Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 11, 2009

T. Morin, Ed.
France Telecom R&D
B. Niven-Jenkins, Ed.
BT
Y. Kamite
NTT Communications
R. Zhang
BT
N. Leymann
Deutsche Telekom
N. Bitar
Verizon

July 10, 2008

Considerations about Multicast for BGP/MPLS VPN Standardization
draft-morin-l3vpn-mvpn-considerations-03

Status of this Memo

Abstract

The current proposal for multicast in BGP/MPLS includes multiple alternative mechanisms for some of the required building blocks of the solution. The aim of this document is to leverage previously documented requirements to identify the key elements and help move forward solution design, toward the definition of a standard having a well defined set of mandatory procedures. The different proposed alternative mechanisms are examined in the light of requirements identified for multicast in L3VPNs, and suggestions are made about which of these mechanisms standardization should favor. Issues related to existing deployments of early implementations are also addressed.

Requirements Language

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].

Table of Contents

1. Introduction

The current proposal for multicast in BGP/MPLS [I-D.ietf-l3vpn-2547bis-mcast] includes multiple alternative mechanisms for some of the required building blocks of the solution. However, it does not identify the core set of mechanisms which must be implemented in order to ensure interoperability. This may lead to a situation where implementations may support different subsets of the available optional mechanisms leading to implementations that do not interoperate, which is a problem for the numerous operators ahaving multi-vendor backbones.

The aim of this document is to leverage the already expressed requirements [RFC4834] and study the properties of each approach, to identify mechanisms that are good candidates for being part of a core set of mandatory mechanisms which can be used to provide a base for interoperable solutions.

This document will go through the different building blocks of the solution and provide recommendations as to which mechanisms should be favored for each building block, while considering the requirements already defined and the goal of a fully-interoperable standard.

Considering the history of the multicast VPN proposals and implementations, the authors also consider it useful to discuss how existing deployments of early implementations [I-D.rosen-vpn-mcast][I-D.raggarwa-l3vpn-2547-mvpn] can fit in the picture, and provide suggestions in this respect.

[This document will evolve to follow key changes in multicast in BGP/MPLS [I-D.ietf-l3vpn-2547bis-mcast] and [I-D.ietf-l3vpn-2547bis-mcast-bgp]. Such changes are for instance, clear statements about compatibility between the different approaches and other optional features, or completed description of procedures that are not currently detailed.]

2. Terminology

Please refer to [I-D.ietf-l3vpn-2547bis-mcast] and [RFC4834].

3. Examining alternatives mechanisms for MVPN functions

3.1. MVPN auto-discovery

The current solution document [I-D.ietf-l3vpn-2547bis-mcast] proposes two different mechanisms for MVPN auto-discovery:

1. BGP-based auto-discovery

2. "PIM/shared tree" : discovery done through the exchange of PIM Hellos by C-PIM instances, accross an MI-PMSI implemented with one shared tree per VPN (using multicast ASM, or MP2MP LDP)

Both solutions address Section 5.2.10 of [RFC4834] which states that "the operation of a multicast VPN solution SHALL be as light as possible and providing automatic configuration and discovery SHOULD be a priority when designing a multicast VPN solution. Particularly the operational burden of setting up multicast on a PE or for a VR/VRF SHOULD be as low as possible".

The key consideration is that PIM-based discovery is only applicable to deployments using a shared tree to instantiate an MI-PMSI (it cannot be applicable to if only P2P or SSM trees are used, because contrary to ASM and MP2MP, building these P2P or SSM trees cannot happen before the autodiscovery has been done), whereas the BGP-based auto-discovery does not place any constraint on the type of multicast trees that would have to be used. BGP-based auto-discovery is independent of the type of P-multicast tree used thus satisfying the requirement in section 5.2.4.1 of [RFC4834] that "a multicast VPN solution SHOULD be designed so that control and forwarding planes are not interdependent".

Additionally, it is to be noted that a number of service providers have chosen to use SSM-based trees for the default MDTs within their current deployments, therefore relying already on some BGP-based auto-discovery.

Moreover, when shared P-tunnels are used, the use of BGP auto-discovery would allow inconsistencies in the addresses/identifiers used for the shared trees to be detected (e.g. the same shared tree identifier being used for different VPNs with distinct BGP route targets). This is particularly attractive in the context of inter-AS VPNs where the impact of any misconfiguration could be magnified and where a single service provider may not operate all the ASs. Note that this technique to detect some misconfiguration cases may not be usable during a transition period from a shared-tree autodiscovery to a BGP-based autodiscovery.

Thus, the recommendation is that implementation of the BGP-based auto-discovery is mandated and should be supported by all mVPN implementations (while PIM/shared-tree based auto-discovery should be optionally considered for migration purpose only).

3.2. S-PMSI Signaling

The current solution document [I-D.ietf-l3vpn-2547bis-mcast] proposes two mechanisms for signaling that multicast flows will be switched to an S-PMSI :

1. a UDP-based TLV protocol specifically for S-PMSI signaling (described in section 7.2.1).

2. a BGP-based mechanism for S-PMSI signaling (described in section 7.2.2).

Section 5.2.10 of [RFC4834] states that "as far as possible, the design of a solution SHOULD carefully consider the number of protocols within the core network: if any additional protocols are introduced compared with the unicast VPN service, the balance between their advantage and operational burden SHOULD be examined thoroughly". The UDP-based mechanism would be an additional protocol in the mvpn stack, which isn't the case for the BGP-based S-PMSI switching signaling, since (a) BGP is identified as a requirement for autodiscovery, and (b) the BGP-based S-PMSI switching signaling procedures are very similar to the autodiscovery procedures.

Furthermore, the BGP-based S-PMSI switching signaling mechanism can be used within MVPNs using either a UI-PMSI or a MI-PMSI while the UDP-based protocol is restricted to use within MVPNs using an MI-PMSI. In practice, this means that, except if shared trees are used, a PE will have to join to all trees of all PEs in a VPN, while in the alternative where BGP-based S-PMSI switching signaling is used, it could delay joining a tree from a PE until traffic from that PE is needed, thus reducing the amount of state maintained on P routers.

