Network Working Group
IETF Internet Draft
Expires: August 2005

Seisho Yasukawa
NTT

Shankar Karuna
Sarveshwar Bandi
Motorola

Adrian Farrel
Old Dog Consulting

February 2005

BGP/MPLS IP Multicast VPNs

draft-yasukawa-l3vpn-p2mp-mcast-01.txt

Status of this Memo

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

Copyright Notice

Abstract

This document describes a solution framework for IP Multicast VPNs. It describes procedures for establishing optimal virtual private IP multicast networks over a provider network. The simple multicast tunnel operation mechanism within a core network provides easy and flexible IP multicast VPN service operation for the service provider. And because the solution can minimize PIM neighbor maintenance over remote PEs, the solution enhances the scalability performance of the multicast VPN service network. This document also describes a P2MP TE LSP based multicast tunnel mechanism which could enhance TE capability and reliability of IP multicast VPNs.

Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Contents

1. Introduction

This document describes a solution framework for IP Multicast VPNs. It describes basic procedures for establishing optimal virtual private IP multicast networks over a provider network by dividing a customer's multicast network into multiple regions and setting independent multicast distribution trees within each customer site and interconnecting these independent trees by provider P2MP MPLS tunnels. In this solution, each PE connected to a customer's site acts as a proxy-source or a proxy-rendezvous point (RP) for the multicast group and a default multicast tunnel is established from a PE which accommodates a multicast source to multiple PEs which act as proxy-source/RP for their receivers in the connected site.

As a result, the solution can construct IP multicast distribution trees that have optimal topologies for IP multicast distribution and avoid using multiple multicast and unicast distribution tunnels in the service provider core during the customer's tree transition phase. This simple multicast tunnel operation mechanism within a core provides easy and flexible IP multicast VPN service operation for the service provider. And, because the solution can terminate each customer's Join/Prune message at PEs, the solution can minimize PIM neighbor maintenance over remote PEs. This enhances the scalability performance of multicast VPN service network. This document also describes a P2MP TE LSP based multicast tunnel mechanism which could enhance TE capability and reliability of IP multicast VPNs.

2. Motivations

2.1. Basic Motivations for IP multicast VPN operation

There are growing demands for providing IP multicast services on top of a BGP/MPLS IP VPN [RFC2547bis] environment. Candidates for these services are, for example, in-house video distribution/conference and data synchronization/distribution between business system servers which usually require strict QoS control and highly reliable network conditions. Therefore it is highly desirable that a solution has the capability to implement some QoS guarantee and protection mechanisms in itself, or has capabilities to operate with conventional QoS guarantee and protection mechanisms.

Because IP multicast VPNs require full meshed multicast distribution between multiple customer sites, and this operation usually requires complicated multicast tree management within a provider core network, it is also highly desirable that a solution provides the service provider with several easy mechanisms to control and manage these IP multicast distributions within the network.

2.2. Motivations for IP multicast VPN protocol design

The basic function of an IP multicast VPN is to enable a multicast source which exists in a customer site to send IP multicast traffic privately over the provider core network to multicast receivers that exist in different customer sites.

To enable this private IP multicast transmission, several solutions have been proposed. Within the proposals, the most significant solution is [MCAST-VPN] which introduces the Multicast Domain (MD) mechanism to interconnect each customer's IP multicast networks over the provider network to enable IP multicast distribution between the sites. An MD is essentially a set of MVRFs associated with interfaces that can send multicast traffic to each other and is equivalent to a multi-access interface from the standpoint of a PIM customer instance. Therefore, in this MD model, a provider wide customer IP multicast network is formed over an MD which transparently exchanges customer's IP multicast control messages between PEs which form IP multicast adjacencies between them. This means that when the customer network runs the PIM protocol, PIM adjacencies are formed between MVRFs at the PEs and periodic PIM Join/Prune messages and Hello messages are transmitted between them.

This features introduces some scalability concerns for the service provider when they operate IP multicast VPNs because today's conventional L3VPNs accommodate a lot of large scale VPNs and it is easily assumed that a majority of these VPNs will introduce IP multicast VPN services in the near future. This would require a huge amount of PIM adjacencies to be maintained over the provider core network and this would reduce the VPN's network performance and increase the difficulties of managing the network.

