       Mandatory Features in a Layer 3 Multicast BGP/MPLS VPN Solution

Abstract

   More that one set of mechanisms to support multicast in a layer 3
   BGP/MPLS VPN has been defined. These are presented in the documents
   that define them as optional building blocks.

   To enable interoperability between implementations, this document
   defines a subset of features that is considered mandatory for a
   multicast BGP/MPLS VPN implementation. This will help implementers
   and deployers understand which L3VPN multicast requirements are best
   satisfied by each option.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

   Specifications for multicast in BGP/MPLS [RFC6513] include multiple
   alternative mechanisms for some of the required building blocks of
   the solution. However, they do not identify which of these
   mechanisms are mandatory to implement in order to ensure
   interoperability. Not defining a set of mandatory-to-implement
   mechanisms leads to a situation where implementations may support
   different subsets of the available optional mechanisms that do not
   interoperate, which is a problem for the numerous operators having
   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 that can be used to provide a base for
   interoperable solutions.

   This document goes through the different building blocks of the
   solution and concludes which mechanisms an implementation is required
   to implement. Section 7 summarizes these requirements.

   Considering the history of the multicast VPN proposals and
   implementations, it is also useful to discuss how existing
   deployments of early implementations [RFC6037] [MVPN] can be
   accommodated and provide suggestions in this respect.

2. Terminology

   Please refer to [RFC6513] and [RFC4834]. As a reminder, in Section
   3.1 of [RFC6513], the "C-" and "P-" notations are used to
   disambiguate between the provider scope and the scope of a specific
   VPN customer; for instance, "C-PIM" designates a PIM protocol
   instance in a VPN or VRF, while "P-PIM" would designate the instance
   of PIM eventually deployed by the provider across its core between P
   and PE routers.

   Other acronyms used in this document include the following:

   o LSP: Label Switched Path

   o P2MP: Point to Multipoint

   o MP2MP: Multipoint to Multipoint

   o GRE: Generic Routing Encapsulation

   o mLDP: Multicast LDP

   o I-PMSI: Inclusive Provider Multiservice Interface

   o MI-PMSI: Multidirectional Inclusive Provider Multiservice
      Interface

   o S-PMSI: Selective Provider Multiservice Interface

   o SSM: Source-Specific Multicast

   o ASM: Any-Source Multicast

   o PIM-SM: PIM Sparse Mode

   o PIM-SSM: PIM Sparse Mode in SSM Mode

   o BIDIR-PIM: Bidirectional PIM

   o AS: Autonomous System

   o ASBR: Autonomous System Border Router

   o VRF: VPN Routing and Forwarding

   o PE: Provider Edge

   o CE: Customer Edge

   o RPA: Rendezvous Point Address

   o RPL: Rendezvous Point Link

   Additionally, 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].

3. Examining Alternative Mechanisms for MVPN Functions

3.1. MVPN Auto-Discovery

   The current solution document [RFC6513] proposes two different
   mechanisms for Multicast VPN (MVPN) auto-discovery:

   1. BGP-based auto-discovery

   2. "PIM/shared P-tunnel": discovery done through the exchange of PIM
       Hellos by C-PIM instances, across an MI-PMSI implemented with one
       shared P-tunnel per VPN (using 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 P-tunnel to instantiate an MI-PMSI (it
is not applicable if only P2P, PIM-SSM, and P2MP LDP/RSVP-TE
P-tunnels are used, because contrary to ASM and MP2MP LDP, building
these types of P-tunnels cannot happen before the auto-discovery has
been done). In contrast, the BGP-based auto-discovery does not place
any constraint on the type of P-tunnel that would have to be used.
BGP-based auto-discovery is independent of the type of P-tunnel 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 P-tunnels for the default multicast
distribution trees 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 P-tunnel to be detected (e.g., the same shared
P-tunnel 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
ASes. Note that this technique to detect some misconfiguration cases
may not be usable during a transition period from a shared-P-tunnel
auto-discovery to a BGP-based auto-discovery.

Thus, the recommendation is that implementation of the BGP-based
auto-discovery is mandated and should be supported by all MVPN
implementations.

3.2. S-PMSI Signaling

The current solution document [RFC6513] proposes two mechanisms for
signaling that multicast flows will be switched to a Selective PMSI
(S-PMSI):

1. a UDP-based TLV protocol specifically for S-PMSI signaling
    (described in Section 7.4.2)

   2. a BGP-based mechanism for S-PMSI signaling (described in Section
       7.4.1)

   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
   auto-discovery and (b) the BGP-based S-PMSI switching signaling
   procedures are very similar to the auto-discovery procedures.

   Furthermore, the UDP-based S-PMSI switching signaling mechanism
   requires an MI-PMSI, while the BGP-based protocol does not. In
   practice, this mean that with the UDP-based protocol, a PE will have
   to join to all P-tunnels of all PEs in an MVPN, while in the
   alternative where BGP-based S-PMSI switching signaling is used, it
   could delay joining a P-tunnel rooted at 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 (see Section 3.5).
   By contrast, while the UDP-based approach for S-PMSI switching
   signaling appears to be usable with segmented inter-AS tunnels, key
   advantages of the segmented approach are lost:

   o ASes are no longer independent in their ability to choose when
      S-PMSIs tunnels will be triggered in their AS (and thus control
      the amount of state created on their P routers).

   o ASes are no longer independent in their ability to choose the
      tunneling technique for the P-tunnels used for an S-PMSI.

   o In an inter-AS option B context, an isolation of ASes is obtained
      as PEs in one AS don't have (direct) exchange of routing
      information with PEs of other ASes. This property is not
      preserved if UDP-based S-PMSI switching signaling is used. By
      contrast, BGP-based C-multicast switching signaling does preserve
      this property.

   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.

