Internet Engineering Task Force
INTERNET DRAFT

Olivier Bonaventure
UCL
Stefaan De Cnodder
Alcatel
Jeffrey Haas
NextHop
Bruno Quoitin
FUNDP
Russ White
Cisco
February, 2003
Expires October, 2003

Controlling the redistribution of BGP routes
<draft-ietf-ptomaine-bgp-redistribution-02.txt>

Status of this Memo

Abstract

This document proposes the redistribution extended community. This new extended community allows a router to influence how a specific route should be redistributed towards a specified set of eBGP speakers. Several types of redistribution communities are proposed. The first type may be used to indicate that a specific route should not be announced over a specified set of eBGP sessions. The second type may be used to indicate that the attached route should only be

announced with the NO_EXPORT community over the specified set of eBGP sessions and the third type may be used to indicate that the attached route should be prepended n times when announced over a specified set of eBGP sessions.

1 Introduction

In today's commercial Internet, many ISPs need to have some control on their inter-domain traffic. In the outgoing direction, this control can be obtained by configuring the BGP routers of the ISP to favor some routes over others by using the LOCAL-PREF attribute. However, due to the asymetry of Internet traffic, most ISPs also need to control their incoming traffic.

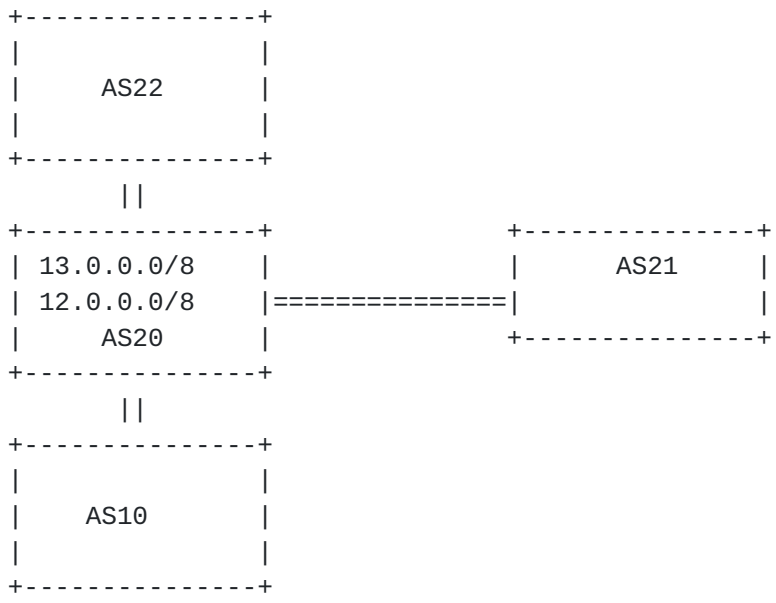

Figure 1: Simple inter-domain topology

In the incoming direction, the only way to influence the traffic flow is to control the redistribution of its routes. Several methods exist and are used in practice [Halabi97,COM-SUR]. In this case, the ISP needs to influence the redistribution and the selection of its own routes by remote ISPs. Since the default configuration of many BGP routers is to select the route with the smallest AS path length, a common technique is to artificially increase the length of the AS path for some announced routes. For example, in figure 1, if AS20 wanted to indicate that it prefers to receive its traffic towards subnet 13.0.0.0/8 through its link with AS22, then it would announce this prefix as usual on this link to AS22 and announce a prefix with the AS20:AS20:AS20:AS20 path to AS21 and AS10. If AS10 and AS21 use the AS path length to select the best BGP route, they will prefer the

shorter route received by AS22. This requires manual configuration of the BGP routers, but path prepending is very frequently used today on the Internet [Huston01]. In some cases, the configuration burden can be reduced by using the BGP communities attribute.

