Workgroup: Network Working Group
Internet-Draft: draft-ietf-idr-rpd-16
Published: 14 February 2023
Intended Status: Standards Track
Expires: 18 August 2023
Authors: Z. Li
Huawei
L. Ou
China Telcom Co., Ltd.
Y. Luo
China Telcom Co., Ltd.
G. Mishra
Verizon Inc.
H. Chen
Futurewei
H. Wang
Huawei

BGP Extensions for Routing Policy Distribution (RPD)

Abstract

It is hard to adjust traffic in a traditional IP network from time to time through manual configurations. It is desirable to have a mechanism for setting up routing policies, which adjusts traffic automatically. This document describes BGP Extensions for Routing Policy Distribution (BGP RPD) to support this with a controller.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

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 18 August 2023.

Copyright Notice

Table of Contents

1. Introduction

Providers have the requirement to adjust their business traffic from time to time in a number of cases including:

*Link congestion and overload caused by a network failure such as a link or node failure, or a live event such as a world cup.

*Poor network transmission quality as the result of traffic delay or loss in some part of a network.

*Some unused network resources such as idle links because of business changes or network additions.

To adjust the traffic flowing to a destination (or adjust traffic for short) is to move the traffic from a overloaded path to another lightly used path. The move keeps the quality of the traffic transmission and uses the network resources optimally.

It is difficult to adjust traffic in a traditional IP network where an operator configures routing policies using command lines or configuration files. Traffic can only be adjusted device by device. All the routers that the traffic traverses need to be reconfigured.

Using a configuration automation system for adjusting traffic affects network performance when the number of routers the traffic may traverse is big. The system has to keep its connections live to all these routers. This consumes network resources.

It is desirable to have an automatic mechanism for setting up routing policies to adjust traffic, which is simple and efficient. This document describes extensions to BGP for Routing Policy Distribution (RPD) for this mechanism with a controller.

2. Terminology

The following terminology is used in this document.

*AFI: Address Family Identifier

*SAFI: Subsequent Address Family Identifier

*MED: Multiple Exit Discriminator (or MULTI_EXIT_DISC)

*RPD: Routing Policy Distribution

3. An Example of Traffic Adjustment

Figure 1 illustrates a simple scenario, where RPD is used by a controller with a Route Reflector (RR) to adjust traffic automatically.

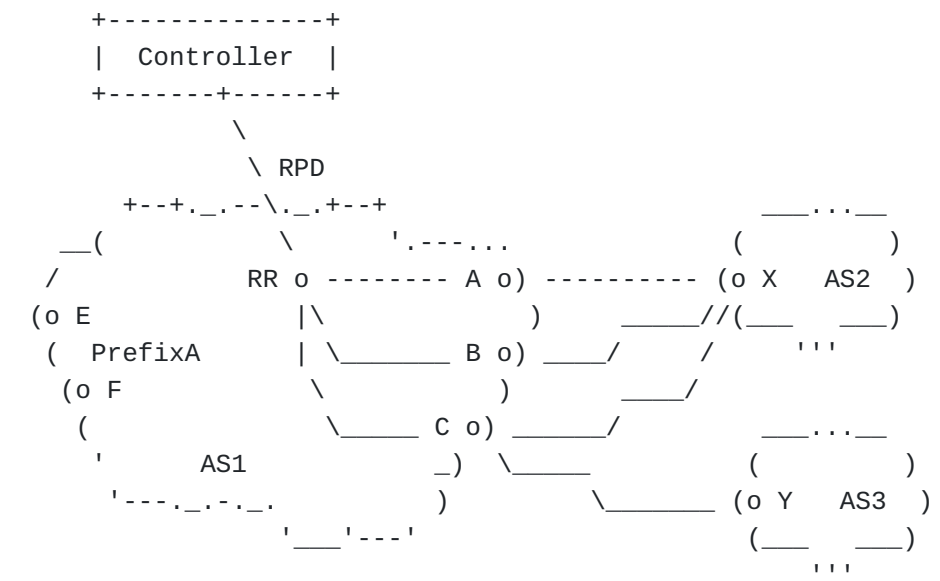

Figure 1: Controller with RR Adjusts Traffic

AS1, AS2 and AS3 belong to provider P1, P2 and P3 respectively. Routers A, B and C are in AS1. Router X is in AS2. There is a BGP session between X and each of routers A, B and C. Router Y is in AS3. There is a BGP session between Y and router C.

AS1 has an IP address prefix named PrefixA, which is advertised to AS2 from AS1. Provider P1 of AS1 wants to adjust the traffic to PrefixA from AS2 automatically. For the traffic to PrefixA from AS2 via link X--A, once link X--A gets congested, P1 wants to move the traffic to link X--B, which is lightly used.

The controller peers with the RR using a BGP session. There is a BGP session between the RR and each of routers A, B and C in AS1, which are shown in the figure. Other sessions in AS1 are not shown in the figure.

The controller obtains the information about traffic flows including the traffic flow to PrefixA. When it decides that the traffic to PrefixA needs to be moved from link X--A to link X--B from the information, it sends a RPD routing policy to A or B for changing MED attribute in the IP route with PrefixA, which is advertised to AS2. Router X in AS2 moves the traffic to link X--B after receiving the IP route with PrefixA having the changed attribute. (Note: how the controller gets the information and makes decision is out of scope of this document).

Suppose that MED of the IP unicast route with PrefixA sent to X by A, B and C is 50, 100 and 150 respectively. To move the traffic to PrefixA in AS1 from link X--A to X--B, the controller sends a RPD routing policy to A. After receiving the RPD routing policy, router

A sends the IP unicast route with PrefixA in AS1 to router X in AS2 and changes the MED to 160 before sending the IP route.

