Network Working Group   A. Karan
Internet-Draft   C. Filsfils
Intended status: Informational   IJ. Wijnands, Ed.
Expires: November 13, 2015   Cisco Systems, Inc.
                                                        B. Decraene
                                                             Orange
                                                       May 12, 2015

Multicast only Fast Re-Route
draft-ietf-rtgwg-mofrr-07

Abstract

   As IPTV deployments grow in number and size, service providers are
   looking for solutions that minimize the service disruption due to
   faults in the IP network carrying the packets for these services.
   This document describes a mechanism for minimizing packet loss in a
   network when node or link failures occur. Multicast only Fast Re-
   Route (MoFRR) works by making simple enhancements to multicast
   routing protocols such as PIM and mLDP.

Status of This Memo

   This Internet-Draft is submitted in full conformance with the
   provisions of BCP 78 and BCP 79.

   Internet-Drafts are working documents of the Internet Engineering
   Task Force (IETF). Note that other groups may also distribute
   working documents as Internet-Drafts. The list of current Internet-
   Drafts is at http://datatracker.ietf.org/drafts/current/.

   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."

   This Internet-Draft will expire on November 13, 2015.

Copyright Notice

Table of Contents

1. Introduction

Different solutions have been developed and deployed to improve
service guarantees, both for multicast video traffic and Video on
Demand traffic. Most of these solutions are geared towards finding
an alternate path around one or more failed network elements (link,
node, path failures).

This document describes a mechanism for minimizing packet loss in a
network when node or link failures occur. Multicast only Fast Re-
Route (MoFRR) works by making simple changes to the way selected
routers use multicast protocols such as PIM and mLDP. No changes to
the protocols themselves are required. With MoFRR, in many cases,
multicast routing protocols don't necessarily have to depend on or

have to wait on unicast routing protocols to detect network failures,
see Section 5

On a Merge Point MoFRR logic determines a primary Upstream Multicast
Hop (UMH) and a secondary UMH and joins the tree via both
simultaneously. Data packets are received over the primary and
secondary paths. Only the packets from the primary UMH are accepted
and forwarded down the tree, the packets from the secondary UMH are
discarded. The UMH determination is different for PIM and mLDP and
explained in Section 4. When a failure is detected on the path to
the primary UMH, the repair occurs by changing the secondary UMH into
the primary and the primary into the secondary. Since the repair is
local, it is fast - greatly improving convergence times in the event
of node or link failures on the path to the primary UMH.

1.1. 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].

1.2. Terminology

MoFRR: Multicast only Fast Re-Route.

ECMP: Equal Cost Multi-Path.

mLDP: Multi-point Label Distribution Protocol.

PIM: Protocol Independent Multicast.

UMH: Upstream Multicast Hop, a candidate next-hop that can be used
   to reach the root of the tree.

tree: Either a PIM (S,G)/(*,G) tree or a mLDP P2MP or MP2MP LSP.

OIF: Outgoing InterFace, an interface used to forward multicast
   packets down the tree towards the receivers. Either a PIM
   (S,G)/(*,G) tree or a mLDP P2MP or MP2MP LSP.

LFA: Loop Free Alternate as defined in [RFC5286] In unicast Fast
   ReRoute, this is an alternate next-hop which can be used to reach
   a unicast destination without using the protected link or node.

Merge Point: A router that joins a multicast stream via two
   divergent upstream paths.

RPF: Reverse Path Forwarding.

RP: Rendezvous Point.

LSR: Label Switched Router.

BFD: Bidirectional Forwarding Detection.

IGP: Interior Gateway Protocol.

MVPN: Multicast Virtual Private Networks.

2. Basic Overview

The basic idea of MoFRR is for a Merge Point router to join a
multicast tree via two divergent upstream paths in order to get
maximum redundancy. The determination of this alternate upstream is
defined in Section 3.

In order to maximize robustness against any failure, the two paths
should be as diverse as possible. Ideally, they should not merge
upstream. Sometimes the topology guarantees maximal redundancy,
other times additional configuration or techniques are needed to
enforce it. See Section 6 for more discussion on the applicability
of MoFRR depending on the network topology.

