

Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: August 19, 2014

Yiqun Cai  
Microsoft  
Sri Vallepalli  
Heidi Ou  
Cisco Systems, Inc.  
Andy Green  
British Telecom  
February 15, 2014

PIM Designated Router Load Balancing  
draft-ietf-pim-drlb-03.txt

Abstract

On a multi-access network, one of the PIM routers is elected as a Designated Router (DR). On the last hop network, the PIM DR is responsible for tracking local multicast listeners and forwarding traffic to these listeners if the group is operated in PIM SM. In this document, we propose a modification to the PIM SM protocol that allows more than one of these last hop routers to be selected so that the forwarding load can be distributed to and handled among these routers.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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."

This Internet-Draft will expire on August 19, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

## Table of Contents

1. Terminology . . . . .	3
2. Introduction . . . . .	3
3. Applicability . . . . .	6
4. Functional Overview . . . . .	6
4.1. GDR Candidates . . . . .	7
4.2. Hash Mask . . . . .	7
4.3. PIM Hello Options . . . . .	8
5. Hello Option Formats . . . . .	9
5.1. PIM DR Load Balancing Capability (DRLBC) Hello Option . . . . .	9
5.2. PIM DR Load Balancing GDR (DRLBGDR) Hello Option . . . . .	10
6. Protocol Specification . . . . .	11
6.1. PIM DR Operation . . . . .	11
6.2. PIM GDR Candidate Operation . . . . .	11
6.3. PIM Assert Modification . . . . .	12
7. IANA Considerations . . . . .	14
8. Security Considerations . . . . .	14
9. Acknowledgement . . . . .	14
10. References . . . . .	14
10.1. Normative Reference . . . . .	14
10.2. Informative References . . . . .	14
Authors' Addresses . . . . .	15

## 1. Terminology

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 [RFC2119].

With respect to PIM, this document follows the terminology that has been defined in [RFC4601].

This document also introduces the following new acronyms:

- o GDR: GDR stands for "Group Designated Router". For each multicast group, a hash algorithm (described below) is used to select one of the routers as a GDR. The GDR is responsible for initiating the forwarding tree building for the corresponding group.
- o GDR Candidate: a last hop router that has potential to become a GDR. A GDR Candidate must have the same DR priority and must run the same GDR election hash algorithm as the DR router. It must send and process received new PIM Hello Options as defined in this document. There might be more than one GDR Candidate on a LAN. But only one can become GDR for a specific multicast group.

## 2. Introduction

On a multi-access network such as an Ethernet, one of the PIM routers is elected as a DR. The PIM DR has two roles in the PIM protocol. On the first hop network, the PIM DR is responsible for registering an active source with the Rendezvous Point (RP) if the group is operated in PIM SM. On the last hop network, the PIM DR is responsible for tracking local multicast listeners and forwarding to these listeners if the group is operated in PIM SM.

Consider the following last hop network in Figure 1:

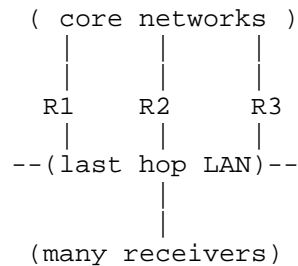


Figure 1: Last Hop Network

Assume R1 is elected as the Designated Router. According to [RFC4601], R1 will be responsible for forwarding to the last hop LAN. In addition to keeping track of IGMP and MLD membership reports, R1 is also responsible for initiating the creation of source and/or shared trees towards the senders or the RPs.

Forcing sole data plane forwarding responsibility on the PIM DR proves a limitation in the protocol. In comparison, even though an OSPF DR, or an IS-IS DIS, handles additional duties while running the OSPF or IS-IS protocols, they are not required to be solely responsible for forwarding packets for the network. On the other hand, on a last hop LAN, only the PIM DR is asked to forward packets while the other routers handle only control traffic (and perhaps drop packets due to RPF failures). The forwarding load of a last hop LAN is concentrated on a single router.

This leads to several issues. One of the issues is that the aggregated bandwidth will be limited to what R1 can handle towards this particular interface. These days, it is very common that the last hop LAN usually consists of switches that run IGMP/MLD or PIM snooping. This allows the forwarding of multicast packets to be restricted only to segments leading to receivers who have indicated their interest in multicast groups using either IGMP or MLD. The emergence of the switched Ethernet allows the aggregated bandwidth to exceed, some times by a large number, that of a single link. For example, let us modify Figure 1 and introduce an Ethernet switch in Figure 2.

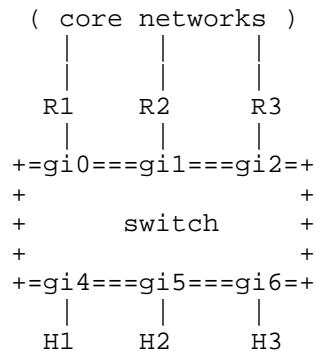


Figure 2: Last Hop Network with Ethernet Switch

Let us assume that each individual link is a Gigabit Ethernet. Each router, R1, R2 and R3, and the switch have enough forwarding capacity to handle hundreds of Gigabits of data.

Let us further assume that each of the hosts requests 500 mbps of data and different traffic is requested by each host. This represents a total 1.5 gbps of data, which is under what each switch or the combined uplink bandwidth across the routers can handle, even under failure of a single router.

On the other hand, the link between R1 and switch, via port gi0, can only handle a throughput of 1gbps. And if R1 is the only router, the PIM DR elected using the procedure defined by RFC 4601, at least 500 mbps worth of data will be lost because the only link that can be used to draw the traffic from the routers to the switch is via gi0. In other words, the entire network's throughput is limited by the single connection between the PIM DR and the switch (or the last hop LAN as in Figure 1).

The problem may also manifest itself in a different way. For example, R1 happens to forward 500 mbps worth of unicast data to H1, and at the same time, H2 and H3 each requests 300 mbps of different multicast data. Once again packet drop happens on R1 while in the mean time, there is sufficient forwarding capacity left on R2 and R3 and link capacity between the switch and R2/R3.

Another important issue is related to failover. If R1 is the only forwarder on the last hop network, in the event of a failure when R1 goes out of service, multicast forwarding for the entire network has to be rebuilt by the newly elected PIM DR. However, if there was a way that allowed multiple routers to forward to the network for different groups, failure of one of the routers would only lead to

disruption to a subset of the flows, therefore improving the overall resilience of the network.

In this document, we propose a modification to the PIM protocol that allows more than one of these routers, called Group Designated Router (GDR) to be selected so that the forwarding load can be distributed to and handled by a number of routers.

### 3. Applicability

The proposed change described in this specification applies to PIM SM last hop routers only.

It does not alter the behavior of a PIM DR on the first hop network. This is because the source tree is built using the IP address of the sender, not the IP address of the PIM DR that sends the registers towards the RP. The load balancing between first hop routers can be achieved naturally if an IGP provides equal cost multiple paths (which it usually does in practice). And distributing the load to do registering does not justify the additional complexity required to support it.

### 4. Functional Overview

In the existing PIM DR election, when multiple last hop routers are connected to a multi-access network (for example, an Ethernet), one of them is selected to act as PIM DR. The PIM DR is responsible for sending Join/Prune messages towards the RP or source. To elect the PIM DR, each PIM router on the network examines the received PIM Hello messages and compares its DR priority and IP address with those of its neighbors. The router with the highest DR priority is the PIM DR. If there are multiple such routers, their IP addresses are used as the tie-breaker, as described in [RFC4601].

