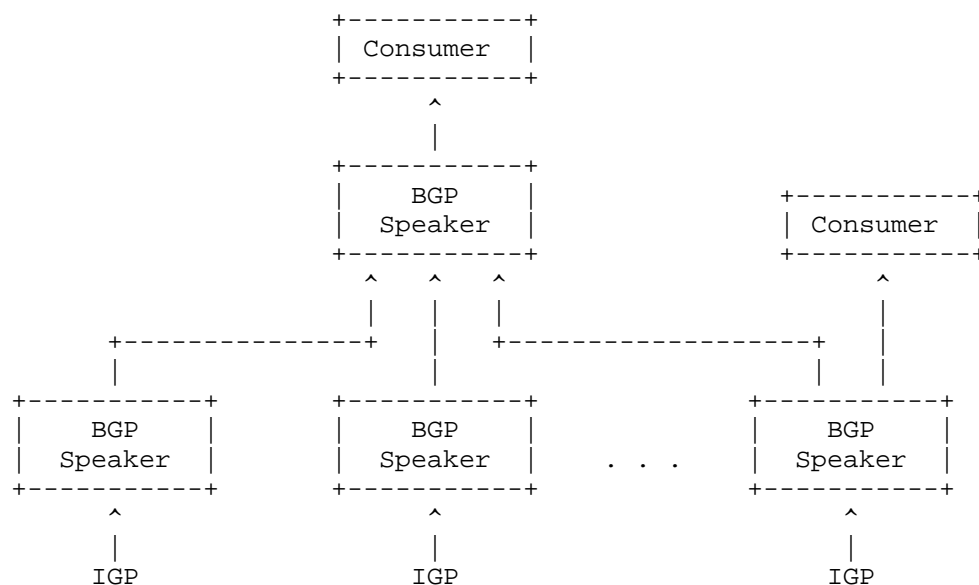


Figure 1: TE Link State info collection

A BGP Speaker may apply configurable policy to the information that it distributes. Thus, it may distribute the real physical topology from the LSDB or the TED. Alternatively, it may create an abstracted topology, where virtual, aggregated nodes are connected by virtual paths. Aggregated nodes can be created, for example, out of multiple routers in a POP. Abstracted topology can also be a mix of physical and virtual nodes and physical and virtual links. Furthermore, the BGP Speaker can apply policy to determine when information is updated to the consumer so that there is reduction of information flow from the network to the consumers. Mechanisms through which topologies can be aggregated or virtualized are outside the scope of this document

## 2. Motivation and Applicability

This section describes use cases from which the requirements can be derived.

### 2.1. MPLS-TE with PCE

As described in [RFC4655] a PCE can be used to compute MPLS-TE paths within a "domain" (such as an IGP area) or across multiple domains (such as a multi-area AS, or multiple ASes).

- o Within a single area, the PCE offers enhanced computational power that may not be available on individual routers, sophisticated policy control and algorithms, and coordination of computation across the whole area.
- o If a router wants to compute a MPLS-TE path across IGP areas its own TED lacks visibility of the complete topology. That means that the router cannot determine the end-to-end path, and cannot even select the right exit router (Area Border Router - ABR) for an optimal path. This is an issue for large-scale networks that need to segment their core networks into distinct areas, but which still want to take advantage of MPLS-TE.

Previous solutions used per-domain path computation [RFC5152]. The source router could only compute the path for the first area because the router only has full topological visibility for the first area along the path, but not for subsequent areas. Per-domain path computation uses a technique called "loose-hop-expansion" [RFC3209], and selects the exit ABR and other ABRs or AS Border Routers (ASBRs) using the IGP computed shortest path topology for the remainder of the path. This may lead to sub-optimal paths, makes alternate/back-up path computation hard, and might result in no TE path being found when one really does exist.

The PCE presents a computation server that may have visibility into more than one IGP area or AS, or may cooperate with other PCEs to perform distributed path computation. The PCE obviously needs access to the TED for the area(s) it serves, but [RFC4655] does not describe how this is achieved. Many implementations make the PCE a passive participant in the IGP so that it can learn the latest state of the network, but this may be sub-optimal when the network is subject to a high degree of churn, or when the PCE is responsible for multiple areas.

The following figure shows how a PCE can get its TED information using the mechanism described in this document.

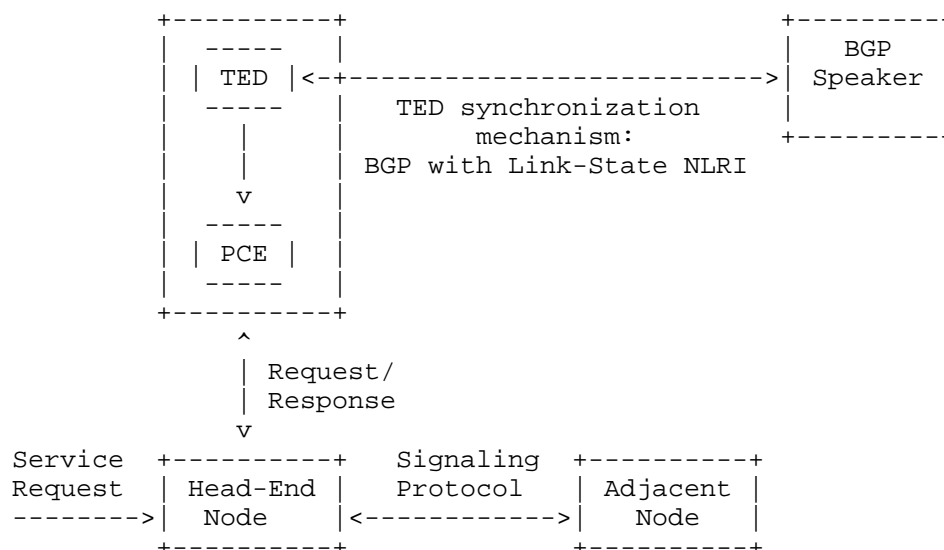


Figure 2: External PCE node using a TED synchronization mechanism

The mechanism in this document allows the necessary TED information to be collected from the IGP within the network, filtered according to configurable policy, and distributed to the PCE as necessary.

## 2.2. ALTO Server Network API

An ALTO Server [RFC5693] is an entity that generates an abstracted network topology and provides it to network-aware applications over a web service based API. Example applications are p2p clients or trackers, or CDNs. The abstracted network topology comes in the form of two maps: a Network Map that specifies allocation of prefixes to Partition Identifiers (PIDs), and a Cost Map that specifies the cost between PIDs listed in the Network Map. For more details, see [I-D.ietf-alto-protocol].

ALTO abstract network topologies can be auto-generated from the physical topology of the underlying network. The generation would typically be based on policies and rules set by the operator. Both prefix and TE data are required: prefix data is required to generate ALTO Network Maps, TE (topology) data is required to generate ALTO Cost Maps. Prefix data is carried and originated in BGP, TE data is originated and carried in an IGP. The mechanism defined in this document provides a single interface through which an ALTO Server can retrieve all the necessary prefix and network topology data from the underlying network. Note an ALTO Server can use other mechanisms to get network data, for example, peering with multiple IGP and BGP

Speakers.

The following figure shows how an ALTO Server can get network topology information from the underlying network using the mechanism described in this document.

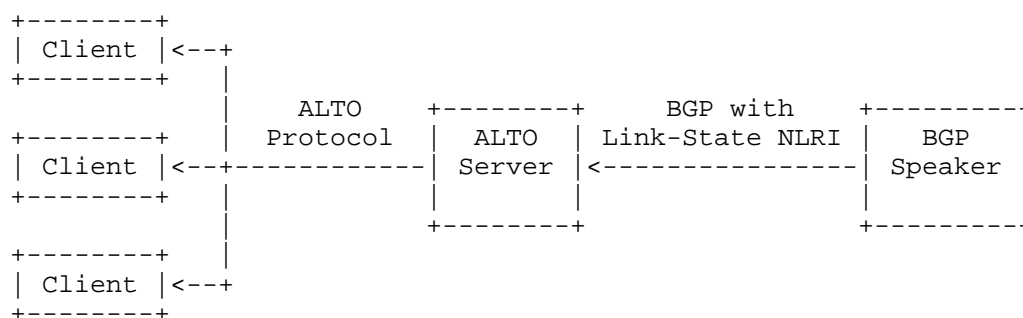


Figure 3: ALTO Server using network topology information

### 3. Carrying Link State Information in BGP

This specification contains two parts: definition of a new BGP NLRI that describes links, nodes and prefixes comprising IGP link state information, and definition of a new BGP path attribute (BGP-LS attribute) that carries link, node and prefix properties and attributes, such as the link and prefix metric or auxiliary Router-IDs of nodes, etc.

#### 3.1. TLV Format

Information in the new link state NLRIs and attributes is encoded in Type/Length/Value triplets. The TLV format is shown in Figure 4.

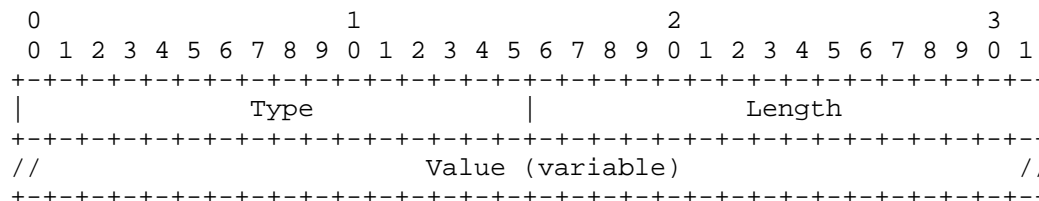


Figure 4: TLV format

The Length field defines the length of the value portion in octets (thus a TLV with no value portion would have a length of zero). The TLV is not padded to four-octet alignment. Unrecognized types are

preserved and propagated. In order to compare NLRI's with unknown TLVs all TLVs MUST be ordered in ascending order. If there are more TLVs of the same type, then the TLVs MUST be ordered in ascending order of the TLV value within the set of TLVs with the same type. All TLVs that are not specified as mandatory are considered optional.

### 3.2. The Link State NLRI

The MP\_REACH and MP\_UNREACH attributes are BGP's containers for carrying opaque information. Each Link State NLRI describes either a node, a link or a prefix.

All non-VPN link, node and prefix information SHALL be encoded using AFI 16388 / SAFI 71. VPN link, node and prefix information SHALL be encoded using AFI 16388 / SAFI 128.

In order for two BGP speakers to exchange Link-State NLRI, they MUST use BGP Capabilities Advertisement to ensure that they both are capable of properly processing such NLRI. This is done as specified in [RFC4760], by using capability code 1 (multi-protocol BGP), with an AFI 16388 / SAFI 71 and AFI 16388 / SAFI 128 for the VPN flavor.

The format of the Link State NLRI is shown in the following figure.

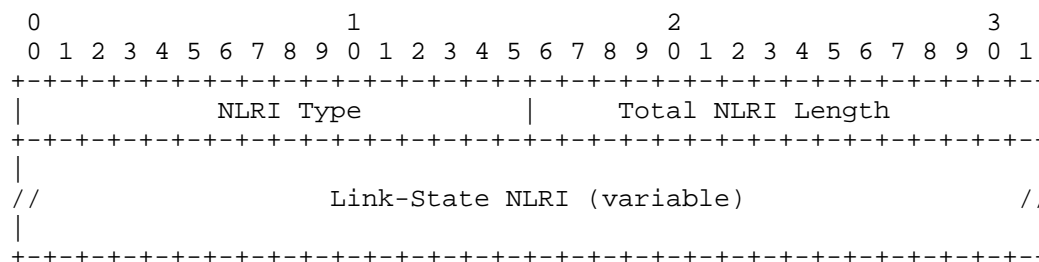


Figure 5: Link State AFI 16388 / SAFI 71 NLRI Format

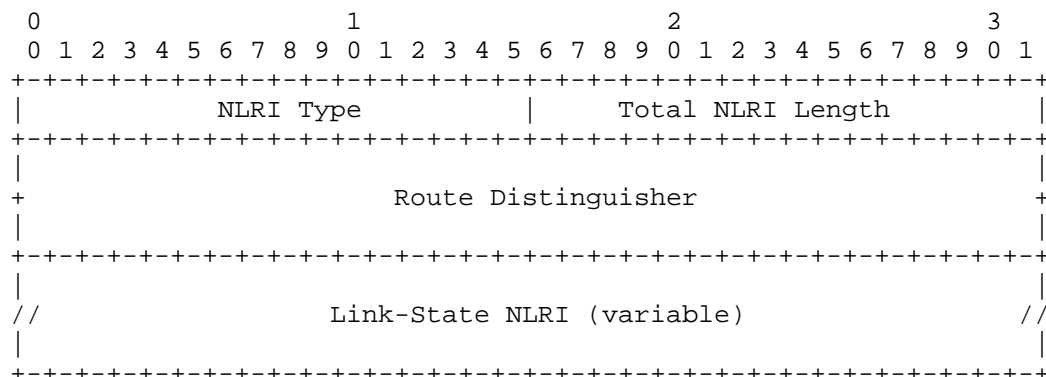


Figure 6: Link State VPN AFI 16388 / SAFI 128 NLRI Format

The 'Total NLRI Length' field contains the cumulative length, in octets, of rest of the NLRI not including the NLRI Type field or itself. For VPN applications it also includes the length of the Route Distinguisher.

The 'NLRI Type' field can contain one of the following values:

Type = 1: Node NLRI

Type = 2: Link NLRI

Type = 3: IPv4 Topology Prefix NLRI

Type = 4: IPv6 Topology Prefix NLRI

The Node NLRI (NLRI Type = 1) is shown in the following figure.

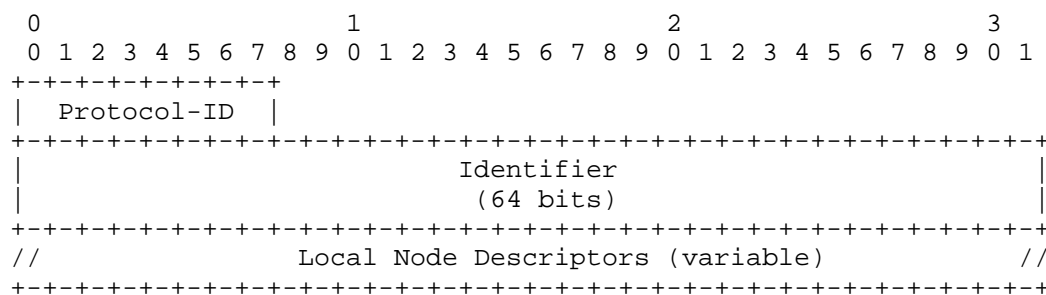


Figure 7: The Node NLRI format

The Link NLRI (NLRI Type = 2) is shown in the following figure.



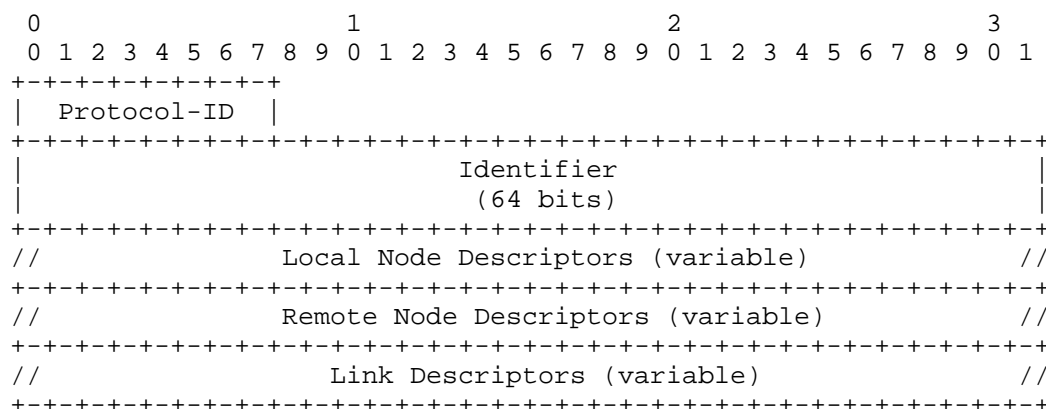


Figure 8: The Link NLRI format

The IPv4 and IPv6 Prefix NLRIs (NLRI Type = 3 and Type = 4) use the same format as shown in the following figure.

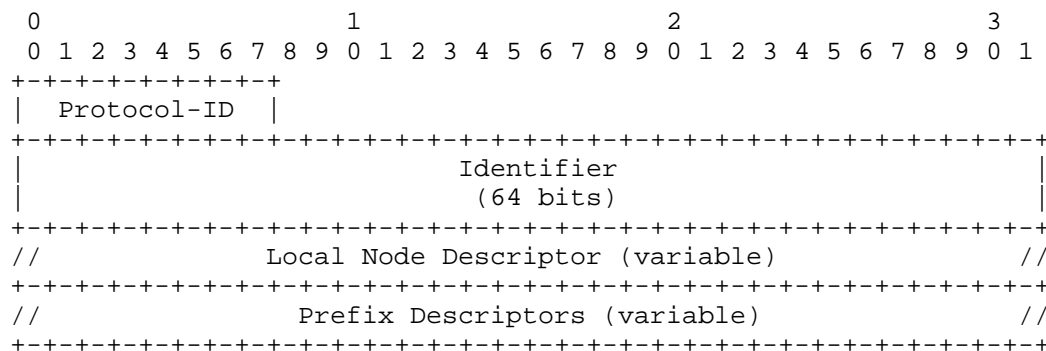


Figure 9: The IPv4/IPv6 Topology Prefix NLRI format

The 'Protocol-ID' field can contain one of the following values:

Protocol-ID = 0: Unknown, The source of NLRI information could not be determined

Protocol-ID = 1: IS-IS Level 1, The NLRI information has been sourced by IS-IS Level 1

Protocol-ID = 2: IS-IS Level 2, The NLRI information has been sourced by IS-IS Level 2

Protocol-ID = 3: OSPF, The NLRI information has been sourced by OSPF

Protocol-ID = 4: Direct, The NLRI information has been sourced from local interface state

Protocol-ID = 5: Static, The NLRI information has been sourced by static configuration

Both OSPF and IS-IS may run multiple routing protocol instances over the same link. See [RFC6822] and [RFC6549]. These instances define independent "routing universes". The 64-Bit 'Identifier' field is used to identify the "routing universe" where the NLRI belongs. The NLRIs representing IGP objects (nodes, links or prefixes) from the same routing universe MUST have the same 'Identifier' value; NLRIs with different 'Identifier' values MUST be considered to be from different routing universes. Table 1 lists the 'Identifier' values that are defined as well-known in this draft.

Identifier	Routing Universe
0	L3 packet topology
1	L1 optical topology

Table 1: Well-known Instance Identifiers

Each Node Descriptor and Link Descriptor consists of one or more TLVs described in the following sections.

### 3.2.1. Node Descriptors

Each link is anchored by a pair of Router-IDs that are used by the underlying IGP, namely, 48 Bit ISO System-ID for IS-IS and 32 bit Router-ID for OSPFv2 and OSPFv3. An IGP may use one or more additional auxiliary Router-IDs, mainly for traffic engineering purposes. For example, IS-IS may have one or more IPv4 and IPv6 TE Router-IDs [RFC5305], [RFC6119]. These auxiliary Router-IDs MUST be included in the link attribute described in Section 3.3.2.

It is desirable that the Router-ID assignments inside the Node Descriptor are globally unique. However there may be Router-ID spaces (e.g. ISO) where no global registry exists, or worse, Router-IDs have been allocated following private-IP RFC 1918 [RFC1918] allocation. We use Autonomous System (AS) Number and BGP-LS Identifier in order to disambiguate the Router-IDs, as described in Section 3.2.1.1.

## 3.2.1.1. Globally Unique Node/Link/Prefix Identifiers

One problem that needs to be addressed is the ability to identify an IGP node globally (by "global", we mean within the BGP-LS database collected by all BGP-LS speakers that talk to each other). This can be expressed through the following two requirements:

(A) The same node must not be represented by two keys (otherwise one node will look like two nodes).

(B) Two different nodes must not be represented by the same key (otherwise, two nodes will look like one node).

We define an "IGP domain" to be the set of nodes (hence, by extension links and prefixes), within which, each node has a unique IGP representation by using the combination of Area-ID, Router-ID, Protocol, Topology-ID, and Instance ID. The problem is that BGP may receive node/link/prefix information from multiple independent "IGP domains" and we need to distinguish between them. Moreover, we can't assume there is always one and only one IGP domain per AS. During IGP transitions it may happen that two redundant IGPs are in place.

In section Section 3.2.1.4 a set of sub-TLVs is described, which allows to specify a flexible key for any given Node/Link information such that global uniqueness of the NLRI is ensured.

## 3.2.1.2. Local Node Descriptors

The Local Node Descriptors TLV contains Node Descriptors for the node anchoring the local end of the link. This is a mandatory TLV in all three types of NLRIs. The length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 3.2.1.4.

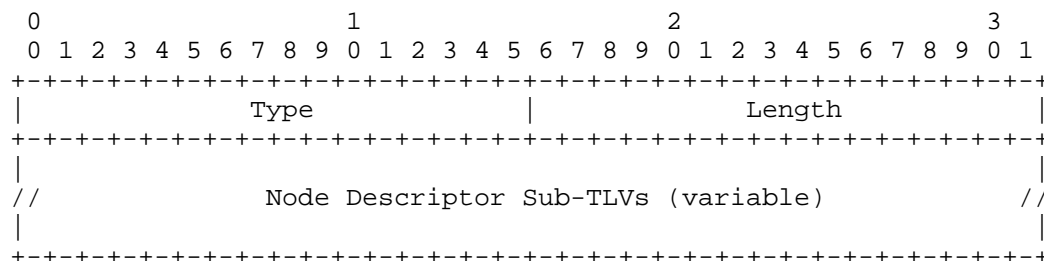


Figure 10: Local Node Descriptors TLV format

### 3.2.1.3. Remote Node Descriptors

The Remote Node Descriptors contains Node Descriptors for the node anchoring the remote end of the link. This is a mandatory TLV for link NLRIs. The length of this TLV is variable. The value contains one or more Node Descriptor Sub-TLVs defined in Section 3.2.1.4.