The RPD routing policy includes:

*Peer IP = the IP address of router X,

*Match conditions: prefix matching PrefixA exact and AS_PATH matching AS1, and

*Action: set MED to 160.

After receiving the RPD routing policy, router A sets the MED to 160 for the IP unicast route with PrefixA in AS1 and sends the IP unicast route to router X. The IP unicast route sent to X from A, B and C has MED 160, 100 and 150 respectively. Router X sends the traffic to PrefixA using link X--B since MED 100 from B is the smallest.

4. Protocol Extensions

This document specifies a solution using a new <AFI, SAFI>[RFC4760] with the BGP Wide Community [I-D.ietf-idr-wide-bgp-communities] for encoding and distributing a routing policy. This routing policy is called a RPD routing policy.

4.1. Using a New <AFI, SAFI>

A new <AFI, SAFI> pair is defined, where the Routing Policy SAFI has codepoint 75, and the AFI MUST be IPv4(1) or IPv6(2). This new pair is called RPD <AFI, SAFI>.

The RPD <AFI, SAFI> uses a new Network Layer Reachability Information (NLRI) defined as follows:

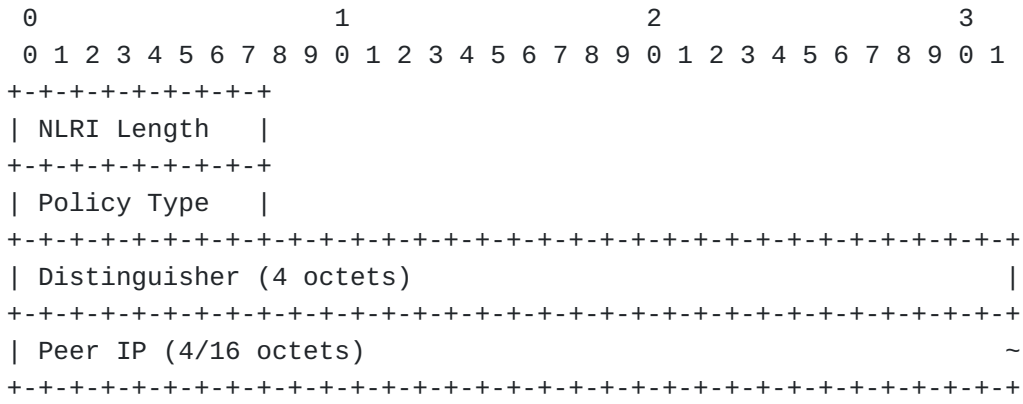

Figure 2: NLRI of RPD <AFI, SAFI>

Where:

NLRI Length: 1 octet represents the length of NLRI in octets as defined in [RFC4760]. If the Length is anything other than 9 or 21 octets, the NLRI is corrupt and the enclosing UPDATE message MUST be ignored.

Policy Type: 1 octet indicates the type of a policy. 1 is for Export policy. If the Policy Type is any other value, the NLRI is corrupt and the enclosing UPDATE message MUST be ignored.

Distinguisher: 4 octet unsigned integer that uniquely identifies the content/policy. It is used to sort/order the polices from the lower to higher Distinguisher. They are applied in ascending order. A policy with a lower/smaller Distinguisher is applied before the policies with a higher/larger Distinguisher.

Peer IP: 4/16 octet value indicates IPv4/IPv6 peers. Its default value is 0. If the value is not a valid IP address and not 0, the NLRI is corrupt and the enclosing UPDATE message MUST be ignored.

The NLRI of RPD <AFI, SAFI> is carried in an MP_REACH_NLRI attribute in a BGP UPDATE message. The "Length of Next Hop Network Address" field of the MP_REACH_NLRI attribute MUST be set to zero.

The RPD routing policies in the UPDATE messages received are stored under the RPD <AFI, SAFI>. Before advertising an IPv4/IPv6 Unicast route (IP route for short), a BGP speaker MUST apply the routing policies to the route.

The content of the Routing Policy is encoded in a BGP Wide Community.

4.2. Atoms

This section defines three Atoms. For your reference, the format of the Atoms is illustrated below:

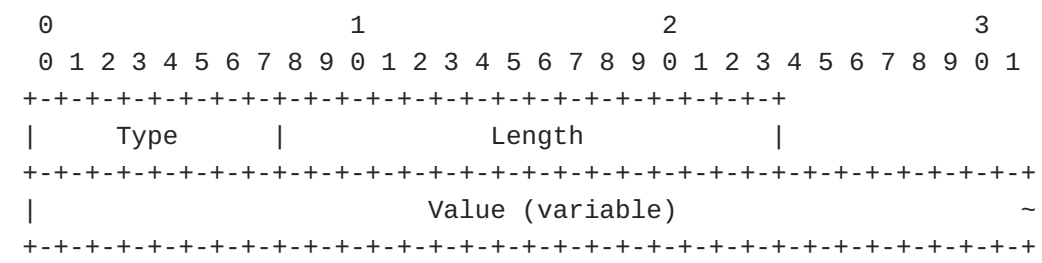

Figure 3: Format of Atoms TLVs

4.2.1. Atom Type TBD1, The Route Attributes (RouteAttr)

A RouteAttr Atom TLV (or RouteAttr Atom for short) specifies one or two groups of conditions. The first group of conditions states a set of IPv4/IPv6 address prefix ranges. The second group identifies a list of route attributes. The Atom has the following format.

