Network Working Group R. Aggarwal
Internet Draft Juniper Networks
Expiration Date: January 2006

Y. Kamite
NTT Communications

Luyuan Fang
AT&T

July 2005

Propagation of VPLS IP Multicast Group Membership Information

draft-raggarwa-l2vpn-vpls-mcast-ctrl-00.txt

Status of this Memo

Abstract

The PEs participating in VPLS need to learn the IP multicast group membership information from remote VPLS sites to enable them to send an IP multicast packet to only those other PEs in the VPLS that have receivers interested in that particular IP multicast packet's multicast source and group. This document describes procedures for propagating multicast control information, learned from local

Private LAN Service (VPLS) sites, to remote VPLS sites. IGMP or PIM snooping is required only on the customer facing interfaces. The procedures do not require IGMP or PIM snooping on the Service Provider backbone links. Instead they use reliable protocol messages to exchange multicast control information between the PEs.

Table of Contents

1. Specification of requirements

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

2. Contributors

Rahul Aggarwal
Yakov Rekhter
Juniper Networks
Yuji Kamite
NTT Communications
Luyuan Fang
AT&T
Chaitanya Kodeboniya
Juniper Networks

3. Introduction

[VPLS-BGP] and [VPLS-LDP] describe a solution for VPLS multicast that relies on ingress replication. [VPLS-MCAST] describes procedures for VPLS multicast that enable the use of multicast trees in the service provider (SP) network.

Irrespective of whether ingress replication or multicast trees are used for sending IP multicast traffic in a VPLS, the PEs participating in VPLS need to learn the IP multicast group membership information from remote VPLS sites to enable them to send an IP multicast packet to only those other PEs in the VPLS that have receivers interested in that particular IP multicast packet's multicast source and group.

By appropriate IGMP or PIM snooping it is possible for the ingress PE to send an IP multicast packet in a VPLS only to the egress PEs that have the receivers for that traffic, rather than to all the PEs in the VPLS instance. While PIM/IGMP snooping allows to avoid the situation where an IP multicast packet is sent to PEs with no receivers, there is a cost for this optimization. Namely, 1) A PE has to maintain (S,G) state for all the (S,G) of all the VPLSs present on the PE. 2) PIM snooping has to be done not only on the CE-PE interfaces, but on Pseudo-Wire (PW) interfaces as well, which in turn introduces a non-negligeable overhead on the PE. It is desirable to reduce this overhead when IGMP/PIM snooping is used.

This document describes procedures for propagating IP multicast group membership information, learned from local Virtual Private LAN Service (VPLS) sites, to remote VPLS sites. IGMP or PIM snooping is required only on the customer facing interfaces. The procedures do not require IGMP or PIM snooping on the Service Provider backbone links. Instead they use reliable protocol messages to exchange multi-

cast control information between the PEs.

This document uses the prefix 'C' to refer to the customer control or data packets and 'P' to refer to the provider control or data packets.

4. Propagating Multicast Control Information

PEs participating in VPLS need to learn the <C-S, C-G> information for two reasons:

1. With ingress replication [VPLS-BGP, VPLS-LDP], this allows a PE to send the IP multicast packet for a <C-S, C-G> only to other PEs in the VPLS instance, that have receivers interested in that particular <C-S, C-G>. This eliminates flooding.

2. It allows the construction of Aggregate Data Trees [VPLS-MCAST].

There are two components for a PE to learn the <C-S, C-G> information in a VPLS:

1. Learning the <C-S, C-G> information from the locally homed Virtual Switch Instances (VSIs).
2. Learning the <C-S, C-G> information from the remote VSIs.

4.1. IGMP/PIM Snooping

In order to learn the <C-S, C-G> information from the locally homed VSIs a PE needs to implement IGMP/PIM snooping on the PE-CE interfaces. This is because there is no PIM adjacency between the locally homed CEs and the PE. IGMP/PIM snooping has to be used to build the database of C-Joins that are being sent by the customer for a particular VSI. This also requires a PE to create a IGMP/PIM instance per VSI for which IGMP/PIM snooping is used. This instance is analogous to the multicast VRF PIM instance that is created for Multicast Virtual Private Networks (MVPNs) [MVPN].

It is conceivable that IGMP/PIM snooping can be used to learn <C-S, C-G> information from remote VSIs by snooping VPLS traffic received over the SP backbone. However IGMP/PIM snooping is computationally expensive. Furthermore the periodic nature of PIM Join/Prune messages implies that snooping PIM messages places even a greater processing burden on a PE. Hence to learn <C-S, C-G> information from remote VSIs, this document proposes the use of a reliable protocol machinery to transport <C-S, C-G> information over the SP infrastruc-

ture. This is described in the next sub-section.

4.2. C-Multicast Control Information Propagation in the SP

A C-Join/Prune message for <C-S, C-G> coming from a customer, that is snooped by a PE, has to be propagated to the remote PE that can reach C-S. One way to do this is to forward the C-Join/Prune as a multicast data packet and let the egress PEs perform IGMP/PIM snooping over the pseudo-wire. However PIM is a soft state protocol and periodically re-transmits C-Join/Prune messages. This places a big burden on a PE while snooping PIM messages. It is not possible to eliminate this overhead for snooping messages received over the customer facing interfaces. However it is possible to alleviate this overhead over SP facing interfaces. This is done by converting snooped PIM C-Join/Prune messages to reliable protocol messages over the SP network. These reliable protocol messages are then sent to the remote PEs.

