Network Working Group
Internet Draft
Expiration Date April 2006

Robert Raszuk
Keyur Patel
Chandra Appanna
David Ward
Cisco Systems, Inc

October 2005

BGP Aggregate Withdraw
draft-raszuk-aggr-withdraw-00.txt

Status of this Memo

Abstract

This document proposes a scheme that allows a BGP speaker to withdraw multiple NLRIs that share a set of properties more efficiently by just specifying the shared properties among them.

1. Introduction

This document proposes a scheme that allows a BGP speaker to withdraw multiple NLRIs that share a set of properties more efficiently by just specifying the shared properties among them.

One area where this kind of feature is particularly important is 2547. The growth and success of 2547 VPN deployments forces operators and vendors to seek much more efficient and scalable mechanisms for vpn prefix management in VPN networks.

This draft introduces new BGP attribute called MP_AGGREGATE_WITHDRAW attribute which allows BGP to withdraw multiple NLRIs in a single message thereby reducing significantly the load on routers, number of BGP update messages and convergence time.

MP_AGGREGATE_WITHDRAW can also be used to implement Graceful Shutdown functionality to allow rerouting of traffic before the BGP session is down.

This mechanism is applicable to and works for any BGP AFI/SAFI.

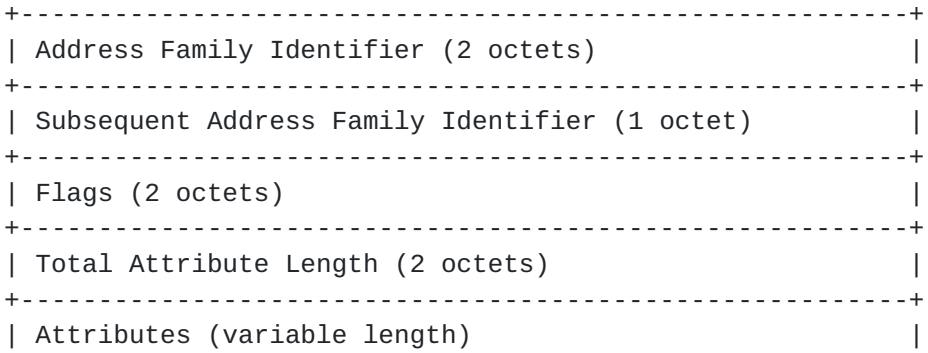
2. Specification of Requirements

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

3. MP_AGGREGATE_WITHDRAW Attribute (Type Code TBD by IANA)

This is an optional non-transitive attribute that can be used for the purpose of aggregating multiple unfeasible NLRIs to be removed from service.

The attribute is encoded as shown below:



```
+-------------------------------------------------------+
| TLVs (optional & variable length)                     |
+-------------------------------------------------------+
```

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 RFC 1700 (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.

Flags:

This 2-octet unsigned integer indicates Flags value for the the MP_AGGREGATE_WITHDRAW. The flags are defined as:

0x01 Withdraw paths that match all attributes
0x02 Withdraw paths that match any one or more attributes
0x04 Set to one only when TLVs are present

Total Attribute Length:

This 2-octet unsigned integer indicates the total length of the Path Attributes field in octets. Its value allows the length of the Network Layer Reachability field to be determined as specified below.

A value of 0 indicates that neither the Network Layer Reachability Information field, nor the Path Attribute field is present in this UPDATE message.

Attributes:

For format description refer to [BGP4].

TLVs:

In the case where there is a need to send other information then those carried in BGP attributes to uniquely identify the NLRIs to be withdrawn we define a TLV field.

The following TLV format has been defined:

Type One octet field set to value of given TLV.

Length	One octet field that indicates the length of the value portion in octets.
Reserved	One octet field reserved for future flags
Value	Description of the value carried in given TLV

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

Only one or zero of TLV value per MP_AGGREGATE_WITHDRAW attribute should be present. If the TLV value is present alone (no attributes) the match should happen on this value alone.

