Internet Engineering Task Force
INTERNET DRAFT

Olivier Bonaventure
FUNDP
Stefaan De Cnodder
Alcatel
Jeffrey Haas
NextHop
Bruno Quoitin
FUNDP
Russ White
Cisco
February, 2002
Expires August, 2002

Controlling the redistribution of BGP routes
<draft-bonaventure-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 to the specified set of eBGP speakers. The second type may be used to indicate that the attached route should only be announced with the NO_EXPORT community to the specified set of eBGP

speakers and the third type may be used to indicate that the attached route should be prepended n times when announced to the specified set of eBGP speakers.

1 Introduction

In today's commercial Internet, many ISPs need to have some control on their interdomain 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 assymetry of Internet traffic, most ISPs also need to control their incoming traffic.

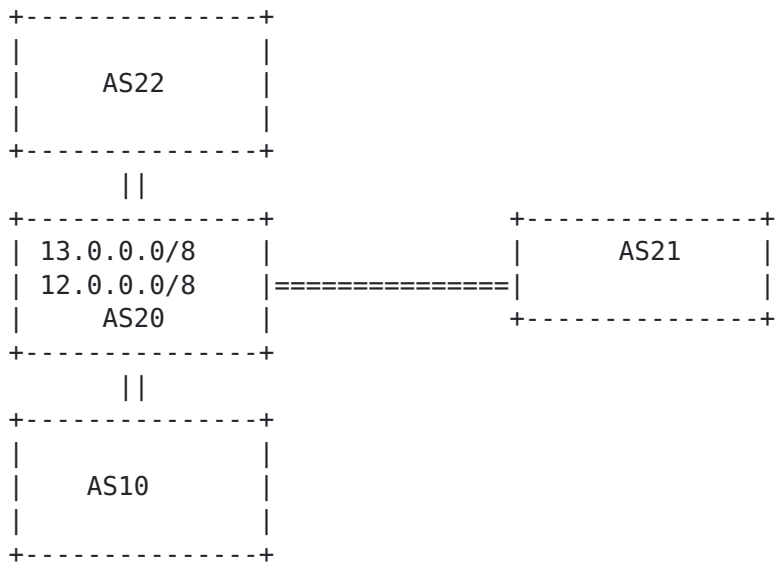

Figure 1: Simple interdomain topology

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 [Hal97,QuB02]. 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 rely only on the AS path length to select the best BGP route, they will prefer the shorter route received by AS22. This requires a manual

configuration of the BGP routers, but path prepending is very frequently used today on the Internet [Hus01]. 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 four times AS20 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 would will then have 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 [QuB02]. This survey reveals the following :

- Many different AS 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 AS within a given geographical area (e.g. a European AS could want to prohibit a route from being announced to US peers).
- Many AS 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 AS.
- Several AS 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 AS have difficulties of providing all these facilities while still relying on their assigned set of BGP community values. For example, some AS have chosen to reuse several BGP community values corresponding to 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 and 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 or in peering agreements.

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 attribute called the redistribution community. If a redistribution policy applies to a long list of BGP speakers, then it will be encoded in several redistribution communities.

2.1 The redistribution community

The extended communities attribute is defined in [RTR01]. This attribute allows a BGP router to attach a set of extended communities to an UPDATE message. Each extended community value is encoded

as an eight octets quantity with a two octets type field and a 6 octets value field. Several types of extended community values are defined in [RTR01]. 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 [RTR01]. The high-order bit is cleared (type assigned by IANA). Since the redistribution community is used for signalling purposes between two AS's, the bit6 is set meaning that the extended community is non-transitive across ASes. This is important to ensure that communities used to affect the redistribution of routes by a given AS are not unnecessirally distributed over the entire Internet as it is often the case today [QuB02]. 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 peers. 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 to be assigned by IANA.

- 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.

- 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 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.

The BGP Speakers Filter

The BGP_Speakers_Filter field is used to specify the eBGP 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 eBGP speakers that will be affected by the specified action. There are two methods to specify the affected eBGP speakers. The first method is to explicitly list all those speakers inside the BGP_Speakers_Filters field of 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 eBGP speakers 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 to be assigned by IANA. Values 0x40-0x7f are vendor specific.

BGP_Speakers_Filter types

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

- The BGP_Speakers_Filter value contains two two octets 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.

