

Internet Engineering Task Force
Internet-Draft
Updates: 5575 (if approved)
Intended status: Standards Track
Expires: May 9, 2014

J. Haas, Ed.
Juniper Networks
November 5, 2013

Clarification of the Flowspec Redirect Extended Community
draft-haas-idr-flowspec-redirect-rt-bis-00

Abstract

This document clarifies the formatting of the the BGP Flowspec Redirect Extended Community, originally documented in RFC 5575.

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 May 9, 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. IANA Considerations	4
3. Security Considerations	4
4. Acknowledgements	4
5. Normative References	5
Author's Address	5

1. Introduction

Dissemination of Flow Specification Rules [RFC5575], commonly known as BGP Flowspec, provided for a BGP Extended Community [RFC4360] that served to redirect traffic to a VRF routing instance that matched the flow specification NLRI. In that RFC, the Redirect Extended Community was documented as follows:

```
: +-----+-----+-----+
: | type   | extended community | encoding |
: +-----+-----+-----+
: | 0x8008 | redirect           | 6-byte Route Target |
: +-----+-----+-----+
:
: [...]
:
: Redirect: The redirect extended community allows the traffic to be
: redirected to a VRF routing instance that lists the specified
: route-target in its import policy. If several local instances
: match this criteria, the choice between them is a local matter
: (for example, the instance with the lowest Route Distinguisher
: value can be elected). This extended community uses the same
: encoding as the Route Target extended community [RFC4360].
: [...]
:
: 11. IANA Considerations
: [...]
:
: The following traffic filtering flow specification rules have been
: allocated by IANA from the "BGP Extended Communities Type -
: Experimental Use" registry as follows:
: [...]
:
: 0x8008 - Flow spec redirect
```

The IANA registry of BGP Extended Communities clearly identifies communities of specific formats. For example, "Two-octet AS Specific Extended Community" [RFC4360], "Four-octet AS Specific Extended Community" [RFC5668] and "IPv4 Address Specific Extended Community" [RFC4360]. Route Targets [RFC4360] identify this format in the high-order (Type) octet of the Extended Community and set the value of the low-order (Sub-Type) octet to 0x02. The Value field of the Route Target Extended Community is intended to be interpreted in the context of its format.

Since the Redirect Extended Community only registered a single code-point in the IANA BGP Extended Community registry, a common interpretation of the Redirect Extended Community's "6-byte Route

Target" has been to look for any matching Route Target sharing the same Value portion of its Extended Community. Thus, multiple Route Targets provisioned in a router's VRFs might match even though the format was different.

This "Value wildcard" behavior does not matched deployed implementations of BGP Flowspec. Deployed implementations of BGP Flowspec use the following formatting for the Redirect Extended Community:

type	extended community	encoding
0x8008	redirect AS-2byte	2-octet AS, 4-octet Value
0x8108	redirect IPv4	4-octet IPv4 Address, 2-octet Value
0x8208	redirect AS-4byte	4-octet AS, 2-octet Value

It should be noted that the low-order nybble of the Redirect's Type field corresponds to the Route Target Extended Community format field (Type). (See [RFC4360], Secs. 3.1, 3.2 and [RFC5668], Sec. 2.) The low order octet (Sub-Type) of the Redirect Extended Community remains 0x08, contrasted to 0x02 for Route Targets.

2. IANA Considerations

IANA is requested to update the "BGP Extended Communities Type - Experimental Use" registry as follows:

```
0x8008 - Flow spec redirect AS-2byte
0x8108 - Flow spec redirect IPv4
0x8208 - Flow spec redirect AS-4byte
```

3. Security Considerations

This document introduces no additional security considerations than those already covered in [RFC5575].

4. Acknowledgements

The contents of this document was raised as part of implementation discussions of BGP Flowspec with the following individuals:

Andrew Karch (Cisco)

Robert Raszuk (NTT I3)

Adam Simpson (Alcatel-Lucent)

Matthieu Texier (Arbor Networks)

Kaliraj Vairavakkalai (Juniper)

5. Normative References

- [RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.
- [RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, August 2009.
- [RFC5668] Rekhter, Y., Sangli, S., and D. Tappan, "4-Octet AS Specific BGP Extended Community", RFC 5668, October 2009.

Author's Address

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

Email: jhaas@juniper.net

Network Working Group
Internet Draft
Intended Status: Standards Track
Expiration Date: April 17, 2014

D. Walton
A. Retana
E. Chen
Cisco Systems
J. Scudder
Juniper Networks
October 16, 2013

Advertisement of Multiple Paths in BGP

draft-ietf-idr-add-paths-09.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on April 17, 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.

Abstract

In this document we propose a BGP extension that allows the advertisement of multiple paths for the same address prefix without the new paths implicitly replacing any previous ones. The essence of the extension is that each path is identified by a path identifier in addition to the address prefix.

1. Introduction

The BGP specification [RFC4271] defines an "Update-Send Process" to advertise the routes chosen by the Decision Process to other BGP speakers. No provisions are made to allow the advertisement of multiple paths for the same address prefix, or Network Layer Reachability Information (NLRI). In fact, a route with the same NLRI as a previously advertised route implicitly replaces the previous advertisement.

In this document we propose a BGP extension that allows the advertisement of multiple paths for the same address prefix without the new paths implicitly replacing any previous ones. The essence of the extension is that each path is identified by a path identifier in addition to the address prefix.

1.1. Specification of Requirements

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

2. How to Identify a Path

As defined in [RFC4271], a path refers to the information reported in the path attribute field of an UPDATE message. As the procedures specified in [RFC4271] allow only the advertisement of one path for a particular address prefix, a path for an address prefix from a BGP peer can be keyed on the address prefix.

In order for a BGP speaker to advertise multiple paths for the same address prefix, a new identifier (termed "Path Identifier" hereafter) needs to be introduced so that a particular path for an address prefix can be identified by the combination of the address prefix and the Path Identifier.

The assignment of the Path Identifier for a path by a BGP speaker is purely a local matter. However, the Path Identifier MUST be assigned in such a way that the BGP speaker is able to use the (prefix, path identifier) to uniquely identify a path advertised to a neighbor. A BGP speaker that re-advertises a route MUST generate its own Path Identifier to be associated with the re-advertised route. A BGP speaker that receives a route SHOULD NOT assume that the identifier carries any particular semantics; it SHOULD be treated as an opaque value.

3. Extended NLRI Encodings

In order to carry the Path Identifier in an UPDATE message, the existing NLRI encodings are extended by prepending the Path Identifier field, which is of four-octets.

For example, the NLRI encodings specified in [RFC4271, RFC4760] are extended as the following:

```
+-----+
| Path Identifier (4 octets) |
+-----+
| Length (1 octet)         |
+-----+
| Prefix (variable)        |
+-----+
```

and the NLRI encoding specified in [RFC3107] is extended as the following:

```
+-----+
| Path Identifier (4 octets) |
+-----+
| Length (1 octet)         |
+-----+
| Label (3 octets)         |
+-----+
| ...                       |
+-----+
| Prefix (variable)        |
+-----+
```

The usage of the extended NLRI encodings is specified in the Operation section.

4. ADD-PATH Capability

The ADD-PATH Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is specified in the IANA Considerations section of this document. The Capability Length field of this capability is variable. The Capability Value field consists of one or more of the following tuples:

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

The meaning and use of the fields are as follows:

Address Family Identifier (AFI):

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

Subsequent Address Family Identifier (SAFI):

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

Send/Receive:

This field indicates whether the sender is (a) willing to

receive multiple paths from its peer (value 1), (b) would like to send multiple paths to its peer (value 2), or (c) both (value 3) for the <AFI, SAFI>.

5. Operation

The Path Identifier specified in the previous section can be used to advertise multiple paths for the same address prefix without subsequent advertisements replacing the previous ones. Apart from the fact that this is now possible, the route advertisement rules of [RFC4271] are not changed. In particular, a new advertisement for a given address prefix and a given path identifier replaces a previous advertisement for the given address prefix and the given path identifier.

A BGP speaker that is willing to receive multiple paths from its peer, or would like to send multiple paths to its peer, SHOULD advertise the ADD-PATH Capability to the peer using BGP Capabilities advertisement [RFC5492].

A BGP speaker MUST follow the existing procedures in generating an UPDATE message for a particular <AFI, SAFI> to a peer unless the BGP speaker advertises the ADD-PATH Capability to the peer indicating its desire to send multiple paths for the <AFI, SAFI>, and also receives the ADD-PATH Capability from the peer indicating its willingness to receive multiple paths for the <AFI, SAFI>, in which case the speaker MUST generate a route update for the <AFI, SAFI> based on the combination of the address prefix and the Path Identifier, and use the extended NLRI encodings specified in this document. The peer SHALL act accordingly in processing an UPDATE message related to a particular <AFI, SAFI>.

A BGP speaker SHOULD include the bestpath when more than one path are advertised to a neighbor unless the bestpath is a path received from that neighbor.

When deployed as a provider edge router or a peering router that interacts with external neighbors, a BGP speaker usually advertises at most one path to the internal neighbors in a network. In the case the speaker is configured to advertise multiple paths to the internal neighbors, it should include the Edge_Discriminator attribute defined in [FAST-CONV] in order to make the route selection consistent inside the network.

As the Path Identifiers are locally assigned, and may or may not be persistent across a control plane restart of a BGP speaker, an implementation SHOULD take special care so that the underlying

forwarding plane of a "Receiving Speaker" as described in [RFC4724] is not affected during the graceful restart of a BGP session.

6. Applications

The BGP extension specified in this document can be used by a BGP speaker to advertise multiple paths in certain applications. The availability of the additional paths can help reduce or eliminate persistent route oscillations [RFC3345]. It can also help with optimal routing and routing convergence in a network. The applications are detailed in separate documents.

7. Deployment Considerations

The extension proposed in this document provides a mechanism for a BGP speaker to advertise multiple paths over a BGP session. Care needs to be taken in its deployment to ensure consistent routing and forwarding in a network, the details of which will be described in separate application documents.

8. IANA Considerations

IANA has assigned capability number 69 for the ADD-PATH Capability described in this document. This registration is in the BGP Capability Codes registry.

9. Security Considerations

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

10. Acknowledgments

We would like to thank David Cook and Naiming Shen for their contributions to the design and development of the extension.

Many people have made valuable comments and suggestions, including Rex Fernando, Eugene Kim, Danny McPherson, Dave Meyer, Pradosh Mohapatra, Keyur Patel, Robert Raszuk, Eric Rosen, Srihari Sangli, Dan Tappan, and Mark Turner.

11. References

11.1. Normative References

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

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.

[RFC4760] Bates, T., Chandra, R., Rekhter, Y., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC3107] Rekhter, R. and E. Rosen, "Carrying Label Information in BGP-4," RFC 3107, May 2001.

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

[RFC4724] Sangli, S., E. Chen, R. Fernando, J. Scudder, and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, January 2007.

[FAST-CONV] Mohapatra, P., R. Fernando, C. Filsfils, R. Raszuk, "Fast Connectivity Restoration Using BGP Add-path", Work in Progress, March 2011.

11.2. Informative References

[RFC3345] McPherson, D., V. Gill, D. Walton, and A. Retana, "Border Gateway Protocol (BGP) Persistent Route Oscillation Condition", RFC 3345, August 2002.

12. Authors' Addresses

Daniel Walton

Email: dwalton76@gmail.com

Alvaro Retana
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709

Email: aretana@cisco.com

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

Email: enkechen@cisco.com

John Scudder
Juniper Networks

Email: jgs@juniper.net

Network Working Group
Internet Draft
Intended Status: Proposed Standard
Expires: November 23, 2013

Pradosh Mohapatra
Cumulus Networks

Rex Fernando
Eric C. Rosen
Cisco Systems, Inc.

James Uttaro
ATT

May 23, 2013

The Accumulated IGP Metric Attribute for BGP

draft-ietf-idr-aigp-10.txt

Abstract

Routing protocols that have been designed to run within a single administrative domain ("IGPs") generally do so by assigning a metric to each link, and then choosing as the installed path between two nodes the path for which the total distance (sum of the metric of each link along the path) is minimized. BGP, designed to provide routing over a large number of independent administrative domains ("autonomous systems"), does not make its path selection decisions through the use of a metric. It is generally recognized that any attempt to do so would incur significant scalability problems, as well as inter-administration coordination problems. However, there are deployments in which a single administration runs several contiguous BGP networks. In such cases, it can be desirable, within that single administrative domain, for BGP to select paths based on a metric, just as an IGP would do. The purpose of this document is to provide a specification for doing so.

Status of this Memo

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

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

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

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

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

Copyright and License Notice

Copyright (c) 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
2	Introduction	3
3	AIGP Attribute	5
3.1	Applicability Restrictions and Cautions	6
3.2	Restrictions on Sending/Receiving	6
3.3	Creating and Modifying the AIGP Attribute	7
3.3.1	Originating the AIGP Attribute	7
3.3.2	Modifications by the Originator	8
3.3.3	Modifications by a Non-Originator	8
4	Decision Process	10
4.1	When a Route has an AIGP Attribute	10
4.2	When the Route to the Next Hop has an AIGP attribute ..	11
5	Deployment Considerations	12
6	IANA Considerations	12
7	Security Considerations	12
8	Acknowledgments	12
9	Authors' Addresses	13
10	Normative References	13
11	Informative References	13

1. Specification of requirements

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

2. Introduction

There are many routing protocols that have been designed to run within a single administrative domain. These are known collectively as "Interior Gateway Protocols" (IGPs). Typically, each link is assigned a particular "metric" value. The path between two nodes can then be assigned a "distance", which is the sum of the metrics of all the links that belong to that path. An IGP selects the "shortest" (minimal distance) path between any two nodes, perhaps subject to the constraint that if the IGP provides multiple "areas", it may prefer the shortest path within an area to a path that traverses more than one area. Typically the administration of the network has some

routing policy which can be approximated by selecting shortest paths in this way.