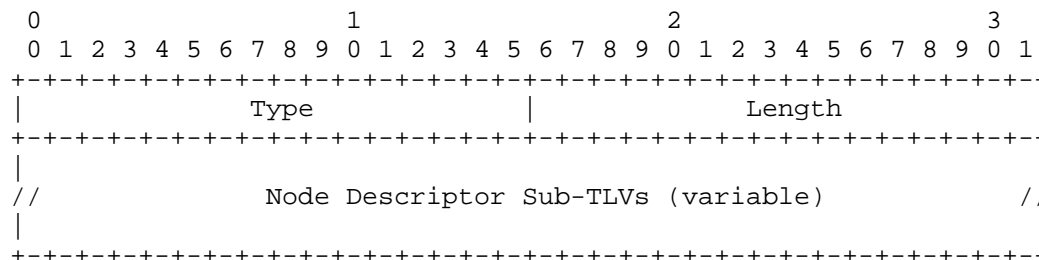


Figure 11: Remote Node Descriptors TLV format

#### 3.2.1.4. Node Descriptor Sub-TLVs

The Node Descriptor Sub-TLV type codepoints and lengths are listed in the following table:

Sub-TLV Code Point	Description	Length
512	Autonomous System	4
513	BGP-LS Identifier	4
514	Area-ID	4
515	IGP Router-ID	Variable

Table 2: Node Descriptor Sub-TLVs

The sub-TLV values in Node Descriptor TLVs are defined as follows:

Autonomous System: opaque value (32 Bit AS Number)

BGP-LS Identifier: opaque value (32 Bit ID). In conjunction with ASN, uniquely identifies the BGP-LS domain. The combination of ASN and BGP-LS ID MUST be globally unique. All BGP-LS speakers within an IGP flooding-set (set of IGP nodes within which an LSP/LSA is flooded) MUST use the same ASN, BGP-LS ID tuple. If an IGP domain consists of multiple flooding-sets, then all BGP-LS speakers within the IGP domain SHOULD use the same ASN, BGP-LS ID tuple. The ASN, BGP Router-ID tuple (which is globally unique [RFC6286] ) of one of the BGP-LS speakers within the flooding-set

(or IGP domain) may be used for all BGP-LS speakers in that flooding-set (or IGP domain).

Area ID: It is used to identify the 32 Bit area to which the NLRI belongs. Area Identifier allows the different NLRIs of the same router to be discriminated.

IGP Router ID: opaque value. This is a mandatory TLV. For an IS-IS non-Pseudonode, this contains 6 octet ISO node-ID (ISO system-ID). For an IS-IS Pseudonode corresponding to a LAN, this contains 6 octet ISO node-ID of the "Designated Intermediate System" (DIS) followed by one octet nonzero PSN identifier (7 octet in total). For an OSPFv2 or OSPFv3 non-"Pseudonode", this contains 4 octet Router-ID. For an OSPFv2 "Pseudonode" representing a LAN, this contains 4 octet Router-ID of the designated router (DR) followed by 4 octet IPv4 address of the DR's interface to the LAN (8 octet in total). Similarly, for an OSPFv3 "Pseudonode", this contains 4 octet Router-ID of the DR followed by 4 octet interface identifier of the DR's interface to the LAN (8 octet in total). The TLV size in combination with protocol identifier enables the decoder to determine the type of the node.

There can be at most one instance of each sub-TLV type present in any Node Descriptor. The TLV ordering within a Node descriptor MUST be kept in order of increasing numeric value of type. This needs to be done in order to compare NLRIs, even when an implementation encounters an unknown sub-TLV. Using stable sorting an implementation can do binary comparison of NLRIs and hence allow incremental deployment of new key sub-TLVs.

#### 3.2.1.5. Multi-Topology ID

The Multi-Topology ID (MT-ID) TLV carries one or more IS-IS or OSPF Multi-Topology IDs for a link, node or prefix.

Semantics of the IS-IS MT-ID are defined in RFC5120, Section 7.2 [RFC5120]. Semantics of the OSPF MT-ID are defined in RFC4915, Section 3.7 [RFC4915]. If the value in the MT-ID TLV is derived from OSPF, then the upper 9 bits MUST be set to 0. Bits R are reserved, SHOULD be set to 0 when originated and ignored on receipt.

The format of the MT-ID TLV is shown in the following figure.

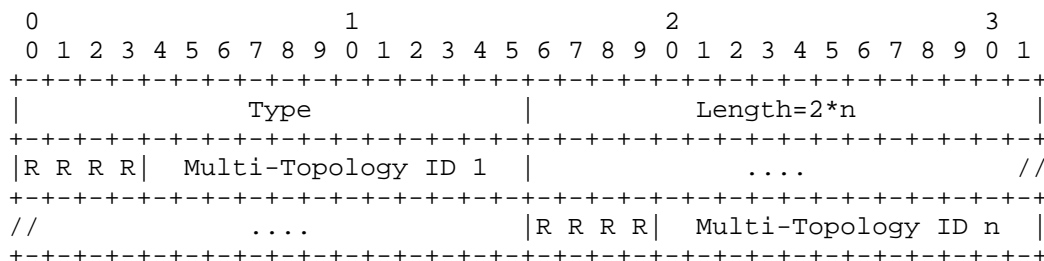


Figure 12: Multi-Topology ID TLV format

where Type is 263, Length is 2\*n and n is the number of MT-IDs carried in the TLV.

The MT-ID TLV MAY be present in a Link Descriptor, a Prefix Descriptor, or in the BGP-LS attribute of a node NLRI. In Link or Prefix Descriptor, only one MT-ID TLV containing only the MT-ID of the topology where the link or the prefix belongs is allowed. In the BGP-LS attribute of a node NLRI, one MT-ID TLV containing the array of MT-IDs of all topologies where the node belongs can be present.

### 3.2.2. Link Descriptors

The 'Link Descriptor' field is a set of Type/Length/Value (TLV) triplets. The format of each TLV is shown in Section 3.1. The 'Link descriptor' TLVs uniquely identify a link among multiple parallel links between a pair of anchor routers. A link described by the Link descriptor TLVs actually is a "half-link", a unidirectional representation of a logical link. In order to fully describe a single logical link two originating routers advertise a half-link each, i.e. two link NLRIs are advertised for a given point-to-point link.

The format and semantics of the 'value' fields in most 'Link Descriptor' TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305], [RFC5307] and [RFC6119]. Although the encodings for 'Link Descriptor' TLVs were originally defined for IS-IS, the TLVs can carry data sourced either by IS-IS or OSPF.

The following TLVs are valid as Link Descriptors in the Link NLRI:

TLV Code Point	Description	IS-IS TLV/Sub-TLV	Value defined in:
258	Link Local/Remote Identifiers	22/4	[RFC5307]/1.1
259	IPv4 interface address	22/6	[RFC5305]/3.2
260	IPv4 neighbor address	22/8	[RFC5305]/3.3
261	IPv6 interface address	22/12	[RFC6119]/4.2
262	IPv6 neighbor address	22/13	[RFC6119]/4.3
263	Multi-Topology Identifier	---	Section 3.2.1.5

Table 3: Link Descriptor TLVs

### 3.2.3. Prefix Descriptors

The 'Prefix Descriptor' field is a set of Type/Length/Value (TLV) triplets. 'Prefix Descriptor' TLVs uniquely identify an IPv4 or IPv6 Prefix originated by a Node. The following TLVs are valid as Prefix Descriptors in the IPv4/IPv6 Prefix NLRI:

TLV Code Point	Description	Length	Value defined in:
263	Multi-Topology Identifier	variable	Section 3.2.1.5
264	OSPF Route Type	1	Section 3.2.3.1
265	IP Reachability Information	variable	Section 3.2.3.2

Table 4: Prefix Descriptor TLVs

#### 3.2.3.1. OSPF Route Type

OSPF Route Type is an optional TLV that MAY be present in Prefix NLRIs. It is used to identify the OSPF route-type of the prefix. It is used when an OSPF prefix is advertised in the OSPF domain with multiple different route-types. The Route Type TLV allows to discriminate these advertisements. The format of the OSPF Route Type TLV is shown in the following figure.

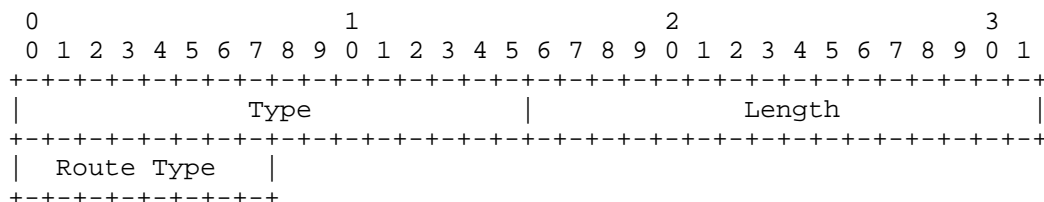


Figure 13: OSPF Route Type TLV Format

where the Type and Length fields of the TLV are defined in Table 4. The OSPF Route Type field values are defined in the OSPF protocol, and can be one of the following:

Intra-Area (0x1)

Inter-Area (0x2)

External 1 (0x3)

External 2 (0x4)

NSSA 1 (0x5)

NSSA 2 (0x6)

### 3.2.3.2. IP Reachability Information

The IP Reachability Information is a mandatory TLV that contains one IP address prefix (IPv4 or IPv6) originally advertised in the IGP topology. Its purpose is to glue a particular BGP service NLRI via virtue of its BGP next-hop to a given Node in the LSDB. A router SHOULD advertise an IP Prefix NLRI for each of its BGP Next-hops. The format of the IP Reachability Information TLV is shown in the following figure:

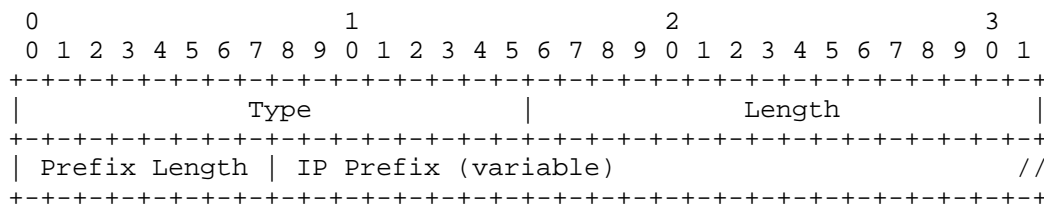


Figure 14: IP Reachability Information TLV Format

The Type and Length fields of the TLV are defined in Table 4. The following two fields determine the address-family reachability



information. The 'Prefix Length' field contains the length of the prefix in bits. The 'IP Prefix' field contains the most significant octets of the prefix; i.e., 1 octet for prefix length 1 up to 8, 2 octets for prefix length 9 to 16, 3 octets for prefix length 17 up to 24 and 4 octets for prefix length 25 up to 32, etc.

### 3.3. The LINK\_STATE Attribute

This is an optional, non-transitive BGP attribute that is used to carry link, node and prefix parameters and attributes. It is defined as a set of Type/Length/Value (TLV) triplets, described in the following section. This attribute SHOULD only be included with Link State NLRIs. This attribute MUST be ignored for all other address-families.

#### 3.3.1. Node Attribute TLVs

Node attribute TLVs are the TLVs that may be encoded in the BGP-LS attribute with a node NLRI. The following node attribute TLVs are defined:

TLV Code Point	Description	Length	Value defined in:
263	Multi-Topology Identifier	variable	Section 3.2.1.5
1024	Node Flag Bits	1	Section 3.3.1.1
1025	Opaque Node Properties	variable	Section 3.3.1.5
1026	Node Name	variable	Section 3.3.1.3
1027	IS-IS Area Identifier	variable	Section 3.3.1.2
1028	IPv4 Router-ID of Local Node	4	[RFC5305]/4.3
1029	IPv6 Router-ID of Local Node	16	[RFC6119]/4.1

Table 5: Node Attribute TLVs

##### 3.3.1.1. Node Flag Bits TLV

The Node Flag Bits TLV carries a bit mask describing node attributes. The value is a variable length bit array of flags, where each bit represents a node capability.

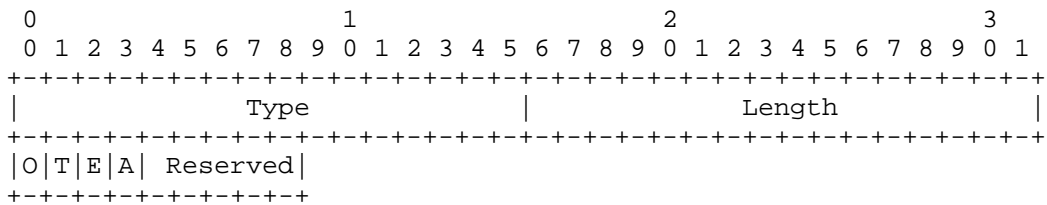


Figure 15: Node Flag Bits TLV format

The bits are defined as follows:

Bit	Description	Reference
'O'	Overload Bit	[RFC1195]
'T'	Attached Bit	[RFC1195]
'E'	External Bit	[RFC2328]
'A'	ABR Bit	[RFC2328]
Reserved	Reserved for future use	

Table 6: Node Flag Bits Definitions

3.3.1.2. IS-IS Area Identifier TLV

An IS-IS node can be part of one or more IS-IS areas. Each of these area addresses is carried in the IS-IS Area Identifier TLV. If more than one Area Addresses are present, multiple TLVs are used to encode them. The IS-IS Area Identifier TLV may be present in the LINK\_STATE attribute only with the Link State Node NLRI.

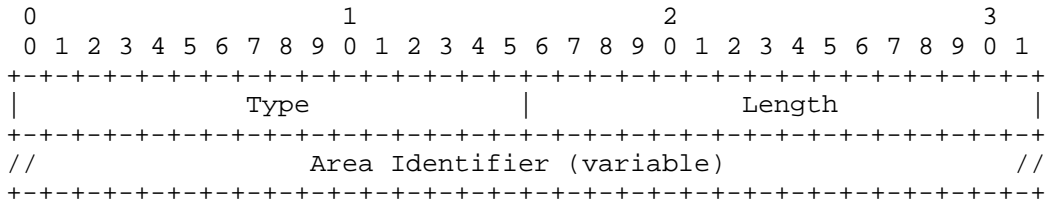


Figure 16: IS-IS Area Identifier TLV Format

3.3.1.3. Node Name TLV

The Node Name TLV is optional. Its structure and encoding has been borrowed from [RFC5301]. The value field identifies the symbolic name of the router node. This symbolic name can be the FQDN for the router, it can be a subset of the FQDN, or it can be any string

operators want to use for the router. The use of FQDN or a subset of it is strongly recommended.

The Value field is encoded in 7-bit ASCII. If a user-interface for configuring or displaying this field permits Unicode characters, that user-interface is responsible for applying the ToASCII and/or ToUnicode algorithm as described in [RFC3490] to achieve the correct format for transmission or display.

Although [RFC5301] is a IS-IS specific extension, usage of the Node Name TLV is possible for all protocols. How a router derives and injects node names for e.g. OSPF nodes, is outside of the scope of this document.

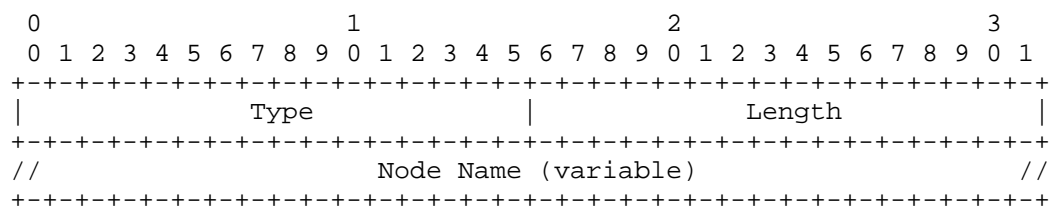


Figure 17: Node Name format

#### 3.3.1.4. Local IPv4/IPv6 Router-ID

The local IPv4/IPv6 Router-ID TLVs are used to describe auxiliary Router-IDs that the IGP might be using, e.g., for TE and migration purposes like correlating a Node-ID between different protocols. If there is more than one auxiliary Router-ID of a given type, then each one is encoded in its own TLV.

#### 3.3.1.5. Opaque Node Attribute TLV

The Opaque Node attribute TLV is an envelope that transparently carries optional node attribute TLVs advertised by a router. An originating router shall use this TLV for encoding information specific to the protocol advertised in the NLRI header Protocol-ID field or new protocol extensions to the protocol as advertised in the NLRI header Protocol-ID field for which there is no protocol neutral representation in the BGP link-state NLRI. A router for example could use this extension in order to advertise the native protocols node attribute TLVs, such as the OSPF Router Informational Capabilities TLV defined in [RFC4970], or the IGP TE Node Capability Descriptor TLV described in [RFC5073].

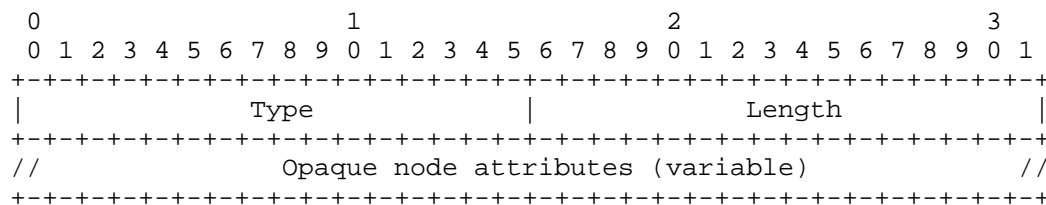


Figure 18: Opaque Node attribute format

### 3.3.2. Link Attribute TLVs

Link attribute TLVs are TLVs that may be encoded in the BGP-LS attribute with a link NLRI. Each 'Link Attribute' is a Type/Length/Value (TLV) triplet formatted as defined in Section 3.1. The format and semantics of the 'value' fields in some 'Link Attribute' TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305] and [RFC5307]. Other 'Link Attribute' TLVs are defined in this document. Although the encodings for 'Link Attribute' TLVs were originally defined for IS-IS, the TLVs can carry data sourced either by IS-IS or OSPF.

The following 'Link Attribute' TLVs are valid in the LINK\_STATE attribute:

TLV Code Point	Description	IS-IS TLV/Sub-TLV	Defined in:
1028	IPv4 Router-ID of Local Node	134/---	[RFC5305]/4.3
1029	IPv6 Router-ID of Local Node	140/---	[RFC6119]/4.1
1030	IPv4 Router-ID of Remote Node	134/---	[RFC5305]/4.3
1031	IPv6 Router-ID of Remote Node	140/---	[RFC6119]/4.1
1088	Administrative group (color)	22/3	[RFC5305]/3.1
1089	Maximum link bandwidth	22/9	[RFC5305]/3.3
1090	Max. reservable link bandwidth	22/10	[RFC5305]/3.5
1091	Unreserved bandwidth	22/11	[RFC5305]/3.6
1092	TE Default Metric	22/18	[RFC5305]/3.7

1093	Link Protection Type	22/20	[RFC5307]/1.2
1094	MPLS Protocol Mask	---	Section 3.3.2.2
1095	Metric	---	Section 3.3.2.3
1096	Shared Risk Link Group	---	Section 3.3.2.4
1097	Opaque link attribute	---	Section 3.3.2.5
1098	Link Name attribute	---	Section 3.3.2.6

Table 7: Link Attribute TLVs

## 3.3.2.1. IPv4/IPv6 Router-ID

The local/remote IPv4/IPv6 Router-ID TLVs are used to describe auxiliary Router-IDs that the IGP might be using, e.g., for TE purposes. All auxiliary Router-IDs of both the local and the remote node MUST be included in the link attribute of each link NLRI. If there are more than one auxiliary Router-ID of a given type, then multiple TLVs are used to encode them.

## 3.3.2.2. MPLS Protocol Mask TLV

The MPLS Protocol TLV carries a bit mask describing which MPLS signaling protocols are enabled. The length of this TLV is 1. The value is a bit array of 8 flags, where each bit represents an MPLS Protocol capability.

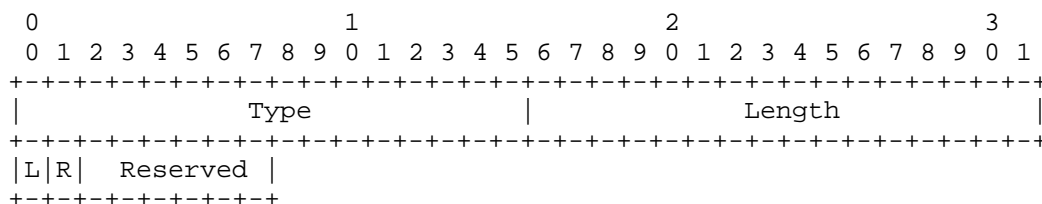


Figure 19: MPLS Protocol TLV

The following bits are defined:

Bit	Description	Reference
'L'	Label Distribution Protocol (LDP)	[RFC5036]
'R'	Extension to RSVP for LSP Tunnels (RSVP-TE)	[RFC3209]
'Reserved'	Reserved for future use	

Table 8: MPLS Protocol Mask TLV Codes

## 3.3.2.3. Metric TLV

The IGP Metric TLV carries the metric for this link. The length of this TLV is variable, depending on the metric width of the underlying protocol. IS-IS small metrics have a length of 1 octet (the two most significant bits are ignored). OSPF metrics have a length of two octets. IS-IS wide-metrics have a length of three octets.

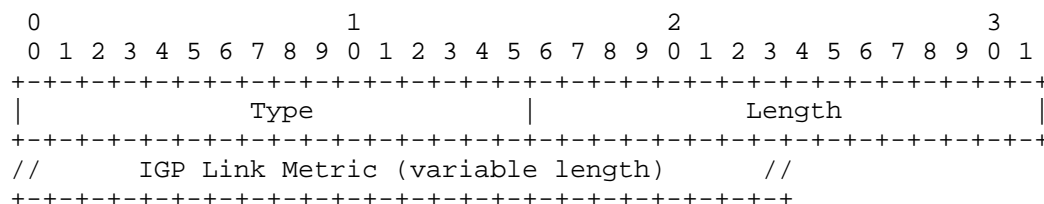


Figure 20: Metric TLV format

## 3.3.2.4. Shared Risk Link Group TLV

The Shared Risk Link Group (SRLG) TLV carries the Shared Risk Link Group information (see Section 2.3, "Shared Risk Link Group Information", of [RFC4202]). It contains a data structure consisting of a (variable) list of SRLG values, where each element in the list has 4 octets, as shown in Figure 21. The length of this TLV is 4 \* (number of SRLG values).