S-PMSI switching signaling approaches can also be compared in an inter-AS context (see Section 3.5). The proposed BGP-based approach for S-PMSI switching signaling provides a good fit with both the segmented and non-segmented inter-AS approaches (seeSection 3.5). By contrast the UDP-based approach for S-PMSI switching signaling appears to be usable with segmented inter-AS tunnels, but in that case key advantages of the segmented approach are lost :

- o there is no more an independence of ASes to choose when S-PMSIs tunnels will be triggered in their AS (and thus control the amount of state created on their P routers), and with which tunneling technique they will be built

- o in an inter-AS option B context, an isolation of ASes is obtained as PEs don't have visibility of, nor exchange with, PEs of other ASes. This property can be preserved if the segmented inter-AS

approach and BGP-based S-PMSI switching signaling are used, but it is not preserved if UDP-based switching signaling is used.

Given all the above, it is the recommendation of the authors that BGP is the preferred solution for S-PMSI switching signaling and should be supported by all implementations.

It is identified that, if nothing prevents a fast-paced creation of S-PMSI, then S-PMSI switching signaling with BGP would possibly impact the Route Reflectors used for mVPN routes. However is it also identified that such a fast-paced behavior would have an impact on P and PE routers resulting from S-PMSI tunnels signaling, which will be the same independently of the S-PMSI signaling approach that is used, and which it is certainly best to avoid by setting up proper mechanisms.

The UDP-based S-PMSI switching signaling protocol can also be considered, as an option, given that this protocol has been in deployment for some time. Implementations supporting both protocols would be expected to provide a per-VRF configuration knob to allow an implementation to use the UDP-based TLV protocol for S-PMSI switching signaling for specific VRFs in order to support the coexistence of both protocols (for example during migration scenarios). Apart from such migration-facilitating mechanisms, the authors specifically do not recommend extending the already proposed UDP-based TLV protocol to new types of P-multicast trees.

3.3. PE-PE Transmission of C-Multicast Routing

The current solution document [I-D.ietf-l3vpn-2547bis-mcast] proposes multiple mechanisms for PE-PE transmission of customer multicast routing information:

1. Full per-MVPN PIM peering across an MI-PMSI (described in section 3.4.1.1).

2. Lightweight PIM peering across an MI-PMSI (described in section 3.4.1.2)

3. The unicasting of PIM C-Join/Prune messages (described in section 3.4.1.3)

4. The use of BGP for carrying C-Multicast routing (described in section 3.4.2).

3.3.1. PE-PE signaling scalability

Scalability being one of the core requirements for multicast VPN, it is useful to compare the proposed C-multicast routing mechanisms from this perspective : Section 4.2.4 of [RFC4834] recommends that "a multicast VPN solution SHOULD support several hundreds of PEs per multicast VPN, and MAY usefully scale up to thousands" and section 4.2.5 states that "a solution SHOULD scale up to thousands of PEs having multicast service enabled".

Scalability with an increased number of VPNs per PE, or with an increased number of multicast state per VPN, are also important, but are not focused on in this section since we didn't identify differences between the different approaches for these matters : all others things equal, the load on PE due to C-multicast routing increases roughly linearly with the number of VPNs per PE, and with the number of multicast state per VPN.

This section thus presents conclusions related to PE-PE signaling scalability, while Appendix A contains more detailed explanations on the differences in ways of handling the C-multicast routing load, between the PIM-based approaches and the BGP-based approach, along with quantified evaluations of the amount of state and messages with the different approaches.

At high scales of multicast deployment, the first and third mechanisms require the PEs to maintain a large number of PIM adjacencies with other PEs of the same multicast VPN (which implies the regular exchange PIM Hellos with each other) and to refresh C-Join/Prune states, thus limiting the scalability of these approaches.

The third mechanism would reduce the amount of C-Join/Prune processing for a given multicast flow for PEs that are not the upstream neighbor for this flow, but would require "explicit tracking" state to be maintained by the upstream PE. It also isn't compatible with the "Join suppression" mechanism. A possible way to reduce the amount of signaling with this approach would be the use of a PIM refresh-reduction mechanism. Such a mechanism, based on TCP, is being considered by the PIM WG ([I-D.farinacci-pim-port]) ; its use in a multicast VPN context hasn't yet been described in [I-D.ietf-l3vpn-2547bis-mcast], but it is expected that this approach would provide a scalability similar with the BGP-based approach used without leveraging RR to process the PE-PE C-multicast routing. [TBC, when/if, this is further described in [I-D.ietf-l3vpn-2547bis-mcast]].

The second mechanism would operate in a similar manner to full per-

MVPN PIM peering except that PIM Hello messages are not transmitted and PIM C-Join/Prune refresh-reduction would be used, thereby improving scalability, but this approach has yet to be fully described. In any case, it seems that it only improves one thing among the things that will impact scalability with an increased number of PEs.

The first and second mechanisms can leverage the "Join suppression" behavior and thus improve the processing burden of an upstream PE, sparing the processing of a Join refresh message for each remote PE joined to a multicast stream. This improvement requires all PEs of a multicast VPN to process all PIM Join and Prune messages sent by any other PE participating in the same multicast VPN whether they are the upstream PE or not.

The fourth mechanism (the use of BGP for carrying C-Multicast routing) would have a comparable drawback of requiring all PEs to process a BGP C-multicast route only interesting a specific upstream PE. For this reason the C-multicast routing approach can leverage the Route-Target constraint mechanisms, which specifically allows only the interested upstream PE to receive a BGP C-multicast route. When RT-constraints are used the fourth mechanism reduces the total amount of message processing load put on the PEs for customer multicast routing to the minimum (by avoiding any processing by "unrelated" PEs, that are not the joining PE nor the upstream PE, and by avoiding the use of refreshes), and inherits BGP features that are expected to improve scalability (for instance, providing a means to offload some of the processing burden associated with client multicast routing onto one or many BGP route-reflectors). This advantage has a cost (the maintenance of a amount of state linear with the number of PEs joined to a stream), but when route reflectors are used, this cost is spread among the route reflectors.

