Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2014

A. Atlas, Ed.
R. Kebler
Juniper Networks
IJ. Wijnands
Cisco Systems, Inc.
A. Csaszar
G. Enyedi
Ericsson

July 12, 2013

An Architecture for Multicast Protection Using Maximally Redundant Trees

draft-atlas-rtgwg-mrt-mc-arch-02

Abstract

Failure protection is desirable for multicast traffic, whether signaled via PIM or mLDP. Different mechanisms are suitable for different use-cases and deployment scenarios. This document describes the architecture for global protection (aka multicast live-live) and for local protection (aka fast-reroute).

The general methods for global protection and local protection using alternate-trees are dependent upon the use of Maximally Redundant Trees. Local protection can also tunnel traffic in unicast tunnels to take advantage of the routing and fast-reroute mechanisms available for IP/LDP unicast destinations.

The failures protected against are single link or node failures. While the basic architecture might support protection against shared risk group failures, algorithms to dynamically compute MRTs supporting this are for future study.

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 January 13, 2014.

Copyright Notice

Table of Contents

1. Introduction 4
1.1. Maximally Redundant Trees (MRTs) 4
1.2. MRTs and Multicast 6
2. Terminology 6
3. Use-Cases and Applicability 8
4. Global Protection: Multicast Live-Live 9
4.1. Creation of MRMTs 10
4.2. Traffic Self-Identification 11
4.2.1. Merging MRMTs for PIM if Traffic Doesn't Self-Identify 12
4.3. Convergence Behavior 13
4.4. Inter-area/level Behavior 14
4.4.1. Inter-area Node Protection with 2 border routers . . . 15
4.4.2. Inter-area Node Protection with > 2 Border Routers . . 16
4.5. PIM 17
4.5.1. Traffic Handling: RPF Checks 17
4.6. mLDP 17
5. Local Repair: Fast-Reroute 17
5.1. PLR-driven Unicast Tunnels 18
5.1.1. Learning the MPs 19
5.1.2. Using Unicast Tunnels and Indirection 19
5.1.3. MP Alternate Traffic Handling 20
5.1.4. Merge Point Reconvergence 21
5.1.5. PLR termination of alternate traffic 21
5.2. MP-driven Unicast Tunnels 21
5.3. MP-driven Alternate Trees 22
6. Acknowledgements 23
7. IANA Considerations 23
8. Security Considerations 23
9. Appendix A 23
9.1. MP-driven Alternate Trees 23
9.1.1. PIM details for Alternate-Trees 26
9.1.2. mLDP details for Alternate-Trees 26
9.1.3. Traffic Handling by PLR 26
9.2. Methods Compared for PIM 27
9.3. Methods Compared for mLDP 27
10. References 27
10.1. Normative References 27
10.2. Informative References 28
Authors' Addresses 29

1. Introduction

This document describes how the algorithms in [I-D.enyedi-rtgwg-mrt-frr-algorithm], which are used in [I-D.ietf-rtgwg-mrt-frr-architecture] for unicast IP/LDP fast-reroute, can be used to provide protection for multicast traffic. It specifically applies to multicast state signaled by PIM[RFC4601] or mLDP[RFC6388]. There are additional protocols that depend upon these (e.g. VPLS, mVPN, etc.) and consideration of the applicability to such traffic will be in a future version.

In this document, global protection is used to refer to the method of having two maximally disjoint multicast trees where traffic may be sent on both and resolved by the receiver. This is similar to the ability with RSVP-TE LSPs to have a primary and a hot standby, except that it can operate in 1+1 mode. This capability is also referred to as multicast live-live and is a generalized form of that discussed in [I-D.ietf-rtgwg-mofrr]. In this document, local protection refers to the method of having alternate ways of reaching the pre-identified merge points upon detection of a local failure. This capability is also referred to as fast-reroute.

This document describes the general architecture, framework, and trade-offs of the different approaches to solving these general problems. It will recommend how to generally provide global protection and local protection for mLDP and PIM traffic. Where protocol extensions are necessary, they will be defined in separate documents as follows.

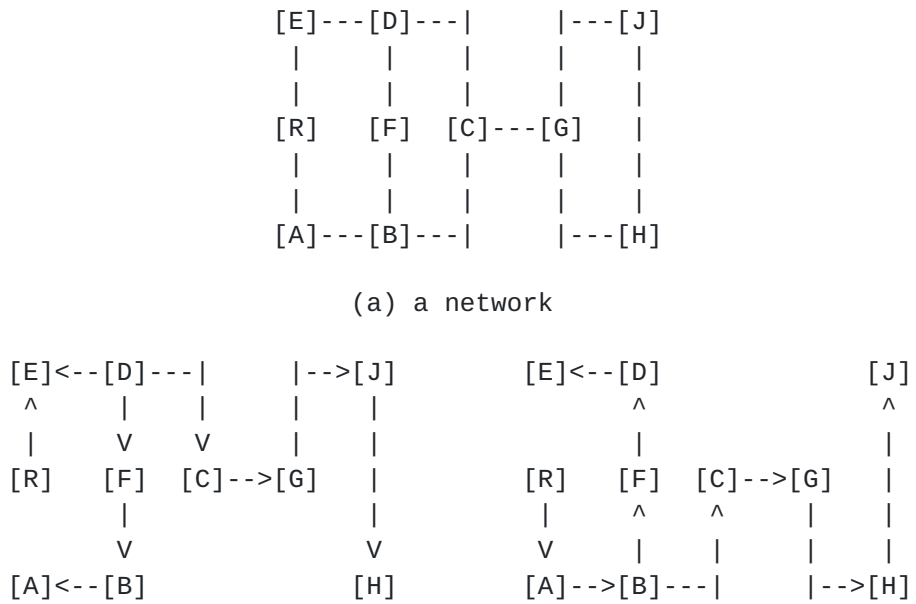
- Global 1+1 Protection Using PIM
- Global 1+1 Protection Using mLDP
- Local Protection Using mLDP: [I-D.wijnands-mpls-mldp-node-protection]This document describes how to provide node-protection and the necessary extensions using targeted LDP session.
- Local Protection Using PIM