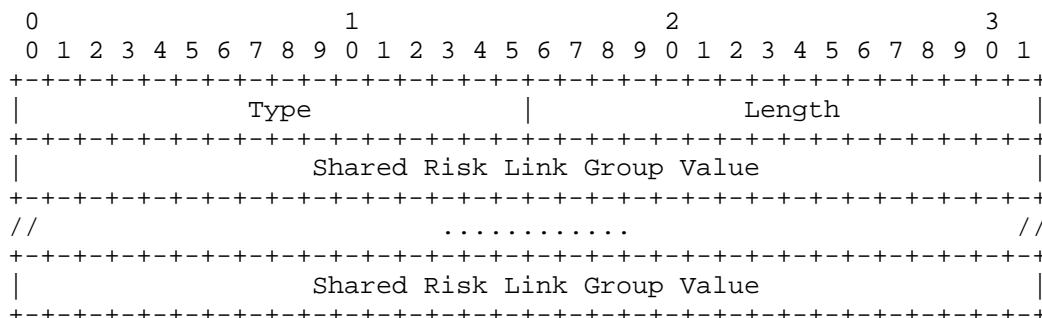


Figure 21: Shared Risk Link Group TLV format

Note that there is no SRLG TLV in OSPF-TE. In IS-IS the SRLG information is carried in two different TLVs: the IPv4 (SRLG) TLV (Type 138) defined in [RFC5307], and the IPv6 SRLG TLV (Type 139) defined in [RFC6119]. In Link State NLRI both IPv4 and IPv6 SRLG information are carried in a single TLV.

### 3.3.2.5. Opaque Link Attribute TLV

The Opaque link attribute TLV is an envelope that transparently carries optional link attribute TLVs advertised by a router. An originating router shall use this TLV for encoding information specific to the protocol advertised in the NLRI header Protocol-ID field or new protocol extensions to the protocol as advertised in the NLRI header Protocol-ID field for which there is no protocol neutral representation in the BGP link-state NLRI.

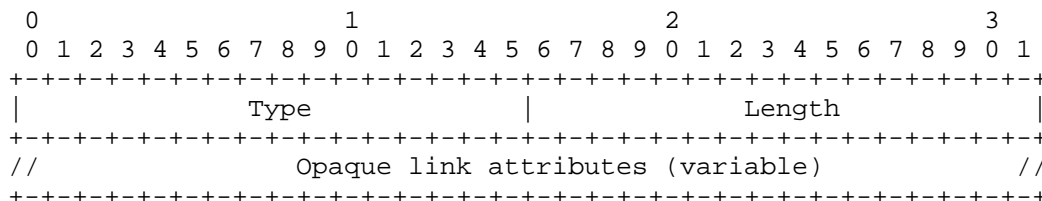


Figure 22: Opaque link attribute format

### 3.3.2.6. Link Name TLV

The Link Name TLV is optional. The value field identifies the symbolic name of the router link. This symbolic name can be the FQDN for the link, it can be a subset of the FQDN, or it can be any string operators want to use for the link. The use of FQDN or a subset of it is strongly recommended.

The Value field is encoded in 7-bit ASCII. If a user-interface for configuring or displaying this field permits Unicode characters, that user-interface is responsible for applying the ToASCII and/or ToUnicode algorithm as described in [RFC3490] to achieve the correct format for transmission or display.

How a router derives and injects link names is outside of the scope of this document.

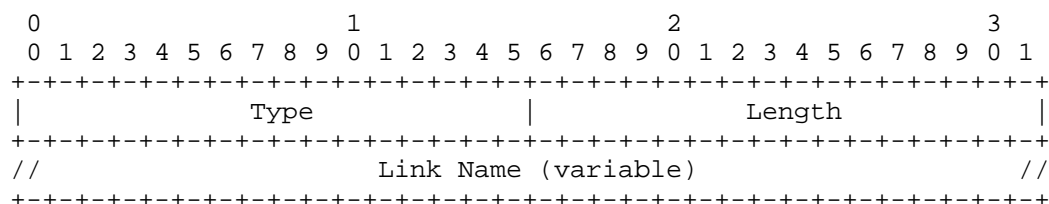


Figure 23: Link Name format

### 3.3.3. Prefix Attribute TLVs

Prefixes are learned from the IGP topology (IS-IS or OSPF) with a set of IGP attributes (such as metric, route tags, etc.) that MUST be reflected into the LINK\_STATE attribute. This section describes the different attributes related to the IPv4/IPv6 prefixes. Prefix Attributes TLVs SHOULD be used when advertising NLRI types 3 and 4 only. The following attributes TLVs are defined:

TLV Code Point	Description	Length	Reference
1152	IGP Flags	1	Section 3.3.3.1
1153	Route Tag	4*n	Section 3.3.3.2
1154	Extended Tag	8*n	Section 3.3.3.3
1155	Prefix Metric	4	Section 3.3.3.4
1156	OSPF Forwarding Address	4	Section 3.3.3.5
1157	Opaque Prefix Attribute	variable	Section 3.3.3.6

Table 9: Prefix Attribute TLVs



3.3.3.1.    IGP Flags TLV

IGP Flags TLV contains IS-IS and OSPF flags and bits originally assigned to the prefix.  The IGP Flags TLV is encoded as follows:

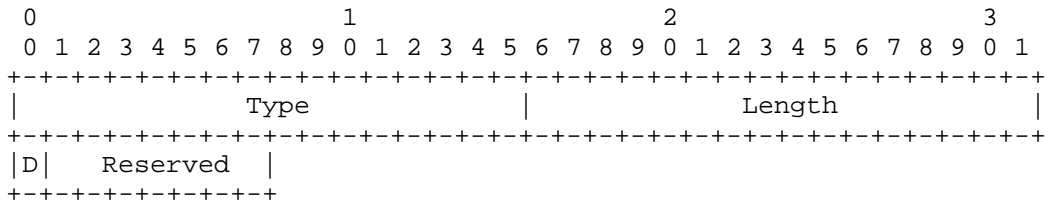


Figure 24: IGP Flag TLV format

The value field contains bits defined according to the table below:

Bit	Description	Reference
'D'	IS-IS Up/Down Bit	[RFC5305]
Reserved	Reserved for future use.	

Table 10: IGP Flag Bits Definitions

3.3.3.2.    Route Tag

Route Tag TLV carries original IGP TAGs (IS-IS [RFC5130] or OSPF) of the prefix and is encoded as follows:

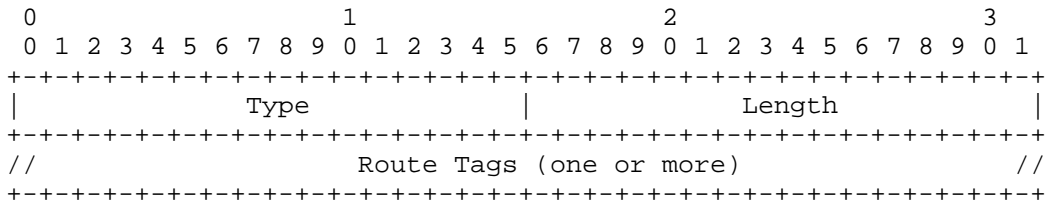


Figure 25: IGP Route TAG TLV format

Length is a multiple of 4.

The value field contains one or more Route Tags as learned in the IGP topology.

## 3.3.3.3.    Extended Route Tag

Extended Route Tag TLV carries IS-IS Extended Route TAGs of the prefix [RFC5130] and is encoded as follows:

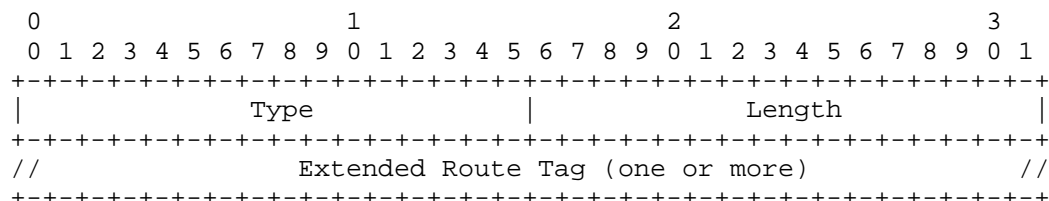


Figure 26: Extended IGP Route TAG TLV format

Length is a multiple of 8.

The 'Extended Route Tag' field contains one or more Extended Route Tags as learned in the IGP topology.

## 3.3.3.4.    Prefix Metric TLV

Prefix Metric TLV carries the metric of the prefix as known in the IGP topology [RFC5305]. The attribute is mandatory and can only appear once.

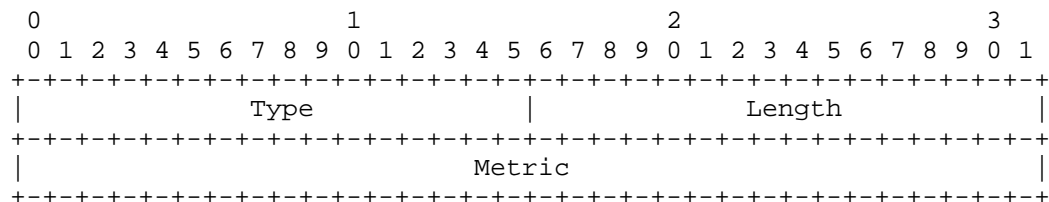


Figure 27: Prefix Metric TLV Format

Length is 4.

## 3.3.3.5.    OSPF Forwarding Address TLV

OSPF Forwarding Address TLV [RFC2328] carries the OSPF forwarding address as known in the original OSPF advertisement. Forwarding address can be either IPv4 or IPv6.

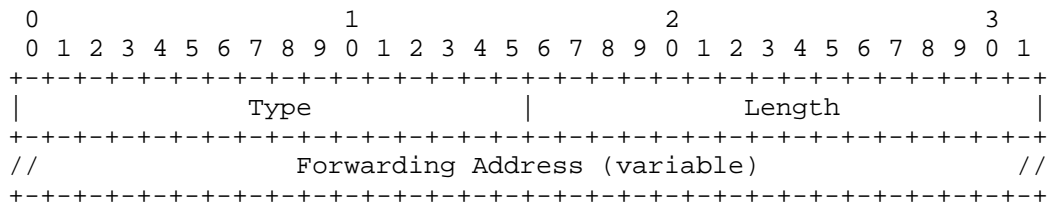


Figure 28: OSPF Forwarding Address TLV Format

Length is 4 for an IPv4 forwarding address an 16 for an IPv6 forwarding address.

### 3.3.3.6. Opaque Prefix Attribute TLV

The Opaque Prefix attribute TLV is an envelope that transparently carries optional prefix attribute TLVs advertised by a router. An originating router shall use this TLV for encoding information specific to the protocol advertised in the NLRI header Protocol-ID field or new protocol extensions to the protocol as advertised in the NLRI header Protocol-ID field for which there is no protocol neutral representation in the BGP link-state NLRI.

The format of the TLV is as follows:

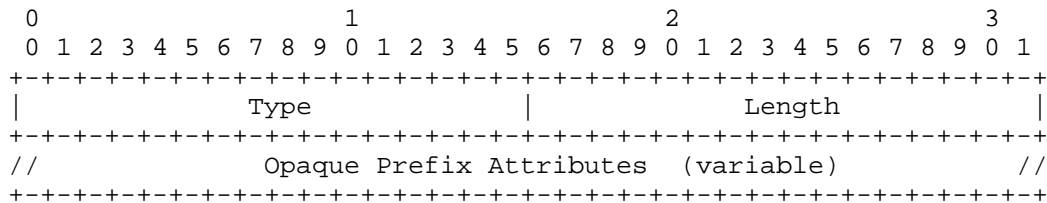


Figure 29: Opaque Prefix Attribute TLV Format

Type is as specified in Table 9 and Length is variable.

### 3.4. BGP Next Hop Information

BGP link-state information for both IPv4 and IPv6 networks can be carried over either an IPv4 BGP session, or an IPv6 BGP session. If IPv4 BGP session is used, then the next hop in the MP\_REACH\_NLRI SHOULD be an IPv4 address. Similarly, if IPv6 BGP session is used, then the next hop in the MP\_REACH\_NLRI SHOULD be an IPv6 address. Usually the next hop will be set to the local end-point address of the BGP session. The next hop address MUST be encoded as described in [RFC4760]. The length field of the next hop address will specify the next hop address-family. If the next hop length is 4, then the

next hop is an IPv4 address; if the next hop length is 16, then it is a global IPv6 address and if the next hop length is 32, then there is one global IPv6 address followed by a link-local IPv6 address. The link-local IPv6 address should be used as described in [RFC2545]. For VPN SAFI, as per custom, an 8 byte route-distinguisher set to all zero is prepended to the next hop.

The BGP Next Hop attribute is used by each BGP-LS speaker to validate the NLRI it receives. However, this specification doesn't mandate any rule regarding the re-write of the BGP Next Hop attribute.

### 3.5. Inter-AS Links

The main source of TE information is the IGP, which is not active on inter-AS links. In some cases, the IGP may have information of inter-AS links ([RFC5392], [RFC5316]). In other cases, for injecting a non-IGP enabled link into the BGP link-state RIB, an implementation MUST support configuration of either 'Static' or 'Direct' links.

### 3.6. Router-ID Anchoring Example: ISO Pseudonode

Encoding of a broadcast LAN in IS-IS provides a good example of how Router-IDs are encoded. Consider Figure 30. This represents a Broadcast LAN between a pair of routers. The "real" (=non pseudonode) routers have both an IPv4 Router-ID and IS-IS Node-ID. The pseudonode does not have an IPv4 Router-ID. Node1 is the DIS for the LAN. Two unidirectional links (Node1, Pseudonode 1) and (Pseudonode1, Node2) are being generated.

The link NRLI of (Node1, Pseudonode1) is encoded as follows: the IGP Router-ID TLV of the local node descriptor is 6 octets long containing ISO-ID of Node1, 1920.0000.2001; the IGP Router-ID TLV of the remote node descriptor is 7 octets long containing the ISO-ID of Pseudonode1, 1920.0000.2001.02. The BGP-LS attribute of this link contains one local IPv4 Router-ID TLV (TLV type 1028) containing 192.0.2.1, the IPv4 Router-ID of Node1.

The link NRLI of (Pseudonode1, Node2) is encoded as follows: the IGP Router-ID TLV of the local node descriptor is 7 octets long containing the ISO-ID of Pseudonode1, 1920.0000.2001.02; the IGP Router-ID TLV of the remote node descriptor is 6 octets long containing ISO-ID of Node2, 1920.0000.2002. The BGP-LS attribute of this link contains one remote IPv4 Router-ID TLV (TLV type 1030) containing 192.0.2.2, the IPv4 Router-ID of Node2.

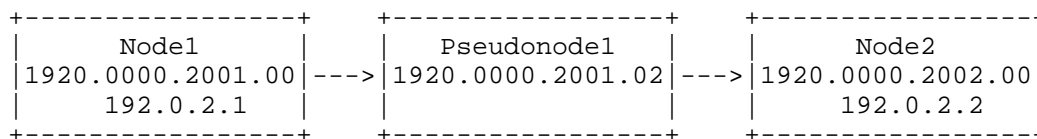


Figure 30: IS-IS Pseudonodes

### 3.7. Router-ID Anchoring Example: OSPFv2 to IS-IS Migration

Graceful migration from one IGP to another requires coordinated operation of both protocols during the migration period. Such a coordination requires identifying a given physical link in both IGPs. The IPv4 Router-ID provides that "glue" which is present in the node descriptors of the OSPF link NLRI and in the link attribute of the IS-IS link NLRI.

Consider a point-to-point link between two routers, A and B, that initially were OSPFv2-only routers and then IS-IS is enabled on them. Node A has IPv4 Router-ID and ISO-ID; node B has IPv4 Router-ID, IPv6 Router-ID and ISO-ID. Each protocol generates one link NLRI for the link (A, B), both of which are carried by BGP-LS. The OSPFv2 link NLRI for the link is encoded with the IPv4 Router-ID of nodes A and B in the local and remote node descriptors, respectively. The IS-IS link NLRI for the link is encoded with the ISO-ID of nodes A and B in the local and remote node descriptors, respectively. In addition, the BGP-LS attribute of the IS-IS link NLRI contains the the TLV type 1028 containing the IPv4 Router-ID of node A; TLV type 1030 containing the IPv4 Router-ID of node B and TLV type 1031 containing the IPv6 Router-ID of node B. In this case, by using IPv4 Router-ID, the link (A, B) can be identified in both IS-IS and OSPF protocol.

## 4. Link to Path Aggregation

Distribution of all links available in the global Internet is certainly possible, however not desirable from a scaling and privacy point of view. Therefore an implementation may support link to path aggregation. Rather than advertising all specific links of a domain, an ASBR may advertise an "aggregate link" between a non-adjacent pair of nodes. The "aggregate link" represents the aggregated set of link properties between a pair of non-adjacent nodes. The actual methods to compute the path properties (of bandwidth, metric) are outside the scope of this document. The decision whether to advertise all specific links or aggregated links is an operator's policy choice. To highlight the varying levels of exposure, the following deployment examples are discussed.

## 4.1. Example: No Link Aggregation

Consider Figure 31. Both AS1 and AS2 operators want to protect their inter-AS {R1,R3}, {R2, R4} links using RSVP-FRR LSPs. If R1 wants to compute its link-protection LSP to R3 it needs to "see" an alternate path to R3. Therefore the AS2 operator exposes its topology. All BGP TE enabled routers in AS1 "see" the full topology of AS and therefore can compute a backup path. Note that the decision if the direct link between {R3, R4} or the {R4, R5, R3} path is used is made by the computing router.

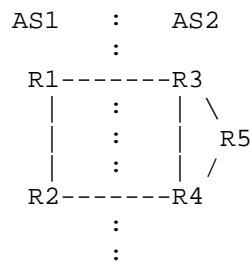


Figure 31: No link aggregation

## 4.2. Example: ASBR to ASBR Path Aggregation

The brief difference between the "no-link aggregation" example and this example is that no specific link gets exposed. Consider Figure 32. The only link which gets advertised by AS2 is an "aggregate" link between R3 and R4. This is enough to tell AS1 that there is a backup path. However the actual links being used are hidden from the topology.

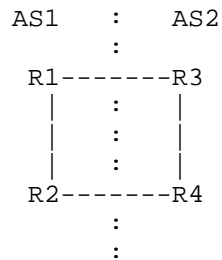


Figure 32: ASBR link aggregation

#### 4.3. Example: Multi-AS Path Aggregation

Service providers in control of multiple ASes may even decide to not expose their internal inter-AS links. Consider Figure 33. AS3 is modeled as a single node which connects to the border routers of the aggregated domain.

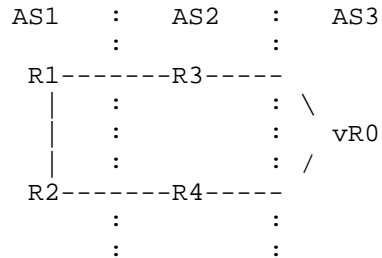


Figure 33: Multi-AS aggregation

#### 5. IANA Considerations

This document requests a code point from the registry of Address Family Numbers. As per early allocation procedure this is AFI 16388.

This document requests a code point from the registry of Subsequent Address Family Numbers. As per early allocation procedure this is SAFI 71.

This document requests a code point from the BGP Path Attributes registry.

This document requests creation of a new registry for node anchor, link descriptor and link attribute TLVs. Values 0-255 are reserved. Values 256-65535 will be used for Codepoints. The registry will be initialized as shown in Table 11. Allocations within the registry will require documentation of the proposed use of the allocated value and approval by the Designated Expert assigned by the IESG (see [RFC5226]).

Note to RFC Editor: this section may be removed on publication as an RFC.

#### 6. Manageability Considerations

This section is structured as recommended in [RFC5706].

## 6.1. Operational Considerations

### 6.1.1. Operations

Existing BGP operational procedures apply. No new operation procedures are defined in this document. It is noted that the NLRI information present in this document purely carries application level data that has no immediate corresponding forwarding state impact. As such, any churn in reachability information has different impact than regular BGP updates which need to change forwarding state for an entire router. Furthermore it is anticipated that distribution of this NLRI will be handled by dedicated route-reflectors providing a level of isolation and fault-containment between different NLRI types.

### 6.1.2. Installation and Initial Setup

Configuration parameters defined in Section 6.2.3 SHOULD be initialized to the following default values:

- o The Link-State NLRI capability is turned off for all neighbors.
- o The maximum rate at which Link State NLRIs will be advertised/withdrawn from neighbors is set to 200 updates per second.

### 6.1.3. Migration Path

The proposed extension is only activated between BGP peers after capability negotiation. Moreover, the extensions can be turned on/off an individual peer basis (see Section 6.2.3), so the extension can be gradually rolled out in the network.

### 6.1.4. Requirements on Other Protocols and Functional Components

The protocol extension defined in this document does not put new requirements on other protocols or functional components.

### 6.1.5. Impact on Network Operation

Frequency of Link-State NLRI updates could interfere with regular BGP prefix distribution. A network operator MAY use a dedicated Route-Reflector infrastructure to distribute Link-State NLRIs.

Distribution of Link-State NLRIs SHOULD be limited to a single admin domain, which can consist of multiple areas within an AS or multiple ASes.



#### 6.1.6. Verifying Correct Operation

Existing BGP procedures apply. In addition, an implementation SHOULD allow an operator to:

- o List neighbors with whom the Speaker is exchanging Link-State NLRIs

#### 6.2. Management Considerations

##### 6.2.1. Management Information

##### 6.2.2. Fault Management

TBD.

##### 6.2.3. Configuration Management

An implementation SHOULD allow the operator to specify neighbors to which Link-State NLRIs will be advertised and from which Link-State NLRIs will be accepted.

An implementation SHOULD allow the operator to specify the maximum rate at which Link State NLRIs will be advertised/withdrawn from neighbors

An implementation SHOULD allow the operator to specify the maximum number of Link State NLRIs stored in router's RIB.

An implementation SHOULD allow the operator to create abstracted topologies that are advertised to neighbors; Create different abstractions for different neighbors.

An implementation SHOULD allow the operator to configure a 64-bit instance ID.

An implementation SHOULD allow the operator to configure a pair of ASN and BGP-LS identifier per flooding set the node participates in.

##### 6.2.4. Accounting Management

Not Applicable.

