                  Autonomous System Confederations for BGP

Status of This Memo

   This document specifies an Internet standards track protocol for the
   Internet community, and requests discussion and suggestions for
   improvements. Please refer to the current edition of the "Internet
   Official Protocol Standards" (STD 1) for the standardization state
   and status of this protocol. Distribution of this memo is unlimited.

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 that 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.

   This document obsoletes RFC 3065.

Table of Contents

1. Introduction

   As originally 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 Internal BGP (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, such as [RFC2796] and
   [RFC1863] (made historic by [RFC4223]). 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, and obsoletes, [RFC3065], which is
   itself a revision of [RFC1965]. 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 [RFC2119].

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 [RFC4456] 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 field with the two
   following values defined:

   Value   Segment Type

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

     2   AS_SEQUENCE: ordered set of autonomous systems that 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 that 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:

AS_PATH is a well-known mandatory attribute. This attribute
identifies the autonomous systems through which routing information
carried in this UPDATE message has passed. The components of this
list can be AS_SETs, AS_SEQUENCEs, AS_CONFED_SETs or
AS_CONFED_SEQUENCES.

When a BGP speaker propagates a route it learned from another BGP
speaker's UPDATE message, it modifies 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 updates 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 prepends its own Member-AS
      number as the last element of the sequence (put it in the
      leftmost position with respect to the position of octets in the
      protocol message). If the act of prepending will cause an
      overflow in the AS_PATH segment (i.e., more than 255 ASs), it
      SHOULD prepend a new segment of type AS_CONFED_SEQUENCE and
      prepend its own AS number to this new segment.

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

   3) if the AS_PATH is empty, the local system creates a path
      segment of type AS_CONFED_SEQUENCE, places its own Member-AS
      Number into that segment, and places that segment into the
      AS_PATH.

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, 3 or 4.

   2) if the first path segment of the remaining AS_PATH is of type
      AS_SEQUENCE, the local system prepends its own AS Confederation
      Identifier as the last element of the sequence (put it in the
      leftmost position with respect to the position of octets in the
      protocol message). If the act of prepending will cause an
      overflow in the AS_PATH segment (i.e., more than 255 ASs), it
      SHOULD prepend a new segment of type AS_SEQUENCE and prepend
      its own AS number to this new segment.

   3) if the first path segment of the remaining AS_PATH is of type
      AS_SET, the local system prepends a new path segment of type
      AS_SEQUENCE to the AS_PATH, including its own AS Confederation
      Identifier in that segment.

   4) if the remaining AS_PATH is empty, the local system creates a
      path segment of type AS_SEQUENCE, places its own AS
      Confederation Identifier into that segment, and places that
      segment into the AS_PATH.

When a BGP speaker originates a route then:

a) the originating speaker includes its own AS Confederation
   Identifier in a path segment, of type AS_SEQUENCE, in 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 AS Confederation Identifier of
   the originating speaker's autonomous system will be the only entry
   the path segment, and this path segment will be the only segment
   in the AS_PATH attribute.

b) the originating speaker includes its own Member-AS Number in a
   path segment, of type AS_CONFED_SEQUENCE, in 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. In this case, the Member-AS Number of the
   originating speaker's autonomous system will be the only entry the
   path segment, and this path segment will be the only segment in
   the AS_PATH attribute.

c) the originating speaker includes 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).

Whenever the modification of the AS_PATH attribute calls for
including or prepending the AS Confederation Identifier or Member-AS
Number of the local system, the local system MAY include/prepend more
than one instance of that value in the AS_PATH attribute. This is
controlled via local configuration.

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 that contains AS_CONFED_SEQUENCE or AS_CONFED_SET
segments from a neighbor that 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 that 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 Interior Gateway Protocol (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 that 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 that is in the same confederation (not necessarily the same
      Member-AS) as "internal".

6. Compatibility Considerations

   All BGP speakers participating as members 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 [RFC3345] 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
   [RFC2385] and [BGP-VULN].

9. Acknowledgments

   The general concept of BGP confederations was taken from IDRP's
   Routing Domain Confederations [ISO10747]. Some of the introductory
   text in this document was taken from [RFC2796].

   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. References

10.1. Normative References

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

   [RFC1965] Traina, P., "Autonomous System Confederations for BGP",
              RFC 1965, June 1996.

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

   [RFC3065] Traina, P., McPherson, D., and J. Scudder, "Autonomous
              System Confederations for BGP", RFC 3065, February 2001.

10.2. Informative References

   [ISO10747] Kunzinger, C., Editor, "Inter-Domain Routing Protocol",
              ISO/IEC 10747, October 1993.

   [RFC1863] Haskin, D., "A BGP/IDRP Route Server alternative to a full
              mesh routing", RFC 1863, October 1995.

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

   [RFC3345] McPherson, D., Gill, V., Walton, D., and A. Retana,
              "Border Gateway Protocol (BGP) Persistent Route
              Oscillation Condition", RFC 3345, August 2002.

   [RFC4223] Savola, P., "Reclassification of RFC 1863 to Historic",
              RFC 4223, October 2005.

   [RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC
              4272, January 2006.

   [RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route
              Reflection: An Alternative to Full Mesh Internal BGP
              (IBGP)", RFC 4456, April 2006.

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.

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 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].

Authors' Addresses

   Paul Traina
   Blissfully Retired
   Email: bgp-confederations@st04.pst.org

   Danny McPherson
   Arbor Networks
   EMail: danny@arbor.net

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

Full Copyright Statement

Intellectual Property

Acknowledgement