If nothing prevents a fast-paced creation of S-PMSI, then S-PMSI
switching signaling with BGP would possibly impact the route
reflectors (RRs) used for MVPN routes. However, 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 independent of the
S-PMSI signaling approach that is used and which 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 (VPN Routing and Forwarding)
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 co-existence 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-tunnels.

3.3. PE-PE Exchange of C-Multicast Routing

The current solution document [RFC6513] proposes multiple mechanisms
for PE-PE exchange of customer multicast routing information
(C-multicast routing):

   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 C-Multicast Routing 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 amount of multicast state per VPN, are also important but
are not focused on in this section since we didn't identify
differences between the various 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 amount of multicast state per VPN.

This section presents conclusions related to PE-PE C-multicast
routing scalability. Appendix A provides more detailed explanations
on the differences in ways PIM-based approaches and the BGP-based
approach handle the C-multicast routing load, along with quantified
evaluations of the amount of state and messages with the different
approaches. Many points made in this section are detailed in
Appendix A.1.

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 of PIM Hellos with each other) and to
periodically refresh C-Join/Prune states, resulting in an increased
processing cost when the number of PEs increases (as detailed in
Appendix A.1). The second approach is less subject to this, and the
fourth approach is not subject to this.

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 specified by the PIM IETF Working Group ([PIM-PORT]); its use
in a multicast VPN context is not described in [RFC6513], but it is
expected that this approach will provide a scalability similar to the
BGP-based approach without RRs.

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 when the number of PEs
increases.

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, Section 16 of [RFC6514] recommends the use of
the Route Target constrained BGP distribution [RFC4684] mechanisms,
which eliminate this drawback by having only the interested upstream
PE receive a BGP C-multicast route. Specifically, when Route Target
constrained BGP distribution is used, the fourth mechanism reduces
the total amount of the C-multicast routing processing load put on
the PEs by avoiding any processing of customer multicast routing
information on the "unrelated" PEs that are neither the joining PE
nor the upstream PE.

Moreover, the fourth mechanism further reduces the total amount of
message processing load by avoiding the use of periodic refreshes and
by inheriting BGP features that are expected to improve scalability
(for instance, providing a means to offload some of the processing
burden associated with customer multicast routing onto one or many
BGP route reflectors). The advantages of the fourth mechanism come
at a cost of maintaining an amount of state linear with the number of
PEs joined to a stream. However, the use of route reflectors allows
this cost to be spread among multiple route reflectors, thus
eliminating the need for a single route reflector to maintain all
this state.

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 reflector infrastructure to the MVPN
routing load by using, for example:

o 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), and

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

The most straightforward approach is to consider the introduction of
route reflectors dedicated to the MVPN service and dimension them

according to the need of that service (but doing so is not required
and is left as an operator engineering decision).

3.3.2. PE-CE Multicast Routing Exchange Scalability

The overhead associated with the PE-CE exchange of C-multicast
routing is independent of the choice of the mechanism used for the
PE-PE C-multicast routing. Therefore, the impact of the PE-CE
C-multicast routing overhead on the overall system scalability is
independent of the protocol used for PE-PE signaling and is therefore
not relevant when comparing the different approaches proposed for the
PE-PE C-multicast routing. This is true even if in some operational
contexts, the PE-CE C-multicast routing overhead is a significant
factor in the overall system overhead.

3.3.3. Scalability of P Routers

The first and second mechanisms are restricted to use within
multicast VPNs that use an MI-PMSI, thereby necessitating:

o the use of a P-tunnel technique that allows shared P-tunnels (for
   example, PIM-SM in ASM mode or MP2MP LDP), or

o the use of one P-tunnel 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 P-tunnels are used, only necessitates the
use of one P-tunnel per VPN per PE attached to a site with a
multicast source or Rendezvous Point (RP) (or with a candidate
Bootstrap Router (BSR), if BSR is used).

In cases where there are fewer PEs connected with sources than the
total number of PEs, the fourth mechanism improves the amount of
state maintained by P routers compared to the amount required to
build an MI-PMSI with P2MP P-tunnels. Such cases are expected to be
frequent for multicast VPN deployments (see Section 4.2.4.1 of
[RFC4834]).

3.3.4. Impact of C-Multicast Routing on Inter-AS Deployments

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 and 5.2.8 of [RFC4834].

In an inter-AS option B context, an isolation of ASes is obtained as
PEs in one AS don't have (direct) exchange of routing information
with PEs of other ASes. This property is not preserved if PIM-based

PE-PE C-multicast routing is used. By contrast, the fourth option
(BGP-based C-multicast routing) does preserve this property.

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 tunnels,
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.

3.3.5. Security and Robustness

BGP supports MD5 authentication of its peers for additional security,
thereby possibly directly benefiting 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 yet been fully described [RFC5796] 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). In both cases,
mechanisms will be required to arbitrate resources (e.g., processing
priorities). In the case of PIM-based procedures, this arbitration
occurs between the different control plane routing instances in the
PE. 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 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.2, and the fourth approach (BGP-based) used in
conjunction with the Route Target Constraint mechanisms [RFC4684] is
the only one having a cost for join/leave operations independent of
the number of PEs in the VPN (with one exception detailed in

   Appendix A.2) and state maintenance 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 timeout, 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 (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 data plane to
   carry the MVPN control plane and thus may benefit from this first
   level of separation to solve (b).

3.3.6. 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 Upstream Multicast Hop selection
algorithm is used, which allows <n> distinct PEs to be selected as
the Upstream Multicast Hop 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, introduce minimal
   delay, thereby possibly having a join latency as good as possible
   depending on implementation efficiency.

o Under degraded conditions (packet loss, congestion, or 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, which may introduce an
   additional delay depending on BGP and TCP implementation but which
   would typically result, under degraded conditions (such packet
   loss, congestion, or 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 benefits 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 to 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
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.