BGP, as distinguished from the IGPs, was designed to run over an arbitrarily large number of administrative domains ("autonomous systems", or "ASes") with limited coordination among the various administrations. BGP does not make its path selection decisions based on a metric; there is no such thing as an "inter-AS metric". There are two fundamental reasons for this:

- The distance between two nodes in a common administrative domain may change at any time due to events occurring in that domain. These changes are not propagated around the Internet unless they actually cause the border routers of the domain to select routes with different BGP attributes for some set of address prefixes. This accords with a fundamental principle of scaling, viz., that changes with only local significance must not have global effects. If local changes in distance were always propagated around the Internet, this principle would be violated.
- A basic principle of inter-domain routing is that the different administrative domains may have their own policies, which do not have to be revealed to other domains, and which certainly do not have to be agreed to by other domains. Yet the use of inter-AS metric in the Internet would have exactly these effects.

There are, however, deployments in which a single administration runs a network which has been sub-divided into multiple, contiguous ASes, each running BGP. There are several reasons why a single administrative domain may be broken into several ASes (which, in this case, are not really "autonomous".) It may be that the existing IGPs do not scale well in the particular environment; it may be that a more generalized topology is desired than could be obtained by use of a single IGP domain; it may be that a more finely grained routing policy is desired than can be supported by an IGP. In such deployments, it can be useful to allow BGP to make its routing decisions based on the IGP metric, so that BGP chooses the "shortest" path between two nodes, even if the nodes are in two different ASes within that same administrative domain. We will refer to the set of ASes in a common administrative domain as an "AIGP Administrative Domain".

There are in fact some implementations that already do something like this, using BGP's MULTI_EXIT_DISC (MED) attribute to carry a value based on IGP metrics. However, that doesn't really provide IGP-like "shortest path" routing, as the BGP decision process gives priority to other factors, such as the AS_PATH length. Also, the standard procedures for use of the MED do not ensure that the IGP metric is

- A value field containing zero or more octets.

This document defines only a single such TLV, the "AIGP TLV". The AIGP TLV is encoded as follows:

- Type: 1
- Length: 11
- Accumulated IGP Metric.

The value field of the AIGP TLV is always 8 bytes long. IGP metrics are frequently expressed as 4-octet values, and this ensures that the AIGP attribute can be used to hold the sum of an arbitrary number of 4-octet values.

3.1. Applicability Restrictions and Cautions

This document only considers the use of the AIGP attribute in networks where each router uses tunneling of some sort to deliver a packet to its BGP next hop. Use of the AIGP attribute in networks that do not use tunneling is outside the scope of this document.

If a Route Reflector supports the AIGP attribute, but some of its clients do not, then the routing choices that result may not all reflect the intended routing policy.

3.2. Restrictions on Sending/Receiving

An implementation that supports the AIGP attribute MUST support a per-session configuration item, AIGP_SESSION, that indicates whether the attribute is enabled or disabled for use on that session.

- The default value of AIGP_SESSION, for EBGP sessions, MUST be "disabled".
- The default value of AIGP_SESSION, for IBGP and confederation-EBGP sessions, MUST be "enabled."

The AIGP attribute MUST NOT be sent on any BGP session for which AIGP_SESSION is disabled.

If an AIGP attribute is received on a BGP session for which AIGP_SESSION is disabled, the attribute MUST be treated exactly as if it were an unrecognized non-transitive attribute. That is, "it MUST be quietly ignored and not passed along to other BGP peers" (see

[BGP], section 5).

3.3. Creating and Modifying the AIGP Attribute

3.3.1. Originating the AIGP Attribute

An implementation that supports the AIGP attribute MUST support a configuration item, AIGP_ORIGINATE, that enables or disables its creation and attachment to routes. The default value of AIGP_ORIGINATE MUST be "disabled".

A BGP speaker MUST NOT add the AIGP attribute to any route whose path leads outside the "AIGP administrative domain" to which the BGP speaker belongs. It may be added only to routes that satisfy one of the following conditions:

- The route is a static route that is being redistributed into BGP
- The route is an IGP route that is being redistributed into BGP
- The route is an IBGP-learned route whose AS_PATH attribute is empty.
- The route is an EBGP-learned route whose AS_PATH contains only ASes that are in the same AIGP Administrative Domain as the BGP speaker.

A BGP speaker R MUST NOT add the AIGP attribute to any route for which R does not set itself as the next hop.

It SHOULD be possible to set AIGP_ORIGINATE to "enabled for the routes of a particular IGP that are redistributed into BGP" (where "a particular IGP" might be "OSPF" or "ISIS"). Other policies determining when and whether to originate an AIGP attribute are also possible, depending on the needs of a particular deployment scenario.

When originating an AIGP attribute for a BGP route to address prefix P, the value of the attribute is set according to policy. There are a number of useful policies, some of which are in the following list:

- When a BGP speaker R is redistributing into BGP an IGP route to address prefix P, the IGP will have computed a "distance" from R to P. This distance MAY be assigned as the value of AIGP attribute.

- A BGP speaker R may be redistributing into BGP a static route to address prefix P, for which a "distance" from R to P has been configured. This distance MAY be assigned as the value of AIGP attribute.
- A BGP speaker R may have received and installed a BGP-learned route to prefix P, with next hop N. Or it may be redistributing a static route to P, with next hop N. Then:
 - * If R has an IGP route to N, the IGP-computed distance from R to N MAY be used as the AIGP attribute value of the route to P.
 - * If R has a BGP route to N, and an AIGP attribute value has been computed for that route (see section 3.3.3), that value MAY be used as the AIGP attribute value of the route to P.

3.3.2. Modifications by the Originator

If BGP speaker R is the originator of the AIGP attribute of prefix P, and at some point the "distance" from R to P changes, R SHOULD issue a new BGP update containing the new value of the AIGP attribute. (Here we use the term "distance" to refer to whatever value the originator assigns to the AIGP attribute, however it is computed; see section 3.3.1.) However, if the difference between the new distance and the distance advertised in the AIGP attribute is less than a configurable threshold, the update MAY be suppressed.

3.3.3. Modifications by a Non-Originator

Suppose a BGP speaker R1 receives a route with an AIGP attribute whose value is A, and a Next Hop whose value is R2. Suppose also that R1 is about to redistribute that route on a BGP session that is enabled for sending/receiving the attribute.

If R1 does not change the Next Hop of the route, then R1 MUST NOT change the AIGP attribute value of the route.

If R1 changes the Next Hop of the route from R2 to R1, and if R1's route to R2 is an IGP-learned route, or a static route that does not require recursive next hop resolution, then R1 must increase the value of the AIGP attribute by adding to A the distance from R1 to R2. This distance is either the IGP-computed distance from R1 to R2, or some value determined by policy. However, A MUST be increased by a non-zero amount.

Note that if R1 and R2 above are EBGp neighbors, and there is a direct link between them on which no IGP is running, then when R1 changes the next hop of a route from R2 to R1, the AIGP metric value MUST be increased by a non-zero amount. The amount of the increase SHOULD be such that it is properly comparable to the IGP metrics. E.g., if the IGP metric is a function of latency, then the amount of the increase should be a function of the latency from R1 to R2.

If R1 changes the Next Hop of the route from R2 to R1, and if R1's route to R2 is a BGP-learned route, or a static route that requires recursive next hop resolution, then the AIGP attribute value needs to be increased in several steps, according to the following procedure. (Note that this procedure is ONLY used when recursive next hop resolution is needed.)

1. Let Xattr be the new AIGP attribute value.
2. Initialize Xattr to A.
3. Set the XNH to R2.
4. Find the route to XNH.
5. If the route to XNH does not require recursive next hop resolution, get the distance D from R1 to XNH. (Note that this condition cannot be satisfied the first time through this procedure.) If D is above a configurable threshold, set the AIGP attribute value to Xattr+D. If D is below a configurable threshold, set the AIGP attribute value to Xattr. In either case, exit this procedure.
6. If the route to XNH is a BGP-learned route, and the route does NOT have an AIGP attribute, then exit this procedure and do not pass on any AIGP attribute.
7. If the route to XNH is a BGP-learned route, and the route has an AIGP attribute value of Y, then set Xattr=Xattr+Y, and set XNH to the next hop of this route. (The intention here is that Y is the AIGP value of the route as it was received by R1, without having been modified by R1.)
8. Go to step 4.

The AIGP value of a given route depends on (a) the AIGP values of all the next hops that are recursively resolved during this procedure, and (b) the IGP distance to any next hop that is not recursively resolved. Any change due to (a) in any of these values MUST trigger a new AIGP computation for that route. Whether a change due to (b)

triggers a new AIGP computation depends upon whether the change in IGP distance exceeds a configurable threshold.

If the AIGP attribute is carried across several ASes, each with its own IGP domain, it is clear that these procedures are unlikely to give a sensible result if the IGPs are different (e.g., some OSPF and some IS-IS), or if the meaning of the metrics is different in the different IGPs (e.g., if the metric represents bandwidth in some IGP domains but represents latency in others). These procedures also are unlikely to give a sensible result if the metric assigned to inter-AS BGP links (on which no IGP is running) or to static routes is not comparable to the IGP metrics. All such cases are outside the scope of the current document.

4. Decision Process

Support for the AIGP attribute involves several modifications to the tie breaking procedures of the BGP "phase 2" decision described in [BGP], section 9.1.2.2. These modifications are described below in sections 4.1 and 4.2.

In some cases, the BGP decision process may install a route without executing any tie breaking procedures. This may happen, e.g., if only one route to a given prefix has the highest degree of preference (as defined in [BGP] section 9.1.1). In this case, the AIGP attribute is not considered.

In other cases, some routes may be eliminated before the tie breaking procedures are invoked, e.g., routes with AS-PATH attributes indicating a loop, or routes with unresolvable next hops. In these cases, the AIGP attributes of the eliminated routes are not considered.

4.1. When a Route has an AIGP Attribute

Assuming that the BGP decision process invokes the tie breaking procedures, the procedures in this section MUST be executed BEFORE any of the tie breaking procedures described in [BGP] section 9.1.2.2 are executed.

If any routes have an AIGP attribute, remove from consideration all routes that do not have an AIGP attribute.

If router R is considering route T, where T has an AIGP attribute,

- then R must compute the value A, defined as follows: set A to the sum of (a) T's AIGP attribute value and (b) the IGP distance from R to T's next hop.
- remove from consideration all routes that are not tied for the lowest value of A.

4.2. When the Route to the Next Hop has an AIGP attribute

Suppose that a given router R1 is comparing two BGP-learned routes, such that either:

- the two routes have equal AIGP attribute values, or else
- neither of the two routes has an AIGP attribute.

The BGP decision process as specified in [BGP] makes use, in its tie breaker procedures, of "interior cost", defined as follows:

"interior cost of a route is determined by calculating the metric to the NEXT_HOP for the route using the Routing Table."

Suppose route T has a next hop of N. We modify the notion of the "interior cost" from node R1 to node N as follows:

- Let R2 be the BGP next hop of the route to N, after all recursive resolution of the next hop is done. Let m be the IGP distance (or in the case of a static route, the configured distance) from R1 to R2.
- If the installed route to N has an AIGP attribute, set A to the AIGP value of the route to N, computing the AIGP value of the route according to the procedure of section 3.3.3.
- If the installed route to N does not have an AIGP value, set A to 0.
- The "interior cost" of route T is the quantity A+m.

5. Deployment Considerations

Using the AIGP attribute to achieve a desired routing policy will be more effective if each BGP speaker can use it to choose from among multiple routes. Thus is it highly recommended that the procedures of [BESTEXT] and [ADDPATH] be used in conjunction with the AIGP Attribute.

If a Route Reflector does not pass all paths to its clients, then it will tend to pass the paths for which the IGP distance from the Route Reflector itself to the next hop is smallest. This may result in a non-optimal choice by the clients.

6. IANA Considerations

IANA has assigned the codepoint 26 in the "BGP Path Attributes" registry to the AIGP attribute.

IANA shall create a registry for "BGP AIGP Attribute Types". 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". Type 1 should be defined as "AIGP", and should refer to this document.

7. Security Considerations

The spurious introduction, though error or malfeasance, of an AIGP attribute, could result in the selection of paths other than those desired.

Improper configuration on both ends of an EBGp connection could result in an AIGP attribute being passed from one service provider to another. This would likely result in an unsound selection of paths.

8. Acknowledgments

The authors would like to thank Waqas Alam, Rajiv Asati, Clarence Filsfils, Robert Raszuk, Yakov Rekhter, Eric Rosenberg, Samir Saad, John Scudder, and Shyam Sethuram for their input.

9. Authors' Addresses

Rex Fernando
Cisco Systems, Inc.
170 Tasman Drive
San Jose, CA 95134
Email: rex@cisco.com

Pradosh Mohapatra
Cumulus Networks
Email: pmohapat@cumulusnetworks.com

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

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

10. Normative References

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

11. Informative References

[ADDPATH] "Fast Connectivity Restoration Using BGP Add-Path", P. Mohapatra, R. Fernando, C. Filsfils, R. Raszuk, draft-pmohapat-idr-fast-conn-restore-03.txt, January 2013.

