Network Working Group
Internet-Draft
Intended status: Informational
Expires: August 28, 2008

T. Morin
G. Cauchie
France Telecom - Orange Labs

February 25, 2008

Multicast blackhole mitigation with PIM adjacency conditions on routing announcements

draft-morin-mboned-mcast-blackhole-mitigation-01

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with Section 6 of BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on August 28, 2008.

Copyright Notice

Copyright (C) The IETF Trust (2008).

Abstract

In a network running PIM-SM, multicast traffic can fail to be properly routed and forwarded during some period of time after a topology change, when unicast routing converges to a new path using a new next-hop before multicast routing adjacencies are fully setup with this new next-hop. At this point, a blackhole appears in the multicast topology and persists until the PIM adjacency become fully setup. This document describes different procedures aiming at

avoiding such blackholes, focusing on the use of non-congruent unicast routing that takes into account the status of PIM adjacencies.

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 3
2. Terminology 3
3. Problem statement 3
   3.1. Ineffective PIM Adjacency 3
   3.2. Problem description 4
   3.3. Partial solutions 4
   3.4. Applicability to SSM and non-SSM multicast 5
4. PIM-adjacency-based non-congruent MRIB 5
   4.1. Strategy description 5
   4.2. Criterias 6
      4.2.1. Prune links on which PIM is not enabled 6
      4.2.2. Prune links with no received PIM Hello 6
      4.2.3. Prune links with no sent PIM Hello 7
      4.2.4. Other criteria 7
5. Alternative proposal 7
6. Interoperability considerations 7
7. Recommandations 8
8. IANA Considerations 8
9. Security Considerations 8
10. Acknowledgements 8
11. References 9
Authors' Addresses 10
Intellectual Property and Copyright Statements 11

1. Introduction

In a network running PIM-SM [RFC4601], multicast traffic can fail to be routed during some period of time after a topology change, when unicast routing protocols converge before the PIM-SM multicast routing protocol adjacencies are fully setup, causing blackholing of multicast traffic. This document describes different procedures aiming at avoiding such blackholes.

Section 3 describes the problem statement, and following Section 4 and Section 5 describe possible solutions to mitigate theses blackholes.

2. Terminology

This document is essentially using terminology from unicast routing protocols and from the PIM-SM specifications [RFC4601].

Moreover, throughout this document the "MT-IGP" term designates an IGP able to carry information about multiple network topology. Example of such IGP are: MT-ISIS [I-D.ietf-isis-wg-multi-topology], MI-ISIS [I-D.previdi-isis-mi], MT-OSPFv2 [RFC4915], MT-OSPFv3 [I-D.ietf-ospf-mt-ospfv3], MI-OSPF [I-D.acee-ospf-multi-instance].

3. Problem statement

3.1. Ineffective PIM Adjacency

In a number of situations, the status of a link can be "up" without the PIM adjacency being fully setup. Possible causes for this kind of situations are :

- o PIM is unconfigured on one or both ends of the link (not yet configured, or misconfigured)

- o a bug or hardware failure is inducing unreliability or delay in sending or receiving PIM Hello messages

- o no PIM Hello has been received yet by one of the routers, but that router would require having received a PIM Hello before sending PIM Join to a neighbor

On this last point, note that it seems that [RFC4601] does not clearly tell if a router should or should not delay sending PIM Joins on a link toward a router from which it hasn't yet received any PIM Hello message. The procedures described in [RFC4601] are so that it

can take up to <Triggered_Hello_Delay> seconds (up to 5s, by default) before a neighbor will send a PIM Hello triggered by a Hello sent by a new neighbor on the link.

Operational experience shows that all of these situations occur, even frequently for some of them.

A similar problem affects the LDP protocol, for which a solution is also being studied[I-D.ietf-mpls-igp-sync].

3.2. Problem description

The generic problem targeted by this document is when the unicast routing protocol starts to consider a link as available for unicast routing (e.g.. a new link, or a flapping link coming up) but the PIM adjacency on this link is ineffective (see section above), causing PIM Joins to fail to be propagated further, and resulting in a blackhole for the corresponding multicast traffic.

This problem can occur in a few distinct situations:

- o IGP case : in an autonomous system, the IGP can disseminate information relative to a new link before PIM-SM is fully setup on this link

- o EGP case : when a link between two ASBRs comes up, the EGP (typically BGP) can possibly advertise the new routes received from the neighbor ASBR before PIM is fully effective on this link

- o multicast BGP/MPLS VPN case : similar to the EGP case

Partial solutions to the issues presented above do exist based on implementation or unicast routing tuning, but fail to cover all plausible blackhole causes.