However, the fourth mechanism is specific in that it offers the possibility of offloading customer multicast routing processing onto one or more BGP Route Reflector(s). When this is used, there is a drawback of increasing the processing load placed on the route reflector infrastructure. In the higher scale scenarios, it may be required to adapt the route relector infrastructure to the mVPN routing load by using, for example:

- a separation of resources for unicast and multicast VPN routing : using dedicated mVPN Route Reflector(s) (or using dedicated mVPN BGP sessions or dedicated mVPN BGP instances) ;

- the deployment of additional route reflector resources, for example increasing the processing resources on existing route reflectors or deployment of additional route reflectors.

Among the above, the most straightforward approach is to consider the introduction of route reflectors dedicated to the mVPN service and dimension them accordingly to the need of that service (but doing so is not required and is left as an operator engineering decision).

3.3.2. P-routers scalability

Mechanisms (1) and (2) are restricted to use within multicast VPNs that use an MI-PMSI, thereby necessitating:

the use of a P-multicast tree technique that allows shared trees (for example PIM-SM in ASM mode or MP2MP LDP)

or the use of one P-multicast tree per PE per VPN, even for PEs that do not have sources in their directly attached sites for that VPN.

By comparison, the fourth mechanism doesn't impose either of these restrictions, and when P2MP trees are used only necessitates the use of one tree per VPN per PE attached to a site with a multicast source or RP (or with a candidate BSR, if BSR is used).

In cases where there are less PEs connected with sources than the total amount of PEs, it improves the amount of state maintained by P-routers compared to the amount required to build an MI-PMSI with P2MP trees. Such cases are expected to be typical for multicast VPN deployments (see sections 4.2.4.1 of [RFC4834]).

3.3.3. Impact of C-multicast routing on Inter-AS deployments

Furthermore, co-existence with unicast inter-AS VPN options, and an equal level of security for multicast and unicast including in an inter-AS context, are specifically mentioned in sections 5.2.6, 5.2.8 and 5.2.12 of [RFC4834].

In an inter-AS option B context, an isolation of ASes is obtained as PEs don't have visibility of, nor exchange with, PEs of other ASes. This property can be preserved if the segmented inter-AS approach and BGP-based C-multicast routing is used, but it is not preserved if PIM-based signaling is used.

By comparison, the fourth option (the use of BGP for carrying C-Multicast routing) does not have any of the above limitations related to inter-AS deployments.

Additionally, the authors note that the proposed BGP-based approach for C-multicast routing provides a good fit with both the segmented and non-segmented inter-AS approaches. By contrast, though the PIM-

based C-multicast routing is usable with segmented inter-AS trees, the inter-AS scalability advantage of the approach is lost, since PEs in an AS will see the C-multicast routing activity of all other PEs of all other ASes.

3.3.4. Security and robustness

BGP supports MD5 authentication of its peers for additional security, thereby possibly benefit directly to multicast VPN customer multicast routing, whether for intra-AS or inter-AS communications. By contrast, with a PIM-based approach, no mechanism providing a comparable level of security to authenticate communications between remote PEs has been yet fully described yet [I-D.ietf-pim-sm-linklocal][], and in any case would require significant additional operations for the provider to be usable in a multicast VPN context.

The robustness of the infrastructure, especially the existing infrastructure providing unicast VPN connectivity, is key. The C-multicast routing function, especially under load, will compete with the unicast routing infrastructure. With the PIM-based approaches, the unicast and multicast VPN routing functions are expected to only compete in the PE, for control plane processing resources. In the case of the BGP-based approach, they will compete on the PE for processing resources, and in the route reflectors (supposing they are used for mVPN routing). It is identified that in both cases, mechanisms will be required to arbitrate resources (e.g. processing priorities). In the case of PIM-based procedures, between the different control plane routing instances in the PE. And in the case of the BGP-based approach, this is likely to require using distinct BGP sessions for multicast and unicast (e.g. through the use of dedicated mVPN BGP route reflectors, or to the use of a distinct session with an existing route reflector).

Multicast routing is dynamic by nature, and multicast VPN routing has to follow the VPN customers multicast routing events. The different approaches can be compared on how they are expected to behave in scenarios where multicast routing in the VPNs is subject to an intense activity. Scalability of each approach under such a load is detailed in Appendix A, and the fourth approach (BGP-based) is the only one having a O(1) cost for join/leave operations, and with which state maintenance is not concentrated on the upstream PE.

On the other hand, while the BGP-based approach is likely to suffer a slowdown under a load that is greater than the available processing resources (because of possibly congested TCP sockets), the PIM-based approaches would react to such a load by dropping messages, with failure-recovery obtained through message refreshes. Thus, the BGP-

based approach could result in a degradation of join/leave latency performance typically spread evenly across all multicast streams being joined in that period, while the PIM-based approach could result in increased join/leave latency, for some random streams, by a multiple of the time between refreshes (e.g. tens of seconds), and possibly in some states the adjacency may time-out resulting in disruption of multicast streams.

The behavior of the PIM-based approach under such a load is also harder to predict, given that the performance of the "Join suppression" mechanism (an important mechanism for this approach to scale) will itself be impeded by delays in Join processing. For these reasons, the BGP-based approach would be able to provide a smoother degradation and more predictable behavior under a highly dynamic load.

In fact, both an "evenly spread degradation" and an "unevenly spread larger degradation" can be problematic, and what seems important is the ability for the VPN backbone operator to (a) limit the amount of multicast routing activity that can be triggered by a multicast VPN customer, and to (b) provide the best possible independence between distinct VPNs. It seems that both of these can be addressed through local implementation improvements, and that both the BGP-based and PIM-based approaches could be engineered to provide (a) and (b). It can be noted though that the BGP approach proposes ways to dampen C-multicast route withdrawals and/or advertisements, and thus already describes a way to provide (a), while nothing comparable has yet been described for the PIM-based approaches (even though it doesn't appear difficult). The PIM-based approaches rely on a per VPN dataplane to carry the mVPN control plane, and thus may benefit from this first level of separation to solve (b).

3.3.5. C-multicast VPN join latency

Section 5.1.3 of [RFC4834] states that "the group join delay [...] is also considered one important QoS parameter. It is thus RECOMMENDED that a multicast VPN solution be designed appropriately in this regard". In a multicast VPN context, the "group join delay"of interest is the time between a CE sending a PIM Join to its PE and the first packet of the corresponding multicast stream being received by the CE.