[BESTEXT], "Advertisement of the Best External Route in BGP", P. Marques, R. Fernando, E. Chen, P. Mohapatra, H. Gredler, draft-ietf-idr-best-external-05.txt, January 2012.

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

IDR
Internet-Draft
Intended status: Standards Track
Expires: August 2, 2018

S. Shah

K. Patel
Arrcus, Inc
S. Bajaj
Viptela
L. Tomotaki
Verizon
M. Boucadair
Orange
January 29, 2018

Inter-domain Traffic Conditioning Agreement (TCA) Exchange Attribute
draft-ietf-idr-sla-exchange-13.txt

Abstract

Network administrators typically enforce Quality of Service (QoS) policies according to Traffic Conditioning Agreement (TCA) with their providers. The enforcement of such policies often relies upon vendor-specific configuration language. Both learning of TCA, either thru TCA documents or via some other out-of-band method, and translating them to vendor specific configuration language is a complex, often manual, process and prone to errors.

This document specifies an optional transitive attribute to signal TCA parameters in-band, across administrative boundaries (considered as Autonomous Systems (AS)), thus simplifying and facilitating some of the complex provisioning tasks in situations where BGP is available as a routing protocol.

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 <https://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on August 2, 2018.

Copyright Notice

Copyright (c) 2018 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 (<https://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. Terminology	4
3. QoS Attribute Definition	5
3.1. QoS Attribute SubType	6
3.2. TCA SubType	7
3.3. TCA Content for ADVERTISE TCA Event	9
3.3.1. Supported IPFIX identifiers for Traffic Class Elements	12
3.3.2. Traffic Class Service types and respective TLVs . . .	15
4. Originating TCA Notification	22
4.1. TCA Contexts	23
4.1.1. TCA Advertisement for Point-to-Point Connection . . .	23
4.1.2. TCA Advertisement for Destination AS Multiple Hops Away	24
5. QoS Attribute Handling at Forwarding Nodes	24
5.1. BGP Node Capable of Processing QoS Attribute	24
5.2. QoS Attribute Handling at Receiver	25
6. Error Handling	25
7. Deployment Considerations	25
8. IANA Considerations	27
9. Security Considerations	28
10. Acknowledgements	29
11. References	29
11.1. Normative References	29
11.2. Informative References	30
Authors' Addresses	31

1. Introduction

Typically there is a contractual Traffic Conditioning Agreement (TCA) for Quality of Service (QoS) established between a customer and a provider or between providers [RFC7297]. This QoS TCA defines the nature of the various traffic classes and services needed within each traffic class. The contract may include full line-rate or sub line-rate without additional traffic classes, or it may contain additional traffic classes and service definitions for those traffic classes. Finer granular traffic classes may be based on some standard code points (e.g., based on DSCP (Differentiated Services Code Point)) or specific set of prefixes.

Once the contractual QoS TCA is established, QoS TCA parameters are enforced in some or all participating devices by deriving those parameters into configuration information on respective devices. The network administrator translates the QoS TCA to QoS policies using router (vendor) specific provisioning language. In a multi-vendor network, translating TCAs into technology-specific and vendor-specific configuration requires the network administrator to consider specific configuration of each vendor. There does not exist any standard protocol to translate TCA agreements into technical clauses and configurations and thus both the steps of out of band learning of negotiated TCA and provisioning them in a vendor specific language can be complex and error-prone.

TCA parameters may have to be exchanged through organizational boundaries, thru TCA documents or via some other off-band method, to an administrator provisioning actual devices. For example, to provide voice services, the provider may negotiate QoS parameters (like min/max rates) for such traffic classified under the EF (Expedited Forwarding) codepoint in Diffserv-enabled [RFC2475] networks. The Administrator at the CE (Customer Edge) not only will have to know that provider's service for voice traffic is EF-based but will also have to know how to implement DSCP EF classification rule along with Low Latency Service, and possibly min/max rate enforcement for the optimal use of bandwidth, as per vendor specific provisioning language.

The Inter-domain exchange of QoS TCA policy described in this document does not require any specific method for the provider in establishing TCAs. It only requires that the provider wishes to send the QoS TCA policy via BGP UPDATE [RFC4271] messages from the provider to a set of receivers (BGP peers). In reaction to, a receiving router may translate that to relevant QoS policy definition on the device. The TCA negotiation and assurance is outside the scope of this document.

This document defines a new optional BGP transitive attribute, referred as QoS Attribute, which has as one of its sub-types the TCA policy. The BGP node of the originating AS sends this QoS Attribute, for prefixes this QoS TCA Policy applies to, in a BGP UPDATE message that will be distributed to a list of destination ASes. The QoS TCA policy can be for inbound traffic to the advertising AS or outbound traffic from the advertising AS, or both.

Protocols and data models are being created to standardize setting routing configuration parameters within networks. YANG data models [RFC6020] are being developed so that NETCONF ([RFC6241]) or RESTCONF [RFC8040] can set these standardize in configuration mechanisms. For ephemeral state, the I2RS protocol is being developed to set ephemeral state. While these protocols provide valid configuration within a domain or across domains, some providers desire to exchange QoS parameters in- band utilizing BGP peering relationships. This is similar to the distribution of Flow Specification information via BGP peering relationships (see [RFC5575] and [RFC7674]).

2. Terminology

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

This document makes use of the following terms:

- o BGP Speaker: A functional component on a BGP capable device that functions as per BGP specification.
- o BGP peers: BGP Speakers adjacent to each other.
- o QoS Attribute Speaker: A functional component on a BGP capable device that produces and/or processes content of the QoS Attribute. A device that is QoS Attribute Speaker is also always a BGP Speaker. However, a BGP Speaker not necessarily always a QoS Attribute Speaker.
- o QoS Attribute content is produced and processed outside the function of the BGP Speaker and thus content of the QoS Attribute is completely opaque to the BGP Speaker. At BGP capable device where QoS Attribute content is produced, length and value of the QoS Attribute is passed from QoS Attribute Speaker to the BGP Speaker where BGP Speaker inserts the attribute into the BGP UPDATE message with appropriate attribute flags, attribute type, and length and value passed from the QoS Attribute Speaker. Similarly, a BGP capable device when receives QoS Attribute in the

BGP UPDATE message, BGP Speaker extracts QoS Attribute value from the message and passes it to the QoS Attribute Speaker where QoS Attribute Speaker processes the content from that passed down value. How the content of the QoS Attribute is passed from the QoS Attribute Speaker to the BGP Speaker and vice versa is implementation specific.

In the context of use of QoS Attribute for TCA parameters exchange, following roles are defined further within the scope of the QoS Attribute Speaker.

- o TCA Producer: This is a QoS Attribute Speaker that produces QoS Attribute for the TCA SubType.
- o TCA Consumer: This is a QoS Attribute Speaker that is intended receiver of QoS Attribute with the TCA SubType.

3. QoS Attribute Definition

The QoS Attribute is an optional transitive attribute (TBD - attribute code to be assigned by IANA) which is applicable to the Source AS and NLRIs advertised in the BGP UPDATE message this attribute is included in. The format of the QoS Attribute is shown in Figure 1.

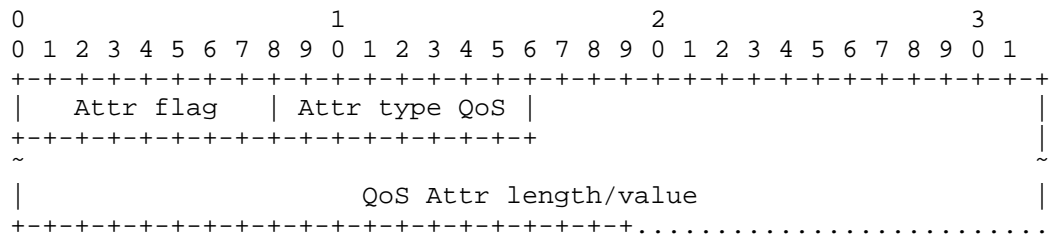


Figure 1: QoS attribute

Attribute flags - 8-bits field

highest order bit (bit 0) - MUST be set to 1, since this is an optional attribute

2nd higher order bit (bit 1) - MUST be set to 1, since this is a transitive attribute

The content of the QoS Attribute is further specified with flags, applicable to QoS Attribute content, and a SubType in a TLV form.

3.1. QoS Attribute SubType

The Value field of the QoS Attribute contains the following:

QoS Attribute flags

Tuple (SubType of the QoS Attribute, SubType length, SubType value)

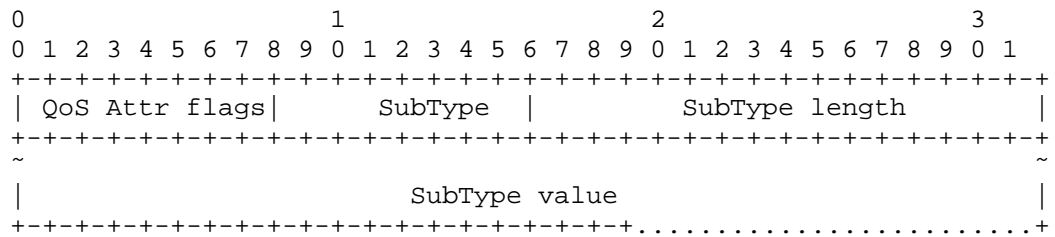


Figure 2: Format of QoS Attribute

QoS Attr flags - 8-bits field

All bits of this field are currently un-used. The space is provided for future use. All bits MUST be set to zero when sent. The values (0x01-0xFF) are reserved, and MUST be ignored when received.

SubType - 8-bits field with following values:

0x00 = reserved

0x01 = TCA

0x02 - 0xf0 = reserved for future use (Standards Action)

0xf1 - 0xff = Private use

The only SubType of the QoS Attribute defined in this specification is the TCA SubType.

SubType length - 16-bits field that specifies length of the SubType value in number of octets.

SubType value - variable length field, as expressed in SubType

length, that contains information about a specified SubType. For the TCA SubType the information is about sender and receiver(s), and TCA parameters as described in Section 3.2.

3.2. TCA SubType

Format of the TCA SubType Value field is shown in Figure 3.

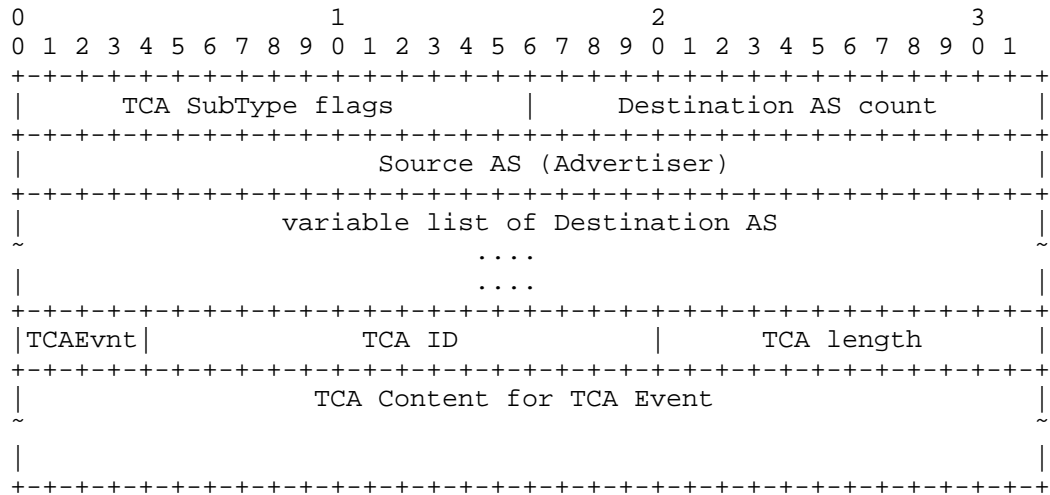


Figure 3: Format of the TCA SubType of the QoS attribute

TCA SubType flags - 16-bits field

Currently un-used. All bits in this field MUST be set to 0. The field is defined for the future use.

Destination AS count - 16-bits field that specifies count of destination ASNs present in the Destination AS list.

If this count is 0 then that is an error condition which should be handled as described in Section 6.

Source AS - 32-bits field AS number space as defined in [RFC6793]

This is the AS where TCA Content is originated from. The Source AS MUST be of the same AS that is originating TCA ID and TCA Content.

The Source AS value of 0 is illegal and thus should be considered an error which should be handled as described in Section 6.

Destination AS list - variable length field that holds as many ASN. identifiers, each is 32-bits AS number space is defined in [RFC6793], as specified in the Destination AS count field.

List of ASNs for which the TCA is relevant to, each of which is a 32-bit number.

TCA Event - 4-bits field with following values:

0x0 = reserved

0x1 = ADVERTISE

0x2 to 0xf = Reserved for future use

The only TCA Event defined in this specification is ADVERTISE.

TCA ID - 16-bits field that specifies identifier which is unique in the scope of Source AS.

The significance of a TCA ID is in the context of the source that is advertising TCA Content. The TCA ID is not globally unique but it MUST be unique within the source AS.

The TCA ID applies to aggregate traffic to prefixes for a given AFI/SAFI that share the same Source AS and TCA ID.