3.3. Partial solutions

A first solution is to relax how PIM behaves and do not require a router to having received or sent a Hello to/from a neighbor before accepting/sending a Join from/to the neighbor. This solution somehow partially defeats the purpose of PIM Hello procedures, that include advertisements of parameters and capabilities between neighbors, and would fail if the use of Hellos is extended to carry information that impacts how Join or Prune are sent, but still appears to be effective given the current use of Hello messages. It remains however a partial solution to the problem exposed here, since it doesn't cover the misconfiguration cases.

A second solution is to improve the PIM procedures for setting up an adjacency, so that the delay before sending a triggered hello or replying to a hello of a new neighbor is reduced to zero (current specifications use a random timer between 0 and Triggered_hello_delay, which defaults to 5s). That solution is an improvement upon the previous one, but still remains a partial solution to the problem described in this document, since it doesn't cover misconfiguration cases.

3.4. Applicability to SSM and non-SSM multicast

The PIM-SM multicast routing protocol is using unicast routes (the MRIB) to decide how PIM Join messages should be routed toward the source (in the SPT case) or the RP (in the RPT case). In this document we will only describe the SSM/SPT case, for the sake of clarity, but all statements equally apply to the non-SSM/RPT case, using the RP instead of the multicast source.

4. PIM-adjacency-based non-congruent MRIB

4.1. Strategy description

Most unicast routing protocols have been extended, or are being extended to support semantic extensions to advertise unicast routes that will be used specifically for multicast routing : MT-ISIS [I-D.ietf-isis-wg-multi-topology], MI-ISIS [I-D.previdi-isis-mi], MT-OSPF [RFC4915], MP-BGP SAFI 2 [RFC4760], MP-BGP SAFI 129 VPNv4 routes [I-D.ietf-l3vpn-2547bis-mcast-bgp]. When one or more of these extensions are in use, PIM-SM will be configured to use the distinct RIB populated by them as its MRIB (the RIB used for RPF lookups). In such a case, the MRIB is said to be "non-congruent" to the unicast RIB.

The issues described in this document can be fully or partially addressed by the use of a non-congruent MRIB that takes into account the status of the PIM-SM adjacencies. A link or the associated next hop won't populate the MRIB based only on the status of the link for the unicast routing protocol, but only if conditions are verified on the status of the PIM adjacency. More specifically:

- in the case of a link where a link-state protocol (typically an IGP) is used, the link (and the routes directed routed through it) will not be advertised in the multicast-specific topology if it is known that the PIM adjacency is not ready.

- in the case of a link where a distance vector protocol (e.g.. BGP) is used, the routes received from a neighbor on that link,

that pertain to the multicast-specific RIB (e.g.. SAFI 2 routes in the case of BGP) will not be re-advertised.

Different facts can be taken into account by the local router to decide whether the PIM adjacency is ready or not. Different routers in a domain can have a different policy for this decision process : the more specific is the information taken into account, the more black-holes are expected to be avoided ; but in any case, there will always be some improvements over just using non-congruent routing.

4.2. Criterias

This sections details the different facts that can be taken into account to decide that a PIM adjacency is effective or not.

4.2.1. Prune links on which PIM is not enabled

This is the most basic criteria, which consist in including in the multicast topology, only links on which PIM is administratively enabled.

It typically allows to cover cases of PIM misconfiguration where PIM is not enabled on both sides of a link.

4.2.2. Prune links with no received PIM Hello

This consists in not including in the multicast topology, links on which no PIM Hello was received yet (or for which the PIM Neighbor Liveness Timer expired).

This is particularly relevant if the local PIM-SM implementation requires having received PIM Hellos from a neighbor before sending a Join toward this neighbor.

4.2.2.1. IGP on LAN specific case

An IGP like OSPF or ISIS elects a node that will be responsible of creating a pseudo nodes for the LAN and each node on the LAN will then announce an adjacency with this pseudo-node. There is not such notion as a pseudo-node in PIM-SM, PIM adjacencies on a LAN being setup between all PIM routers on the LAN.

In the case of an IGP on a LAN the "prune links with no received PIM Hello" criteria translates into requiring the router that created the pseudo node to prune links between the pseudo-node and routers from which it didn't receive a Hello (or for which the Neighbor Liveness Timer expired).

