L3VPN Working Group
Internet-Draft
Updates: 6513,6514 (if approved)
Intended status: Standards Track
Expires: April 18, 2015

E. Rosen, Ed.
Juniper Networks, Inc.
K. Subramanian
Cisco Systems, Inc.
J. Zhang
Juniper Networks, Inc.

October 15, 2014

Ingress Replication Tunnels in Multicast VPN
draft-rosen-l3vpn-ir-02

Abstract

RFCs 6513, 6514, and other RFCs describe procedures by which a Service Provider may offer Multicast VPN service to its customers. These procedures create point-to-multipoint (P2MP) or multipoint-to-multipoint trees across the Service Provider's backbone. One type of P2MP tree that may be used is known as an "Ingress Replication (IR) tunnel". In an IR tunnel, a parent node need not be "directly connected" to its child nodes. When a parent node has to send a multicast data packet to its child nodes, it does not use layer 2 multicast, IP multicast, or MPLS multicast to do so. Rather, it makes n individual copies, and then unicasts each copy, through an IP or MPLS unicast tunnel, to exactly one child node. While the prior MVPN specifications allow the use of IR tunnels, those specifications are not always very clear or explicit about how the MVPN protocol elements and procedures are applied to IR tunnels. This document updates RFCs 6513 and 6514 by adding additional details that are specific to the use of IR tunnels.

Status of This Memo

Table of Contents

1. Introduction

RFCs 6513, 6514, and others describe procedures by which a Service Provider (SP) may offer Multicast VPN (MVPN) service to its customers. These procedures create point-to-multipoint (P2MP) or multipoint-to-multipoint (MP2MP) tunnels, called "P-tunnels" (Provider-tunnels), across the SP's backbone network. Customer multicast traffic is carried through the P-tunnels.

A number of different P-tunnel technologies are supported. One of the supported P-tunnel technologies is known as "ingress replication" or "unicast replication". We will use the acronym "IR" to refer to this P-tunnel technology.

An IR P-tunnel is a P2MP tree, but a given node on the tree is not necessarily "directly attached" to its parent node or to its child nodes. To send a multicast data packet from a parent node to one of its child nodes, the parent node encapsulates the packet and then unicasts it (through a P2P or MP2P MPLS LSP or a unicast IP tunnel) to the child node. If a node on an IR tree has n child nodes, and has a multicast data packet that must be sent along the tree, the parent node makes n individual copies of the data packet, and then sends each copy, through a unicast tunnel, to exactly one child node. No lower layer multicast technology is used when sending traffic from a parent node to a child node; multiple copies of the packet may therefore be sent out a single interface.

With the single exception of IR, the P-tunnel technologies supported by the MVPN specifications are pre-existing IP multicast or MPLS multicast technologies. Each such technology has its own set of specifications, its own setup and maintenance protocols, its own syntax for identifying specific multicast trees, and its own procedures for enabling a router to be added to or removed from a particular multicast tree. For IR P-tunnels, on the other hand, there is no prior specification for setting up and maintaining the P2MP trees; the procedures and protocol elements used for setting up and maintaining the P2MP trees are specified in the MVPN specifications themselves, and all the signaling/setup is done by using the BGP A-D (Auto-Discovery) routes that are defined in [RFC6514]. (The unicast tunnels used to transmit multicast data from one node to another in an IR P-tunnel may of course have their own setup and maintenance protocols, e.g., [RFC5036], [RFC3209].)

Since the transmission of a multicast data packet along an IR P-tunnel is done by transmitting the packet through a unicast tunnel, previous RFCs sometimes speak of an IR P-tunnel as "consisting of" a set of unicast tunnels. However, that way of speaking is not quite accurate. For one thing, it obscures the fact that an IR P-tunnel is

really a P2MP tree, whose nodes must maintain multicast state in both the control and data planes. For another, it obscures the fact the unicast tunnels used by a particular IR P-tunnel need not be specific to that P-tunnel; a single unicast tunnel can carry the multicast traffic of many different IR P-tunnels (and can also carry unicast traffic as well).

In this document, we provide a clearer and more explicit conceptual model for IR P-tunnels, clarifying the relationship between an IR P-tunnel and the unicast tunnels that are used for data transmission along the IR P-tunnel.