TCA Length - 12-bits field that specifies the length of the TCA Content. The length is expressed in octets. The TCA Content is optional for an advertised TCA ID. If the TCA Content need not be there, the TCA length field MUST be set to zero in such a case.

TCA Content - A variable length field (optional field)

The TCA Content field contains TCA parameters relevant to specified TCA SubType. Since the only defined TCA SubType is ADVERTISE, this specification describes TCA Content only for the ADVERTISE TCA Event.

If TCA Content field exists in a BGP UPDATE message that contains the QoS Attribute with a TCA SubType for TCA Event ADVERTISE, format of the TCA Content is as described in Section 3.3.

If the TCA Content field does not exist, then the advertised message refers to TCA Content advertised in the previous message for the same TCA ID. If there does not exist any prior TCA Content to relate to the advertised TCA ID, then receiver, TCA Consumer, can ignore the TCA advertisement and it may simply update Destination AS count and Destination AS list.

The lack of a valid prior TCA Content field does not make this attribute invalid, so the QoS Attribute MUST be forwarded as a valid BGP optional transitive attribute.

3.3. TCA Content for ADVERTISE TCA Event

The only TCA Event described in this specification is ADVERTISE. The format of TCA Content for the ADVERTISE Event is shown in Figure 4.

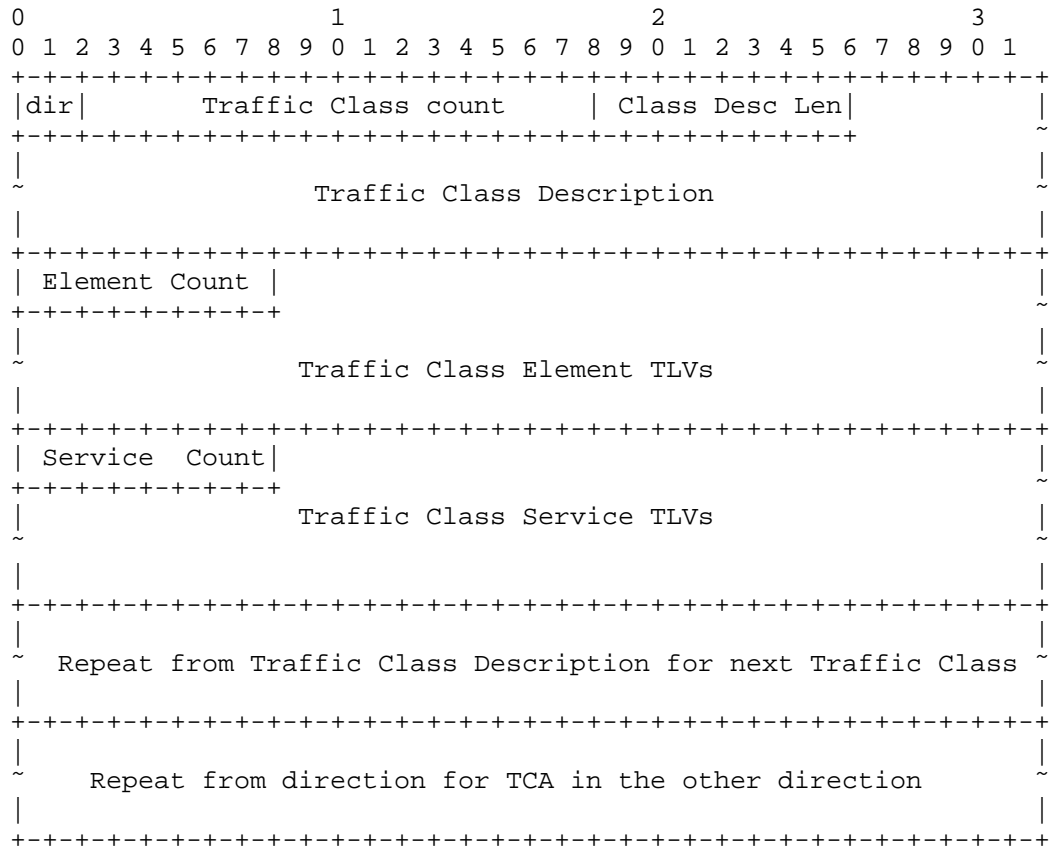


Figure 4: TCA-Content for ADVERTISE TCA Event

TCA Content contains Traffic Class TLVs that is a set of Traffic Class Elements (Classifiers) and Traffic Class Service TLVs for a list of Traffic Classes specified by Traffic Class count. This Traffic Class TLVs MUST be specified for one direction first and then optionally followed by the specification for the other direction.

dir (Direction) - 2-bits field that specifies Direction of the traffic TCA is applicable to. The following values are defined:

0x0 = reserved

0x1 = incoming, traffic to source AS from destination AS

0x2 = outgoing, traffic from source AS towards destination AS

0x3 = for future use

Traffic Class (Classifier Group) count - 16 bits field that specifies number of Traffic Classes.

The value of zero (0x00) in this field is a special value which means no TCA for the traffic in a specified direction. When Traffic Class count is 0, for a specific direction, the rest of the TCA Content fields MUST NOT be encoded, for that specific direction.

Traffic Class Description Len - 8-bits field that specifies the length of the Traffic Class Description field. The length is expressed in octets.

The value of zero in this field indicates that no Traffic Class Description field follows.

Traffic Class Description - variable length field, as expressed in The Traffic Class Description Len field, MUST carry UTF-8 encoded ([RFC3629]) description.

Traffic Class Elements (Classifier) Count - 8-bits field that specifies the count of Traffic Class Elements.

The value zero (0x00) means there are no Traffic Class Elements in the traffic class, and thus the Traffic Class is to classify rest of the traffic not captured otherwise by other Traffic Classes in the set for a specified direction.

Traffic Class that has 0 elements MUST be presented last in the advertised list of Traffic Classes for a specific Direction.

Otherwise it is considered an error condition which should be handled as described in Section 6.

The QoS Attribute advertised from a specific source MUST NOT have more than one such Traffic Classes (Traffic Class with 0 elements count). If there are more than one such Traffic Classes present then it is an error condition which should be handled as described in Section 6.

Traffic Class Element TLVs - (optional) variable length field holding as many TLVs specified by the Traffic Class Elements Count field. Each TLV has the following format:

IPFIX Element Identifier - 8-bits field that specifies IPFIX Identifiers listed in Table 1.

Length of Value field - 8-bits field that specifies the length, expressed in octets, of the value field.

Value - A variable field that specifies a value appropriate for the IPFIX Element Identifier. It is an error, if the value field does not contain the appropriate format, which should be handled as described in Section 6. Only the IPFIX elements shown in Table 1 are supported.

Any Traffic Class Element advertised in the QoS Attribute only applies to the advertised AFI/SAFI NLRI within the BGP UPDATE message the QoS Attribute is contained in. If a receiver, TCA Consumer, receives a BGP UPDATE message with QoS Attribute for an unsupported AFI/SAFI then TCA Consumer MAY ignore advertised TCA. TCA Consumer MAY update only Destination AS count and Destination AS list, and then QoS Attribute and rest of the BGP UPDATE message MUST be forwarded as per QoS Attribute and BGP protocol specification.

Traffic Class Service Count - 8-bits field that specifies count of Traffic Class Service TLVs.

A value of zero is a special value indicating "no bounded service" (a.k.a., Best Effort (BE)).

Traffic Class Service TLVs - (optional) variable length field with the following format for the TLVs

Traffic Class Service type - 16-bits field that specifies a service type. Each service type is detailed in Section 3.3.2. The list of available service types are,

0x00 = reserved
 0x01 = COMMITTED_TSPEC
 0x02 = PEAK_TSPEC
 0x03 = COMMITTED_IN_PROFILE_MARKING
 0x04 = COMMITTED_OUT_PROFILE_MARKING
 0x05 = PEAK_OUT_PROFILE_MARKING
 0x06 = DROP_THRESHOLD
 0x07 = RELATIVE_PRIORITY
 0x08 = EFFECTIVE_MAX_RATE

Length of Value field - 08-bits field that specifies the length of the value field. The length of the value is expressed in octets.

Value - a variable length field that specifies the value appropriate for each of the Service Types. It is an error, if this field does not contain the appropriate format, which should be handled as described in Section 6. The format of the value for each of the service types is described in Section 3.3.2

3.3.1. Supported IPFIX identifiers for Traffic Class Elements

IPFIX [RFC7012] has well defined identifier set for a large number of packet attributes; an IPFIX IANA registry maintains values for packet classifier attributes (<<https://www.ietf.org/assignments/ipfix/ipfix.xml#ipfix-information-elements>> ipfix.xml#ipfix-information-elements). Only the IPFIX attributes listed in Table 1 are supported. Any new attribute to be supported by TCA SubType MUST be a Standards Action as described in IANA section.

ID	Name	Context
195	ipDiffServCodePoint	Indicates the value of the marking used in the link(s) between the TCA Consumer and TCA Producer domains.
203	mplsTopLabelExp	Indicates the value of the marking used in the link(s) between the TCA Consumer and

		TCA Producer domains.
244	dot1qPriority	Indicates the value of the marking used in the link(s) between the TCA Consumer and TCA Producer domains.
8	sourceIPv4Address	Indicates the source IPv4 address of an aggregate traffic over a connection subject to a TCA; the direction is being explicitly indicated in the ADVERTISE Event message.
27	sourceIPv6Address	Indicates the source IPv6 address of an aggregate traffic over a connection subject to a TCA; the direction is being explicitly indicated in the ADVERTISE Event message.
9	sourceIPv4PrefixLength	Indicates the length of the source IPv4 prefix.
29	sourceIPv6PrefixLength	Indicates the length of the source IPv6 prefix.
44	sourceIPv4Prefix	Indicates the source IPv4 prefix of an aggregate traffic over a connection subject to a TCA; the direction is being explicitly indicated in the ADVERTISE Event message.
170	sourceIPv6Prefix	Indicates the source IPv6 prefix of an aggregate traffic over a connection subject to a TCA; the direction is being explicitly indicated in the ADVERTISE Event message.
12	destinationIPv4Address	Indicates the destination IPv4 address of an aggregate traffic over a connection

		subject to a TCA; the direction is being explicitly indicated in the ADVERTISE Event message.
28	destinationIPv6Address	Indicates the destination IPv6 address of an aggregate traffic over a connection subject to a TCA; the direction is being explicitly indicated in the ADVERTISE Event message.
13	destinationIPv4PrefixLength	Indicates the length of the destination IPv4 prefix.
30	destinationIPv6PrefixLength	Indicates the length of the destination IPv6 prefix.
45	destinationIPv4Prefix	Indicates the destination IPv4 prefix of an aggregate traffic over a connection subject to a TCA; the direction is being explicitly indicated in the ADVERTISE Event message.
169	destinationIPv6Prefix	Indicates the destination IPv6 prefix of an aggregate traffic over a connection subject to a TCA; the direction is being explicitly indicated in the ADVERTISE Event message.
4	protocolIdentifier	Indicates whether any or a specific protocol for the traffic class.
7	sourceTransportPort	This parameter is used only for protocols with port identifiers. It indicates the source port number for the transport protocol identified by "protocolIdentifier".
11	destinationTransportPort	This parameter is used only for protocols with port

rate indicates the minimum rate, measured in octets of IP datagrams per second (a.k.a, bytes per second), that the service advertiser is providing for a given class of traffic on advertiser's hop. Note that it does not necessarily translate to a minimum rate service to the receiver of a TCA unless the traffic class definition clearly represents a sole receiver of a TCA.

Parameter (b): indicates maximum burst size, measured in octets of IP datagram size. Since queuing delay can be considered a function of burst size (b) and committed-rate (r), in presence of non-zero parameter (r), parameter (b) represents bounded delay for the Traffic Class. This delay is a single hop queuing delay when TCA is to be implemented at the resource constrained bottleneck. In other words this burst size can be considered as a buffer size. Value of 0 for parameter (b) means the advertiser does not mandate specific bounded delay.

3.3.2.2. PEAK_TSPEC

The PEAK_TSPEC TLV definition:

Type - 0x01

Length - 8-bits field that specifies length, expressed in octets, of the value field. The length of the value field MUST be specified to be 8 octets to hold the value defined as per format below.

Value - PEAK_TSPEC value consists of the (r), (b) parameters as described in Invocation Information section of [RFC2212] and shown in Figure 5. Note that inheriting the definition of TSPEC (Traffic SPECification) here does not enable RFC2212 functionality. Only the format of the Traffic Specification is used in this specification.

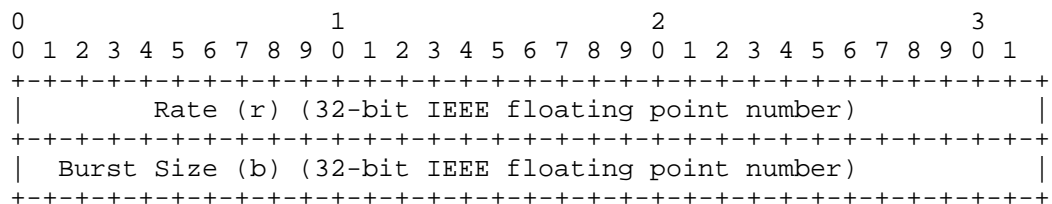


Figure 6: Traffic Class PEAK_TSPEC

Format of Parameters (r) and (b): are 32-bit IEEE floating point

numbers. Positive infinity is represented as an IEEE single precision floating-point number with an exponent of all ones and a sign mantissa of all zeros. The format of IEEE floating-point numbers is further summarized in [RFC4506].