A further MVPN proposal [MVPN-RAGGARWA] addresses three scalability concerns for this conventional [MCAST-VPN] solution as fundamental issues to be resolved. The first addressed concern is the overhead of PIM neighbor adjacencies. The second concern is the overhead of

periodic PIM Join/Prune messages, and the last concern is the amount of state in the SP core. It is highly desirable for any solution to address these concerns.

Another concern which is not yet addressed is suboptimal IP multicast distribution which could easily occur in an IP multicast VPN environment. Most customers want to operate their own private IP multicast networks over multiple customer sites which are usually widely separated from each other over the provider network, and it is easily assumed that the customer runs PIM [PIM-SM] with only a very limited number of RPs in sites which may be located remotely from the site which a multicast source belongs to. In this case, during shared tree distribution, multicast packets must first be sent to the RP's site and then must to be sent back to receivers' sites even in the case where some receivers belong to the same site as the source. This kind of multiple transmission over the provider network is suboptimal and wastes network resource. Moreover some receivers would suffer severe transmission delays and some multicast application would not tolerate this.

In addition to these undesirable conditions, a customer's IP multicast distribution pattern changes drastically when the distribution tree is transferred from a shared tree to a source specific tree. This would also cause drastic changes in the multicast distribution pattern in the provider core and would introduce unstable conditions for the operation of the core network and consequently for the VPNs. There is no mechanism for a service provider to limit this kind of undesirable condition and to control the multicast distribution pattern. Therefore it is highly desirable for a solution to address these concerns. It is desirable for a solution to provide the service provider with mechanisms which can avoid this kind of suboptimal IP multicast distribution within their core networks and which can allow control of multicast distribution within their core networks by eliminating the necessity of using and changing multiple multicast tunnels and by providing a minimum number of stable multicast tunnels which are easily managed by the service provider.

Because some IP multicast VPN applications require bandwidth guarantees, delay-constrained multicast distribution paths, and highly reliable paths, it is desirable for a solution to have capabilities to setup a bandwidth guaranteed multicast distribution path, to explicitly control routes of the distribution path, and to accommodate global and fast local repair mechanisms.

3. IP Multicast VPN Framework

3.1. Basic network model

This document utilizes the same network model as BGP/IP MPLS VPNs [RFC2547bis] for realizing IP multicast VPNs. A provider configures whether a particular VPN is multicast-enabled or not. All the PEs that contain customer sites belonging to the same VPN with multicast enabled on them will be connected using a "Multicast Distribution Tree (MDT)" in the provider's backbone.

As deployed in BGP/IP MPLS VPNs [RFC2547bis] and [MCAST-VPN], each CE router is a multicast routing adjacency of a PE router, but CE routers at different sites do NOT become multicast routing adjacencies of each other.

Unlike the [MCAST-VPN] model, this document proposes the use of BGP for both discovering IP multicast VPN membership and exchanging IP multicast routing information in a given IP multicast VPN.

As for the membership discovery, all the PE routers advertise their IP multicast VPN membership to other PE routers using BGP so that each PE router in the network has a complete view of the IP multicast VPN membership of the other PE routers.

After this information exchange, the Default MDT for a Multicast Domain is constructed automatically. Note that this Default MDT construction occurs when the PEs in the domain come up and advertise their membership of a multicast enabled VPN, and the construction does not depend on the existence of multicast traffic in the domain.

One further difference is that the MDTs are usually created by P2MP TE LSPs by running a P2MP TE signaling protocol in the backbone. Therefore, it may not be necessary to run IP multicast routing protocols in the core to support IP multicast VPNs (dependent on how the P2MP TE trees are managed).

As for multicast routing information exchange, this document proposes BGP to carry the information. Therefore, being different from the [MCAST-VPN] model, the customer's IP multicast instances of a particular VPN running on the PE do not form routing adjacencies

with each other. This means that the customer's IP multicast control packets are always terminated at PEs and independent multicast domains are formed within each customer site which is a member of the IP multicast VPN. In this model, the customer's IP multicast control messages, such as PIM Join/Prune messages, are converted to BGP messages and these messages are exchanged between PEs so that IP multicast routing can be enabled between each independent IP multicast domains.

