INTERNET-DRAFT   Paul Traina
                                                   Blissfully Retired
                                                     Danny McPherson
                                                       Arbor Networks
                                                         John Scudder
                                                     Juniper Networks
Expires: August 2007   February 2007

Autonomous System Confederations for BGP
<draft-ietf-idr-rfc3065bis-06.txt>

Status of this Memo

Copyright Notice

Abstract

   The Border Gateway Protocol (BGP) is an inter-autonomous system
   routing protocol designed for Transmission Control Protocol/Internet
   Protocol (TCP/IP) networks. BGP requires that all BGP speakers
   within a single autonomous system (AS) must be fully meshed. This
   represents a serious scaling problem that has been well documented in
   a number of proposals.

   This document describes an extension to BGP which may be used to
   create a confederation of autonomous systems that is represented as a
   single autonomous system to BGP peers external to the confederation,
   thereby removing the "full mesh" requirement. The intention of this
   extension is to aid in policy administration and reduce the
   management complexity of maintaining a large autonomous system.

Table of Contents

1. Introduction

As currently defined, BGP requires that all BGP speakers within a
single AS must be fully meshed. The result is that for n BGP
speakers within an AS n*(n-1)/2 unique IBGP sessions are required.
This "full mesh" requirement clearly does not scale when there are a
large number of IBGP speakers within the autonomous system, as is
common in many networks today.

This scaling problem has been well documented and a number of
proposals have been made to alleviate this [RFC 1863, RFC 2796].
This document presents another alternative alleviating the need for a
"full mesh" and is known as "Autonomous System Confederations for
BGP", or simply, "BGP Confederations". It has also been observed
that BGP Confederations may provide improvements in routing policy
control.

This document is a revision of [RFC 3065], which is itself a revision
to [RFC 1965]. It includes editorial changes, terminology
clarifications and more explicit protocol specifications based on
extensive implementation and deployment experience with BGP
Confederations.

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 RFC 2119 [RFC 2119].

1.2. Terminology

AS Confederation

   A collection of autonomous systems represented and advertised
   as a single AS number to BGP speakers that are not members of
   the local BGP confederation.

AS Confederation Identifier

   An externally visible autonomous system number that identifies
   a BGP confederation as a whole.

Member Autonomous System (Member-AS)

   An autonomous system that is contained in a given AS
   confederation. Note that "Member Autonomous System" and
   "Member-AS" are used entirely interchangeably throughout
   this document.

Member-AS Number

   An autonomous system number identifier visible only within
   a BGP confederation, and used to represent a Member-AS
   within that confederation.

2. Discussion

   It may be useful to subdivide autonomous systems with a very large
   number of BGP speakers into smaller domains for purposes of
   controlling routing policy via information contained in the BGP
   AS_PATH attribute. For example, one may choose to consider all BGP
   speakers in a geographic region as a single entity.

   In addition to potential improvements in routing policy control, if
   techniques such as those presented here or in [RFC 2796] are not
   employed, [BGP-4] requires BGP speakers in the same autonomous system
   to establish a full mesh of TCP connections among all speakers for
   the purpose of exchanging exterior routing information. In
   autonomous systems the number of intra-domain connections that need
   to be maintained by each border router can become significant.

   Subdividing a large autonomous system allows a significant reduction
   in the total number of intra-domain BGP connections, as the
   connectivity requirements simplify to the model used for inter-domain
   connections.

   Unfortunately, subdividing an autonomous system may increase the
   complexity of routing policy based on AS_PATH information for all
   members of the Internet. Additionally, this division increases the
   maintenance overhead of coordinating external peering when the
   internal topology of this collection of autonomous systems is
   modified.

   Therefore, division of an autonomous system into separate systems may
   adversely affect optimal routing of packets through the Internet.

   However, there is usually no need to expose the internal topology of

this divided autonomous system, which means it is possible to regard
a collection of autonomous systems under a common administration as a
single entity or autonomous system, when viewed from outside the
confines of the confederation of autonomous systems itself.

3. AS_CONFED Segment Type Extension

Currently, BGP specifies that the AS_PATH attribute is a well-known
mandatory attribute that is composed of a sequence of AS path
segments. Each AS path segment is represented by a triple <path
segment type, path segment length, path segment value>.