Parameter (r): indicates peak-rate of the traffic class. This rate indicates the maximum rate, measured in octets of IP datagrams per second (a.k.a, bytes per second), that the service advertiser is providing for a given class of traffic on advertiser's hop.

Parameter (b): indicates maximum burst size, measured in octets of IP datagram size.

When PEAK_TSPEC TLV is advertised, COMMITTED_TSPEC TLV MUST be present in the advertisement. Advertisement of PEAK_TSPEC TLV without COMMITTED_TSPEC TLV MUST be considered an error condition which should be handled as described in Section 6. If committed-rate of the TCA is 0 then rate advertised in the COMMITTED_TSPEC shall be 0. Note that existence of COMMITTED_TSPEC in TCA advertisement is not mandatory nor is it a mandate that COMMITTED_TSPEC and PEAK_TSPEC must always go together. COMMITTED_TSPEC TLV is optional but only when there is no PEAK_TSPEC TLV present in the advertised TCA.

PEAK_TSPEC TLV with rate value of 0 MUST be considered an error condition which should be handled as described in Section 6.

3.3.2.3. COMMITTED_IN_PROFILE_MARKING

This Traffic Class Service Type defines action performed, by the TCA Producer, on packets that are compliant to the committed-rate specified in the COMMITTED_TSPEC TLV. If committed-rate specified in the COMMITTED_TSPEC TLV is 0 then TLV for this Traffic Class Service Type SHOULD NOT be advertised. COMMITTED_IN_PROFILE_MARKING TLV SHOULD be ignored by the TCA Consumer if there does not exist COMMITTED_TSPEC TLV for the specified direction, or committed-rate specified in the COMMITTED_TSPEC TLV is 0.

The COMMITTED_IN_PROFILE_MARKING TLV definition:

Type - 0x03

Length - 8-bits field that specifies length, expressed in octets, of the value field. The length of the value field MUST be specified to be 2 octets to hold the value defined as per format below.

Value - contains the Marking code-point type and value

Marking code-point type - 8-bits IPFIX Element Identifier.

Marking code-point value - 8-bits code-point number.

The marking code-point type of 0x00 is a drop identifier. When marking code-point type value is 0x00 (that is drop), the marking code-point value in this case has no meaning and thus the value in this field should be ignored.

The following table lists the supported IPFIX Identifiers. Any value other than 0 or identifier from the following table is an error condition which should be handled as described in Section 6.

ID	Name
195	ipDiffServCodePoint
203	mplsTopLabelExp
244	dot1qPriority

Table 2

3.3.2.4. COMMITTED_OUT_PROFILE_MARKING

This Traffic Class Service Type defines action performed, at the TCA Producer, on packets that are not compliant to the committed-rate specified in the COMMITTED_TSPEC TLV, and compliant to rate specified in the PEAK_TSPEC TLV if PEAK_TSPEC TLV exists.

The COMMITTED_OUT_PROFILE_MARKING TLV definition:

Type - 0x04

Length - 8-bits field that specifies length, expressed in octets, of the value field. The length of the value field MUST be specified to be 2 octets to hold the value defined as per format below.

Value - contains the Marking code-point type and value

Marking code-point type - 8-bits IPFIX Element Identifier

Marking code-point value - 8-bits code-point number

The marking code-point type of 0x00 is a drop identifier. When marking code-point type value is 0x00 (that is drop), the marking code-point value in this case has no meaning and thus the value in this field should be ignored.

Table 2 lists the supported IPFIX Identifiers. Any value other than 0 or identifier from the Table 2 is an error condition which should be handled as described in Section 6.

3.3.2.5. PEAK_OUT_PROFILE_MARKING

This Traffic Class Service Type defines action performed, at the TCA Producer, on packets that are not compliant to the max-rate specified in the PEAK_TSPEC TLV. PEAK_OUT_PROFILE_MARKING TLV SHOULD be ignored by the TCA Consumer if there does not exist PEAK_TSPEC TLV for the specified direction.

The PEAK_OUT_PROFILE_MARKING TLV definition:

Type - 0x06

Length - 8-bits field that specifies length, expressed in octets, of the value field. The length of the value field MUST be specified to be 2 octets to hold the value defined as per format below.

Value - contains the Marking code-point type and value

Marking code-point type - 8-bits IPFIX Element Identifier

Marking code-point value - 8-bits code-point number

The marking code-point type of 0x00 is a drop identifier. When marking code-point type value is 0x00 (that is drop), the marking code-point value in this case has no meaning and thus the value in this field should be ignored.

Table 2 lists the supported IPFIX Identifiers. Any value other than 0 or identifier from the Table 2 is an error condition which should be handled as described in Section 6.

3.3.2.6. DROP_THRESHOLD

The DROP_THRESHOLD TLV definition:

Type - 0x07

Length - 8-bits field that specifies length, expressed in octets, of the value field.

Value - Count of drop thresholds, followed by content for each drop threshold in the form of (code-point type, count of code-points, list of code-points, threshold value).

Count of drop thresholds - 8-bits field that specifies number of drop thresholds specified in this TLV. Content of each drop threshold is to follow following format

Code-point type - 8-bits IPFIX Element Identifier from the list shown in Table 6.

Count of code-points - 8-bits field that specifies number of code-point values to follow for a specified code-point type.

List of code-points - each code-point value is specified in size of 8 bits and thus total size for this field is 8 bits multiplied by as many number of code-points specified.

Burst value - This is a fixed size 32-bits IEEE floating point number that specifies burst value in unit of bytes.

All advertised drop thresholds, for a specific traffic class, are applicable to a single queue associated with that traffic class. A threshold for a set of code-points is a logical marker where an arrived packet is to be dropped if overall depth of a queue is beyond a threshold of a code-point set a packet is classified into. Choice of dropping discipline is implementation specific. If a packet can not be classified into any of the advertised code-point set then that means the TCA Producer is not defining any specific dropping behavior and thus dropping behavior is subject to implementation specific of the TCA Consumer.

ID	Name
195	ipDiffServCodePoint
203	mplsTopLabelExp
244	dot1qPriority

Table 3

3.3.2.7. RELATIVE_PRIORITY

The RELATIVE_PRIORITY TLV definition:

Type - 0x08

Length - 8-bits field that specifies length, expressed in octets, of the value field. Given supported range of priority values in this specification, the length of the value field MUST be limited to and thus MUST be specified exactly as 1 octet.

Value - A value from range of 0 - 255. Lower the value means higher the priority

Relative priority indicates scheduling priority of this traffic class. Voice traffic, for example, which requires lowest latency compared to any other traffic, may have lowest value advertised in relative priority. For two different traffic classification groups where one classification group may be considered more important than the other, but from a scheduling perspective does not require to be distinguished with a different priority, relative priority for those classification groups should be advertised with the same value.

A higher priority class of traffic to be served without pre-empted by lower priority class of traffic for more than a packet time at the configured rate.

For a system that implements WRR only (i.e., no priority queuing), it is possible to use a hierarchical WRR scheduling to achieve a behavior close to priority queueing where a root scheduling node has two child nodes. One child node is a queue assigned with a maximum possible value of a weight and advertised rate of highest priority Traffic Class as output bandwidth. The other child node is a scheduling node serving group of rest other advertised Traffic Classes (in the form of queues or yet another level of hierarchical WRR scheduler). Note that implementation specifics are out of the scope of this specification and this is an example to highlight how relative priority attribute can be relevant and treated by a system that implements only WRR. A system may choose to implement alternate methods to achieve a similar behavior.

3.3.2.8. EFFECTIVE_MAX_RATE

The EFFECTIVE_MAX_RATE TLV definition:

Type - 0x02

Length - 8-bits field that specifies length, expressed in octets, of the value field. The length of the value field MUST be specified to be 5 octets to hold the value defined as per format below.

Value - Contains value of rate and per packet overhead

Aggregate max rate - 32-bits IEEE floating point number

Per packet overhead - 8-bits specifying value of overhead octets

Aggregate max rate indicates rate measured based on combined octets of packet's IP datagram size and advertised per packet overhead.

A packet traversing from the TCA Producer to the TCA Consumer or vice-versa may see packet overhead, additional octets on top of IP datagram size, difference between the Producer and the Consumer sent or received over a physical link. In cases, where advertised TCA is for a Consumer where total traffic between Consumer and Producer is to be capped to a specific sub-rate of a physical link, due to packet overhead differences between Producer and Consumer, sum of traffic from each TRAFFIC CLASS may overrun that total cap causing undesired behavior. In such cases, Producer can explicitly notify this TLV in advertised TCA.

4. Originating TCA Notification

The QoS Attribute for the TCA SubType MUST only be added to the BGP UPDATE message at the node that is TCA Producer. Any QoS Attribute Speaker, in the path to the TCA Consumer MUST NOT modify content of that attribute except modification of the Destination AS list.

QoS Attribute with the TCA SubType SHOULD NOT be advertised periodically just for the purpose of KEEPALIVE between TCA Producer and TCA Consumer. Some sort of TCA policy change, at the TCA Producer, may be considered as a trigger for the advertisement.

For any modified TCA policy at the TCA Producer, the TCA Producer MUST re-advertise the entire set of TCA parameters. There is no provision to advertise partial set of TCA parameters. Announcing a TCA ID different from an earlier advertised one, for the same prefix and from the same Source AS, indicates Source AS is advertising new TCA Content to replace the previous one advertised with the same TCA ID.

In order to withdraw a given TCA between TCA Producer and TCA Consumer, the TCA Produced MUST sent TCA Content with the same TCA ID, AS Source, and NLRI prefix, as were used to advertise earlier TCA parameters, and the Traffic Class count MUST be set to 0.

4.1. TCA Contexts

4.1.1. TCA Advertisement for Point-to-Point Connection

In certain cases, the advertisement of a TCA is intended to relate to aggregate traffic over a point-to-point connection between a specific destination and a specific source. A point-to-point connection may be a physical link or a virtual link (e.g. a tunnel). In such cases, a BGP UPDATE message with source AS number and NLRI prefix as an IP address of a TCA Producer can uniquely identify physical/virtual link in order to establish the context for the advertised TCA for that point to point link.

In the simplest case where Provider (e.g., PE) and Customer (e.g., CE) devices are directly connected via a physical link and have only a single link between them, the CE can uniquely identify the forwarding link to the PE with the following:

- o AS number of the PE,
- o NLRI prefix being an IP address of the PE, that is the next hop address from CE to PE.

The TCA advertised in the QoS Attribute in the BGP UPDATE message sent from the PE to a CE, along with the PE's AS number and PE's IP address, establishes TCA context for the aggregate traffic through CE-to-PE link.

The TCA advertised in the QoS Attribute in the BGP UPDATE message from PE to CE, with PE's AS number and any other prefix, means TCA for that specific prefix based traffic, a subset of traffic through CE-to-PE link.

Even though this example is in the context of IP prefixes, QoS Attribute's TCA exchange does not have to be limited to the IP address family (IPv4 and IPv6). TCA advertisement is generic to all forms of NLRI types that are supported by the BGP specification (like IPv4, IPv6, VPN-IPv4, VPN-IPv6).

When BGP UPDATE message with the QoS Attribute, containing TCA SubType, is triggered for a point-to-point connection (physical or logical), the Source AS number in the TCA SubType SHOULD be set to

TCA Producer's AS number and destination AS number SHOULD be set to AS number of BGP peer's that is targeted TCA Consumer.

4.1.2. TCA Advertisement for Destination AS Multiple Hops Away

When advertised TCA is not for the BGP peer of a TCA Producer, the Source AS field, in the TCA SubType, MUST be set. The list of destination AS(es) also MUST be set, in the TCA SubType, to avoid flooding of the QoS Attribute data in the network beyond those destinations. Destination AS(es) is a list of TCA Consumers the advertised TCA is intended for.

If a new prefix is learned and traffic with this new prefix is subject to TCA parameters that have already been advertised before for other existing prefixes, then the BGP UPDATE for this new prefix MAY include QoS Attribute containing just a TCA ID that was advertised earlier. This BGP UPDATE message does not require to have the whole TCA Content. The TCA ID is sufficient to relate TCA parameters to new advertised prefixes.

5. QoS Attribute Handling at Forwarding Nodes

The propagation of the QoS Attribute in the BGP UPDATE messages depends on the rules detailed in the following sub-sections.

5.1. BGP Node Capable of Processing QoS Attribute

If a BGP peer is also a QoS Attribute Speaker, it MAY process the QoS Attribute. If BGP UPDATE message has a QoS Attribute with a list of destination ASes, QoS Attribute Speaker MAY trim the list and adjust the count of the destination AS to exclude ones that are not required in further announcement of BGP UPDATE messages.

A QoS Attribute Speaker MUST drop TCA SubType from the QoS Attribute, if there are no more ASes left in the QoS Attribute's destination list. The rest of the QoS Attribute contents may be forwarded if there exist other SubTypes of QoS Attribute and forwarding rules meet other SubTypes requirements. If there is no other SubTypes in that QoS Attribute content then QoS Attribute Speaker MUST drop the entire QoS Attribute all together. BGP Speaker MAY announce further other attributes and NLRI information, if they meet rules defined by other attributes and BGP specification.

Except extracting the entire TCA SubType of the QoS Attribute and trimming the list of Destination AS list, all other content MUST NOT be modified by any QoS Attribute Speaker or BGP Speaker in the path of a BGP UPDATE message.