3.3.7. Conclusion on C-Multicast Routing

The first and fourth approaches are relevant contenders for
C-multicast routing. Comparisons from a theoretical standpoint lead
to identification of some advantages as well as possible drawbacks in
the fourth approach. Comparisons from a practical standpoint are
harder to make: since only reduced deployment and implementation
information is available for the fourth approach, advantages would be
seen in the first approach that has been applied through multiple
deployments and shown to be operationally viable.

Moreover, the first mechanism (full per-MVPN PIM peering across an
MI-PMSI) is the mechanism used by [RFC6037]; therefore, it is
deployed and operating in MVPNs today. The fourth approach may or
may not end up being preferred for a said deployment, but because the
first approach has been in deployment for some time, the support for
this mechanism will in any case be helpful to facilitate an eventual
migration from a deployment using mechanism close to the first
approach.

Consequently, at the present time, implementations are recommended to
support both the fourth (BGP-based) and first (full per-MVPN PIM
peering) mechanisms. Further experience on deployments of the fourth
approach is needed before some best practices can be defined. In the
meantime, this recommendation would enable a service provider to
choose between the first and the fourth mechanisms, without this
choice being constrained by vendor implementation choices. A service
provider can also take into account the peculiarities of its own
deployment context by pondering the weight of the different factors
into account.

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 layer 3 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 multicast VPN standards should recommend that
   implementations support multicast VPNs and 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 P-tunnels. In addition, mLDP and GRE/
   IP-ASM-Multicast implementations may also support the building of
   MP2MP multicast P-tunnels. The use of MP2MP P-tunnels may provide
   some scaling benefits to the service provider as only a single MP2MP
   P-tunnel 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 P-tunnel will require additional configuration to define
   the same P-tunnel identifier or multicast ASM group address in all
   PEs (it has been noted that some auto-configuration could be possible

for MP2MP P-tunnels, but this is not currently supported by the auto-
discovery procedures). (It has been noted that C-multicast routing
schemes not covered in [RFC6513] could expose different advantages of
MP2MP multicast P-tunnels; this is out of the 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 data plane 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 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 it has segmented
    into a number of ASes.

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

3. A service provider may wish to interconnect its 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 histories, and the service provider
may not have unified (or may be unwilling to unify) 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 ASes 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 co-existence 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 [RFC6513] 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 ASes, 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 that are then 'stitched' together by
    the ASBRs (described in Section 8.2).

(Note that an inter-AS deployment can alternatively rely on Option A
-- so-called "back-to-back" VRFs -- that option is not considered in
this section given that it can be used without any inter-AS-specific
mechanism.)

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. [RFC6037] only supports the non-segmented
inter-AS solution; therefore, the non-segmented inter-AS solution is
likely to be useful to some operators for backward compatibility and
during migration from [RFC6037] to [RFC6513].

The following is a comparison matrix between the "segmented inter-AS
P-tunnels" and "non-segmented inter-AS P-tunnels" approaches:

o Scalability for I-PMSIs: The "segmented inter-AS P-tunnels"
   approach is more scalable, because of the ability of an ASBR to
   aggregate multiple intra-AS P-tunnels used for I-PMSI within its
   own AS into one inter-AS P-tunnel to be used by other ASes. Note
   that the I-PMSI scalability improvement brought by the "segmented
   inter-AS P-tunnels" approach is higher when segmented P-tunnels
   have a granularity of source AS (see item below).

o Scalability for S-PMSIs: The "segmented inter-AS P-tunnels"
   approach, when used with the BGP-based C-multicast routing
   approach, provides flexibility in how the bandwidth/state trade-
   off is handled, to help with scalability. Indeed, in that case,
   the trade-off made for a said (C-S,C-G) in a downstream AS can be
   made more in favor of scalability than the trade-off made by the
   neighbor upstream AS, thanks to the ability to aggregate one or
   more S-PMSIs of the upstream AS in one I-PMSI tunnel in a
   downstream AS.

o Configuration at ASBRs: Depending on whether segmented P-tunnels
   have a granularity of source ASBR or source AS, the "segmented
   inter-AS P-tunnels" approach would require respectively the same
   or additional configuration on ASBRs as the "non-segmented
   inter-AS P-tunnels" approach.

o Independence of tunneling technology from one AS to another: The
   "segmented inter-AS P-tunnels" approach provides this; the "non-
   segmented inter-AS P-tunnels" approach does not.

o Facilitated coexistence with, and migration from, existing
   deployments and lighter engineering in some scenarios: The "non-
   segmented inter-AS P-tunnels" approach provides this; the
   "segmented inter-AS P-tunnels" approach does not.

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, given
the different elements reminded above, 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 coexistence with, and migration from,
existing deployments, and 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.

3.6. BIDIR-PIM Support

   In BIDIR-PIM, the packet-forwarding rules have been improved over
   PIM-SM, allowing traffic to be passed up the shared tree toward the
   RPA. To avoid multicast packet looping, BIDIR-PIM uses a mechanism
   called the designated forwarder (DF) election, which establishes a
   loop-free tree rooted at the RPA. Use of this method ensures that
   only one copy of every packet will be sent to an RPA, even if there
   are parallel equal cost paths to the RPA. To avoid loops, the DF
   election process enforces a consistent view of the DF on all routers
   on network segment, and during periods of ambiguity or routing
   convergence, the traffic forwarding is suspended.

   In the context of a multicast VPN solution, a solution for BIDIR-PIM
   support must preserve this property of similarly avoiding packet
   loops, including in the case where multicast VRFs in a given MVPN
   don't have a consistent view of the routing to C-RPL/C-RPA (Customer-
   RPL/Customer-RPA, i.e., RPL/RPA of a Bidir customer PIM instance).

   Section 11 of the current MVPN specification [RFC6513] defines three
   methods to support BIDIR-PIM, as RECOMMENDED in [RFC4834]:

   1. Standard DF election procedure over an MI-PMSI

   2. VPN Backbone as the RPL (Section 11.1)

   3. Partitioned Sets of PEs (Section 11.2)

   Method (1) is naturally applied to deployments using "Full per-MVPN
   PIM peering across an MI-PMSI" for C-multicast routing, but as
   indicated in [RFC6513], Section 11, the DF election may not work well
   in an MVPN environment, and an alternative to DF election would be
   desirable.