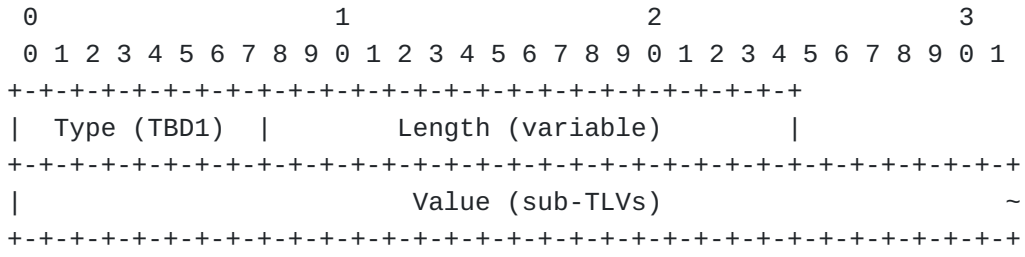

Figure 4: Format of RouteAttr Atom TLV

The Type for RouteAttr Atom is TBD1.

In RouteAttr Atom, four sub-TLVs are defined: IPv4 Address Prefix Range List, IPv6 Address Prefix Range List, AS_PATH RegEx, and Community List sub-TLV. The first two state IPv4 and IPv6 address prefix ranges respectively. The last two identify AS_PATH and Community attributes respectively. Each of these sub-TLVs has the format as follows.

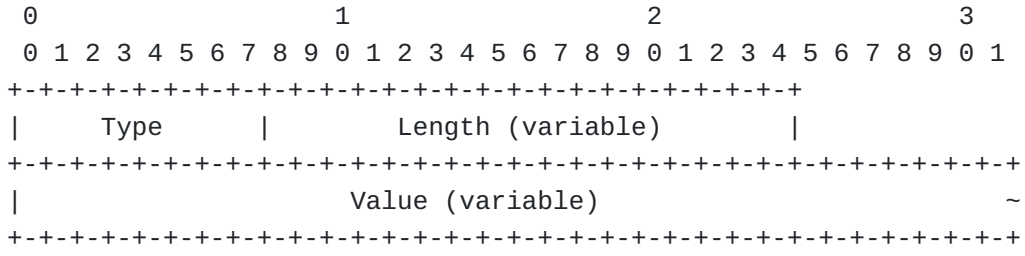

Figure 5: Format of sub-TLV in RouteAttr Atom

4.2.1.1. IPv4 Address Prefix Range List sub-TLV

The IPv4 Address Prefix Range List sub-TLV contains a list of IPv4 address prefix ranges. Each range describes an IPv4 address prefix or group of Pv4 address prefixes and is represented by a tuple <M-Type, IPv4 Address, Prefix Length, PL-Lower-Bound, PL-Upper-Bound>, where PL is short for prefix length. Its format is illustrated below:

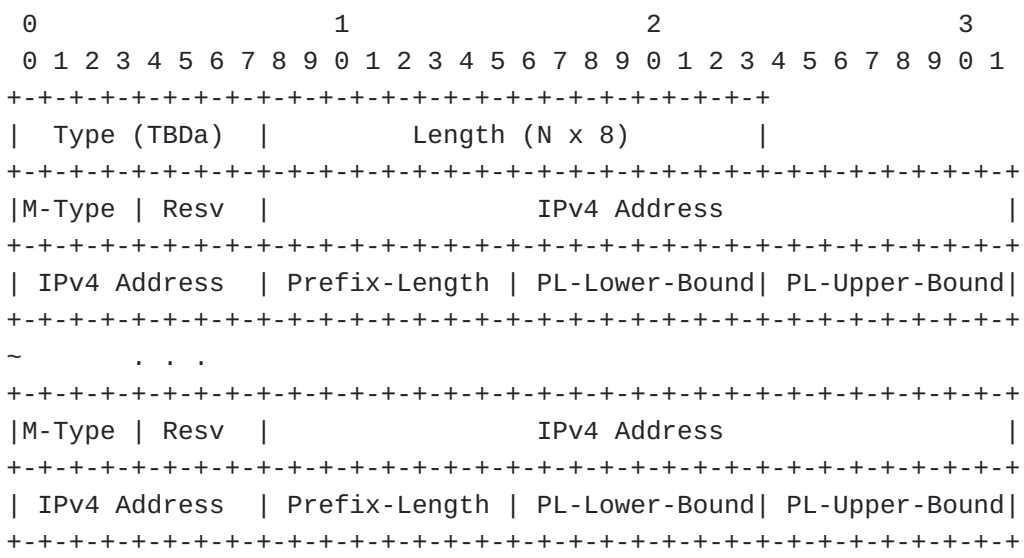

Figure 6: Format of IPv4 Address Prefix Range List sub-TLV

Type: The Type for IPv4 Address Prefix Range List is TBDa.

Length: N x 8, where N is the number of IPv4 address prefix ranges in the sub-TLV. If Length is not a multiple of 8, the Atom is corrupt and the enclosing UPDATE message MUST be ignored.

Resv: 4 bits. They MUST be sent as zero and ignored on receipt.

IPv4 Address: 4 octets that describe an IPv4 prefix. This field, together with the Prefix-Length follows the same semantics as the NLRI encoding from [RFC4271], except that the trailing bits in the IPv4 Address fill the 4-octet field.

Prefix-Length: 1 octet field that represents the Prefix Length of the IPv4 Address, as specified in [RFC4271].

PL-Lower-Bound: 1-octet field that represents the lower bound of the IPv4 Address's prefix length. This field MUST be greater than, or equal to, the Prefix-Length, or be 0. If this field is less than the Prefix-Length and not 0, the enclosing UPDATE message MUST be ignored.

PL-Upper-Bound: 1-octet field that represents the upper bound of the IPv4 Address's prefix length. This field MUST be greater than, or equal to, the Prefix-Length, or be 0. If this field is less than the Prefix-Length and not 0, the enclosing UPDATE message MUST be ignored.