In order to share forwarding load among last hop routers, besides the normal PIM DR election, the GDR is also elected on the last hop multi-access network. There is only one PIM DR on the multi-access network, but there might be multiple GDR Candidates.

For each multicast group, a hash algorithm is used to select one of the routers to be the GDR. Hash Masks are defined for Source, Group and RP separately, in order to handle PIM ASM/SSM. The masks are announced in PIM Hello by DR as a DR Load Balancing GDR (DRLBGDR) Hello Option. Besides that, a DR Load Balancing Capability (DRLBC) Hello Option, which contains hash algorithm type, is also announced by router interfaces which have this specification supported. Last

hop routers who are with the new DRLBC Option, and with the same GDR election hash algorithm and the same DR priority as the PIM DR are GDR Candidates.

A hash algorithm based on the announced Source, Group or RP masks allows one GDR to be assigned to a corresponding multicast group, and that GDR is responsible for initiating the creation of the multicast forwarding tree for the group.

#### 4.1. GDR Candidates

GDR is the new concept introduced by this specification. GDR Candidates are routers eligible for GDR election on the LAN. To become a GDR Candidate, a router MUST support this specification, have the same DR priority and run the same GDR election hash algorithm as the DR on the LAN.

For example, assume there are 4 routers on the LAN: R1, R2, R3 and R4, which all support this specification on the LAN. R1, R2 and R3 have the same DR priority while R4's DR priority is less preferred. In this example, R4 will not be eligible for GDR election, because R4 will not become a PIM DR unless all of R1, R2 and R3 go out of service.

Further assume router R1 wins the PIM DR election, and R1, R2 run the same hash algorithm for GDR election, while R3 runs a different one. Then only R1 and R2 will be eligible for GDR election, R3 will not.

As a DR, R1 will include its own Load Balancing Hash Masks, and also the identity of R1 and R2 (the GDR Candidates) in its DRLBGDR Hello Option.

#### 4.2. Hash Mask

A Hash Mask is used to extract a number of bits from the corresponding IP address field (32 for v4, 128 for v6), and calculate a hash value. A hash value is used to select a GDR from GDR Candidates advertised by PIM DR. For example, 0.255.0.0 defines a Hash Mask for an IPv4 address that masks the first, the third and the fourth octets.

There are three Hash Masks defined,

- o RP Hash Mask
- o Source Hash Mask
- o Group Hash Mask

The Hash Masks MUST be configured on the PIM routers that can

potentially become a PIM DR.

A simple Modulo hash algorithm will be discussed in this document. However, to allow other hash algorithm to be used, a 4-bytes "Hash Algorithm Type" field is included in DRLBC Hello Option to specify the hash algorithm used by a last hop router.

If different hash algorithm types are advertised among last hop routers, only last hop routers running the same hash algorithm as the DR (and having the same DR priority as the DR) are eligible for GDR election.

For ASM groups, a hash value is calculated using the following Modulo algorithm:

o  $\text{hashvalue\_RP} = (((\text{RP\_address} \& \text{RP\_hashmask}) \gg N) \& 0\text{xFFFF}) \% M$

RP\_address is the address of the RP defined for the group. N is the number of zeros, counted from the least significant bit of the RP\_hashmask. For example, for a given IPv4 RP\_hashmask 0.255.0.0, N will be 16. M is the number of GDR Candidates as described above.

If RP\_hashmask is 0, a hash value is also calculated using the group Hash Mask in a similar fashion.

o  $\text{hashvalue\_Group} = (((\text{Group\_address} \& \text{Group\_hashmask}) \gg N) \& 0\text{xFFFF}) \% M$

For SSM groups, a hash value is calculated using both the source and group Hash Mask

o  $\text{hashvalue\_SG} = (((\text{Source\_address} \& \text{Source\_hashmask}) \gg N_S) \& 0\text{xFFFF}) \wedge (((\text{Group\_address} \& \text{Group\_hashmask}) \gg N_G) \& 0\text{xFFFF}) \% M$

#### 4.3. PIM Hello Options

When a last hop PIM router sends a PIM Hello from an interface with this specification support, it includes a new option, called "Load Balancing Capability (DRLBC)".

Besides this DRLBC Hello Option, the elected PIM DR also includes a new "DR Load Balancing GDR (DRLBGDR) Hello Option". The DRLBGDR Hello Option consists of three Hash Masks as defined above and also the sorted addresses of all GDR Candidates on the last hop network.

The elected PIM DR uses DRLBC Hello Option advertised by all routers on the last hop network to compose its DRLBGDR. The GDR Candidates



use DRLBGDR Hello Option advertised by PIM DR to calculate hash value.

## 5. Hello Option Formats

### 5.1. PIM DR Load Balancing Capability (DRLBC) Hello Option

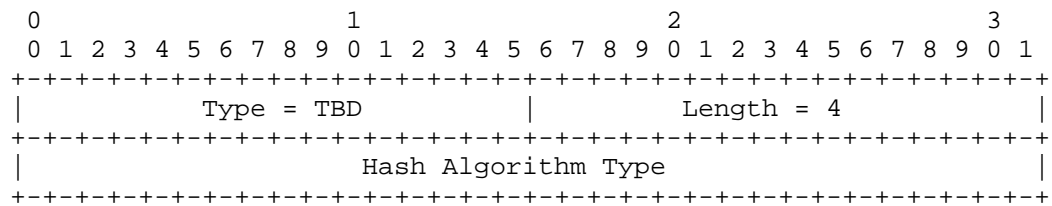


Figure 3: Capability Hello Option

Type: TBD.

Length: 4 octets

Hash Algorithm Type: 0 for Modulo hash algorithm

This DRLBC Hello Option SHOULD be advertised by last hop routers from interfaces which support this specification.

## 5.2. PIM DR Load Balancing GDR (DRLBGDR) Hello Option

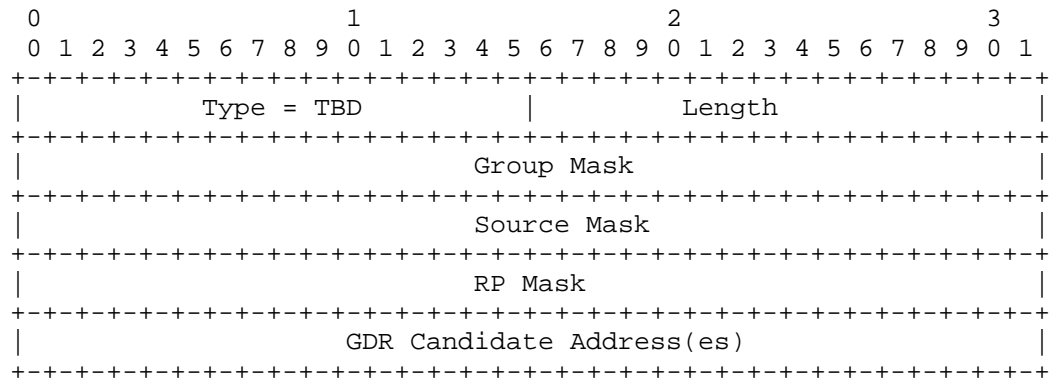


Figure 4: GDR Hello Option

Type: TBD

Length:

Group Mask (32/128 bits): Mask

Source Mask (32/128 bits): Mask

RP Mask (32/128 bits): Mask

All masks MUST be in the same address family, with the same length.