##### 6.2.5. Performance Management

An implementation SHOULD provide the following statistics:

- o Total number of Link-State NLRI updates sent/received
- o Number of Link-State NLRI updates sent/received, per neighbor
- o Number of errored received Link-State NLRI updates, per neighbor
- o Total number of locally originated Link-State NLRIs

#### 6.2.6. Security Management

An operator SHOULD define ACLs to limit inbound updates as follows:

- o Drop all updates from Consumer peers

### 7. TLV/Sub-TLV Code Points Summary

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

TLV Code Point	Description	IS-IS TLV/ Sub-TLV	Value defined in:
256	Local Node Descriptors	---	Section 3.2.1.2
257	Remote Node Descriptors	---	Section 3.2.1.3
258	Link Local/Remote Identifiers	22/4	[RFC5307]/1.1
259	IPv4 interface address	22/6	[RFC5305]/3.2
260	IPv4 neighbor address	22/8	[RFC5305]/3.3
261	IPv6 interface address	22/12	[RFC6119]/4.2
262	IPv6 neighbor address	22/13	[RFC6119]/4.3
263	Multi-Topology ID	---	Section 3.2.1.5
264	OSPF Route Type	---	Section 3.2.3
265	IP Reachability Information	---	Section 3.2.3
512	Autonomous System	---	Section 3.2.1.4
513	BGP-LS Identifier	---	Section 3.2.1.4
514	Area ID	---	Section 3.2.1.4
515	IGP Router-ID	---	Section 3.2.1.4
1024	Node Flag Bits	---	Section 3.3.1.1
1025	Opaque Node Properties	---	Section 3.3.1.5
1026	Node Name	variable	Section 3.3.1.3
1027	IS-IS Area Identifier	variable	Section 3.3.1.2
1028	IPv4 Router-ID of Local Node	134/---	[RFC5305]/4.3
1029	IPv6 Router-ID of Local Node	140/---	[RFC6119]/4.1
1030	IPv4 Router-ID of Remote Node	134/---	[RFC5305]/4.3
1031	IPv6 Router-ID of Remote Node	140/---	[RFC6119]/4.1
1088	Administrative group (color)	22/3	[RFC5305]/3.1
1089	Maximum link bandwidth	22/9	[RFC5305]/3.3
1090	Max. reservable link bandwidth	22/10	[RFC5305]/3.5
1091	Unreserved bandwidth	22/11	[RFC5305]/3.6

1092	TE Default Metric	22/18	[RFC5305]/3.7
1093	Link Protection Type	22/20	[RFC5307]/1.2
1094	MPLS Protocol Mask	---	Section 3.3.2.2
1095	Metric	---	Section 3.3.2.3
1096	Shared Risk Link Group	---	Section 3.3.2.4
1097	Opaque link attribute	---	Section 3.3.2.5
1098	Link Name attribute	---	Section 3.3.2.6
1152	IGP Flags	---	Section 3.3.3.1
1153	Route Tag	---	[RFC5130]
1154	Extended Tag	---	[RFC5130]
1155	Prefix Metric	---	[RFC5305]
1156	OSPF Forwarding Address	---	[RFC2328]
1157	Opaque Prefix Attribute	---	Section 3.3.3.6

Table 11: Summary Table of TLV/Sub-TLV Codepoints

## 8. Security Considerations

Procedures and protocol extensions defined in this document do not affect the BGP security model. See [I-D.ietf-karp-routing-tcp-analysis] for details.

A BGP Speaker SHOULD NOT accept updates from a Consumer peer.

An operator SHOULD employ a mechanism to protect a BGP Speaker against DDOS attacks from Consumers.

## 9. Contributors

We would like to thank Robert Varga for the significant contribution he gave to this document.

## 10. Acknowledgements

We would like to thank Nischal Sheth, Alia Atlas, David Ward, Derek Yeung, Murtuza Lightwala, John Scudder, Kaliraj Vairavakkalai, Les Ginsberg, Liem Nguyen, Manish Bhardwaj, Mike Shand, Peter Psenak, Rex Fernando, Richard Woundy, Steven Luong, Tamas Mondal, Waqas Alam, Vipin Kumar, Naiming Shen, Balaji Rajagopalan and Yakov Rekhter for

their comments.

## 11. References

### 11.1. Normative References

- [RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.
- [RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, March 1999.
- [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.
- [RFC3490] Faltstrom, P., Hoffman, P., and A. Costello, "Internationalizing Domain Names in Applications (IDNA)", RFC 3490, March 2003.
- [RFC4202] Kompella, K. and Y. Rekhter, "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, October 2005.
- [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.
- [RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, June 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.
- [RFC5130] Previdi, S., Shand, M., and C. Martin, "A Policy Control Mechanism in IS-IS Using Administrative Tags", RFC 5130, February 2008.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5301] McPherson, D. and N. Shen, "Dynamic Hostname Exchange Mechanism for IS-IS", RFC 5301, October 2008.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5307] Kompella, K. and Y. Rekhter, "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, October 2008.
- [RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, February 2011.
- [RFC6286] Chen, E. and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, June 2011.
- [RFC6822] Previdi, S., Ginsberg, L., Shand, M., Roy, A., and D. Ward, "IS-IS Multi-Instance", RFC 6822, December 2012.

## 11.2. Informative References

- [I-D.ietf-alto-protocol]  
Alimi, R., Penno, R., and Y. Yang, "ALTO Protocol", draft-ietf-alto-protocol-13 (work in progress), September 2012.
- [I-D.ietf-karp-routing-tcp-analysis]  
Jethanandani, M., Patel, K., and L. Zheng, "Analysis of BGP, LDP, PCEP and MSDP Issues According to KARP Design Guide", draft-ietf-karp-routing-tcp-analysis-07 (work in progress), April 2013.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S.

Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.

- [RFC5073] Vasseur, J. and J. Le Roux, "IGP Routing Protocol Extensions for Discovery of Traffic Engineering Node Capabilities", RFC 5073, December 2007.
- [RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.
- [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.
- [RFC5693] Seedorf, J. and E. Burger, "Application-Layer Traffic Optimization (ALTO) Problem Statement", RFC 5693, October 2009.
- [RFC5706] Harrington, D., "Guidelines for Considering Operations and Management of New Protocols and Protocol Extensions", RFC 5706, November 2009.
- [RFC6549] Lindem, A., Roy, A., and S. Mirtorabi, "OSPFv2 Multi-Instance Extensions", RFC 6549, March 2012.

#### Authors' Addresses

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

Email: hannes@juniper.net

Jan Medved  
Cisco Systems, Inc.  
170, West Tasman Drive  
San Jose, CA 95134  
US

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

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

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

Adrian Farrel  
Juniper Networks, Inc.  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
US

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

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

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





Internet Engineering Task Force  
Internet-Draft  
Intended status: Standards Track  
Expires: January 15, 2014

H. Jeng  
AT&T  
J. Haas  
Y. Rekhter  
J. Zhang  
Juniper Networks  
July 14, 2013

Multicast Geo-Distribution Control  
draft-rekhter-geo-distribution-control-03

Abstract

Consider a content provider that wants to deliver a particular content to a set of customers/subscribers, where the provider and the subscribers are connected by an IP service provider. This document covers two areas needed to accomplish this:

1. Providing the content provider with the information of whether it can use the multicast connectivity service provided by the IP service provider to deliver a particular content to a particular set of subscribers, and
2. Providing the content provider with a mechanism to restrict delivery of a given content to a particular set of the subscribers.

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 15, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1. Specification of Requirements . . . . .	3
1.1. Introduction . . . . .	3
1.2. Overview of Operations . . . . .	4
2. IANA Considerations . . . . .	5
3. Security Considerations . . . . .	5
4. Acknowledgements . . . . .	6
5. Normative References . . . . .	6
Authors' Addresses . . . . .	6

## 1. Specification of Requirements

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

### 1.1. Introduction

Consider a content provider that wants to deliver a particular content to a set of customers/subscribers, where the provider and the subscribers are connected by an IP service provider. This document covers two areas needed to accomplish this:

1. Providing the content provider with the information of whether it can use the multicast connectivity service provided by the IP service provider to deliver a particular content to a particular set of subscribers, and
2. Providing the content provider with a mechanism to restrict delivery of a given content to a particular set of the subscribers.

For the purpose of this document we assume that a content provider consists of one or more Content Servers, and one or more Content Distribution Controllers. While this document assumes communication between Content Servers and Content Distribution Controllers, the procedures for implementing such communication is outside the scope of this document.

Content Servers are connected to one or more IP service provider (ISP) that can offer both multicast and unicast connectivity service to the subscribers of the content provider. Content provider uses this ISP(s) to deliver content to its subscribers.

Subscribers are connected to the Edge Routers (ERs) of the ISP. Note that the multicast connectivity service provided by the ISP extends all the way to the ERs. Such service could be provided by either deploying IP multicast natively, or with some tunneling mechanism like AMT, or by a combination of both within the ISP. However, between the ERs and the subscribers there may, or may not be multicast connectivity.

In the case where a particular subscriber of a given content provider does not have multicast connectivity to its ER, the content provider would use IP unicast service provided by the ISP to transmit the particular content to that subscriber.

A subscriber may want to access a particular content that is not

available to that subscriber due to policy reasons. When that subscriber would have received that content via unicast connectivity, the Content Distribution Controller, or the Content Servers, or both may enforce the policy to not deliver the content. However, when the content would be delivered via multicast connectivity it may be possible for the subscriber to receive the content by illicitly participating in the multicast signaling for that content.

To prevent a subversion of the intent of this content delivery policy, a mechanism is provided to make this policy available to devices participating in multicast signaling.

## 1.2. Overview of Operations

An ISP, using the procedures described in Multicast Distribution Reachability Signaling [MDRS], provides a content provider, and specifically Content Distribution Controller(s) of that content provider, with the information of whether a particular subscriber of that content provider has multicast connectivity to an ER of that ISP with the information of whether a particular group of subscribers can receive multicast content.

For each content provided by a content provider, the content provider maintains a list of subscribers who are either excluded or allowed to receive the content. For the purpose of maintaining this list this document assumes that subscribers are grouped into "zones" based on IP addresses, so that exclusion/inclusion uniformly applies to all the subscribers within a given zone. Procedures by which subscribers are grouped into zones are outside the scope of this document. However, this document assumes that this grouping is done consistently by both the content provider and the ISP(s) that the content provider uses for delivering its content.

To enforce the exclusion/inclusion policies, the content provider uses procedures described in Multicast Distribution Control Signaling [MDCS].

For each content provided by a content provider, the content provider selects a particular multicast channel (S, G) for distributing this content using multicast connectivity service. Procedures by which the content provider selects a particular multicast channel, and maintains the mapping are outside the scope of this document.

Subscribers are connected to the Edge Routers (ERs) of the ISP. Note that when multicast connectivity service provided is by the ISP, that service extends all the way to the ERs. Such service could be provided by either deploying IP multicast natively, or with some tunneling mechanism like AMT, or a combination of both within the

ISP. However, between the ERs and the subscribers there may, or may not be multicast connectivity.

When a subscriber wants to receive the particular content from its content provider, the subscriber issues a request for this content to the Content Distribution Controller of the provider. When the Content Distribution Controller receives the request, the Content Distribution Controller uses the information carried in the request (e.g., IP address of the subscriber) to determine the zone of the subscriber, and based on that zone to determine whether the subscriber can receive this content.

If the Content Distribution Controller determines that the subscriber can receive the content, then based on the information provided by the multicast distribution reachability signaling the Content Distribution Controller determines whether the subscriber can receive this content using multicast connectivity service, and if yes, then returns to the subscriber the multicast channel selected for distributing the content.

If the Content Distribution Controller determines that the subscriber can receive the content, but can not receive the content using multicast connectivity service, the Content Distribution Controller returns to the subscriber the information needed to receive this content using unicast connectivity service.

If the content would have been delivered to the subscriber via multicast connectivity, but the Content Distribution Controller had determined the subscriber was not permitted access to this content, then this policy may need to be enforced by the Edge Routers or upstream multicast routers to prevent illicit access of this content. This policy is enforced by utilizing filtering information distributed using Multicast Distribution Control Signaling [MDCS].

Specification of the procedures for communication between subscribers and Content Distribution Controllers are outside the scope of this document.

## 2. IANA Considerations

This document introduces no IANA Considerations.

## 3. Security Considerations

TBD

#### 4. Acknowledgements

The authors would like to thank Han Nguyen for his contributions to this document.

#### 5. Normative References

- [MDCS] Jeng, H., Haas, J., Rekhter, Y., and J. Zhang, "Multicast Distribution Control Signaling", draft-rekhter-mdcs-00.txt (work in progress), 2013.
- [MDRS] Jeng, H., Haas, J., Rekhter, Y., and J. Zhang, "Multicast Distribution Reachability Signaling", draft-rekhter-mdrs-00.txt (work in progress), 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

#### Authors' Addresses

Huajin Jeng  
AT&T

Email: [hj2387@att.com](mailto:hj2387@att.com)

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

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

Yakov Rekhter  
Juniper Networks  
1194 N. Mathida Ave.  
Sunnyvale, CA 94089  
US

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

Jeffrey (Zhaohui) Zhang  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
US

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





Internet Engineering Task Force  
Internet-Draft  
Updates: 5575 (if approved)  
Intended status: Standards Track  
Expires: January 15, 2014

H. Jeng  
AT&T  
J. Haas  
Y. Rekhter  
J. Zhang  
Juniper Networks  
July 14, 2013

Multicast Distribution Control Signaling  
draft-rekhter-mdcs-00

Abstract

This document describes a mechanism whereby the BGP Flow Specification NLRI format may be utilized to distribute multicast Control Plane filters. This mechanism is called Multicast Distribution Control Signaling (MDCS).

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 15, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	3
2. Specification of Requirements . . . . .	3
2.1. Multicast Distribution Control Signaling . . . . .	3
2.2. An example of configuration on ERs . . . . .	6
3. Summary of Updates to BGP Flowspec . . . . .	7
4. IANA Considerations . . . . .	7
5. Security Considerations . . . . .	7
6. Acknowledgements . . . . .	7
7. References . . . . .	7
7.1. Normative References . . . . .	7
7.2. Informative References . . . . .	8
Authors' Addresses . . . . .	8

## 1. Introduction

Consider a content provider that wants to deliver a particular content to a set of customers/subscribers, where the provider and the subscribers are connected by an IP service provider and the content is distributed using multicast connectivity. The content provider may wish to restrict delivery of the content to a subset of the subscribers in a centralized fashion.

For the purpose of this document we assume that a content provider consists of one or more Content Servers, and one or more Content Distribution Controllers. While this document assumes communication between Content Servers and Content Distribution Controllers, the procedures for implementing such communication is outside the scope of this document.

Content Servers are connected to one or more IP service providers (ISPs) that are offering multicast delivery of the content to the subscribers of the content provider. Content providers use these ISPs to deliver content to their subscribers.

Subscribers are connected to the Edge Routers (ERs) of the ISP. Note that the multicast connectivity service provided by the ISP extends all the way to the ERs. Such service could be provided by either deploying IP multicast natively, or with some tunneling mechanism like AMT, or by a combination of both within the ISP. However, between the ERs and the subscribers there may, or may not be multicast connectivity.

For further information, see [geo-dist].

## 2. Specification of Requirements

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

### 2.1. Multicast Distribution Control Signaling

Multicast distribution control signaling is intended to enforce exclusion/inclusion policies of a content provider, and specifically to prevent a subscriber from accessing a particular multicast channel carrying a particular content provided by the content provider if the subscriber obtained the information about this channel through some illegitimate means.

Multicast distribution control signaling for a particular content is

originated by Content Distribution Controller(s), and uses BGP Flow Spec [RFC5575] as follows:

For a particular content carried over a particular (S, G) multicast flow the Content Distribution Controller responsible for that content originates a BGP Flow Spec route. This route is carried using BGP multi-protocol capabilities [RFC4760] with AFI 1 (for IPv4) or 2 (for IPv6), and the MCAST-FLOWSPEC SAFI. The NLRI of the route carries S in the Source Prefix component (with length of 32 for IPv4 or 128 for IPv6), and G in the Destination Prefix component (with length of 32 for IPv4 or 128 for IPv6).

This route is ultimately propagated to the ER of the ISP connected to the content provider.

An ER that receives BGP Flow Spec routes carrying the multicast distribution control information applies it to PIM and/or IGMP messages the ER receives from the subscribers connected to that ER. (Note that such IGMP messages may be encapsulated in MDT messages.) Specifically, the ER, based on the information received in the BGP Flow Spec routes, decides whether to accept (or reject) a particular PIM or IGMP Join received on one of its subscriber's ports, as follows:

As a Content Distribution Controller originates a BGP Flow Spec route for a particular (S, G) multicast flow, such a route will carry one or more Route Targets [RFC4360], which will ultimately control inclusion/exclusion of that flow on individual ports of ERs that receive this route.

Each subscriber port on an ER is associated with one or more zones. For each zone that a port belongs to, the port is provisioned with two sets of RTs associated with that zone - the inclusion set is for allowing to accept PIM or IGMP Join for some content (or to be more precise for the (S, G) flow that carries that content), and the exclusion set is for disallowing to accept PIM or IGMP Joins for some other content. All those RTs (of all subscribers ports) control import of BGP Flow Spec routes by the ER.

Note that the RTs associated with the subscriber port are ordered. This permits configurations that accommodate include or exclude policies of zones of differing geographic size or overlap. See below for an example.

If the RTs carried by a given BGP Flow Spec route carrying multicast distribution control signaling match the inclusion set of RTs associated with a given port on an ER, then PIM or IGMP Joins for the (S, G) carried in the route and received from the subscriber(s)

connected to that port SHOULD be accepted by the ER. If the RTs carried by the route match the exclusion set, then PIM or IGMP Joins for the (S, G) carried in the route MUST NOT be accepted when received from the subscriber(s) connected to that port. (See example section below.)

Each subscriber port on an ER is provisioned with the default inclusion/exclusion policy that controls acceptance (or rejection) of PIM or IGMP Join messages in the absence of any multicast distribution control signaling. In the former case, in the absence of any multicast distribution signaling, subscribers connected to that port may receive any multicast flow. In the latter case, in the absence of any multicast distribution control signaling, subscribers connected to that port may receive no multicast flows. BGP Flow Spec routes that carry multicast distribution control signaling modify such default behavior.

Once a Content Distribution Controller determines that a particular (S, G) multicast stream no longer used to carry a particular content, the Content Distribution Controller withdraws the BGP Flow Spec route that carries multicast distribution control information for that content.

Note that while [RFC5575] uses the information carried in BGP Flow Spec routes for the purpose of Data Plane filtering, this document uses this information for the purpose of filtering multicast Control Plane traffic (PIM or IGMP).

To constrain the distribution of BGP Flow Spec routes that carry multicast distribution control information to only the relevant ERs, the ERs MAY originate Route Target Constraint (RTC) routes that carry the RTs that control import of the BGP Flow Spec routes on these ERs.

To constrain the import of these RTC routes to only the Content Distribution Controllers, the Content Distribution Controllers are configured with one or more RTs. These RTs control import by the Content Distribution Controller(s) of the RTC routes originated by the ERs. Furthermore, the Content Distribution Controllers MAY themselves originate RTC routes that carry the import RT(s) configured on these Content Distribution Controllers, and that control import of RTC routes by these Content Distribution Controllers.

This document assumes that if a given content provider has multiple Content Distribution Controllers, then all of these Controllers are provisioned with the same RT(s) that control import of the RTC routes originated by the ERs. Furthermore, this document assumes that if a given ISP is providing (multicast) connectivity service to more than

one content provider, then the RTC routes originated by any of the ERs of that ISP MUST carry the set union of the import RTs used by the Content Distribution Controllers of all of these content providers.

RTs carried by routes with AFI 1 and MCAST-FLOWSPEC SAFI SHOULD NOT be re-used by routes with any other AFI and/or SAFI. Likewise, RTs carried by routes with AFI 2 and MCAST-FLOWSPEC SAFI SHOULD NOT be re-used by routes with any other AFI and/or SAFI. Furthermore, RTs carried by routes with AFI 1 and SAFI 132 (AFI/SAFI used by RTC routes) SHOULD NOT be re-used by routes with any other AFI and/or SAFI.

Note that while [RFC4684] uses RTC routes to constrain distribution of VPN-IP routes [RFC4364], this document uses RTC routes to constrain distribution of BGP Flow Spec routes, and also to (recursively) constrain distribution of RTC routes themselves.

## 2.2. An example of configuration on ERs

Consider an ER in Manhattan that has a port that is provisioned with the following import RTs:

```
<include-manhattan, exclude-manhattan, include-nyc, exclude-  
nyc, include-east, exclude-east, include-usa, exclude-usa>
```

When the ER receives a Flow Spec route with <exclude-nyc, include-manhattan, include-usa> RTs, the ER first try to match "include-manhattan" or "exclude-manhattan" (the first ones on the list) - and the result is "include-manhattan". Therefore, the (S, G) carried in the Flow Spec route is allowed on that port of the ER.

Consider another ER in Boston that has a port that is provisioned with the following import RTs:

```
<include-cambridge, exclude-cambridge, include-bos, exclude-  
bos, include-east, exclude-east, include-usa, exclude-usa>
```

The above mentioned Flow Spec route will be imported (due to the include-usa RT), and will result in the (S, G) carried in the flow Spec route to be allowed on that port of the ER.

Now consider a different Flow Spec route with the <exclude-usa, include-bos, include-nyc, exclude-manhattan> RTs. The (S, G) carried in the route will be disallowed in Manhattan, allowed in Boston, and allowed in Queens (as the route will match the "include-nyc" RT).

### 3. Summary of Updates to BGP Flowspec

As described above, this document makes small changes to the BGP Flow Specification mechanism when carried using the MCAST-FLOWSPEC SAFI:

- o Destination addresses will contain a multicast group rather than a unicast destination.
- o Flow specification routes for this SAFI are used for filtering multicast Control Plane traffic rather than the matching multicast traffic itself.
- o Flow specification routes for this SAFI will carry one or more Route Target extended communities.
- o Flow specification component types not applicable to signaling multicast Control Plane traffic MUST be ignored. E.g.: ICMP type, ICMP code, TCP flags, Fragment.

### 4. IANA Considerations

This document defines a new BGP Subsequent Address Family Identifier (SAFI) value, MCAST-FLOWSPEC. The authors request assignment of a value from the First Come, First Served portion of this registry.

### 5. Security Considerations

TBD

### 6. Acknowledgements

The authors would like to thank Han Nguyen for his contributions to this document.

### 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.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, 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.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, August 2009.

## 7.2. Informative References

- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [geo-dist] Jeng, H., Haas, J., Rekhter, Y., and J. Zhang, "Multicast Geo Distribution Control", draft-rekhter-geo-distribution-control-03.txt (work in progress), 2013.