1.1. Maximally Redundant Trees (MRTs)

Maximally Redundant Trees (MRTs) are described in [I-D.enyedi-rtgwg-mrt-frr-algorithm]; here we only give a brief description about the concept. A pair of MRTs is a pair of directed spanning trees (red and blue tree) with a common root, directed so that each node can be reached from the root on both trees. Moreover, these trees are redundant, since they are constructed so that no

single link or single node failure can separate any node from the root on both trees, unless that failed link or node is splitting the network into completely separated components (e.g. the link or node was a cut-edge or cut-vertex).

Although for multicast, the arcs (directed links) are directed away from the root instead of towards the root, the same MRT computations are used and apply. This is similar to how multicast uses unicast routing's next-hops as the upstream-hops. Thus this definition slightly differs from the one presented in [I-D.enyedi-rtgwg-mrt-frr-algorithm], since the arcs are directed away and not towards the root. When we need two paths towards a given destination and not two away from it (e.g. for unicast detours for local repair solutions), we only need to reverse the arcs from how they are used for the unicast routing case; thus constructing MRTs towards or away from the root is the same problem. A pair of MRTs is depicted in Figure 1.



(a) a network

(b) Blue MRT of root R

(c) Red MRT of root R

Figure 1: A network and two MRTs found in it

It is important to realize that this redundancy criterion does not imply that, after a failure, either of the MRTs remains intact, since a node failure must affect any spanning tree. Redundancy here means that there will be a set of nodes, which can be reached along the blue MRT, and there will be another set, which remains reachable along the red MRT. As an example, suppose that node F goes down; that would separate B and A on the blue MRT and D and E on the red

MRT. Naturally, it is possible that the intersection of these two sets is not empty, e.g. C, G, H and J will remain reachable on both MRTs. Additionally, observe that a single link can be used in both of the trees in different directions, so even a link failure can cut both trees. In this example, the failure of link F<->B leads to the same reachability sets.

Finally, it is critical to recall that a pair of MRTs is always constructed together and they are not SPTs. While it would be useful to have an algorithm that could find a redundant pair for a given tree (e.g. for the SPT), that is impossible in general. Moreover, if there is a failure and at least one of the trees change, the other tree may need to change as well. Therefore, even if a node still receives the traffic along the red tree, it cannot keep the old red tree, and construct a blue pair for it; there can be reconfiguration in cases when traditional shortest-path-based-thinking would not expect it. To converge to a new pair of disjoint MRTs, it is generally necessary to update both the blue MRT and the red MRT.

The two MRTs provide two separate forwarding topologies that can be used in addition to the default shortest-path-tree (SPT) forwarding topology (usually MT-ID 0). There is a Blue MRT forwarding topology represented by one MT-ID; similarly there is a Red MRT forwarding topology represented by a different MT-ID. Naturally, a multicast protocol is required to use the forwarding topologies information to build the desired multicast trees. The multicast protocol can simply request appropriate upstream interfaces, but include the MT-ID when needed.

1.2. MRTs and Multicast

Maximally Redundant Trees (MRT) provide two advantages for protecting multicast traffic. First, for global protection, MRTs are precisely what needs to be computed to have maximally redundant multicast distribution trees. Second, for local repair, MRTs ensure that there will protection to the merge points; the certainty of a path from any merge point to the PLR that avoids the failure node allows for the creation of alternate trees.

A known disadvantage of MRT, and redundant trees in general, is that the trees do not necessarily provide shortest detour paths. Modeling is underway to investigate and compare the MRT lengths for the different algorithm options [I-D.enyedi-rtgwg-mrt-frr-algorithm].

2. Terminology

2-connected: A graph that has no cut-vertices. This is a graph that requires two nodes to be removed before the network is partitioned.

2-connected cluster: A maximal set of nodes that are 2-connected.

2-edge-connected: A network graph where at least two links must be removed to partition the network.

ADAG: Almost Directed Acyclic Graph - a graph that, if all links incoming to the root were removed, would be a DAG.

block: Either a 2-connected cluster, a cut-edge, or an isolated vertex.

cut-link: A link whose removal partitions the network. A cut-link by definition must be connected between two cut-vertices. If there are multiple parallel links, then they are referred to as cut-links in this document if removing the set of parallel links would partition the network.

cut-vertex: A vertex whose removal partitions the network.

DAG: Directed Acyclic Graph - a graph where all links are directed and there are no cycles in it.

GADAG: Generalized ADAG - a graph that is the combination of the ADAGs of all blocks.

