Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 14, 2015

T. Morin, Ed.
S. Litkowski
Orange
K. Patel
Cisco Systems
J. Zhang
R. Kebler
J. Haas
Juniper Networks

February 10, 2015

Multicast VPN state damping
draft-ietf-bess-multicast-damping-00

Abstract

This document describes procedures to damp multicast VPN routing state changes and control the effect of the churn due to the multicast dynamicity in customer site. The procedures described in this document are applicable to BGP-based multicast VPN and help avoid uncontrolled control plane load increase in the core routing infrastructure. New procedures are proposed inspired from BGP unicast route damping principles, but adapted to multicast.

Requirements Language

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

Table of Contents

1. Introduction

In a multicast VPN [RFC6513] deployed with BGP-based procedures [RFC6514], when receivers in VPN sites join and leave a said multicast group or channel through multicast membership control protocols (IGMP, MLD), multicast routing protocols accordingly adjust

multicast routing states and P-multicast tree states, to forward or prune multicast traffic to these receivers.

In VPN contexts, providing isolation between customers of a shared infrastructure is a core requirement resulting in stringent expectations with regards to risks of denial of service attacks. Hence, mechanisms need to be put in place to ensure that the load put on the BGP control plane, and on the P-tunnel setup control plane, remains under control regardless of the frequency at which multicast memberships changes are made by end hosts. By nature multicast memberships change based on the behavior of multicast applications running on end hosts, hence the frequency of membership changes can legitimately be much higher than the typical churn of unicast routing states. Section 16 of [RFC6514] specifically spells out the need for damping the activity of C-multicast and Leaf Auto-discovery routes.

This document describes procedures, remotely inspired from existing BGP route damping, aimed at protecting these control planes while at the same time avoiding negative effects on the service provided, although at the expense of a minimal increase in average of bandwidth use in the network.

The base principle is described in Section 3. Existing mechanisms that could be relied upon are discussed in Section 4. Section 5 details the procedures introduced by these specifications.

Section 6 provide specific details related to the damping of multicast VPNs P-tunnel state.

Finally, Section 7 discusses operational considerations related to the proposed mechanism.

2. Terminology

TBC

3. Overview

The procedures described in this document allows the network operator to configure multicast VPN PEs so that they can delay the propagation of multicast state prune messages, when faced with a rate of multicast state dynamicity exceeding a certain configurable threshold. Assuming that the number of multicast states that can be created by a receiver is bounded, delaying the propagation of multicast state pruning results in setting up an upper bound to the average frequency at which the router will send state updates to an upstream router.

From the point of view of a downstream router, such as a CE, this approach has no impact: the multicast routing states changes that it solicits to its PE will be honored without any additional delay. Indeed the propagation of joins is not impacted by the proposed defined procedures, and having the upstream router delay state prune propagation to its own upstream does not affect what traffic is sent to the downstream router. In particular, the amount of bandwidth used on the PE-CE link downstream to a PE applying this damping technique is not increased.

This approach increases the average bandwidth utilization on a link upstream to a PE applying this technique, such as a PE-PE link: indeed, a said multicast flow will be forwarded for a longer time than if no damping was applied. That said, it is expected that this technique will allow to meet the goals of protecting the multicast routing infrastructure control plane without a significant average increase of bandwidth; for instance, damping events happening at a frequency higher than one event per X second, can be done without increasing by more than X second the time during which a multicast flow is present on a link.

To be practical, such a mechanism requires configurability, in particular, needs to offer means to control when damping is triggered, and to allow delaying a multicast state Prune for a time increasing with the churn of this multicast state.

4. Existing mechanisms

This section describes mechanisms that could be considered to address the issue, but that end up appearing as not suitable or not efficient enough.

4.1. Rate-limiting of multicast control traffic

[RFC4609] examines multicast security threats and among other things the risk described in Section 1. A mechanism relying on rate-limiting PIM messages is proposed in section 5.3.3 [RFC4609], but has the identified drawbacks of impacting the service delivered and having side-effects on legitimate users.

4.2. Existing PIM, IGMP and MLD timers

In the context of PIM multicast routing protocols [RFC4601], a mechanism exists that in some context may offer a form of de facto damping mechanism for multicast states. Indeed, when active, the prune override mechanism consists in having a PIM upstream router introduce a delay ("prune override interval") before taking into account a PIM Prune message sent by a downstream neighbor.