GDR Address (32/128 bits): Address(es) of GDR Candidate(s)

All addresses must be in the same address family. The addresses are sorted from high to low. The order is converted to the ordinal number associated with each GDR candidate in hash value calculation. For example, addresses advertised are R3, R2, R1, the ordinal number assigned to R3 is 0, to R2 is 1 and to R1 is 2. If "Interface ID" option (type 31) presents in a GDR Candidate's PIM Hello message, and the "Router ID" portion is non-zero,

\* For IPv4, the "GDR Candidate Address" will be set directly to "Router ID".

\* For IPv6, the "GDR Candidate Address" will be set to the IPv4-IPv6 translated address of "Router ID", as described in [RFC4291], that is the "Router-ID" is appended to the prefix of 96-bits zeros.

If the "Interface ID" option is not present in a GDR Candidate's PIM Hello message, or if the "Interface ID" option is present, but "Router ID" field is zero, the "GDR Candidate Address" will be the IPv4 or IPv6 source address from PIM Hello message.

This DRLBGDR Hello Option SHOULD only be advertised by the elected PIM DR.

## 6. Protocol Specification

### 6.1. PIM DR Operation

The DR election process is still the same as defined in [RFC4601]. A DR that has this specification enabled on the interface, advertises the new LBGRD Hello Option, which contains value of masks from user configuration, followed by a sorted list of addresses of all GDR Candidates. Moreover, same as non-DR routers, DR also advertises DRLBC Hello Option to indicate its capability of supporting this specification and the type of its GDR election hash algorithm.

If a PIM DR receives a neighbor Hello with DRLBGRD Option, the PIM DR SHOULD ignore the TLV.

If a PIM DR receives a neighbor DRLBC Hello Option, which contains the same hash algorithm type as the DR, and the neighbor has the same DR priority as the DR, PIM DR SHOULD consider the neighbor as a GDR Candidate and insert the neighbor's address into the sorted list of DRLBGRD Option.

### 6.2. PIM GDR Candidate Operation

When an IGMP join is received, without this proposal, router R1 (the PIM DR) will handle the join and potentially run into the issues described earlier. Using this proposal, a hash algorithm is used to determine which router is going to be responsible for building forwarding trees on behalf of the host.

The algorithm works as follows, assuming the router in question is X, which is a GDR Candidate, and its ordinal number assigned implicitly by PIM DR in DRLBGDR Hello Option is Ox:

- o If the group is ASM, and the RP Hash Mask announced by the PIM DR is not zero, calculate the value of hashvalue\_RP. If hashvalue\_RP is equal to Ox, X becomes the GDR.

For example, X with IPv4 address 10.1.1.3, receives a DRLBGDR Hello Option from the DR, which announces RP Hash Mask 0.255.0.0, and a

list of GDR Candidates, sorted by IP addresses from high to low, 10.1.1.3, 10.1.1.2 and 10.1.1.1. The ordinal number assigned to those addresses would be 0 for 10.1.1.3 (X), 1 for 10.1.1.2, and 2 for 10.1.1.1. Assume there are 2 RPs: RP1 172.3.10.10 for Group1 and RP2 172.2.10.10 for Group2. Following the modulo hash algorithm

$$\text{hashvalue\_RP} = (((\text{RP\_address} \& \text{RP\_hashmask}) \gg N) \& 0\text{xFFFF}) \% M$$

Here N is 16 for 0.255.0.0, and M is 3 for the total number of GDR Candidates. The hashvalue\_RP for RP1 172.3.10.10 is 0, matches the ordinal number assigned to X. X will be the GDR for Group1, which uses 172.3.10.10 as the RP. The hashvalue\_RP for RP2 172.2.10.10 is 2, which is different from X's ordinal number, hence, X will not be GDR for Group2.

- o If the group is ASM, and the RP Hash Mask announced by the PIM DR is zero, obtain the value of hashvalue\_Group. Compare hashvalue\_Group with 0x, to decide if X is the GDR.
- o If the group is SSM, then use hashvalue\_SG to determine if X is the GDR.

If X is the GDR for the group, X will be responsible for building the forwarding tree.

A router interface where this protocol is enabled advertises DRLBC Hello Option in its PIM Hello, even if the router may not be a GDR Candidate.

A GDR Candidate may receive a DRLBGDR Hello Option from PIM DR, with different Hash Masks from those configured on it, The GDR Candidate must use the Hash Masks advertised by the PIM DR to calculate the hash value.

A GDR Candidate may receive a DRLBGDR Hello Option from a non-DR PIM router. The GDR Candidate must ignore such DRLBGDR Hello Option.

A GDR Candidate may receive a Hello from the elected PIM DR, and the PIM DR does not support this specification. The GDR election described by this specification will not take place, that is only the PIM DR joins the multicast tree.

### 6.3. PIM Assert Modification

It is possible that the identity of the GDR might change in the middle of an active flow. Examples this could happen include:

1. When a new PIM router comes up

## 2. When a GDR restarts

When the GDR changes, existing traffic might be disrupted. Duplicates or packet loss might be observed. To illustrate the case, consider the following scenario: there are two streams G1 and G2. R1 is the GDR for G1, and R2 is the GDR for G2. When R3 comes up online, it is possible that R3 becomes GDR for both G1 and G2, hence R2 starts to build the forwarding tree for G1 and G2. If R1 and R2 stop forwarding before R3 completes the process, packet loss might occur. On the other hand, if R1 and R2 continue forwarding while R3 is building the forwarding trees, duplicates might occur.

This is not a typical deployment scenario but it still might happen. Here we describe a mechanism to minimize the impact. The motivation is that we want to minimize packet loss. And therefore, we would allow a small amount of duplicates and depend on PIM Assert to minimize the duplication.

When the role of GDR changes as above, instead of immediately stopping forwarding, R1 and R2 continue forwarding to G1 and G2 respectively, while at the same time, R3 build forwarding trees for G1 and G2. This will lead to PIM Asserts.

Due to the introduction of GDR, this document suggests the following modification to the Assert packet: if a router enables this specification on its downstream interface, but it is not a GDR, it would adjust its Assert metric to (PIM\_ASSERT\_INFINITY - 1).

Using the above example, assume R1 and R3 agree on the new GDR, which is R3. R1 will set its Assert metric as (PIM\_ASSERT\_INFINITY - 1). That will make R3, which has normal metric in its Assert as the Assert winner.

For G2, assume it takes a little bit longer time for R2 to find out that R3 is the new GDR and still thinks itself being the GDR while R3 already has assumed the role of GDR. Since both R2 and R3 think they are GDRs, they further compare the metric and IP address. If R3 has the better routing metric, or same metric but better tie-breaker, the result will be consistent with GDR selection. If unfortunately, R2 has the better metric or same metric but better tie-breaker R2 will become the Assert winner and continues to forward traffic. This will continue until:

1. The next PIM Hello option from DR is seen that selects R3 as the GDR.
  2. R3 will build the forwarding tree and send an Assert.
- The process continues until R2 agrees to the selection of R3 as being the GDR, and set its own Assert metric to (PIM\_ASSERT\_INFINITY - 1), which will make R3 the Assert winner. During the process, we will see intermittent duplication of traffic but packet loss will be

minimized. In the unlikely case that R2 never relinquishes its role as GDR (while every other router thinks otherwise), the proposed mechanism also helps to keep the duplication to a minimum until manual intervention takes place to remedy the situation.