A Merge Point router should only accept and forward on one of the
upstream paths at a time in order to avoid duplicate packet
forwarding. The selection of the primary and secondary UMH is done
by the MoFRR logic and normally based on unicast routing to find loop
free candidates. This is described in Section 4.

Note, the impact of additional amount of data on the network is
mitigated when tree membership is densely populated. When a part of
the network has redundant data flowing, join latency for new joining
members is reduced because its likely a tree Merge Point is not far
away.

3. Determination of the secondary UMH

The secondary UMH is a Loop Free Alternate (LFA) as per [RFC5286].

3.1. ECMP-mode MoFRR

If the IGP installs two ECMP paths to the source, then as per
[RFC5286] the LFA is a primary Next-hop. If the Multicast tree is
enabled for ECMP-Mode MoFRR, the router installs them as primary and
secondary UMH. Before the failure, only packets received from the
primary UMH path are processed while packets received from the
secondary UMH are dropped.

The selected primary UMH SHOULD be the same as if the MoFRR extension
was not enabled.

If more than two ECMP paths exist, one is selected as primary and
another as secondary UMH. The selection of the primary and secondary
is a local decision. Information from the IGP link-state topology
could be leveraged to optimize this selection such that the primary
and secondary path are maximal divergent and don't lead to the same
upstream node. Note that MoFRR does not restrict the number of UMH
paths that are joined. Implementations may use as many paths as are
configured.

3.2. Non-ECMP-mode MoFRR

A router X configured for non-ECMP-mode MoFRR for a Multicast tree
joins a primary path to its primary UMH and a secondary path to its
LFA UMH. In order to prevent control-plane loops a router MUST stop
joining the secondary UMH if this UMH is the only member in the OIF
list.

To illustrate the reason for this rule, let's consider the example in
FIG3. If PE1 and PE2 have received an IGMP request for a Multicast
tree, they will both join the primary path on their plane and a
secondary path to the neighbor PE. If their receivers would leave at
the same time, it could be possible for the Multicast tree on PE1 and
PE2 to never get deleted as each PE refresh each other via the
secondary path joins (remember that a secondary path join is not
distinguishable from a primary join).

4. Upstream Multicast Hop Selection

An Upstream Multicast Hop (UMH) is a candidate next-hop that can be
used to reach the root of the tree. This is normally based on
unicast routing to find loop free candidate(s). With MoFRR
procedures we select a primary and a backup UMH. The procedures for
determining the UMH are different for PIM and mLDP. See below;

4.1. PIM

The UMH selection in PIM is also known as the Reverse Path Forwarding
(RPF) procedure. Based on a unicast route lookup on either the
Source address or Rendezvous Point (RP) [RFC4601], an upstream
interface is selected for sending the PIM Joins/Prunes AND accepting
the multicast packets. The interface the packets are received on is
used to pass or fail the RPF check. If packets are received on an
interface that was not selected by the RPF procedure, or not the
primary, the packets are discarded.

4.2. mLDP

The UMH selection in mLDP also depends on unicast routing, but the
difference with PIM is that the acceptance of multicast packets is
based on MPLS labels and independent of the interface the packet is
received on. Using the procedures as defined in [RFC6388] an
upstream Label Switched Router (LSR) is elected. The upstream LSR
that was elected for a Label Switched Path (LSP) gets a unique local
MPLS Label allocated. Multicast packets are only forwarded if the
MPLS label matches the MPLS label that was allocated for that LSPs
(primary) upstream LSR.

5. Detecting Failures

Once the two paths are established, the next step is detecting a
failure on the primary path to know when to switch to the backup
path. This is a local issue but this section explore some
possibilities.

The first (and simplest) option is to detect the failure of the local
interface as it it's done for unicast Fast ReRoute. Detection can be
performed using the loss of signal or the loss of probing packets
(e.g. BFD). This option can be used in combination with the other
options as documented below. Just like for unicast fast reroute,
50msec switch-over is possible.

A second option consists of comparing the packets received on the
primary and secondary streams but only forwarding one of them -- the
first one received, no matter which interface it is received on.
Zero packet loss is possible for RTP-based streams.