RFC 6514 defines a protocol element called a "tunnel identifier", which for most P-tunnel technologies is used to identify a P-tunnel (i.e., to identify a P2MP or MP2MP tree). However, when IR P-tunnels are used, this protocol element does not identify an IR P-tunnel. In some cases it identifies one of the P-tunnel's constituent unicast tunnels, and in other cases it is not used to identify a tunnel at all. In this document, we provide an explicit specification for how IR P-tunnels are actually identified.

Some of the MVPN specifications use phrases like "join the identified P-tunnel", even though there has up to now not been an explicit specification of how to identify an IR P-tunnel, of how a router joins such a P-tunnel, or of how a router prunes itself from such a P-tunnel. In this document, we make these procedures more explicit.

RFC 6514 does provide a method for binding an MPLS label to a P-tunnel, but does not discuss the label allocation policies that are needed for correct operation when the P-tunnel is an IR P-tunnel. Those policies are discussed in this document.

This document does not provide any new protocol elements or procedures; rather it makes explicit just how a router is to use the protocol elements and procedures of [RFC6513] and [RFC6514] to identify an IR P-tunnel, to join an IR P-tunnel, and to prune itself from an IR P-tunnel. This document also discusses the MPLS label allocation policies that need to be supported when binding MPLS labels to IR P-tunnels, and the timer policies that need to be supported when switching a customer multicast flow from one P-tunnel to another. As the material in this document must be understood in order to properly implement IR P-tunnels, this document is considered to update [RFC6513] and [RFC6514]. This document also discusses the application of "seamless multicast" [SMLS-MC] and "extranet" [MVPN-XNET] procedures to IR P-tunnels.

This draft does not discuss the use of IR P-tunnels to support a VPN customer's use of BIDIR-PIM. [C-BIDIR-IR] explains how to adapt the

procedures of [RFC6513], [RFC6514], and [MVPN-BIDIR] so that a customer's use of BIDIR-PIM can be supported by IR P-tunnels.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL", when and only when appearing in all capital letters, are to be interpreted as described in [RFC2119].

2. What is an IR P-tunnel?

An IR P-tunnel is a P2MP tree. Its nodes are BGP speakers that support the MVPN procedures of [RFC6514] and related RFCs. In general, the nodes of an IR P-tunnel are either PE routers, ASBRs, or (if [SMLS-MC] is supported) ABRs. (MVPN procedures are sometimes used to support non-MVPN, or "global table" multicast; one way of doing this is defined in [SMLS-MC]. In such a case, IR P-tunnels can be used outside the context of MVPN.)

MVPN P-tunnels may be either "segmented" or "non-segmented" (as these terms are defined in [RFC6513] and [RFC6514]).

A "non-segmented" IR P-tunnel is a two-level P2MP tree, consisting only of a root node and a set of nodes that are children of the root node. When used in an MVPN context, the root is an ingress PE, and the child nodes of the root are the egress PEs.

In a segmented P-tunnel, IR may be used for some or all of the segments. If a particular segment of a segmented P-tunnel uses IR, then the root of that segment may have child nodes that are ABRs or ASBRs, rather than egress PEs.

As with any type of P2MP tree, each node of an IR P-tunnel holds "multicast state" for the P-tunnel. That is, each node knows the identity of its parent node on the tree, and each node knows the identities of its child nodes on the tree. In the MVPN specs, the "parent" node is also known as the "Upstream Multicast Hop" or "UMH".

What distinguishes an IR P-tunnel from any other kind of P2MP tree is the method by which a data packet is transmitted from a parent node to a child node. To transmit a multicast data packet from a parent node to a child node along a particular IR P-tunnel, the parent node does the following:

- o It labels the packet with a label (call it a "P-tunnel label") that the child node has assigned to that P-tunnel,

- o It then places the packet in a unicast encapsulation and unicasts the packet to the child node. That is, the parent node sends the

packet through a "unicast tunnel" to a particular child node. This unicast tunnel need not be specially created to be part of the IR P-tunnel; it can be any P2P or MP2P unicast tunnel that will get the packets from the parent node to the child node. A single such unicast tunnel may be carrying multicast data packets of several different P2MP trees, and may also be carrying unicast data packets.

The parent node repeats this process for each child node, creating one copy for each child node, and sending each copy through a unicast tunnel to corresponding child node. It does not use layer 2 multicast, IP multicast, or MPLS multicast to transmit packets to its child nodes. As a result, multiple copies of each packet may be sent out a single interface; this may happen, e.g., if that interface is the next hop interface, according to unicast routing, from the parent node to several of the child nodes.