4.2.3. Prune links with no sent PIM Hello

This consists in not including in the multicast topology, the links on which no PIM Hello was sent yet.

This is particularly relevant during the randomized period of time between a Hello being received by a new neighbor and a triggered Hello being sent.

4.2.4. Other criteria

Other criteria may be considered, like not including links the received PIM Hello would lack the indication of a required Hello option, or an unknown PIM protocol version, etc.

5. Alternative proposal

An alternative mechanism to the one described in previous section would consist in using infinite metrics for a link until where PIM is enabled until the PIM adjacency is fully setup. This alternative has the advantage of not requiring the use of a multi-topology IGP (in the IGP case) or of SAFI 2 routing (in the EGP case), but has the drawback of impacting unicast traffic, causing perturbations of unicast routing during the period of time where the PIM adjacency is not fully setup yet. For this reason, this alternative is not favored and not described further in this document.

6. Interoperability considerations

Both mechanisms described in this memo do not cause any interoperability issue.

The decision to use or not use non-congruent unicast routing for multicast routing has to be made consistently across a routing domain, but the decision to apply or not apply the specific procedures described in this document, and the policy to decide whether or not a link will be used for multicast routing, can be a purely local procedure. The worst situation that could occur is not seeing any improvement over the initial situation where no specific procedure is used, and where blackholing of multicast traffic is more likely to occur.

For these reason, this document is only intended as a guideline for implementors, and only target the informational status.

7. Recommandations

To prevent misconfiguration cases, it is RECOMMENDED to make procedures in 4.2.1 best practice (Prune links with no received PIM Hello from the multicast-RPF-specific unicast topology) when a non-congruent multicast routing technique is used.

It is RECOMMENDED to make procedures in 4.2.2 and 4.2.3 best practice when a non-congruent multicast routing technique is used, if the plain PIM procedures of [RFC4601] for sending and receiving Hello are used by at least one router on the link, and they need not be used if improved Hello adjacency procedures, such as described in Section 3.3, are known to be used by all routers on a link. Updating [RFC4601] to improve the PIM adjacency setup time should strongly be considered.

Logging is RECOMMENDED when a router choses to not include a link for multicast routing.

Operator feedback through the router management interface is RECOMMENDED, both in the management interface section related to unicast protocols which are used to implement the non-congruency, and in the section related to PIM-SM.

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 in this document are believed to not introduce any specific issue introduced compared to multicast routing done with PIM-SM [RFC4601].

10. Acknowledgements

The authors want to thank Bruno Decraene and Toerless Eckert for discussions that helped to shape this proposal.

11. References

[I-D.acee-ospf-multi-instance]
Lindem, A., Roy, A., and S. Mirtorabi, "OSPF Multi-Instance Extensions", draft-acee-ospf-multi-instance-01 (work in progress), February 2008.

[I-D.ietf-isis-wg-multi-topology]
Przygienda, T., "M-ISIS: Multi Topology (MT) Routing in IS-IS", draft-ietf-isis-wg-multi-topology-12 (work in progress), November 2007.

[I-D.ietf-l3vpn-2547bis-mcast-bgp]
Aggarwal, R., "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-bgp-04 (work in progress), November 2007.

[I-D.ietf-mpls-igp-sync]
Jork, M., "LDP IGP Synchronization", draft-ietf-mpls-igp-sync-00 (work in progress), September 2007.

[I-D.ietf-ospf-mt-ospfv3]
Mirtorabi, S. and A. Roy, "Multi-topology routing in OSPFv3 (MT-OSPFv3)", draft-ietf-ospf-mt-ospfv3-03 (work in progress), July 2007.

[I-D.previdi-isis-mi]
Previdi, S., "IS-IS Multi-instance", draft-previdi-isis-mi-00 (work in progress), May 2007.

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

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

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, June 2007.

Authors' Addresses

Thomas Morin
France Telecom - Orange Labs
2 avenue Pierre Marzin
Lannion 22307
France

Email: thomas.morin@orange-ftgroup.com

Gregory Cauchie
France Telecom - Orange Labs
38-40 rue du General Leclerc
Issy Les Moulineaux 92794
France

Email: gregory.cauchie@orange-ftgroup.com

Full Copyright Statement

Copyright (C) The IETF Trust (2008).

This document is subject to the rights, licenses and restrictions contained in BCP 78, and except as set forth therein, the authors retain all their rights.

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Intellectual Property

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).