A third option assumes a minimum known packet rate for a given data
stream. If a packet is not received on the primary RPF within this
time frame, the router assumes primary path failure and switches to
the secondary RPF interface. 50msec switch-over may be possible for
high rate stream (e.g. IP TV where SD video has a continuous inter-
packet gap of ~ 3msec) but in general the delay is dependant on the
rate of the multicast stream.

A fourth option leverages the significant improvements of the IGP
convergence speed. When the primary path to the source is withdrawn
by the IGP, the MoFRR-enabled router switches over to the backup
path, the UMH is changed to the secondary UMH. Since the secondary
path is already in place, and assuming it is disjoint from the
primary path, convergence times would not include the time required
to build a new tree and hence are smaller. Sub-second to sub-200msec
switch-over should be possible.

6. MoFRR applicability

   MoFRR applicability is topology dependent. The applicability is the
   same as LFA FRR which is discussed in [RFC6571].

   The following section will discuss MoFRR applicability to dual-plane
   network topologies.

6.1. Dual-Plane Topology

   MoFRR works best in dual-planes topologies as illustrated in the
   figures below. MoFRR may be enabled on any router in the network.
   In the figures below, MoFRR is shown enabled on the Provider Edge
   (PE) routers to illustrate one way in which the technology may be
   deployed.

```
                         S
                  P     / \ P
                       /   \
               ^     G1     R1  ^
              P     /         \  P
                   /           \
              G2----------R2     ^
              | \          | \  P
          ^   |  \         |  \
          P   |   G3----------R3
              |   |         |    |
              |   |         |    |  ^
             G4---|------R4   |  P
             ^  \  |         \  |
            P    \ |          \ |
              G5----------R5
             ^    |             | ^
            P     |             | P
                  |             |
                 Gi            Ri
                  \ \__      ^  /|
                   \   \   S1/ | ^
                  ^ \   ^\   /  |P2
                 P1 \ S2\_/__  |
                     \    /   \|
                      PE1      PE2
```
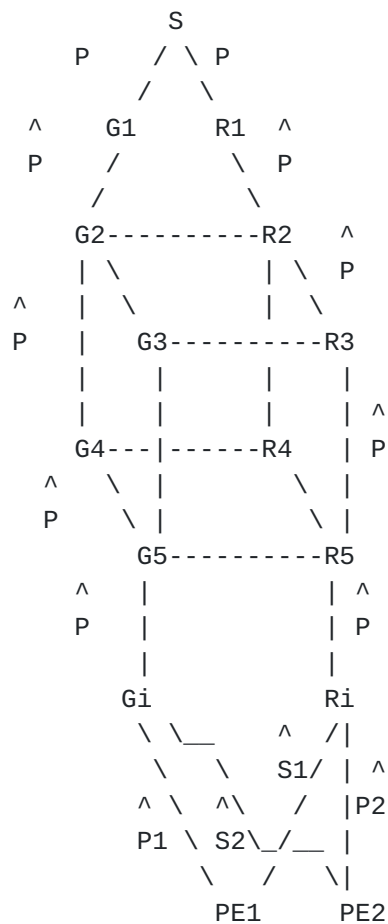
P = Primary path
S = Secondary path

        FIG1. Two-Plane Network Design

   The topology has two planes, a primary plane and a secondary plane
   that are fully disjoint from each other all the way into the POPs.
   This two plane design is common in service provider networks as it
   eliminates single point of failures in their core network. The links
   marked P indicate the normal (Primary) path of how the PIM joins flow
   from the POPs towards the source of the network. Multicast streams,
   especially for the densely watched channels, typically flow along
   both the planes in the network anyway.

   The only change MoFRR adds to this is on the links marked S where the
   PE routers join a secondary path to their secondary ECMP UMH. As a
   result of this, each PE router receives two copies of the same
   stream, one from the primary plane and the other from the secondary
   plane. As a result of normal UMH behavior, the multicast stream
   received over the primary path is accepted and forwarded to the

downstream receivers. The copy of the stream received from the
secondary UNH is discarded.