This preserves two important features of BGP/IP MPLS VPNs [RFC2547bis]; "Separation of controlling and forwarding plane in the provider core" and "Distribution of customer's routing information via provider's routing facility". These are very important preservations for the SP to preserve its network management model while allowing it to control/engineer the customer's data traffic and control traffic over the network.

Multicast data packets from within a VPN are received from a CE router by an ingress PE router, the ingress PE then encapsulates the mutlicast packets and forwards them along the Default MDT to all the PE routers connected to sites of the given VPN. Every PE router attached to a site of the given VPN thus receives all multicast packets from within that VPN. If a particular PE router is not on the path to any receiver of that multicast group, the PE simply discards that packet.

In the same way as [MCAST-VPN], this document proposes to set up Data MDTs to accommodate a large amount of traffic being sent to a particular multicast group but where that group does not have receivers at all the VPN sites.

3.2. Proxy-Source/RP model

Section 2.2 of [RFC3446] describes the need to run multiple RPs in the scenarios where the topological location of RP, Source and receivers are unpredictable. This document proposes a solution along similar lines.

This document proposes that all PEs which are members of a given VPN act as proxy Source/RPs for a given IP multicast group.
Approving all the PEs to be a proxy Source/RP for the group, each

customer site can form an independent IP multicast tree within the site regardless of multicast tree formations in other sites.

To accomplish this model, a PE that is either directly connected to the multicast source in the VPN or connected via a CE MUST act as a proxy-RP for the receivers in the same site when the customer network runs PIM-SM protocol in the VPN.

If a customer wants to run the RP within his network, in such a case the customer can configure one RP on each of his sites and can use an anycast-RP and MSDP based mechanism [RFC3446]. This solution can be extended to support such an approach and is presently out of scope of this document.

A PE that is either directly connected to an RP in the VPN or connected via a CE MUST act as a Source/RP for the receivers in the same site, and a PE that is either directly connected to multicast receivers for that multicast group or connected via a CE MUST act as a Source/RP for the receivers in the same site.

A P2MP TE LSP or a MP2MP TE LSP which is described in a later section is established from an ingress PE to multiple egress PEs to form a default MDT. The setup of default MDT is triggered after the VPN membership auto-discovery phase.

Therefore, each egress PE which acts as a proxy-Source/RP can always receive IP multicast data traffic via this LSP tunnel.

A P2MP TE LSP is established from an ingress PE to egress PEs which are interested in receiving the multicast data dynamically to form a data MDT. The ingress PE sets up and modifies a data MDT dynamically by receiving BGP messages which convey egress PEs interests for joining/leaving the VPN membership. Therefore, each egress PE which acts as a proxy-Source/RP can receive IP multicast data traffic via this LSP tunnel on demand.

3.3. Proxy-Source/RP function

To make each PE interconnected to a customer site act as a proxy-RP,

this document assumes that some RP discovery mechanisms, such as static configuration or Bootstrap Router, indicates all of the PIM router existing in the connected customer site to perform the same group-to-RP (PE) mapping. This ensures that all the PIM router in the customers site utilize the PE as RP.

The PE which act as RP and interconnected to IP multicast source must handle PIM register message operations, including decapsulation and sending source specific join message and PIM register stop message and must convert PIM register message to IP multicast source active information.

The PE which acts as RP and is interconnected to multicast receivers must terminate PIM Join/Prune messages and convert this information to IP multicast registration information.

3.4. Auto-Discovery of VPN membership

This document assumes MDT SAFI [MDT-SAFI] to discover VPN membership. In this model, a MDT group-address is defined per Multicast Domain and all the PEs that are configured with MVRFs belonging to same IP multicast VPN discover each other thorough this SAFI.

3.5. Default MDT configuration

After VPN membership discovery, each PE which is a member of an IP multicast VPN will trigger the formation of a P2MP tree towards every other PE that has Multicast VRFs for the same Multicast Domain. The P2MP TE LSPs [P2MP-RSVP] are established by assigning a MDT group-address to the P2MP ID of the SESSION object, and assigning the initiator PE's address to the Tunnel Sender Address of the SENDER_TEMPLATE object. The destination PEs are designated as leaves of the P2MP tree and are encoded as recipients in P2P Sub-LSP Objects.

