Network Working Group
Internet Draft
Expiration Date: August 2000

Tony Bates
Cisco Systems
Ravi Chandra
Siara Systems
Dave Katz
Juniper Networks
Yakov Rekhter
Cisco Systems

Multiprotocol Extensions for BGP-4

draft-ietf-idr-bgp4-multiprotocol-v2-04.txt

1. Status of this Memo

2. Abstract

Currently BGP-4 [BGP-4] is capable of carrying routing information only for IPv4 [IPv4]. This document defines extensions to BGP-4 to enable it to carry routing information for multiple Network Layer protocols (e.g., IPv6, IPX, etc...). The extensions are backward compatible - a router that supports the extensions can interoperate with a router that doesn't support the extensions.

3. Overview

The only three pieces of information carried by 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 associated a particular Network Layer protocol with NLRI. To identify individual Network Layer protocols this document uses Address Family, as defined in [RFC1700].

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.

4. 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.

(c) to allow a given router to report some or all of the Subnetwork Points of Attachment (SNPAs) that exist within the local system

The attribute is encoded as shown below:

```
      +---------------------------------------------------------+
      | Address Family Identifier (2 octets)                    |
      +---------------------------------------------------------+
      | Subsequent Address Family Identifier (1 octet)          |
      +---------------------------------------------------------+
      | Length of Next Hop Network Address (1 octet)            |
      +---------------------------------------------------------+
      | Network Address of Next Hop (variable)                  |
      +---------------------------------------------------------+
      | Number of SNPAs (1 octet)                               |
      +---------------------------------------------------------+
      | Length of first SNPA(1 octet)                           |
      +---------------------------------------------------------+
      | First SNPA (variable)                                   |
      +---------------------------------------------------------+
      | Length of second SNPA (1 octet)                         |
      +---------------------------------------------------------+
      | Second SNPA (variable)                                  |
      +---------------------------------------------------------+
      | ...                                                     |
      +---------------------------------------------------------+
      | Length of Last SNPA (1 octet)                           |
      +---------------------------------------------------------+
      | Last SNPA (variable)                                    |
      +---------------------------------------------------------+
```

```
      | Network Layer Reachability Information (variable)       |
      +---------------------------------------------------------+
```

The use and meaning of these fields are as follows:

Address Family Identifier:

This field carries the identity of the Network Layer protocol associated with the Network Address that follows. Presently defined values for this field are specified in RFC1700 (see the Address Family Numbers section).

Subsequent Address Family Identifier:

This field provides additional information about the type of the Network Layer Reachability Information carried in the attribute.

Length of Next Hop Network Address:

A 1 octet field whose value expresses the length of the "Network Address of Next Hop" field as 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

Number of SNPAs:

A 1 octet field which contains the number of distinct SNPAs to be listed in the following fields. The value 0 may be used to indicate that no SNPAs are listed in this attribute.

Length of Nth SNPA:

A 1 octet field whose value expresses the length of the "Nth SNPA of Next Hop" field as measured in semi-octets

Nth SNPA of Next Hop:

A variable length field that contains an SNPA of the router whose Network Address is contained in the "Network Address of Next Hop" field. The field length is an integral number of octets in length, namely the rounded-up integer value of one half the SNPA length expressed in semi-octets; if the SNPA

contains an odd number of semi-octets, a value in this field will be padded with a trailing all-zero semi-octet.

Network Layer Reachability Information:

A variable length field that lists NLRI for the feasible routes that are being advertised in this 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 border router that should be used as the next hop to the destinations listed in the MP_NLRI attribute in the UPDATE message. When advertising a MP_REACH_NLRI attribute to an external peer, a router may use one of its own interface addresses in the next hop component of the attribute, provided the external peer to which the route is being advertised shares a common subnet with the next hop address. This is known as a "first party" next hop. A BGP speaker can advertise to an external peer an interface of any internal peer router in the next hop component, provided the external peer to which the route is being advertised shares a common subnet with the next hop address. This is known as a "third party" next hop information. A BGP speaker can advertise any external peer router in the next hop component, provided that the Network Layer address of this border router was learned from an external peer, and the external peer to which the route is being advertised shares a common subnet with the next hop address. This is a second form of "third party" next hop information.