   The advantage of methods (2) and (3) is that they do not require
   running the DF election procedure among PEs.

   Method (2) leverages the fact that in BIDIR-PIM, running the DF
   election procedure is not needed on the RPL. This approach thus has
   the benefit of simplicity of implementation, especially in a context

where BGP-based C-multicast routing is used. However, it has the
drawback of putting constraints on how BIDIR-PIM is deployed, which
may not always match the requirements of MVPN customers.

Method (3) treats an MVPN as a collection of sets of multicast VRFs,
all PEs in a set having the same reachability information towards
C-RPA but distinct from PEs in other sets. Hence, with this method,
C-Bidir packet loops in MVPN are resolved by the ability to partition
a VPN into disjoint sets of VRFs, each having a distinct view of the
converged network. The partitioning approach to BIDIR-PIM requires
either upstream-assigned MPLS labels (to denote the partition) or a
unique MP2MP LSP per partition. The former is based on PE
Distinguisher Labels that have to be distributed using auto-discovery
BGP routes, and their handling requires the support for upstream
assigned labels and context label lookups [RFC5331]. The latter,
using MP2MP LSP per partition, does not have these constraints but is
restricted to P-tunnel types supporting MP2MP connectivity (such as
mLDP [RFC6388]).

This approach to C-Bidir can work with PIM-based or BGP-based
C-multicast routing procedures and is also generic in the sense that
it does not impose any requirements on the BIDIR-PIM service
offering.

Given the above considerations, method (3) "Partitioned Sets of PEs"
is the RECOMMENDED approach.

In the event where method (3) is not applicable (lack of support for
upstream assigned labels or for a P-tunnel type providing MP2MP
connectivity), then method (1) "Standard DF election procedure over
an MI-PMSI" and (2) "VPN Backbone as the RPL" are RECOMMENDED as
interim solutions, (1) having the advantage over (2) of not putting
constraints on how BIDIR-PIM is deployed and the drawbacks of only
being applicable when PIM-based C-multicast is used and of possibly
not working well in an MVPN environment.

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 Multicast
Source Discovery Protocol (MSDP) or the processing of PIM Registers
on the PE. Securing the PE routers for such activity requires
special care and 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 a VRF MSDP instance is not required on
any PE router and where all the RPs are deployed within the
customers' networks or CEs.

5. Avoiding Duplicates

It is recommended that implementations support the procedures
described in Section 9.1.1 of [RFC6513] "Discarding Packets from
Wrong PE", allowing fully avoiding duplicates.

6. 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 are 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.

7. Summary of Recommendations

The following list summarizes conclusions on the mechanisms that
define the set of mandatory-to-implement mechanisms in the context of
[RFC6513].

Note well that the implementation of the non-mandatory alternative
mechanisms is not precluded.

Recommendations are:

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

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

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

o that implementations use the "Partitioned Sets of PEs" approach
   for BIDIR-PIM support;

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

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

o that implementations MUST support deployments when the 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

o that implementations support the procedures described in Section
   9.1.1 of [RFC6513].

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.

   Additional credit is due to Maria Napierala for co-authoring
   Section 3.6 on BIDIR-PIM Support.

10. References

10.1. Normative References

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

   [RFC6513]   Rosen, E., Ed. and R. Aggarwal, Ed., "Multicast in MPLS/
               BGP IP VPNs", RFC 6513, February 2012.

   [RFC6514]   Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP
               Encodings and Procedures for Multicast in MPLS/BGP IP
               VPNs", RFC 6514, February 2012.

10.2. Informative References

   [MVPN]   Aggarwal, R., "Base Specification for Multicast in BGP/
               MPLS VPNs", Work in Progress, June 2004.

   [PIM-PORT] Farinacci, D., Wijnands, I., Venaas, S., and M.
               Napierala, "A Reliable Transport Mechanism for PIM", Work
               in Progress, October 2011.

   [RFC4364]   Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private
               Networks (VPNs)", RFC 4364, February 2006.

   [RFC4684]   Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk,
               R., Patel, K., and J. Guichard, "Constrained Route
               Distribution for Border Gateway Protocol/MultiProtocol
               Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual
               Private Networks (VPNs)", RFC 4684, November 2006.

   [RFC4834]   Morin, T., Ed., "Requirements for Multicast in Layer 3
               Provider-Provisioned Virtual Private Networks (PPVPNs)",
               RFC 4834, April 2007.

   [RFC5331]   Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream
               Label Assignment and Context-Specific Label Space",
               RFC 5331, August 2008.

   [RFC5796]   Atwood, W., Islam, S., and M. Siami, "Authentication and
               Confidentiality in Protocol Independent Multicast Sparse
               Mode (PIM-SM) Link-Local Messages", RFC 5796, March 2010.

   [RFC6037]   Rosen, E., Cai, Y., and IJ. Wijnands, "Cisco Systems'
               Solution for Multicast in BGP/MPLS IP VPNs", RFC 6037,
               October 2010.

   [RFC6388]   Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas,
               "Label Distribution Protocol Extensions for Point-to-
               Multipoint and Multipoint-to-Multipoint Label Switched
               Paths", RFC 6388, November 2011.

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 an MVPN context, this has to be done for each MVPN,
   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 or stop forwarding the traffic of a
   said multicast stream toward the remote PEs, on some PMSI tunnel.

   When a Join message is received by a PE, this PE knows that it should
   be sending traffic for the corresponding multicast group of the
   corresponding MVPN. However, 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 there 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, and they differ, among other things, in how a PE
   determines 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 on a LAN
      Prunes itself from a multicast tree, all other PEs on that LAN
      check their own state to known if they are on the tree, in which
      case they send a PIM Join message on that LAN to override the
      Prune. Thus, for each PIM Prune message, all PE routers on the
      LAN work to let the upstream PE determine the answer to the "did
      the last receiver leave?" question.