Since data traveling along an IR P-tunnel is always unicast from parent node to child node, it can be convenient to think of an IR P-tunnel as a P2MP tree whose arcs are unicast tunnels. However, it is important to understand that the unicast tunnels need not be specific to any particular IR P-tunnel. If R1 is the parent node of R2 on two different IR P-tunnels, a single unicast tunnel from R1 to R2 may be used to carry data along both IR P-tunnels. All that is required is that when the data packets arrive at R2, R2 will see the "P-tunnel label" at the top of the packets' label stack; R2's further processing of the packets will depend upon that label. Note that the same unicast tunnel between R1 and R2 may also be carrying unicast data packets.

Typically the unicast tunnels are the Label Switched Paths (LSPs) that already exist to carry unicast traffic; either MP2P LSPs created by LDP [RFC5036] or P2P LSPs created by RSVP-TE [RFC3209]. However, any other kind of unicast tunnel may be used. A unicast tunnel may have an arbitrary number of intermediate routers; those routers do not maintain any multicast state for the IR P-tunnel, and in general are not even aware of its existence.

As with all other P-tunnel types, IR P-tunnels may be used as Inclusive P-tunnels or as Selective P-tunnels.

3. How are IR P-tunnels Identified?

There are four MVPN BGP route types in which P-tunnels can be identified: Intra-AS I-PMSI A-D routes, Inter-AS I-PMSI A-D routes, S-PMSI A-D routes, and Leaf A-D routes. (These route types are all defined in [RFC6514]).

Whenever it is necessary to identify a P-tunnel in a route of one of these types, a "PMSI Tunnel Attribute" (PTA) is added to the route. As defined in [RFC6514] section 5, the PTA contains four fields: "Tunnel Type", "MPLS Label", "Tunnel Identifier", and "Flags". [RFC6514] defines only one bit in the "Flags" field, the "Leaf Information Required" bit.

If a route identifies an IR P-tunnel, the "Tunnel Type" field of its

PTA is set to the value 6, meaning "Ingress Replication".

Most types of P-tunnel are associated with specific protocols that are used to set up and maintain tunnels of that type. For example, if the "Tunnel Type" field is set to 2, meaning "mLDP P2MP LSP", the associated setup protocol is mLDP [mLDP]. The associated setup protocol always has a method of identifying the tunnels that it sets up. For example, mLDP uses a "FEC element" to identify a tree. If the "Tunnel type" field is set to 3, meaning "PIM SSM Tree", the associated setup protocol is PIM, and "(S,G)" is used to identify the tree. In these cases, the "Tunnel Identifier" field of the PTA carries a tree identifier as defined by the setup protocol used for the particular tunnel type.

IR P-tunnels, on the other hand, are entirely setup and maintained by the use of BGP A-D routes, and are not associated with any other setup protocol. (The unicast tunnels used to transmit multicast data along an IR P-tunnel may have their own setup and maintenance protocols, of course.) Further, the identifier of an IR P-tunnel does not appear in the PTA at all. Rather, the P-tunnel identifier is in the "Network Layer Reachability Information" (NLRI) field of the A-D routes that are used to advertise and to setup the P-tunnel.

When an IR P-tunnel is identified in an S-PMSI A-D route, an Intra-AS I-PMSI A-D route, or an Inter-AS I-PMSI A-D route (we will refer to these three route types as "advertising A-D routes"), its identifier is hereby defined to be the NLRI of that route. See sections 4.1, 4.2, and 4.3 of [RFC6514] for the specification of these NLRIs. Note that the P-tunnel identifier includes the "route type" and "length" octets of the NLRI.

An advertising A-D route is considered to identify an IR P-tunnel only if it carries a PTA whose "Tunnel Type" field is set to "IR".

When an IR P-tunnel is identified in an S-PMSI A-D route or in an Inter-AS I-PMSI A-D route, the "Leaf Info Required" bit of the Flags field of the PTA MUST be set.

In an advertising A-D route:

o If the "Leaf Info Required" bit of the Flags field of the PTA is set, then the "Tunnel Identifier" field of the PTA has no significance whatsoever, and MUST be ignored upon reception.

Note that, per RFC6514, the length of the "Tunnel Identifier" field is variable, and is inferred from the length of the PTA. Even when this field is of no significance, its length MUST be the length of an IP address in the address space of the SP's backbone, as specified in section 4.2 of [RFC6515]. In this case, it is RECOMMENDED that it be set to a routable address of the router that constructed the PTA. (While it might make more sense to allow or even require the field to be omitted entirely, that might raise issues of backwards compatibility with implementations that were designed prior to the publication of this document.)