This mechanism has not been designed specifically for the purpose of damping multicast state, but as a means to allow PIM to operate on multi-access networks. See [RFC4601] section 4.3.3. However, when active, this mechanism will prevent a downstream router to produce multicast routing protocol messages that would cause, for a said multicast state, the upstream router to send to its own upstreamrouter, multicast routing protocol messages at a rate higher than 1/[prune override interval], thus providing de-facto a form of damping.

Similarly, the IGMP and MLD multicast membership control protocols can provide a similar behavior, under the right conditions.

These mechanisms are not considered suitable to meet the goals spelled out in Section 1, the main reasons being that:

- o when enabled these mechanisms require additional bandwidth on the local link on which the effect of a Prune is delayed (in our case the PE-CE link)

- o when enabled these mechanisms require disabling explicit tracking, even though enabling this feature may otherwise be desired

- o on certain implementations, these mechanisms are incompatible with behavior that cannot be turned off

- o they do not provide a suitable level of configurability

- o they do not provide a way to discriminate between multicast flows based on estimation of their dynamicity

4.3. BGP Route Damping

The procedures defined in [RFC2439] and [RFC7196] for BGP route flap damping are useful for operators who want to control the impact of unicast route churn on the routing infrastructure, and offer a standardized set of parameters to control damping.

These procedures are not directly relevant in a multicast context, for the following reasons:

- o they are not specified for multicast routing protocol in general

- o even in contexts where BGP routes are used to carry multicast routing states (e.g. [RFC6514]), these procedures do not allow to implement the principle described in this document, the main reason being that a damped route becomes suppressed, while the

target behavior would be to keep advertising when damping is triggered on a multicast route

However, the set of parameters standardized to control the thresholds of the exponential decay mechanism can be relevantly reused. This is the approach proposed for the procedures described in this document (Section 5). Motivations for doing so is to help the network operator deploy this feature based on consistent configuration parameter, and obtain predictable results, without the drawbacks of exposed in Section 4.1 and Section 4.2.

5. Procedures for multicast state damping

5.1. PIM procedures

This section describes procedures for multicast state damping satisfying the goals spelled out in Section 1. This section spells out procedures for (S,G) states in the PIM-SM protocol ([RFC4601] ; they apply unchanged for such states created based on multicast group management protocols (IGMP [RFC3376], MLD [RFC3810]) on downstream interfaces. The same procedures are applied to (*,G) states in the context of PIM-SM ASM groups (damping is not applied to (S,G,Rpt) Prune state).

The following notions introduced in [RFC2439] are reused in these procedures:

figure-of-merit: a number reflecting the current estimation of past recent activity of an (S,G) multicast routing state, which evolves based on routing events related to this state and based an exponential decay algorithm ; the activation or inactivation of damping on the state is based on this number ; this number is associated to the upstream state machine for (S,G)

cutoff-threshold: value of the *figure-of-merit* over which damping is applied (configurable parameter)

reuse-threshold: value of the *figure-of-merit* under which damping stops being applied (configurable parameter)

decay-half-life: period of time used to control how fast is the exponential decay of the *figure-of-merit* (configurable parameter)

Additionally to these values, a configurable "*increment-factor*" parameter is introduced, that controls by how much the *figure-of-merit* is incremented on multicast state update events.

Section Section 7.3 proposes default and maximum values for the configurable parameters.

On reception of updated multicast membership or routing information on a downstream interface I for a said (S,G) state, that results in a change of the state of the PIM downstream state machine (see section 4.5.3 of [RFC4601]), a router implementing these procedures MUST:

o apply unchanged procedures for everything relating to what multicast traffic ends up being sent on downstream interfaces, including interface I

o increasing the *figure-of-merit* for the (S,G) by the *increment-factor* (updating the *figure-of-merit* based on the decay algorithm must be done prior to this increment)

o update the damping state for the (S,G) state: damping becomes active on the state if the recomputed *figure-of-merit* is above the configured *cutoff-threshold*

o if damping is inactive on (S,G) state, update the upstream state machine as usual (as per section 4.5.7 of [RFC4601])

o if damping becomes active for the (S,G) state:

  * if the received message has caused the upstream state machine to transition to Joined state, update the upstream state machine for (S,G) (applying usual PIM procedures in section 4.5.7 of [RFC4601], including sending a PIM Join to the upstream neighbor)

  * if the received message has caused the upstream state machine to transition to NotJoined state, do not update the upstream state machine for (S,G)

  * then freeze the upstream state machine in Joined state, and and setup a trigger to update it once damping later becomes inactive again. The effect is that in the meantime, PIM Join messages will be sent as refreshes to the upstream neighbor, but no PIM Prune message will be sent.

o if damping was already active: do not update the upstream state machine for (S,G) (the upstream state machine was frozen after processing the previous message)

Once the *figure-of-merit* for (S,G) damping state decays to a value below the configured *reuse-threshold*, the upstream state machine for (S,G) is recomputed based on states of downstream state machines,

eventually leading to a PIM Join or Prune message to be sent to the upstream neighbor.

Same techniques as the ones described in [RFC2439] can be applied to determine when the figure-of-merit value is recomputed based on the exponential decay algorithm and the configured *decay-half-life*.

Given the specificity of multicast applications, it is REQUIRED for the implementation to let the operator configure the *decay-half-life* in seconds, rather than in minutes. When the recomputation is done periodically, the period should be low enough to not significantly delay the inactivation of damping on a multicast state beyond what the operator wanted to configure (i.e. for a half-life of 10s, recomputing the *figure-of-merit* each minute would result in a multicast state to remained damped for a much longer time than what the parameters are supposed to command).

PIM implementations typically follow [RFC4601] suggestion that "implementations will only maintain state when it is relevant to forwarding operations - for example, the 'NoInfo' state might be assumed from the lack of other state information, rather than being held explicitly" (Section 4.1 of [RFC4601]). To properly implement implement damping procedures, an implementation MUST keep an explicit (S,G) state as long as damping is active on an (S,G). Once an (S,G) state expires, and damping becomes inactive on this state, its associated *figure-of-merit* and damping state are removed as well.

Note that these procedures:

- do not impact PIM procedures related to refreshes or expiration of multicast routing states: PIM Prune messages triggered by the expiration of the (S,G) keep-alive timer, are not suppressed or delayed, and the reception of Join messages not causing transition of state on the downstream interface does not lead to incrementing the *figure-of-merit*;
- do not impact the PIM assert mechanism, in particular PIM Prune messages triggered by a change of the PIM assert winner on the upstream interface, are not suppressed or delayed;
- do not impact PIM Prune messages that are sent when the RPF neighbor is updated for a said multicast flow;
- do not impact PIM Prune messages that are sent in the context of switching between a Rendez-vous Point Tree and a Shortest Path Tree.

Note also that no action is triggered based on the reception of PIM Prune messages (or corresponding IGMP/MLD messages) that relate to non-existing (S,G) state, in particular, no *figure-of-merit* or damping state is created in this case.

5.2. Procedures for multicast VPN state dampening

The procedures described in Section 5.1 can be applied in the VRF PIM-SM implementation (in the "C-PIM instance"), with the corresponding action to suppressing the emission of a Prune(S,G) message being to not withdraw the C-multicast Source Tree Join (C-S,C-G) BGP route. Implementation of [RFC6513] relying on the use of PIM to carry C-multicast routing information MUST support this technique.

In the context of [RFC6514] where BGP is used to distribute C-multicast routing information, the following procedure is proposed as an alternative and consists in applying damping in the BGP implementation, based on existing BGP damping mechanism, applied to C-multicast Source Tree Join routes and Shared Tree Join routes (and as well to Leaf A-D routes - see Section 6), and modified to implement the behavior described in Section 3 along the following guidelines:

- o not withdrawing (instead of not advertising) damped routes
- o providing means to configure the half-life in seconds if that option is not already available
- o using parameters for the exponential decay that are specific to multicast, based on default values and multicast specific configuration

While these procedures would typically be implemented on PE routers, in a context where BGP Route Reflectors are used it can be considered useful to also be able to apply damping on RRs as well. Additionally, for mVPN Inter-AS deployments, it can be needed to protect one AS from the dynamicity of multicast VPN routing events from other ASes. In that perspective, it is RECOMMENDED for implementations to support damping mVPN C-multicast routes directly into BGP, without relying on the PIM-SM state machine.