It is to be noted that the C-multicast routing procedures will only impact the group join latency of a said multicast stream for the first receiver that is located across the provider backbone from the multicast source-connected PE (or the first <n> receivers in the specific case where a specific UMH selection algorithm is used, that allows <n> distinct UMH to be selected by distinct downstream PEs).

The different approaches proposed seem to have different characteristics in how they are expected to impact join latency:

- o the PIM-based approaches minimize the number of control plane processing hops between a new receiver-connected PE and the source-connected PE, and being datagram-based introduces minimal delay, thereby possibly having a join latency as good as possible depending on implementation efficiency

- o under degraded conditions (packet loss, congestion, high control plane load) the PIM-based approach may impact the latency for a given multicast stream in an all or nothing manner : if a C-multicast routing PIM Join packet is lost, latency can reach a high time (a multiple of the periodicity of PIM Join refreshes)

- o the BGP-based approach uses TCP exchanges, that may introduce an additional delay depending on BGP and TCP implementation, but which would typically result, under degraded conditions (such packet loss, congestion, high control plane load), in a comparably lower increase of latency spread more evenly across the streams

- o as shown in Appendix A, the BGP-based approach is particular in that it removes load from all the PEs (without putting this load on the upstream PE for a stream); this improvement of background load can bring improved performance when a PE acts as the upstream PE for a stream, and thus benefit join latency

This qualitative comparison of approaches shows that the BGP-based approach is designed for a smoother degradation of latency under degraded conditions such as packet loss, congestion, or high control plane load. On the other hand, the PIM-based approaches seem to structurally be able to reach the shorter "best-case" group join latency (especially compared to deployment of the BGP-based approach where route-reflectors are used).

Doing a quantitative comparison of latencies is not possible without referring to specific implementations and benchmarking procedures, and would possibly expose different conclusions, especially for best-case group join latency for which performance is expected vary with PIM and BGP implementations. We can also note that improving a BGP implementation for reduced latency of route processing would not only benefit multicast VPN group join latency, but the whole BGP-based routing, which means that the need for good BGP/RR performance is not specific to multicast VPN routing.

Last, C-multicast join latency will be impacted by the overall load put on the control plane, and the scalability of the C-multicast routing approach is thus to be taken into account. As explained in

sections Section 3.3.1 and Appendix A, the BGP-based approach will provide the best scalability with an increased number of PEs per VPN, thereby benefiting group join latency in such higher scale scenarios.

3.3.6. Extranet

An illustrative example of the benefit brought by using a C-multicast routing approach close to the technique for unicast VPN routing is how the "extranet" feature can be implemented : when BGP-based mechanisms are used, the already defined and well understood BGP route target import/export semantics are just reused and applied to BGP mVPN routes. By contrast, it is not specified how implementing the same feature would be done in the context of other C-multicast routing mechanisms, and thus unclear how this would bring a comparable consistency benefit, or if it is possible without significant engineering trade-offs given that their control plane is tied to a specific MI-PMSI tunnel. [to be updated when Extranet is described for approaches other than the BGP-based approaches]

Note that the support for the Extranet feature is stated as a MUST in sections 5.1.6 of [RFC4834].

3.3.7. Conclusion on C-multicast routing

The fourth approach (BGP-based) for customer multicast routing clearly presents some advantages over the PIM-based alternatives. However it has yet to be deployed within an operational mVPN, and only limited experience exists with its implementations. By contrast, PIM-based mechanisms lack many of these benefits and have identified limitations in how they can handle customer multicast routing load in higher-scale scenarios. Despite these, experience in multiple deployments shows that the "Full PIM peering" approach is operationally viable.

Consequently, at the present time and until there is experience with all of the proposed mechanisms it is not clear which of the above mechanisms should be recommended as the preferred solution to implementers. It would appear prudent for implementations to consider supporting both the fourth (BGP-based) and first (full per-MPVN PIM peering) mechanisms. Further experience on both implementations is likely to be required before some best practice can be defined.

The first mechanism (full per-MVPN PIM peering across an MI-PMSI) is the mechanism used by [I-D.rosen-vpn-mcast] and therefore it is deployed and operating in MVPNs today. The authors recognize that because full per-MVPN PIM peering has been in deployment for some time, the support for this mechanism may be helpful for backwards

compatibility and in order to facilitate migration towards the BGP-based approach.

Moreover to improve the clarity of the proposed specifications, if the hello suppression and refresh-reduction procedures are not fully specified and the benefit they can bring well identified, the authors would recommend that the proposals for lightweight PIM peering across an MI-PMSI (the second mechanism) and for the unicasting of PIM C-Join/Prune messages (the third mechanism) be removed from the final revision of [I-D.ietf-l3vpn-2547bis-mcast].

3.4. Encapsulation techniques for P-multicast trees

In this section the authors will not make any restricting recommendations since the appropriateness of a specific provider core data plane technology will depend on a large number of factors, for example the service provider's currently deployed unicast data plane, many of which are service provider specific.

However, implementations should not unreasonably restrict the data plane technology that can be used, and should not force the use of the same technology for different VPNs attached to a single PE. Initial implementations may only support a reduced set of encapsulation techniques and data plane technologies but this should not be a limiting factor that hinders future support for other encapsulation techniques, data plane technologies or interoperability.

Section 5.2.4.1 of [RFC4834] states "In a multicast VPN solution extending a unicast L3 PPVPN solution, consistency in the tunneling technology has to be favored: such a solution SHOULD allow the use of the same tunneling technology for multicast as for unicast. Deployment consistency, ease of operation and potential migrations are the main motivations behind this requirement."

Current unicast VPN deployments use a variety of LDP, RSVP-TE and GRE/IP-Multicast for encapsulating customer packets for transport across the provider core of VPN services. In order to allow the same encapsulations to be used for unicast and multicast VPN traffic, it is recommended that multicat VPN standards should recommend implementations to support for multicast VPNs, all the P2MP variants of the encapsulations and signaling protocols that they support for unicast and for which some multipoint extension is defined, such as mLDP, P2MP RSVP-TE and GRE/IP-multicast.