## 7. IANA Considerations

Two new PIM Hello Option Types are required to be assigned to the DR Load Balancing messages. [HELLO-OPT], this document recommends 34(0x22) as the new "PIM DR Load Balancing Capability Hello Option", and 35(0x23) as the new "PIM DR Load Balancing GDR Hello Option".

## 8. Security Considerations

Security of the PIM DR Load Balancing Hello message is only guaranteed by the security of PIM Hello message, so the security considerations for PIM Hello messages as described in PIM-SM [RFC4601] apply here.

## 9. Acknowledgement

The authors would like to thank Steve Simlo, Taki Millonis for helping with the original idea, Bill Atwood for review comments, Stig Venaas, Toerless Eckert and Rishabh Parekh for helpful conversation on the document.

## 10. References

### 10.1. Normative Reference

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

### 10.2. Informative References

- [RFC3973] Adams, A., Nicholas, J., and W. Siadak, "Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised)", RFC 3973, January 2005.
- [RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano,

"Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

[RFC6395] Gulrajani, S. and S. Venaas, "An Interface Identifier (ID) Hello Option for PIM", RFC 6395, October 2011.

[RFC4291] Hinden, R. and L. S., "IP Version 6 Addressing Architecture", RFC 6890, February 2006.

[HELLO-OPT]  
IANA, "PIM Hello Options", PIM-HELLO-OPTIONS per RFC4601 <http://www.iana.org/assignments/pim-hello-options>, March 2007.

#### Authors' Addresses

Yiqun Cai  
Microsoft  
La Avenida  
Mountain View, CA 94043  
USA

Email: [yiqunc@microsoft.com](mailto:yiqunc@microsoft.com)

Sri Vallepalli  
Cisco Systems, Inc.  
Tasman Drive  
San Jose, CA 95134  
USA

Email: [svallepa@cisco.com](mailto:svallepa@cisco.com)

Heidi Ou  
Cisco Systems, Inc.  
Tasman Drive  
San Jose, CA 95134  
USA

Email: [hou@cisco.com](mailto:hou@cisco.com)

Andy Green  
British Telecom  
Adastral Park  
Ipswich IP5 2RE  
United Kingdom

Email: andy.da.green@bt.com





L3VPN Working Group  
Internet Draft  
Intended Status: Standards Track  
Expires: August 14, 2014

Jeffrey Zhang  
Lenny Giuliano  
Juniper Networks, Inc.

Eric C. Rosen  
Karthik Subramanian  
Cisco Systems, Inc.

Dante J. Pacella  
Verizon

Jason Schiller  
Google

February 14, 2014

## Global Table Multicast with BGP-MVPN Procedures

draft-zzhang-l3vpn-mvpn-global-table-mcast-03.txt

### Abstract

RFC6513, RFC6514, and other RFCs describe protocols and procedures which a Service Provider (SP) may deploy in order offer Multicast Virtual Private Network (Multicast VPN or MVPN) service to its customers. Some of these procedures use BGP to distribute VPN-specific multicast routing information across a backbone network. With a small number of relatively minor modifications, the very same BGP procedures can also be used to distribute multicast routing information that is not specific to any VPN. Multicast that is outside the context of a VPN is known as "Global Table Multicast", or sometimes simply as "Internet multicast". In this document, we describe the modifications that are needed to use the MVPN BGP procedures for Global Table Multicast.

### Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

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/lid-abstracts.txt>.

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

#### Copyright and License Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction .....	4
2	Adapting MVPN Procedures to GTM .....	6
2.1	Use of Route Distinguishers .....	7
2.2	Use of Route Targets .....	7
2.3	UMH-eligible Routes .....	9
2.3.1	Routes of SAFI 1, 2 or 4 with MVPN ECs .....	10
2.3.2	MVPN ECs on the Route to the Next Hop .....	11
2.3.3	Non-BGP Routes as the UMH-eligible Routes .....	12
2.3.4	Why SFS Does Not Apply to GTM .....	13
2.4	Inclusive and Selective Tunnels .....	14
2.5	I-PMSI A-D Routes .....	14
2.5.1	Intra-AS I-PMSI A-D Routes .....	14
2.5.2	Inter-AS I-PMSI A-D Routes .....	15
2.6	S-PMSI A-D Routes .....	15
2.7	Leaf A-D Routes .....	15
2.8	Source Active A-D Routes .....	15
2.9	C-multicast Source/Shared Tree Joins .....	16
3	Differences from other MVPN-like GTM Procedures .....	17
4	IANA Considerations .....	18
5	Security Considerations .....	18
6	Additional Contributors .....	19
7	Acknowledgments .....	19
8	Authors' Addresses .....	20
9	References .....	21
9.1	Normative References .....	21
9.2	Informative References .....	21

## 1. Introduction

[RFC4364] specifies architecture, protocols, and procedures that a Service Provider (SP) can use to provide Virtual Private Network (VPN) service to its customers. In that architecture, one or more Customer Edge (CE) routers attach to a Provider Edge (PE) router. Each CE router belongs to a single VPN, but CE routers from several VPNs may attach to the same PE router. In addition, CEs from the same VPN may attach to different PEs. BGP is used to carry VPN-specific information among the PEs. Each PE router maintains a separate Virtual Routing and Forwarding table (VRF) for each VPN to which it is attached.

[RFC6513] and [RFC6514] extend the procedures of [RFC4364] to allow the SP to provide multicast service to its VPN customers. The customer's multicast routing protocol (e.g., PIM) is used to exchange multicast routing information between a CE and a PE. The PE stores a given customer's multicast routing information in the VRF for that customer's VPN. BGP is used to distribute certain multicast-related control information among the PEs that attach to a given VPN, and BGP may also be used to exchange the customer multicast routing information itself among the PEs.

While this multicast architecture was originally developed for VPNs, it can also be used (with a small number of modifications to the procedures) to distribute multicast routing information that is not specific to VPNs. The purpose of this document is to specify the way in which BGP MVPN procedures can be adapted to support non-VPN multicast.

Multicast routing information that is not specific to VPNs is stored in a router's "global table", rather than in a VRF; hence it is known as "Global Table Multicast" (GTM). GTM is sometimes more simply called "Internet multicast". However, we will avoid that term because it suggests that the multicast data streams are available on the "public" Internet. The procedures for GTM can certainly be used to support multicast on the public Internet, but they can also be used to support multicast streams that are not public, e.g., content distribution streams offered by content providers to paid subscribers. For the purposes of this document, all that matters is that the multicast routing information is maintained in a global table rather than in a VRF.

This architecture does assume that the network over which the multicast streams travel can be divided into a "core network" and one or more non-core parts of the network, which we shall call "attachment networks". The multicast routing protocol used in the attachment networks may not be the same as the one used in the core,

so we consider there to be a "protocol boundary" between the core network and the attachment networks. We will use the term "Protocol Boundary Router" (PBR) to refer to the core routers that are at the boundary. We will use the term "Attachment Router" (AR) to refer to the routers that are not in the core but that attach to the PBRs.

This document does not make any particular set of assumptions about the protocols that the ARs and the PBRs use to exchange unicast and multicast routing information with each other. For instance, multicast routing information could be exchanged between an AR and a PBR via PIM, IGMP, or even BGP. Multicast routing also depends on an exchange of routes that are used for looking up the path to the root of a multicast tree. This routing information could be exchanged between an AR and a PBR via IGP, via EBGP, or via IBGP ([RFC6368]). Note that if IBGP is used, the [RFC6368] "push/pop procedures" are not necessary.