Recently, several large ISPs have gone one step further by defining BGP communities that allow their customers to influence the redistribution of their routes. For example, in figure 1, AS20 could configure its BGP routers to always prepend AS20 four times when they announce via eBGP a route received from one of AS20's customers with a special community attribute. For this, AS20 needs to publish the specific BGP communities that it supports and its customers need to configure their router appropriately. If AS20 needs to define a new BGP community or change an existing one, it must inform all its customers may need to update the configuration of their routers.

A more detailed survey of the utilization of the BGP community attribute by ISPs may be found in [COM-SUR]. This survey reveals the following:

- Many different ASes define their own BGP community values to allow their customers/peers to indicate that a particular route should not be propagated towards a specific AS, towards the routers attached to a specific IX, or towards ASes within a given geographical area (e.g. a European AS could want to prohibit a route from being announced to US peers).
- Many ASes define their own BGP community values to allow their peers or customers to indicate that an announced route should be prepended when announced towards a specific AS, IX or set of ASes.
- Several ASes define their own BGP community attribute to indicate that a given route should only be redistributed towards a specified AS.

Furthermore, this survey also reveals that some ASes have difficulty providing all these facilities while still relying on their assigned set of BGP community values. For example, some ASes have chosen to reuse several BGP community values in the private AS space (i.e. community values 64512:00 - 65534:65535) to be able to define structured communities that allow their customers to influence the redistribution of their routes. Some of these community values appear in BGP tables on the global Internet.

Although the survey shows that these BGP communities are widely used today to provide such facilities, this is far from the best solution. Requiring each AS to select its own values for the BGP communities and to document these values in the routing registries is not very efficient because it forces the BGP routers to be configured manually based on information found in these registries, peering agreements,

or elsewhere.

In this document, we define a new type of BGP extended community. By using a set BGP extended communities with a precise syntax, we support most of the current utilizations of the BGP community without relying unnecessarily on manual configuration of the BGP routers. We believe that reducing the manual configuration of these routers would be very useful for the stability and the performance of the global Internet.

2 Controlled redistribution of BGP routes

This document defines a method to allow a BGP speaker to influence how its peers will redistribute its own routes. For this, the BGP speaker may define for each announced route a redistribution policy that controls how this route will be redistributed. This is done by defining a set of allowed or requested operations and a list of BGP speakers. The list of BGP speakers can be specified by listing either the BGP speakers that are covered by the redistribution policy or those that are not covered by this policy. The current version of this document supports the following operations:

- the attached route should not be announced to the specified BGP speakers

- the attached route should only be announced to the specified BGP speakers

- the attached route should be announced with the NO_EXPORT attribute to the specified BGP speakers

- the attached route should be prepended n times when announced to to the specified BGP speakers

The redistribution policies are encoded in a special type of extended community called the redistribution community. If a redistribution policy applies to several of BGP speakers, then it will be encoded in several redistribution communities.

2.1 The redistribution community

The extended communities attribute is defined in [EXT-COM]. This

attribute allows a BGP router to attach a set of extended communi ties to an UPDATE message. Each extended community value is encoded as an eight octets quantity with a one or two octets type field and a six or seven octet value field. Several types of extended commu nity values are defined in [EXT-COM]. This document proposes a new well-known extended community: the redistribution community.

The redistribution community is composed of a one octet type field (regular type). It is encoded as defined in [EXT-COM]. The high- order bit is cleared (type assigned by IANA). Since the redistribu tion community is only used for signalling purposes between two neighbor AS's, bit6 is set meaning that the extended community is non-transitive across ASes. This is important to ensure that com munities used to affect the redistribution of routes by a given AS are not unnecessarily distributed over the entire Internet as it is often the case today [COM-SUR]. The remaining 6 lower-order bits are to be defined by IANA (TBDTBD notation in figure 1).

Figure 1: Encoding of the redistribution community

The remaining 7 octets of the redistribution community indicate how a router will advertise the received route to its peers. This requires two pieces of information: a filter to select a subset of BGP speakers and an action that indicates how the attached route should be redistributed to the selected BGP speakers. The high- order octet indicates the action to be taken and the 6 remaining octets define the filter.

