PIM-SM DR Priority Auto-Adjustment

draft-xu-pim-drpriority-auto-adjustment-03

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and 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 February 24, 2011.

Copyright Notice

Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents...
Internet-Draft     PIM-SM DR Priority Auto-Adjustment August 2010

carefully, as they describe your rights and restrictions with respect to this document.

Abstract

This document defines some mechanisms for automatically adjusting the Protocol Independent Multicast (PIM) Designed Router (DR) priority according to the status of the tracked interface. This approach can be used to realize DR auto-switchover in case the DR loses the route to multicast source or Rendezvous Point (RP).

Conventions used in this document

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. Problem Statement............................................3
2. DR Priority Auto-adjustment Mechanism........................4
3. Security Considerations.....................................5
4. IANA Considerations.........................................5
5. Acknowledgments.............................................6
6. References................................................6
   6.1. Normative References...................................6
   6.2. Informative References................................6
Authors’ Addresses............................................6
1. Problem Statement

In the Protocol Independent Multicast – Sparse Mode (PIM-SM) specification [RFC4601], a Designated Routers (DR) plays a very important role on the shared-media LAN like Ethernet which may have more than one PIM-SM speaking routers and one or more hosts. One of the PIM-SM speaking routers is designated aka elected (per interface) to send (a) PIM join/prune messages upon receiving Internet Group Management Protocol (IGMP) [RFC3376] membership report/leave messages from multicast receivers/hosts, or (b) PIM Register packets to a Rendezvous Point (RP) on behalf of the multicast senders. DR election happens on all interfaces, LAN or otherwise.

To achieve DR redundancy, multiple PIM routers are usually deployed in a shared-media LAN like Ethernet. As depicted in the following figure, two PIM routers R1, R2 are connected to the same LAN where a multicast receiver is located. Assume R1 is elected as the DR for that LAN. Now R1’s upstream link to the outside network is down, as a result, R1 will lose the reachability to the multicast senders and/or RPs. If R1 doesn’t quit the DR role, the multicast service for the receiver will be broken even though R2 may still have reachability to the multicast senders and/or RPs.

In most cases, the recovery of the multicast forwarding tree is dependent on the unicast route convergence. However, in a topology such as the one illustrated above, the (LAN) interface between R1 and R2 is treated passive from the routing protocol perspective for obvious reasons (flooding, scale, spoofing etc.), mooting the unicast routing convergence applicability.

While it is possible that a failover static route can be configured on these two routers instead of running IGP, that’s to say, each PIM router is configured with a default route with its next-hop pointed to the opposite side, and this default route will not take effect until the reachability to the outside network is lost. However, this approach has a few side-effects. Assume these two routers are...
connected to the same Internet Service Provider (ISP) network via different links, once the default route learnt from the ISP is withdrawn, the forwarding loop will occur between them upon receiving a packet whose best route is the configured default route. Another side-effect with this approach is when the upstream link to the outside network recovers (from down to up), the RPF interface will change. As a result, the transient forwarding interruption due to SPT/RPT rebuilding will occur. Although this transient interruption issue due to the unicast route change can be alleviated by using some make-before-break mechanism (e.g., by borrowing some ideas from RPT→SPT switch mechanism, the old upstream link will not be pruned until the stream is received from the new upstream link.), the cost is relatively high for this scenario.

2. DR Priority Auto-adjustment Mechanism

A PIM router could exactly cease the DR role by decreasing its DR priority once it loses routes to the RPs or multicast sources. Since the addresses of the RPs or the multicast sources may not be known to it, the PIM router could simply determine whether or not it has lost the route to the RPs or multicast sources just by tracking the status of its upstream link. This approach is much similar to the interface tracking mechanism of the Virtual Router Redundancy Protocol (VRRP) [RFC2338]. Taken the above scenario as an example, R1 will track its upstream link to the outside network, once the upstream link is down, its PIM DR priority will be reduced automatically so as to cease the DR role. As a result, R2 will take over the DR role. To speed up the DR switchover further, as soon as the DR priority adjustment occurs, a new hello message containing the current DR priority will be triggered.

In most cases, two last-hop routers each with an upstream link are enough for redundancy purpose. In case there are more than one upstream links on a PIM router, all of these interfaces should be tracked for the DR priority adjustment. As long as there is a route to the outside network, it is not necessary for the DR switchover to take place. However, if the number of uplink interfaces is large, the network administrator can actually choose part of uplinks which are to be tracked for the DR adjustment.

Once the failed upstream link on R1 recovers (from down to up), its DR priority can be either recovered to the original value or not according to the pre-configured preemption policy. If preemption is not allowed, its DR priority will remain unchanged so as to avoid unnecessary DR switchover. As a result, the transient multicast forwarding interruption due to multicast forwarding tree rebuilding can be avoided. Once the current DR fails or its DR priority changes,
the non-DRs SHOULD restore theirs DR priorities to their original value provided they have got route to the uplinks for competing in a new round of DR election.

In fact, if VRRP is run on the PIM routers and the VRRP has enabled the upstream link tracking mechanism, the PIM DR failover mechanism could be coupled with the VRRP so as to reuse the upstream link tracking mechanism of VRRP. One option is to synchronize the PIM DR priority value to the VRRP priority value always. In this way, the PIM DR and the VRRP master will always run on an identical router if the VRRP Preempt_Mode is set to True. Another option is to make the PIM DR and the VRRP master run on an identical router anyway (i.e., regardless whether or not the VRRP Preempt_Mode is True). To achieve this goal, the PIM DR priority of the VRRP master router SHOULD always be set to a higher fixed value than that of the VRRP slave router automatically.

Note that these approaches are fully compatible with the current PIM specification. There is no need to change the PIM protocol on wire at all since they are just internal implementations.

Although the PIM router could cease the DR role by declaring its death directly, e.g., sending a hello message with a zero HoldTime, or keeping silent till it times out, this approach has a few limitations in some specific scenarios. Taking the above scenario as an example, assume there is another multicast sender directly connected to R1 via another interface. Since the receiver and this sender are connected to the same PIM router, the receiver ought to be able to receive the multicast stream from this sender. However, in case there is another multicast source in the outside network using the same address (i.e., in anycast source scenario), a problem due to PIM assert failure will occur since join/prune or assert messages from a non-neighbor will be discarded silently according to the PIM specification.

3. Security Considerations

The DR priority auto-adjustment mechanisms described in this document does not introduce any new security risk.

4. IANA Considerations

There is no IANA consideration for this document.
5. Acknowledgments

The authors would like to thank Liu Hui for her valuable comments. Thanks should also be given to Stig Venaas who mentioned the idea of coupling the PIM DR with the VRRP master tightly.

6. References

6.1. Normative References


6.2. Informative References


Authors’ Addresses

Xiaohu Xu
Huawei Technologies,
No.3 Xinxi Rd., Shang-Di Information Industry Base,
Hai-Dian District, Beijing 100085, P.R. China
Phone: +86 10 82836073
Email: xuxh@huawei.com

Ganesan SriManikandan
Extreme Networks
Temple Steps, 184-187, (8th floor)
Anna Salai, Saidapet,
Chennai - 600 115
India
Email: sganesan@extremenetworks.com

Rajiv Asati
Cisco Systems
7025-4 Kit Creek Rd
PO Box 14987
Research Triangle Park, NC 27709
USA
Email: rajiva@cisco.com

Prashant Jhingran
Cisco Systems
SEZ Unit, Cessna Business Park,
Sarjapur Marathalli Outer Ring Road,
Bangalore - 560 103
India
Email: pjhingra@cisco.com