- If the "Leaf Info Required" bit is not set, the "Tunnel Identifier" field of the PTA does have significance, but it does not identify the IR P-tunnel. The use of the PTA's "Tunnel Identifier" field in this case is discussed in Section 5 of this document.

Note that according to the above definition, there is no way for two different advertising A-D routes (i.e., two advertising A-D routes with different NLRIs) to advertise the same IR P-tunnel. In the terminology of [RFC6513], an IR P-tunnel can instantiate only a single PMSI. If an ingress PE, for example, wants to bind two customer multicast flows to a single IR P-tunnel, it must advertise that tunnel in an I-PMSI A-D route or in an S-PMSI A-D route whose NLRI contains wildcards ([RFC6625]).

When an IR P-tunnel is identified in a Leaf A-D route, its identifier is the "route key" field of the route's NLRI. See section 4.4 of [RFC6514].

A Leaf A-D route is considered to identify an IR P-tunnel only if it carries a PTA whose "Tunnel Type" field is set to "IR". In this type of route, the "Tunnel Identifier" field of the PTA does have significance, but it does not identify the IR P-tunnel. The use of the PTA's "Tunnel Identifier" field in this case is discussed in Section 5.

4. How to Join an IR P-tunnel

The procedures for joining an IR P-tunnel depend upon whether the P-tunnel has been previously advertised, and if so, upon how the P-tunnel was advertised. Note that joining an unadvertised P-tunnel is only possible when using the "Global Table Multicast" procedures of [SMLS-MC].

4.1. Advertised P-tunnels

The procedures in this section apply when the P-tunnel to be joined has been advertised in an S-PMSI A-D route, an Inter-AS I-PMSI A-D route, or an Intra-AS I-PMSI A-D route.

The procedures for joining an advertised IR P-tunnel depend upon whether the A-D route that advertises the P-tunnel has the "Leaf Info Required" bit set in its PTA.

4.1.1. If the 'Leaf Info Required Bit' is Set

The procedures in this section apply when the P-tunnel to be joined has been advertised in a route whose PTA has the "Leaf Info Required Bit" set.

The router joining a particular IR P-tunnel must determine its UMH for that P-tunnel. If the route that advertised the P-tunnel contains a P2MP Segmented Next Hop Extended Community, the UMH is determined from the value of this community (see [SMLS-MC]). Otherwise the UMH is determined from the route's next hop (see [RFC6514]).

Once the UMH is determined, the router joining the IR P-tunnel originates a Leaf A-D route. The NLRI of the Leaf A-D route MUST contain the tunnel identifier (as defined in Section 3 above) as its "route key". The UMH MUST be identified by attaching an "IP Address Specific Route Target" (or an "IPv6 Address Specific Route Target") to the Leaf A-D route. The IP address of the UMH appears in the "global administrator" field of the Route Target (RT). Details can be found in [RFC6514] and [SMLS-MC].

The Leaf A-D route MUST also contain a PTA whose fields are set as follows:

- o The "Tunnel Type" field is set to "IR".
- o The "Tunnel Identifier" field is set as described in Section 5 of this document.
- o The "MPLS Label" field is set to a non-zero value. This is the "P-tunnel label". The value must be chosen so as to satisfy various constraints, as discussed in Section 6 this document.

4.1.2. If the 'Leaf Info Required Bit' is Not Set

The procedures in this section apply when the P-tunnel to be joined has been advertised in a route whose PTA does not have the "Leaf Info Required Bit" set. This can only be the case if the P-tunnel was advertised in an Intra-AS I-PMSI A-D route.

If an IR P-tunnel is advertised in the Intra-AS I-PMSI A-D routes originated by the PE routers of a given MVPN, the Intra-AS I-PMSI can be thought of as being instantiated by a set of IR P-tunnels. Each PE is the root of one such P-tunnel, and the other PEs are children of the root. A PE simultaneously joins all these P-tunnels by originating (if it hasn't already done so) an Intra-AS I-PMSI A-D route with a PTA whose fields are set as follows:

- o The "Tunnel Type" field is set to "IR".
- o The "Tunnel Identifier" field is set as described in Section 5 of this document.
- o The "MPLS Label" field MUST be set to a non-zero value. This label value will be used by the child node to associate a received packet with the I-PMSI of a particular MVPN. The MPLS label allocation policy must be such as to ensure that the binding from label to I-PMSI is one-to-one.