In order for each PE to forward received IP multicat data packets to the appropriate MVRF, each PE which is a member of the IP multicast VPN MUST associate the P2MP TE LSPs with a proper MVRF by assigned P2MP Labels. These associations are configured when a PE assigns P2MP

Labels to the P2MP TE LSPs. Therefore, in this model, PHP operation MUST be strictly prohibited.

A MP2MP TE LSP can be also utilized for establishing a Default MDT. With the help of Route Reflector functionality in MP-BGP the PE's interested in this Multicast Domain are learnt at the node that runs as the Route reflector and these PE's are configured as leaves for P2MP TE LSP with the route reflector node as root.

After an RP has set up a P2MP TE LSP to the leaves, the leaves setup a P2P TE LSPs to the RP. In this way, a MP2MP TE LSP is established between PEs which are members of the Multicast Domain. This MP2MP based Default MDT configuration mechanism is useful but it is out of scope of this document. The detailed mechanism and procedure will be defined in the another [MP2MP-TE-MPLS] document.

3.6 Exchanging IP multicast register information

This documents proposes exchanging two kinds of IP multicast register information between PEs via BGP.

A new Subsequent-Address Family called Source Activate (SA) SAFI is newly defined to announce the activation of a particular customer's IP multicast data stream. This SAFI is used by a PE which is either directly connected to an IP multicast source or connected via a CE to inform other PEs located in different sites that a particular customer's (S,G) IP multicast data stream has become active and this active IP multicast data can be accessed via this announcing PE. Therefore, a PE which is either directly connected to an IP multicast source or connected via a CE and acts as proxy RP and sends this information via BGP after it confirms establishment of a source specific IP multicast tree between the Designated Router and itself by sending a Register-Stop message and receiving a Null-Register Message.

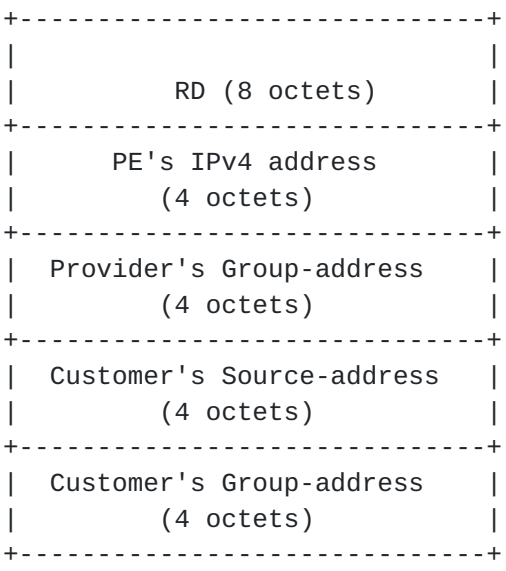
Another Subsequent-Address Family called JOIN SAFI is also newly defined to announce the interest of a particular PE to join and prune a particular customer's IP multicast data stream. This SAFI is used by PEs which are either directly connected to IP multicast receivers or connected via CEs to inform a PE which is either directly connected to an IP multicast source or connected via a CE that they

are interested in receiving a particular customer's (S,G) IP multicast data stream by announcing JOIN SAFI information. Therefore PEs which are either directly connected to IP multicast receivers or connected via CEs and which act as proxy-Source/RPs send this information via BGP after they have received SA information and have confirmed that IP multicast receivers in their site are also interested in receiving/leaving this IP multicast data stream by receiving customer Join/Prune message.

3.6.1. Source Activate (SA) SAFI

This SAFI is used to announce to all the other PEs about a particular customer (S,G) data stream becoming active.

SA SAFI:



PE's IPv4 address: IP address of the PE that has detected a multicast source in the customer network.

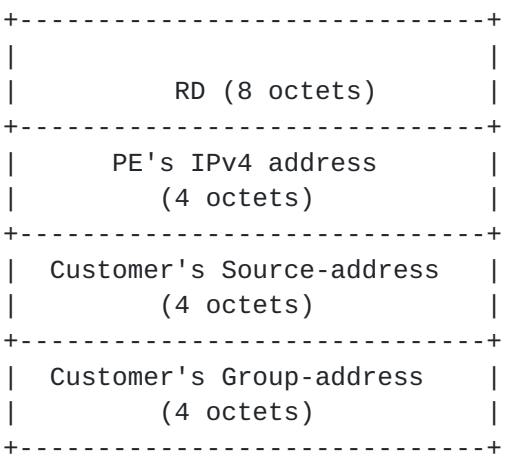
Provider's Group-address: The PE assigns a multicast group address from an address range that is independent of the customer multicast group addresses. This multicast group address is used by the PE's that have interested receivers to be able to associate a Data MDT associated with data stream corresponding to Customer's Source-address and Customer's Group-address

Customer's Source-address: Address of customer's multicast source.
Customer's Group-address : Address of customer's multicast group address.