Maximally Redundant Trees (MRT): A pair of trees where the path from any node X to the root R along the first tree and the path from the same node X to the root along the second tree share the minimum number of nodes and the minimum number of links. Each such shared node is a cut-vertex. Any shared links are cut-links. Any RT is an MRT but many MRTs are not RTs.

Maximally Redundant Multicast Trees (MRMT): A pair of multicast trees built of the sub-set of MRTs that is needed to reach all interested receivers.

network graph: A graph that reflects the network topology where all links connect exactly two nodes and broadcast links have been transformed into the standard pseudo-node representation.

Redundant Trees (RT): A pair of trees where the path from any node X to the root R along the first tree is node-disjoint with the path from the same node X to the root along the second tree. These can be computed in 2-connected graphs.

Merge Point (MP): For local repair, a router at which the alternate traffic rejoins the primary multicast tree. For global protection, a router which receives traffic on multiple trees and must decide which stream to forward on.

Point of Local Repair (PLR): The router that detects a local failure and decides whether and when to forward traffic on appropriate alternates.

MT-ID: Multi-topology identifier. The default shortest-path-tree topology is MT-ID 0.

MultiCast Ingress (MCI): Multicast Ingress, the node where the multicast stream enters the current transport technology (MPLS-mLDP or IP-PIM) domain. This maybe the router attached to the multicast source, the PIM Rendezvous Point (RP) or the mLDP Root node address.

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

Stream Selection: The process by which a router determines which of the multiple primary multicast streams to accept and forward. The router can decide on a packet-by-packet basis or simply per-stream. This is done for global protection 1+1 and described in [I-D.ietf-rtgwg-mofrr].

MultiCast Egress (MCE): Multicast Egress, a node where the multicast stream exists the current transport technology (MPLS-mLDP or IP-PIM) domain. This is usually a receiving router that may forward the multicast traffic on towards receivers based upon IGMP or other technology.

3. Use-Cases and Applicability

Protection of multicast streams has gained importance with the use of multicast to distribute video, including live video such as IP-TV. There are a number of different scenarios and uses of multicast that require protection. A few preliminary examples are described below.

- When video is distributed via IP or MPLS for a cable application, it is desirable to have global protection 1+1 so that the customer-perceived impact is limited. A QAM can join two multicast groups and determine which stream to use based upon the stream quality. A network implementing this may be custom-engineered for this particular purpose.

- In financial markets, stock ticker data is distributed via multicast. The loss of data can have a significant financial impact. Depending on the network, either global protection 1+1 or local protection can minimize the impact.

- Several solutions exist for updating software or firmwares of a large number of end-user or operator-owned networking equipment that are based on IP multicast. Since IP multicast is based on datagram transport so taking care of lost data is cumbersome and decreases the advantages offered by multicast. Solutions may rely on sending the updates several times: a properly protected network may result in that less repetitions are required. Other solutions rely on the recipent asking for lost data segments explicitly on-demand. A network failure could cause data loss for a significant number of receivers, which in turn would start requesting the the lost data in a burst that could overload the server. Properly engineered multicast fast reroute would minimise such impacts.

- Some providers offer multicast VPN services to their customers. SLAs between the customer and provider may set low packet loss requirements. In such cases interruptions longer than the outage timescales targeted by FRR could cause direct financial losses for the provider.

Global protection 1+1 uses maximally redundant multicast trees (MRMTs) to simultaneously distribute a multicast stream on both MRMTs. The disadvantage is the extra state and bandwidth requirements of always sending the traffic twice. The advantage is that the latency of each MRMT can be known and the receiver can select the best stream.

Local protection provides a patch around the fault while the multicast tree reconverges. When PLR replication is used, there is no extra multicast state in the network, but the bandwidth requirements vary based upon how many potential merge-points must be provided. When alternate-trees are used, there is extra multicast state but the bandwidth requirements on a link can be minimized to no more than once for the primary multicast tree traffic and once for the alternate-tree traffic.

4. Global Protection: Multicast Live-Live

In MoFRR [I-D.ietf-rtgwg-mofrr], the idea of joining both a primary and a secondary tree is introduced with the requirement that the primary and secondary trees be link and node disjoint. This works well for networks where there are dual-planes, as explained in [I-D.ietf-rtgwg-mofrr]. For other networks, it is still desirable to

have two disjoint multicast trees and allow a receiver to join both and make its own decision about which traffic to accept.

Using MRTs gives the ability to guarantee that the two trees are as disjoint as possible and dynamically recomputed whenever the topology changes. The MRTs used are rooted at the MultiCast Ingress (MCI). One multicast tree is created using the Blue MRT forwarding topology. The second multicast tree is created using the Red MRT forwarding topology. This can be accomplished by specifying the appropriate MT-ID associated with each forwarding topology.

There are four different aspects of using MRTs for 1+1 Global Protection that are necessary to consider. They are as follows.

1. Creation of the maximally redundant multicast trees (MRMTs) based upon the forwarding topologies.

2. Traffic Identification: How to handle traffic when the two MRMTs overlap due to a cut-vertex or cut-link.

3. Convergence: How to converge after a network change and get back to a protected state.

4. Inter-area/inter-level Behavior: How to compute and use MRMTs when the multicast source is outside the area/level and how to provide border-router protection.

4.1. Creation of MRMTs

The creation of the two maximally redundant multicast trees occurs as described below. This assumes that the next-hops to the MCI associated with the Blue and Red forwarding topologies have already been computed and stored.