The Action octet is encoded as follow:

- The high and the second order bits (Bit7 and Bit6) are reserved and set to zero in this document

- Bit5-3 are the Action type

- Bit2-0 are the Action parameters

Action types

This document defines three types of actions (values 000b - 010b). Values 011b-111b are reserved for future use and are to be assigned

by IANA by IETF consensus as defined in [RFC2434].

- 000b Prepend. This action means that the AS number of the announcing router should be prepended when announcing the attached route to the BGP speakers covered by the redistribution policy. The action parameter indicates how many times the AS number should be prepended. Using an action parameter of 000 with this type, although legal, will not cause any additional prepending of the route's AS PATH.

- 001b No_Export. This action means that the NO_EXPORT community should be inserted when announcing the attached route to the BGP speakers covered by the redistribution policy. This action type does not require a parameter. The action parameter should be set to zero by the sender and ignored by the receiver.

- 010b Do not announce. This action means that the route should not be announced to the BGP speakers covered by the redistribution pol icy. This action type does not require a parameter. The action parameter should be set to zero by the sender and ignored by the receiver.

The BGP Speakers Filter

The BGP_Speakers_Filter field is used to specify the BGP speakers that will be affected by the specified action. It is composed of a one octet type field and a five octets value field.

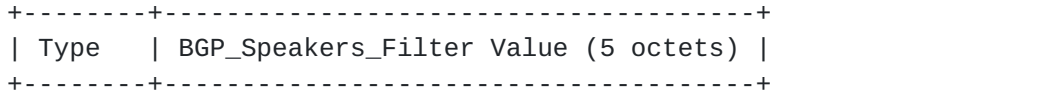

Figure 2: Encoding of the BGP_Speakers_Filter field

The BGP_Speakers_Filter field is used to specify the BGP speakers that will be affected by the specified action. There are two meth ods to specify the affected BGP speakers. The first method is to explicitly list all those speakers inside the BGP_Speakers_Filters field of the redistribution communities. In this case, the high order bit of the type field of the BGP_Speakers_Filter field is set to 1. The second method is to explicitly list only the BGP speak ers that will not be affected by the specified action. In this case, the high order bit of the BGP_Speakers_Filter type field shall be set to 0. The 7 low order bits of the BGP_Speakers_Filter type field are used to indicate the type of BGP speakers included

in the five low order octets of the BGP_Speakers_Filter field. This document defines four types of BGP_Speakers_Filters (values 0x01-0x04). Value 0x00 is reserved and values 0x05-0x3f are reserved for future use and are to be assigned by IANA by IETF con sensus as defined in [RFC2434]. Values 0x40-0x7f are vendor spe cific and are assigned by IANA on a first come, first serve basis.

BGP_Speakers_Filter types

- The BGP_Speakers_Filter value contains a two octet AS number (type 0x01)

- The BGP_Speakers_Filter value contains two two octet AS numbers type 0x02)

- The BGP_Speakers_Filter value contains a CIDR prefix/length pair (type 0x03)

- The BGP_Speakers_Filter value contains a four octets AS number (type 0x04)

The BGP_Speakers_Filter value shall be encoded as follows. If this field contains a two octet AS number, the AS number shall be placed in the two low order octets. The three high order octets shall be set to zero upon transmission and ignored upon reception.

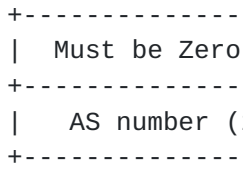

Figure 3: BGP speakers filter containing a single two octet AS number

If the BGP_Speakers_Filter value contains two two octet AS numbers, one of the AS numbers should be placed in the two low order octets. The other AS number should be placed in the next two higher order octets and the last octet shall be set to zero upon transmission and ignored upon reception.

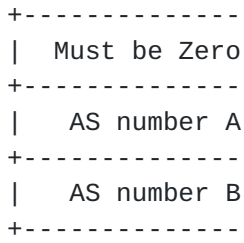

Figure 4: BGP speakers filter containing two distinct two octet AS number

If the BGP_Speakers_Filter value contains a four octet AS number, the AS number shall be placed in the four low order octets. The high order octet shall be set to zero upon transmission and ignored upon reception.