All three of the above encapsulation techniques support the building of P2MP multicast trees. In addition mLDP and GRE/IP-ASM-Multicast implementations may also support the building of MP2MP multicast

trees. The use of MP2MP trees may provide some scaling benefits to the service provider as only a single MP2MP tree need be deployed per VPN, thus reducing by an order of magnitude the amount of multicast state that needs to be maintained by P routers. This gain in state is at the expense of bandwidth optimization, since sites that do not have multicast receivers for multicast streams sourced behind a said PE group will still receive packets of such streams, leading to non-optimal bandwidth utilization across the VPN core. One thing to consider is that the use of MP2MP multicast tree will require additional configuration to define the same tree identifier or multicast ASM group address in all PEs (it has been noted that some auto-configuration could be possible for MP2MP trees, but this it is not currently supported by the auto-discovery procedures). [It has been noted that C-multicast routing schemes not covered in [I-D.ietf-l3vpn-2547bis-mcast] could expose different advantages of MP2MP multicast trees - this is out of scope of this document]

MVPN services can also be supported over a unicast VPN core through the use of ingress PE replication whereby the ingress PE replicates any multicast traffic over the P2P tunnels used to support unicast traffic. While this option does not require the service provider to modify their existing P routers (in terms of protocol support) and does not require maintaining multicast-specific state on the P routers in order for the service provider to be able deploy a multicast VPN service, the use of ingress PE replication obviously leads to non-optimal bandwidth utilization and it is therefore unlikely to be the long term solution chosen by service providers. However ingress PE replication may be useful during some migration scenarios or where a service provider considers the level of multicast traffic on their network to be too low to justify deploying multicast specific support within their VPN core.

All proposed approaches for control plane and dataplane can be used to provide aggregation amongst multicast groups within a VPN and amongst different multicast VPNs, and potentially reduce the amount of state to be maintained by P routers. However the latter -- the aggregation amongst different multicast VPNs will require support for upstream-assigned labels on the PEs. Support for upstream-assigned labels may require changes to the data plane processing of the PEs and this should be taken into consideration by service providers considering the use of aggregate S-PMSI tunnels for the specific platforms that the service provider has deployed.

3.5. Inter-AS deployments options

There are a number of scenarios that lead to the requirement for inter-AS multicast VPNs, including:

1. a service provider may have a large network that they have segmented into a number of ASs.

2. a service provider's multicast VPN may consist of a number of ASs due to acquisitions and mergers with other service providers.

3. a service provider may wish to interconnect their multicast VPN platform with that of another service provider.

The first scenario can be considered the "simplest" because the network is wholly managed by a single service provider under a single strategy and is therefore likely to use a consistent set of technologies across each AS.

The second scenario may be more complex than the first because the strategy and technology choices made for each AS may have been different due to their differing history and the service provider may not have (or may be unwilling to) unified the strategy and technology choices for each AS.

The third scenario is the most complex because in addition to the complexity of the second scenario, the ASs are managed by different service providers and therefore may be subject to a different trust model than the other scenarios.

Section 5.2.6 of [RFC4834] states that "a solution MUST support inter-AS multicast VPNs, and SHOULD support inter-provider multicast VPNs", "considerations about coexistence with unicast inter-AS VPN Options A, B and C (as described in section 10 of [RFC4364]) are strongly encouraged" and "a multicast VPN solution SHOULD provide inter-AS mechanisms requiring the least possible coordination between providers, and keep the need for detailed knowledge of providers' networks to a minimum - all this being in comparison with corresponding unicast VPN options".

Section 8 of [I-D.ietf-l3vpn-2547bis-mcast] addresses these requirements by proposing two approaches for mVPN inter-AS deployments:

1. Non-segmented inter-AS tunnels where the multicast tunnels are end-to-end across ASes, so even though the PEs belonging to a given MVPN may be in different ASs the ASBRs play no special role and function merely as P routers (described in section 8.1).

2. Segmented inter-AS tunnels where each AS constructs its own separate multicast tunnels which are then 'stitched' together by the ASBRs (described in section 8.2).

Section 5.2.6 of [RFC4834] also states "Within each service provider the service provider SHOULD be able on its own to pick the most appropriate tunneling mechanism to carry (multicast) traffic among PEs (just like what is done today for unicast)". The segmented approach is the only one capable of meeting this requirement.

The segmented inter-AS solution would appear to offer the largest degree of deployment flexibility to operators. However the non-segmented inter-AS solution can simplify deployment in a restricted number of scenarios and [I-D.rosen-vpn-mcast] only supports the non-segmented inter-AS solution and therefore the non-segmented inter-AS solution is likely to be useful to some operators for backward compatibility and during migration from [I-D.rosen-vpn-mcast] to [I-D.ietf-l3vpn-2547bis-mcast].

The applicability of segmented or non-segmented inter-AS tunnels to a given deployment or inter-provider interconnect will depend on a number of factors specific to each service provider. However, due to the additional deployment flexibility offered by segmented inter-AS tunnels, it is the recommendation of the authors that all implementations should support the segmented inter-AS model. Additionally, the authors recommend that implementations should consider supporting the non-segmented inter-AS model in order to facilitate co-existence with existing deployments, and as a feature to provide a lighter engineering in a restricted set of scenarios, although it is recognized that initial implementations may only support one or the other.

4. Co-located RPs

Section 5.1.10.1 of [RFC4834] states "In the case of PIM-SM in ASM mode, engineering of the RP function requires the deployment of specific protocols and associated configurations. A service provider may offer to manage customers' multicast protocol operation on their behalf. This implies that it is necessary to consider cases where a customer's RPs are outsourced (e.g., on PEs). Consequently, a VPN solution MAY support the hosting of the RP function in a VR or VRF."

However, customers who have already deployed multicast within their networks and have therefore already deployed their own internal RPs are often reluctant to hand over the control of their RPs to their service provider and make use of a co-located RP model, and providing RP-collocation on a PE will require the activation of MSDP or the processing of PIM Registers on the PE. Securing the PE routers for such activity requires special care, additional work, and will likely rely on specific features to be provided by the routers themselves.

The applicability of the co-located RP model to a given MVPN will thus depend on a number of factors specific to each customer and service provider.

It is therefore the recommendation that implementations should support a co-located RP model, but that support for a co-located RP model within an implementation should not restrict deployments to using a co-located RP model : implementations MUST support deployments when activation of a PIM RP function (PIM Register processing and RP-specific PIM procedures) or VRF MSDP instance is not required on any PE router and where all the RPs are deployed within the customers' networks or CEs.

