Network Working Group
Request for Comments: 3065
Obsoletes: 1965
Category: Standards Track

P. Traina
Juniper Networks, Inc.
D. McPherson
Amber Networks, Inc.
J. Scudder
Cisco Systems, Inc.

February 2001

Autonomous System Confederations for BGP

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.

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].

2. 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 [3,5]. This document represents another alternative in alleviating the need for a "full mesh" and is known as "Autonomous System Confederations for BGP", or simply, "BGP Confederations". It can also be said the BGP Confederations MAY provide improvements in routing policy control.

This document is a revision of RFC 1965 [4] and it includes editorial changes, clarifications and corrections based on the deployment experience with BGP Confederations. These revisions are summarized in Appendix A.

3. Terms and Definitions

AS Confederation

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

AS Confederation Identifier

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

Member-AS

An autonomous system that is contained in a given AS confederation.

Member-AS Number

An autonomous system number visible only internal to a BGP confederation.

4. 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 [5] 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.

Finally, dividing a large AS may unnecessarily increase the length of the sequence portions of the AS_PATH attribute. Several common BGP implementations can use the number of "AS hops" required to reach a given destination as part of the path selection criteria. While this is not an optimal method of determining route preference, given the lack of other in-band information, it provides a reasonable default behavior which is widely used across the Internet. 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.

5. 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 [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 ASs a route in the UPDATE message has traversed
2	AS_SEQUENCE: ordered set of ASs a route in the UPDATE message has traversed

This document reserves two additional segment types:

3	AS_CONFED_SEQUENCE: ordered set of Member AS Numbers in the local confederation that the UPDATE message has traversed
4	AS_CONFED_SET: unordered set of Member AS Numbers in the local confederation that the UPDATE message has traversed

6. Operation

A member of a BGP confederation will use its AS Confederation ID in all transactions with peers that are not members of its confederation. This confederation identifier is considered to be 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 will use its Member AS Number in all transactions with peers that are members of the same confederation as the given router.

A BGP speaker receiving an AS_PATH attribute containing an autonomous system matching its own confederation 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.

6.1. AS_PATH Modification Rules

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 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 autonomous system confederation, then 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 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 confederation identifier 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 current autonomous system 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, that segment and any immediately following segments of the type AS_CONFED_SET or AS_CONFED_SEQUENCE are 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 confederation ID 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 of type AS_SET the local system shall prepend a new path segment of type AS_SEQUENCE to the AS_PATH, including its own confederation ID 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 located in its own Member AS Number. (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-AS 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 autonomous system 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 autonomous system number of the originating speaker's member confederation will be the only entry in the AS_PATH attribute).

7. Common Administration Issues

It is reasonable for member ASs of a confederation to share a common administration and IGP information for the entire confederation.

It shall be legal for a BGP speaker to advertise an unchanged NEXT_HOP and MULTI_EXIT_DISCRIMINATOR (MED) attribute to peers in a neighboring AS within the same confederation. In addition, the restriction against sending the LOCAL_PREFERENCE attribute to peers in a neighboring AS within the same confederation is removed. 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].

8. Compatability Considerations

All BGP speakers participating in 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 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. Security Considerations

This extension to BGP does not change the underlying security issues inherent in the existing BGP, such as those defined in [6].

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 [5].

The authors would like to acknowledge Bruce Cole of Juniper Networks 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 of Cisco Systems, Inc., for his feedback.

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

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

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

13. Authors’ Addresses

Paul Traina
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089 USA

Phone: +1 408 745-2000
EMail: pst+confed@juniper.net

Danny McPherson
Amber Networks, Inc.
48664 Milmont Drive
Fremont, CA 94538

Phone: +1 510.687.5226
EMail: danny@ambernetworks.com

John G. Scudder
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134

Phone: +1 734.669.8800
EMail: jgs@cisco.com

Appendix A: Comparison with RFC 1965

The most notable change from [1] is that of reversing the values AS_CONFED_SEQUENCE(4) and AS_CONFED_SET(3) to those defined in section "AS_CONFED Segment Type Extension". The reasoning for this is that in the initial implementation, which was already widely deployed, they were implemented backwards from [4], and as such, subsequent implementations implemented them backwards as well. In order to foster interoperability and compliance with deployed implementations, they've therefore been changed here as well.

The "Compatibility Discussion" was removed and incorporated into other discussions in the document. Also, the mention of hierarchical confederations is removed. The use of the term "Routing Domain Identifier" was replaced with Member AS Number.

Finally, the "Deployment Considerations" section was expanded a few subtle grammar changes were made and a bit more introductory text was added.

Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.