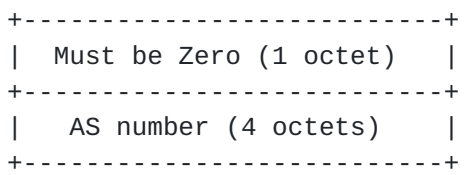

Figure 5: BGP speakers filter containing a single four octets AS number

If the BGP_Speakers_Filter value contains a CIDR prefix/length pair, it should be encoded as shown below:

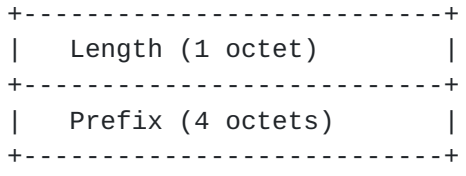

Figure 6: BGP speakers filter containing a CIDR prefix/length pair

The Length field indicates the length in bits of the IP address prefix. A length of zero indicates a prefix that matches all IP addresses. The Prefix field contains IP address prefixes followed by enough trailing bits with a value of zero to make the end of the field fall on a four octets boundary.

2.2 Utilization of the redistribution communities

A router may, depending on its policy, add one or several redistri bution communities to a route originated by itself or received from another BGP speaker over an iBGP or an eBGP session. In practice, it can be expected that the redistribution communities will often be attached by the originator of the route will as this is an attempt of the route originator to do some form of inter-domain traffic engineering. In practice, it can also be expected that most utilizations of the redistribution communities will only require to attach a small number of those communities to a given route.

When a router attaches one or several redistribution communities to a route, it must ensure that two of the included redistribution communities do not conflict. This is necessary to ensure that the redistribution communities will be processed in a deterministic manner by the remote BGP peer. When several redistribution

communities contain the same action type and parameter, then all the BGP_Speakers_filters of those communities must have the same high order bit in the BGP_Speakers_Filter type. A BGP router that receives a route containing invalid redistribution communities for a given action type and parameter should ignore all the redistribu tion communities concerning this action type and parameter. This ignore action must be logged.

2.3 Operations

This document defines the procedures to support the redistribution communities that are non-transitive extended communities of type 01TBDTBD. An implementation that understands the redistribution communities should discard and ignore upon receipt the extended communities of the form 00TBDTBD (i.e. same type as a redistribu tion community but transitive).

The redistribution communities defined in this document only affect the redistribution of the associated route over eBGP sessions. The redistribution communities do not affect the redistribution of routes over iBGP sessions or between the sub-ASes of a confedera tion.

When a router receives a route with redistribution communities, it should apply the operations specified by these communities when redistributing the route over eBGP sessions. Since the redistribu tion communities defined by this document are non-transitive, a router will remove the received redistribution communities when redistributing those routes over eBGP sessions. Of course, nothing prevents this router from adding its own redistribution communities to this route before redistributing it.

A router should apply the policies defined by the redistribution communities to the routes that is has selected for advertisement from its Adj-Rib-Out based on its own policy. A route that contains redistribution communities should be processed as follows:

First, the BGP speaker should build for each action type and param eter contained in the redistribution communities attached to the route a list of the target BGP speakers contained in the BGP_Speak ers_filters for this action type. In the remainder of this sec tion, we will use the wordings "an eBGP session is affected by action type x" to indicate that either of the following is true when the BGP_Speakers_Filters contain AS numbers:

- the AS number of the remote BGP peer appears inside one of the BGP_Speakers_Filter of the redistribution communities with action x and the high order bit of the BGP_Speakers_Filter type is set to one

- the AS number of the remote BGP peer does not appear inside any of the BGP_Speakers_Filter of the redistribution communities with action x and the high order bit of the BGP_Speakers_Filter type is set to zero

When the BGP_Speakers_Filter contains a CIDR prefix/length, we will use the wordings "an eBGP session is affected by action type x" to indicate that either of the following is true:

- the IP address of at least one of the endpoints of the eBGP ses sion belongs to the CIDR prefix specified inside one of the BGP_Speakers_Filter of the redistribution communities with action x and the high order bit of the BGP_Speakers_Filter type is set to one

- the IP addresses of the local and the remote endpoints of the eBGP session do not belong to the CIDR prefix specified inside one of the BGP_Speakers_Filter of the redistribution communities with action x and the high order bit of the BGP_Speakers_Filter type is set to zero

Then, when a route is about to be redistributed over an eBGP ses sion, the router first checks if this session is affected by action type "Do not announce". If this is the case, the route is not announced over this eBGP session. Otherwise, the router checks the other action types as follows.

- If the eBGP session is affected by action type "No export" then the well-known community NO_EXPORT is attached to the route.

- If the eBGP session is affected by one or more actions of type "Prepend", then the AS-Path of the route shall be prepended n times (with the AS number of the router redistributing the route) where n is the smallest parameter of the matched "Prepend" actions.

Otherwise the route is announced over the eBGP session.

2.4 Filtering and precedence

In order to allow operators to control the implementation of their policies, a BGP implementation that supports the redistribution communities should allow the operator to determine, on a per ses sion basis whether redistribution communities are permitted or denied on this session. For example, an AS could elect to receive redistribution communities from its customers, but not on a shared-cost peering session. A BGP implementation may provide additional details in the filtering of the redistribution communities, but this is implementation dependent and goes beyond this specifica tion.

A BGP implementation that supports both the normal (extended) com munities and the redistribution communities should allow the opera tor to adjust the order in which these types of communities are processed. The default precedence should be to first process the redistribution communities before processing the other manually defined (extended) communities.

3 IANA considerations

This document requests the attribution of a new BGP extended commu nities type [EXT-COM] field from IANA.

The redistribution community contains two fields that are to be maintained by IANA. The first field is the Action type that is part of the Action octet defined in section 2.1 shall be maintained by IANA. This document defines the utilization of action types 000b - 010b ; action types 011b - 111b are reserved for future use and are to be assigned by IANA by IETF consensus as defined in [RFC2434]. The second field are the seven low order bits of the Type octet of the BGP_Speakers_Filter defined in section 2.1. This document defines four types of BGP_Speakers_Filters (values 0x01-0x04). Value 0x00 is reserved and values 0x05-0x3f are reserved for future use and are to be assigned by IANA by IETF consensus as defined in [RFC2434]. Values 0x40-0x7f are vendor specific and are assigned by IANA on a first come, first serve basis.

4 Security considerations

This extension to BGP does not change the underlying security issues of the extended community attribute.

5 Conclusion

This document has proposed a new type of extended communities

called the redistribution communities. These redistribution commu nities can be used by a BGP router to influence the redistribution of some of its routes by its peers. Three types of redistribution communities have been proposed. The first type may be used to indicate that a specific route should not be announced over the specified set of eBGP sessions. The second type may be used to indicate that the attached route should only be announced with the NO_EXPORT community over the specified set of eBGP sessions and the third type may be used to indicate that the attached route should be prepended n times when announced over the specified set of eBGP sessions.

Acknowledgements

This work was partially funded by the European Commission, within the ATRIUM IST project. We would like to thank Bart Peirens, Alvaro Retana and Andrew Partan for their comments.

References

[Halabi97] B. Halabi. Internet Routing Architectures. Cisco Press, 1997.

[Huston01] G. Huston. AS1221 BGP table statistics. available from http://www.telstra.net/ops/bgp/, 2001.

[COM-SUR] B. Quoitin, O. Bonaventure, A survey of the utilization of the BGP community attribute, Internet draft, draft-quoitin-bgp-comm-survey-00.txt, work in progress, February 2002

[Quoitin02] B. Quoitin, An implementation of the BGP redistribution communities in zebra, Technical report Infonet-TR-2002-03, February 2002, http://www.infonet.fundp.ac.be/doc/tr/Infonet-TR-2002-03.html

[EXT-COM] S. Sangli, D. Tappan, and Y. Rekhter. BGP extended com munities attribute. Internet draft, draft-ietf-idr-bgp-ext-communi ties-05.txt, work in progress, May 2002.