## Authors' Addresses

Huajin Jeng  
AT&T

Email: [hj2387@att.com](mailto:hj2387@att.com)

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

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

Yakov Rekhter  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
US

Email: yakov@juniper.net

Jeffrey (Zhaohui) Zhang  
Juniper Networks  
1194 N. Mathilda Ave.  
Sunnyvale, CA 94089  
US

Email: zzhang@juniper.net



Internet Engineering Task Force  
Internet-Draft  
Intended status: Standards Track  
Expires: January 15, 2014

H. Jeng  
AT&T  
J. Haas  
Y. Rekhter  
J. Zhang  
Juniper Networks  
July 14, 2013

Multicast Distribution Reachability Signaling  
draft-rekhter-mdrs-00

Abstract

This document describes a mechanism whereby a subscriber's Internet service provider may signal in BGP the ability of the subscriber network to receive content using multicast connectivity. This mechanism is called Multicast Distribution Reachability Signaling (MDRS).

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 15, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must

include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

#### Table of Contents

1. Introduction . . . . .	3
2. Specification of Requirements . . . . .	3
2.1. Multicast Distribution Reachability Signaling . . . . .	4
3. IANA Considerations . . . . .	5
4. Security Considerations . . . . .	5
5. Acknowledgements . . . . .	5
6. Normative References . . . . .	5
Authors' Addresses . . . . .	5

## 1. Introduction

Consider a content provider that wants to deliver a particular content to a set of customers/subscribers, where the provider and the subscribers are connected by an IP service provider. This content provider can deliver its content via unicast connectivity or, if supported by the subscriber network, multicast connectivity. A mechanism is required to determine if the subscriber network supports delivery of content to subscribers via multicast connectivity.

This document describes a mechanism whereby the subscriber's Internet service provider may signal in BGP the ability of the subscriber network to receive the content using multicast connectivity. This mechanism is called Multicast Distribution Reachability Signaling (MDRS).

For the purpose of this document we assume that a content provider consists of one or more Content Servers, and one or more Content Distribution Controllers. While this document assumes communication between Content Servers and Content Distribution Controllers, the procedures for implementing such communication is outside the scope of this document.

Content Servers are connected to one or more IP service providers (ISPs) that can offer both multicast and unicast connectivity service to the subscribers of the content provider. Content providers use these ISPs to deliver content to their subscribers.

Subscribers are connected to the Egress Routers (ERs) of the ISP. Note that the multicast connectivity service provided by the ISP extends all the way to the ERs. Such service could be provided by either deploying IP multicast natively, or with some tunneling mechanism like AMT, or by a combination of both within the ISP. However, between the ERs and the subscribers there may, or may not be multicast connectivity.

In the case where a particular subscriber of a given content provider does not have multicast connectivity to its ER, the content provider would use IP unicast service provided by the ISP to transmit the particular content to that subscriber.

## 2. Specification of Requirements

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

## 2.1. Multicast Distribution Reachability Signaling

Multicast distribution reachability signaling is responsible for giving a content provider, and specifically Content Distribution Controller(s) of the content provider the information of whether a particular subscriber of that content provider has multicast connectivity to an ER of an ISP that the content provider uses for distributing its content.

This document assumes that each ER can determine the multicast reachability status for each of the subscriber connected to that ER. Procedures by which an ER accomplishes this are outside the scope of this document.

To indicate whether a given ER has multicast reachability to a subscriber (be that either a native multicast or AMT) this document uses BGP as follows. An ER originates into IBGP routes for the subscribers connected to that ER for which the ER has multicast reachability. These routes are carried using BGP multi-protocol capabilities [RFC4760] with AFI 1 or 2, and the MCAST-REACH SAFI. The NLRI field in the MP\_REACH\_NLRI/MP\_UNREACH\_NLRI attribute of these routes contains subscribers' IP addresses encoded as IP address prefixes. The value of the AFI field in the MP\_REACH\_NLRI/MP\_UNREACH\_NLRI attribute of these routes determines whether subscribers' addresses are IPv4 or IPv6 (AFI 1 indicates IPv4 addresses, AFI 2 indicates IPv6 addresses).

A Content Distribution Controller, when it receives such routes, uses them to determine whether the content could be delivered to the subscribers via the ISP who owns the ERs using the multicast connectivity service provided by the ISP.

To constrain the flow of BGP routes that carry multicast distribution reachability information such routes carry a particular Route Target (RT) Extended Community [RFC4360], and Content Distribution Controller(s) are provisioned to import routes with such a RT.

RTs carried by routes with AFI 1 and MCAST-REACH SAFI SHOULD NOT be re-used by routes with any other AFI and/or SAFI. Likewise, RTs carried by routes with AFI 2 and MCAST-REACH SAFI SHOULD NOT be re-used by routes with any other AFI and/or SAFI.

To facilitate such constrained distribution of multicast distribution reachability information one MAY use Constrained Route Distribution [RFC4684].

### 3. IANA Considerations

This document defines a new BGP Subsequent Address Family Identifier (SAFI) value, MCAST-REACH. The authors request assignment of a value from the First Come, First Served portion of this registry.

### 4. Security Considerations

TBD

### 5. Acknowledgements

The authors would like to thank Han Nguyen for his contributions to this document.

### 6. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.
- [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.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

### Authors' Addresses

Huajin Jeng  
AT&T

Phone:  
Email: [hj2387@att.com](mailto:hj2387@att.com)



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

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

Yakov Rekhter  
Juniper Networks  
1194 N. Mathida Ave.  
Sunnyvale, CA 94089  
US

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

Jeffrey (Zhaohui) Zhang  
Juniper Networks  
1194 N. Mathida Ave.  
Sunnyvale, CA 94089  
US

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



Internet Engineering Task Force  
Internet-Draft  
Intended status: Standards Track  
Expires: January 13, 2014

J. Uttaro  
AT&T  
E. Chen  
Cisco Systems  
B. Decraene  
Orange  
J. Scudder  
Juniper Networks  
July 12, 2013

Support for Long-lived BGP Graceful Restart  
draft-uttaro-idr-bgp-persistence-02

Abstract

In this document we introduce a new BGP capability termed "Long-lived Graceful Restart Capability" so that stale routes can be retained for a longer time upon session failure. In addition a new BGP community "LLGR\_STALE" is introduced for marking stale routes retained for a longer time. We also specify that such long-lived stale routes be treated as the least-preferred, and their advertisements be limited to BGP speakers that have advertised the new capability. Use of this extension is not advisable in all cases, and we provide guidelines to help determine if it is.

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 13, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1. Introduction . . . . .	3
1.1. Requirements Language . . . . .	4
2. Definitions . . . . .	4
3. Protocol Extensions . . . . .	5
3.1. Long-lived Graceful Restart Capability . . . . .	5
3.2. LLGR_STALE Community . . . . .	6
3.3. NO_LLGR Community . . . . .	6
4. Operation . . . . .	7
4.1. Use of Graceful Restart Capability . . . . .	7
4.2. Session Resets . . . . .	7
4.3. Processing LLGR_STALE Routes . . . . .	9
4.4. Route Selection . . . . .	10
4.5. Multicast VPN . . . . .	10
4.6. Errors . . . . .	10
4.7. Optional Partial Deployment Procedure . . . . .	10
4.8. Procedures When BGP is the PE-CE Protocol in a VPN . . . . .	11
5. Deployment Considerations . . . . .	12
5.1. When BGP is the PE-CE Protocol in a VPN . . . . .	13
5.2. Risks of Depreferencing Routes . . . . .	13
6. Security Considerations . . . . .	14
7. Examples of Operation . . . . .	16
8. Acknowledgements . . . . .	18
9. Contributors . . . . .	18
10. IANA Considerations . . . . .	19
11. References . . . . .	19
11.1. Normative References . . . . .	19
11.2. Informative References . . . . .	20
Authors' Addresses . . . . .	20

## 1. Introduction

Historically, routing protocols in general and BGP in particular have been designed with a focus on correctness, where a key part of "correctness" is for each network element's forwarding state to converge toward the current state of the network as quickly as possible. For this reason, the protocol was designed to remove state advertised by routers which went down (from a BGP perspective) as quickly as possible. Over time, this has been relaxed somewhat, notably by BGP Graceful Restart [RFC4724]; however, the paradigm has remained one of attempting to rapidly remove "stale" state from the network.

Over time, two phenomena have arisen that call into question the underlying assumptions of this paradigm. The first is the widespread adoption of tunneled forwarding infrastructures, for example MPLS. Such infrastructures eliminate the risk of some types of forwarding loops that can arise in hop-by-hop forwarding, and thus reduce one of the motivations for strong consistency between forwarding elements. The second is the increasing use of BGP as a transport for data less closely associated with packet forwarding than was originally the case. Examples include the use of BGP for autodiscovery (VPLS [RFC4761]) and filter programming (FLOWSPEC [RFC5575]). In these cases, BGP data takes on a character more akin to configuration than to traditional routing.

The observations above motivate a desire to offer network operators the ability to choose to retain BGP data for a longer period than has hitherto been possible when the BGP control plane fails for some reason. Although the semantics of BGP Graceful Restart [RFC4724] are close to those desired, several gaps exist, most notably in maximum time for which "stale" information can be retained -- Graceful Restart imposes a 4095 second upper bound.

In this document we introduce a new BGP capability termed "Long-lived Graceful Restart Capability" so that stale information can be retained for a longer time across a session reset. We also introduce a new BGP community, "LLGR\_STALE", to mark such information. Such stale information is to be treated as least-preferred, and its advertisement limited to BGP speakers that support the new capability. Where possible, we reference the semantics of BGP Graceful Restart [RFC4724] rather than specifying similar semantics in this document.

The expected deployment model for this extension is that it will only be invoked for certain address families. This is discussed in more detail in the Deployment Considerations section (Section 5). When used, its use may be combined with that of traditional Graceful

Restart, in which case it is invoked only after the traditional Graceful Restart interval has elapsed, or it may be invoked immediately. Apart from the potential to greatly extend the timer, the most obvious difference between Long-Lived and traditional Graceful Restart is that in the Long-Lived version, routes are "depreferenced", that is, treated as least-preferred, whereas in the traditional version, route preference is not affected. The design choice to treat Long-Lived Stale routes as least-preferred was informed by the expectation that they might be retained for a (potentially) almost unbounded period of time, whereas in the traditional Graceful Restart case, stale routes are retained for only a brief interval. In the GR case, the tradeoff between advertising new route status (at the cost of routing churn) and not advertising it (at the cost of suboptimal or incorrect route selection) is resolved in favor of not advertising, and in the LLGR case, it is resolved in favor of advertising new state.

### 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. Definitions

**Depreference, Depreferenced:** A route is said to be depreferenced if it has its route selection preference reduced in reaction to some event.

**GR:** Abbreviation for "Graceful Restart" [RFC4724], also sometimes referred to herein as "conventional Graceful Restart" or "conventional GR" to distinguish it from the "Long-lived Graceful Restart" defined by this document.

**Helper:** Or "helper router". During Graceful Restart or Long-lived Graceful Restart, the router that detects a session failure and applies the listed procedures. [RFC4724] refers to this as the "receiving speaker".

**LLGR:** Abbreviation for "Long-lived Graceful Restart".

**LLST:** Abbreviation for "Long-lived Stale Time".

**Route:** We use "route" to mean any information encoded as a BGP NLRI and set of path attributes. As discussed above, the connection between such routes and installation of forwarding state may be quite remote.

### 3. Protocol Extensions

A new BGP capability and two new BGP communities are introduced.

#### 3.1. Long-lived Graceful Restart Capability

The "Long-lived Graceful Restart Capability" is a new BGP capability [RFC5492] that can be used by a BGP speaker to indicate its ability to preserve its state according to the procedures of this document. This capability **MUST** be advertised in conjunction with the Graceful Restart capability [RFC4724], see the "Use of Graceful Restart Capability" section (Section 4.1).

The capability value consists of one or more tuples <AFI, SAFI, Flags, Long-lived Stale Time> as follows:

```

+-----+
| Address Family Identifier (16 bits) |
+-----+
| Subsequent Address Family Identifier (8 bits) |
+-----+
| Flags for Address Family (8 bits) |
+-----+
| Long-lived Stale Time (24 bits) |
+-----+
| ... |
+-----+
| Address Family Identifier (16 bits) |
+-----+
| Subsequent Address Family Identifier (8 bits) |
+-----+
| Flags for Address Family (8 bits) |
+-----+
| Long-lived Stale Time (24 bits) |
+-----+

```

The meaning of the fields are as follows:

Address Family Identifier (AFI), Subsequent Address Family Identifier (SAFI):

The AFI and SAFI, taken in combination, indicate that the BGP speaker has the ability to preserve its forwarding state for the address family during a subsequent BGP restart. Routes may be explicitly associated with a particular AFI and SAFI using the encoding of [RFC4760] or implicitly associated with <AFI=IPv4, SAFI=Unicast> if using the encoding of [RFC4271].

## Flags for Address Family:

This field contains bit flags relating to routes that were advertised with the given AFI and SAFI.

```

    0 1 2 3 4 5 6 7
    +---+---+---+---+
    |F|   Reserved   |
    +---+---+---+---+

```

The most significant bit is used to indicate whether the state for routes that were advertised with the given AFI and SAFI has indeed been preserved during the previous BGP restart. When set (value 1), the bit indicates that the state has been preserved. This bit is called the "F bit" since it was historically used to indicate preservation of Forwarding State. Use of the F bit is detailed in the Session Resets section (Section 4.2).

The remaining bits are reserved and MUST be set to zero by the sender and ignored by the receiver.

## Long-lived Stale Time:

This time (in seconds) specifies how long stale information (for the AFI/SAFI) may be retained (possibly in conjunction with the period specified by the "Restart Time" in the Graceful Restart Capability, if present).

## 3.2. LLGR\_STALE Community

We introduce a new BGP community [RFC1997] "LLGR\_STALE" (value: TBD). It can be used to mark stale routes retained for a longer period of time. Such long-lived stale routes are to be handled according to the procedures specified in the Operation section (Section 4).

An implementation MAY allow users to configure policies that accept, reject, or modify routes based on the presence or absence of this community.

## 3.3. NO\_LLGR Community

We introduce a new BGP community "NO\_LLGR" (value: TBD). It can be used to mark routes which a BGP speaker does not want treated according to these procedures, as detailed in the Operation section (Section 4).



An implementation MAY allow users to configure policies that accept, reject, or modify routes based on the presence or absence of this community.

#### 4. Operation

A BGP speaker MAY use BGP Capabilities Advertisements [RFC5492] to advertise the "Long-lived Graceful Restart Capability" to indicate its ability to retain state and perform related procedures specified in this document. The setting of the parameters for an AFI/SAFI depends on the properties of the BGP speaker, network scale, and local configuration.

In the presence of the "Long-lived Graceful Restart Capability", the procedures specified in [RFC4724] and [I-D.ietf-idr-bgp-gr-notification] continue to apply unless explicitly revised by this document.

##### 4.1. Use of Graceful Restart Capability

The Graceful Restart capability MUST be advertised in conjunction with the LLGR capability. If it is not so advertised, the LLGR capability MUST be disregarded. The purpose for mandating that both be used in conjunction is to enable reuse of certain base mechanisms that are common to both "flavors", notably origination, collection and processing of EoR, as well as the finite state machine modifications and connection reset logic introduced by GR.

We observe that if support for conventional Graceful Restart is not desired for the session, the conventional GR phase can be skipped by omitting all AFI/SAFI from the GR capability, advertising a Restart Time of zero, or both. The Session Resets section (Section 4.2) discusses the interaction of conventional and long-lived GR.

##### 4.2. Session Resets

BGP Graceful Restart [RFC4724], updated by [I-D.ietf-idr-bgp-gr-notification], defines conditions under which a BGP session can reset and have its associated routes retained. If such a reset occurs for a session for which the LLGR Capability has also been exchanged, the following procedures apply.

If the Graceful Restart Capability that was received does not list all AFI/SAFI supported by the session, then for those non-listed AFI/SAFI the GR "Restart Time" shall be deemed zero. Similarly, if the received LLGR Capability does not list all AFI/SAFI supported by the session, then for those non-listed AFI/SAFI the "Long-lived Stale

Time" shall be deemed zero.

The following text in Section 4.2 of the GR specification [RFC4724] no longer applies:

If the session does not get re-established within the "Restart Time" that the peer advertised previously, the Receiving Speaker MUST delete all the stale routes from the peer that it is retaining.

and the following procedures are specified instead:

After the session goes down and before the session is re-established, the stale routes for an AFI/SAFI MUST be retained. The interval for which they are retained is limited by the sum of the "Restart Time" in the received Graceful Restart Capability and the "Long-lived Stale Time" in the received Long-lived Graceful Restart Capability. These timers MAY be modified by local configuration.

If the value of the "Restart Time" or the "Long-lived Stale Time" is zero, the duration of the corresponding period would be zero seconds. So, for example, if the "Restart Time" is zero and the "Long-lived Stale Time" is nonzero, only the procedures particular to LLGR would apply. Conversely, if the "Long-lived Stale Time" is zero and the "Restart Time" is nonzero, only the procedures of GR would apply. If both are zero, none of these procedures would apply, only those of the base BGP specification (although EoR would still be used as detailed in [RFC4724]). And finally, if both are nonzero, then the procedures would be applied serially -- first those of GR, then those of LLGR. We observe that during the first interval, while the procedures of GR are in effect, route preference would not be affected, while during the second interval, while LLGR procedures are in effect, routes would be treated as least-preferred as specified elsewhere in this document.

Once the "Restart Time" period ends (including the case that the "Restart Time" is zero), the LLGR period is said to have begun and the following procedures MUST be performed:

- o The helper router MUST start a timer for the "Long-lived Stale Time". If the timer for the "Long-lived Stale Time" expires before the session is re-established, the helper MUST delete all the stale routes from the neighbor that it is retaining.
- o The helper router MUST attach the LLGR\_STALE community for the stale routes being retained. Note that this requirement implies that the routes would need to be readvertised, to disseminate the modified community.

- o If any of the routes from the peer have been marked with the NO\_LLGR community, either as sent by the peer, or as the result of a configured policy, they MUST NOT be retained, but MUST be removed as per the normal operation of [RFC4271].
- o The helper router MUST perform the procedures listed under Section 4.3.

Once the session is re-established, the procedures specified in [RFC4724] apply for the stale routes irrespective of whether the stale routes are retained during the "Restart Time" period or the "Long-lived Stale Time" period. However, in the case of consecutive restarts (i.e., the session goes down before the EoR is received) the previously marked stale routes MUST NOT be deleted before the timer for the "Long-lived Stale Time" expires.

Similarly to [RFC4724], once the session is re-established, if the F bit for a specific address family is not set in the newly received LLGR Capability, or if a specific address family is not included in the newly received LLGR Capability, or if the LLGR and accompanying GR Capability are not received in the re-established session at all, then the Helper MUST immediately remove all the stale routes from the peer that it is retaining for that address family.

If a "Long-lived Stale Time" timer is running for a peer, it MUST NOT be updated (other than by manual operator intervention) until the peer has established and synchronized a new session. The session is termed "synchronized" once the EoR has been received from the peer.

The value of the "Long-lived Stale Time" in the capability received from a neighbor MAY be reduced by local configuration.

While the session is down, the expiration of the "Long-lived Stale Time" timer is treated analogously to the expiration of the "Restart Time" timer in Graceful Restart. However, the timer continues to run once the session has re-established. The timer is not stopped, nor updated, until EoR is received from the peer. If the timer expires during synchronization with the peer, any stale routes that the peer has not refreshed, are removed. If the session subsequently resets prior to becoming synchronized, any remaining routes should be removed immediately.

#### 4.3. Processing LLGR\_STALE Routes

A BGP speaker that has advertised the "Long-lived Graceful Restart Capability" to a neighbor MUST perform the following upon receiving a route from that neighbor with the "LLGR\_STALE" community, or upon attaching the "LLGR\_STALE" community itself per Section 4.2:

- o Treat the route as the least-preferred in route selection (see below). See the Risks of Depreferencing Routes section (Section 5.2) for a discussion of potential risks inherent in doing this.
- o The route SHOULD NOT be advertised to any neighbor from which the Long-lived Graceful Restart Capability has not been received. The exception is described in the Optional Partial Deployment Procedure section (Section 4.7). Note that this requirement implies that such routes should be withdrawn from any such neighbor.
- o The "LLGR\_STALE" community MUST NOT be removed when the route is further advertised.

#### 4.4. Route Selection

In this document, when we refer to treating a route as least-preferred, this means the route MUST be treated as less preferred than any other route that is not so treated. When performing route selection between two routes both of which are least-preferred, normal tie-breaking applies. Note that this would only be expected to happen if the only routes available for selection were least-preferred -- in all other cases, such routes would have been eliminated from consideration.

#### 4.5. Multicast VPN

Special consideration is required if LLGR is to be applied to the Multicast VPN SAFI [RFC6514]. Considerations for Multicast VPNs will be covered in a future revision of this document.

#### 4.6. Errors

If the LLGR capability is received without an accompanying GR capability, the LLGR capability MUST be ignored, that is, the implementation MUST behave as though no LLGR capability had been received.

#### 4.7. Optional Partial Deployment Procedure

Ideally, all routers in an Autonomous System would support this specification before it was enabled. However, to facilitate incremental deployment, stale routes MAY be advertised to neighbors that have not advertised the Long-lived Graceful Restart Capability under the following conditions:

- o The neighbors MUST be internal (IBGP or Confederation) neighbors.
- o The NO\_EXPORT community [RFC1997] MUST be attached to the stale routes.
- o The stale routes MUST have their LOCAL\_PREF set to zero. See the Risks of Depreferencing Routes section (Section 5.2) for a discussion of potential risks inherent in doing this.

If this strategy for partial deployment is used, the network operator should set LOCAL\_PREF to zero for all LLGR routes throughout the Autonomous System. This trades off a small reduction in flexibility (ordering may not be preserved between competing LLGR routes) for consistency between routers which do, and do not, support this specification. Since consistency of route selection can be important for preventing forwarding loops, the latter consideration dominates.

#### 4.8. Procedures When BGP is the PE-CE Protocol in a VPN

In VPN deployments, for example [RFC4364], BGP is often used as a PE-CE protocol. It may be a practical necessity in such deployments to accommodate interoperation with CEs that cannot easily be upgraded to support specifications such as this one. This leads to a problem: in this specification, we take pains to ensure that "stale" routing information will not leak beyond the perimeter of routers that support these procedures, so that it can be depreferenced as expected, and we provide a workaround (Section 4.7) for the case where one or more IBGP routers are not upgraded. However, in the VPN PE-CE case, the protocol in use is EBGP, and our workaround does not work since it relies on the use of LOCAL\_PREF, an IBGP-only path attribute.