5. Existing deployments

Some suggestions provided in this document can be used to incrementally modify currently deployed implementations without hindering these deployments, and without hindering the consistency of the standardized solution by providing optional per-VRF configuration knobs to support modes of operation compatible with currently deployed implementations, while at the same time using the recommended approach on implementations supporting the standard.

In cases where this may not be easily achieved, a recommended approach would be to provide a per-VRF configuration knob that allows incremental per-VPN migration of the mechanisms used by a PE device, which would allow migration with some per-VPN interruption of service (e.g. during a maintenance window).

Mechanisms allowing "live" migration by providing concurrent use of multiple alternatives for a given PE and a given VPN, is not seen as a priority considering the expected implementation complexity associated with such mechanisms. However, if there happen to be cases where they could be viably implemented relatively simply, such mechanisms may help improve migration management.

6. Summary of recommendations

The following list summarizes the authors' recommendations. These recommendations are not intended to prevent the implementation of alternative solutions, rather they are the authors' recommendations for the mechanisms that should be made mandatory in [I-D.ietf-l3vpn-2547bis-mcast] and therefore be supported by all implementations.

It is the authors' recommendation:

- that BGP-based auto-discovery be the mandated solution for auto-discovery ;

- that BGP be the mandated solution for S-PMSI switching signaling ;

- that implementations support both the BGP-based and the full per-MPVN PIM peering solutions for PE-PE transmission of customer multicast routing until further operational experience is gained with both solutions ;

- that implementations implement the P2MP variants of the P2P protocols that they already implement, such as mLDP, P2MP RSVP-TE and GRE/IP-Multicast ;

- that implementations support segmented inter-AS tunnels and consider supporting non-segmented inter-AS tunnels (in order to maintain backwards compatibility and for migration) ;

- implementations MUST support deployments when activation of a PIM RP function (PIM Register processing and RP-specific PIM procedures) or VRF MSDP instance is not required on any PE router.

7. IANA Considerations

This document makes no request to IANA.

[Note to RFC Editor: this section may be removed on publication as an RFC.]

8. Security Considerations

This document does not by itself raise any particular security considerations.

9. Acknowledgements

We would like to thank Adrian Farrel, Eric Rosen, Yakov Rekhter, and Maria Napierala for their feedback that helped shape this document.

10. Informative References

[RFC4834] Morin, T., "Requirements for Multicast in L3 Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4834, April 2007.

[I-D.ietf-l3vpn-2547bis-mcast]
Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-06 (work in progress), October 2006.

[I-D.ietf-l3vpn-2547bis-mcast-bgp]
Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-bgp-05 (work in progress), June 2008.

[I-D.rosen-vpn-mcast]
Rosen, E., "Multicast in MPLS/BGP VPNs", draft-rosen-vpn-mcast-08 (work in progress), December 2004.

[I-D.raggarwa-l3vpn-2547-mvpn]
Aggarwal, R., "Base Specification for Multicast in BGP/ MPLS VPNs", draft-raggarwa-l3vpn-2547-mvpn-00 (work in progress), June 2004.

[I-D.ietf-pim-sm-linklocal]
Atwood, J., "Authentication and Confidentiality in PIM-SM Link-local Messages", draft-ietf-pim-sm-linklocal-02 (work in progress), November 2007.

[I-D.farinacci-pim-port]
Farinacci, D., Wijnands, I., Karan, A., Boers, A., and M. Napierala, "A Reliable Transport Mechanism for PIM", draft-farinacci-pim-port-01 (work in progress), May 2008.

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

Appendix A. Scalability of C-multicast routing processing load

The main role of multicast routing is to let routers determine that they should start or stop forwarding a said multicast stream on a said link. In the multicast VPN context, this has to be made for each VPN, and the associated function is thus named "customer-multicast routing" or "C-multicast routing" and its role is to let PE routers determine that they should start of stop forwarding the traffic of a said multicast stream toward the remote PEs, on some S-PMSI tunnel.

When some "join" message is received by a PE, this PE knows that it should be sending traffic for the corresponding multicast group of

the corresponding VPN. But the reception of a "prune" message from a remote PE is not enough by itself for a PE to know that it should stop forwarding the corresponding multicast traffic : it has to make sure that they aren't any other PEs that still have receivers for this traffic.

There are many ways that the "C-multicast routing" building block can be designed so that a PE can determine when it can stop forwarding a said multicast stream toward other PEs:

PIM LAN Procedures, by default

By default when PIM LAN procedures are used, when a PE Prunes itself from a multicast tree, all other PEs check their own state to known if they are on the tree, in which case they send a PIM Join message to override the Prune. The "did the last receiver leave?" question is thus implicitly replied to by all PE routers, for each PIM Prune message.

PIM LAN Procedures, with explicit tracking :

PIM LAN procedures can use an "explicit tracking" approach, where a PE which is the upstream router for a multicast stream maintains an updated list of all neighbors who are joined to the tree. Thus, when it receives a Leave message from a PIM neighbor, it instantly knows the answer to the "did the last receiver leave?" question.

In this case, the question is replied to by the upstream router alone. The side effect of this "explicit tracking" is that "Join suppression" is not used : the downstream PEs will always send Joins toward the upstream PE, which will have to process them all.

BGP-based C-multicast routing

When BGP-based procedures are used for C-multicast routing, if no BGP route reflector is used, the "did the last receiver leave?" question is answered like in the PIM "explicit tracking" approach. But, when a BGP route reflector is used (which is expected to be the recommended approach), the role of maintaining an updated list of the PE part of a said multicast tree is taken care of by the route reflector(s). Using plain BGP route selection procedures, the route reflector will withdraw a C-multicast Source Tree Join for a said (C-S,C-G) when there is no PE advertising one anymore. In this context, the "did the last receiver leave?" question can be said to be answered by the route-reflector alone.

Furthermore, the BGP route distribution can leverage more than one route reflector : if a hierarchy of route reflectors is used, the "did the last receiver leave?" question is partly answered by each route reflector in the hierarchy.