3.6.2. JOIN SAFI

This SAFI will be used by a PE when it desires to receive data for a particular customer (S,G) data stream.

JOIN SAFI:



PE's IPv4 address: IP address of the PE that has interested receivers.

Customer's Source-address: Address of customer's multicast source.
Customer's Group-address : Address of customer's multicast group address.

3.7 Default MDT operation

This section describes how the proposed solution enables IP multicast VPN operation using a Default MDT assuming a VPN customer runs the PIM-SM protocol within their network.

To establish a Default MDT, MVRFs are first configured on every PE which is a member of a given IP multicast VPN. A unique group address and RD are assigned to that Multicast Domain. After this configuration, each that PE belongs to that Multicast Domain announces its VPN membership to other PEs by BGP using MDT-SAFI.

After this auto-discovery of VPN membership, each PE initiates the formation of a P2MP TE LSP by assigning the group address to that LSP. A P2MP TE LSP is established from the initiator PE to other PEs which are also members of this Multicast Domain. During the P2MP TE LSP establishment each PE that is also egress of the LSP and a member of that Multicast Domain configures an association between P2MP TE LSPs which constitute Default MDT and MVRF for that Multicast Domain by relating assigned MPLS labels to the MVRF. The egress PE uses the group address announced by the PE, which is the root for this P2MP TE LSP, in the MDT-SAFI message [MDT-SAFI].

Note that as described in the previous section, a MP2MP TE LSP could be utilized to establish the Default MDT. But MP2MP TE MPLS is out of scope of this document and its detailed mechanism and procedure will be addressed in another document.

Because in this mechanism, each PE that is member of the Multicast Domain must act as a proxy-RP for that multicast groups. Therefore, all PIM router within a customer's site must be configured by some mechanism to treat a connected PE as its RP and to perform same group-to-proxy-RP mapping. Examples of these mechanism are static configuration and Bootstrap Router mechanism [PIM-BSR] described in the previous section.

Consider the situation where an IP multicast source starts IP multicast distribution. The Designated Router (DR) encapsulates IP multicast data packets and sends these packets as PIM register messages to a PE which now act as proxy-RP for that customer's site. Receiving PIM register messages, the PE sends a source specific Join message to the DR to form a source specific multicast tree between the PE and the DR. The PE simultaneously starts announcing to other PEs the activation of a multicast group via BGP using the SA SAFI. The PE sends a Register stop message to the DR after it starts receiving multicast Data on the source specific tree.

When receivers in another customer site want to Join a given multicast distribution, then the receivers send Join (*,G) messages

to their RP to form a shared multicast distribution tree. Note that a PE which is either directly connected to the receivers or connected via a CE now acts as proxy-RP for the receivers. Therefore a shared multicast distribution tree is formed from proxy-RP (PE) to multiple receivers in a customer's site. And because this PE knows that corresponding multicast data is already activated, the PE sends its IP multicast registry information via BGP using JOIN SAFI. This information is flooded to all the PEs. As a result, an ingress PE which is either directly connected to the IP multicast source or connected via a CE configures its MVRF to start forwarding received IP multicast data packets to a P2MP TE LSP which comprises the Default MDT. In this way, IP multicast data packets are MPLS encapsulated at an ingress proxy-RP (PE) and are tunneled via the P2MP TE LSP to all the egress proxy-Source/RPs (PEs) and these IP multicast data packets are decapsulated at the egress PEs and IP multicast data packets are transmitted to IP multicast receivers in the site if a shared tree is already formed by multicast receivers who are interested in receiving these IP multicast flows.