4. TLV definitions

4.1 Route Distinguisher

In the 2547 VPNs [RFC2547] in the MP_AGGREGATE_WITHDRAW there is a need for unique identification of VPN routes to which attached attributes belong to. This is accomplished by distributing route distinguisher in the following tlv encoding:

Type: One octet field set to value of 1
Length: One octet field set to value of eight
Reserved: One octet field reserved (all zeros)
Value: Eight octet RD value

4.2 TIME_TO_WITHDRAW

This time represents a TIME_TO_WITHDRAW. It is has a value field length of 2 octet. This type represents the time after which the forwarding support will be withdrawn for all reachability associated with the MP_AGGREGATE_WITHDRAW and is a value in seconds.

Type: One octet field set to value of 1
Length: One octet field set to a value of 2
Reserved: One octet field reserved (all zeros)
Value: 2 octet value representing number of seconds

5. MP_AGGREGATE_WITHDRAW Capability

The MP_AGGREGATE_WITHDRAW Capability is a new BGP capability [BGP-CAP] that can be used by a BGP speaker to indicate its ability to receive and send aggregated withdraws.

This capability is defined as follows:

Capability code: TBD by IANA

Capability length: variable

Capability value: Consists of the one or more of the tuples <AFI, SAFI> as follows:

```
+-------------------------------------------------+
| Address Family Identifier (16 bits)             |
+-------------------------------------------------+
| Subsequent Address Family Identifier (8 bits)   |
+-------------------------------------------------+
| ...                                             |
+-------------------------------------------------+
| Address Family Identifier (16 bits)             |
+-------------------------------------------------+
| Subsequent Address Family Identifier (8 bits)   |
+-------------------------------------------------+
```

Address Family Identifier (AFI):

This field carries the identity of the Network Layer protocol for which the Graceful Restart support is advertised. Presently defined values for this field are specified in [RFC1700].

Subsequent Address Family Identifier (SAFI):

This field provides additional information about the type of the Network Layer Reachability Information carried in the attribute. Presently defined values for this field are specified in [RFC1700].

6. Aggregate Withdraw Extended Community Attribute

Aggregate Withdraw Extended Community is a mandatory non-transitive extended community that can be used for the purpose of uniformed marking closed NLRI groups with common fate sharing. The mandatory requirement comes from a fact that an implementation which supports MP_AGGREGATE_WITHDRAW must also support Aggregate Withdraw Extended Community.

Aggregate Withdraw Extended Community attribute is carried in BGP Extended Community Attribute of type code 16.

The Aggregate Withdraw Extended Community attribute is encoded as follows:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|  Type high    |   Type low    |             Value             |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                      Value (cont.)                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

The value of the high-order octet of the type field for the Marker

Community can be 0x43. That indicated first come first served IANA type of assignment, non-transitive, opaque extended community

The value of the low-order octet of the type field for this community is .... (TBD).

The value is a locally significant 6 octet value assigned by bgp speaker to differentiate the routes based on various operator's depended requirements. It's allocation can be fully algorithmic and automatic or it could be assigned some meaningful structure. Being a locally significant it can be overwritten by any BGP speaker.

7. Operation

A BGP speaker receiving an update message with MP_AGGREGATE_WITHDRAW does not support MP_AGGREGATE_WITHDRAW capability, it simply ignores the message and logs the warning.

The BGP speaker implementing MP_AGGREGATE_WITHDRAW capability and receiving an update message with MP_AGGREGATE_WITHDRAW should remove all the NLRIs (paths) that match the attribute and TLV list specified in the MP_AGGREGATE_WITHDRAW attribute for each AFI/SAFI. The matching of the attributes is further qualified by the operation type specified in the flags field associated with the AFi/SAFI and can be logical AND or OR.

The additional TLV value presence is indicated by by the flags field. It's value will always be a logical AND to all other attributes if present.