We can see that answering the "last receiver leaves" question is a

significant proportion of the work that the C-multicast routing building block has to make, and where approaches differ most. The different approaches for handling C-multicast routing can result in a different amount of processing and how this processing is spread among the different functions. These differences can be better estimated by quantifying the amount of message processing and state maintenance.

Though the type of processing, messages and states, may vary with the different approaches, we propose here a rough estimation of the load of PEs, in terms of number of messages processed and number of control plane states maintained : a "message processed" being a message being parsed, a lookup being done, and some action being taken (such has updating a control plane or data plane state), and a "state maintained" being a multicast state kept in the control plane memory of a PE, related to a interface or a PE being subscribed to a multicast stream (we don't compare the data plane states on PE routers, which wouldn't vary between the different options chosen).

The following subsections do such an estimation for each proposed approach for C-multicast routing, for different phases of the following scenario:

o one SSM multicast stream is considered - scalability extrapolation to more than one stream is linear

o only the intra-AS case is concerned (with the segmented inter-AS trees and BGP-based C-multicast routing, #mvpn_PES and #joined_PEs should refer to the PEs of the mVPN in the AS, not to all PEs of the mVPN)

o the scenario is as follows:

   * one PE Joins the multicast stream (because of a new receiver-connected site has sent a Join on the PE-CE link), followed by additional PE that also join the multicast stream, one after the other ; we evaluate the processing required for the addition of each PE

   * some period of time T passes, without any PE joining or leaving (baseline)

   * all PE leaves, one after the other, until the last one leaves ; we evaluate the processing required for the leave of each PE

o the parameters used are:

* #mVPN_PEs : the number of PEs in the mVPN

* #R_PEs : the number of PEs joining the multicast stream

* #RRs : the number of route reflectors

* T_PIM_r : the time between two refreshes of a PIM Join (default is 60s)

The estimation unit used is the "message.equipment" or "m.e", one "message.equipment" being "one equipment processing one message" (10 m.e being "10 equipments processing each one message", or "5 messages each processed by 2 equipments", or "1 message processed by 10 equipment", etc.). Similarly for the amount of control plane state, we count in "state.equipment" or "s.e".

We distinguish three different types of equipments : the upstream PE for the multicast stream, the RR (if any), and the other PEs (which are not the upstream PE). The estimation is a total number of "message.equipment", for each type of equipment.

Additional precisions:

o for PIM, only Join and Prune messages are counted ; the PIM Hellos are not counted since these are not messages that trigger specific action in a typical scenario; message processing related to the PIM Assert mechanism is also not taken into account, because it is only active in transient state

o for BGP, only UPDATE message for mVPN route carrying C-multicast routing information are considered

A.1. PIM LAN procedures, by default

	upstream PE (1)	other PEs (#mvpn_PEs -1)	RR (none)	total
first PE joins	1 m.e	#mVPN_PEs-1 m.e	/	#mVPN_PEs m.e
for *each* additional PE joining	1 m.e	#mvpn_PEs-1 m.e	/	#mvpn_PEs m.e

baseline processing over a period T	T/T_PIMr m.e	(T/T_PIMr) . (#mvpn_PEs -1) m.e	/	(T/T_PIMr) x #mvpn_PEs m.e
for *each* PE leaving	2 m.e	2(#mvpn_PEs-1) m.e	/	2 x #mvpn_PEs m.e
the last PE leaves	1 m.e	#mvpn_PEs-1 m.e	/	#mvpn_PEs m.e
total for #R_PEs PEs	#R_PEs x 2 + T/T_PIMr m.e	(#mvpn_PEs-1) x (#R_PEs) x 2 + T/T_PIMr) . (#mvpn_PEs -1) m.e	0	#mvpn_PEs x (3 x #joined_PEs + T/T_PIMr) m.e
total state maintained	1 s.e	#joined_PE s.e	0	#R_PEs+1 s.e

Amount of messages processed for one multicast tree of one VPN - PIM LAN procedures, by default

We suppose here that the Join suppression and PIM Override mechanisms are fully effective, ie. that a Join sent by a PE is instantly seen by other PEs. Strictly speaking, this is not true, and depending on network delays and timing, there could be cases where more messages are exchanged.

A.2. PIM LAN procedures, with explicit tracking

	upstream PE (1)	other PEs (#mvpn_PEs -1)	RRs (none)	total
first PE joins	1 m.e	1 m.e (see note below)	/	2 m.e
for *each* additional PE joining	1 m.e	1 m.e (see note below)	/	2 m.e

baseline processing over a period T	(T/T_PIM) m.e x #R_PEs m.e	(T/T_PIMr) m.e (see note below)	/	(T/T_PIMres) x #R_PEs m.e
for *each* PE leaving	1 m.e	1 m.e (see note below)	/	2 m.e
the last PE leaves	1 m.e	1 m.e (see note below)	/	2 m.e
total for #R_PEs PEs	#R_PEs (2 + T/T_PIMr) m.e	#R_PEs x (2 + T/T_PIMr) m.e	0	#R_PEs x (4 + T/T_PIMr) m.e
total state maintained	#R_PEs s.e	#R_PEs s.e	0	2 x #R_PEs s.e

Amount of messages processed for one multicast tree of one VPN - PIM LAN procedures, with explicit tracking

Note: in this explicit tracking mode, a said Join or Leave message requires processing only by the upstream PE and the PE sending the message ; indeed, other PEs don't have any action to take ; it is to be noted though that these other PEs will still have to parse the PIM message, which is not non-zero processing. We make here the assumption that this is not significant.

A.3. BGP-based

About RR: we suppose that a message has to be processed by r BGP route reflectors to go from a receiver-connected PE to the source-connected PE. In practice, r depends on how RR are meshed, and would typically be small (max 1,2,3...), and r tends quickly toward 1 (as soon as there is a receiver-connected PEs in each RR cluster).