We observe that the principal motivation for restricting the propagation of "stale" routing information is the desire to prevent it from spreading without limit once it exits the "safe" perimeter. We further observe that VPN deployments are typically topologically constrained, making this concern moot. For this reason, an implementation MAY advertise stale routes over a PE-CE session, when explicitly configured to do so. That is, the second rule listed in Section 4.3 MAY be disregarded in such cases. All other rules continue to apply. Finally, if this exception is used, the implementation SHOULD by default attach the NO\_EXPORT community to the routes in question, as an additional protection against stale routes spreading without limit. Attachment of the NO\_EXPORT community MAY be disabled by explicit configuration, to accommodate exceptional cases.

See further discussion in Section 5.1.

## 5. Deployment Considerations

The deployment considerations discussed in [RFC4724] apply to this document. In addition, network operators are cautioned to carefully consider the potential disadvantages of deploying these procedures for a given AFI/SAFI. Most notably, if used for an AFI/SAFI that conveys traditional reachability information, use of a long-lived stale route could result in a loss of connectivity for the covered prefix. This specification takes pains to mitigate this risk where possible, by making such routes least-preferred and by restricting the scope of such routes to routers that support these procedures (or, optionally, a single Autonomous System, see "Optional Partial Deployment Procedure", above). However, according to the normal rules of IP forwarding a stale more-specific route, that has no non-stale alternate paths available, will still be used instead of a non-stale less-specific route. Networks in which the deployment of these procedures would be especially concerning include those which do not use "tunneled" forwarding (in other words, those using traditional hop-by-hop forwarding).

Implementations **MUST NOT** enable these procedures by default. They **MUST** require affirmative configuration per AFI/SAFI in order to enable them.

The procedures of this document do not alter the route resolvability requirement of [RFC4271] Section 9.1.2.1.. Because of this, it will commonly be the case that "stale" IBGP routes will only continue to be used if the router depicted in the next hop remains resolvable, even if its BGP component is down. Details of IGP fault-tolerance strategies are beyond the scope of this document. In addition to the foregoing, it may be advisable to check the viability of the next hop through other means, see for example [I-D.ietf-idr-bgp-bestpath-selection-criteria]. This may be especially useful in cases where the next hop is known directly at the network layer, notably EBGP.

As discussed in this document, after a BGP session goes down and before the session is re-established, stale routes may be retained for up to two consecutive periods, controlled by the "Restart Time" and the "Long-lived Stale Time", respectively. During the first period routing churn would be prevented but with potential blackholing of traffic. During the second period potential blackholing of traffic may be reduced but routing churn would be visible throughout the network. The setting of the relevant parameters for a particular application should take into account the tradeoffs, the network dynamics and potential failure scenarios. If needed, the first period can be bypassed either by local configuration or by setting the "Restart Time" in the Graceful

Restart Capability to zero and/or not listing the AFI/SAFI in that Capability.

The setting of the F bit (and the "Forwarding State" bit of the accompanying GR capability) depends in part on deployment considerations. The F bit can be understood as an indication that the Helper should flush associated routes (if the bit is left clear). As discussed in the Introduction, an important use case for LLGR is for routes that are more akin to configuration than to traditional routing. For such routes, it may make sense to always set the F bit, regardless of other considerations. Likewise, for control-plane-only entities such as dedicated route reflectors, that do not participate in the forwarding plane, it makes sense to always set the F bit. Overall, the rule of thumb is that if loss of state on the restarting router can reasonably be expected to cause a forwarding loop or black hole, the F bit should be set scrupulously according to whether state has been retained. Specifics of when the F bit is, and is not, set is implementation-dependent and may also be controlled by configuration.

#### 5.1. When BGP is the PE-CE Protocol in a VPN

As discussed in Section 4.8, it may be necessary to advertise stale routes to a CE in some VPN deployments, even if the CE does not support this specification. In that case, the network operator configuring their PE to advertise such routes should notify the operator of the CE receiving the routes, and the CE should be configured to depreferenciate the routes. Typical BGP implementations will be able to do this by matching on the LLGR\_STALE community, and setting the LOCAL\_PREF for matching routes to zero, similar to the procedure described in Section 4.7.

#### 5.2. Risks of Depreferencing Routes

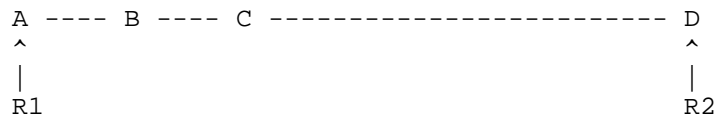
Depreferencing EBGP routes is considered safe, no different from the common practice of applying a routing policy to an EBGP session. However, the same is not always true of IBGP.

Consistent route selection is a fundamental tenet of IBGP correctness and safe operation in hop-by-hop routed networks. When routers within an AS apply different criteria in selecting routes, they can arrive at inconsistent route selections, potentially with the consequence of forming forwarding loops unless some form of tunneled forwarding is used to prevent "core" routers from making a (potentially inconsistent) forwarding decision based on the IP header.

This specification uses the state of a peering session as an input to

the selection criteria, depreferencing routes that are associated with a session that has gone down but have not yet aged out. Since different routers within an AS might have different notions as to whether their respective sessions with a given peer are up or down, they might apply different selection criteria to routes from that peer. This could result in a forwarding loop forming between such routers.

For an example of such a forwarding loop, consider the following simple topology:



In this example, A - D are routers with a full mesh of IBGP sessions between them. The short links have unit cost, the long link has cost 5. Routers A and D are AS border routers, each advertising some route, R, into the AS -- these are denoted R1 and R2 in the diagram. In ordinary operation, it can be seen that routers B and C will select R1 for forwarding, and will forward toward A.

Suppose that the session between A and B goes down for some reason, and stays down long enough for LLGR processing to be invoked on B. Then on B, route R1 will be depreferenced, leading to the selection of R2 by B. However, C will continue to prefer R1. It can be seen that in this case, a forwarding loop for packets destined to R would form between B and C. (We note that other forwarding loop scenarios can be constructed for traditional GR, but are generally considered less severe since GR can remain in effect for a much more limited interval.)

The potential benefits of this specification can outweigh the risks discussed above, as long as care is exercised in deployment. The cardinal rule to be followed is, if a given set of routes are being used within an AS for hop-by-hop forwarding, it is NOT RECOMMENDED to enable LLGR procedures. If tunneled forwarding (such as MPLS) is used within the AS, or if routes are being used for purposes other than hop-by-hop forwarding, less caution is needed, though the operator should still carefully consider the consequences of enabling LLGR.

## 6. Security Considerations

The security implications of the LLGR mechanism defined within in this document are akin to those incurred by the maintenance of stale



routing information within a network. This is particularly relevant when considering the maintenance of routing information that is utilised for service segregation - such as MPLS label entries.

For MPLS VPN services, the effectiveness of the traffic isolation between VPNs relies on the correctness of the MPLS labels between ingress and egress PEs. In particular, when an egress PE withdraws a label L1 allocated to a VPN1 route, this label MUST not be assigned to a VPN route of a different VPN until all ingress PEs stop using the old VPN1 route using L1.

Such a corner case may happen today, if the propagation of VPN routes by BGP messages between PEs takes more time than the label re-allocation delay on a PE. Given that we can generally bound worst case BGP propagation time to a few minutes (for example 2-5), the security breach will not occur if PEs are designed to not reallocate a previous used and withdrawn label before a few minutes.

The problem is made worse with BGP GR between PEs as VPN routes can be stalled for a longer period of time (for example 20 minutes).

This is further aggravated by the BGP LLGR extension proposed in this document as VPN routes can be stalled for a much longer period of time (for example 2 hours, 1 day).

Therefore, to avoid VPN breach, before enabling BGP LLGR, SPs needs to check how fast a given label can be reused by a PE, taking into account:

- o The load of the BGP route churn on a PE (in term of number of VPN label advertised and churn rate).
- o The label allocation policy on the PE (possibly depending upon the size of pool of the VPN labels (which can be restricted by hardware consideration or others MPLS usages), the label allocation scheme (for example per route or per VRF/CE), the re-allocation policy (for example least recently used label...))

Note that [RFC4781] which defines Graceful Restart Mechanism for BGP with MPLS is also applicable to BGP LLGR.

In addition to these considerations, the LLGR mechanism described within this document is considered to be complex to exploit maliciously - in order to inject packets into a topology, there is a requirement to engineer a specific LLGR state between two PE devices, whilst engineering label reallocation to occur in a manner that results in the two topologies overlapping. Such allocation is particularly difficult to engineer (since it is typically an internal

mechanism of an LSR).

## 7. Examples of Operation

For illustrative purposes, we present a few examples of how this specification might be used in practice. These examples are neither exhaustive nor normative.

Consider the following scenario: A border router, ASBR1, has an IBGP peering with a route reflector, RR1, from which it learns routes. It has an EBGP peering with an external peer, EXT, to which it advertises those routes. The external peer has advertised the GR and LLGR Capabilities to ASBR1. ASBR1 is configured to support GR and LLGR on its session with RR1 and EXT. RR1 advertises a GR Restart Time of 1 (second) and a LLST of 3600 (seconds):

Time	Event
t	ASBR1's IBGP session with RR fails. ASBR1 retains RR's routes according to the rules of GR [RFC4724]
t+1	GR Restart Time expires. ASBR1 transitions RR's routes to long-lived stale by attaching the LLGR_STALE community and depreferencing them. However, since it has no backup routes, it continues to make use of them. It re-announces them to EXT with the LLGR_STALE community attached.
t+1+3600	LLST expires. ASBR1 removes RR's stale routes from its own RIB and sends BGP updates to withdraw them from EXT.

Next, imagine the same scenario but suppose RR1 advertised a GR Restart Time of zero, effectively disabling GR. Equally, ASBR1 could have used local configuration to override RR1's offered Restart Time, setting it to a locally-configured value of zero:

Time	Event
t	ASBR1's IBGP session with RR fails. ASBR1 transitions RR's routes to long-lived stale by attaching the LLGR_STALE community and depreferencing them. However, since it has no backup routes, it continues to make use of them. It re-announces them to EXT with the LLGR_STALE community attached.
t+0+3600	LLST expires. ASBR1 removes RR's stale routes from its own RIB and sends BGP updates to withdraw them from EXT.

Next, imagine the original scenario, but consider that the ASBR1-RR1 session comes back up and becomes synchronized 180 seconds after the failure was detected:

Time	Event
t	ASBR1's IBGP session with RR fails. ASBR1 retains RR's routes according to the rules of GR [RFC4724]
t+1	GR Restart Time expires. ASBR1 transitions RR's routes to long-lived stale by attaching the LLGR_STALE community and depreferencing them. However, since it has no backup routes, it continues to make use of them. It re-announces them to EXT with the LLGR_STALE community attached.
t+1+179	Session is reestablished and resynchronized. ASBR1 removes the LLGR_STALE community from RR1's routes and re-announces them to EXT with the LLGR_STALE community removed.

Finally, imagine the original scenario, but consider that EXT has not advertised the LLGR Capability to ASBR1:

Time	Event
t	ASBR1's IBGP session with RR fails. ASBR1 retains RR's routes according to the rules of GR [RFC4724]
t+1	GR Restart Time expires. ASBR1 transitions RR's routes to long-lived stale by attaching the LLGR_STALE community and depreferencing them. However, since it has no backup routes, it continues to make use of them. It withdraws them from EXT.
t+1+3600	LLST expires. ASBR1 removes RR's stale routes from its own RIB.

## 8. Acknowledgements

We would like to thank Roberto Fragassi, John Medamana, Han Nguyen, Jeffrey Haas, Nabil Bitar, Nicolai Leymann, Pranav Mehta, Saikat Ray, Martin Djernaes and Eric Rosen for their valuable inputs and contributions to the discussions and solutions.

## 9. Contributors

Clarence Filsfils  
Cisco Systems  
Brussels 1000  
Belgium

Email: cf@cisco.com

Pradosh Mohapatra  
Cumulus Networks

Email: pmohapat@cumulusnetworks.com

Yakov Rekhter  
Juniper Networks

Email: yakov@juniper.net

Rob Shakir  
BT

Email: rob.shakir@bt.com

Adam Simpson  
Alcatel-Lucent  
600 March Road  
Ottawa, Ontario K2K 2E6  
Canada

Email: adam.simpson@alcatel-lucent.com

## 10. IANA Considerations

This document defines a new BGP capability - Long-lived Graceful Restart Capability. The Capability Code needs to be assigned by IANA.

This document introduces a new BGP community "LLGR\_STALE" for marking the long-lived stale routes, and another community "NO\_LLGR" to indicate that stale routes should not be retained. These community values need to be assigned by IANA.

## 11. References

### 11.1. Normative References

- [I-D.ietf-idr-bgp-gr-notification]  
Patel, K., Fernando, R., Scudder, J., and J. Haas,  
"Notification Message support for BGP Graceful Restart",  
draft-ietf-idr-bgp-gr-notification-01 (work in progress),  
April 2013.
- [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.
- [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.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.
- [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.

#### 11.2. Informative References

- [I-D.ietf-idr-bgp-bestpath-selection-criteria] Asati, R., "BGP Bestpath Selection Criteria Enhancement", draft-ietf-idr-bgp-bestpath-selection-criteria-06 (work in progress), February 2013.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4781] Rekhter, Y. and R. Aggarwal, "Graceful Restart Mechanism for BGP with MPLS", RFC 4781, January 2007.
- [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

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

Email: [jul738@att.com](mailto:jul738@att.com)

Enke Chen  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

Email: enkechen@cisco.com

Bruno Decraene  
Orange  
38-40 Rue de General Leclerc  
92794 Issy Moulineaux cedex 9  
France

Email: bruno.decraene@orange.com

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

Email: jgs@juniper.net





Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 10, 2014

J. Uttaro  
AT&T  
S. Ray  
Cisco Systems  
P. Mohapatra  
Cumulus Networks  
July 09, 2013

One Administrative Domain  
draft-uttaro-idr-oad-00

Abstract

The notional premise that different Autonomous Systems belong to different administrative authorities may not always hold. A single administrative authority may instantiate services on and across multiple ASes. A customer accessing those services can reasonably expect that attributes such as LOCAL\_PREF that influence routing be applicable even across different ASes. This document describes a mechanism to do so.

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 10, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

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 . . . . .	2
1.1. Requirements Language . . . . .	3
1.2. Terminology . . . . .	3
2. Motivation . . . . .	3
2.1. One Administrative Domain . . . . .	3
3. ATTR_SET_STACK attribute . . . . .	6
4. Example Scenarios . . . . .	8
4.1. Single provider scenario . . . . .	8
4.2. Dual provider scenario . . . . .	11
5. Configuration Management . . . . .	12
6. Operational Considerations . . . . .	13
7. Acknowledgments . . . . .	13
8. IANA Considerations . . . . .	13
9. Security Considerations . . . . .	13
10. Normative References . . . . .	14
Authors' Addresses . . . . .	14

## 1. Introduction

One of the basic assumptions of Internet deployment is that different Autonomous Systems (ASes) belong to different administrative authorities that use independent policies. Therefore, attributes such as LOCAL\_PREF are not sent across AS boundary. As networks have evolved, such an assumption may not always hold. A single administrative authority such as a Service Provider (SP) may own equipments in multiple ASes and may instantiate services on and across multiple ASes. As a result, an SP customer's end-points may be connected to multiple ASes even though the customer expects the SP

network to behave as a single "domain". For instance, a customer utilizing LOCAL\_PREF to influence routing expects that the expressed routing preference be preserved at all of their endpoints whether or not they are connected to same or different ASes. This expectation is reasonable since the ASes, being under the same administrative authority, use consistent policies and LOCAL\_PREF set in one AS would be comparable in another AS (when designed to be so). To facilitate such control, this document proposes an approach where non-transitive attributes are tunneled across ASes and are interpreted at traffic ingress points.

### 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. Terminology

One Administrative Domain (OAD):

A collection of autonomous systems (ASes) that are managed by a single administrative entity. They do not appear any different to ASes that belong to a separate administration.

## 2. Motivation

### 2.1. One Administrative Domain

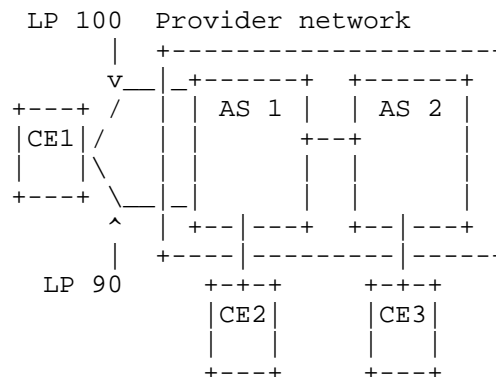


Figure 1: Typical OAD network

Today a large SP network often consists of multiple ASes, for instance, reflecting the SP's internal management structure. The SP

provides services across those ASes to its customers. Some of the sites of a given customer may be connected to one AS whereas some of the other sites of the same customer may be connected to another AS. However, for these customers, the SP network is a single entity. In many instances, the customer desires the routing behavior between two of its sites be uniform whether or not these sites are in the same AS or in different ASes.

Figure 1 provides a typical example of a VPN customer. A customer site with equipment, CE1, is dual-homed to the provider in AS1. A second site of the customer with CE2 is also connected to AS1. A third site of the same customer with CE3 is connected to AS2. CE1 advertises a route. The customer sets different LOCAL\_PREF for its two links to the provider network and thereby chooses one of the links as the primary path. CE2 receives the LOCAL\_PREFs and correctly uses the preferred link for forwarding. However, CE3 doesn't receive the LOCAL\_PREFs since LOCAL\_PREF is not sent across ASes. So CE3 might start to load balance the traffic to CE1 over both links, or might use the non-preferred link solely.

In this scenario, the two ASes are contiguous and under the same administrative domain. So it is desirable that the SP customer be able to use the simple mechanism of setting LOCAL\_PREF to influence routing decisions irrespective of the internal design of the provider network. In other words, it is desirable to make the OAD behave essentially as one AS.

The SP may be able to solve the issue by mapping LOCAL\_PREF to a community in AS1, allowing the community to go across the AS boundary and finally reverse mapping the community to LOCAL\_PREF in AS2. However, an approach like that is narrow in scope and is difficult to manage in a large network.

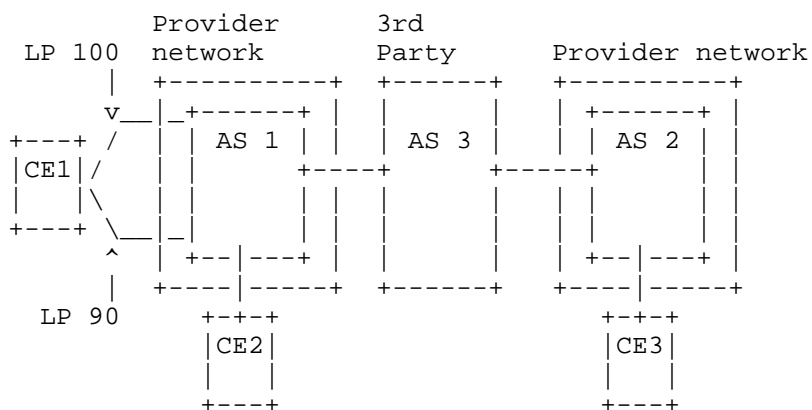


Figure 2: Non-adjacent OAD network

Multiple ASes under the same administrative authority may not always be contiguous. Figure 2 shows a scenario where two ASes, AS1 and AS2, that belong to the same provider, are separated by an AS that is owned by a third party. Such a scenario may arise due to merger of two SPs. While the mechanism proposed in this draft would work in the same way, caution must be exercised in exposing internal parameters of the provider network to a 3rd party transit AS.

We acknowledge that one can consider fixing the problem described here by merging the ASes into one AS (i.e., by renumbering them to one ASN). However, in many cases that is not a viable option. Instead, the solution described here allows an OAD consisting of multiple ASes to essentially behave as a single AS.

### 3. ATTR\_SET\_STACK attribute

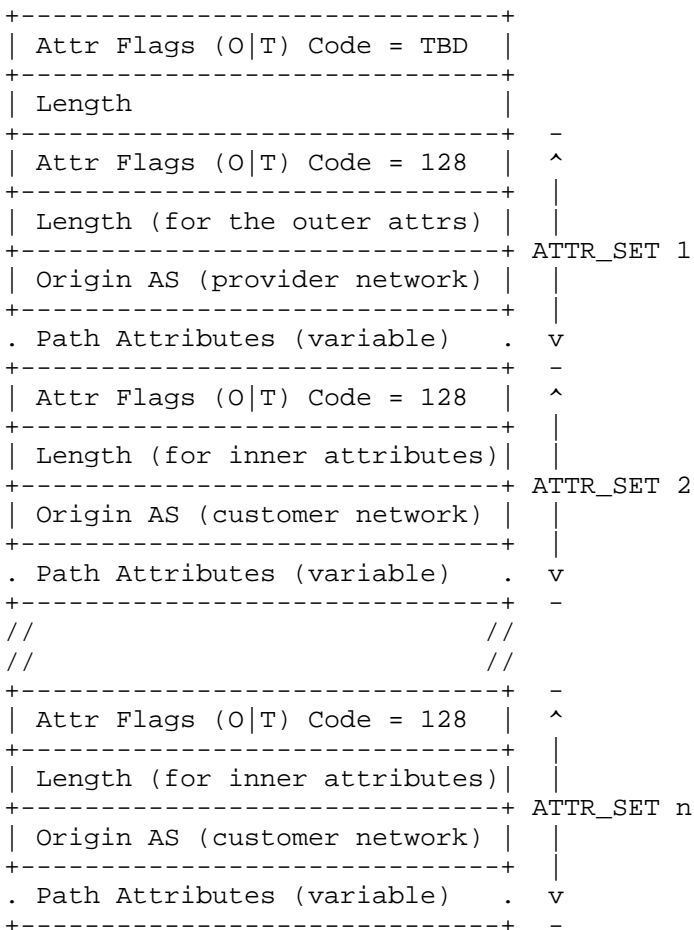


Figure 3: ATTR SET STACK

The problem described in Section 2 arises because non-transitive attributes that crucially influence routing decisions are dropped at AS boundaries. The key idea is to carry these non-transitive attributes to the traffic ingress point. BGP already supports attribute tunneling by using the ATTR\_SET attribute that transparently carries multiple attributes that need to be preserved across AS boundaries ([RFC6368]). However, ATTR\_SET can carry only one set of attributes. As shown in the examples later on, a solution for the present problem needs to carry two sets of attributes, (i) the attribute set for the edge (PE to CE connection, to address the

problem described in [RFC6368]), and (ii) the attribute set for the core (PE to RR connection). Moreover, a mechanism is needed to differentiate the set of attributes for the core from the set of attributes for the edge. Such distinction is needed even if, say, only the attributes for the core is present.