When a router detects a routing failure on the path to its primary
UMH, it will switch to the secondary UMH and accept packets for that
stream. If the failure is repaired the router may switch back. The
primary and secondary UMHs have only local context and not end-to-end
context.

As one can see, MoFRR achieves the faster convergence by pre-building
the secondary multicast tree and receiving the traffic on that
secondary path. The example discussed above is a simple case where
there are two ECMP paths from each PE device towards the source, one
along the primary plane and one along the secondary. In cases where
the topology is asymmetric or is a ring, this ECMP nature does not
hold, and additional rules have to be taken into account to choose
when and where to join the secondary path.

MoFRR is appealing in such topologies for the following reasons:

1. Ease of deployment and simplicity: the functionality is only
    required on the PE devices although it may be configured on all
    routers in the topology. Furthermore, each PE device can be
    enabled separately, there is no need for a network wide
    coordination in order to deploy MoFRR. Inter-operability testing
    is not required as there are no PIM or mLDP protocol change.

2. End-to-end failure detection and recovery: any failure along the
    path from the source to the PE can be detected and repaired with
    the secondary disjoint stream.(see Section 5 options 2, 3, 4)

3. Capacity Efficiency: as illustrated in the previous example, the
    Multicast trees corresponding to IPTV channels cover the backbone
    and distribution topology in a very dense manner. As a
    consequence, the secondary path graft into the normal Multicast
    trees (ie. trees signaled by PIM or mLDP without MoFRR extension)
    at the aggregation level and hence do not demand any extra
    capacity either on the distribution links or in the backbone.
    They simply use the capacity that is normally used, without any
    duplication. This is different from conventional FRR mechanisms
    which often duplicate the capacity requirements when the backup
    path crosses links/nodes which already carry the primary/normal
    tree and hence twice as much capacity is required.

4. Loop free: the secondary path join is sent on an ECMP disjoint
    path. By definition, the neighbor receiving this request is
    closer to the source and hence will not cause a loop.

The topology we just analyzed is very frequent and can be modelled as per Fig2. The PE has two ECMP disjoint paths to the source. Each ECMP path uses a disjoint plane of the network.

```
                    Source
                    /     \
                 Plane1  Plane2
                   |       |
                   A1      A2
                     \    /
                       PE
```
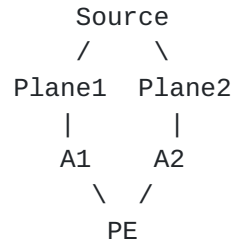
         FIG2. PE is dual-homed to Dual-Plane Backbone

Another frequent topology is described in Fig 3. PEs are grouped by pairs. In each pair, each PE is connected to a different plane. Each PE has one single shortest-path to a source (via its connected plane). There is no ECMP like in Fig 2. However, there is clearly a way to provide MoFRR benefits as each PE can offer a disjoint secondary path to the other plane PE (via the disjoint path).

MoFRR secondary neighbor selection process needs to be extended in this case as one cannot simply rely on using an ECMP path as secondary neighbor. This extension is referred to as non-ecmp extension and is described in Section 3.2.


```
                    Source
                    /     \
                 Plane1  Plane2
                   |       |
                   A1      A2
                   |       |
                 PE1----PE2
```
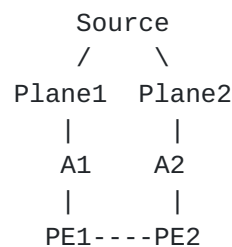
         FIG3. PEs are connected in pairs to Dual-Plane Backbone

6.2. Other Topologies

As mentioned in section Section 6.1, MoFRR works best in dual-plane topologies. If MoFRR is applied to none dual-plane networks, its possible that the secondary path is effected by the same failure that effected the primary path. In that case, there is no guarentee that the backup path will provide an un-interupted traffic flow of packets without loss or duplication.

6.3. Capacity Planning for MoFRR

   The previous section has described two very frequent designs (Fig 2
   and Fig 3) which provide maximum MoFRR benefits.

   Designers with topologies different than Fig2 and 3 can still benefit
   from MoFRR thanks to the use of capacity planning tools.