1. A receiving router determines that it wants to join both the Blue tree and the Red tree. The details on how it does this decision are not covered in this document and could be based on configuration, additional protocols, etc.

2. The router selects among the Blue next-hops an Upstream Multicast Hop (UMH) to reach the MCI node. The router joins the tree towards the selected UMH including a multi-topology id (MT-ID) identifying the Blue MRT.

3. The router selects among the Red next-hops an Upstream Multicast Hop (UMH) to reach the MCI node. The router joins the tree towards the selected UMH including a multi-topology id (MT-ID) identifying the Red MRT.

4. When a router receives a tree setup request specifying a particular MT-ID (e.g. Color), then the router selects among the Color next-hops to the MCI a UMH node, creates the necessary multicast state, and joins the tree towards the UMH node.

4.2. Traffic Self-Identification

Two maximally redundant trees will share any cut-vertices and cut-links in the network. In the multicast global protection 1+1 case, this means that the potential single failures of the other nodes and links in the network are still protected against. If each cut-vertex cannot associate traffic to a particular MRMT, then the traffic would be incorrectly replicated to both MRMT resulting in complete duplication of traffic. An example of such MRTs is given earlier in Figure 1 and repeated below in Figure 2, where there are two cut-vertices C and G and a cut-link C<->G.

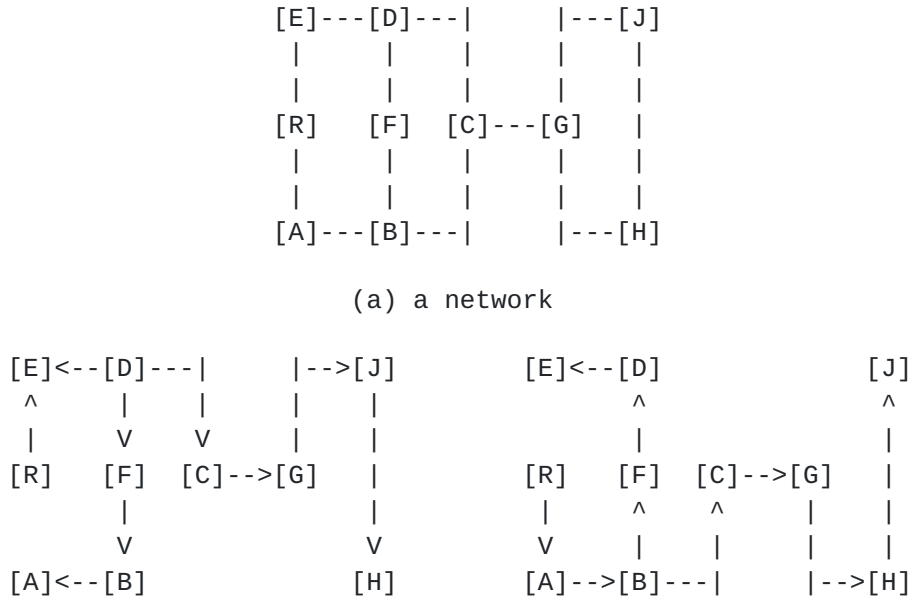


Figure 2: A network and two MRTs found in it

In this example, traffic from the multicast source R to a receiver G, J, or H will cross link C<->G on both the Blue and Red MRMTs. When this occurs, there are several different possibilities depending upon protocol.

mLDP: Different label bindings will be created for the Blue and Red MRMTs. As specified in [I-D.iwijnand-mpls-mldp-multi-topology], the P2MP FEC Element will use the MT IP Address Family to encode the Root node address and MRT T-ID. Each MRMT will therefore have a different P2MP FEC Element and be assigned an independent label.

PIM: There are three different ways to handle IP traffic forwarded based upon PIM when that traffic will overlap on a link.

A. Different Groups: If different multicast groups are used for each MRMT, then the traffic clearly indicates which MRMT it belongs to. In this case, traffic on the Blue MRMT would use multicast group G-blue and traffic on the Red MRMT would use multicast group G-red.

B. Different Source Loopbacks: Another option is to use different IP addresses for the source S, so S might announce S-red and S-blue. In this case, traffic on the Blue MRMT would have an IP source of S-blue and traffic on the Red MRMT would have an IP source of S-red.

C. Stream Selection and Merging: The third option, described in Section 4.2.1, is to have a router that gets (S,G) Joins for both the Blue MT-ID and the Red MT-ID merge those into a single tree. The router may need to select which upstream stream to use, just as if it were a receiving router.

There are three options presented for PIM. The most appropriate will depend upon deployment scenario as well as router capabilities.

4.2.1. Merging MRMTs for PIM if Traffic Doesn't Self-Identify

When traffic doesn't self-identify, the cut-vertices must follow specific rules to avoid traffic duplication. This section describes that behavior which allows the same (S,G) to be used for both the Blue MT-ID and Red MT-ID (e.g. when the traffic doesn't self-identify as to its MT-ID).

The behavior described in this section differs from the conflict resolution described in [RFC6420] because these rules apply to the Global Protection 1+1 case. Specifically, it is not sufficient for a upstream router to pick only one of the two MT-IDs to join because that does not maximize the protection provided.