If the the TIME_TO_WITHDRAW is sent in the MP_AGGREGATE_WITHDRAW, it must be interpreted by the receiveing BGP speaker as the minimum duration for which the sending BGP speaker will preserve forwarding of reachability already announced prior to receiving this MP_AGGREGATE_WITHDRAW. The purpose of TIME_TO_WITHDRAW is to allow the implentation of Graceful Shutdown funtionality whereby the receiving BGP speaker is provided a some time to reconverge before the sending BGP speaker is no longer available for forwarding traffic.

In the event of an AFI/SAFI being in the MP_AGGREGATE_WITHDRAW attribute that is not supported as per the initial capability negotiation, a BGP Notification message with the notification code set to UNSUPPORTED_AFI_SAFI should be sent and the session should be terminated.

8. Deployment Considerations

8.1 Sessions to all CEs in a vrf goes down or is being shutdown.

Today: All vrf routes are send within MP_UNREACH

New: A single message with RD lists all export RTs which were under given vrf is being send.

8.2 Sessions to one CE in a vrf goes down or is being shutdown.

Today: All routes from a given CE are send within MP_UNREACH

New: A single message with marker extended community and optionally an RD under given vrf is being send.

8.3 A subset of routes or all routes of given AFI/SAFI marked with a unique community or an attribute

Today: It would require to send all route in an MP_UNREACH attribute

New: Just one msg with MP_AGGREGATE_WITHDRAW listing this unique attribute would be sufficient

8.4 A BGP next hop on NLRIs with a single path goes down

Today: It would require to send all routes in an MP_UNREACH attribute

New: Just one msg with MP_AGGREGATE_WITHDRAW listing this next hop will be sufficient.

9. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing BGP [RFC2385].

10. Acknowledgments

The authors would like to thank Dan Tappan and Shyam Suri for their suggestions and feedback.

11.IANA Considerations

This document defines new BGP MP_AGGREGATE_WITHDRAW attribute. New attribute code should be introduced using the Standards Action process defined in [RFC-2434].

12. Normative References

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

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

13. Informative References

[RFC1771] Rekhter, Y., and T. Li, "A Border Gateway Protocol 4

(BGP-4)", RFC 1771, March 1995.

[IDR-BGP4] Rekhter, Y., and T. Li, "A Border Gateway Protocol 4 (BGP-4)", Work in Progress (draft-ietf-idr-bgp4-21.txt), April 2003.

[RFC2858] Bates, T., Rekhter, Y., Chandra, R., Katz, D., "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000

[RFC2547] Rosen, E., Rekhter Y., "BGP/MPLS VPNs", RFC2547, March 1999

[RFC2434] Narten, T., and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", RFC 2434/BCP 0026, October, 1998.

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August, 1998.

[RFC1700] Reynolds, J., and Postel, J., "Assigned Numbers", STD 2, RFC 1700, October 1994. See also: http://www.iana.org/numbers.html.

14. Authors' Addresses

Robert Raszuk
Cisco Systems, Inc.
170 West Tasman Dr
San Jose, CA 95134
raszuk@cisco.com

Keyur Patel
Cisco Systems, Inc.
170 West Tasman Dr
San Jose, CA 95134
keyupate@cisco.com

Chandra Appanna
Cisco Systems, Inc.
170 West Tasman Dr
San Jose, CA 95134
achandra@cisco.com

David Ward
Cisco Systems, Inc.
170 West Tasman Dr
San Jose, CA 95134
wardd@cisco.com

15. IPR Notices

The IETF takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Information on the IETF's procedures with respect to rights in standards-track and standards-related documentation can be found in BCP-11. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementors or users of this specification can be obtained from the IETF Secretariat.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this standard. Please address the information to the IETF Executive Director.

16. Terms of Use

Cisco has a pending patent which relates to the subject matter of this Internet Draft. If a standard relating to this subject matter is adopted by IETF and any claims of any issued Cisco patents are necessary for practicing this standard, any party will be able to obtain a license from Cisco to use any such patent claims under openly specified, reasonable, non-discriminatory terms to implement and fully comply with the standard.

17. Full Copyright Notice