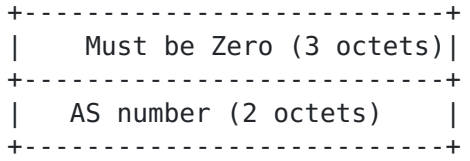

Figure 3 : BGP speakers filter containing a single two octets AS number

If the BGP_Speakers_Filter value contains two two octets 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.

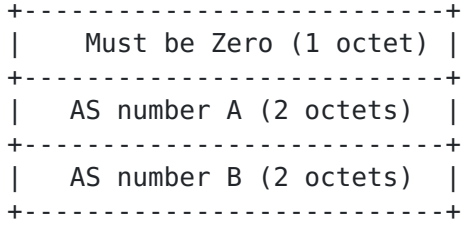

Figure 4 : BGP speakers filter containing two distinct two octets AS number

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.

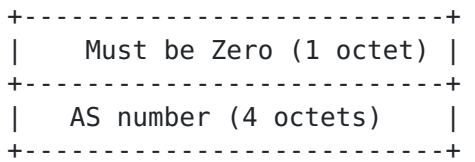

Figure 5 : BGP speakers filter containing a single four octets AS number

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

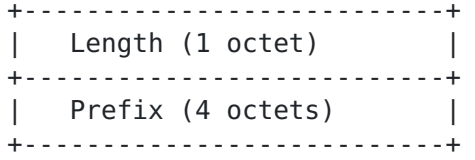

Figure 6 : BGP speakers filter containing a CIDR prefix/length pair

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 any number of redistribution communities to a route originated by itself or received from another BGP speaker with iBGP or eBGP. 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 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 redistribution communities concerning this action type and parameter.

In practice, it can be expected that only the originator of the route will attach the redistribution communities 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.

2.3 Operations

The redistribution communities defined in this document only affect the redistribution of the associated route to eBGP peers. The redistribution communities do not affect the redistribution of routes via iBGP or between the sub-ASs of a confederation.

When a router receives a route with redistribution communities, it should apply the operations specified by these communities when redistributing the route to eBGP peers. Since the redistribution communities defined by this document are non-transitive, a router will remove the received redistribution communities when redistributing the route to eBGP peers. 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 parameter contained in the redistribution communities attached to the route a list of the target BGP speakers contained in the BGP_Speakers_filters for this action type. In the remainder of this section, we will use the wordings "a BGP speaker P is affected by action type x with parameter" to indicate that either of the following is true :

- P 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

- P 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

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

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

- If peer P is affected by one or more actions of type "Prepend", then the AS-Path of the route shall be prepended n times where n is the smallest parameter of the matched "Prepend" actions.

Then the route is announced to peer P.

3 IANA considerations

This document requests the attribution of a new BGP extended communities type field from IANA. In addition, this document proposes that IANA maintains the action types and the BGP speakers filter types values defined in section 2.

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 communinities called the redistribution communities. These redistribution communities 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 to the specified set of eBGP speakers. The second type may be used to indicate that the attached route should only be announced with the NO_EXPORT community to the specified set of eBGP speakers and the third type may be used to indicate that the attached route should be prepended n times when announced to the specified set of eBGP speakers.

Acknowledgements

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

References

[Hal97] B. Halabi. Internet Routing Architectures. Cisco Press, 1997.

[Hus01] G. Huston. AS1221 BGP table statistics. available from http://www.telstra.net/ops/bgp/, 2001.

[QuB02] 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

[Quo02] B. Quoitin, An implementation of the BGP redistribution communities in zebra, Technical report Infonet-TR-2002-03, February 2002, to appear

[RTR01] S. Sangli, D. Tappan, and Y. Rekhter. BGP extended communities attribute. Internet draft,draft-ietf-idr-bgp-ext-communities-01.txt, work in progress, August 2001.

Authors' Addresses

Olivier Bonaventure, Bruno Quoitin
Infonet group (FUNDP)
Rue Grandgagnage 21, B-5000 Namur, Belgium
Email: Olivier.Bonaventure@info.fundp.ac.be, 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

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 configuration 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 redistribution communities defined in this document. In this case, a better solution is to rely on a direct support for the redistribution communities inside the BGP implementation itself as discussed in [Quo02]. With a BGP implementation supporting directly the redistribution communities, a few lines of configuration will be sufficient to enable or disable some or all of the redistribution communites for each peer.