M-Type: 4-bit field specifying the IPv4 address prefix range format type. The values are specified below.

M-Type = 0:

The IPv4 address prefix described corresponds to the IPv4 Address with the specified Prefix-Length. PL-Lower-Bound and PL-Upper-Bound MUST be sent as zero and ignored on receipt.

M-Type = 1: Describes a set of IPv4 address prefixes that correspond to the IPv4 Address/Prefix-Length combination and a prefix length greater than or equal to PL-Lower-Bound. PL-Upper-Bound MUST be sent as zero and ignored on receipt.

M-Type = 2: Describes a set of IPv4 address prefixes that correspond to the IPv4 Address/Prefix-Length combination and a prefix length less than or equal to PL-Upper-Bound. PL-Lower-Bound MUST be sent as zero and ignored on receipt.

M-Type = 3: Describes a set of IPv4 address prefixes that correspond to the IPv4 Address/Prefix-Length combination and a prefix length greater than or equal to PL-Lower-Bound and less than or equal to PL-Upper-Bound.

For example, tuple <M-Type=0, IPv4 Address = 10.1.0.0, Prefix-Length = 16, PL-Lower-Bound = 0, PL-Upper-Bound = 0> represents 10.1.0.0/16.

<M-Type=1, IPv4 Address = 10.1.1.0, Prefix-Length = 24, PL-Lower-Bound = 28, PL-Upper-Bound = 0> represents the set of IPv4 address prefixes that correspond to 10.1.1.0/24 with a prefix length greater than, or equal to, 28 bits (up to and including 32 bits). That is that it represents any IPv4 address prefix that matches 10.1.1.0/24 and 28 <= whose prefix length <= 32.

<M-Type=2, IPv4 Address = 10.1.1.0, Prefix-Length = 24, PL-Lower-Bound = 0, PL-Upper-Bound = 26> represents the set of IPv4 address prefixes that correspond to 10.1.1.0/24 with a prefix length less than, or equal to, 26 bits (up to and including 24 bits). That is that it represents any IPv4 address prefix that matches 10.1.1.0/24 and 24 <= whose prefix length <= 26.

<M-Type=3, IPv4 Address = 10.1.1.0, Prefix-Length = 24, PL-Lower-Bound = 26, PL-Upper-Bound = 30> represents the set of IPv4 address prefixes that correspond to 10.1.1.0/24 with a prefix length greater than, or equal to, 26 bits, and less than, or equal to, 30 bits. That is that it represents any IPv4 address prefix that matches 10.1.1.0/24 and 26 <= whose prefix length <= 30.

4.2.1.2. IPv6 Address Prefix Range List sub-TLV

Similarly, an IPv6 Address Prefix Range List sub-TLV contains a list of IPv6 address prefix ranges. Each range describes an IPv6 address prefix or group of IPv6 address prefixes and is represented by a

tuple <M-Type, IPv6 Address, Prefix Length, PL-Lower-Bound, PL-Upper-Bound>. Its format is illustrated below:

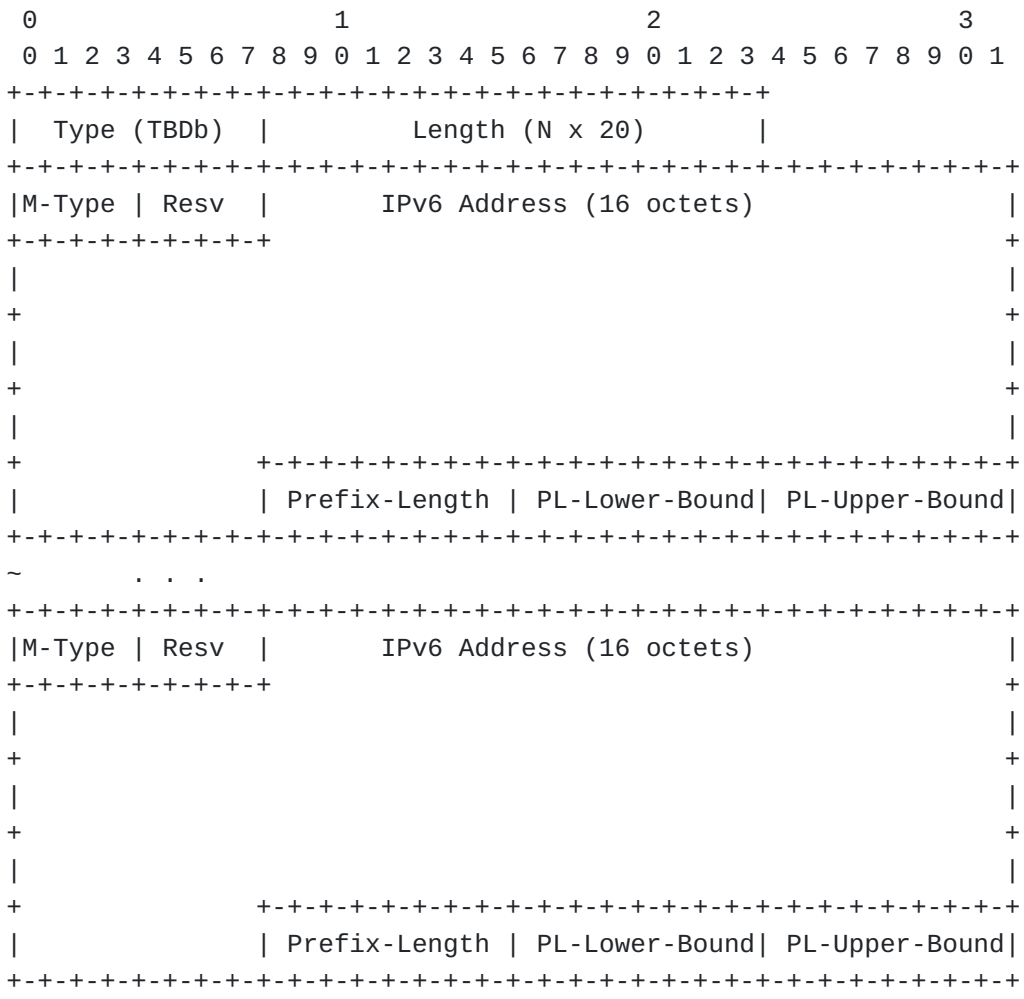

Figure 7: Format of IPv6 Address Prefix Range List sub-TLV

Type: The Type for IPv6 Address Prefix Range List is TBDb.

Length: N x 20, where N is the number of IPv6 address prefix ranges in the sub-TLV. If Length is not a multiple of 20, the enclosing UPDATE message MUST be ignored.

The other fields are similar to those described in Section 4.2.1.1.

For example, tuple <M-Type=0, IPv6 Address = 2001:db8:0:0:0:0:0:0, Prefix-Length = 32, PL-Lower-Bound = 0, PL-Upper-Bound = 0> represents 2001:db8:0:0:0:0:0:0/32.

<M-Type=1, IPv6 Address = 2001:db8:0:0:0:0:0:0, Prefix-Length = 32, PL-Lower-Bound = 32, PL-Upper-Bound = 0> represents the set of IPv6 address prefixes that correspond to 2001:db8:0:0:0:0:0:0/32 with a

prefix length greater than, or equal to, 32 bits (up to and including 128 bits).

<M-Type=2, IPv6 Address = 2001:db8:0:0:0:0:0:0, Prefix-Length = 32, PL-Lower-Bound = 0, PL-Upper-Bound = 64> represents the set of IPv6 address prefixes that correspond to 2001:db8:0:0:0:0:0:0/32 with a prefix length less than, or equal to, 64 bits (up to and including 32 bits).

<M-Type=3, IPv6 Address = 2001:db8:0:0:0:0:0:0, Prefix-Length = 32, PL-Lower-Bound = 48, PL-Upper-Bound = 64> represents the set of IPv6 address prefixes that correspond to 2001:db8:0:0:0:0:0:0/32 with a prefix length greater than, or equal to, 48 bits, and less than, or equal to, 64 bits.

4.2.1.3. AS_PATH RegEx sub-TLV

An AS_PATH RegEx sub-TLV represents any AS_PATH specified by a regular expression [RegExIEEE]. Its format is illustrated below:

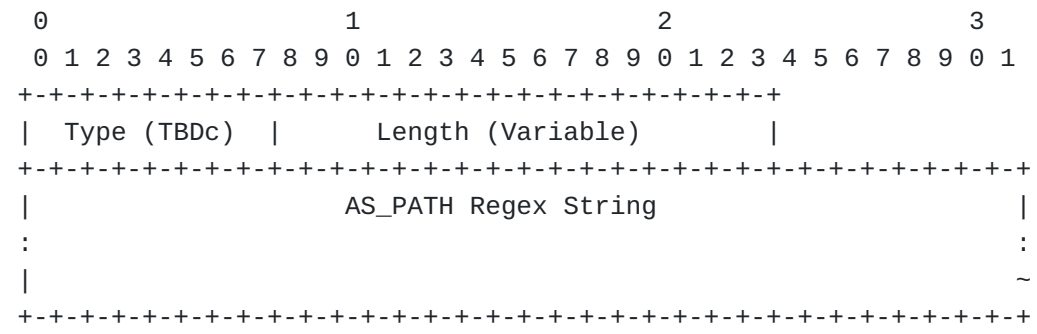

Figure 8: Format of AS_PATH RegEx sub-TLV

Type: The Type for AS_PATH RegEx is TBDc.

Length: Variable, maximum is 1024.

AS_PATH Regex String: It is a regular expression as defined in [RegExIEEE].

For example, regular expression "12345$" represents any AS_PATH that end with 12345.

4.2.1.4. Community List sub-TLV

A Community List sub-TLV represents a list of communities in the BGP COMMUNITIES defined by [RFC1997]. Its format is illustrated below:

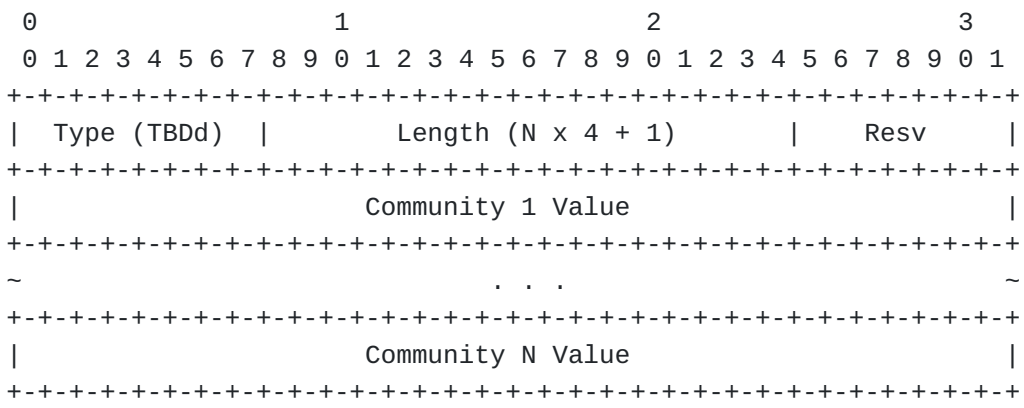

Figure 9: Format of Community List sub-TLV

Type: The Type for Community List is TBDd.