   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 by having the upstream PE maintain an up-to-
      date list of the PEs that are joined to the tree, thus making it
      possible to instantly know the answer to the "did the last
      receiver leave?" question whenever a PE leaves the said multicast
      tree.

      However, when a BGP route reflector is used (which is expected to
      be the recommended approach), the role of maintaining an updated
      list of the PEs that are part of a said multicast tree is taken
      care of by the route reflector(s). Using BGP procedures, a route
      reflector that had advertised a C-multicast Source Tree Join route
      for a said (C-S,C-G) to other route reflectors before will
      withdraw this route when there is no of its clients PEs

advertising this route anymore. Similarly, a route reflector that
had advertised this route to its client PEs before will withdraw
this route when its (other) client PEs and its route reflectors
peers are no longer advertising this route. In this context, the
"did the last receiver leave?" question can be said to be answered
by the route reflector(s).

Furthermore, the BGP route distribution can leverage more than one
route reflector: if multiple route reflectors are used with PEs
being distributed (as clients) among these route reflectors, the
"did the last receiver leave?" question is partly answered by each
of these route reflectors.

We can see that the "last receiver leaves" question is a part of the
work that the C-multicast routing building block has to address, and
the different approaches significantly differ. The different
approaches for handling C-multicast routing can indeed 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" is a message
being parsed, a lookup being done, and some action being taken (such
as, for instance, updating a control plane or data plane state or
discarding the information in the message). A "state maintained" is
a multicast state kept in the control plane memory of a PE, related
to an interface or a PE being subscribed to a multicast stream (note
that a state will be counted on an equipment as many times as the
number of protocols in which it is present, e.g., two times when
present both as a PIM state and a BGP route). Note that here we
don't compare the data plane states on PE routers, which wouldn't
vary between the different options chosen.

A.1. Scalability with an Increased Number of PEs

The following sections evaluate the processing and state maintenance
load for an increasingly high number of PEs in a VPN.

A.1.1. SSM Scalability

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.

o Only the intra-AS case is concerned (with the segmented inter-AS
   tunnels and BGP-based C-multicast routing, #mvpn_PE and #R_PE
   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
      a number of additional PEs that also join the same multicast
      stream, one after the other; we evaluate the processing
      required for the addition of each PE.

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

   * All PEs leave, 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_PE: the number of PEs in the MVPN

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

   * #RR: 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" corresponds to "one equipment processing one
message" (10 m.e being "10 equipments processing each one message",
"5 messages each processed by 2 equipments", or "1 message processed
by 10 equipment", etc.). Similarly, for the amount of control plane
state, the unit used is "state.equipment" or "s.e". This accounts
for the fact that a message (or a state) can be processed (or
maintained) by more than one node.

We distinguish three different types of equipments: the upstream PE
for the considered multicast stream, the RR (if any), and the other
PEs (which are not the upstream PE).

The numbers or orders of magnitude given in the tables in the
following subsections are totals across all equipments of a same
type, for each type of equipment, in the "m.e" and "s.e" units
defined above.

Additionally:

o For PIM, only Join and Prune messages are counted:

   * The load due to PIM Hellos can be easily computed separately
      and only depends on the number of PEs in the VPN.

   * Message processing related to the PIM Assert mechanism is also
      not taken into account, for the sake of simplicity.

o For BGP, all advertisements and withdrawals of C-multicast Source
   Tree Join routes are considered (Source-Active auto-discovery
   routes are not used in an SSM context); following the
   recommendation in Section 16 of [RFC6514], the case where the
   Route Target Constraint mechanisms [RFC4684] is not used is not
   covered.

(Note that for all options provided for C-multicast routing, the
procedures to set up and maintain a shortest path tree toward the
source of an SSM group are the same as the procedures used to set up
and maintain a shortest path tree toward an RP or a non-SSM source;
the results of this section are thus re-used in Appendix A.1.2.)

A.1.1.1. PIM LAN Procedures, by Default

	upstream PE (1)	other PEs (total across (#mvpn_PE-1) PEs)	RR (none)	total across all equipments
first PE joins	1 m.e	#mvpn_PE-1 m.e	/	#mvpn_PE m.e
for *each* additional PE joining	1 m.e	#mvpn_PE-1 m.e	/	#mvpn_PE m.e
baseline processing over a period T	T/T_PIM_r m.e	(T/T_PIM_r) . (#mvpn_PE-1) m.e	/	(T/T_PIM_r) x #mvpn_PE m.e
for *each* PE leaving	2 m.e	2(#mvpn_PE-1) m.e	/	2 x #mvpn_PE m.e
the last PE leaves	1 m.e	#mvpn_PE-1 m.e	/	#mvpn_PE m.e
total for #R_PE PEs	#R_PE x 2 + T/T_PIM_r m.e	(#mvpn_PE-1) x (#R_PE) x 2 + T/T_PIM_r) . (#mvpn_PE-1) m.e	0	#mvpn_PE x (3 x #R_PE + T/T_PIM_r) m.e
total state maintained	1 s.e	#R_PE s.e	0	#R_PE+1 s.e

Messages Processing and State Maintenance - PIM LAN Procedures, by
Default

We suppose here that the PIM Join suppression and Prune Override
mechanisms are fully effective, i.e., that a Join or Prune message
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, and the number given in
this table is a lower bound to the number of PIM messages exchanged.

A.1.1.2. BGP-Based C-Multicast Routing

   The following analysis assumes that BGP route reflectors (RRs) are
   used, and no hierarchy of RRs (note that the analysis also assumes
   that Route Target Constraint mechanisms are used).