When receivers in the customer's site want to switch multicast distribution trees from a shared tree to a source specific tree, the receivers send source specific Join (S,G) message to source node. In this case, a source node is located in another site and therefore this message is always transmitted to the PE which now acts as the proxy-RP for that multicast group. Receiving this source specific Join message, the PE changes IP multicast distribution to the source specific tree because the PE starts acting as proxy-Source from that point of time. Note that this multicast tree switch over does not cause any additional JOIN SAFI transmission by the PE because the PE can already receive IP multicast data and can act as proxy- Source without changing the multicast distribution operation over the provider core network. In this way, each customer's IP multicast domain switches IP multicast tree from shared tee to source specific tree independently without affecting the multicast distribution within a provider core network and another customer's sites.

3.8. Data MDT operation

This document proposes to setup Data MDTs in addition to a Default MDT when IP multicast transmission over a Default MDT is judged as inefficient. The difference from conventional approaches [MCAST-VPN] [MVPN-RAGGARWA] is this document assumes that Data MDTs are constructed by [P2MP-RSVP]. Therefore, this proposal can construct QoS guaranteed and highly reliable Data MDTs which would be better

for a particular class of IP multicast traffic.

This section describes how the proposed solution enables IP multicast VPN operation using the Data MDT assuming a VPN customer runs the PIM-SM protocol within their network. In this document, we assume two kinds of Data MDT construction model; Static configuration model and Traffic driven model.

In the static configuration model, we assume that IP multicast flows which are transmitted by Data MDTs are pre-defined and mutually agreed by the SP and its customers. This means that multicast group addresses assigned for IP multicast flows which use Data MDT are pre-defined and configured on each PE's MVRF. Therefore an ingress PE which is registered by this multicast group address can setup, modify and tear-down an appropriate P2MP TE LSP on demand. When several downstream PEs which act as proxy-RP for interconnected customer sites detect the existence of receivers which are interested in joining a particular multicast distribution by receiving Join messages, then the PEs report their interest to join the group by BGP using the JOIN SAFI. After receiving these reports, an ingress PE establishes a P2MP TE LSP to the egress leaves which have reported interest. After this operation, when a new PE uses the BGP JOIN SAFI to report to the ingress PE its interest to join the the group, the ingress PE initiates Grafting and expands the P2MP TE LSP to reach that new PE. When an existing PE detects that no more receivers are connected to a customer's site by receiving Prune messages, the PE withdraws the corresponding IP multicast registration by using BGP withdraw message specifying the corresponding JOIN SAFI.

Then the ingress PE initiates Pruning and cuts out the unnecessary leaf from the P2MP TE LSP. In this way, this proposal can setup, modify and tear-down Data MDT on demand basis.

In the traffic driven model, an ingress PE monitors incoming IP multicast data packets and when it detects some data flow exceeds a pre-determined threshold, then the ingress PE immediately establishes a new P2MP TE LSP to reach the receivers' PEs because the ingress PE recognizes which PEs are interested in joining this group via BGP. After establishment, the ingress PE switches corresponding IP multicast data packets from the Default MDT to the Data MDT. In this way, Data MDT based IP multicast transmission is performed on demand in this model. After this Data MDT establishment, this model follows exactly the same operations as the static configuration model when the Data MDT is modified.

In order to associate the Data MDT with the appropriate MVRF, the egress PEs utilize the PE Group-address announced by the PE (which is the root of the P2MP TE LSP) via the SA SAFI.

3.9. Inter-Site Scaling and Site Interdependencies

As described in section 3.8, customer sites may be members of the same multicast VPN yet operate in near independence. That is, each site has its own RP and may choose to use a source-specific tree based at the PE, or a shared distribution tree. This choice is private to each customer site, and is not communicated to other sites (nor would it provide any useful information if it were communicated).

Clearly, also, the number of leaves downstream of an egress PE does not affect the way in which the traffic is managed at upstream nodes. That is, neither the source site nor the MPLS P2MP TE LSP in the SP's network are in any way different if there is one or one hundred receivers at a customer site downstream of an egress PE. The PE represents a fixed point in the multicast tree that must receive data and is an egress of the MPLS P2MP TE LSP.

