Network Working Group
Internet-Draft
Intended status: Informational
Expires: September 14, 2011

A. Karan
C. Filsfils
D. Farinacci
Cisco Systems, Inc.
B. Decraene
France Telecom
N. Leymann
U. Joorde
Deutsche Telekom
T. Telkamp
Cariden Technologies, Inc.

March 13, 2011

Multicast only Fast Re-Route
draft-karan-mofrr-01

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 draft 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.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

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

This draft 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. 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.

MoFRR involves transmitting a multicast join message from a receiver towards a source on a primary path and transmitting a secondary multicast join message from the receiver towards the source on a backup path. Data packets are received from the primary and secondary paths. The redundant packets are discarded at topology merge points using RPF checks. When a failure is detected on the primary path, the repair occurs by changing the interface on which packets are accepted to the secondary interface. Since the repair is local, it is fast - greatly improving convergence times in the event of node or link failures on the primary path.

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.

Primary Join : Multicast join message sent from receiver towards the source on the primary path.

Secondary Join : Multicast join message sent from receiver towards the source on the secondary path.

2. Basic Overview

MoFRR uses standard PIM JOIN/PRUNE messages to set up a primary and a secondary multicast forwarding path by establishing a primary and a secondary RPF interface on each router that receives a PIM join. The outgoing interface list remains the same.

Data packets are received from the primary and backup paths. Redundant packets received on the secondary RPF interface are discarded because of an RPF failure. When the router detects a forwarding failure in the primary path, it changes RPF to the secondary path and immediately has packets available to forward out each outgoing interface.

The primary and secondary MoFRR forwarding paths should not use the same nodes or links. This may be configured or determined by computations described in this document.

Note, the impact of additional amount of data on the network is mitigated when group membership is densely populated. When a part of the network has redundant data flowing, join latency for new joining members is reduced because joins don't have to propagate far to get to on-tree routers.

3. Topologies for MoFRR

MoFRR works best in topologies illustrated in the figure 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.

3.1. Dual-Plane Topology

```
                         S
                   PJ   / \ PJ
                       /   \
                ^    G1     R1  ^
               PJ    /       \  PJ
                    /         \
                   G2----------R2   ^
                   | \         | \  PJ
               ^   |  \        |  \
               PJ  |   G3----------R3
                   |    |      |    |
                   |    |      |    | ^
                   G4---|------R4   | PJ
                 ^   \  |        \  |
                 PJ   \ |         \ |
                       G5----------R5
                   ^   |           | ^
                   PJ  |           | PJ
                       |           |
                      Gi           Ri
                       \ \__    ^  /|
                        \   \  SJ1/ | ^
                       ^ \  ^\   /  |PJ2
                      PJ1 \SJ2\_/__ |
                           \   /   \|
                            PE1     PE2
   PJ = Primary Join
   SJ = Secondary Join

         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 PJ indicate the normal 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 anyways.

The only change MoFRR adds to this is on the links marked SJ where the PE routers send a secondary PIM joins to their ECMP neighbor towards the source. 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 multicast RPF checks the multicast stream received over the primary path is accepted and forwarded to the downstream links. The copy of the stream received

on the secondary path is discarded.

When a router detects a routing failure on its primary RPF interface, it will switch to the secondary RPF interface and accept packets on that stream. If the failure is repaired the router may switch back. The primary and secondary path 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 send the secondary PIM joins.

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. PEs not enabled for MoFRR do not see any change or degradation. Inter-operability testing is not required as there is no PIM 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.

3. Capacity Efficiency: as illustrated in the previous example, the PIM trees corresponding to IPTV channels cover the backbone and distribution topology in a very dense manner. As a consequence, the secondary joins graft into the normal PIM trees (ie. trees signaled by PIM 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 (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 PIM join is sent on an ECMP disjoint path. By definition, the neighbor receiving this secondary PIM join is closer to the source and hence will not send a PIM join back.

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 later in the document.

```
                             Source
                             /    \
                         Plane1  Plane2
                            |      |
                           A1     A2
                            |      |
                           PE1----PE2
```

FIG3. PEs are connected in pairs to Dual-Plane Backbone

4. 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.

A first 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 second 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 switchover is possible.

A third 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 RPF interface is changed to the secondary RPF interface. 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. Realistic availability requirements (sub-second to sub-200msec) should be possible.

A fourth option consists in leveraging connected link failure. This option makes sense when MoFRR is deployed across the network (not only at PE).

5. ECMP-mode MoFRR

If the IGP installs two ECMP paths to the source and if the (S, G) PIM state is enabled for ECMP-Mode MoFRR, the router installs them as primary RPF and secondary RPF. It sends a PIM join to both RPF entries. Only packets receive from the primary RPF entry are processed. Packets received from the secondary RPF are dropped (equivalent to an RPF failure).

The selected primary RPF interface should be the same as if MoFRR extension was not enabled.

If more than two ECMP paths exist, two are selected as primary and secondary RPF interfaces. Information from the IGP link-state topology could be leveraged to optimize this selection.

Note, MoFRR does not restrict the number of paths on which joins are sent. Implementations may use as many paths as are configured.

6. Non-ECMP-mode MoFRR

