Network Working Group
Internet Draft
Expiration Date: September 2011

Tony Bates (Skype)
Ravi Chandra (Cisco Systems)
Dave Katz (Juniper Networks)
Yakov Rekhter (Juniper Networks)

March 28, 2011

Multiprotocol Extensions for BGP-4

draft-ietf-idr-rfc4760bis-01.txt

Status of this Memo

Copyright Notice

Abstract

This document defines extensions to BGP-4 to enable it to carry routing information for multiple Network Layer protocols (e.g., IPv6, IPX, L3VPN, etc...). The extensions are backward compatible - a router that supports the extensions can interoperate with a router that doesn't support the extensions.

1. Specification of Requirements

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

2. Overview

The only three pieces of information carried by BGP-4 [BGP-4] that are IPv4 specific are (a) the NEXT_HOP attribute (expressed as an IPv4 address), (b) AGGREGATOR (contains an IPv4 address), and (c) NLRI (expressed as IPv4 address prefixes). This document assumes that any BGP speaker (including the one that supports multiprotocol capabilities defined in this document) has to have an IPv4 address (which will be used, among other things, in the AGGREGATOR attribute). Therefore, to enable BGP-4 to support routing for multiple Network Layer protocols, the only two things that have to be added to BGP-4 are (a) the ability to associate a particular Network Layer protocol with the next hop information, and (b) the ability to associate a particular Network Layer protocol with NLRI. To identify individual Network Layer protocols associated with the next hop information and semantics of NLRI, this document uses a combination of Address Family, as defined in [IANA-AF], and Subsequent Address Family (as described in this document).

One could further observe that the next hop information (the information provided by the NEXT_HOP attribute) is meaningful (and necessary) only in conjunction with the advertisements of reachable destinations - in conjunction with the advertisements of unreachable destinations (withdrawing routes from service), the next hop information is meaningless. This suggests that the advertisement of reachable destinations should be grouped with the advertisement of the next hop to be used for these destinations, and that the advertisement of reachable destinations should be segregated from the advertisement of unreachable destinations.

To provide backward compatibility, as well as to simplify introduction of the multiprotocol capabilities into BGP-4, this

document uses two new attributes, Multiprotocol Reachable NLRI (MP_REACH_NLRI) and Multiprotocol Unreachable NLRI (MP_UNREACH_NLRI). The first one (MP_REACH_NLRI) is used to carry the set of reachable destinations together with the next hop information to be used for forwarding to these destinations. The second one (MP_UNREACH_NLRI) is used to carry the set of unreachable destinations. Both of these attributes are optional and non-transitive. This way, a BGP speaker that doesn't support the multiprotocol capabilities will just ignore the information carried in these attributes and will not pass it to other BGP speakers.

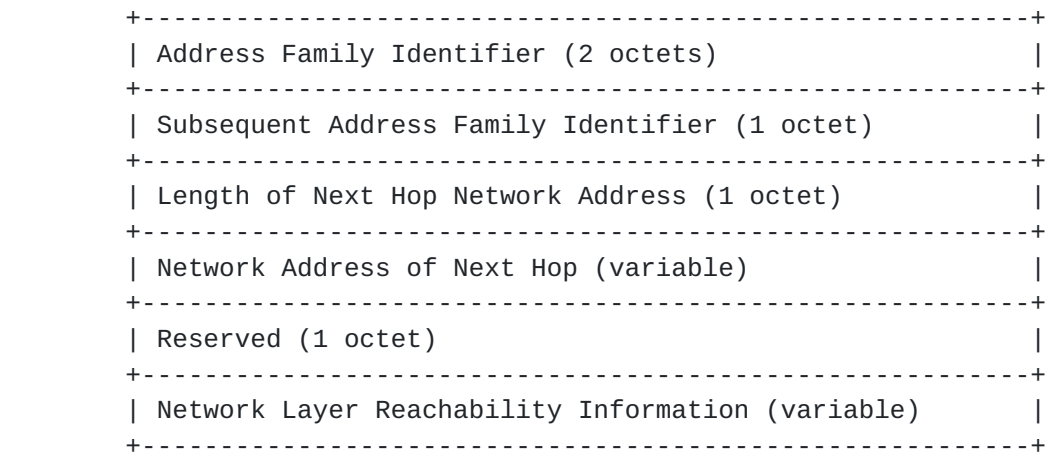
3. Multiprotocol Reachable NLRI - MP_REACH_NLRI (Type Code 14):

This is an optional non-transitive attribute that can be used for the following purposes:

(a) to advertise a feasible route to a peer

(b) to permit a router to advertise the Network Layer address of the router that should be used as the next hop to the destinations listed in the Network Layer Reachability Information field of the MP_NLRI attribute.