Length: N x 4 + 1, where N is the number of communities. If Length is not a multiple of 4 plus 1, the Atom is corrupt and the enclosing UPDATE message MUST be ignored.

Resv: 1 octet. These bits MUST be sent as zero and ignored on receipt.

Community Value: The Community List contains a list of Community Values. Each Community Value is a 4-octet field for a community defined by [RFC1997].

4.2.2. Atom Type TBD2, The MED Change

A MULTI_EXIT_DISC (MED) Change Atom indicates an action to change the MED. Its format is illustrated as a TLV (Type Length Value) below. The Value field consists of an OP field of 1 octet and an Argument field of 4 octets.

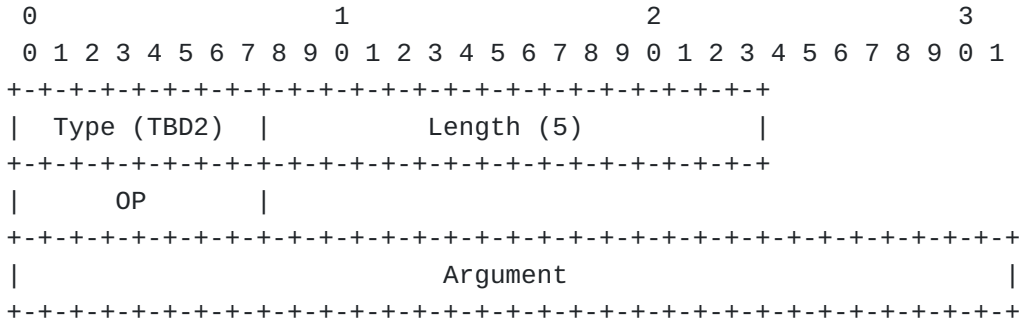

Figure 10: Format of MED Change Atom

Type:
The Type for MED Change Atom is TBD2.

Length: 5. If Length is any other value, the Atom is corrupt and the enclosing UPDATE message MUST be ignored.

Argument: 4 octet unsigned integer.

OP: 1 octet. Three values are defined:

OP = 0: assign the Value of the Argument to the existing MED. If the MED attribute does not exist for an IP route, add a MED attribute with the value.

OP = 1: add the Value of the Argument to the existing MED. If the sum is greater than the maximum allowed value, use that maximum value instead. If the MED attribute does not exist for an IP route, the action specified by the Atom to the route is not taken.

OP = 2: subtract the Value of the Argument from the existing MED. If the result is less than 0, use 0 instead. If the MED attribute does not exist for an IP route, the action specified by the Atom to the route is not taken.

If OP is any other value, the Atom is corrupt and the enclosing UPDATE message MUST be ignored.

4.2.3. Atom Type TBD3, The AS_PATH Change

An AS_PATH Change Atom indicates an action to change the AS_PATH. Its format is illustrated below:

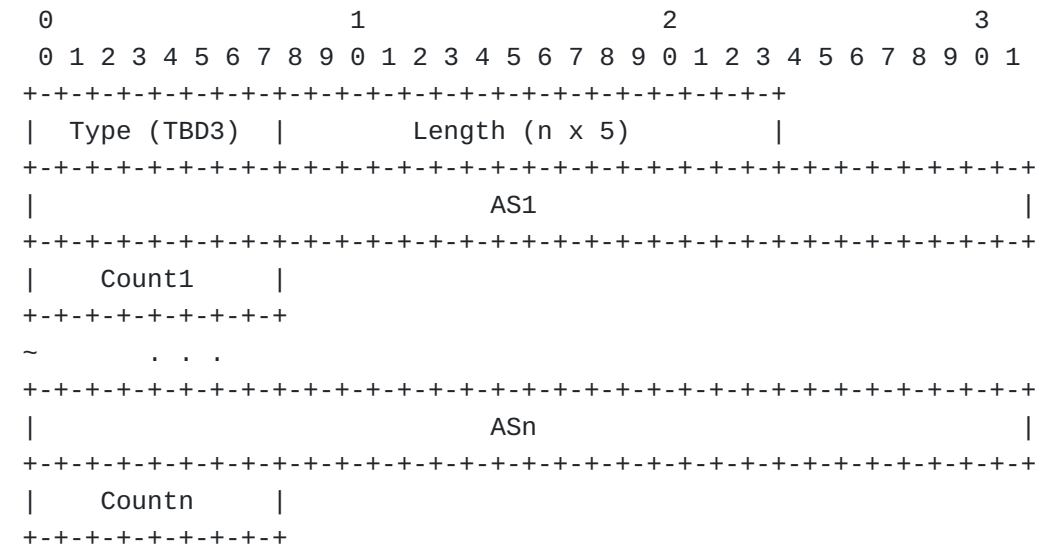

Figure 11: Format of AS_PATH Change Atom

Type:
The Type for AS_PATH Change Atom is TBD3.

Length: n x 5. If Length is not a multiple of 5, the Atom is corrupt and the enclosing UPDATE message MUST be ignored.

AS and Count: The Atom contains a list of AS and Count pairs. Each AS and Count pair has an AS field of 4 octets for an AS number and a Count field of 1 octet for an unsigned integer indicating the number of times the AS number repeats.

The sequence of AS numbers specified by the Atom is added to the existing AS_PATH. The AS numbers SHOULD be local AS numbers.