5.2. QoS Attribute Handling at Receiver

Once QoS Attribute with the TCA SubType is received at intended receiver (TCA Consumer) , processing of advertised TCA Content is optional for the TCA Consumer. TCA Consumer MAY just trim the Destination AS list as per rules described in this specification, without processing any other content of the Attribute. If Receiver chooses to process advertised TCA content, it may encounter errors beyond the ones described in this document, errors like unavailability of resources if Receiver chooses to implement policies for advertised TCA. In such a case Receiver MAY simply log a message. QoS attribute still MUST be forwarded as per rules defined in this document and rest of the BGP UPDATE message MUST be processed as per BGP specification. If intended receiver is not a QoS Attribute Speaker than BGP Speaker MUST forward this attribute without any change if rest of the BGP UPDATE message also meets forwarding rules as per BGP specification.

When BGP UPDATE messages are triggered only as a result of TCA policy change, propagating BGP UPDATE message beyond intended TCA Consumers is not necessary. If the TCA Consumer device implementations are capable of policy based filtering, it may implement a policy to filter such BGP UPDATE messages based on prefixes and QoS Attribute containing TCA SubType.

6. Error Handling

Error conditions, while processing of the QoS Attribute content, MUST be handled with the approach of attribute discard as described in [RFC7606]. Processing of QoS Attribute content is done by QoS Attribute Speaker and thus in case of errors, resulting in attribute discard, QoS Attribute Speaker SHOULD convey such indication to the BGP Speaker and rest of the BGP message SHOULD be processed by the BGP Speaker as per BGP specification.

7. Deployment Considerations

One of the use cases is for a provider to advertise contracted TCA parameters to a Customer Edge (CE) in cases where eBGP is deployed between PE and CE. The TCA parameters may already be provisioned by the provider on the PE device (facing CE). This provisioned TCA parameters are then advertised thru proposed QoS Attribute to the CE device. The CE device may read the QoS Attribute and TCA SubType content to implement the QoS policy on the device.

Contracted TCA from PE to CE may be full line-rate or sub line-rate or finer granular controlled services. The advertised TCA can be useful when contracted service is sub-rate of a link and/or when for

finer granular traffic classes that are controlled (e.g. voice, video services may be capped to certain rate).

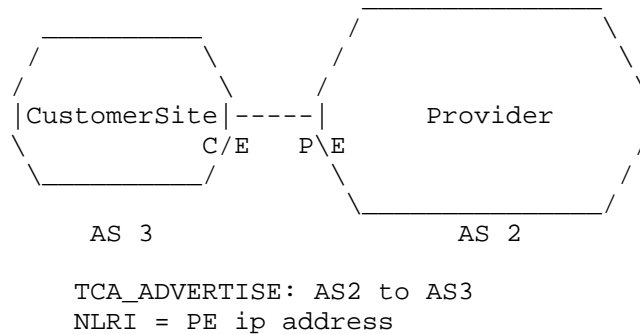


Figure 7: - Example 1

Another use case can be to advertise TCAs among different network sites within one Enterprise network. In Hub and Spoke deployments, Administrator may define TCAs at spoke and advertise QoS TCA parameters to the Hub thru BGP updates. In Figure 7, each spoke (AS1 and AS2) are connected to Hub (AS3) via a VPN tunnel. As shown in Figure 7, AS2 can advertise TCA to AS3 in the context of that tunnel ip address.

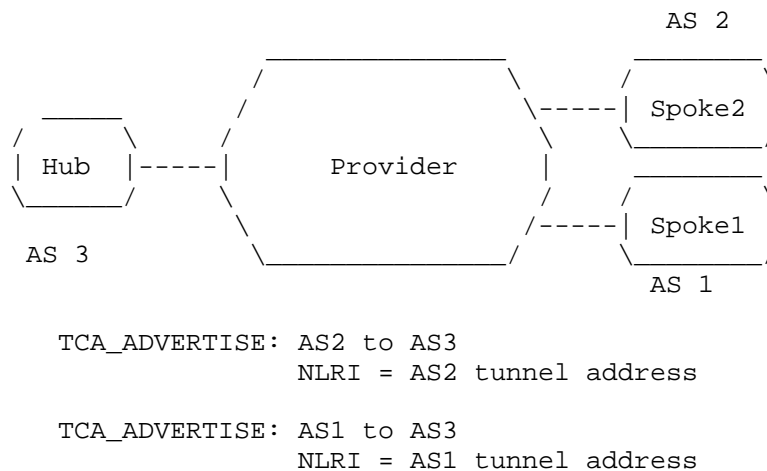


Figure 7 - Example 2

Deployment options are not limited to involving CEs, PE-to-CE or CE-to-CE, only. For any contract between two providers, TCA parameters may be advertised from one to the other.

8. IANA Considerations

This document defines a new BGP optional transitive path attribute, called QoS Attribute. IANA action is required to allocate a new code-point in the BGP path Attributes registry.

IANA is requested to create a registry for QoS Attribute SubTypes. This is a registry of 1 octet value, divided into two pools. One pool of numbers to be assigned on a standards action/early allocation basis. The initial assignments are as shown below. The other pool is for the private use, available range for which is as shown below.

QoS Attribute SubTypes

=====

Reserved	0x00
TCA	0x01
Reserved	0x02-0xf0 (Standards Action)
Private use	0xf1-0xff

IANA is requested to create a registry for QoS Attribute TCA Event Types. This is a registry of 4-bits value, divided into two pools. One pool of numbers to be assigned on a standards action/early allocation basis. One pool of numbers to be assigned on a standards action/early allocation basis. The initial assignments are as shown below. The other pool is for the private use, available range for which is as shown below.

QoS Attribute TCA Event Types

=====

Reserved	0x0
ADVERTISE	0x1
Reserved	0x2 - 0xc (Standards Action)
Private use	0xd - 0xf

IANA is requested to create a registry to define QoS Attribute TCA Direction. This is the direction in forwarding path, advertised QoS TCA is applicable to. This is a 2-bit registry. Values for QoS Attribute TCA direction are:

QoS Attribute TCA Direction

=====

Reserved	0x0
To source AS from destination AS	0x1
From source AS to destination AS	0x2
Reserved (Standards Action)	0x3

QoS Attribute TCA Traffic Class Element Types will be referring to existing IPFIX IANA types as listed in Table 1. While IPFIX registry

is maintained by IANA out of scope of this specification, the use of IPFIX identifiers for this specification are limited to what is described in Table 1. Any new addition of IPFIX identifiers to this table should be a Standards Action.

IANA is requested to create a registry for QoS Attribute TCA Traffic Class Service Types. This is a registry of 2 octet values, to be assigned on a standards action/early allocation basis. The initial assignments are:

Traffic Class Service Type	Value
=====	=====
Reserved	0x00
COMMITTED_TSPEC	0x01
PEAK_TSPEC	0x02
COMMITTED_IN_PROFILE_MARKING	0x03
COMMITTED_OUT_PROFILE_MARKING	0x04
PEAK_OUT_PROFILE_MARKING	0x05
DROP_THRESHOLD	0x06
RELATIVE_PRIORITY	0x07
EFFECTIVE_MAX_RATE	0x08
Standards Action	0x09 - 0x3FFF
FCFS	0x4000 - 0x4FF0

9. Security Considerations

BGP security vulnerabilities analysis is documented in [RFC4272], while BGP-related security considerations are discussed in [RFC4271]. Also, the reader may refer to [RFC7132] for more details about BGP path threat model. Means to prevent route hijacking SHOULD be enabled. Such means include RPKI based origin validation [RFC7115] and BGP Path validation (e.g., [I-D.ietf-sidr-bgpsec-protocol]). Rest of the content in this section discusses additional privacy and security considerations that are applicable to the attribute defined in this document.

The information conveyed in the QoS Attribute TCA SubType reveals sensitive data that should not be exposed publicly to non-authorized parties. Deployment considerations mainly target use of QoS Attribute and TCA SubType in managed networks and those where a trust relationship is in place (Customer to Provider, or Provider to Provider). Administrators MUST disable this attribute to be sent to a remote peer which whom no trust relationship is in place. Both TCA Producer and Consumer SHOULD NOT publish valid TCA IDs to non-authorized nodes.

The attribute may be advertised by a misbehaving node to communicate TCA parameters that are not aligned with the TCA agreements. The enforcement of TCA parameters is outside the scope of this document.

The attribute defined in this document may be used by a misbehaving node for denial-of-service (e.g., inadequately rate-limit or drop some critical traffic). As a mitigation, a BGP peer **MUST** accept this attribute only from trusted BGP peers. For example, ACLs may be configured to identify the trusted ASes that are allowed to send the attribute. Further, administrators of a TCA Consumer's domain are **RECOMMENDED** to generate TCA ID using pseudo-random schemes [RFC4086]. Using robust TCA IDs make it hard to guess a valid TCA.

10. Acknowledgements

Thanks to Fred Baker, David Black, Sue Hares, Benoit Claise and Alvaro Retana for their suggestions and to Christian Jacquenet, Ken Briley, Rahul Patel, Fred Yip, Lou Berger, Brian Carpenter, Bertrand Duvivier, Bruno Decraene, David Black, and Ron Bonica for the review.

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC2212] Shenker, S., Partridge, C., and R. Guerin, "Specification of Guaranteed Quality of Service", RFC 2212, DOI 10.17487/RFC2212, September 1997, <<https://www.rfc-editor.org/info/rfc2212>>.
- [RFC3629] Yergeau, F., "UTF-8, a transformation format of ISO 10646", STD 63, RFC 3629, DOI 10.17487/RFC3629, November 2003, <<https://www.rfc-editor.org/info/rfc3629>>.
- [RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <<https://www.rfc-editor.org/info/rfc4271>>.
- [RFC4506] Eisler, M., Ed., "XDR: External Data Representation Standard", STD 67, RFC 4506, DOI 10.17487/RFC4506, May 2006, <<https://www.rfc-editor.org/info/rfc4506>>.

- [RFC7012] Claise, B., Ed. and B. Trammell, Ed., "Information Model for IP Flow Information Export (IPFIX)", RFC 7012, DOI 10.17487/RFC7012, September 2013, <<https://www.rfc-editor.org/info/rfc7012>>.
- [RFC7115] Bush, R., "Origin Validation Operation Based on the Resource Public Key Infrastructure (RPKI)", BCP 185, RFC 7115, DOI 10.17487/RFC7115, January 2014, <<https://www.rfc-editor.org/info/rfc7115>>.
- [RFC7606] Chen, E., Ed., Scudder, J., Ed., Mohapatra, P., and K. Patel, "Revised Error Handling for BGP UPDATE Messages", RFC 7606, DOI 10.17487/RFC7606, August 2015, <<https://www.rfc-editor.org/info/rfc7606>>.

11.2. Informative References

- [I-D.ietf-sidr-bgpsec-protocol] Lepinski, M. and K. Sriram, "BGPsec Protocol Specification", draft-ietf-sidr-bgpsec-protocol-22 (work in progress), January 2017.
- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, DOI 10.17487/RFC2475, December 1998, <<https://www.rfc-editor.org/info/rfc2475>>.
- [RFC4086] Eastlake 3rd, D., Schiller, J., and S. Crocker, "Randomness Requirements for Security", BCP 106, RFC 4086, DOI 10.17487/RFC4086, June 2005, <<https://www.rfc-editor.org/info/rfc4086>>.
- [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <<https://www.rfc-editor.org/info/rfc4272>>.
- [RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, DOI 10.17487/RFC5575, August 2009, <<https://www.rfc-editor.org/info/rfc5575>>.
- [RFC6020] Bjorklund, M., Ed., "YANG - A Data Modeling Language for the Network Configuration Protocol (NETCONF)", RFC 6020, DOI 10.17487/RFC6020, October 2010, <<https://www.rfc-editor.org/info/rfc6020>>.

- [RFC6241] Enns, R., Ed., Bjorklund, M., Ed., Schoenwaelder, J., Ed., and A. Bierman, Ed., "Network Configuration Protocol (NETCONF)", RFC 6241, DOI 10.17487/RFC6241, June 2011, <<https://www.rfc-editor.org/info/rfc6241>>.
- [RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, DOI 10.17487/RFC6793, December 2012, <<https://www.rfc-editor.org/info/rfc6793>>.
- [RFC7132] Kent, S. and A. Chi, "Threat Model for BGP Path Security", RFC 7132, DOI 10.17487/RFC7132, February 2014, <<https://www.rfc-editor.org/info/rfc7132>>.
- [RFC7297] Boucadair, M., Jacquenet, C., and N. Wang, "IP Connectivity Provisioning Profile (CPP)", RFC 7297, DOI 10.17487/RFC7297, July 2014, <<https://www.rfc-editor.org/info/rfc7297>>.
- [RFC7674] Haas, J., Ed., "Clarification of the Flowspec Redirect Extended Community", RFC 7674, DOI 10.17487/RFC7674, October 2015, <<https://www.rfc-editor.org/info/rfc7674>>.
- [RFC8040] Bierman, A., Bjorklund, M., and K. Watsen, "RESTCONF Protocol", RFC 8040, DOI 10.17487/RFC8040, January 2017, <<https://www.rfc-editor.org/info/rfc8040>>.