The PBRs are not necessarily "edge" routers, in the sense of [RFC4364]. For example, they may be both be Autonomous System Border Routers (ASBR). As another example, an AR may be an "access router" attached to a PBR that is an OSPF Area Border Router (ABR). Many other deployment scenarios are possible. However, the PBRs are always considered to be delimiting a "backbone" or "core" network. A multicast data stream from an AR is tunneled over the core network from an Ingress PBR to one or more Egress PBRs. Multicast routing information that a PBR learns from the ARs attached to it is stored in the PBR's global table. The PBRs use BGP to distribute multicast routing and auto-discovery information among themselves. This is done following the procedures of [RFC6513], [RFC6514], and other MVPN specifications, as modified in this document.

In general, PBRs follow the same MVPN/BGP procedures that PE routers follow, except that these procedures are adapted to be applicable to the global table rather than to a VRF. Details are provided in subsequent sections of this document.

By supporting GTM using the BGP procedures designed for MVPN, one obtains a single control plane that governs the use of both VPN and non-VPN multicast. Most of the features and characteristics of MVPN carry over automatically to GTM. These include scaling, aggregation, flexible choice of tunnel technology in the SP network, support for both segmented and non-segmented tunnels, ability to use wildcards to identify sets of multicast flows, support for the Any Source Multicast (ASM), Single Source Multicast (SSM), and Bidirectional (bidir) multicast paradigms, support for both IPv4 and IPv6 multicast flows over either an IPv4 or IPv6 SP infrastructure, support for unsolicited flooded data (including support for BSR as RP-to-group mapping protocols), etc.

This document not only uses MVPN procedures for GTM, but also, insofar as possible, uses the same protocol elements, encodings, and formats. The BGP Updates for GTM thus use the same Subsequent Address Family Identifier (SAFI), and have the same Network Layer Reachability Information (NLRI) format, as the BGP Updates for MVPN.

Details for supporting MVPN (either IPv4 or IPv6 MVPN traffic) over an IPv6 backbone network can be found in [RFC6515]. The procedures and encodings described therein are also applicable to GTM.

The document [SEAMLESS-MCAST] extends [RFC6514] by providing procedures that allow tunnels through the core to be "segmented" at ABRs within the core. The ABR segmentation procedures are also applicable to GTM as defined in the current document. In general, the MVPN procedures of [SEAMLESS-MCAST], adapted as specified in the current document, are applicable to GTM.

The document [SEAMLESS-MCAST] also defines a set of procedures for GTM. Those procedures are different from the procedures defined in the current document, and the two sets of procedures are not interoperable with each other. The two sets of procedures can co-exist in the same network, as long as they are not applied to the same multicast flows or to the same multicast group addresses. See section 3 for more details.

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 [RFC2119].

## 2. Adapting MVPN Procedures to GTM

In general, PBRs support Global Table Multicast by using the procedures that PE routers use to support VPN multicast. For GTM, where [RFC6513] and [RFC6514] talk about the "PE-CE interface", one should interpret that to mean the interface between the AR and the PBR. For GTM, where [RFC6513] and [RFC6514] talk about the "backbone" network, one should interpret that to mean the part of the network that is delimited by the PBRs.

A few adaptations to the procedures of [RFC6513] and [RFC6514] need to be made. Those adaptations are described in the following subsections.

## 2.1. Use of Route Distinguishers

The MVPN procedures require the use of BGP routes, defined in [RFC6514], that have a SAFI value of 5 ("MCAST-VPN"). We refer to these simply as "MCAST-VPN routes". [RFC6514] defines the Network Layer Reachability Information (NLRI) format for MCAST-VPN routes. The NLRI field always begins with a "Route Type" octet, and, depending on the route type, may be followed by a "Route Distinguisher" (RD) field.

When a PBR originates an MCAST-VPN route in support of GTM, the RD field (for those routes types where it is defined) of that route's NLRI MUST be set to zero (i.e., to 64 bits of zero). Since no VRF may have an RD of zero, this allows "MCAST-VPN" routes that are "about" GTM to be distinguished from MCAST-VPN routes that are about VPNs.

## 2.2. Use of Route Targets

The MVPN procedures require all MCAST-VPN routes to carry Route Targets (RTs). When a PE router receives an MCAST-VPN route, it processes the route in the context of a particular VRF if and only if the route is carrying an RT that is configured as one of that VRF's "import RTs".

There are two different "kinds" of RT used in MVPN.

- One kind of RT is carried only by the following MCAST-VPN route types: C-multicast Shared Tree Joins, C-multicast Source Tree Joins, and Leaf A-D routes. This kind of RT identifies the PE router that has been selected by the route's originator as the "Upstream PE" or as the "Upstream Multicast Hop" (UMH) for a particular (set of) multicast flow(s). Per [RFC6514] and [RFC6515], this RT must be an IPv4-address-specific or IPv6-address-specific Extended Community (EC), whose "Global Administrator" field identifies the Upstream PE or the UMH. If the Global Administrator field identifies the Upstream PE, the "Local Administrator" field identifies a particular VRF in that PE.

The GTM procedures of this document require the use of this type of RT, in exactly the same situations where it is used in the MVPN specification. However, one adaptation is necessary: the "Local Administrator" field of this kind of RT MUST always be set to zero, thus implicitly identifying the global table, rather than identifying a VRF. We will refer to this kind of RT as a "PBR-identifying RT".



- The other kind of RT is the conventional RT first specified in [RFC4364]. It does not necessarily identify a particular router by address, but is used to constrain the distribution of VPN routes, and to ensure that a given VPN route is processed in the context of a given VRF if and only if the route is carrying an RT that has been configured as one of that VRF's "import RTs".

Whereas every VRF must be configured with at least one import RT, there is heretofore no requirement to configure any RTs for the global table of any router. As stated above, this document makes the use of PBR-identifying RTs mandatory for GTM. This document makes the use of non-PBR-identifying RTs OPTIONAL for GTM.

The procedures for the use of RTs in GTM are the following:

- If the global table of a particular PBR is NOT configured with any import RTs, then a received MCAST-VPN route is processed in the context of the global table only if it is carrying no RTs, or if it is carrying a PBR-identifying RT whose Global Administrator field identifies that PBR.
- The global table in each PBR MAY be configured with (a) a set of export RTs to be attached to MCAST-VPN routes that are originated to support GTM, and (b) with a set of import RTs for GTM.

If the global table of a given PBR has been so configured, the PBR will process a received MCAST-VPN route in the context of the global table if and only if the route carries an RT that is one of the global table's import RTs, or if the route carries a PBR-identifying RT whose global administrator field identifies the PBR.

If the global tables are configured with RTs, care must be taken to ensure that the RTs configured for the global table are distinct from any RTs used in support of MVPN (except in the case where it is actually intended to create an "extranet" [MVPN-extranet] in which some sources are reachable in global table context while others are reachable in VPN context.)

The "RT Constraint" procedures of [RFC4684] MAY be used to constrain the distribution of MCAST-VPN routes (or other routes) that carry RTs that have been configured as import RTs for GTM. (This includes the PBR-identifying RTs.)