In [BGP-4], the path segment type is a 1-octet long field with the
two following values defined:

Value   Segment Type

  1   AS_SET: unordered set of autonomous systems a route in
           the UPDATE message has traversed

  2   AS_SEQUENCE: ordered set of autonomous systems a route
           in the UPDATE message has traversed

This document specifies two additional segment types:

  3   AS_CONFED_SEQUENCE: ordered set of Member Autonomous
           Systems in the local confederation that the UPDATE message
           has traversed

  4   AS_CONFED_SET: unordered set of Member Autonomous Systems
           in the local confederation that the UPDATE message has
           traversed

4. Operation

A member of a BGP confederation MUST use its AS Confederation
Identifier in all transactions with peers that are not members of its
confederation. This AS confederation identifier is the "externally
visible" AS number and this number is used in OPEN messages and
advertised in the AS_PATH attribute.

A member of a BGP confederation MUST use its Member-AS Number in all

transactions with peers that are members of the same confederation as
the local BGP speaker.

A BGP speaker receiving an AS_PATH attribute containing an autonomous
system matching its own AS Confederation Identifier SHALL treat the
path in the same fashion as if it had received a path containing its
own AS number.

A BGP speaker receiving an AS_PATH attribute containing an
AS_CONFED_SEQUENCE or AS_CONFED_SET which contains its own Member-AS
Number SHALL treat the path in the same fashion as if it had received
a path containing its own AS number.

4.1. AS_PATH Modification Rules

When implementing BGP Confederations Section 5.1.2 of [BGP-4] is
replaced with the following text:

When a BGP speaker propagates a route which it has learned from
another BGP speaker's UPDATE message, it SHALL modify the route's
AS_PATH attribute based on the location of the BGP speaker to which
the route will be sent:

a) When a given BGP speaker advertises the route to another BGP
   speaker located in its own Member-AS, the advertising speaker
   SHALL NOT modify the AS_PATH attribute associated with the
   route.

b) When a given BGP speaker advertises the route to a BGP speaker
   located in a neighboring autonomous system that is a member of
   the local confederation, the advertising speaker SHALL update
   the AS_PATH attribute as follows:

   1) if the first path segment of the AS_PATH is of type
      AS_CONFED_SEQUENCE, the local system SHALL prepend its own
      Member-AS Number as the last element of the sequence (put
      it in the leftmost position).

   2) if the first path segment of the AS_PATH is not of type
      AS_CONFED_SEQUENCE the local system SHALL prepend a new path
      segment of type AS_CONFED_SEQUENCE to the AS_PATH, including
      its own Member-AS Number in that segment.

c) When a given BGP speaker advertises the route to a BGP speaker
   located in a neighboring autonomous system that is not a member of

        the local confederation, the advertising speaker SHALL update the
        AS_PATH attribute as follows:

        1) if any path segments of the AS_PATH are of the type
           AS_CONFED_SEQUENCE or AS_CONFED_SET, those segments MUST
           be removed from the AS_PATH attribute, leaving the sanitized
           AS_PATH attribute to be operated on by steps 2 or 3.

        2) if the first path segment of the remaining AS_PATH is of type
           AS_SEQUENCE, the local system SHALL prepend its own
           AS Confederation Identifier as the last element of the sequence
           (put it in the leftmost position).

        3) if there are no path segments following the removal of the
           first AS_CONFED_SET/AS_CONFED_SEQUENCE segments, or if the
           first path segment of the remaining AS_PATH is not of type
           AS_SEQUENCE the local system SHALL prepend a new path segment
           of type AS_SEQUENCE to the AS_PATH, including its own AS
           Confederation Identifier in that segment.

     When a BGP speaker originates a route:

     a) the originating speaker SHALL include an empty AS_PATH attribute
        in all UPDATE messages sent to BGP speakers residing within the
        same Member-AS. (An empty AS_PATH attribute is one whose length
        field contains the value zero).

     b) the originating speaker SHALL include its own Member-AS Number in
        an AS_CONFED_SEQUENCE segment of the AS_PATH attribute of all
        UPDATE messages sent to BGP speakers located in neighboring
        Member Autonomous Systems that are members of the local
        confederation (i.e., the originating speaker's Member-AS Number
        will be the only entry in the AS_PATH attribute).

     c) the originating speaker SHALL include its own AS Confederation
        Identifier in an AS_SEQUENCE segment of the AS_PATH attribute of
        all UPDATE messages sent to BGP speakers located in neighboring
        autonomous systems that are not members of the local
        confederation. (In this case, the originating speaker's AS
        Confederation Identifier will be the only entry in the AS_PATH
        attribute).

[5]. Error Handling

     A BGP speaker MUST NOT transmit updates containing AS_CONFED_SET or

AS_CONFED_SEQUENCE attributes to peers that are not members of the
local confederation.

It is an error for a BGP speaker to receive an update message with an
AS_PATH attribute which contains AS_CONFED_SEQUENCE or AS_CONFED_SET
segments from a neighbor which is not located in the same
confederation. If a BGP speaker receives such an update message, it
SHALL treat the message as having a malformed AS_PATH according to
the procedures of [BGP-4] Section 6.3 ("UPDATE message error
handling").

