INTERNET-DRAFT Paul Traina

Danny McPherson
Arbor Networks
John Scudder
Cisco Systems
Expires: November 2004 May 2004

Autonomous System Confederations for BGP
<draft-ietf-idr-rfc3065bis-02.txt>

Status of this Document

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

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].

This document is a product of the . Comments should be addressed to the authors, or the mailing list at

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 [3,6]. 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 [5], which is itself a revision to RFC 1965 [4]. It includes editorial changes, terminology clarifications and more explicit protocol specifications based on deployment experience with BGP Confederations. These revisions are summarized in Appendices A and B.

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.

3. 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 [6] are not employed, [1] 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.

4. AS_CONFED Segement 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 [1], 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

5. 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.

5.1. AS_PATH Modification Rules

When implementing BGP Confederations Section 5.1.2 of [1] 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 autonomous system, the advertising speaker SHALL 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).

6. 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 [1] 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 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 [1] Section 6.3 ("Update message error handling").

7. Common Administration Issues

It is reasonable for Member Autonomous Systems of a confederation to share a common administration and IGP information for the entire confederation.

7.1. 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.

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.

7.2. 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 [1].

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.

8. 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.

9. 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 [9] to cause permanent oscillation between candidate routes when using the tie breaking rules required by BGP [1]. 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.

10. Intellectual Property

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.

11. Acknowledgments

The general concept of BGP confederations was taken from IDRP's Routing Domain Confederations [2]. Some of the introductory text in this document was taken from [6].

The authors would like to acknowledge Bruce Cole for his implementation feedback and extensive analysis of the limitations of the protocol extensions described in this document and [5]. We would also like to acknowledge Srihari Ramachandra, Alex Zinin, Naresh Kumar Paliwal, Jeffrey Haas 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.

12. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing BGP, such as those defined in [7].

13. References

[1] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

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

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

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

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

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

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

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

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

14. Authors' Addresses

Paul Traina
EMail: pst+confed@spamcatcher.bogus.com

Danny McPherson
Arbor Networks
EMail: danny@arbor.net

John G. Scudder
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
Phone: +1 734.302.4128
EMail: jgs@cisco.com

15. Full Copyright Statement