The NLRI and the RTs of the originated I-PMSI A-D route are set as specified in [RFC6514].

Note that if a set of IR P-tunnels is joined in this manner, the "discard from the wrong PE" procedures of [RFC6513] section 9.1.1 cannot be applied to that P-tunnel. Thus duplicate prevention on such IR P-tunnels requires the use of either Single Forwarder Selection ([RFC6513] section 9.1.2) or native PIM procedures ([RFC6513] section 9.1.3).

4.2. Unadvertised P-tunnels

In [SMLS-MC], a procedure is defined for "Global Table Multicast", in which a P-tunnel can be joined even if the P-tunnel has not been previously advertised. See the sections of that document entitled "Leaf A-D Route for Global Table Multicast" and "Constructing the Rest of the Leaf A-D Route". The route key of the Leaf A-D route has the form of the "S-PMSI Route-Type Specific NLRI" in this case, and that should be considered to be the P-tunnel identifier. Note that the procedure for finding the UMH is different in this case; the UMH is the next hop of the best UMH-eligible route towards the "ingress

PE". See the section of that document entitled "Determining the Upstream ABR/PE/ASBR (Upstream Node)".

5. The PTA's 'Tunnel Identifier' Field

If the "Tunnel Type" field of a PTA is set to "IR", its "Tunnel Identifier" field is significant only when one of the following two conditions holds:

- o The PTA is carried by a Leaf A-D route, or
- o The "Leaf Information Required" bit of the "Flags" field of the PTA is not set.

If one of these conditions holds, then the "Tunnel Identifier" field must contain a routable IP address of the originator of the route. (See [RFC6514] sections 9.2.3.2.1 and 9.2.3.4.1 for the detailed specification of the contents of this field.) This address is used by the UMH to determine the unicast tunnel that it will use in order to send data, along the IR P-tunnel identified by the route key, to the originator of the Leaf A-D route.

The means by which the unicast tunnel is determined from this IP address is outside the scope of this document. The means by which the unicast tunnel is set up and maintained is also outside the scope of this document.

Section 4 of [RFC6515] MUST be applied when a PTA is carried in a Leaf A-D route, and describes how to determine whether the "Tunnel Identifier" field carries an IPv4 or an IPv6 address.

If neither of the above conditions hold, then the "Tunnel Identifier" field is of no significance, and MUST be ignored upon reception.

6. The PTA's 'MPLS Label' Field

When a PTA is carried by an S-PMSI A-D route or an Inter-AS I-PMSI A-D route, and the "Tunnel Type" field is set to "IR", the "MPLS Label" field is of no significance. In this case, it SHOULD be set to zero upon transmission and MUST be ignored upon reception.

The "MPLS Label" field is significant only when the PTA appears either in a Leaf A-D route or in an Intra-AS I-PMSI A-D route that does not have the "Leaf Information Required" bit set. In these cases, the MPLS label is the label that the originator of the route is assigning to the IR P-tunnel(s) identified by the route's NLRI. (That is, the MPLS label assigned in the PTA is what we have called the "P-tunnel label".)

6.1. Leaf A-D Route Originated by an Egress PE

As previously stated, when a Leaf A-D route is used to join an IR P-tunnel, the "route key" of the Leaf A-D route is the P-tunnel identifier.

We now define the notion of the "root of an IR P-tunnel".

- o If the identifier of an IR P-tunnel is of the form of an S-PMSI NLRI, the "root" of the P-tunnel is the router identified in the "Originating Router's IP Address" field of that NLRI.

- o If the identifier of an IR P-tunnel is of the form specified in Section "Leaf A-D Route for Global Table Multicast" of [SMLS-MC], the "root" of the P-tunnel is the router identified in the "Ingress PE's IP Address" field of that NLRI.

- o If the identifier of an IR P-tunnel is of the form of an Intra-AS I-PMSI NLRI, the "root" of the P-tunnel is the router identified in the "Originating Router's IP Address" field of that NLRI.

- o If the identifier of an IR P-tunnel is of the form of an Inter-AS I-PMSI NLRI, the "root" of the P-tunnel is same as the identifier of the P-tunnel, i.e., the combination of an RD and an AS.