   Such tools are able to simulate the ability of each PE to build two
   disjoint branches of the same tree. This for hundreds of PEs and
   hundreds of sources.

   This allows to assess the MoFRR protection coverage of a given
   network, for a set of sources.

   If the protection coverage is deemed insufficient, the designer can
   use such tool to optimize the topology (add links, change IGP
   metrics).

6.4. PE nodes

   Many Service Providers devise their topology such that PEs have
   disjoint paths to the multicast sources. MoFRR leverages the
   existence of these disjoint paths without any PIM or mLDP protocol
   modification. Interoperability testing is thus not required. In
   such topologies, MoFRR only needs to be deployed on the PE devices.
   Each PE device can be enabled one by one.

6.5. Other Applications

   While all the examples in this document show the MoFRR applicability
   on PE devices, it is clear that MoFRR could be enabled on aggregation
   or core routers.

   MoFRR can be popular in Data Center network configurations. With the
   advent of lower cost ethernet and increasing port density in routers,
   there is more meshed connectivity than ever before. When using a
   3-level access, distribution, and core layers in a Data Center, there
   is a lot of inexpensive bandwidth connecting the layers. This will
   lend itself to more opportunities for ECMP paths at multiple layers.
   This allows for multiple layers of redundancy protecting link and
   node failure at each layer with minimal redundancy cost.

   Redundancy costs are reduced because only one packet is forwarded at
   every link along the primary and secondary data paths so there is no
   duplication of data on any link thereby providing make-before-break
   protection at a very small cost.

A MoFRR router only accepts packets from the primary path and
discards packets from the secondary path. For that reason,
management applications (like ping and mtrace) will not work when
verifying the secondary path.

The MoFRR principle may be applied to MVPNs.

7. IANA Considerations

This document makes no request of IANA.

8. Security Considerations

There are no security considerations for this design other than what
is already in the main PIM specification [RFC4601] and mLDP
specification [RFC6388].

9. Acknowledgments

Thanks to Dave Oran and Alvaro Retana for their review and comments
on this document.

The authors would like to especially acknowledge the contribution
from Dino Farinacci, John Zwiebel and Greg Shepherd for the genesis
of the MoFRR concept.

10. Contributor Addresses

Below is a list of other contributing authors in alphabetical order:

   Dino Farinacci
   Email: farinacci@gmail.com

   Wim Henderickx
   Alcatel-Lucent
   Copernicuslaan 50
   Antwerp 2018
   Belgium
   Email: wim.henderickx@alcatel-lucent.com

   Uwe Joorde
   Deutsche Telekom
   Dahlweg 100
   D-48153 Muenster
   Germany
   Email: Uwe.Joorde@telekom.de

   Nicolai Leymann
   Deutsche Telekom
   Winterfeldtstrasse 21
   Berlin 10781
   DE
   Email: N.Leymann@telekom.de

   Jeff Tantsura
   Ericsson
   300 Holger Way
   San Jose CA 95134
   USA
   Email: jeff.tantsura@ericsson.com

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.

   [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast
              Reroute: Loop-Free Alternates", RFC 5286, September 2008.

11.2. Informative References

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

   [RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas,
              "Label Distribution Protocol Extensions for Point-to-
              Multipoint and Multipoint-to-Multipoint Label Switched
              Paths", RFC 6388, November 2011.

   [RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B.,
              Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free
              Alternate (LFA) Applicability in Service Provider (SP)
              Networks", RFC 6571, June 2012.

Authors' Addresses

   Apoorva Karan
   Cisco Systems, Inc.
   3750 Cisco Way
   San Jose CA, 95134
   USA

   Email: apoorva@cisco.com

   Clarence Filsfils
   Cisco Systems, Inc.
   De kleetlaan 6a
   Diegem BRABANT 1831
   Belgium

   Email: cfilsfil@cisco.com

   IJsbrand Wijnands (editor)
   Cisco Systems, Inc.
   De Kleetlaan 6a
   Diegem 1831
   BE

   Email: ice@cisco.com

   Bruno Decraene
   Orange
   38-40 rue du General Leclerc
   Issy Moulineaux Cedex 9, 92794
   FR

   Email: bruno.decraene@orange.com