As described in [RFC6420], a router that receives (S,G) Joins for both the Blue MT-ID and the Red MT-ID can merge the set of downstream interfaces in its forwarding entry. Unlike the procedures defined in [RFC6420], the router must send a Join upstream for each MT-ID. If a

router has different upstream interfaces for these MRMTs, then the router will need to do stream selection and forward the selected stream to its outgoing interfaces, just as if it were an MCE. The stream selection methods of detecting failures and handle traffic discarding are described in [I-D.ietf-rtgwg-mofrr].

This method does not work if the MRMTs merge on a common LAN with different upstream routers. In this case, the traffic cannot be distinguished on the LAN and will result in duplication on the LAN. The normal PIM Assert procedure would stop one of the upstream routers from transmitting duplicates onto the LAN once it is detected. This, in turn, may cause the duplicate stream to be pruned back to the source. Thus, end-to-end protection in this case of the MRMTs converging on a single LAN with different upstream interfaces can only be accomplished by the methods of traffic self-identification.

4.3. Convergence Behavior

It is necessary to handle topology changes and get back to having two MRMTs that provide global protection. To understand the requirements and what can be computed, recall the following facts.

a. It is not generally possible to compute a single tree that is maximally redundant to an existing tree.

b. The pair of MRTs must be computed simultaneously.

c. After a single link or node failure, there is one set of nodes that can be reached from the root on the Blue MRMT and a second set of nodes that can be reached from the root on the Red MRMT. If the failure wasn't a cut-vertex or cut-edge, all nodes will be in at least one of these two sets.

To gracefully converge, it is necessary to never have a router where both its red MRMT and blue MRMT are broken. There are three different ways in which this could be done. These options are being more fully explored to see which is most practical and provides the best set of trade-offs.

Ordered Convergence When a single failure occurs, each receiver determines whether it was affected or unaffected. First, the affected receivers identify the broken MRMT color (e.g. blue) and join the MRMT via their new UMH for that MRT color. Once the affected receivers receive confirmation that the new MRMT has been successfully created back to the MCI, then the affected receivers switch to using that MRMT. The affected receivers tear down the old broken MRMT state and join the MRMT via their new UMH for the

other MRT color (e.g. red). Finally, once the affected receivers receive confirmation that the new MRMT has been successfully created back to the MCI, the affected receivers can tear down the old working MRMT state. Once the affected receivers have updated their state, the unaffected receivers need to also do the same staging - first joining the MRMT via their new UMH for the Blue MRT, waiting for confirmation, switching to using traffic from the Blue MRMT, tearing down the old Blue MRMT state, joining the MRMT via their new UMH for the Red MRT, waiting for confirmation, and tearing down the old Red MRMT state. There are complexities remaining, such as determining how an Unaffected Receiver decides that the Affected Receivers are done. When the topology change isn't a failure, all receivers are unaffected and the same process can apply.

Protocol Make-Before-Break In the control plane, a router joins the tree on the new Blue topology but does not stop receiving traffic on the old Blue topology. Once traffic is observed from the new Blue UMH, then the router accepts traffic on the new Blue UMH and removes the old Blue UMH. This behavior can happen simultaneously with both Blue and Red forwarding topologies. An advantage is that it works regardless of the type of topology change and existing traffic streams aren't broken. Another advantage is that the complexity is limited and this method is well understood. The disadvantage is that the number of traffic-affecting events depends upon the number of hops to the MCI.

Multicast Source Make-Before-Break On a topology change, routers would create new MRMTs using new MRT forwarding state and leaving the old MRMTs as they are. After the new MRMTs are complete, the multicast source could switch from sending on the old MRMTs to sending on the new MRMTs. After a time, the old MRMTs could be torn down. There are a number of details to still investigate.

4.4. Inter-area/level Behavior

A source outside of the IGP area/level can be treated as a proxy node. When the join request reaches a border router (whether ABR for OSPF or LBR for ISIS), that border router needs to determine whether to use the Blue or Red forwarding topology in the next selected area/level.