Note that if a P-tunnel is segmented, the root of the P-tunnel, by this definition, is actually the root of the entire P-tunnel, not the root of the local segment.

In order to apply the procedures of RFC 6513 Section 9.1.1 ("Discarding Packets from Wrong PE"), the following condition MUST be met by the MPLS label allocation policy:

> Suppose an egress PE originates two Leaf A-D routes, each with a different route key in its NLRI, and each with a PTA specifying a "Tunnel Type" of "IR". Thus each of the Leaf A-D routes identifies a different IR P-tunnel. Suppose further that each of those IR P-tunnels has a different root. Then the egress PE MUST NOT specify the same MPLS label in both PMSI Tunnel attributes.

That is, to apply the "Discarding Packets from the Wrong PE" duplicate prevention procedures ([RFC6513] section 9.1.1), the same MPLS label MUST NOT be assigned to two IR P-tunnels that have different roots.

If segmented P-tunnels are in use, the above rule is necessary but not sufficient to prevent a PE from forwarding duplicate data to the CEs. For various reasons, a given egress PE or egress ABR or egress

ASBR may decide to change its parent node, on a given segmented P-tunnel, from one router to another. It does this by changing the RT of the Leaf A-D route that it originated in order to join that P-tunnel. Once the RT is changed, there may be a period of time during which the old parent node and the new parent node are both sending data of the same multicast flow. To ensure that the egress node not forward duplicate data, whenever the egress node changes the RT that it attaches to a Leaf A-D route, it MUST also change the "MPLS Label" specified in the Leaf A-D route's PTA. This allows the egress router to distinguish between packets arriving on a given P-tunnel from the old parent and packets arriving on that same P-tunnel from the new parent. At any given time, a router MUST consider itself to have only a single parent node on a given P-tunnel, and MUST discard traffic that arrives on that P-tunnel from a different parent node.

If extranet functionality [MVPN-XNET] is not implemented in a particular egress PE, or if an egress PE is provisioned with the knowledge that extranet functionality is not needed, the PE may adopt the policy of assigning a label that is unique for the ordered triple <root, parent node, egress VRF>. This will enable the egress PE to apply the duplicate prevention procedures discussed above, and to determine the VRF to which an arriving packet must be directed.

However, this policy is not sufficient to support the "Discard Packets from the Wrong P-tunnel" procedures that are specified in [MVPN-XNET]. To support those procedures, the labels specified in the PTA of Leaf A-D routes originated by a given egress PE MUST be unique for the ordered triple <root, root RD, parent node>, where the "root RD" is taken from the RD field of the IR P-tunnel identifier. (All forms of IR P-tunnel identifier contain an embedded "RD" field.) This policy is also sufficient for supporting non-extranet cases, but in some cases may result in the use of more labels than the policy of the previous paragraph.

6.2. Leaf A-D Route Originated by an Intermediate Node

When a P-tunnel is segmented, there will be "intermediate nodes" (nodes that have a parent and also have children on the P-tunnel). Each intermediate node is a leaf node of an "upstream segment" and a parent node of a "downstream segment". The intermediate node "splices" together the two segments, so that data it receives on the upstream segment gets transmitted on the downstream segment. If either the upstream or downstream segments (or both) are instantiated by IR, the need to do this splicing places certain constraints on the MPLS label allocation policy.

6.2.1. Upstream and Downstream Segments are IR Segments

An intermediate node N (i.e., a node that has a parent and also has children) on an IR P-tunnel may originate a Leaf A-D route with a particular route key as a result of receiving a Leaf A-D route with that same route key. This will happen only if the received Leaf A-D route carries an IP address specific RT whose Global Administrator field identifies node N.

Suppose intermediate node N originates two Leaf A-D routes, one whose route key is K1, and one whose route key is K2, where K1 != K2. In general, the respective PTAs of these Leaf A-D routes MUST specify distinct non-zero MPLS labels, such that it is possible to map uniquely from the specified label value to a single IR P-tunnel (call this the "uniqueness rule"). There is one exception to this rule; the exception is specified below.

Consider the set of Leaf A-D routes with route key K1 or route key K2 such that:

- o N has received these Leaf A-D routes and has them currently installed.

- o Each of these Leaf A-D routes carries an IP Address Specific Route Target that identifies N in its Global Administrator field.