In [RFC6513], the UMH-eligible routes (see section 5.1 of [RFC6513], "Eligible Routes for UMH Selection") are generally routes of SAFI 128 (Labeled VPN-IP routes) or 129 (VPN-IP multicast routes), and are required to carry RTs. These RTs determine which VRFs import which

such routes. However, for GTM, when the UMH-eligible routes may be routes of SAFI 1, 2, or 4, the routes are not required to carry RTs. This document does NOT specify any new rules for determine whether a SAFI 1, 2, or 4 route is to be imported into the global table of any PBR.

### 2.3. UMH-eligible Routes

[RFC6513] section 5.1 defines procedures by which a PE router determines the "C-root", the "Upstream Multicast Hop" (UMH), the "Upstream PE", and the "Upstream RD" of a given multicast flow. (In non-VPN multicast documents, the UMH of a multicast flow at a particular router is generally known as the "RPF neighbor" for that flow.) It also defines procedures for determining the "Source AS" of a particular flow. Note that in GTM, the "Upstream PE" is actually the "Upstream PBR".

The definition of the C-root of a flow is the same for GTM as for MVPN.

For MVPN, to determine the UMH, Upstream PE, Upstream RD, and Source AS of a flow, one looks up the C-root of the flow in a particular VRF, and finds the "UMH-eligible" routes (see section 5.1.1 of [RFC6513]) that "match" the C-root. From among these, one is chosen as the "selected UMH route".

For GTM, the C-root is of course looked up in the global table, rather than in a VRF. For MVPN, the UMH-eligible routes are routes of SAFI 128 or 129. For GTM, the UMH-eligible routes are routes of SAFI 1, SAFI 4, or SAFI 2. If the global table has imported routes of SAFI 2, then these are the UMH-eligible routes. Otherwise, routes of SAFI 1 or SAFI 4 are the UMH-eligible routes. For the purpose of UMH determination, if a SAFI 1 route and a SAFI 4 route contain the same IP prefix in their respective NLRI fields, then the two routes are considered by the BGP bestpath selection process to be comparable.

[RFC6513] defines procedures for determining which of the UMH-eligible routes that match a particular C-root is to become the "Selected UMH route". With one exception, these procedures are also applicable to GTM. The one exception is the following. Section 9.1.2 of [RFC6513] defines a particular method of choosing the Upstream PE, known as "Single Forwarder Selection" (SFS). This procedure MUST NOT be used for GTM (see section 2.3.4 for an explanation of why the SFS procedure cannot be applied to GTM).

In GTM, the "Upstream RD" of a multicast flow is always considered to

be zero, and is NOT determined from the Selected UMH route.

The MVPN specifications require that when BGP is used for distributing multicast routing information, the UMH-eligible routes MUST carry the VRF Route Import EC and the Source AS EC. To determine the Upstream PE and Source AS for a particular multicast flow, the Upstream PE and Source AS are determined, respectively, from the VRF Route Import EC and the Source AS EC of the Selected UMH route for that flow. These ECs are generally attached to the UMH-eligible routes by the PEs that originate the routes.

In GTM, there are certain situations in which it is allowable to omit the VRF Route Import EC and/or the Source AS EC from the UMH-eligible routes. The following sub-sections specify the various options for determining the Upstream PBR and the Source AS in GTM.

The procedures in sections 2.3.1 MUST be implemented. The procedures in sections 2.3.2 and 2.3.3 are OPTIONAL to implement. It should be noted that while the optional procedures may be useful in particular deployment scenarios, there is always the potential for interoperability problems when relying on OPTIONAL procedures.

#### 2.3.1. Routes of SAFI 1, 2 or 4 with MVPN ECs

If the UMH-eligible routes have a SAFI of 1, 2 or 4, then they MAY carry the VRF Route Import EC and/or the Source AS EC. If the selected UMH route is a route of SAFI 1, 2 or 4 that carries the VRF Route Import EC, then the Upstream PBR is determined from that EC. Similarly, if the selected UMH route is a route of SAFI 1, 2, or 4 route that carries the Source AS EC, the Source AS is determined from that EC.

When the procedure of this section is used, a PBR that distributes a UMH-eligible route to other PBRs is responsible for ensuring that the VRF Route Import and Source AS ECs are attached to it.

If the selected UMH-eligible route has a SAFI of 1, 2 or 4, but is not carrying a VRF Route Import EC, then the Upstream PBR is determined as specified in section 2.3.2 or 2.3.3 below.

If the selected UMH-eligible route has a SAFI of 1, 2 or 4, but is not carrying a Source AS EC, then the Source AS is considered to be the local AS.

### 2.3.2. MVPN ECs on the Route to the Next Hop

Some service providers may consider it to be undesirable to have the PBRs put the VRF Route Import EC on all the UMH-eligible routes. Or there may be deployment scenarios in which the UMH-eligible routes are not advertised by the PBRs at all. The procedures described in this section provide an alternative that can be used under certain circumstances.

The procedures of this section are OPTIONAL.

In this alternative procedure, each PBR MUST originate a BGP route of SAFI 1, 2 or 4 to itself. This route MUST carry a VRF Route Import EC that identifies the PBR. The address that appears in the Global Administrator field of that EC MUST be the same address that appears in the NLRI and in the Next Hop field of that route. This route MUST also carry a Source AS EC identifying the AS of the PBR.

Whenever the PBR distributes a UMH-eligible route for which it sets itself as next hop, it MUST use this same IP address as the Next Hop of the UMH-eligible route that it used in the route discussed in the prior paragraph.

When the procedure of this section is used, then when a PBR is determining the Selected UMH Route for a given multicast flow, it may find that the Selected UMH Route has no VRF Route Import EC. In this case, the PBR will look up (in the global table) the route to the Next Hop of the Selected UMH route. If the route to the Next Hop has a VRF Route Import EC, that EC will be used to determine the Upstream PBR, just as if the EC had been attached to the Selected UMH Route.

If recursive route resolution is required in order to resolve the next hop, the Upstream PBR will be determined from the first route with a VRF Route Import EC that is encountered during the recursive route resolution process. (The recursive route resolution process itself is not modified by this document.)

The same procedure can be applied to find the Source AS, except that the Source AS EC is used instead of the VRF Route Import EC.

Note that this procedure is only applicable in scenarios where it is known that the Next Hop of the UMH-eligible routes is not be changed by any router that participates in the distribution of those routes; this procedure MUST NOT be used in any scenario where the next hop may be changed between the time one PBR distributes the route and another PBR receives it. The PBRs have no way of determining dynamically whether the procedure is applicable in a particular deployment; this must be made known to the PBRs by provisioning.

Some scenarios in which this procedure can be used are:

- all PBRs are in the same AS, or
- the UMH-eligible routes are distributed among the PBRs by a Route Reflector (that does not change the next hop), or
- the UMH-eligible routes are distributed from one AS to another through ASBRs that do not change the next hop.

If the procedures of this section are used in scenarios where they are not applicable, GTM will not function correctly.

### 2.3.3. Non-BGP Routes as the UMH-eligible Routes

In particular deployment scenarios, there may be specific procedures that can be used, in those particular scenarios, to determine the Upstream PBR for a given multicast flow.

Suppose the PBRs neither put the VRF Route Import EC on the UMH-eligible routes, nor do they distribute BGP routes to themselves. It may still be possible to determine the Upstream PBR for a given multicast flow, using specific knowledge about the deployment.

For example, suppose it is known that all the PBRs are in the same OSPF area. It may be possible to determine the Upstream PBR for a given multicast flow by looking at the link state database to see which router is attached to the flow's C-root.