```
                    SourceS
                     /    \
                    /      \
                   Backbone
                  |        |
                  |        |
                  |        |
                  X--------N
```

Fig5. Non-ECMP-Mode MoFRR

X is configured for MoFRR for state (S, G)
R(X) is Xs RPF to S
N is a neighbor of X
R(N) is Ns RPF to S
xs represents the IGP metric from X to S
ns represents the IGP metric from N to S
xn represents the IGP metric from X to N

A router X configured for non-ECMP-mode MoFRR for (S, G) sends a primary PIM join to its primary RPF R(X) and a secondary PIM Join to a neighbor N if the following three conditions are met.

C1: xs < xn + ns
C2: ns < nx + xs
C3: X cannot send a secondary join to N if N is the only member of the OIF list

The first condition ensures that N is not on the primary branch from X to S.

The second condition ensures that X is not on the primary branch from N to S.

These two conditions ensure that at least locally the two paths are disjoint.

The third condition is required to break control-plane loops which could occur in some scenarios.

For example in FIG3, if PE1 and PE2 have received an igmp request for (S, G), they will both send a primary PIM join on their plane and a secondary PIM join to the neighbor PE. If their receivers would leave at the same time, it could be possible for the (S, G) states on PE1 and PE2 to never get deleted as each PE refresh each other via the secondary PIM joins (remember that a secondary PIM join is not distinguishable from a primary PIM join. MoFRR does not require any PIM protocol modification).

A control-plane loop occurs when two nodes keep a state forever due to the secondary joins they send to each other. This forever condition is not acceptable as no real receiver is connected to the nodes (directly via IGMP or indirectly via PIM). Rule 3 prevents this case as it prevents the mutual refresh of secondary joins and it applies it in the specific case where there is no real receiver connected.

6.1. Variation

Rule R3 can be removed if Rule 2 is restricted as follows:

R2p: ns < xs

This ensures that X only sends a secondary join to a neighbor N who is strictly closer to the source than X is. By reciprocity, N will thus never be able to send an sedondary join to the same source via X. The strictly smaller than is key here.

Note that this non-ECMP-mode MoFRR variation does not support the square topology and hence is less preferred.

7. Keep It Simple Principle

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 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. PEs not enabled for MoFRR do not see any change or degradation.

Multicast streams with Tight SLA requirements are often characterized by a continuous high packet rate (SD video has a continuous interpacket gap of ~ 3msec). MoFRR simply leverages the stream characteristic to detect any failures along the primary branch and switch-over on the secondary branch in a few 10s of msec.

8. Capacity Planning for MoFRR

As for LFA FRR (draft-ietf-rtgwg-lfa-applicability-00), MoFRR applicability is topology dependent.

In this document, we have 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 benefits 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).

9. 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.

The MoFRR behavior described for PIM are immediately applicable to MLDP. Alternate methods to detect failures such as MPLS-OAM or BFD may be considered.

The MoFRR principle may be applied to MVPNs.

The MoFRR principle may be applied to mLDP [I-D.ietf-mpls-ldp-p2mp]. The reader may simply switch the term secondary-PIM-Join by secondary-Label-Map message.

10. Security Considerations

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

11. Acknowledgments

The authors would like to thank John Zwiebel, Greg Shepherd and Dave Oran for their review of the draft.

12. References

12.1. Normative References

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

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

12.2. 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.

[I-D.ietf-mpls-ldp-p2mp] Minei, I., Wijnands, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-12 (work in progress), February 2011.

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

Dino Farinacci
Cisco Systems, Inc.
425 East Tasman Drive
San Jose CA, 95134
USA

Email: dino@cisco.com

Bruno Decraene
France Telecom
38-40 rue du General Leclerc
Issy Moulineaux cedex 9, 92794
FR

Email: bruno.decraene@orange-ftgroup.com

Nicolai Leymann
Deutsche Telekom
Winterfeldtstrasse 21
Berlin 10781
DE

Email: N.Leymann@telekom.de

Uwe Joorde
Deutsche Telekom
Winterfeldtstrasse 21
Berlin 10781
DE

Email: N.Leymann@telekom.de

Thomas Telkamp
Cariden Technologies, Inc.
888 Villa Street, Suite 500
Mountain View CA, 94041
USA

Email: telkamp@cariden.com