Now suppose that all the Leaf A-D routes in this set have the same originating router, and that the PTAs of these Leaf A-D routes all specify the same MPLS label. Suppose further that N's UMH for K1 is the same as N's UMH for K2. In this particular case, N MAY specify the same MPLS label in the PTA of Leaf A-D route it originates for K1 as in the PTA of he route it originates for K2. However, if at any future time these conditions no longer hold, N must reoriginate at least one of the Leaf A-D routes with a different label so that the "uniqueness rule" holds.

6.2.2. Only One Segment is IR

To handle the case where an intermediate node, call it N, is splicing together two P-tunnel segments, only one of which is IR, it is necessary to generalize the rules of the preceding sub-section.

Suppose N is a leaf node of two (upstream) P-tunnel segments, call them U1 and U2. Suppose also that N is a parent node of two (downstream) P-tunnel segments, call them D1 and D2. And suppose that N needs to splice U1 to D1, and U2 to D2.

To follow the uniqueness rule of Section 6.2.1 of this document, N must assign a different MPLS label to U1 than it assigns to U2. How this assignment is made depends, of course, on the control protocol used to set up U1 and U2.

There is one case in which the uniqueness rule need not be followed. Suppose that there is a node M such that (a) M is N's only child node on D1, and (b) M is N's only child node on D2. M will have advertised to N a label L1 bound to D1, and a label L2 bound to D2. If (and for as long as) L1==L2, then N MAY violate the uniqueness rule by advertising to its parent node for U1 the same MPLS label it advertises to its parent node for U2.

Section 6.2.1 of this document specifies in detail the way this requirement is applied when the upstream and downstream segments are all IR segments.

6.3. Intra-AS I-PMSI A-D Route

When a router joins a set of IR P-tunnels using the procedures of Section 4.1.2 of this document, the procedures of section 9.1.1 of [RFC6513] cannot be applied, no matter what the label allocation policy is. In this case, the ingress PE is the same as the UMH, but it is not possible to assign a label uniquely to a particular ingress PE or UMH. However, the label in the MPLS label field of the PTA MUST NOT appear in the MPLS label field of the PTA carried by any other route originated by the same router.

7. How A Child Node Prunes Itself from an IR P-tunnel

If a particular IR P-tunnel was joined via the procedures of Section 4.1.2 of this document, a router can prune itself from the P-tunnel by withdrawing the Intra-AS I-PMSI A-D route it used to join the P-tunnel. This is not usually done unless the router is removing itself entirely from a particular MVPN.

The procedures in the remainder of this section apply when a router joined a particular IR P-tunnel by originating a Leaf A-D route (as described in Section 4.1.1 or Section 4.2 of this document).

If a router no longer has a need to receive any multicast data from a given IR P-tunnel, it may prune itself from the P-tunnel by withdrawing the Leaf A-D route it used to join the tunnel. This is done, e.g., if the router no longer needs any of the flows traveling over the P-tunnel, or if all the flows the router does need are being received over other P-tunnels.

A router that is attached to a particular IR P-tunnel via a particular parent node may determine that it needs to stay joined to that P-tunnel, but via a different parent node. This can happen, for example, if there is a change in the Next Hop or the P2MP Segmented Next Hop Extended Community of the S-PMSI A-D route in which that P-tunnel was advertised. In this case, the router changes the Route Target of the Leaf A-D route it used to join the IR P-tunnel, so that the Route Target now identifies the new parent node.

A parent node must notice when a child node has been pruned from a particular tree, as this will affect the parent node's multicast data state. Note that the pruning of a child node may appear to the parent node as the explicit withdrawal of a Leaf A-D route, or it may appear as a change in the Route Target of a Leaf A-D route. If the Route Target of a particular Leaf A-D route previously identified a particular parent node, but changes so that it no longer does so, the effect on the multicast state of the parent node is the same as if the Leaf A-D route had been explicitly withdrawn.

8. Parent Node Actions Upon Receiving Leaf A-D Route

These actions are detailed in [RFC6514] and [SMLS-MC]. Two points of clarification are made:

- If a router R1 receives and installs a Leaf A-D route originated by router R2, R1's multicast state is affected only if the Leaf A-D route carries an "IP Address Specific RT" (or "IPv6 Address Specific RT") whose "global administrator" field identifies R1.

  (This is as specified in [RFC6514] and [SMLS-MC].) If a Leaf A-D route's RT does not identify R1, but then changes so that it does identify R1, R1 must take the same actions it would take if the Leaf A-D route were newly received.