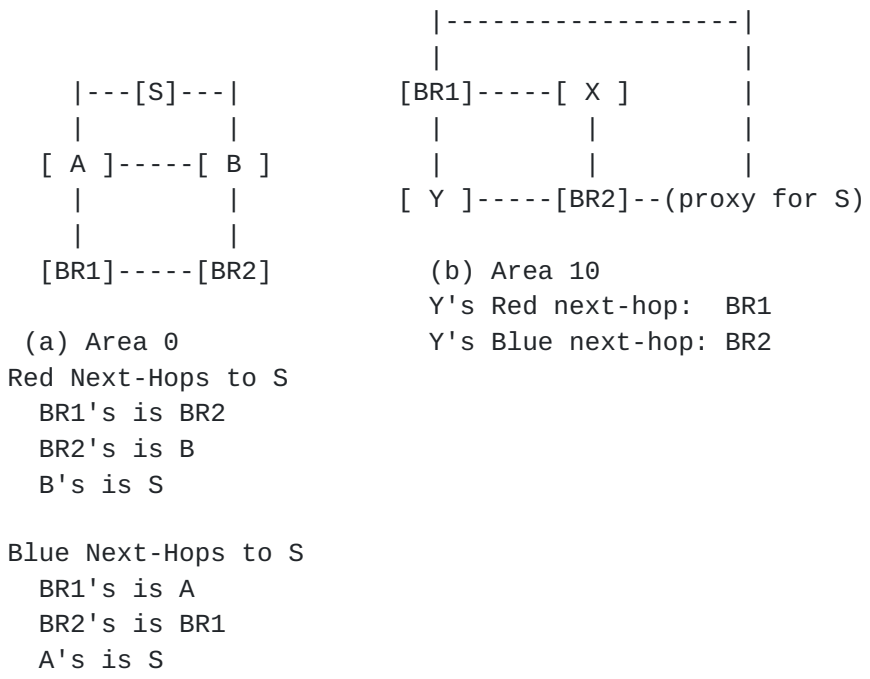


Figure 3: Inter-area Selection - next-hops towards S

Achieving maximally node-disjoint trees across multiple areas is hard due to the information-hiding and abstraction. If there is only one border router, it is trivial but protection of the border router is not possible. With exactly 2 border routers, inter-area/level node protection is reasonably straightforward but can require that the BR rewrite the (S,G) for PIM. With more than 2 border routers, inter-area node protection is possible at the cost of additional bandwidth and border router complexity. These two solutions are described in the following sub-sections.

4.4.1. Inter-area Node Protection with 2 border routers

If there are exactly two border routers between the areas, then the solution and necessary computation is straightforward. In that specific case, each BR knows that only the other BR must specifically be avoided in the second area when a forwarding topology is selected. As described in [I-D.enyedi-rtgwg-mrt-frr-algorithm], it is possible for a node X to determine whether the Red or Blue forwarding topology should be used to reach a node D while avoiding another node Y.

The results of this computation and the resulting changes in MT-ID from Red to Blue or Blue to Red are illustrated in Figure 3. It shows an example where BR1 must modify joins received from Area 10 for the Red MT-ID to use the Blue MT-ID in Area 0. Similarly, BR2 must modify joins received from Area 10 for the Blue MT-ID to use the

Red MT-ID in Area 0.

For mLDP, modifying the MT-ID in the control-plane is all that is needed. For PIM, if the same (S,G) is used for both the Blue MT-ID and the Red MT-ID, then only control-plane changes are needed. However, for PIM, if different group IDs (e.g. G-red and G-blue) or different source loopback addresses (S-red and S-blue) are used, it is necessary to modify the traffic to reflect the MT-ID included in the join message received on that interface. An alternative could be to use an MPLS label that indicates the MT-ID instead of different group IDs or source loopback addresses.

To summarize the necessary logic, when a BR1 receives a join from a neighbor in area N to a destination D in area M on the Color MT-ID, the BR1:

a. Identifies the BR2 at the other end of the proxy node in area N.

b. Determines which forwarding topology may avoid BR2 to reach D in area M. Refer to that as Color-2 MT-ID.

c. Uses Color-2 MT-ID to determine the next-hops to S. When a join is sent upstream, the MT-ID used is that for Color-2.

4.4.2. Inter-area Node Protection with > 2 Border Routers

If there are more than two BRs between areas, then the problem of ensuring inter-area node-disjointness is not solved. Instead, once a request to join the multicast tree has been received by a BR from an area that isn't closest to the multicast source, the BR must join both the Red MT-ID and the Blue MT-ID in the area closest to the multicast source. Regardless of what single link or node failure happens, each BR will receive the multicast stream. Then, the BR can use the stream-selection techniques specified in [I-D.ietf-rtgwg-mofrr] to pick either the Blue or Red stream and forward it to downstream routers in the other area. Each of the BRs for the other area should be attached to a proxy-node representing the other area.

This approach ensures that a BR will receive the multicast stream in the closest area as long as the single link or node failure isn't a single point of failure. Thus, each area or level is independently protected. The BR is required to be able to select among the multicast streams and, if necessary for PIM, translate the traffic to contain the correct (S,G) for forwarding.

4.5. PIM

Capabilities need to be exchanged to determine that a neighbor supports using MRT forwarding topologies with PIM. Additional signaling extensions are not necessary to PIM to support Global Protection. [RFC6420] already defines how to specify an MT-ID as a Join Attribute.

4.5.1. Traffic Handling: RPF Checks

For PIM, RPF checks would still be enabled by the control plane. The control plane can program different forwarding entries on the G-blue incoming interface and on the G-red incoming interface. The other interfaces would still discard both G-blue and G-red traffic.

The receiver would still need to detect failures and handle traffic discarding as is specified in [I-D.ietf-rtgwg-mofrr].

4.6. mLDP

Capabilities need to be exchanged to determine that a neighbor supports using MRT forwarding topologies with mLDP. The basic mechansims for mLDP to support multi-topology are already described in [I-D.iwijnand-mpls-mldp-multi-topology]. It may be desirable to extend the capability defined in this draft to indicate that MRT is or is not supported.

5. Local Repair: Fast-Reroute

Local repair for multicast traffic is different from unicast in several important ways.

- There is more than a single final destination. The full set of receiving routers may not be known by the PLR and may be extremely large. Therefore, it makes sense to repair to the immediate next-hops for link-repair and the next-next-hops for node-repair. These are the potential merge points (MPs).

- If a failure cannot be positively identified as a node-failure, then it is important to repair to the immediate next-hops since they may have receivers attached.

- If a failure cannot be positively identified as a link-failure and node protection is desired, then it is important to repair to the next-next-hops since they may not receive traffic from the immediate next-hops.

- Updating multicast forwarding state may take significantly longer than updating unicast state, since the multicast state is updated tree by tree based on control-plane signaling.