As another example, suppose it is known that the set of PBRs is fully meshed via Traffic Engineering (TE) tunnels. When a PBR looks up, in its global table, the C-root of a particular multicast flow, it may find that the next hop interface is a particular TE tunnel. If it can determine the identify of the router at the other end of that TE tunnel, it can deduce that that router is the Upstream PBR for that flow.

This is not an exhaustive set of examples. Any procedure that correctly determines the Upstream PBR in a given deployment scenario MAY be used in that scenario.

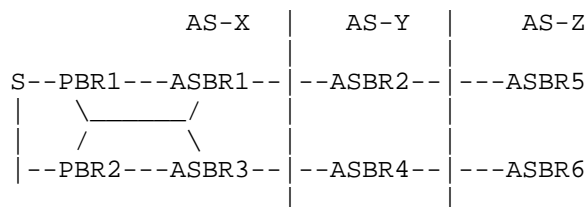
#### 2.3.4. Why SFS Does Not Apply to GTM

To see why the SFS procedure cannot be applied to GTM, consider the following example scenario. Suppose some multicast source S is homed to both PBR1 and PBR2, and suppose that both PBRs export a route (of SAFI 1, 2, or 4) whose NLRI is a prefix matching the address of S. These two routes will be considered comparable by the BGP decision process. A route reflector receiving both routes may thus choose to redistribute just one of the routes to S, the one chosen by the bestpath algorithm. Different route reflectors may even choose different routes to redistribute (i.e., one route reflector may choose the route to S via PBR1 as the bestpath, while another chooses the route to S via PBR2 as the bestpath). As a result, some PBRs may receive only the route to S via PBR1 and some may receive only the route to S via PBR2. In that case, it is impossible to ensure that all PBRs will choose the same route to S.

The SFS procedure works in VPN context as along the following assumption holds: if S is homed to VRF-x in PE1 and to VRF-y in PE2, then VRF-x and VRF-y have been configured with different RDs. In VPN context, the route to S is of SAFI 128 or 129, and thus has an RD in its NLRI. So the route to S via PE1 will not have the same NLRI as the route to S via PE2. As a result, all PEs will see both routes, and the PEs can implement a procedure that ensures that they all pick the same route to S.

That is, the SFS procedure of [RFC6513] relies on the UMH-eligible routes being of SAFI 128 or 129, and relies on certain VRFs being configured with distinct RDs. Thus the procedure cannot be applied to GTM.

One might think that the SFS procedure could be applied to GTM as long as the procedures defined in [ADD-PATH] are applied to the UMH-eligible routes. Using the [ADD-PATH] procedures, the BGP speakers could advertise more than one path to a given prefix. Typically [ADD-PATH] is used to report the n best paths, for some small value of n. However, this is not sufficient to support SFS, as can be seen by examining the following scenario.



In AS-X, PBR1 reports to both ASBR1 and ASBR3 that it has a route to S. Similarly, PBR2 reports to both ASBR1 and ASBR3 that it has a route to S. Using [ADD-PATH], ASBR1 reports both routes to ASBR2, and ASBR3 reports both routes to ASBR4. Now AS-Y sees 4 paths to S. The AS-Z ASBRs will each see eight paths (four via ASBR2 and four via ASBR4). To avoid this explosion in the number of paths, a BGP speaker that uses [ADD-PATH] is usually considered to report only the n best paths. However, there is then no guarantee that the reported set of paths will contain at least one path via PBR1 and at least one path via PBR2. Without such a guarantee, the SFS procedure will not work.

#### 2.4. Inclusive and Selective Tunnels

The MVPN specifications allow multicast flows to be carried on either Inclusive Tunnels or on Selective Tunnels. When a flow is sent on an Inclusive Tunnel of a particular VPN, it is sent to all PEs in that VPN. When sent on a Selective Tunnel of a particular VPN, it may be sent to only a subset of the PEs in that VPN.

This document allows the use of either Inclusive Tunnels or Selective Tunnels for GTM. However, any service provider electing to use Inclusive Tunnels for GTM should carefully consider whether sending a multicast flow to ALL its PBRs would result in problems of scale. There are potentially many more MBRs for GTM than PEs for a particular VPN. If the set of PBRs is large and growing, but most multicast flows do not need to go to all the PBRs, the exclusive use of Selective Tunnels may be a better option.

#### 2.5. I-PMSI A-D Routes

##### 2.5.1. Intra-AS I-PMSI A-D Routes

Per [MVPN-BGP}, there are certain conditions under which is it NOT required for a PE router implementing MVPN to originate one or more Intra-AS I-PMSI A-D routes. These conditions apply as well to PBRs implementing GTM.

In addition, a PBR implementing GTM is NOT required to originate an Intra-AS I-PMSI A-D route if both of the following conditions hold:

- The PBR is not using Inclusive Tunnels for GTM, and

- The distribution of the C-multicast Shared Tree Join and C-multicast Source Tree Join routes is done in such a manner that the next hop of those routes does not change.

Please see also the sections on RD and RT usage.

#### 2.5.2. Inter-AS I-PMSI A-D Routes

There are no GTM-specific procedures for the origination, distribution, and processing of these routes, other than those specified in the sections on RD and RT usage.

#### 2.6. S-PMSI A-D Routes

There are no GTM-specific procedures for the origination, distribution, and processing of these routes, other than those specified in the sections on RD and RT usage.

#### 2.7. Leaf A-D Routes

There are no GTM-specific procedures for the origination, distribution, and processing of these routes, other than those specified in the sections on RD and RT usage.

#### 2.8. Source Active A-D Routes

There are no MANDATORY GTM-specific procedures for the origination, distribution, and processing of these routes, other than those specified in the sections on RD and RT usage.

However, this document defines an OPTIONAL procedure to allow additional constraints on the distribution of the Source Active A-D routes for GTM. If some site has receivers for a particular ASM group G, then it is possible (by the procedures of [RFC6514]) that every PBR attached to site with a source for group G will originate a Source Active A-D route whose NLRI identifies that source and group. These Source Active A-D routes may be distributed to every PBR. If only a relatively small number of PBRs are actually interested in traffic from group G, but there are many sources for group G, this could result in a large number of (S,G) Source Active A-D routes being installed in a large number of PBRs that have no need of them.

For GTM, it is possible to constrain the distribution of (S,G) Source Active A-D routes to those PBRs that are interested in GTM traffic to



group G. This can be done using the following OPTIONAL procedures:

- If a PBR originates a C-multicast Shared Tree Join whose NLRI contains (RD=0,\*,G), then it dynamically creates an import RT for its global table, where the Global Administrator field of the RT contains the group address G, and the Local Administrator field contains zero. (Note that an IPv6-address-specific RT would need to be used if the group address is an IPv6 address.)
- When a PBR creates such an import RT, it uses "RT Constraint" [RFC4684] procedures to advertise its interest in routes that carry this RT.
- When a PBR originates a Source Active A-D route from its global table, it attaches the RT described above.
- When the C-multicast Shared Tree Join is withdrawn, so is the corresponding RT constrain route, and the corresponding RT is removed as an import RT of its global table.

These procedures enable a PBR to automatically filter all Source Active A-D routes that are about multicast groups in which the PBR has no interest.

This procedure does introduce the overhead of distributing additional "RT Constraint" routes, and therefore may not be cost-effective in all scenarios, especially if the number of sources per ASM group is small. This procedure may also result in increased join latency.