Authors' Addresses

Shitanshu Shah

Email: shitanshu_shah@hotmail.com

Keyur Patel
Arrcus, Inc

Email: keyur@arrcus.com

Sandeep Bajaj
Viptela

Luis Tomotaki
Verizon
400 International
Richardson, TX 75081
US

Email: luis.tomotaki@verizon.com

Mohamed Boucadair
Orange
Rennes
35000
France

Email: mohamed.boucadair@orange.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 23, 2014

Z. Li
M. Chen
S. Zhuang
Huawei Technologies
October 20, 2013

An Architecture of Central Controlled Border Gateway Protocol (BGP)
draft-li-idr-cc-bgp-arch-00

Abstract

As the Software Defined Networks (SDN) solution develops, BGP is extended to support central control. This document introduces an architecture of using BGP for central controlling. Some use cases under this new framework are also discussed. For specific use cases, making necessary extensions in BGP are required.

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 23, 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. Terminology	3
3. Architecture	4
3.1. Reference Model	4
3.2. Deployment Mode	4
3.3. Requirement of Protocol Extensions	5
3.3.1. Building Connectivity	5
3.3.2. Roles Auto-Discovery	6
3.3.3. Establishing BGP Sessions	6
3.3.4. Capability Negotiation	6
3.3.5. High Availability	6
3.3.6. Security	7
4. Use Cases	7
4.1. Network Topology Acquisition	7
4.2. Simplifying Network Operation and Maintenance	7
4.3. MPLS Global Label Allocation	8
4.4. RR-Based Traffic Steering	8
4.5. Inter-Controller Applications	9
5. IANA Considerations	9
6. Security Considerations	9
7. References	10
7.1. Normative References	10
7.2. Informative References	10
Authors' Addresses	11

1. Introduction

The Border Gateway Protocol (BGP) defined in [RFC 4271], is well-known as Internet inter-domain routing protocol. Multiprotocol BGP (MP-BGP) framework defined in [RFC 4760], is an extension to BGP that enables BGP to carry routing information for multiple network layers and address families. The current MP-BGP specification can provide a rich service set within a BGP enabled network, such as L2VPN, L3VPN, MVPN, EVPN, etc.

There have been some BGP-based central controlled applications, such as BGP RR [RFC4456]. A route reflector (RR) is a network routing component. The purpose of the RR is concentration in that way all

Client's forwarding routes are exchanged by the central RR Router. It offers an alternative to the logical full-mesh requirement of internal border gateway protocol (IBGP). A RR acts as a focal point for IBGP sessions. Multiple IBGP routers can only peer with a central point, rather than peer with every other router in a full mesh manner. All the other IBGP routers become route reflector clients.

With the emergence of Software Defined Networks (SDN), BGP plays as an important part in a central controlled environment.

1 Building a central controlled framework for Controller and its Clients, BGP can be used to communicate between Controller and its Clients, Controller and other Controllers. The information carried by BGP includes network and service topology, network and service forwarding entries etc.

2 Many new applications are emerging under the centrally-controlled framework, such as network virtualization, global traffic engineering etc. Some new applications bring extension requirements to BGP.

This document defines an architecture of Central Controlled BGP and then use cases under this framework are described. For some use cases requirements of BGP extensions are discussed.

2. Terminology

BGP: Border Gateway Protocol

EVPN: Ethernet VPN

L2VPN: Layer 2 VPN

L3VPN: Layer 3 VPN

MVPN: Multicast VPN

RR: Route Reflector

SDN: Software-Defined Network

S-EVPN: Segment-based EVPN

3. Architecture

3.1. Reference Model

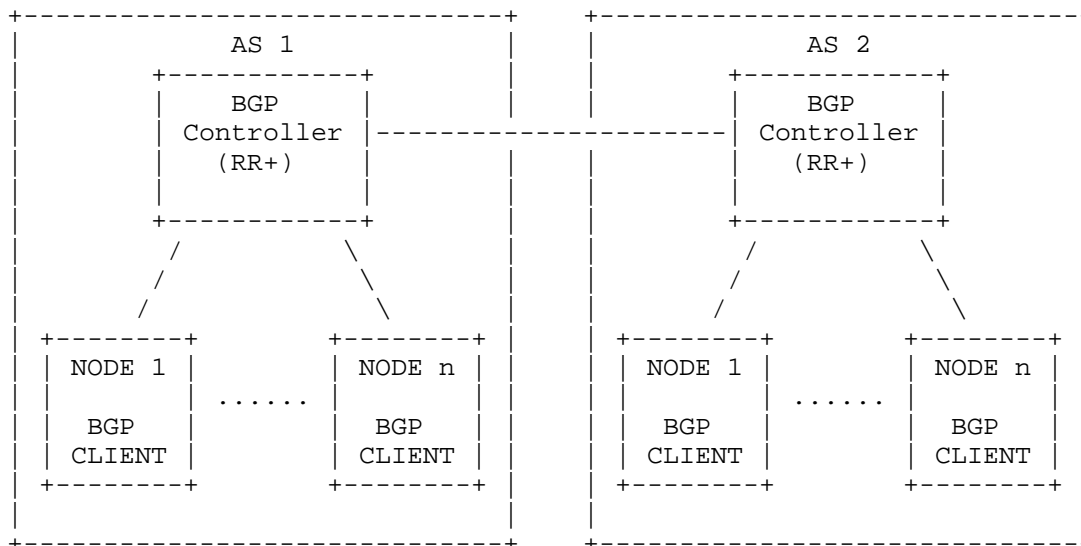


Figure 1: An Architecture of Central Controlled BGP

The figure above depicts a typical architecture of central controlled BGP. It consists of two essential network elements: BGP Controller and BGP Client. BGP Controller controls all the BGP Clients within its administrative domain by communicating with them.

In the above framework, BGP Controller is placed at the same position of traditional Route Reflector (RR) [RFC4456]. Moreover from point of view of implementation BGP Controller can be considered as function-enhanced RR. So in this document, BGP Controller is named RR+ as well.

3.2. Deployment Mode

BGP Controller can run on a general-purpose server or a network device. If BGP Controller runs on a network device, it MUST support both central-controlled functionality and forwarding functionality. Same as BGP Controller, BGP Client can run on a general-purpose server or a network device. It is more meaningful to decouple control plane and forwarding functionality on BGP Client because this manner enables network devices focusing on forwarding functionality. This deployment model is shown in the following figure:

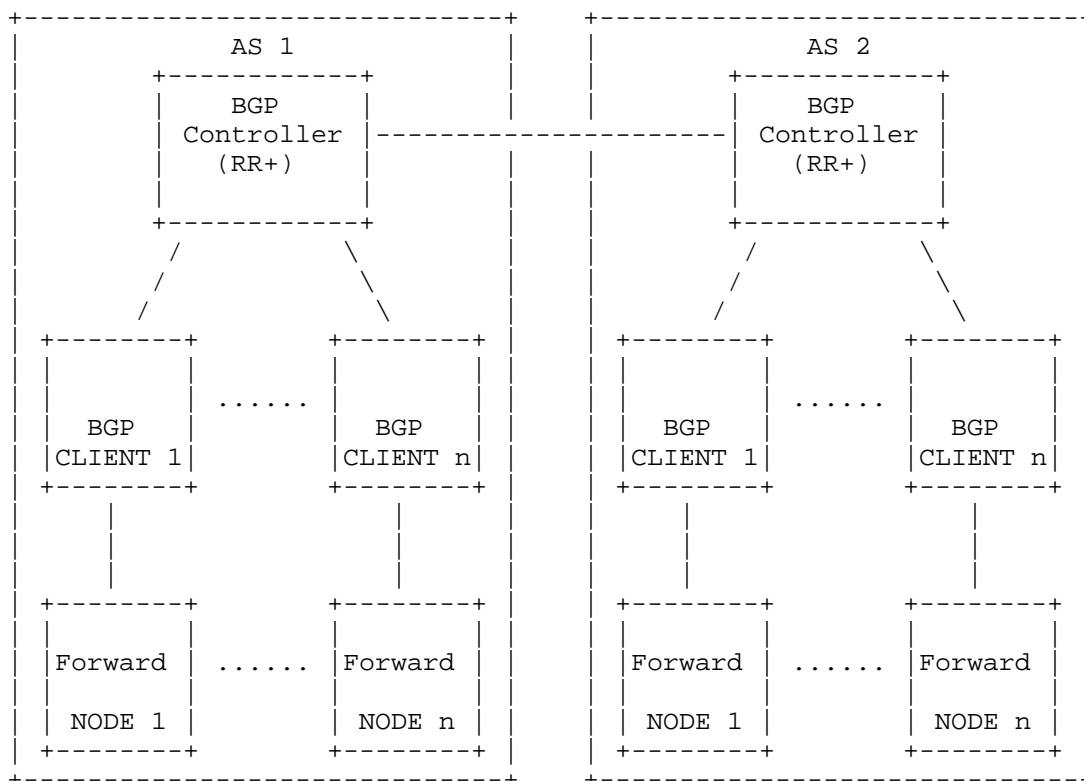


Figure 2: Decoupling BGP Client and Forwarding

In the reference model, there are multiple BGP controllers in multiple ASes. In fact in one AS, there can also be multiple BGP controllers to control different sets of BGP clients and IBGP peers are set up between these BGP controllers. Such application scenario can refer to [I-D.ietf-mppls-seamless-mppls] in which there are multiple IGP areas which runs BGP in one AS. This document focuses on multiple BGP controller in different ASes. The requirement and use cases can also be applied to the case of multiple BGP controller in one AS.

3.3. Requirement of Protocol Extensions

Building a BGP-based Central Controlled Framework needs extensions to IGP, BGP and I2RS etc.

3.3.1. Building Connectivity

Connectivity between BGP Controller and BGP Clients in an AS can be built by extending IGP protocol. In order to simplify network operations, such connectivity SHOULD be automatically established.

3.3.2. Roles Auto-Discovery

A BGP-based Central Controlled Framework consists of two basic roles: BGP Controller and BGP Client. Such roles can be auto-discovered by extending IGP protocol to flooding the role information within an AS. When IGP has finished the flooding process of role information, BGP Controller and BGP Client can establish a BGP session on demand.

3.3.3. Establishing BGP Sessions

For the intra-AS case, when IGP has finished the flooding process of role information within an AS, BGP Controller and BGP Client can automatically establish a BGP session. It is not necessary to establish BGP sessions amongst BGP Clients.

For the inter-AS case, the peer BGP controller SHOULD be specified to establish BGP sessions.

3.3.4. Capability Negotiation

In order for BGP Controller and BGP Client to support BGP-based Central Controlled framework in a friendly way, this document suggests to defines a new BGP Central Control Capability. The Central Control Capability SHOULD be defined as per [RFC5492]. By advertising the BGP Central Control Capability to a peer, a BGP speaker conveys information if it is able to send, receive, and properly handle BGP Central Control related processes.

The existing BGP capabilities should be kept in the Central Controlled framework and other new capabilities should be extended according to new applications based on the Central Controlled framework.

3.3.5. High Availability

In the BGP-based Central Controlled framework, BGP Controller plays a key role. To void one-point-failure of BGP Controller, it is possible to run redundant BGP Controllers for high availability.

With multiple BGP Controllers, it is important to synchronize route processing policy configuration for all of them to perform the exactly same routing decisions. When Primary BGP Controller failed, the Backup BGP Controllers will take over the work of the Primary BGP Controller.

To ensure BGP route persistence in case of occurrence of BGP Controller failure, the new Primary BGP Controller SHOULD perform resynchronization with BGP Clients.

When BGP Client loses connection with Primary BGP Controller, it SHOULD following BGP Graceful Restart routine defined in [RFC 4724] similar as a GR Helper.

3.3.6. Security

In BGP-based Central Controlled framework, it SHOULD be ensured that communications between BGP Controllers and BGP Clients conform to network security policy. The communication key used on BGP Client can be configured through I2RS or other way.

4. Use Cases

In BGP-based Central Controlled framework, new use cases which are difficult to be supported in traditional networks are emerging. In some specific use cases, extension and enhancement of BGP protocol are necessary.

4.1. Network Topology Acquisition

In BGP-based Central Controlled framework, BGP Controller can get the topology of the whole network. Some applications such as ALTO can get network topology information from BGP Controller. The topology information of one AS can also be advertised by the controller to the other BGP Controller in the other AS. BGP has been extended to distribute link-state and traffic engineering information as defined in [I-D.ietf-idr-ls-distribution].

4.2. Simplifying Network Operation and Maintenance

The adoption of the new BGP-based Central Controlled Framework can reduce the complexity and effort of network operation and maintenance by following manners:

1. By using I2RS APIs, it would allow network operator to setup BGP policy configuration from a single central point. This helps avoid manual configuration of BGP policy on multiple BGP Clients and reduce the complexity of BGP policy configuration.

2. For network with VPN service which includes L3VPN, L2VPN, E-VPN etc, BGP Controller COULD store all the VPN user information. Use of I2RS APIs to set L3VPN configuration from BGP Controller would allow network operator no need to configure a VPN many times on different BGP Clients. Furthermore, in the new Central Controlled framework

VPN parameters such as Route Distinguisher (RD), Route Targets (RT) can be automatically generated and the configuration between CE and PE can be generated automatically after the CE is authenticated by the Central Controller. This can simplify network operations and maintenance greatly.

4.3. MPLS Global Label Allocation

MPLS Global Label should be allocated in a central point to guarantee all distributed network nodes can understand meaning of a specific global label in same. The new BGP-based Central Controlled framework is particularly suitable to allocate MPLS Global Label through some necessary MP-BGP extensions.