The attribute is encoded as shown below:



The use and meaning of these fields are as follows:

Address Family Identifier (AFI):

This field in combination with the Subsequent Address Family

Identifier field identifies the set of Network Layer protocols to which the address carried in the Next Hop field must belong, the way in which the address of the next hop is encoded, and the semantics of the Network Layer Reachability Information that follows. If the Next Hop is allowed to be from more than one Network Layer protocol, the encoding of the Next Hop MUST provide a way to determine its Network Layer protocol.

Presently defined values for the Address Family Identifier field are specified in the IANA's Address Family Numbers registry [IANA-AF].

Subsequent Address Family Identifier (SAFI):

This field in combination with the Address Family Identifier field identifies the set of Network Layer protocols to which the address carried in the Next Hop must belong, the way in which the address of the next hop is encoded, and the semantics of the Network Layer Reachability Information that follows. If the Next Hop is allowed to be from more than one Network Layer protocol, the encoding of the Next Hop MUST provide a way to determine its Network Layer protocol.

Length of Next Hop Network Address:

A 1-octet field whose value expresses the length of the "Network Address of Next Hop" field, measured in octets.

Network Address of Next Hop:

A variable-length field that contains the Network Address of the next router on the path to the destination system. The Network Layer protocol associated with the Network Address of the Next Hop is identified by a combination of <AFI, SAFI> carried in the attribute.

Reserved:

A 1 octet field that MUST be set to 0, and SHOULD be ignored upon receipt.

Network Layer Reachability Information (NLRI):

A variable length field that lists NLRI for the feasible routes that are being advertised in this attribute. The semantics of NLRI is identified by a combination of <AFI, SAFI> carried in the attribute.

When the Subsequent Address Family Identifier field is set to one of the values defined in this document, each NLRI is encoded as specified in the "NLRI encoding" section of this document.

The next hop information carried in the MP_REACH_NLRI path attribute defines the Network Layer address of the router that SHOULD be used as the next hop to the destinations listed in the MP_NLRI attribute in the UPDATE message.

The rules for the next hop information are the same as the rules for the information carried in the NEXT_HOP BGP attribute (see Section 5.1.3 of [BGP-4]).

An UPDATE message that carries the MP_REACH_NLRI MUST also carry the ORIGIN and the AS_PATH attributes (both in EBGP and in IBGP exchanges). Moreover, in IBGP exchanges such a message MUST also carry the LOCAL_PREF attribute.

An UPDATE message that carries no NLRI, other than the one encoded in the MP_REACH_NLRI attribute, SHOULD NOT carry the NEXT_HOP attribute. If such a message contains the NEXT_HOP attribute, the BGP speaker that receives the message SHOULD ignore this attribute.

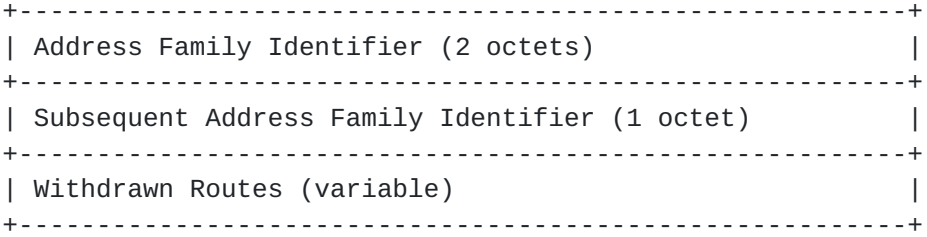
An UPDATE message SHOULD NOT include the same address prefix (of the same <AFI, SAFI>) in more than one of the following fields: WITHDRAWN ROUTES field, Network Reachability Information fields, MP_REACH_NLRI field, and MP_UNREACH_NLRI field. The processing of an UPDATE message in this form is undefined.

This document RECOMMENDS that the MP_REACH_NLRI attribute be placed as the very first path attribute in an UPDATE message.

4. Multiprotocol Unreachable NLRI - MP_UNREACH_NLRI (Type Code 15):

This is an optional non-transitive attribute that can be used for the purpose of withdrawing multiple unfeasible routes from service.

The attribute is encoded as shown below:



The use and the meaning of these fields are as follows:

Address Family Identifier (AFI):

This field in combination with the Subsequent Address Family Identifier field identifies the set of Network Layer protocols to which the address carried in the Next Hop field must belong, the way in which the address of the next hop is encoded, and the semantics of the Network Layer Reachability Information that follows. If the Next Hop is allowed to be from more than one Network Layer protocol, the encoding of the Next Hop MUST provide a way to determine its Network Layer protocol.

Presently defined values for the Address Family Identifier field are specified in the IANA's Address Family Numbers registry [IANA-AF].