Towards this end, this document generalizes the attribute tunneling mechanism by introducing a new attribute called ATTR\_SET\_STACK that carries multiple ATTR\_SETs by stacking them. This approach allows adding multiple ATTR\_SETs as well as preserves the sequence in which they must be used. The attribute is defined as shown in Figure 3.

The 'Length' field of ATTR\_SET\_STACK includes the cumulative length, in octet, of all the ATTR\_SET attributes.

In this document we define the rules for stacking two ATTR\_SET attributes, which are sufficient for the purpose of OAD. We keep the rules open to future additions to support applications that may require more than two ATTR\_SET attributes.

Rules:

- o When an AS border router (ASBR) advertises a route that doesn't have an ATTR\_SET\_STACK attribute to another AS, if allowed by the policy, the ASBR
  - \* Creates an ATTR\_SET\_STACK attribute,
  - \* "Pushes" any existing ATTR\_SET attribute in the ATTR\_SET\_STACK attribute.
  - \* Encodes the current attributes in an ATTR\_SET and "pushes" this ATTR\_SET in the ATTR\_SET\_STACK attribute.

Thus, when there are edge attributes to tunnel, the ASBR creates an ATTR\_SET\_STACK attribute with two ATTR\_SET attributes in it with the ATTR\_SET for the edge attributes at the bottom. When only core attributes are to be tunneled, it creates an ATTR\_SET\_STACK attribute with one ATTR\_SET attribute in it carrying the core (set by PE) attributes.

- o An ingress PE that imports the route "pops" the top ATTR\_SET attributes from the ATTR\_SET\_STACK. If permitted by the local policy, it uses the attributes from it in its best path selection process.

- o When an ingress PE advertises an imported route to a CE, only the bottom ATTR\_SET element is advertised to it (without any ATTR\_SET\_STACK attribute wrapper).
- o If a router receives a route with an ATTR\_SET\_STACK attribute, and it propagates that route to one of its peers, then if the peer is trusted, the peer receives the route with the same ATTR\_SET\_STACK attribute; otherwise the ATTR\_SET\_STACK is removed from the route.

Note that the creation of ATTR\_SET\_STACK is controlled by local policy (discussed later) and SHOULD be done only for trusted peer ASes.

#### 4. Example Scenarios

In this section, we provide some examples of customer accessing VPN service from a provider to illustrate the difference between the existing behavior and the OAD behavior.

##### 4.1. Single provider scenario

This is a simpler case of a customer connected to only one provider network and there is no edge attribute set.

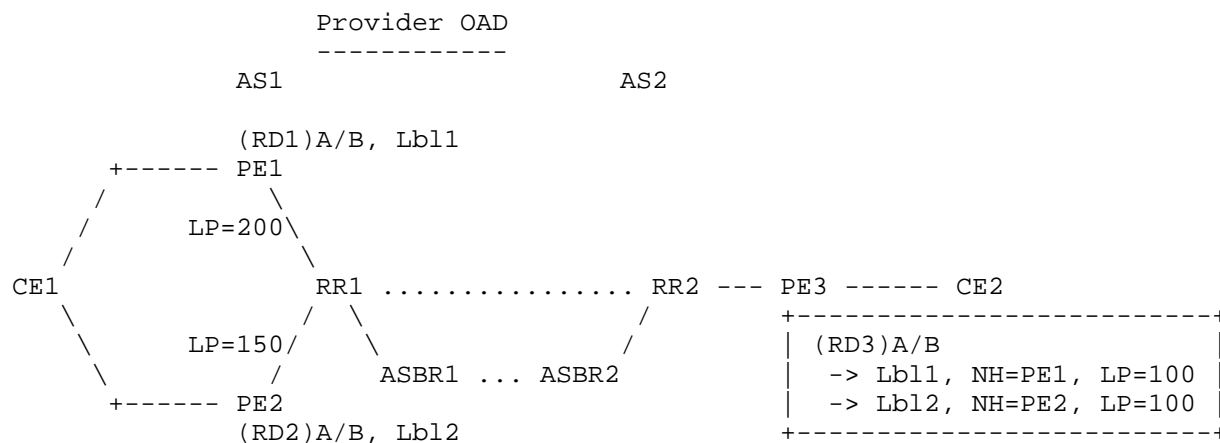


Figure 4: Option C Network (existing behavior)



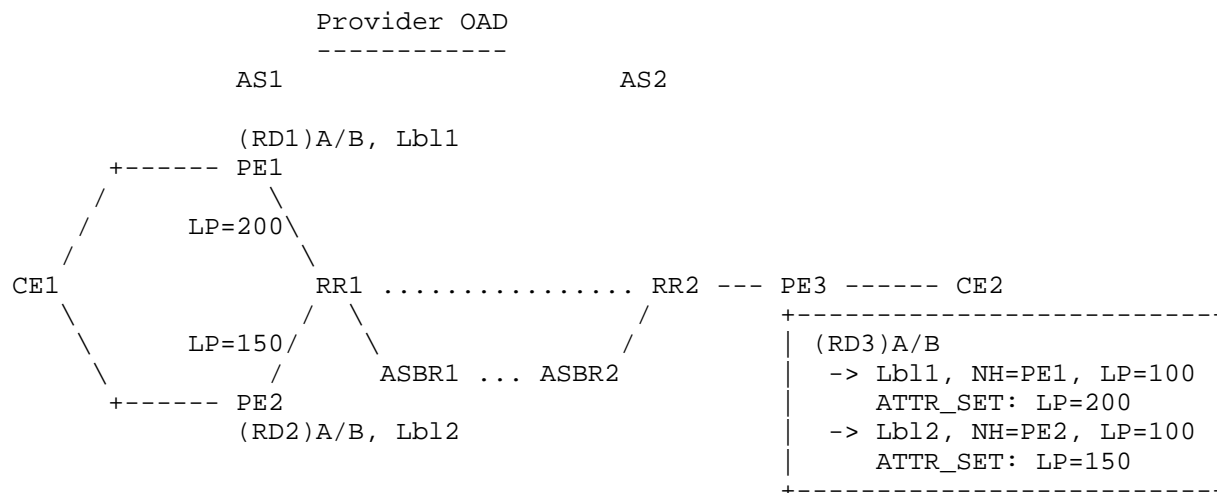


Figure 5: Option C Network (OAD behavior)

As shown in Figure 4, the provider network consisting of two ASes connected by option C technique ([RFC4364]). The customer site with CE1 is dual-homed and advertises prefix A/B to PE1 and PE2. Customer prefers the PE1-CE1 link. This preference is expressed by setting LOCAL\_PREF to 200 on the route advertised by PE1 whereas PE2 sets LOCAL\_PREF to 150. The second customer site with CE2 is connected to PE3 in AS2. Each PE uses a unique RD. So PE3 receives two prefixes: (RD1)A/B and (RD2)A/B, and imports them into (RD3)A/B. Therefore, the prefix (RD3)A/B has two paths. The first path is with nexthop PE1 (in option C, the nexthops remain unchanged), and the second path is with nexthop PE2.

## Existing behavior:

When RR1 sends the routes to RR2, since they are in different ASes, RR1 does not send LOCAL\_PREFs to RR2. So when RR2 sends the routes to PE3, it sends default LOCAL\_PREF (shown as 100). I.e., PE3 loses the route preferences that were set in AS1.

## OAD behavior:

When OAD behavior is turned on on RR1 (and RR2 is added as a trusted peer), when RR1 sends the routes to RR2, it creates an ATTR\_SET\_STACK attribute with one ATTR\_SET in it that contains the LOCAL\_PREF of the route. When PE3 imports the routes into (RD3)A/B, it extracts the LOCAL\_PREFs from the ATTR\_SET\_STACK (which contains only one ATTR\_SET attribute). Therefore, PE3 has both the LOCAL\_PREF set by PE1 and PE2 (coming from the ATTR\_SET\_STACK) and the (default) LOCAL\_PREF set by RR2. As

per the policy set on PE2, the LOCAL\_PREFs coming from AS1 can be used by PE2 for computing best path and hence honor the routing preferences set by the customer. This behavior is depicted in Figure 5.

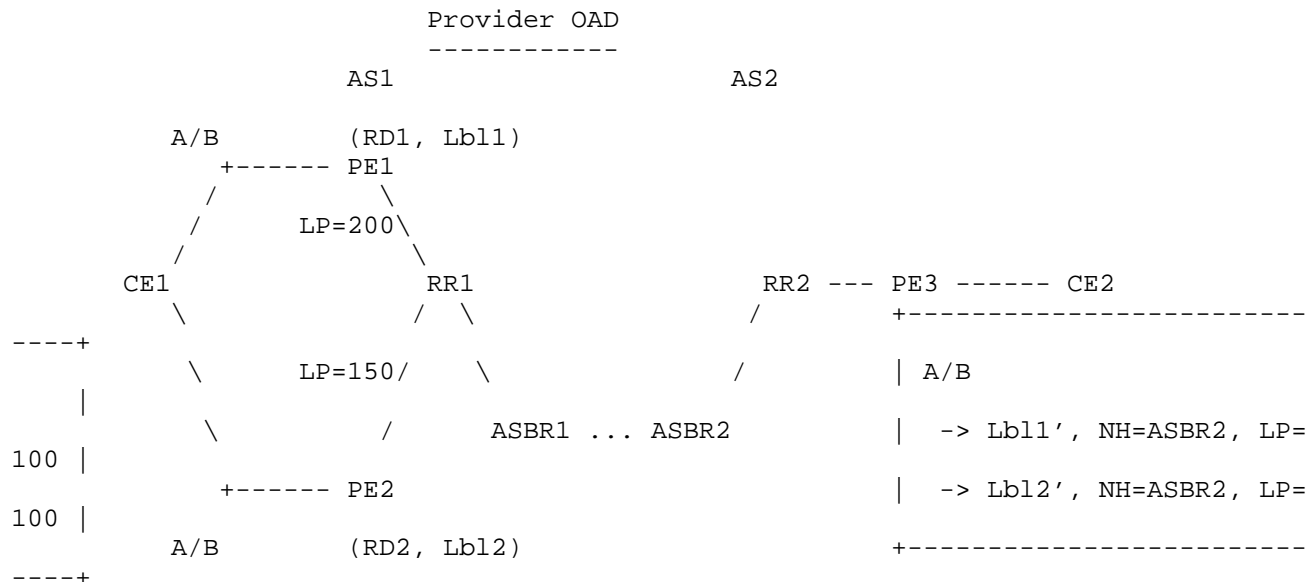


Figure 6: Option B Network (existing behavior)

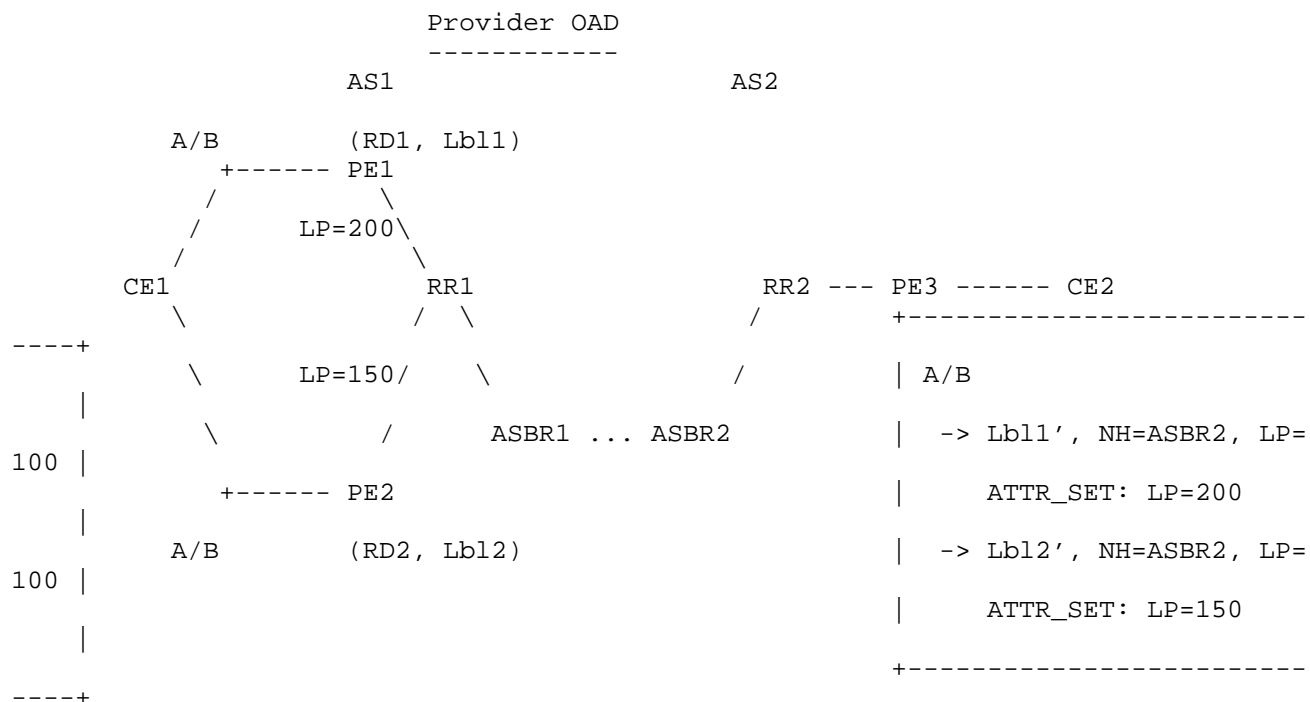


Figure 7: Option B Network (OAD behavior)

Figure 6 shows the same provider network when its two ASes are

connected by option B ([RFC4364]). Similar to the option C case, on PE3, the prefix (RD3)A/B has two paths, but both with nexthop ASBR2.

The VPN label of each route is changed by ASBR2, which allows the packet to ultimately reach PE1 or PE2.

#### Existing behavior:

Similar to option C, ASBR1 does not send LOCAL\_PREFs to ASBR2. So PE3 loses the route preferences that were set in AS1.

#### OAD behavior:

When OAD behavior is turned on on ASBR1 (and ASBR2 is added as a trusted peer), when ASBR1 sends the routes to ASBR2, it creates an ATTR\_SET\_STACK attribute with one ATTR\_SET in it that contains the LOCAL\_PREF of the route. This way PE3 receives both the LOCAL\_PREF set by PE1 and PE2 (coming from the ATTR\_SET\_STACK) and the (default) LOCAL\_PREF set by ASBR2. Therefore PE2 can honor the routing preferences set by the customer.

### 4.2. Dual provider scenario

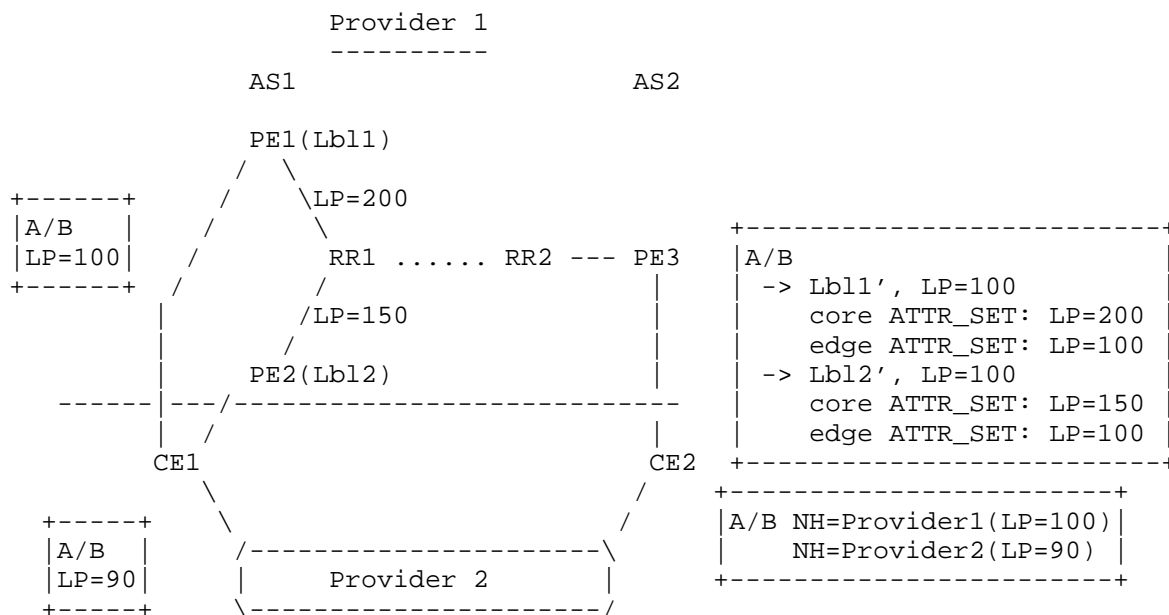


Figure 8: OAD Network in Dual Provider Setup

This example considers the scenario when there is both an edge ATTR\_SET and a core ATTR\_SET. The scenario is shown in Figure 8 where a customer utilizes enterprise VPN service from both Provider 1 and Provider 2. Provider 1 runs an OAD consisting of two ASes, AS1

and AS2, connected by interAS Option B or Option C techniques. To Provider 1, the customer connects one site at AS1 via CE1 and another site at AS2 via CE2. At AS1, CE1 is dual-homed connecting to PE1 and PE2 as IBGP ([RFC6368]) and prefers PE1.

CE1 originates a route, A/B, that it advertises to CE2 via both Provider 1 and Provider 2. CE1 prefers Provider 1 by setting the LOCAL\_PREF attribute to 100 towards Provider 1 and to 90 towards Provider 2. Within Provider 1, since PE1 is preferred by the customer, PE1 advertises A/B to RR1 with LOCAL\_PREF 200 (and label Lbl1) and PE2 advertises A/B with LOCAL\_PREF 150 (and label Lbl2). RR1 preserves both routes since PE1 and PE2 uses different route-distinguishers for the customer VPN route.

In Provider 1's OAD, PE3 receives two routes for A/B: the first one with label Lbl1' and a next-hop that takes the packet to PE1, and the second one with label Lbl2' and a next-hop that takes the packet to PE2.

CE2 receives one route each from Provider 1 (at AS2) and Provider 2. By using the mechanism described in [RFC6368], CE2 sees the LOCAL\_PREF attributes set by CE1 and chooses Provider 1's path and sends traffic to PE3.

Existing behavior:

PE3 does not have any visibility into the LOCAL\_PREFs that PE1 or PE2 has set (as LOCAL\_PREF is non-transitive attribute) and may choose the path with Lbl2' as its bestpath and send traffic to PE2 violating the intent of the customer to receive traffic via PE1.

OAD behavior:

When OAD is turned on, PE3 receives the ATTR\_SET\_STACK attribute containing two ATTR\_SETs: (i) the top ATTR\_SET containing the core attributes (set by PE1 or PE2), (ii) the bottom ATTR\_SET containing the edge attributes that comes from the CE. PE3 extracts the top ATTR\_SET for its own best path computation and sends the bottom ATTR\_SET to CE2. This way PE3 is able to honor the preferences set in AS1.

## 5. Configuration Management

An implementation MUST allow the operator to identify the neighbors that belong to the same OAD, and/or are trusted.

An implementation MUST allow the operator to specify whether the attributes from the ATTR\_SET (within an ATTR\_SET\_STACK) are to be used for best path computation. Note that attributes MUST not be

mixed; i.e., either only the attributes from an ATTR\_SET are used, or no attribute from an ATTR\_SET are used.

## 6. Operational Considerations

When non-transitive attributes such as LOCAL\_PREF are tunneled across AS boundary, the values used for these attributes must be consistent across different ASes in an OAD.

When the originator sends an ATTR\_SET\_STACK attribute to a 3rd party peer AS, even if the peer AS is a transit AS with respect to the provider network, the peer AS may extract the ATTR\_SETs and use them for its own calculations (e.g., if the customer also has a site connected to the 3rd party AS). If the routing policies of the 3rd party AS is not consistent with the originator AS, routing inconsistencies may occur. Therefore, ATTR\_SET\_STACK attribute may be sent to a peer AS only if the peer AS is trusted. In this context, a trusted AS is either in the same OAD, or it is contractually bound to treat the ATTR\_SET\_STACK attribute as an opaque attribute, or its routing policy is consistent with the originator AS.

A route carrying an ATTR\_SET attribute potentially has two sets of non-transitive attributes for possible use: (i) those in the ATTR\_SET, and (ii) those carried by the route. The non-transitive attributes are given a "global" scope when those in the ATTR\_SET are used. Sometimes, however, a "local" scope may be preferred in some ASes in a given OAD, in which case the non-transitive attributes carried by the route are used. Local policy must govern which set of attributes should be used.

## 7. Acknowledgments

## 8. IANA Considerations

IANA shall assign a value from the "BGP Path Attributes" registry, to be called "ATTR\_SET\_STACK", with this document as the reference.

## 9. Security Considerations

The proposed mechanism allows non-transitive attributes to be sent across AS boundary. Sending the non-transitive attributes to non-trusted peers can create routing inconsistencies and other vulnerabilities and MUST not be done.

Procedures and protocol extensions defined in this document do not otherwise affect the BGP security model.

## 10. 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.
- [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.

## Authors' Addresses

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

Email: [uttaro@att.com](mailto:uttaro@att.com)

Saikat Ray  
Cisco Systems  
170 W. Tasman Drive  
San Jose, CA 95134  
USA

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

Pradosh Mohapatra  
Cumulus Networks  
140C S. Whisman Rd  
Mountain View, CA 94041  
USA

Email: [pmohapat@cumulusnetworks.com](mailto:pmohapat@cumulusnetworks.com)

Inter-Domain Routing Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 13, 2014

Q. Wu  
D. Wang  
Huawei  
July 12, 2013

BGP attribute for North-Bound Distribution of Traffic Engineering (TE)  
performance Metric  
draft-wu-idr-te-pm-bgp-00

## Abstract

In order to populate network performance information like link latency, latency variation and packet loss into TED and ALTO server, this document describes extensions to BGP protocol, that can be used to distribute network performance information (such as link delay, delay variation, packet loss, residual bandwidth, and available bandwidth, link utilization, channel throughput).