We make the assumption that RT constraint is used, if not the amount of state and message processing with this approach is similar to the PIM with explicit tracking approach, without the Joins refreshes.

	upstream PE (1)	other PEs (#mvpn_PEs -1)	RRs (#RRs)	total
first PE joins	1 m.e	1 m.e	r m.e	(r+2) m.e
for *each* additional PE joining	0	1 m.e	between 1 and r m.e	between 2 and (r+1) m.e
baseline processing over a period T	0	0	0	0
for *each* PE leaving	0	1 m.e	between 1 and r m.e	between 2 and (r+1) m.e
the last PE leaves	1 m.e	1 m.e	r m.e	(r+2) m.e
total for #R_PEs PEs	2 m.e	#R_PEs x 2 m.e	2 (r+#R_PEs) m.e	2 (2 x #R_PEs + r + 1) m.e
total state maintained	1 s.e	#R_PEs s.e	approx. #R_PEs x #RRs s.e	approx. (#R_PEs x (#RRs+1)) m.e

Amount of messages processed for one multicast tree of one VPN - BGP-based procedures

A.4. Side by side orders of magnitude comparison

This section concludes on the previous section by considering the orders of magnitude when the number of PE in a VPN increases.

	PIM LAN Procedures, default	PIM LAN Procedures, explicit tracking	BGP-based
first PE joins	O(#mVPN_PEs)	O(1)	O(1)
for *each* additional PE joining	O(#mVPN_PEs)	O(1)	O(1)
baseline processing over a period T	(T/T_PIMr) x O(#mvpn_PEs)	(T/T_PIMr) x O(#R_PEs)	0
for *each* PE leaving	O(#mVPN_PEs)	O(1)	O(1)
the last PE leaves	O(#mVPN_PEs)	O(1)	O(1)
total for #R_PEs PEs	O(#mVPN_PEs x #R_PEs) + O(#mVPN_PEs x T/T_PIMr)	O(#R_PEs) x (T/T_PIMr)	O(#R_PEs)
states	O(#R_PEs)	O(#R_PEs)	O(#R_PEs x #RRs)
notes	(processing and state maintenance are essentially done by, and spread amongst, the PEs of the mvpn ; non-upstream PEs have processing to do)	(processing and state maintenance is essentially done on the upstream PE)	(processing and state maintenance is essentially done by, and spread amongst, the RRs)

Amount of messages processed for one multicast tree of one VPN - PIM LAN procedures, with explicit tracking

The conclusions that can be drawn from the above are that:

o the PIM LAN Procedures default approach is particular in that all PEs, including those that are neither upstream nor downstream for a given message have processing to do, which results in a total amount of messages to process which is in O(#mVPN_PEs x #R_PEs), i.e. O(#mVPN_PEs ^ 2) if the proportion of R_PEs is considered constant when the number of PEs increases

o the two PIM-based approach do refreshes of Join messages, this is a linear factor not changing the order of magnitude, but which can be significant for long-lived streams

o the BGP-based approach requires an amount of message processing in O(#R_PEs), lower than the two other approaches, and which is independent of the duration of streams

o state maintenance is in the same order of magnitude for all approaches : O(#R_PEs), but the repartition is different:

   * the PIM LAN Procedure default approach fully spreads, and minimizes, the amount of state (one state per PE)

   * the PIM LAN procedure with explicit tracking, concentrate all state on the upstream PE

   * the BGP-based procedures spread all the state on the set of route reflectors

This quantification of message processing is based on a use case where each PE with a receiver joins and leave once. Drawing scalability-related conclusions for other patterns or frequency of changes of the set of receiver-connected PEs, requires considering the cost of each approach for "a new PE joining" and "a (non-last) PE leaving". From this perspective, the "PIM LAN Procedure default approach" is the most costly one (processing in O(#mVPN_PEs)), whereas the other approaches are in O(1) ; the "PIM LAN Procedures with explicit tracking" reduce the processing to the minimum in that case, the BGP-based approach having a cost increased by a linear factor depending on the number of RRs that will have to parse the message.

Appendix B. Switching to S-PMSI

[the following point was fixed in -07, and is here for reference only]

Section 7.2.2.3 of [I-D.ietf-l3vpn-2547bis-mcast] proposes two approaches for how a source PE can decide when to start transmitting

customer multicast traffic on a S-PMSI:

1. The source PE sends multicast packets for the <C-S, C-G> on both the I-PMSI P-multicast tree and the S-PMSI P-multicast tree simultaneously for a pre-configured period of time, letting the receiver PEs select the new tree for reception, before switching to only the S-PMSI.

2. The source PE waits for a pre-configured period of time after advertising the <C-S, C-G> entry bound to the S-PMSI before fully switching the traffic onto the S-PMSI-bound P-multicast tree.

The first alternative has essentially two drawbacks:

- <C-S,C-G> traffic is sent twice for some period of time, which would appear to be at odds with the motivation for switching to an S-PMSI in order to optimize the bandwidth used by the multicast tree for that stream.

- It is unlikely that the switchover can occur without packet loss or duplication if the transit delays of the I-PMSI P-multicast tree and the S-PMSI P-multicast tree differ.

By contrast, the second alternative has none of these drawbacks, and satisfy the requirement in section 5.1.3 of [RFC4834], which states that "[...] a multicast VPN solution SHOULD as much as possible ensure that client multicast traffic packets are neither lost nor duplicated, even when changes occur in the way a client multicast data stream is carried over the provider network". The second alternative also happen to be the one used in existing deployments.

For these reasons, it is the authors' recommendation to mandate the implementation of the second alternative for switching to S-PMSI.

Authors' Addresses

Thomas Morin (editor)
France Telecom R&D
2 rue Pierre Marzin
Lannion 22307
France

Email: thomas.morin@orange-ftgroup.com

Ben Niven-Jenkins (editor)
BT
208 Callisto House, Adastral Park
Ipswich, Suffolk IP5 3RE
UK

Email: benjamin.niven-jenkins@bt.com

Yuji Kamite
NTT Communications Corporation
Tokyo Opera City Tower
3-20-2 Nishi Shinjuku, Shinjuku-ku
Tokyo 163-1421
Japan

Email: y.kamite@ntt.com

Raymond Zhang
BT
2160 E. Grand Ave.
El Segundo CA 90025
USA

Email: raymond.zhang@bt.com

Nicolai Leymann
Deutsche Telekom
Goslarer Ufer 35
10589 Berlin
Germany

Email: nicolai.leymann@t-systems.com

Nabil Bitar
Verizon
40 Sylvan Road
Waltham, MA 02451
USA

Email: nabil.n.bitar@verizon.com