   Given these assumptions, a message carrying a C-multicast route from
   a downstream PE would need to be processed by the RRs that have that
   PE as their client. Due to the use of Route Target Constraint
   mechanisms [RFC4684], these RRs would then send this message to only
   the RRs that have the upstream PE as a client. None of the other RRs
   and none of the other PEs will receive this message. Thus, for a
   message associated with a given MVPN, the total number of RRs that
   would need to process this message only depends on the number of RRs
   that maintain C-multicast routes for that MVPN and that have either
   the receiver-connected PE or the source-connected PE as their clients
   and is independent of the total number of RRs or the total number of
   PEs.

   In practice, for a given MVPN, a PE would be a client of just 2 RRs
   (for redundancy, an RR cluster would typically have 2 RRs).
   Therefore, in practice the message would need to be processed by at
   most 4 RRs (2 RRs if both the downstream PE and the upstream PE are
   the clients of the same RRs). Thus, the number of RRs that have to
   process a given message is at most 4. Since RRs in different RR
   clusters have a full Internal BGP (iBGP) mesh among themselves, each
   RR in the RR cluster that contains the upstream PE would receive the
   message from each of the RRs in the RR cluster that contains the
   downstream PE. Given 2 RRs per cluster, the total number of messages
   processed by all the RRs is 6.

   Additionally, as soon as there is a receiver-connected PE in each RR
   cluster, the number of RRs processing a C-multicast route tends
   quickly toward 2 (taking into account that a PE peering to RRs will
   be made redundant).

	upstream PE (1)	other PEs (total across (#mvpn_PE-1) PEs)	RRs (#RR)	total across all equipments
first PE joins	2 m.e	2 m.e	6 m.e	10 m.e
for *each* additional PE joining	between 0 and 2 m.e	2 m.e	(at most) 6 m.e tending toward 2 m.e	(at most) 10 m.e tending toward 4 m.e
baseline processing over a period T	0	0	0	0
for *each* PE leaving	between 0 and 2 m.e	2 m.e	(at most) 6 m.e tending toward 2	(at most) 10 m.e tending toward 4 m.e
the last PE leaves	2 m.e	2 m.e	6 m.e	10 m.e
total for #R_PE PEs	at most 2 x #RRs m.e (see note below)	#R_PE x 4 m.e	(at most) 6 x #R_PE m.e (tending toward 2 x #R_PE m.e)	at most 10 x #R_PE + 2 x #RRs m.e (tending toward 6 x #R_PE + #RRs m.e)
total state maintained	4 s.e	2 x #R_PE s.e	approx. 2 #R_PE + #RR x #clusters s.e	approx. 4 #R_PE + #RRx #clusters + 4 m.e

Message Processing and State Maintenance - BGP-Based Procedures

Note on the total of m.e on the upstream PE:

o There are as many "message.equipment"s on the upstream PE as the
   number of times the RRs of the cluster of the upstream PE need to
   re-advertise the C-multicast (C-S,C-G) route; such a re-
   advertisement is not useful for the upstream PE, because the
   behavior of the upstream PE for a said (VPN associated to the
   Route Target, C-S,C-G) will not depend on the precise attributes
   carried by the route (other than the Route Target, of course) but
   will happen in some cases due to how BGP processes these routes.
   Indeed, a BGP peer will possibly re-advertise a route when its
   current best path changes for the said NLRI if the set of
   attributes to advertise also changes.

o Let's look at the different relevant attributes and when they can
   influence when a re-advertisement of a C-multicast route will
   happen:

   * next-hop and originator-id: A new PE joining will not
      mechanically result in a need to re-advertise a C-multicast
      route because as the RR aggregates C-multicast routes with the
      same NLRI received from PEs in its own cluster (Section 11.4 of
      [RFC6514]), the RR rewrites the values of these attributes;
      however, the advertisements made by different RRs peering with
      the RRs in the cluster of the upstream PE may lead to updates
      of the value of these attributes.

   * cluster-list: The value of this attribute only varies between
      clusters, changes of the value of this attributes does not
      "follow" PE advertisements, and only advertisements made by
      different RRs may possibly lead to updates of the value of this
      attribute.

   * local-pref: The value of this attribute is determined locally;
      this is true both for the routes advertised by each PE (which
      could all be configured to use the same value) and for a route
      that results from the aggregation by an RR of the route with
      the same NLRI advertised by the PEs of his cluster (the RRs
      could also be configured to use a local pref independent of the
      local_pref of the routes advertised to him). Thus, this
      attribute can be considered to result in a need to re-advertise
      a C-multicast route.

   * Other BGP attributes do not have a particular reason to be set
      for C-multicast routes in intra-AS, and if they were, an RR
      (or, for attributes relevant for inter-AS, an ASBR) would also
      overwrite these values when aggregating these routes.

   o Given the above, for a said C-multicast Source Tree Join (S,G)
      NLRI, what may force an RR to re-advertise the route with
      different attributes to the upstream PE would be the case of an RR
      of another cluster advertising a route better than its current
      best route, because of the values of attributes specific to that
      RR (next-hop, originator-id, cluster-list) but not because of
      anything specific to the PEs behind that RR. If we consider our
      (#R_PE -1) joining a said (C-S,C-G), one after the other after the
      first PE joining, some of these events may thus lead to a re-
      advertisement to the upstream PE, but the number of times this can
      happen is at worse the number of RRs in clusters having receivers
      (plus one because of the possible advertisement of the same route
      by a PE of the local cluster).

   o Given that we look at scalability with an increased number of PEs
      in this section, we need to consider the possibility that all
      clusters may have a client PE with a receiver. We also need to
      consider that the two RRs of the cluster of the upstream PE may
      need to re-advertise the route. With this in mind, we know that
      2x#RRs is an upper bound to the number of updates made by RRs to
      the upstream PE, for the considered C-multicast route.