Each PE maintains the database of IGMP/PIM <C-S, C-G> entries that are learnt, usign reliable protocol messages, from remote PEs for each VSI. This is in addition to the database of IGMP/PIM <C-S, C-G> entries that are learnt from the local CEs, by snooping as described in the previous sub-section.

Compared to MVPNs there is an additional challenge while propagating snooped PIM C-Join/Prune messages over the SP network for VPLS. If the ingress PE wishes to propagate the C-Join/Prune only to the upstream PE which has reachability to C-S, this upstream PE is not known. This is because the local PE doesn't have a route to reach C-S. This is unlike MVPNs where the route to reach C-S is known from the unicast VPN routing table. This implies that the C-Join/Prune message has to be sent to all the PEs in the VPLS. This document proposes two possible solutions for achieving this and one of these will be eventually picked after discussion in the WG.

4.2.1. Using PIM

The PIM Neighbor discovery and maintenance is based on the VPLS membership information learnt as part of VPLS auto-discovery [BGP-AUTO].
VPLS auto-discovery allows a particular PE to learn which of the other PEs belong to a particular VPLS instance. Each of these PEs can be treated as a neighbor for PIM procedures while sending PIM C-Join/Prune messages to other PEs. The neighbor is considered up as long as the VPLS auto-discovery mechanism does not withdraw the

neighbor membership in the VPLS instance.

The C-Join/Prune messages is sent to all the PEs in the VPLS using unicast PIM messages. The use of unicast PIM implies that there is no PIM Join suppression for P-PIM messages. PIM refresh reduction

mechanisms, that are currently being worked upon in the PIM WG, MUST be used. These mechanisms aim at introducing reliability into PIM protocol messages, thereby reducing the overhead from the current periodic nature of PIM messages. To send the C-Join/Prune message to a particular remote PE, the message is encapsulated in the PW used to reach the PE, for the VPLS that the C-Join/Prune message belongs to.

4.2.2. Using BGP

The use of PIM for propagation of VPLS C-Join/Prune information may have scalability limitations. This is because even after building PIM refresh reduction mechanisms PIM will not have optimized transport when there is one sender and multiple receivers. BGP provides such transport as it has route-reflector machinery. Hence a reasonable option to propagate the C-Join/Prune information is to use BGP.

We describe the information elements needed if BGP were to be used to propagate the VPLS C-Join/Prune information in the SP network. The encoding details will be described in the future.

The following information is required to be advertised by BGP for a VPLS <C-Source, C-Group> for VPLS C-Join propagation and withdrawn by BGP for VPLS C-Prune propagation.

1. The RD configured for the VPLS instance. This is required to uniquely identify the <C-Source, C-Group> as the addresses could overlap between different VPLS instances.
2. The C-Source address. This can be a prefix.
3. The C-Group address. This can be a prefix.

When a PE distributes this information via BGP, it must include the Route Target (RT) Extended Communities attribute. This RT must be an "Import RT" of each VSI in the VPLS. The BGP distribution procedures used by [VPLS-BGP] or [BGP-AUTO] will then ensure that the advertised information gets associated with the right VSIs.

5. Security Considerations

Security considerations discussed in [VPLS-BGP] and [VPLS-LDP] apply to this document.

6. Acknowledgments

Many thanks to Thomas Morin for his support of this work.

7. References

7.1. Normative References

[RFC2119] "Key words for use in RFCs to Indicate Requirement Levels.", Bradner, March 1997

[RFC3107] Y. Rekhter, E. Rosen, "Carrying Label Information in BGP-4", RFC3107.

[VPLS-BGP] K. Kompella, Y. Rekther, "Virtual Private LAN Service", draft-ietf-l2vpn-vpls-bgp-02.txt

[VPLS-LDP] M. Lasserre, V. Kompella, "Virtual Private LAN Services over MPLS", draft-ietf-l2vpn-vpls-ldp-03.txt

[BGP-AUTO] H. Ould-Brahim et al., "Using BGP as an Auto-Discovery Mechanism for Layer-3 and Layer-2 VPNs", draft-ietf-l3vpn-bgpvpn-auto-04.txt

7.2. Informative References

[VPLS-MCAST] R. Aggarwal, Y. Kamite, L. Fang, "VPLS Multicast", draft-raggarwa-l2vpn-vpls-mcast-01.txt

[MVPN] E. Rosen, R. Aggarwal [Editors], "Multicast in BGP/MPLS VPNs", draft-ietf-l3vpn-2547bis-mcast-00.txt

8. Author Information

Rahul Aggarwal
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
Email: rahul@juniper.net

Yakov Rekhter
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
Email: yakov@juniper.net

Yuji Kamite
NTT Communications Corporation
Tokyo Opera City Tower
3-20-2 Nishi Shinjuku, Shinjuku-ku,
Tokyo 163-1421,
Japan
Email: y.kamite@ntt.com

Luyuan Fang
AT&T
200 Laurel Avenue, Room C2-3B35
Middletown, NJ 07748
Phone: 732-420-1921
Email: luyuanfang@att.com

Chaitanya Kodeboniya
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
Email: ck@juniper.net

9. Intellectual Property Statement

10. Full Copyright Statement