- For tunnel-based IP/LDP approaches, neither the PLR nor the MP may be able to specify which interface the alternate traffic will arrive at the MP on. The simplest reason is the unicast forwarding includes the use of ECMP and the path selection is based upon internal router behavior for all paths between the PLR and the MP.

For multicast fast-reroute, there are three different mechanisms that can be used. As long as the necessary signaling is available, these methods can be combined in the same network and even for the same PLR and failure point.

PLR-driven Unicast Tunnels: The PLR learns the set of MPs that need protection. On a failure, the PLR replicates the traffic and tunnels it to each MP using the unicast route. If desired, an RSVP-TE tunnel could be used instead of relying upon unicast routing.

MP-driven Unicast Tunnels: Each MP learns the identity of the PLR. Before failure, each MP independently signals to the PLR the desire for protection and other information to use. On a failure, the PLR replicates the traffic and tunnels it to each MP using the unicast route. If desired, an RSVP-TE tunnel could be used instead of relying upon unicast routing.

MP-driven Alternate Trees: Each MP learns the identity of the PLR and the failure point (node and interface) to be protected against. Each MP selects an upstream interface and forwarding topology where the path will avoid the failure point; each MP signals a join towards that upstream interface to create that state.

Each of these options is described in more detail in their respective sections. Then the methods are compared and contrasted for PIM and for mLDP.

5.1. PLR-driven Unicast Tunnels

With PLR-driven unicast tunnels, the PLR learns the set of merge points (MPs) and, on a locally detected failure, uses the existing unicast routing to tunnel the multicast traffic to those merge points. The failure being protected against may be link or node failure. If unicast forwarding can provide an SRLG-protecting alternate, then SRLG-protection is also possible.

There are five aspects to making this work.

1. PLR needs to learn the MPs and their associated MPLS labels to create protection state.

2. Unicast routing has to offer alternates or have dedicated tunnels to reach the MPs. The PLR encapsulates the multicast traffic and directs it to be forwarded via unicast routing.

3. The MP must identify alternate traffic and decide when to accept and forward it or drop it.

4. When the MP reconverges, it must move to its new UMH using make-before-break so that traffic loss is minimized.

5. The PLR must know when to stop sending traffic on the alternates.

5.1.1. Learning the MPs

If link-protection is all that is desired, then the PLR already knows the identities of the MPs. For node-protection, this is not sufficient. In the PLR-driven case, there is no direct communication possible between the PLR and the next-next-hops on the multicast tree. (For mLDP, when targeted LDP sessions are used, this is considered to be MP-driven and is covered in Section 5.2.)

In addition to learning the identities of the MPs, the PLR must also learn the MPLS label, if any, associated with each MP. For mLDP, a different label should be supplied for the alternate traffic; this allows the MP to distinguish between the primary and alternate traffic. For PIM, an MPLS label is used to identify that traffic is the alternate. The unicast tunnel used to send traffic to the MP may have penultimate-hop-popping done; thus without an explicit MPLS label, there is no certainty that a packet could be conclusively identified as primary traffic or as alternate traffic.

A router must tell its UMH the identity of all downstream multicast routers, and their associated alternate labels, on the particular multicast tree. This clearly requires protocol extensions. The extensions for PIM are given in [I-D.kebler-pim-mrt-protection].

5.1.2. Using Unicast Tunnels and Indirection

The PLR must encapsulate the multicast traffic and tunnel it towards each MP. The key point is how that traffic then reaches the MP. There are basically two possibilities. It is possible that a dedicated RSVP-TE tunnel exists and can be used to reach the MP for just this traffic; such an RSVP-TE tunnel would be explicitly routed

to avoid the failure point. The second possibility is that the packet is tunneled via LDP and uses unicast routing. The second case is explored here.

It is necessary to assume that unicast LDP fast-reroute [I-D.ietf-rtgwg-mrt-frr-architecture][RFC5714][RFC5286] is supported by the PLR. Since multicast convergence takes longer than unicast convergence, the PLR may have two different routes to the MP over time. When the failure happens, the PLR will have an alternate, whether LFA or MRT, to reach the MP. Then the unicast routing converges and the PLR will have a new primary route to the MP. Once the routing has converged, it is important that alternate traffic is no longer carried on the MRT forwarding topologies. This rule allows the MRT forwarding topologies to reconverge and be available for the next failure. Therefore, it is also necessary for the tunneled multicast traffic to move from the alternate route to the new primary route when the PLR reconverges. Therefore, the tunneled multicast traffic should use indirection to obtain the unicast routing's current next-hops to the MP. If physical indirection is not feasible, then when the unicast LIB is updated, the associated multicast alternate tunnel state should be as well.

When the PLR detects a local failure, the PLR replicates each multicast packet, swaps or adds the alternate MPLS label needed by the MP, and finally pushes the appropriate label for the MP based upon the outgoing interface selected by the unicast routing.

For PIM, if no alternate labels are supplied by the MPs, then the multicast traffic could be tunneled in IP. This would require unicast IP fast-reroute.

5.1.3. MP Alternate Traffic Handling

A potential Merge Point must determine when and if to accept alternate traffic. There are two critical components to this decision. First, the MP must know the state of all links to its UMH. This allows the MP to determine whether the multicast stream could be received from the UMH. Second, the MP must be able to distinguish between a normal multicast tree packet and an alternate packet.