Normally the next hop information is chosen such that the shortest available path will be taken. A BGP speaker must be able to support disabling advertisement of third party next hop information to handle imperfectly bridged media or for reasons of policy.

A BGP speaker must never advertise an address of a peer to that peer as a next hop, for a route that the speaker is originating. A BGP speaker must never install a route with itself as the next hop.

When a BGP speaker advertises the route to an internal peer, the advertising speaker should not modify the next hop information associated with the route. When a BGP speaker receives the route via an internal link, it may forward packets to the next hop address if the address contained in the attribute is on a common subnet with the local and remote BGP speakers.

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. If such a message is received from an external peer, the local system shall check whether the leftmost AS in the AS_PATH attribute is equal to the autonomous system number of the peer than sent the message. If that is not the case, the local system shall send the NOTIFICATION message with Error Code UPDATE Message Error, and the Error Subcode set to Malformed AS_PATH.

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.

5. 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:

```
+---------------------------------------------------------+
| Address Family Identifier (2 octets)                    |
+---------------------------------------------------------+
| Subsequent Address Family Identifier (1 octet)          |
+---------------------------------------------------------+
| Withdrawn Routes (variable)                             |
+---------------------------------------------------------+
```

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

Address Family Identifier:

This field carries the identity of the Network Layer protocol associated with the NLRI that follows. Presently defined values for this field are specified in RFC1700 (see the Address Family Numbers section).

Subsequent Address Family Identifier:

This field provides additional information about the type of the Network Layer Reachability Information carried in the attribute.

Withdrawn Routes:

A variable length field that lists NLRI for the routes that are being withdrawn from service. 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.

6. NLRI encoding

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

```
+---------------------------+
|   Length (1 octet)        |
+---------------------------+
|   Prefix (variable)       |
+---------------------------+
```

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.

7. 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

3 - Network Layer Reachability Information used for both unicast and multicast forwarding

8. 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 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".

9. Use of BGP Capability Negotiation

A BGP speaker that uses Multiprotocol Extensions should use the Capability Negotiation 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:

```
0       7      15      23      31
+-------+-------+-------+-------+
|      AFI      | Res.  | SAFI  |
+-------+-------+-------+-------+
```

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

Res. - Reserved (8 bit) field. Should be set to 0 by the sender and 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 Negotiation mechanism) the capability to support that particular <AFI, SAFI> routes.

10. IANA Considerations

As specified in this document, the MPL_REACH_NLRI and MP_UNREACH_NLRI attributes contain the Subsequence Address Family Identifier (SAFI) field. SAFI value 0 is reserved. SAFI values 1, 2, and 3 are assigned in this document. SAFI values 4 through 63 are to be assigned by IANA using the "IETF Consensus" policy defined in RFC2434. SAFI values 64 through 127 are to be assigned by IANA, using the "First Come First Served" policy defined in RFC2434. SAFI values 128 through 255 are vendor-specific, and values in this range are not to be assigned by IANA.

11. Security Considerations

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

12. Acknowledgements

The authors would like to thank members of the IDR Working Group for their review and comments.

13. References

[BGP-CAP] "Capabilities Negotiation with BGP-4", R. Chandra, J. Scudder, draft-ietf-idr-bgp4-cap-neg-05.txt, February 1999

[BGP-4] "A Border Gateway Protocol 4 (BGP-4)", Y. Rekhter & T. Li, RFC1771, March 1995

[Heffernan] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC2385, August 1998.

[IPv4] "Internet Protocol", J. Postel, September 1981

[RFC1700] "Assigned Numbers", J. Reynolds, J. Postel, RFC1700, October 1994 (see also http://www.iana.org/iana/assignments.html)

14. Author Information

Tony Bates
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
email: tbates@cisco.com

Ravi Chandra
Siara Systems Incorporated
1195 Borregas Avenue
Sunnyvale, CA 94089
e-mail: rchandra@siara.com

Dave Katz
Juniper Networks, Inc.
3260 Jay St.
Santa Clara, CA 95054
email: dkatz@jnx.com

Yakov Rekhter
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
email: yakov@cisco.com