MPLS Global Label is defined in [I-D.li-mppls-global-label-framework] and related use cases are defined in [I-D.li-mppls-global-label-usecases].

The extensions of BGP for MPLS global label include:

1. A new BGP Capability called Global Label Capability is suggested to be introduced by following [RFC5492]. BGP Controller can negotiate with BGP client on this new BGP capability.
2. BGP Controller determines the COMMON label space for all its BGP Clients.
3. For each BGP client, BGP Controller allocates different MPLS Global Labels for different services and advertises the MPLS Global Labels to the BGP Client.
4. BGP Client receives the MPLS Global Labels, and generates corresponding MPLS forwarding entries.

Many types of services such as VPLS[RFC4761], Multicast VPN[RFC6514], E-VPN[I-D.ietf-l2vpn-evpn] are based on MP-BGP. So making extensions to MP-BGP to allocate MPLS global label for these services is a nature way from point of network solution. The use cases of MPLS global label defined in [I-D.li-mppls-global-label-usecases] includes S-EVPN, Split Horizon in VPLS MCAST and BGP MVPN, Egress PE protection in VPN, etc.

4.4. RR-Based Traffic Steering

RR-based Traffic Steering (RRTS) defined in [I-D.chen-idr-rr-based-traffic-steering-usecase], is an idea that leverages the BGP route reflection mechanism to realize traffic steering in the network, therefore the operators can conduct specific traffic to traverse

specific path, domains and/or planes as demand. The essential of RRTS is that the concept of traffic engineering is introduced into BGP network. With the new BGP-based Central Controlled framework defined in this document, the operators can steer the network traffic easily. More detailed description could be found in [I-D.chen-idr-rr-based-traffic-steering-usecase].

4.5. Inter-Controller Applications

Information is communicated between the BGP controller to fulfill the inter-AS applications such as inter-AS VPN. Besides the inter-AS applications, there proposes a type of new application to communicate control information only between BGP Controllers to set up service directly and download necessary forwarding entry to the nodes. Thus the BGP sessions between the Controller and the node can be saved and the control functionality on the node can be saved further. This type of model is shown in the following figure. In this model, the service set up between the nodes is proxied by the BGP Controllers.

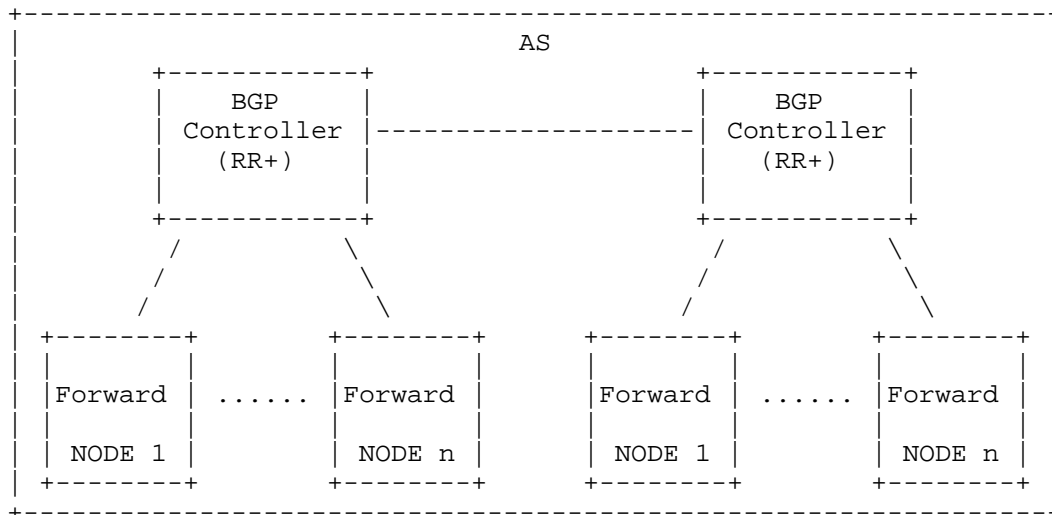


Figure 3: Removing BGP Session between Controller and NODE

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

TBD.

7. References

7.1. Normative References

- [I-D.ietf-l2vpn-evpn]
Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-04 (work in progress), July 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.
- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, April 2006.
- [RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, February 2009.
- [RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [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.

7.2. Informative References

- [I-D.chen-idr-rr-based-traffic-steering-usecase]
Chen, M., Zhuang, S., Zhu, Y., and S. Wang, "Use Cases of Route Reflection based Traffic Steering", draft-chen-idr-rr-based-traffic-steering-usecase-00 (work in progress), July 2013.
- [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.
- [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-04 (work in progress), July 2013.
- [I-D.li-l2vpn-segment-evpn]
Li, Z., Yong, L., and J. Zhang, "Segment-Based EVPN(S-EVPN)", draft-li-l2vpn-segment-evpn-00 (work in progress), July 2013.
- [I-D.li-mpls-global-label-framework]
Li, Z., Zhao, Q., and T. Yang, "A Framework of MPLS Global Label", draft-li-mpls-global-label-framework-00 (work in progress), July 2013.
- [I-D.li-mpls-global-label-usecases]
Li, Z., Zhao, Q., and T. Yang, "Usecases of MPLS Global Label", draft-li-mpls-global-label-usecases-00 (work in progress), July 2013.

Authors' Addresses

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

Email: lizhenbin@huawei.com

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

Email: mach.chen@huawei.com

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

Email: zhuangshunwan@huawei.com

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

Q. Wu
D. Wang
Huawei
S. Previdi
Cisco
H. Gredler
Juniper
S. Ray
Cisco
October 21, 2013

BGP attribute for North-Bound Distribution of Traffic Engineering (TE)
performance Metrics
draft-wu-idr-te-pm-bgp-03

Abstract

In order to populate network performance information like link latency, latency variation, packet loss and bandwidth into Traffic Engineering Database(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, available bandwidth and utilized bandwidth).

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	3
2. Conventions used in this document	4
3. Use Cases	5
3.1. MPLS-TE with PCE	5
3.2. ALTO Server Network API	5
4. Carrying TE Performance information in BGP	7
5. Attribute TLV Details	9
6. Security Considerations	10
7. IANA Considerations	11
8. References	12
8.1. Normative References	12
8.2. Informative References	12
Appendix A. Change Log	13
A.1. draft-wu-idr-te-pm-bgp-03	13
A.2. draft-wu-idr-te-pm-bgp-02	13
Authors' Addresses	14

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-pce-pcep-service-aware]. [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-connection 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 the end to end path computation ([I-D.ietf-pce-pcep-service-aware]) and selection which enable establishing segment overlay tunnel between overlay nodes and stitching them together to compute end to end path.

In order to populate network performance information like link latency, latency variation, packet loss and bandwidth 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, available bandwidth, and utilized bandwidth).

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

In inter-AS path computation, PCE in each AS participant in different IGP. In Hierarchy of PCE, A child PCE must be configured with the address of its parent PCE[RFC6805]. Configuration system is challenged by handling changes in parent PCE identities and coping with failure events, especially when parent PCE and child PCE are not a part of the same routing domain.

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.

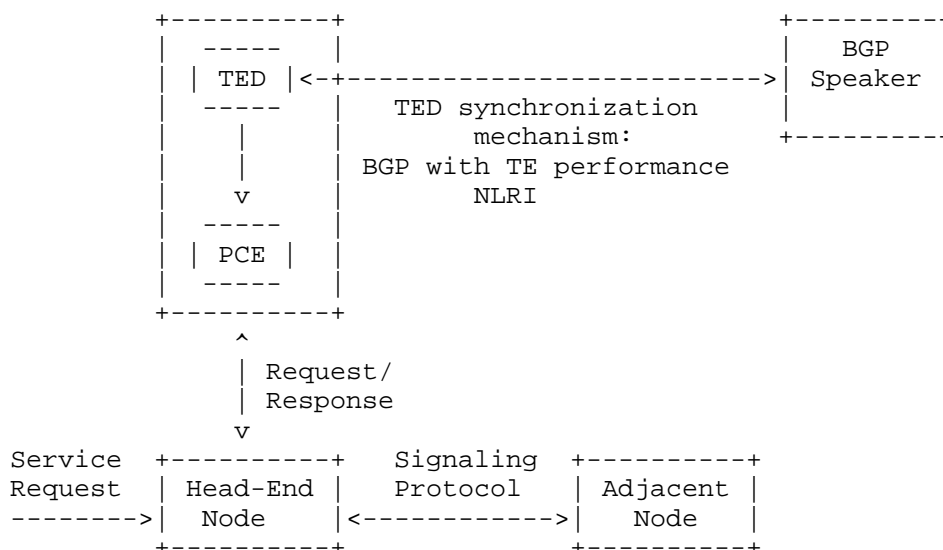


Figure 1: External PCE node using a TED synchronization mechanism

3.2. ALTO Server Network API

The ALTO Server can aggregate information from multiple systems to provide an abstract and unified view that can be more useful to applications.

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.

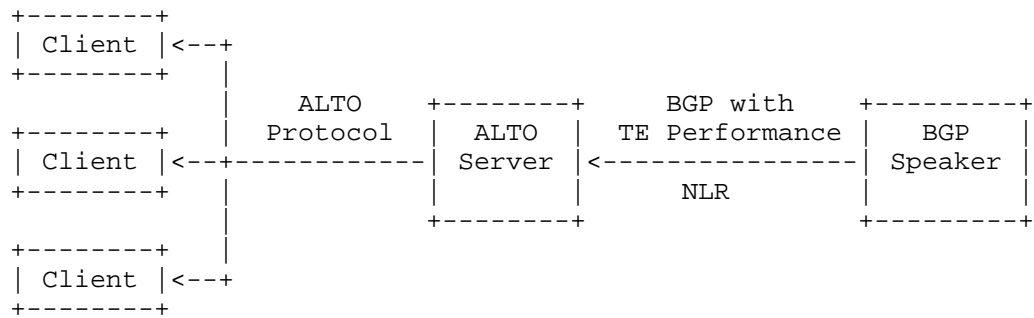


Figure 2: ALTO Server using network performance information

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

Type	Value
TBD1	Unidirectional Link Delay
TBD2	Min/Max Unidirectional Link Delay
TBD3	Unidirectional Delay Variation
TBD4	Unidirectional Packet Loss
TBD5	Unidirectional Residual Bandwidth
TBD6	Unidirectional Available Bandwidth
TBD7	Unidirectional Utilized Bandwidth

As can be seen in the list above, the TLVs described in this document carry different types of network performance information. These TLVs include a bit called the Anomalous (or "A") bit at the left-most bit after length field of each TLV. The other bits in the first octets after length field of each TLV is reserved for future use. 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 exceeds configurable maximum thresholds, a TLV with the A bit set is advertised. These TLVs could be used by the receiving BGP peer 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 falls below a configurable value, that TLV can be re-advertised with the Anomalous bit cleared. In this case, a receiving BGP peer 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 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 [ISIS-TE-METRIC], the bandwidth advertisements, the delay and delay variation advertisements, packetloss defined in this document MUST be encoded in the same unit as one defined in IS-IS Extended IS Reachability sub-TLVs [ISIS-TE-METRIC]. 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 defined in section 3.2.2 of [I-D.ietf-idr-ls-distribution] 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]. 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:
xxxxx	Unidirectional Link Delay	22/xx	[ISIS-TE]/4.1
xxxxx	Min/Max Unidirectional Link Delay	22/xx	[ISIS-TE]/4.2
xxxxx	Unidirectional Delay Variation	22/xx	[ISIS-TE]/4.3
xxxxx	Unidirectional Link Loss	22/xx	[ISIS-TE]/4.4
xxxxx	Unidirectional Residual Bandwidth	22/xx	[ISIS-TE]/4.5
xxxxx	Unidirectional Available Bandwidth	22/xx	[ISIS-TE]/4.6
xxxxx	Unidirectional Utilized Bandwidth	22/xx	[ISIS-TE]/4.7

Table 1: Link Attribute TLVs

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.
- [I-D.ietf-pce-pcep-service-aware]
Dhruv, D., "Extensions to the Path Computation Element Communication Protocol (PCEP) to compute service aware Label Switched Path (LSP)",
ID draft-ietf-pce-pcep-service-aware-01, July 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.

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.

Appendix A. Change Log

Note to the RFC-Editor: please remove this section prior to publication as an RFC.

A.1. draft-wu-idr-te-pm-bgp-03

The following are the major changes compared to previous version 02:

- o Add unidirectional utilized bandwidth metric as the seventh metric Carried in a new BGP attribute.

A.2. draft-wu-idr-te-pm-bgp-02

The following are the major changes compared to previous version 01:

- o Taking out link utilization metric and channel throughput metric from this version and will add link utilization metric back to the update when there was agreement on what measurement unit is used for link utilization.
- o Some additional texts in BGP extension section 4 to explain how to position 'A' bit in the BGP TE performance TLV.
- o Add two editor notes to explain the status of this draft and open issue that need be resolved.
- o Some additional text in the use case sections to clarify how to use these TE performance metrics.

Authors' Addresses

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

Email: sunseawq@huawei.com

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

Email: wangdanhua@huawei.com

Stefano Previdi
Cisco Systems, Inc.
Via Del Serafico 200
Rome 00191
Italy

Email: sprevidi@cisco.com

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

Email: hannes@juniper.net

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

Email: sairay@cisco.com