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

Authors' Addresses

Olivier Bonaventure,
Dept. Computing Science and Engineering
Universite catholique de Louvain (UCL)
Place Sainte-Barbe, 2, B-1348 Louvain-la-Neuve (Belgium)
Email: Olivier.Bonaventure@info.fundp.ac.be
URL : http://www.info.ucl.ac.be/people/OBO

Bruno Quoitin
Infonet group (FUNDP)
Rue Grandgagnage 21
B-5000 Namur
Email: Bruno.Quoitin@info.fundp.ac.be
URL : http://www.infonet.fundp.ac.be

Stefaan De Cnodder
Alcatel
Francis Wellesplein 1
B-2018 Antwerp, Belgium
Email: stefaan.de_cnodder@alcatel.be

Jeffrey Haas
NextHop Technologies
Email: jhaas@nexthop.com

Russ White
Cisco Systems
Email: ruwhite@cisco.com

Full Copyright Statement

Appendix 1 Simple example

The redistribution communities defined in this document can be used in two different ways. A first possible solution would be to rely on the existing support for the extended communities in BGP implementations and to manually configure the redistribution communities defined in this document. This solution could be used today by ISPs to support the redistribution communities (or a subset of those communities) defined in this document instead on defining special community values in their community space and advertising them in the routing registries.

To illustrate a possible configuration with an existing BGP implementation supporting the extended communities, we use a syntax similar to the syntax used by zebra. Let us assume that one route from AS3 has two peerings: one peering with AS2 and one peering with AS1. The configuration below shows how AS3's router could be configured to support the redistribution communities defined in this document. In the configuration in figure A, we show each extended community in hex format for readability reasons and only consider a subset of the redistribution communities. Figure A shows how AS3 would configure its routers to allow to request that a route announced to AS1 would be prepended n times before being announced and to request that a specific route would not be announced to AS2.

```
router bgp 3
   neighbor 172.17.1.1 remote-as 1
   neighbor 172.17.1.1 route-map prepend1_as1 out
   neighbor 172.17.1.2 remote-as 2
   neighbor 172.17.1.2 route-map do_not_announce_as2 out
! Extended community list
! -------------------------
!   action "prepend x times"
!   filter "include AS1"
!
ip extcommunity-list 1 permit 0x4401810000000001
ip extcommunity-list 2 permit 0x4402810000000001
ip extcommunity-list 3 permit 0x4403810000000001
ip extcommunity-list 4 permit 0x4404810000000001
!
! Route-maps
! -------------------------
!   action "prepend x times"
!   filter "include AS1"
!
route-map prepend_as1 permit 10
  match extcommunity 1
```

```
  set as-path prepend 1
!
route-map prepend_as1 permit 20
  match extcommunity 2
  set as-path prepend 2
!
route-map prepend_as1 permit 30
  match extcommunity 3
  set as-path prepend 3
!
route-map prepend_as1 permit 40
  match extcommunity 4
  set as-path prepend 4
!
! Extended community list
! -------------------------
!   action "do not announce"
!   filter "include AS2"
!
ip extcommunity-list 5 permit 0x4410810000000002
!
route-map do_not_announce_as2 deny 10
  match extcommunity 5
!
```

Figure A: Sample configuration

For a router with a small number of peers, such a manual configura tion of the redistribution communities is possible. However, if the routers has many peers, the required configuration file may become very large, especially if one wants to fully support all the redis tribution communities defined in this document. In this case, a bet ter solution is to rely on a direct support for the redistribution communities inside the BGP implementation itself as discussed in [Quoitin02]. With a BGP implementation supporting directly the redis tribution communities, a few lines of configuration will be suffi cient to enable or disable some or all of the redistribution communi ties for each peer.