For this reason, there is no requirement for the routing protocols to report each receiver across the SP network to the ingress site when the receivers are added to or removed from the multicast group. All that is required is that the egress PE reports (using the BGP JOIN SAFI) when the first receiver joins the group so that it becomes a leaf on the MPLS P2MP TE LSP, and indicates when the last receiver at the site leaves the multicast group so that the PE can be pruned from the MPLS P2MP TE LSP. In order to make this optimization each PE must count the number of receivers at its site, but since it is acting as a proxy-source/RP this is easy for it to do. Such an optimization substantially reduces the amount of MP-BGP traffic caused by the VPN multicast group and is RECOMMENDED for all implementations.

Note that an egress PE that makes this optimization may further protect itself against flapping membership of a multicast group. In the case where the membership may frequently vary between no receivers and some receivers an egress PE MAY choose to remain as a leaf of the MPLS P2MP TE LSP for some period of time (controllable by the operator) even when it has no downstream receivers for the multicast traffic. If a local timer expires before any new receivers join the group, the PE should use MP-BGP to report that it no longer

wishes to receive data for the multicast group, and this will result in it being pruned from the MPLS P2MP TE LSP tree. Any traffic received by a PE for a multicast group for which it has no downstream receivers SHOULD be discarded.

3.10 Targeted JOIN SAFI transmission

The advertisement of multicast JOIN-SAFI information in response to a SA-SAFI, as per standard BGP procedures, will be advertised to all the PEs which act as RPs. The JOIN-SAFI is used only by the PE which generated the SA-SAFI and would be dropped on other PEs. Therefore it is highly desirable to avoid sending the BGP JOIN-SAFI messages to PEs which do not require to receive it.

This document proposes a BGP filtering mechanism termed Multicast distribution filtering (MDF) that would help to restrict the advertisement of the JOIN-SAFI to only the PE which advertised the SA-SAFI. Multicast distribution filters will be created dynamically on the downstream PE at the time of receiving the SA-SAFI message. When the downstream PE attempts to respond with a JOIN-SAFI, MDF filter restricts the distribution of the JOIN-SAFI message so that it is sent only to the PE from which SA-SAFI was received.

The MDF BGP mechanism will be detailed in a later revision of this document.

3.11. Multi-Homing scenario

The proposed solution provides a very effective fail over mechanism in the case where the multicast source/receivers are connected to multiple PEs. The PEs which are connected to the customer network announce themselves as the RPs via the bootstrap mechanism. When one of the PE fails the alternate PE becomes RP for this multicast domain and this PE can readily takeover as it is already connected via the default MDT.

Anycast-RP mechanism can also be used to provide RP redundnacy.

3.12 Inter-Provider Backbones

The solution described in this document can be easily extended to Inter-AS operations in line with the model described in 2547bis.

3.12.1 Option A

In VRF-to-VRF connections at the AS border routers (Option A), the PEs associate a sub-interface with a VRF, and use PIM to exchange unlabeled control/Data multicast information with each other. An ASBR PE router, will send a PIM Register message to its EBGP peer on the connected sub-interface on detecting a multicast source (either by receiving a SA-SAFI message from its IBGP peer, or by receiving a register message). The rest of the operations do not need any modifications and are same the as explained in earlier sections of this document for single AS operations.

3.12.2 Option B

In the case of option B, the ASBR PE router exchanges MDT-SAFI, JOIN-SAFI and SA-SAFI messages with its MP-EBGP peers just as it does with its MP-IBGP peers. In the case of MP-EBGP peer, the MDT-SAFI will carry a label. The PE, when forwarding across AS boundaries, encapsulates the multicast data with the label that the downstream PE advertised in the MDT-SAFI. The downstream PE will use this label to map the data to a VRF. After identifying the VRF, the downstream PE can perform the operations as described previously in the draft.

The detailed procedures and modification required to the BGP message will be described in later version of this document.

3.12.3 Option C

In the case of option C, MP-EBGP multihop peering is established for the PEs. The PEs will attempt to form P2MP LSPs across AS boundaries.

3.13 Support for Other PIM Variants

3.13.1 PIM-SSM Support

PIM-SSM can be supported using the procedures described in this draft with a slight change to the sequence of JOIN-SAFI and SA-SAFI BGP messages. In the case of PIM-SSM, the PIM routers connected to the multicast source will not send PIM register messages to the RP (PE) for multicast data being sent to group addresses that fall in the PIM-SSM range. Thus it is not possible for the PE to send SA-SAFI messages for these multicast streams. In such cases, when the downstream PE router receives PIM Join(S,G) from the CE router, it will directly generate JOIN-SAFI message to the upstream PE router, without waiting for the SA-SAFI message.