When not all routers in a deployment have the capability to drop traffic coming from the wrong PE (as spelled out in section 9.1.1 of [RFC6513]), then the withdrawal of a C-multicast route resulting from a change in the UMH SHOULD NOT be damped. An implementation of these specs MUST whether, not damp these withdrawals by default, or alternatively provide a tuning knob to disable then damping of these

withdrawals. Additionally, in such a context, it is RECOMMENDED to *not* enable any multicast VPN route damping on RRs and ASBRs, since these equipments cannot distinguish these events.

The choice to implement damping based on BGP routes or the procedures described in Section 5, is up to the implementor, but at least one of the two MUST be implemented; keeping in mind that in contexts where damping on RRs and ASBRs the BGP approach is RECOMMENDED.

Note well that damping SHOULD NOT be applied to BGP routes of the following sub-types: "Intra-AS I-PMSI A-D Route", "Inter-AS I-PMSI A-D Route", "S-PMSI A-D Route", and "Source Active A-D Route".

6. Procedures for P-tunnel state damping

6.1. Damping mVPN P-tunnel change events

When selective P-tunnels are used (see section 7 of [RFC6513]), the effect of updating the upstream state machine for a said (C-S,C-G) state on a PE connected to multicast receivers, is not only to generate activity to propagate C-multicast routing information to the source connected PE, but also to possibly trigger changes related to the P-tunnels carrying (C-S,C-G) traffic. Protecting the provider network from an excessive amount of change in the state of P-tunnels is required, and this section details how this can be done.

A PE implementing these procedures for mVPN MUST damp Leaf A-D routes, in the same manner as it would for C-multicast routes (see Section 5.2).

A PE implementing these procedures for mVPN MUST damp the activity related to removing itself from a P-tunnel. Possible ways to do so depend on the type of P-tunnel, and local implementation details are left up to the implementor.

The following is proposed as example of how the above can be achieved.

- For P-tunnels implemented with the PIM protocol, this consists in applying multicast state damping techniques described in Section 5.1 to the P-PIM instance, at least for (S,G) states corresponding to P-tunnels.

- For P-tunnels implemented with the mLDP protocol, this consists in applying damping techniques completely similar as the one described in Section 5, but generalized to apply to mLDP states

- For root-initiated P-tunnels (P-tunnels implemented with the P2MP RSVP-TE, or relying on ingress replication), no particular action needs to be implemented to damp P-tunnels membership, if the activity of Leaf A-D route themselves is damped

- Another possibility is to base the decision to join or not join the P-tunnel to which a said (C-S,C-G) is bound, and to advertise or not advertise a Leaf A-D route related to (C-S,C-G), based on whether or not a C-multicast Source Tree Join route is being advertised for (C-S,C-G), rather than by relying on the state of the C-PIM Upstream state machine for (C-S,C-G)

6.2. Procedures for Ethernet VPNs

Specifications exists to support or optimize multicast and broadcast in the context of Ethernet VPNs [RFC7117], relying on the use of S-PMSI and P-tunnels. For the same reasons as for IP multicast VPNs, an implementation of these procedures MUST follow the procedures described in this section.Section 6.1.

7. Operational considerations

7.1. Enabling and configuring multicast damping

In the context of multicast VPNs, these procedures would be enabled on PE routers. Additionally in the case of C-multicast routing based on BGP extensions ([RFC6514]) these procedures can be enabled on ASBRs, and possibly Route Reflectors as well.

7.2. Troubleshooting and monitoring

Implementing the damping mechanisms described in this document should be complemented by appropriate tools to observe and troubleshoot damping activity.

More specifically it is RECOMMENDED to complement the existing interface providing information on multicast states with information on eventual damping of corresponding states (e.g. MRIB states): C-multicast routing states and P-tunnel states.

7.3. Default and maximum values

The following values are RECOMMENDED to adopt as default conservative values:

- increment-factor: 1000

- cutoff-threshold: 3000

o decay-half-life: 10s

o reuse-threshold: 1500

For unicast damping, it is common to set an upper bound to the time during which a route is suppressed. In the case of multicast state damping, which relies on not withdrawing a damped route, it may be desirable to avoid a situation were a multicast flow would keep flowing in a portion of the network for a very large time in the absence of receivers.