It is a error for a BGP speaker to receive an update message from a
confederation peer which is not in the same Member-AS that does not
have AS_CONFED_SEQUENCE as the first segment. If a BGP speaker
receives such an update message, it SHALL treat the message as having
a malformed AS_PATH according to the procedures of [BGP-4] Section
6.3 ("Update message error handling").

5.1. Common Administrative Issues

It is reasonable for Member Autonomous Systems of a confederation to
share a common administration and IGP information for the entire
confederation. It is also reasonable for each Member-AS to run an
independent IGP. In the latter case, the NEXT_HOP may need to be set
using policy (i.e., by default it is unchanged).

5.2. MED and LOCAL_PREF Handling

It SHALL be legal for a BGP speaker to advertise an unchanged
NEXT_HOP and MULTI_EXIT_DISC (MED) attribute to peers in a
neighboring Member-AS of the local confederation.

MEDs of two routes SHOULD only be compared if the first autonomous
systems in the first AS_SEQUENCE in both routes are the same - i.e.,
skip all the autonomous systems in the AS_CONFED_SET and
AS_CONFED_SEQUENCE. An implementation MAY provide the ability to
configure path selection such that MEDs of two routes are comparable
if the first autonomous systems in the AS_PATHs are the same,
regardless of AS_SEQUENCE or AS_CONFED_SEQUENCE in the AS_PATH.

An implementation MAY compare MEDs received from a Member-AS via
multiple paths. An implementation MAY compare MEDs from different

Member Autonomous Systems of the same confederation.

In addition, the restriction against sending the LOCAL_PREF attribute
to peers in a neighboring autonomous system within the same
confederation is removed.

5.3. AS_PATH and Path Selection

Path selection criteria for information received from members inside
a confederation MUST follow the same rules used for information
received from members inside the same autonomous system, as specified
in [BGP-4].

In addition, the following rules SHALL be applied:

1) If the AS_PATH is internal to the local confederation (i.e., there
   are only AS_CONFED_* segments) consider the neighbor AS to be the
   local AS.

2) Otherwise, if the first segment in the path which is not an
   AS_CONFED_SEQUENCE or AS_CONFED_SET is an AS_SEQUENCE, consider
   the neighbor AS to be the leftmost AS_SEQUENCE AS.

3) When comparing routes using AS_PATH length, CONFED_SEQUENCE and
   CONFED_SETs SHOULD NOT be counted.

4) When comparing routes using the internal (iBGP learned) versus
   external (eBGP learned) rules, treat a route that is learned from
   a peer which is in the same confederation (not necessarily the
   same Member-AS) as "internal".

6. Compatability Considerations

All BGP speakers participating as member of a confederation MUST
recognize the AS_CONFED_SET and AS_CONFED_SEQUENCE segment type
extensions to the AS_PATH attribute.

Any BGP speaker not supporting these extensions will generate a
NOTIFICATION message specifying an "UPDATE Message Error" and a sub-
code of "Malformed AS_PATH".

This compatibility issue implies that all BGP speakers participating

   in a confederation MUST support BGP confederations. However, BGP
   speakers outside the confederation need not support these extensions.

7. Deployment Considerations

   BGP confederations have been widely deployed throughout the Internet
   for a number of years and are supported by multiple vendors.

   Improper configuration of BGP confederations can cause routing
   information within an AS to be duplicated unnecessarily. This
   duplication of information will waste system resources, cause
   unnecessary route flaps, and delay convergence.

   Care should be taken to manually filter duplicate advertisements
   caused by reachability information being relayed through multiple
   Member Autonomous Systems based upon the topology and redundancy
   requirements of the confederation.

   Additionally, confederations (as well as route reflectors), by
   excluding different reachability information from consideration at
   different locations in a confederation, have been shown [RFC 3365] to
   cause permanent oscillation between candidate routes when using the
   tie breaking rules required by BGP [BGP-4]. Care must be taken when
   selecting MED values and tie breaking policy to avoid these
   situations.

   One potential way to avoid this is by configuring inter-Member-AS IGP
   metrics higher than intra-Member-AS IGP metrics and/or using other
   tie breaking policies to avoid BGP route selection based on
   incomparable MEDs.

8. Security Considerations

   This extension to BGP does not change the underlying security issues
   inherent in the existing BGP protocol, such as those described in
   [RFC 2385] and [BGP-VULN].

9. Acknowledgments

   The general concept of BGP confederations was taken from IDRP's
   Routing Domain Confederations [ISO 10747]. Some of the introductory
   text in this document was taken from [RFC 2796].