When an upstream PE router receives a JOIN-SAFI message for a

customer multicast group address in the PIM-SSM range, it will trigger PIM Join(S,G) towards the source of multicast data. The SA-SAFI message is triggered by the upstream PE on detecting the multicast data flow. The provider group address advertised in the SA-SAFI message will be used by the downstream PE to map multicast data sent on the data tunnel to the right VPN.

3.13.2 PIM-DM/PIM-BIDIR

The support for these PIM variants will be detailed in future revisions of this document.

3.14 Tunnel Applicability

This solution does not place any restrictions on the technology used to establish MDTs over the provider core network. Possible technologies include RSVP-TE P2MP LSP tunnels [P2MP-SIG], PIM-based tunnels, or GRE and IP-in-IP multicast tunnels.

4. IANA Considerations

[TBD]

5. Security Considerations

Since this document is based on [RFC2547bis] and [P2MP-RSVP], the security considerations involving those drafts apply here as well.

6. Acknowledgements

We would like to thank Antu Chatterjee and Masaaki TAKAGI for their suggestions and contributions to this draft

7. Intellectual Property Considerations

The IETF takes no position regarding the validity or scope of any

8. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3667] Bradner, S., "IETF Rights in Contributions", BCP 78, RFC 3667, February 2004.

[RFC3668] Bradner, S., Ed., "Intellectual Property Rights in IETF Technology", BCP 79, RFC 3668, February 2004.

[RFC2547bis] Rosen, E., et. al. "BGP/MPLS VPNs", draft-ietf-l3vpn-rfc2547bis, work in progress

[PIM-SM] B. Fenner, M. Handley, H. Holbrook, I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", draft-ietf-pim-sm-v2-new-08.txt, work in progress.

9. Informational References

[RFC3446] D. Kim, D. Meyer, H. Kilmer, D. Farinacci, "Anycast Rendevous Point (RP) mechanism using Protocol Independent Multicast (PIM) and Multicast Source Discovery Protocol (MSDP)", January 2003

[RFC2434] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP: 26, RFC 2434, October 1998.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3552] Rescorla E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP: 72, RFC 3552, July 2003.

[P2MP-REQ] S. Yasukawa, et. al., "Requirements for Point to Multipoint Traffic Engineered MPLS LSPs", draft-ietf-mpls-p2mp-requirement, work in progress.

[P2MP-RSVP] R. Aggarwal, et. al., "Extensions to RSVP-TE for Point to Multipoint TE LSPs", draft-raggarwa-mpls-rsvp-te-p2mp, work in progress.

[MDT-SAFI] Nalawade and Sreekantiah, "MDT SAFI", draft-nalawade-idr-mdt-safi-00.txt, February 2004

[MCAST-VPN] Y. Cai, E. Rosen, I. Wijnands, "Multicast in BGP/MPLS IP VPNs", <draft-rosen-vpn-mcast-07.txt>, May 2004.

[MVPN-RAGGARWA] R. Aggarwal, "Multicast in BGP/MPLS VPNs and VPLS", <draft-raggarwa-l3vpn-mvpn-vpls-mcast-00.txt>, February, 2004.

[MP2MP-TE-MPLS] Work in progress

[PIM-BSR] N.Bhaskar, A Gall and Stig Venaas,
"BootStrap Router(BSR) Mechanism for PIM",
<draft-ietf-pim-sm-bsr-04.txt>, July 2004.

10. Authors' Addresses

Seisho Yasukawa
NTT Corporation
9-11, Midori-Cho 3-Chome
Musashino-Shi, Tokyo 180-8585,
Japan
Phone: +81 422 59 4769
Email: yasukawa.seisho@lab.ntt.co.jp

Shankar Karuna
Motorola
Vanenburg IT park, Madhapur,
Hyderabad, India
Email: kshankar@motorola.com

Sarveshwar Bandi
Motorola
Vanenburg IT park, Madhapur,
Hyderabad, India
Email: sarvesh@motorola.com

Adrian Farrel
Old Dog Consulting
EMail: adrian@olddog.co.uk

11. Full Copyright Statement