- It is possible that router R1 will receive and install a Leaf A-D route originated by router R2, where:

  * the route's RT identifies R1,

  * the route's NLRI contains a route key whose first octet indicates that it is identifying a P-tunnel advertised in an S-PMSI A-D route,

  * R1 has neither originated nor installed any such S-PMSI A-D route.

If at some later time, R1 installs the corresponding S-PMSI A-D route, and the Leaf A-D route is still installed, and the Leaf A-D

route's RT still identifies R1, then R1 MUST follow the same procedures it would have followed if the S-PMSI A-D route had been installed before the Leaf A-D route was installed. (I.e., implementers must not assume that events occur in the "usual" or "expected" order.)

9. Use of Timers when Switching UMH

Suppose a child node has joined a particular IR P-tunnel via a particular UMH, and it now determines (for whatever reason) that it needs to change its UMH on that P-tunnel. It does this by modifying the RT of a Leaf A-D route.

It is desirable for such a "switch of UMH" to be done using a "make before break" technique, so that the older UMH does not stop transmitting the packets on the given P-tunnel to the child until the newer UMH has a chance to start transmitting the packets on the given P-tunnel to the child. However, the control plane operation (modifying the RT of the Leaf A-D route) does not permit the child node to first join the P-tunnel at the new UMH, and then later prune itself from the old UMH; a single control plane operation has both effects. Therefore, to achieve "make before break", timers must be used as follows:

1. The old UMH must continue transmitting to the child node for a period of time after it sees the child's Leaf A-D route being withdrawn (or its RT changing to identify a different UMH).

2. The child node must continue to accept packets from the old UMH for a period of time before it starts to accept packets from the new UMH (and discard packets from the old).

Further, the timer in 1 should be longer than the timer in 2. This allows the child to switch from one UMH to another without any loss of data.

10. IANA Considerations

This document contains no actions for IANA.

11. Acknowledgments

The authors wish to thank Yakov Rekhter for his contributions to this work. We also wish to thank Huajin Jeng and Samir Saad for their contributions, and to thank Thomas Morin for pointing out some of the issues that needed further elaboration.

Section 6.1 discusses the importance of having an MPLS label allocation policy that, when ingress replication is used, allows an egress PE to infer the identity of a received packet's ingress PE. This issue was first raised in earlier work by Xu Xiaohu.

12. Security Considerations

No security considerations are raised by this document beyond those already discussed in [RFC6513] and [RFC6514].

13. References

13.1. Normative References

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

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

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

[RFC6515] Aggarwal, R. and E. Rosen, "IPv4 and IPv6 Infrastructure Addresses in BGP Updates for Multicast VPN", RFC 6515, February 2012.

13.2. Informative References

[C-BIDIR-IR]
Zhang, J., Rekhter, Y., and A. Dolganow, "Simulating 'Partial Mesh of MP2MP P-Tunnels' with Ingress

Replication", internet-draft draft-ietf-l3vpn-mvpn-bidir-ingress-replication-01, July 2014.

[MVPN-BIDIR]
Rosen, E., Wijnands, IJ., Cai, Y., and A. Boers, ""MVPN: Using Bidirectional P-Tunnels", internet-draft draft-ietf-l3vpn-mvpn-bidir-08, June 2014.

[MVPN-XNET]
Rekhter, Y. and E. Rosen, "Extranet Multicast in BGP/IP MPLS VPNs", internet-draft draft-ietf-l3vpn-mvpn-extranet-05, July 2014.

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

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC6625] Rosen, E., Rekhter, Y., Hendrickx, W., and R. Qiu, "Wildcards in Multicast VPN Auto-Discovery Routes", RFC 6625, May 2012.

[SMLS-MC] Rekhter, Y., Aggarwal, R., Morin, T., Grosclaude, I., Leymann, N., and S. Saad, "Inter-Area P2MP Segmented LSPs", internet-draft draft-ietf-mpls-seamless-mcast-14, June 2014.

Authors' Addresses

Eric C. Rosen (editor)
Juniper Networks, Inc.
10 Technology Park Drive
Westford, Massachusetts 01886
US

Email: erosen@juniper.net

Karthik Subramanian
Cisco Systems, Inc.
170 Tasman Drive
San Jose, California 95134
US

Email: kartsubr@cisco.com

Jeffrey Zhang
Juniper Networks, Inc.
10 Technology Park Drive
Westford, Massachusetts 01886
US

Email: zzhang@juniper.net