   The authors would like to acknowledge Jeffrey Haas for his extensive
   feedback on this document. We'd also like to thank Bruce Cole,
   Srihari Ramachandra, Alex Zinin, Naresh Kumar Paliwal, Jeffrey Haas,
   Cengiz Alaettinoglu, Mike Hollyman and Bruno Rijsman for their
   feedback and suggestions.

   Finally, we'd like to acknowledge Ravi Chandra and Yakov Rekhter for
   providing constructive and valuable feedback on earlier versions of
   this specification.

[10](). IANA Considerations

This spefication introduces no new IANA considerations and therefore
requires no actions on the part of IANA.

11. References

11.1. Normative References

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

[RFC 1965] Traina, P. "Autonomous System Confederations for BGP",
     RFC 1965, June 1996.

[RFC 3065] Traina, P., McPherson, D. and Scudder, J., "Autonomous
     System Confederations for BGP", RFC 3065, February 2001.

11.2. Informative References

[ISO 10747] Kunzinger, C., Editor, "Inter-Domain Routing Protocol",
     ISO/IEC 10747, October 1993.

[RFC 1863] Haskin, D., "A BGP/IDRP Route Server alternative to a
     full mesh routing", RFC 1863, October 1995.

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

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

[RFC 2796] Bates, T., Chandra, R. and E. Chen, "BGP Route Reflection
     An Alternative to Full Mesh IBGP", RFC 2796, April 2000.

[RFC 3365] McPherson, D., Gill, V., Walton, D., Retana, A., "Border
     Gateway Protocol (BGP) Persistent Route Oscillation Condition",
     RFC 3345, August 2002.

[BGP-VULN] Murphy, S., "BGP Security Vulnerabilities Analysis",
     Internet-Draft, "Work in Progress".

12. Appendix A: Aggregate Routing Information

   As a practical matter, aggregation as discussed in [BGP-4] section
   9.2.2.2 is not generally employed within confederations. However, in
   the event that such aggregation is performed within a confederation,
   the rules of [BGP-4] should be followed, making the necessary
   substitutions between AS_SET and AS_CONFED_SET and similarly,
   AS_SEQUENCE and AS_CONFED_SEQUENCE. Confederation-type segments
   (AS_CONFED_SET and AS_CONFED_SEQUENCE) MUST be kept separate from
   non-confederation segments (AS_SET and AS_SEQUENCE). An
   implementation could also choose to provide a form of aggregation
   wherein non-confederation segments are aggregated as discussed in
   [BGP-4] section 9.2.2.2 and confederation-type segments are not
   aggregated.

   Support for aggregation of confederation-type segments is not
   mandatory.

13. Appendix B: Changes From RFC 3065

   The primary trigger for an update to RFC 3065 was regarding issues
   associated with AS path segment handling, in particular what to do
   when interacting with BGP peers external to a confederation and to
   ensure AS_CONFED_[SET|SEQUENCE] segment types are not propagated to
   peers outside of a confederation.

   As such, the "Error Handling" section above was added and applies not
   only to explicitly call attention to BGP Confederation speakers, but
   to all BGP speakers.

   Other changes are mostly trivial and surrounding some clarification
   and consistency in terminology and denoting that
   AS_CONFED_[SET|SEQUENCE] Segment Type handling should be just as it
   is in the base BGP specification [BGP-4].

14. Authors' Addresses

      Paul Traina
      Blissfully Retired
      Email: bgp-confederations <possibly at> st04.pst.org

      Danny McPherson
      Arbor Networks
      EMail: danny@arbor.net

      John G. Scudder
      Juniper Networks
      EMail: jgs@juniper.net

Intellectual Property Statement

      Intellectual Property Rights 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; nor does it represent that it has
      made any independent effort to identify any such rights. Information
      on the procedures with respect to rights in RFC documents can be
      found in BCP 78 and BCP 79.

      Copies of IPR disclosures made to the IETF Secretariat 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 implementers or users of this
      specification can be obtained from the IETF on-line IPR repository at
      http://www.ietf.org/ipr.

      The IETF invites any interested party to bring to its attention any
      copyrights, patents or patent applications, or other proprietary
      rights that may cover technology that may be required to implement
      this standard. Please address the information to the IETF at
      ietf-ipr@ietf.org.

Copyright Statement

      This document is subject to the rights, licenses and restrictions
      contained in BCP 78, and except as set forth therein, the authors
      retain all their rights.

      This document and the information contained herein are provided on an
      "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS

OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST
AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES,
EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT
THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY
IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR
PURPOSE.

Acknowledgment