The proposed default maximum value for the figure-of-merit is 20x<increment-factor>, i.e. 20000 with the proposed default increment-factor of 1000.

The following values are proposed as maximums:

o decay half-life: 60s

o cutoff-threshold: 50000

8. IANA Considerations

This document makes no request of IANA.

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

9. Security Considerations

The procedures defined in this document do not introduce additional security issues not already present in the contexts addressed, and actually aim at addressing some of the identified risks without introducing as much denial of service risk as some of the mechanisms already defined.

The protection provided relates to the control plane of the multicast routing protocols, including the components implementing the routing protocols and the components responsible for updating the multicast forwarding plane.

The procedures describe are meant to provide some level of protection for the router on which they are enabled by reducing the amount of routing state updates that it needs to send to its upstream neighbor or peers, but do not provide any reduction of the control plane load related to processing routing information from downstream neighbors. Protecting routers from an increase in control plane load due to activity on downstream interfaces toward core routers (or in the

context of BGP-based mVPN C-multicast routing, BGP peers) shall rely upon the activation of damping on corresponding downstream neighbors (or BGP peers) and/or at the edge of the network. Protecting routers from an increase in control plane load due to activity on customer-facing downstream interfaces or downstream interfaces to routers in another administrative domain, is out of the scope of this document and should rely upon already defined mechanisms (see [RFC4609]).

To be effective the procedures described here must be complemented by configuration limiting the number of multicast states that can be created on a multicast router through protocol interactions with multicast receivers, neighbor routers in adjacent ASes, or in multicast VPN contexts with multicast CEs. Note well that the two mechanism may interact: state for which Prune has been requested may still remain taken into account for some time if damping has been triggered and hence result in otherwise acceptable new state from being successfully created.

Additionally, it is worth noting that these procedures are not meant to protect against peaks of control plane load, but only address averaged load. For instance, assuming a set of multicast states submitted to the same Join/Prune events, damping can prevent more than a certain number of Join/Prune messages to be sent upstream in the period of time that elapses between the reception of Join/Prune messages triggering the activation of damping on these states and when damping becomes inactive after decay.

10. Acknowledgements

We would like to thank Bruno Decraene and Lenny Giuliano for discussions that helped shape this proposal. We would also like to thank Yakov Rekhter and Eric Rosen for their reviews and helpful comments. Thanks to Wim Henderickx for his comments and support of this proposal.

11. References

11.1. Normative References

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

[RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping", RFC 2439, November 1998.

[RFC3376] Cain, B., Deering, S., Kouvelas, I., Fenner, B., and A. Thyagarajan, "Internet Group Management Protocol, Version 3", RFC 3376, October 2002.

[RFC3810] Vida, R. and L. Costa, "Multicast Listener Discovery Version 2 (MLDv2) for IPv6", RFC 3810, June 2004.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[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.

[RFC7117] Aggarwal, R., Kamite, Y., Fang, L., Rekhter, Y., and C. Kodeboniya, "Multicast in Virtual Private LAN Service (VPLS)", RFC 7117, February 2014.

[RFC7196] Pelsser, C., Bush, R., Patel, K., Mohapatra, P., and O. Maennel, "Making Route Flap Damping Usable", RFC 7196, May 2014.

11.2. Informative References

[RFC4609] Savola, P., Lehtonen, R., and D. Meyer, "Protocol Independent Multicast - Sparse Mode (PIM-SM) Multicast Routing Security Issues and Enhancements", RFC 4609, October 2006.

Authors' Addresses

Thomas Morin (editor)
Orange
2, avenue Pierre Marzin
Lannion 22307
France

Email: thomas.morin@orange.com

Stephane Litkowski
Orange
France

Email: stephane.litkowski@orange.com

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: keyupate@cisco.com

Jeffrey (Zhaohui) Zhang
Juniper Networks Inc.
10 Technology Park Drive
Westford, MA 01886
USA

Email: zzhang@juniper.net

Robert Kebler
Juniper Networks Inc.
10 Technology Park Drive
Westford, MA 01886
USA

Email: rkebler@juniper.net

Jeff Haas
Juniper Networks

Email: jhaas@juniper.net