Subsequent Address Family Identifier (SAFI):

This field in combination with the Address Family Identifier field identifies the set of Network Layer protocols to which the address carried in the Next Hop must belong, the way in which the address of the next hop is encoded, and the semantics of the Network Layer Reachability Information that follows. If the Next Hop is allowed to be from more than one Network Layer protocol, the encoding of the Next Hop MUST provide a way to determine its Network Layer protocol.

Withdrawn Routes Network Layer Reachability Information:

A variable-length field that lists NLRI for the routes that are being withdrawn from service. The semantics of NLRI is identified by a combination of <AFI, SAFI> carried in the attribute.

When the Subsequent Address Family Identifier field is set to one of the values defined in this document, each NLRI is encoded as specified in the "NLRI encoding" section of this document.

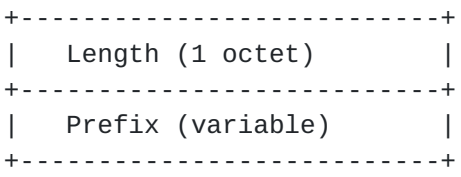
An UPDATE message that contains the MP_UNREACH_NLRI is not required to carry any other path attributes.

This document RECOMMENDS that the MP_UNREACH_NLRI attribute be placed as the very first path attribute in an UPDATE message, unless the UPDATE message also carries the MP_REACH_NLRI attribute, in which case it is RECOMMENDED to place the MP_UNREACH_NLRI right after the

MP_REACH_NLRI attribute.

5. NLRI encoding

The optional Network Layer Reachability information is encoded as one or more 2-tuples of the form <length, prefix>, whose fields are described below:



The use and the meaning of these fields are as follows:

a) Length:

The Length field indicates the length, in bits, of the address prefix. A length of zero indicates a prefix that matches all (as specified by the address family) addresses (with prefix, itself, of zero octets).

b) Prefix:

The Prefix field contains an address prefix followed by enough trailing bits to make the end of the field fall on an octet boundary. Note that the value of trailing bits is irrelevant.

6. Subsequent Address Family Identifier

This document defines the following values for the Subsequent Address Family Identifier field carried in the MP_REACH_NLRI and MP_UNREACH_NLRI attributes:

1 - Network Layer Reachability Information used for unicast forwarding

2 - Network Layer Reachability Information used for multicast forwarding

An implementation MAY support all, some, or none of the Subsequent

Address Family Identifier values defined in this document.

7. Error Handling

If a BGP speaker receives from a neighbor an UPDATE message that contains the MP_REACH_NLRI or MP_UNREACH_NLRI attribute, and if the speaker determines that the attribute is incorrect, the speaker MUST delete all the BGP routes received from that neighbor whose AFI/SAFI is the same as the one carried in the incorrect MP_REACH_NLRI or MP_UNREACH_NLRI attribute. For the duration of the BGP session over which the UPDATE message was received, the speaker then SHOULD ignore all the subsequent routes with that AFI/SAFI received over that session.

In addition, the speaker MAY terminate the BGP session over which the UPDATE message was received. The session SHOULD be terminated with the Notification message code/subcode indicating "UPDATE Message Error"/"Optional Attribute Error".

If a BGP speaker receives from a neighbor an UPDATE message that contains a valid MP_REACH_NLRI or MP_UNREACH_NLRI attribute, and if the speaker determines that the attribute has no NLRI, the speaker MUST NOT treat this UPDATE message as a BGP error, and specifically MUST NOT terminate the BGP session over which the UPDATE was received, and MUST NOT ignore all the subsequent routes received over that session with the AFI/SAFI carried in the attribute. This is irrespective of whether the received message contains any non-empty Withdrawn Routes, and/or non-empty Network Layer Reachability Information fields.

8. Redistribution from one <AFI, SAFI> to another

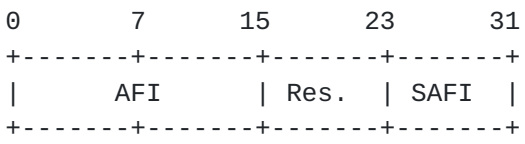
A router SHALL NOT redistribute routing information received over one particular combination of <AFI, SAFI> into another <AFI, SAFI> unless explicitly configured. The implications of doing such redistribution are numerous and serious, but outside the scope of this document.

If a router is explicitly configured to redistribute routing information received over one particular combination of <AFI, SAFI> over another <AFI, SAFI>, then when redistributing the information the router MUST set NEXT_HOP to self.

9. Use of BGP Capability Advertisement

A BGP speaker that uses Multiprotocol Extensions SHOULD use the Capability Advertisement procedures [BGP-CAP] to determine whether the speaker could use Multiprotocol Extensions with a particular peer.