A.1.1.3. Side-by-Side Orders of Magnitude Comparison

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

	PIM LAN Procedures	BGP-based
first PE joins (in m.e)	O(#mvpn_PE)	O(1)
for *each* additional PE joining (in m.e)	O(#mvpn_PE)	O(1)
baseline processing over a period T (in m.e)	(T/T_PIM_r) x O(#mvpn_PE)	0
for *each* PE leaving (in m.e)	O(#mvpn_PE)	O(1)
the last PE leaves (in m.e)	O(#mvpn_PE)	O(1)
total for #R_PE PEs (in m.e)	O(#mvpn_PE x #R_PE) + O(#mvpn_PE x T/T_PIM_r)	O(#R_PE)
states (in s.e) notes	O(#R_PE) (processing and state maintenance are essentially done by, and spread amongst, the PEs of the MVPN; non-upstream PEs have processing to do)	O(#R_PE) (processing and state maintenance is essentially done by, and spread amongst, the RRs)

        Comparison of Orders of Magnitude for Message Processing and State
                 Maintenance (Totals across All Equipments)

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

o In the PIM-based approach, any message will be processed by all
   PEs, including those that are neither upstream nor downstream for
   the message; as a result, the total number of messages to process
   is in O(#mvpn_PE x #R_PE), i.e., O(#mvpn_PE ^ 2) if the proportion
   of receiver PEs is considered constant when the number of PEs
   increases. The refreshes of Join messages introduce 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_PE) lower than the PIM-based approach. The amount is
   independent of the duration of streams.

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

   * The PIM-based approach fully spreads, and minimizes, the amount
      of state (one state per PE).

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

A.1.2. ASM Scalability

The conclusions in Appendix A.1.1 are reused in this section, for the
parts that are common to the setup and maintenance of states related
to a source tree or a shared tree.

When PIM-SM is used in a VPN and an ASM multicast group is joined by
some PEs (#R_PEs) with some sources sending toward this multicast
group address, we can note the following:

PEs will generally have to maintain one shared tree, plus one source
tree for each source sending toward G; each tree resulting in an
amount of processing and state maintenance similar to what is
described in the scenario in Appendix A.1.1, with the same
differences in order of magnitudes between the different approaches
when the number of PEs is high.

An exception to this is when, for a said group in a VPN among the PIM
instances in the customer routers and VRFs, none would switch to the
shortest path tree (SPT) (SwitchToSptDesired always false): in that
case, the processing and state maintenance load is the one required
for maintenance of the shared tree only. It has to be noted that
this scenario is dependent on customer policy. To compare the
resulting load in that case, between PIM-based approaches and the

BGP-based approach configured to use inter-site shared trees, the
scenario in Appendix A.1.1 can be used with #R_PEs joining a (C-*,
C-G) ASM group instead of an SSM group, and the same differences in
order of magnitude remain true. In the case of the BGP-based
approach used without inter-site shared trees, we must take into
account the load resulting from the fact that to build the C-PIM
shared tree, each PE has to join the source tree to each source;
using the notations of Appendix A.1.1, this adds an amount of load
(total load across all equipments) that is proportional to #R_PEs and
the number of sources. The order of magnitude with an increasing
number of PEs is thus unchanged, and the differences in order of
magnitude also remain the same.

Additionally, to the maintenance of trees, PEs have to ensure some
processing and state maintenance related to individual sources
sending to a multicast group; the related procedures and behaviors
largely may differ depending on which C-multicast routing protocol is
used, how it is configured, how the multicast source discovery
mechanism is used in the customer VPN, and which SwitchToSptDesired
policy is used. However, the following can be observed:

o When BGP-based C-multicast routing is used:

   * Each PE will possibly have to process and maintain a BGP
      Source-Active auto-discovery route for (some or all) sources of
      an ASM group. The number of Source-Active auto-discovery
      routes will typically be one but may be related to the number
      of upstream PEs in the following cases: when inter-site shared
      trees are used and simultaneously more than one PE is used as
      the upstream PE for SPT (C-S,C-G) trees and when inter-site
      shared trees are used and there are multiple PEs that are
      possible upstream for this (S,G).

   * This results in message processing and state maintenance (total
      across all the equipments) linearly dependent on the number of
      PEs in the VPN (#mvpn_PE) for each source, independent of the
      number of PEs joined to the group.

   * Depending on whether or not inter-site shared trees are used,
      on the SwitchToSptDesired policy in the PIM instances in the
      customer routers and VRFs, and on the relative locations of
      sources and RPs, this will happen for all (S,G) of an ASM group
      or only for some of them and will be done in parallel to the
      maintenance of shared and/or source trees or at the first join
      of a PE on a source tree.

      o When PIM-based C-multicast routing is used, depending on the
         SwitchToSptDesired policy in the PIM instances in the customer
         routers and VRFs and depending on the relative locations of
         sources and RPs, there are:

         * Possible control plane state transitions triggered by the
            reception of (S,G) packets. Such events would induce
            processing on all PEs joined to G.

         * Possible PIM Assert messages specific to (S,G). This would
            induce a message processing on each PE of the VPN for each PIM
            Assert message.

   Given the above, the additional processing that may happen for each
   individual source sending to the group, beyond the maintenance of
   source and shared trees, does not change the order of magnitude
   identified above.

A.2. Cost of PEs Leaving and Joining

   The quantification of message processing in Appendix A.1.1 is done
   based on a use case where each PE with receivers has joined and left
   once. Drawing scalability-related conclusions for other patterns of
   changes of the set of receiver-connected PEs can be done by
   considering the cost of each approach for "a new PE joining" and "a
   PE leaving".

   For the "PIM LAN Procedure" approach, in the case of a single SSM or
   SPT tree, the total amount of message processing across all nodes
   depends linearly on the number of PEs in the VPN when a PE joins such
   a tree.