4.3. Community Value in BGP Wide Community

[I-D.ietf-idr-wide-bgp-communities] defines the Type 1 BGP Community Container, the BGP Wide Community. It contains a Community Value of 4 octets indicating what set of actions a router is requested to take upon reception of an IP route matching the conditions in this community. This section specifies two Community Values:

*MATCH AND SET ATTR (TBDx)

*MATCH AND NOT ADVERTISE (TBDy)

4.3.1. MATCH AND SET ATTR (TBDx)

For the BGP Wide Community with Community Value MATCH AND SET ATTR (TBDx), its Targets TLV MUST contain a RouteAttr Atom, its Parameters TLV MUST include a MED Change Atom and/or a AS_PATH Change Atom. The RouteAttr Atom MUST contain an IPv4/IPv6 (IP for short) Address Prefix Range List and may contain a Community List and/or AS_PATH sub-TLVs. The Prefix Range List states a set of IP address prefix ranges. The Community List and/or AS_PATH identify a set of path attributes.

After a BGP speaker receives the BGP Wide Community in a BGP UPDATE message for it, the speaker extracts the routing policy from the BGP Wide Community. For any IP route to a peer of the speaker, if the IP address prefix of the route is in any prefix range stated by the Prefix Range List and the route has the attributes identified by the Community List and/or AS_PATH, then the attributes of the IP route are modified per the actions specified by the MED Change and/or AS_PATH Change Atom before sending it to the peer.

4.3.2. MATCH AND NOT ADVERTISE (TBDy)

For the BGP Wide Community with Community Value MATCH AND NOT ADVERTISE (TBDy), its Targets TLV MUST contain a RouteAttr Atom. The

Atom has the same contents and semantic as the one described in Section 4.3.1.

After a BGP speaker receives the BGP Wide Community in a BGP UPDATE message for it, the speaker extracts the routing policy from the BGP Wide Community. For any IP route to a peer of the speaker, if the IP address prefix of the route is in any prefix range stated by the Prefix Range List and the route has the attributes identified by the Community List and/or AS_PATH, then the IP route will not be advertised to the peer.

5. Operational Considerations

To adjust the traffic flowing to an AS with a controller, an operator needs to create a BGP RPD session between the controller and a RR in the AS. This session SHOULD be independent of routing information. The controller can distribute a RPD routing policy to any BGP speaker in the AS using this session. The speaker applies the policy to the IP routes to be sent to its peers as specified.

For the session between the controller and the RR, some existing mechanisms such as BGP Graceful Restart (GR) [RFC4724] and BGP Long-lived Graceful Restart (LLGR) SHOULD be used to let the RR keep the RPD routing policies from the controller for some time. With support of "Long-lived Graceful Restart Capability" [I-D.ietf-idr-long-lived-gr], the RPD routing policies can be retained for a longer time after the controller fails. When the controller recovers from its failure within the graceful period, the RR still have the RPD routing policies from the controller before the failure.

For the sessions between the speaker and its peers, the mechanisms mentioned above are not necessary. When the speaker goes down, the traffic to the AS through the speaker from its peers needs take another path without going through the speaker. The peers withdraw the routes from the speaker and adjust (reroute) the traffic to use another path without the speaker. This is expected.

For the traffic to an IP address prefix in the AS from an neighbor AS, the operator needs make sure that the traffic can be adjusted through changing the MED and/or AS_PATH attribute in the IP route with the prefix to be sent to the neighbor AS.

In a BGP speaker, there are routing policies from different sources, including RPD and others such as configuration and PCE. The speaker applies all the policies as needed. It applies the RPD routing policies after applying the other routing policies. In order to adjust traffic using RPD routing policies with MED change and/or

AS_PATH change, the operator needs make sure that the RPD policies are not superseded by any policy from other sources.

When a RPD routing policy is to be applied by a BGP speaker to only one of its peers, the Peer field SHOULD be the IP address of this peer. After receiving the RPD routing policy, the BGP speaker applies the policy to the IP routes to be sent to this peer.

When a RPD routing policy is to be applied by a BGP speaker to all its peers in some of its neighbor ASs, the Autonomous System Number (ANS) List Atom can be used in the Targets TLV to select these neighbor ASs while the Peer field is 0. After receiving the RPD routing policy, the BGP speaker applies the policy to the IP routes to be sent to the peers in these selected neighbor ASs.

When a RPD routing policy is to be applied by a BGP speaker to some of its peers, the IP Prefix List Atom can be used in the Targets TLV to select these peers while the Peer field is 0. After receiving the RPD routing policy, the BGP speaker applies the policy to the IP routes to be sent to these selected peers.

There are already lots of existing policies configured on the routers in an operational network. There are different types of policies, which include security, management and control policies. These policies are relatively stable. However, the policies for adjusting traffic are dynamic. Whenever the traffic through a path is not expected, the policies to adjust the traffic for that path are configured on the related routers. Some users would like to separate the stable policies from the dynamic ones even though they have configuration automation systems (including YANG models). In this case, RPD with a controller (RPD for short) should be considered over others. Using RPD, the stable policies and dynamic ones are separated from users' view.

When the number of routers to be configured for adjusting traffic is big and keeping all the connections live between a configuration automation system and these routers affects network performance, RPD should be considered over this system. Using RPD, there is one connection between the controller and a RR in an AS. There is almost no impact on the network performance.

When it takes a long time for a configuration automation system to adjust traffic, RPD should be considered over this system. Using RPD, the policies for adjusting traffic are distributed to the related routers and applied in routing speed.

6. IANA Considerations

6.1. Existing Assignments

IANA has assigned the Routing Policy SAFI of value 75 from the registry "Subsequent Address Family Identifiers (SAFI) Parameters".