## 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 13, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of



the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Introduction . . . . .	2
2. Conventions used in this document . . . . .	3
3. Use Cases . . . . .	3
3.1. MPLS-TE with PCE . . . . .	3
3.2. ALTO Server Network API . . . . .	3
4. Carrying TE Performance information in BGP . . . . .	4
5. Attribute TLV Details . . . . .	5
5.1. Link Utilization TLV . . . . .	6
5.2. Channel Throughput TLV . . . . .	7
6. Security Considerations . . . . .	8
7. IANA Considerations . . . . .	8
8. References . . . . .	8
8.1. Normative References . . . . .	8
8.2. Informative References . . . . .	9
Authors' Addresses . . . . .	9

## 1. Introduction

As specified in [RFC4655], a Path Computation Element (PCE) is an entity that is capable of computing a network path or route based on a network graph, and of applying computational constraints during the computation. In order to compute an end to end path, the PCE needs to have a unified view of the overall topology. [I.D-ietf-idr-ls-distribution] describes a mechanism by which links state and traffic engineering information can be collected from networks and shared with external components using the BGP routing protocol. This mechanism can be used by both PCE and ALTO server to gather information about the topologies and capabilities of the network.

With the growth of network virtualization technology, the needs for inter-connecting between various overlay technologies (e.g. Enterprise BGP/MPLS IP VPNs) in the Wide Area Network (WAN) become important. The Network performance or QoS requirements such as latency, limited bandwidth, packet loss, and jitter, are all critical factors that must be taken into account in path computation and selection to establish segment overlay tunnel between overlay nodes and stitch them together to compute end to end path.

In order to populate network performance information like link latency, latency variation and packet loss into TED and ALTO server, this document describes extensions to BGP protocol, that can be used to distribute network performance information (such as link delay, delay variation, packet loss, residual bandwidth, and available bandwidth, link utilization, channel throughput).

## 2. Conventions used in this document

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

## 3. Use Cases

### 3.1. MPLS-TE with PCE

The following figure shows how a PCE can get its TE performance information beyond that contained in the LINK\_STATE attributes [I.D -ietf-idr-ls-distribution] using the mechanism described in this document.

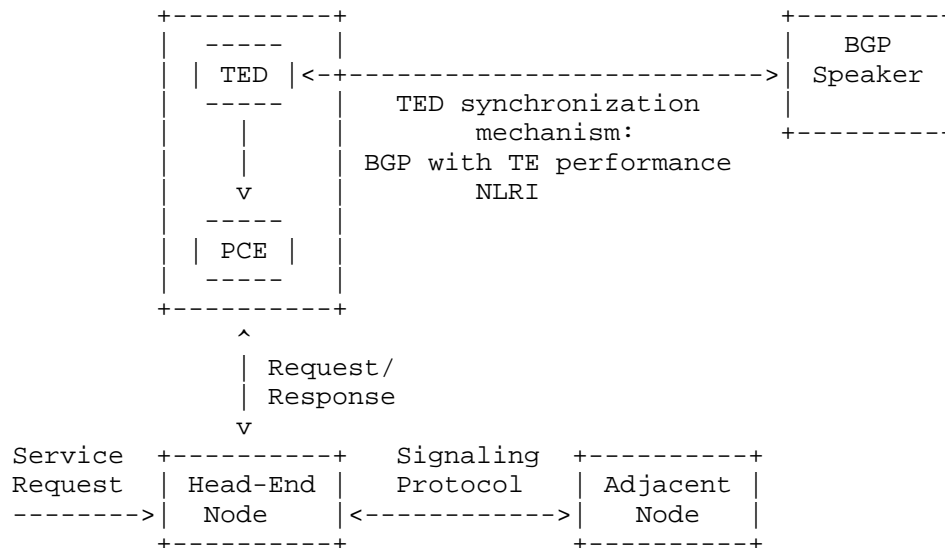


Figure 1: External PCE node using a TED synchronization mechanism

### 3.2. ALTO Server Network API

The following figure shows how an ALTO Server can get TE performance information from the underlying network beyond that contained in the LINK\_STATE attributes [I.D-ietf-idr-ls-distribution] using the mechanism described in this document.

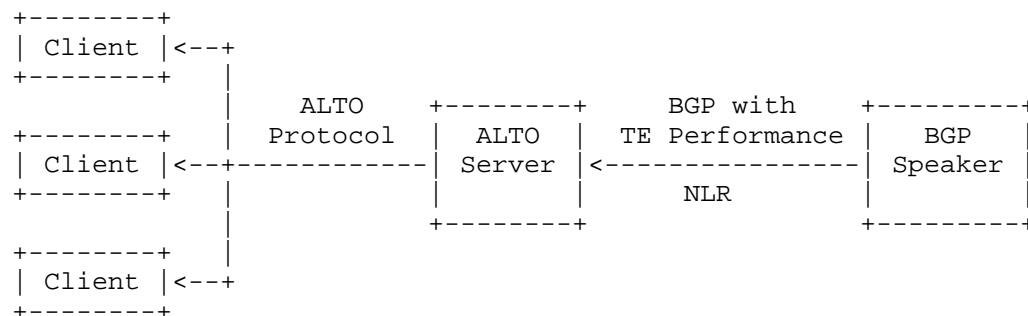


Figure 2: ALTO Server using network performance information

#### 4. Carrying TE Performance information in BGP

This document proposes new BGP TE performance TLVs that can be announced as attribute in the BGP-LS NLRI (defined in [I.D-ietf-idr-ls-distribution]) to distribute network performance information. The extensions in this document build on the ones provided in BGP-LS [I.D-ietf-idr-ls-distribution] and BGP-4 [RFC4271].

BGP-LS NLRI defined in [I.D-ietf-idr-ls-distribution] has nested TLVs which allow the BGP-LS NLRI to be readily extended. This document proposes several additional TLVs as its attributes:

Type	Value
TBD1	Unidirectional Link Delay
TBD2	Unidirectional Delay Variation
TBD3	Unidirectional Packet Loss
TBD4	Unidirectional Residual Bandwidth
TBD5	Unidirectional Available Bandwidth
TBD6	Link Utilization
TBD7	Channel Throughput

As can be seen in the list above, the TLVs described in this document carry different types of network performance information. Many (but not all) of the TLVs include a bit called the Anomalous (or "A") bit. When the A bit is clear (or when the TLV does not include an A bit), the TLV describes steady state link performance. This information could conceivably be used to construct a steady state performance topology for initial tunnel path computation, or to verify alternative failover paths.

When network performance downgrades and falls below configurable link-local thresholds a TLV with the A bit set is advertised. These TLVs could be used by the receiving node to determine whether to redirect failing traffic to a backup path, or whether to calculate an entirely new path. If link performance improves later and exceeds a configurable minimum value (i.e., threshold), that TLV can be re-advertised with the Anomalous bit cleared. In this case, a receiving node can conceivably do whatever re-optimization (or fallback) it wishes to do (including nothing).

Note that when a TLV does not include the A bit, that sub-TLV cannot be used for failover purposes. The A bit was intentionally omitted from some TLVs to help mitigate oscillations.

Consistent with existing ISIS TE specifications [RFC5305][ISIS-TE-METRIC], the bandwidth advertisements defined in this document MUST be encoded as IEEE floating point values. The delay and delay variation advertisements defined in this draft MUST be encoded as integer values. Delay values MUST be quantified in units of microseconds, packet loss MUST be quantified as a percentage of packets sent, and bandwidth MUST be sent as bytes per second. All values (except residual bandwidth) MUST be calculated as rolling averages where the averaging period MUST be a configurable period of time.

## 5. Attribute TLV Details

Link attribute TLVs are TLVs that may be encoded in the BGP-LS attribute with a link NLRI. Each 'Link Attribute' is a Type/Length/Value (TLV) triplet formatted as defined in Section 3.1 of [I-D.ietf-idr-ls-distribution]. The format and semantics of the 'value' fields in some 'Link Attribute' TLVs correspond to the format and semantics of value fields in IS-IS Extended IS Reachability sub-TLVs, defined in [RFC5305] and . Although the encodings for 'Link Attribute' TLVs were originally defined for IS-IS, the TLVs can carry data sourced either by IS-IS or OSPF.

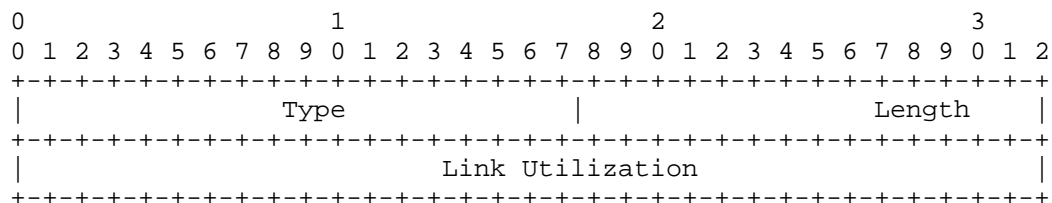
The following 'Link Attribute' TLVs are valid in the LINK\_STATE attribute:

TLV Code Point	Description	IS-IS TLV/Sub-TLV	Defined in:
xxxx	Unidirectional Link Delay	22/xx	[ISIS-TE-METRIC]/4.1
xxxx	Min/Max Unidirectional Link Delay	22/xx	[ISIS-TE-METRIC]/4.2
xxxx	Unidirectional Delay Variation	22/xx	[ISIS-TE-METRIC]/4.3
xxxx	Unidirectional Link Loss	22/xx	[ISIS-TE-METRIC]/4.4
xxxx	Unidirectional Residual Bandwidth	22/xx	[ISIS-TE-METRIC]/4.5
xxxx	Unidirectional Available Bandwidth	22/xx	[ISIS-TE-METRIC]/4.6
xxxx	Link Utilization	----	section 5.1
xxxx	Channel Throughput	----	section 5.2

Table 1: Link Attribute TLVs

### 5.1. Link Utilization TLV

This TLV advertises the average link utilization between two directly connected IS-IS neighbors. The link utilization advertised by this sub-TLV MUST be the utilization percentage per interval from the local neighbor to the remote one. The format of this sub-TLV is shown in the following diagram:



where:

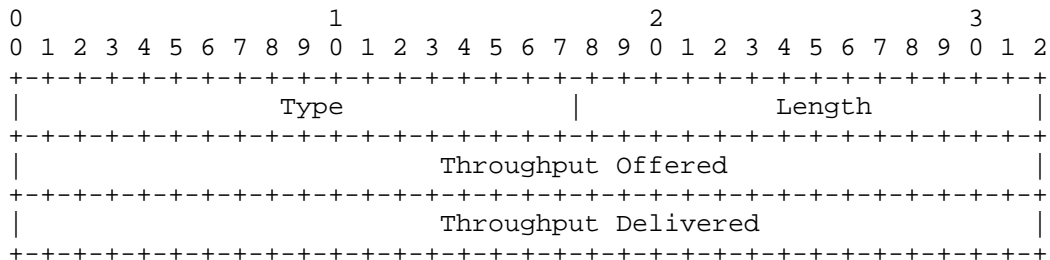
Type: TBA

Length: 4

Link Utilization. This 24-bit field carries the average link utilization over a configurable interval. A commonly used time interval is 5 minutes, and this interval has been sufficient to support network operations and design for some time. link utilization can be calculated by counting the IP-layer (or other layer) octets received over a time interval and dividing by the theoretical maximum number of octets that could have been delivered in the same interval(see section 6.4 of [RFC6703]). If there is no value to send (unmeasured and not statically specified), then the sub-TLV should not be sent or be withdrawn.

## 5.2. Channel Throughput TLV

This TLV advertises the average Channel Throughput between two directly connected IS-IS neighbors. The channel throughput advertised by this sub-TLV MUST be the throughput between the local neighbor and the remote one. The format of this sub-TLV is shown in the following diagram:



where:

Type: TBA

Length: 8

Throughput offered: This 24-bit field carries the average throughput offered over a configurable interval. Throughput offered can be calculated by counting the number of units successfully transmitted in the interval (See section 2.3 of [RFC6374]). If there is no value to send (unmeasured and not statically specified), then

the sub-TLV should not be sent or be withdrawn.

Throughput delivered: This 24-bit field carries the average throughput delivered over a configurable interval. Throughput delivered can be calculated by counting the number of units successfully received in the interval (See section 2.3 of [RFC6374]). If there is no value to send (unmeasured and not statically specified), then the sub-TLV should not be sent or be withdrawn.

## 6. Security Considerations

This document does not introduce security issues beyond those discussed in [I.D-ietf-idr-ls-distribution] and [RFC4271].

## 7. IANA Considerations

IANA maintains the registry for the TLVs. BGP TE Performance TLV will require one new type code per TLV defined in this document.

## 8. References

### 8.1. Normative References

- [I-D.ietf-idr-ls-distribution]  
Gredler, H., "North-Bound Distribution of Link-State and TE Information using BGP", ID draft-ietf-idr-ls-distribution-03, May 2013.
- [ISIS-TE-METRIC]  
Giacalone, S., "ISIS Traffic Engineering (TE) Metric Extensions", ID draft-ietf-isis-te-metric-extensions-00, June 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", March 1997.
- [RFC4271] Rekhter, Y., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC5305] Li, T., "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC6374] Frost, D., "Packet Loss and Delay Measurement for MPLS Networks ", RFC 6374, September 2011.

[RFC6703] Morton, A., "Reporting IP Network Performance Metrics: Different Points of View ", RFC 6703, August 2012.

## 8.2. Informative References

[ALTO] Yang, Y., "ALTO Protocol", ID <http://tools.ietf.org/html/draft-ietf-alto-protocol-16>, May 2013.

[RFC4655] Farrel, A., "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.

## Authors' Addresses

Qin Wu  
Huawei  
101 Software Avenue, Yuhua District  
Nanjing, Jiangsu 210012  
China

Email: [sunseawq@huawei.com](mailto:sunseawq@huawei.com)

Danhua Wang  
Huawei  
101 Software Avenue, Yuhua District  
Nanjing, Jiangsu 210012  
China

Email: [wangdanhua@huawei.com](mailto:wangdanhua@huawei.com)



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: January 16, 2014

Z. Li  
L. Zhang  
Huawei Technologies  
July 15, 2013

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

## Abstract

As the BGP is deployed in a single Autonomous System for network convergence such as Seamless MPLS, it is desirable for BGP to carry more information to help select routing more intelligently. It can reduce the cost proposed by complex policy control design on BGP routes and adapt to network change easily. This document proposed a new path attribute for BGP routes that can record the next hop path for the route to help BGP route selection and network management.

## 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 16, 2014.

## Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

## Table of Contents

1. Introduction . . . . .	2
2. Motivation . . . . .	3
2.1. Complexity of Route Selection . . . . .	3
2.2. New Role of BGP in Seamless MPLS Network . . . . .	4
3. Definition of NEXTHOP_PATH ATTRIBUTE . . . . .	4
4. Process of NEXTHOP_PATH ATTRIBUTE . . . . .	5
4.1. Creating and Modifying the NEXTHOP_PATH Attribute . . . . .	5
4.2. Decision Process . . . . .	6
5. IANA Considerations . . . . .	7
6. Security Considerations . . . . .	7
7. Normative References . . . . .	7
Authors' Addresses . . . . .	8

## 1. Introduction

[I-D.ietf-mpls-seamless-mpls] describes an architecture which can be used to extend MPLS networks to integrate access and aggregation networks into a single MPLS domain ("Seamless MPLS"). As the mobile backhaul service is deployed widely, the requirement of the integration of mobile backhaul networks and core networks has been proposed. For the reason of scalability, the Seamless MPLS network tends to be divided into multiple IGP areas for access, aggregation, and core networks and IBGPs runs among Area Border Routers (ABRs) which should act as inline RRs to reflect the labeled BGP routes or BGP VPN routes to remote BGP peers with next hop self (NHS).

As the BGP is used in a single Autonomous System for network convergence, it is desirable for BGP to carry more information to help select routing more intelligently. It can reduce the cost proposed by complex policy control design on BGP routes and adapt to network change easily.

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.

## 2. Motivation

### 2.1. Complexity of Route Selection

In the Seamless MPLS network, Area Border Routers (ABRs) which run IBGP should act as inline RRs to reflect the labeled BGP routes or BGP VPN routes to remote BGP peers with next hop self (NHS). Each ABR should process route selection which needs complex route policy to control the BGP route distribution in the Seamless MPLS network , as shown below:

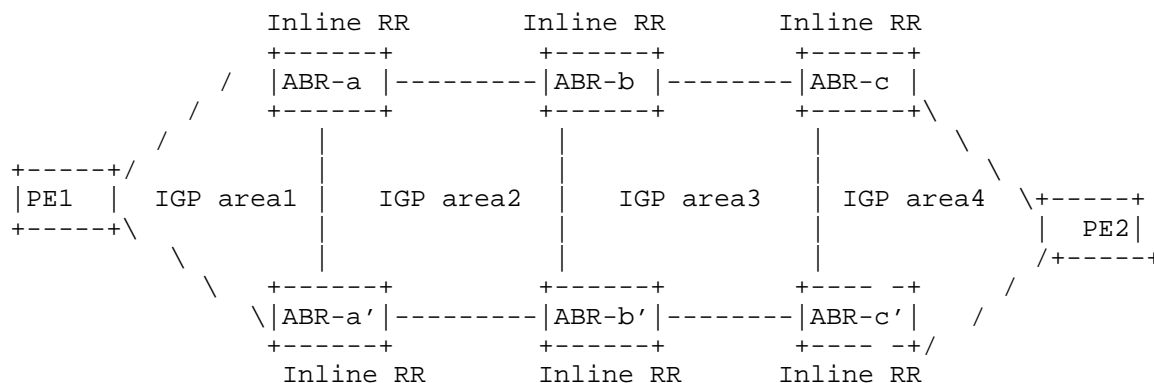


Figure 1 Seamless MPLS Network with Multiple IGP Areas

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.

Here is an example for complex route policy. ABR-b may receive routes of the same prefix originated from PE1 from different three IBGP peers, ABR-a, ABR-a', ABR-c, and ABR-c'. The route policy should guarantee that the route from ABR-a or ABR-a' is selected as

the best one. At the same time, routes of the same prefix originated from PE2 may be received from ABR-a, ABR-a', ABR-c, and ABR-c'. The route policy should guarantee that the route from ABR-c or ABR-c' is selected as the best one. To satisfy the different best route selection requirements, each IBGP speaker has to configure complex route policy.

## 2.2. New Role of BGP in Seamless MPLS Network

When Seamless MPLS makes integration of mobile backhaul networks and core networks, BGP in Seamless MPLS network act more like an "Interior Gateway Protocol (IGP)". As the whole Seamless MPLS network is in a single AS for uniform administration, the security requirement proposed for traditional BGP can be reduced. At the same time some path attributes for BGP route such as AS\_PATH is no use in this scenario. As the BGP is deployed from implementing network convergence, it is desirable for BGP to carry more information to help select route more intelligently. It can simplify policy control design on BGP routes and adapt to network change easily. Moreover the additional path information may facilitate the network operation and maintenance. [I-D.ietf-idr-aigbp] is the example which can help BGP route selection by advertising IGP metric information with BGP route. In this document, we propose a new method to record next hop list for the BGP route, which can be used for automatic BGP route selection and facilitating network operation and maintenance. The new attribute, NEXTHOP\_PATH ATTRIBUTE, is defined for the BGP route to record the next hop path. It can work as AS\_PATH ATTRIBUTE.

## 3. Definition of NEXTHOP\_PATH ATTRIBUTE

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

```

0                               1
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Attr. Flags |Attr. Type Code|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Figure 2 NEXTHOP\_PATH ATTRIBUTE Type definition

Attr. Flags

SHOULD be optional transitive

Attr. Type Code

NEXTHOP\_PATH 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 segmen is shown in the figure 3.



- 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 version.
- 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.

#### 4. Process of NEXTHOP\_PATH ATTRIBUTE

The NEXTHOP\_PATH ATTRIBUTE defined here is an optional transitive BGP Path Attribute, the process of this attribute MUST accord with the procedures in [RFC4271].

#### 4.1. Creating and Modifying the NEXTHOP\_PATH Attribute

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

1. If the route is originated in this BGP peaker
  - \* If the NEXTHOP\_PATH ATTRIBUTE is supported, the NEXTHOP\_PATH 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 ATTRIBUTE.

- \* If the NEXTHOP\_PATH ATTRIBUTE is not supported, the route will be distributed without NEXTHOP\_PATH ATTRIBUTE.

2. if the route is received from one BGP speaker's UPDATE message

- \* If the NEXTHOP\_PATH ATTRIBUTE is NULL and the local BGP speaker support NEXTHOP\_PATH ATTRIBUTE, when the route is propagated to another IBGP speaker with next hop self (NHS ), the NEXTHOP\_PATH 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 ATTRIBUTE
- \* If the NEXTHOP\_PATH ATTRIBUTE is non-NULL and the local BGP speaker support NEXTHOP\_PATH 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 ATTRIBUTE is NULL and the local BGP speaker support NEXTHOP\_PATH 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 ATTRIBUTE.
- \* If the NEXTHOP\_PATH ATTRIBUTE is non-NULL and the local BGP speaker support NEXTHOP\_PATH 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 ATTRIBUTE, it SHOULD keep the NEXTHOP\_PATH ATTRIBUTE unchanged whether the route is distribute with next hop self or not.

4.2. Decision Process

Support for the NEXTHOP\_PATH ATTRIBUTE involves several modifications to the tie breaking procedures of the "phase 2" decision of BGP route selection, described in section 9.1.2.2 of [RFC4271].

If the NEXTHOP\_PATH ATTRIBUTE of a BGP route contains a next hop path loop, the BGP route MUST be excluded from the Phase 2 decision

function. The next hop path loop detection is done by scanning the full next hop path (as specified in the NEXTHOP\_PATH ATTRIBUTE), and checking if the local BGP speaker appears in the next hop path.

The NEXTHOP\_PATH ATTRIBUTE can be used for BGP route selection. The priority of the NEXTHOP\_PATH ATTRIBUTE for route selection is the same as the AS\_PATH attribute.

When a route is received from different IBGP speakers, if the best route cannot be acquired through the higher priority rules, the NEXTHOP\_PATH ATTRIBUTE SHOULD be used for route selection, and the route with least nexthops will be selected. If the lengths of the next hop lists are the same, the rest rules SHOULD be used for route selection.

## 5. IANA Considerations

IANA need to assign the codepoint in the "BGP Path Attributes" registry to the NEXTHOP\_PATH 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 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 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 ATTRIBUTE, it can parse the NEXTHOP\_PATH ATTRIBUTE to get the inner network architecture of the other network.

In order to prevent this possible security problem, the NEXTHOP\_PATH ATTRIBUTE capability should be disabled for specific BGP speaker, such as EBGp. This can reduce the security risk.

## 7. Normative References

[I-D.ietf-idr-aigp]

Mohapatra, P., Fernando, R., Rosen, E., and J. Uttaro,  
"The Accumulated IGP Metric Attribute for BGP", draft-  
ietf-idr-aigp-10 (work in progress), May 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-03 (work in progress), May 2013.

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

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

#### 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