   For the "BGP-based" approach:

   o In the case of a single SSM tree, the total amount of message
      processing across all nodes is independent of the number of PEs,
      for "a new PE" joining and "a PE leaving"; it also depends on how
      route reflectors are meshed, but not on linear dependency.

   o In the case of an SPT tree for an ASM group, BGP has additional
      processing due to possible Source-Active auto-discovery routes:

      * When BGP-based C-multicast routing is used with inter-site
         shared trees, for the first PE joining (and the last PE
         leaving) a said SPT, the processing of the corresponding
         Source-Active auto-discovery routes results in a processing
         cost linearly dependent on the number of PEs in the VPN. For

subsequent PEs joining (and non-last PE leaving), there is no
processing due to advertisement or withdrawal of Source-Active
auto-discovery routes.

   * When BGP-based C-multicast routing is used without inter-site
      shared trees, the processing of Source-Active auto-discovery
      routes for an (S,G) happens independently of PEs joining and
      leaving the SPT for (S,G).

In the case of a new PE having to join a shared tree for an ASM group
G, we see the following:

o The processing due to the PE joining the shared tree itself is the
   same as the processing required to set up an SSM tree, as
   described before (note that this does not happen when BGP-based
   C-multicast routing is used without inter-site shared trees).

o For each source for which the PE joins the SPT, the resulting
   processing cost is the same as one SPT tree, as described before.

   * The conditions under which a PE will join the SPT for a said
      (C-S,C-G) are the same between the BGP-based with inter-site
      shared tree approach and the PIM-based approach, and depend
      solely on the SwitchToSptDesired policy in the PIM instances in
      the customer routers in the sites connected to the PE and/or in
      the VRF.

   * The conditions under which a PE will join the SPT for a said
      (C-S,C-G) differ between the BGP-based without inter-site
      shared trees approach and the PIM-based approach.

   * The SPT for a said (S,G) can be joined by the PE in the
      following cases:

      + as soon as one router, or the VPN VRF on the PE, has
         SwitchToSptDesired(S,G) being true

      + when BGP-based routing is used and configured to not use
         inter-site shared trees

   * Said differently, the only case where the PE will not join the
      SPT for (S,G) is when all routers in the sites of the VPN
      connected to the PE, or the VPN VRF itself, will never have
      SwitchToSptDesired(S,G) being true, with the additional
      condition that inter-site shared trees are used when BGP-based
      C-multicast routing is used.

Thus, when one PE joins a group G to which n sources are sending
traffic, we note the following with regards to the dependency of the
cost (in total amount of processing across all equipments) to the
number of PEs:

o In the general case (where any router in the site of the VPN
   connected to the PE, or the VRF itself, may have
   SwitchToSptDesired(S,G) being true):

   * For the "PIM LAN Procedure" approach, the cost is linearly
      dependent on the number of PEs in the VPN and linearly
      dependent on the number of sources.

   * For the "BGP-based" approach, the cost is linearly dependent on
      the number of sources, and, in the sub-case of the BGP-based
      approach used with inter-site shared trees, is also dependent
      on the number of PEs in the VPN only if the PE is the first to
      join the group or the SPT for some source sending to the group.

o Else, under the assumption that routers in the sites of the VPN
   connected to the PE, and the VPN VRF itself, will never have the
   policy function SwitchToSptDesired(S,G) being possibly true, then:

   * In the case of the PIM-based approach, the cost is linearly
      dependent on the number of PEs in the VPN, and there is no
      dependency on the number of sources.

   * In the case of the BGP-based approach with inter-site shared
      trees, the cost is linearly dependent on the number of RRs, and
      there is no dependency on the number of sources.

   * In the case of the BGP-based approach without inter-site shared
      trees, the cost is linearly dependent on the number of RRs and
      on the number of sources.

Hence, with the PIM-based approach, the overall cost across all
equipments of any PE joining an ASM group G is always dependent on
the number of PEs (same for a PE that leaves), while the BGP-based
approach has a cost independent of the number of PEs. An exception
is the first PE joining the ASM group for the BGP-based approach used
without inter-site shared trees; in that case, there is a dependency
with the number of PEs.

On the dependency with the number of sources, without making any
assumption on the SwitchToSptDesired policy on PIM routers and VRFs
of a VPN, we see that a PE joining an ASM group may induce a
processing cost linearly dependent on the number of sources. Apart
from this general case, under the condition where the

   SwitchToSptDesired is always false on all PIM routers and VRFs of the
   VPN, then with the PIM-based approach, and with the BGP-based
   approach used with inter-site shared trees, the cost in amount of
   messages processed will be independent of the number of sources (it
   has to be noted that this condition depends on customer policy).

Appendix B. Switching to S-PMSI

   (The following point was fixed in a draft version of the document
   that became [RFC6513] and is here for reference only.)

   In early versions of the document that became [RFC6513], two
   approaches were proposed 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 had essentially two drawbacks:

   o (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.

   o 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
   satisfies 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
   happens to be the one used in existing deployments.

   Consistent with this analysis, only the second alternative is
   discussed in [RFC6513].

Authors' Addresses

    Thomas Morin (editor)
    France Telecom - Orange
    2 rue Pierre Marzin
    Lannion 22307
    France
    EMail: thomas.morin@orange.com

    Ben Niven-Jenkins (editor)
    BT
    208 Callisto House, Adastral Park
    Ipswich, Suffolk IP5 3RE
    UK
    EMail: ben@niven-jenkins.co.uk

    Yuji Kamite
    NTT Communications Corporation
    Granpark Tower
    3-4-1 Shibaura, Minato-ku
    Tokyo 108-8118
    Japan
    EMail: y.kamite@ntt.com

    Raymond Zhang
    Alcatel-Lucent
    777 Middlefield Rd.
    Mountain View, CA 94043
    USA
    EMail: raymond.zhang@alcatel-lucent.com

    Nicolai Leymann
    Deutsche Telekom
    Winterfeldtstrasse 21-27
    10781 Berlin
    Germany
    EMail: n.leymann@telekom.de

    Nabil Bitar
    Verizon
    60 Sylvan Road
    Waltham, MA 02451
    USA
    EMail: nabil.n.bitar@verizon.com