## 2.9. C-multicast Source/Shared Tree Joins

[RFC6514] section 11.1.3 has the following procedure for determining the IP-address-specific RT that is attached to a C-multicast route: (a) determine the upstream PE, RD, AS, (b) find the proper Inter-AS or Intra-AS I-PMSI A-D route based on (a), (c) find the next hop of that A-D route, (d) base the RT on that next hop.

However, for GTM, in environments where it is known a priori that that the next hop of the C-multicast Source/Shared Tree Joins does not change during the distribution of those routes, the proper procedure for creating the IP-address-specific RT is to just put the IP Address of the Upstream PBR in the Global Administrator field of the RT. In other scenarios, the procedure of the previous paragraph (as modified by this document's sections on "RD usage" and "RT usage") is applied by the PBRs.

### 3. Differences from other MVPN-like GTM Procedures

The document [SEAMLESS-MCAST] also defines a procedure for GTM that is based on the BGP procedures that were developed for MVPN.

However, the GTM procedures of [SEAMLESS-MCAST] are different than and are NOT interoperable with the procedures defined in this document.

The two sets of procedures can co-exist in the same network, as long as they are not applied to the same multicast flows or to the same ASM multicast group addresses.

Some of the major differences between the two sets of procedures are the following;

- The [SEAMLESS-MCAST] procedures for GTM do not use C-multicast Shared Tree Joins or C-multicast Source Tree Joins at all. The procedures of this document use these C-multicast routes for GTM, setting the RD field of the NLRI to zero.
- The [SEAMLESS-MCAST] procedures for GTM use Leaf A-D routes instead of C-multicast Shared/Source Tree Join routes. Leaf A-D routes used in that manner can be distinguished from Leaf A-D routes used as specified in [RFC6514] by means of the NLRI format; [SEAMLESS-MCAST] defines a new NLRI format for Leaf A-D routes. Whether a given Leaf A-D route is being used according to the [SEAMLESS-MCAST] procedures or not can be determined from its NLRI. (See [SEAMLESS-MCAST] section "Leaf A-D Route for Global Table Multicast".)
- The Leaf A-D routes used by the current document contain an NLRI that is in the format defined in [RFC6514], NOT in the format as defined in [SEAMLESS-MCAST]. The procedures assumed by this document for originating and processing Leaf A-D routes are as specified in [RFC6514], NOT as specified in [SEAMLESS-MCAST].
- The current document uses an RD value of zero in the NLRI in order to indicate that a particular route is "about" a Global Table Multicast, rather than a VPN multicast. No other semantics are inferred from the fact that RD is zero. [SEAMLESS-MCAST] uses two different RD values in its GTM procedures, with semantic differences that depend upon the RD values.
- In order for both sets of procedures to co-exist in the same network, the PBRs MUST be provisioned so that for any given IP group address in the global table, all egress PBRs use the same set of procedures for that group address (i.e., for group G,

either all egress PBRs use the GTM procedures of this document or all egress PBRs use the GTM procedures of [SEAMLESS-MCAST].

#### 4. IANA Considerations

This document has no IANA considerations.

#### 5. Security Considerations

The security considerations of this document are primarily the security considerations of the base protocols, as discussed in [RFC6514], [RFC4601], and [RFC5294].

This document makes use of a BGP SAFI (MCAST-VPN routes) that was originally designed for use in VPN contexts only. It also makes use of various BGP path attributes and extended communities (VRF Route Import Extended Community, Source AS Extended Community, Route Target Extended Community) that were originally intended for use in VPN contexts. If these routes and/or attributes leak out into "the wild", multicast data flows may be distributed in an unintended and/or unauthorized manner.

Internet providers often make extensive use of BGP communities (ie, adding, deleting, modifying communities throughout a network). As such, care should be taken to avoid deleting or modifying the VRF Route Import Extended Community and Source AS Extended Community. Incorrect manipulation of these ECs may result in multicast streams being lost or misrouted.

The procedures of this document require certain BGP routes to carry IP multicast group addresses. Generally such group addresses are only valid within a certain scope. If a BGP route containing a group address is distributed outside the boundaries where the group address is meaningful, unauthorized distribution of multicast data flows may occur.

## 6. Additional Contributors

Zhenbin Li  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China  
Email: lizhenbin@huawei.com

Wei Meng  
ZTE Corporation  
No.50 Software Avenue, Yuhuatai District  
Nanjing  
China  
Email: meng.wei2@zte.com.cn, vally.meng@gmail.com

Cui Wang  
ZTE Corporation  
No.50 Software Avenue, Yuhuatai District  
Nanjing  
China  
Email: wang.cuil@zte.com.cn

Shunwan Zhuang  
Huawei Technologies  
Huawei Bld., No.156 Beiqing Rd.  
Beijing 100095  
China  
Email: zhuangshunwan@huawei.com

## 7. Acknowledgments

The authors and contributors would like to thank Rahul Aggarwal, Huajin Jeng, Hui Ni, Yakov Rekhter, and Samir Saad for their contributions to this work.

## 8. Authors' Addresses

Lenny Giuliano  
Juniper Networks  
2251 Corporate Park Drive  
Herndon, VA 20171  
US  
Email: lenny@juniper.net

Dante J. Pacella  
Verizon  
Verizon Communications  
22001 Loudoun County Parkway  
Ashburn, VA 20147  
US  
Email: dante.j.pacella@verizonbusiness.com

Eric C. Rosen  
Cisco Systems, Inc.  
1414 Massachusetts Avenue  
Boxborough, MA, 01719  
US  
Email: erosen@cisco.com

Jason Schiller  
Google  
1818 Library Street  
Suite 400  
Reston, VA 20190  
US  
Email: jschiller@google.com

Karthik Subramanian  
Cisco Systems, Inc.  
170 Tasman Drive  
San Jose, CA, 95134  
US  
Email: kartsubr@cisco.com

Jeffrey Zhang  
Juniper Networks  
10 Technology Park Dr.  
Westford, MA 01886  
US  
Email: zzhang@juniper.net

## 9. References

### 9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4364], Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks", RFC 4364, February 2006.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.
- [RFC6515] Aggarwal, R., and E. Rosen, "IPv4 and IPv6 Infrastructure Addresses in BGP Updates for Multicast VPN", RFC 6515, February 2012.

### 9.2. Informative References

- [ADD-PATH] "Advertisement of Multiple Paths in BGP", D. Walton, A. Retana, E. Chen, J. Scudder, draft-ietf-idr-add-paths-09.txt, October 2013.
- [RFC6368] Marques, P., Raszuk, R., Patel, K., Kumaki, K., and T. Yamagata, "Internal BGP as the Provider/Customer Edge Protocol for BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 6368, September 2011.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC4684] P. Marques, et. al., "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684,

November 2006.

[RFC5294] Savola, P. and J. Lingard, "Host Threats to Protocol Independent Multicast (PIM)", RFC 5294, August 2008.

[MVPN-extranet] Rekhter, Y. and E. Rosen (editors), "Extranet Multicast in BGP/IP MPLS VPNs", draft-ietf-l3vpn-mvpn-extranet-03.txt, January 2014

[SEAMLESS-MCAST] Rekhter, Y., Aggarwal, R., Morin, T., Grosclaude, I., Leymann, N., and S. Saad, "Inter-Area P2MP Segmented LSPs", draft-ietf-mpls-seamless-mcast-09.txt, December 2013