6.2. BGP Wide Community Community Types

IANA is requested to assign from the registry "Registered Type 1 BGP Wide Community Community Types" the following values:

Community Type Value	Description	Reference
TBDx (0x80000018 suggested)	MATCH AND SET ATTR	This document
TBDy (0x80000019 suggested)	MATCH AND NOT ADVERTISE	This document

6.3. BGP Community Container Atom Types

IANA is requested to assign from the registry "BGP Community Container Atom Types" as follows:

Type Value	Name	Reference
TBD1 (0x09 suggested)	RouteAttr	This document
TBD2 (0x0A suggested)	MED Change	This document
TBD3 (0x0B suggested)	AS_PATH Change	This document
TBDa (0x0C suggested)	IPv4 Prefix Range List	This document
TBDb (0x0D suggested)	IPv6 Prefix Range List	This document
TBDc (0x0E suggested)	AS-Path RegEx	This document
TBDd (0x0F suggested)	Community List	This document

7. Security Considerations

All the security considerations for base BGP [RFC4271][RFC4272] and BGP Wide Community [I-D.ietf-idr-wide-bgp-communities] apply to the BGP extensions defined in this document.

This document depends on the BGP Multiprotocol extension [RFC4760], which states that the extension does not change the underlying security issues inherent in the existing BGP. It does not fundamentally change the security behavior of BGP deployments. It may be observed that the RPD is used only within a well-defined scope, for example, within a single AS or a set of ASes that are administrated by a single service provider.

This document defines two community values in the BGP Wide Community to distribute and apply routing policies. One is MATCH AND SET ATTR (TBDx) and the other is MATCH AND NOT ADVERTISE (TBDy). Using the former changes one or more best IP routes distributed by BGP and redirects a certain traffic flows in a network. Using the latter drops one or more IP routes distributed by BGP and redirects some traffic flows in a network. The potential effects of the distribution and use of a undesired routing policy from a (rogue) router include causing network congestions and reducing the quality of the services. They can also have the effect of dropping traffic. Note that a rogue node can use these to attack the network, but a misconfigured policy could have the same effect. It is necessary to prevent a (rogue) router from advertising an incorrect or undesired routing policy through BGP sessions. The risk can be mitigated by using the techniques such as those discussed in [RFC5925] to help authenticate BGP sessions.

Note that a typical RPD deployment requires a BGP session between a controller and a route reflector in a network administrated by a single service provider. The controller distributes RPD routing policies to some routers in the network through this BGP session. There is concern that a rogue controller might be introduced into the network. The rogue controller may inject false RPD routing policies or take over and change existing RPD routing policies. This corresponds to a rogue BGP speaker entering the network, or a route reflector being subverted. It is strongly recommended that the techniques such as those in [RFC5925] be used to secure this BGP session, the route reflector be configured with the identity of the controller, and software loads on the controller be protected.

8. References

8.1. Normative References

[I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S., and P. Jakma, "BGP Community Container Attribute", Work in Progress, Internet-Draft, draft-ietf-idr-wide-bgp-communities-08, 11 July 2022, <https://www.ietf.org/archive/id/draft-ietf-idr-wide-bgp-communities-08.txt>.

[RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <https://www.rfc-editor.org/info/rfc1997>.

[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>.

[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>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/rfc/rfc4760>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

8.2. Informative References

[I-D.ietf-idr-long-lived-gr] Uttaro, J., Chen, E., Decraene, B., and J. Scudder, "Support for Long-lived BGP Graceful Restart", Work in Progress, Internet-Draft, draft-ietf-idr-long-lived-gr-03, 6 December 2022, <https://www.ietf.org/archive/id/draft-ietf-idr-long-lived-gr-03.txt>.

[RegExIEEE] The Open Group., "IEEE Std 1003.1-2017 (Revision of IEEE Std 1003.1-2008)", 2018.

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

DOI 10.17487/RFC4724, January 2007, <https://www.rfc-editor.org/info/rfc4724>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <https://www.rfc-editor.org/info/rfc5925>.

[RFC9012] Patel, K., Van de Velde, G., Sangli, S., and J. Scudder, "The BGP Tunnel Encapsulation Attribute", RFC 9012, DOI 10.17487/RFC9012, April 2021, <https://www.rfc-editor.org/info/rfc9012>.

Acknowledgments

The authors would like to thank Acee Lindem, Jeff Haas, Jie Dong, Lucy Yong, Qiandeng Liang, Zhenqiang Li, Robert Raszuk, Donald Eastlake, Ketan Talaulikar, and Jakob Heitz for their comments to this work.

Contributors

The following people have substantially contributed to the definition of the BGP RPD and to the editing of this document:

Sujian Lu
Tencent
Email: jasonlu@tencent.com

Shunwan Zhuang
Huawei
Email: zhuangshunwan@huawei.com

Peng Zhou
Huawei
Email: Jewpon.zhou@huawei.com

Authors' Addresses

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

Email: lizhenbin@huawei.com

Liang Ou

China Telcom Co., Ltd.
109 West Zhongshan Ave,Tianhe District
Guangzhou
510630
China

Email: ouliang@chinatelecom.cn

Yujia Luo
China Telcom Co., Ltd.
109 West Zhongshan Ave,Tianhe District
Guangzhou
510630
China

Email: luoyuj@sdu.edu.cn

Gyan S. Mishra
Verizon Inc.
13101 Columbia Pike
Silver Spring, MD 20904
United States of America

Phone: 301 502-1347
Email: gyan.s.mishra@verizon.com

Huaimo Chen
Futurewei
Boston, MA,
United States of America

Email: Huaimo.chen@futurewei.com

Haibo Wang
Huawei
Huawei Bld., No.156 Beiqing Rd.
Beijing
100095
China

Email: rainsword.wang@huawei.com