The fields in the Capabilities Optional Parameter are set as follows. The Capability Code field is set to 1 (which indicates Multiprotocol Extensions capabilities). The Capability Length field is set to 4. The Capability Value field is defined as:



The use and meaning of this field is as follow:

AFI - Address Family Identifier (16 bit), encoded the same way as in the Multiprotocol Extensions

Res. - Reserved (8 bit) field. MUST be set to 0 by the sender and MUST be ignored by the receiver.

SAFI - Subsequent Address Family Identifier (8 bit), encoded the same way as in the Multiprotocol Extensions.

A speaker that supports multiple <AFI, SAFI> tuples includes them as multiple Capabilities in the Capabilities Optional Parameter.

To have a bi-directional exchange of routing information for a particular <AFI, SAFI> between a pair of BGP speakers, each such speaker MUST advertise to the other (via the Capability Advertisement mechanism) the capability to support that particular <AFI, SAFI> routes.

10. IANA Considerations

As specified in this document, the MP_REACH_NLRI and MP_UNREACH_NLRI attributes contain the Subsequence Address Family Identifier (SAFI) field. The SAFI name space is defined in this document. The IANA registered and maintains values for the SAFI namespace as follows:

- SAFI values 1 and 2 are assigned in this document.

- SAFI value 3 is reserved. It was assigned by RFC 2858 for a use that was never fully implemented, so it is deprecated by this document.

- SAFI values 5 through 63 are to be assigned by IANA using either the Standards Action process, defined in [RFC2434], or the Early IANA Allocation process, defined in [RFC4020].

- SAFI values 67 through 127 are to be assigned by IANA, using the "First Come First Served" policy, defined in RFC 2434.

- SAFI values 0 and 255 are reserved.

- SAFI values 128 through 240 are part of the previous "private use" range. At the time of approval of this document, the unused values were provided to IANA by the Routing Area Director. These unused values, namely, 130, 131, 135 through 139, and 141 through 240, are considered reserved in order to avoid conflicts.

- SAFI values 241 through 254 are for "private use", and values in this range are not to be assigned by IANA.

11. Comparison with RFC4760

This document explicitly spells out that receiving an UPDATE message that carried MP_REACH_NLRI or MP_UNREACH_NLRI attribute, with the attribute carrying no NLRI, must not be treated as an error.

This document also recommends that that the MP_REACH_NLRI and MP_UNREACH_NLRI attributes be placed as the very first path attributes in an UPDATE in this case.

12. Comparison with RFC2858

This document makes the use of the next hop information consistent with the information carried in the NEXT_HOP BGP path attribute.

This document removes the definition of SAFI 3, and deprecates SAFI 3.

This document changes partitioning of the SAFI space. Specifically, in RFC 2858 SAFI values 128 through 240 were part of the "private use" range. This document specifies that of this range, allocations that are currently in use are to be recognized by IANA, and that unused values, namely 130, 131, 135 through 139, and 141 through 240, should be considered reserved.

This document renames the Number of SNPAs field to Reserved, and removes the rest of the SNPA-related information from the MP_REACH_NLRI attribute.

13. Comparison with RFC 2283

This document restricts the MP_REACH_NLRI attribute to carry only a single instance of <AFI, SAFI, Next Hop Information, ...>.

This document restricts the MP_UNREACH_NLRI attribute to carry only a single instance of <AFI, SAFI, ...>.

This document clarifies handling of an UPDATE message that carries no NLRI, other than the one encoded in the MP_REACH_NLRI attribute.

This document clarifies error handling in the presence of MP_REACH_NLRI or MP_UNREACH_NLRI attributes.

This document specifies the use of BGP Capabilities Advertisements in conjunction with multi-protocol extensions.

Finally, this document includes the "IANA Consideration" section.

14. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing BGP.

15. Acknowledgements

The authors would like to thank members of the IDR Working Group for their review and comments. We also acknowledge comments from Keyur Patel.

16. Normative References

[BGP-CAP] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 3392, November 2002.

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

[IANA-AF] "Address Family Numbers", Reachable from http://www.iana.org/numbers.html

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

[RFC2434] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 2434, October 1998.

[RFC4020] Kompella, K. and A. Zinin, "Early IANA Allocation of Standards Track Code Points", BCP 100, RFC 4020, February 2005.

17. Authors' Addresses

Tony Bates
Skype

Ravi Chandra
Cisco Systems
EMail: rchandra@cisco.com

Dave Katz
Juniper Networks, Inc.
EMail: dkatz@juniper.net

Yakov Rekhter
Juniper Networks, Inc.
EMail: yakov@juniper.net