The logic is similar for PIM and mLDP, but in PIM there is only one RPF-interface or interface of interest to the UMH. In mLDP, all the directly connected interfaces to the UMH are of interest. When the MP detects a local failure, if that interface was the last connected to the UMH and used for the multicast group, then the MP must rapidly switch from accepting the normal multicast tree traffic to accepting the alternate traffic. This rapid change must happen within the same approximately 50 milliseconds that the PLR switching to send traffic

on the alternate takes and for the same reasons. It does no good for the PLR to send alternate traffic if the MP doesn't accept it when it is needed.

The MP can identify alternate traffic based upon the MPLS label. This will be the alternate label that the MP supplied to its UMH for this purpose.

5.1.4. Merge Point Reconvergence

After a failure, the MP will want to join the multicast tree according to the new topology. It is critical that the MP does this in a way that minimizes the traffic disruption. Whenever paths change, there is also the possibility for a traffic-affecting event due to different latencies. However, traffic impact above that should be avoided.

The MP must do make-before-break. Until the MP knows that its new UMH is fully connected to the MCI, the MP should continue to accept its old alternate traffic. The MP could learn that the new UMH is sufficient either via control-plane mechanisms or data-driven. In the latter case, the reception of traffic from the new UMH can trigger the change-over. If the data-driven approach is used, a time-out to force the switch should apply to handle multicast trees that have long quiet periods.

5.1.5. PLR termination of alternate traffic

The PLR sends traffic on the alternates for a configurable time-out. There is no clean way for the next-hop routers and/or next-next-hop routers to indicate that the traffic is no longer needed.

If better control were desired, each MP could tell its UMH what the desired time-out is. The UMH could forward this to the PLR as well. Then the PLR could send alternate traffic to different MPs based upon the MP's individual timer. This would only be an advantage if some of the MPs were expected to have a longer multicast reconvergence time than others - either due to load or router capabilities.

5.2. MP-driven Unicast Tunnels

MP-driven unicast tunnels are only relevant for mLDP where targeted LDP sessions are feasible. For PIM, there is no mechanism to communicate beyond a router's immediate neighbors; these techniques could work for link-protection, but even then there would not be a way of requesting that the PLR should stop sending traffic.

There are three differences for MP-driven unicast tunnels from PLR-

driven unicast tunnels.

1. The MPs learn the identity of the PLR from their UMH. The PLR does not learn the identities of the MPs.

2. The MPs create direct connections to the PLR and communicate their alternate labels.

3. When the MPs have converged, each explicitly tells the PLR to stop sending alternate traffic.

The first means that a router communicates its UMH to all its downstream multicast hops. Then each MP communicates to the PLR(s) (1 for link-protection and 1 for node-protection) and indicates the multicast tree that protection is desired for and the associated alternate label.

When the PLR learns about a new MP, it adds that MP and associated information to the set of MPs to be protected. On a failure, the PLR does the same behavior as for the PLR-driven unicast tunnels.

After the failure, the MP reconverges using make-before-break. Then the MP explicitly communicates to the PLR(s) that alternate traffic is no longer needed for that multicast tree. When the node-protecting PLR hasn't changed for a MP, it may be necessary to withdraw the old alternate label, which tells the PLR to stop transmitting alternate traffic, and then provide a new alternate label.

5.3. MP-driven Alternate Trees

In this document we have defined different solutions to achieve fast convergence for multicast link and node protection based on MRTs. At a high level these solutions can be separated in Local and Global protections. Alternate Trees, which is a Local protection schema, initially looked like an attractive solution for Multicast node protection since it avoids replicating the packet by the PLR to each of the receivers of the protected node and waisting bandwidth. However, this comes at the expense of extra multicast state and complexity. In order to mitigate the extra multicast state its possible to aggregate the Alternate Trees by creating an Alternate Tree per protected node and reuse it for all the multicast trees going through this node. This further complicates the procedures and upstream assigned labels are required to de-aggregate the trees. With aggregation we are also introducing an unwanted side effect. The receiver population of the aggregated trees will very likely not be the same. That means multicast packets will be forwarded on the Alternate Tree to node(s) that may not have a receiver(s) for the

protected tree. The more protected trees are aggregated, the higher the risk of forwarding unwanted multicast packets, this leads again to waisting bandwidth.

Considering the complexity of this solution and the unwanted side-effects, the authors of this document believe its better to solve Multicast node protection using a Global protection schema, as documented in Section 4. The solution previously defined in this section has been move to Appendix A (Section 9).

6. Acknowledgements

The authors would like to thank Kishore Tiruveedhula, Santosh Esale, and Maciek Konstantynowicz for their suggestions and review.

7. IANA Considerations

This doument includes no request to IANA.

8. Security Considerations

This architecture is not currently believed to introduce new security concerns.

9. Appendix A

9.1. MP-driven Alternate Trees

For some networks, it is highly desirable not to have the PLR perform replication to each MP. PLR replication can cause substantial congestion on links used by alternates to different MPs. At the same time, it is also desirable to have minimal extra state created in the network. This can be resolved by creating alternate-trees that can protect multiple multicast groups as a bypass-alternate-tree. An alternate-tree can also be created per multicast group, PLR and failure point.

It is not possible to merge alternate-trees for different PLRs or for different neighbors. This is shown in Figure 4 where G can't select an acceptable upstream node on the alternate tree that doesn't violate either the need to avoid C (for PLR A) or D (for PLR B).

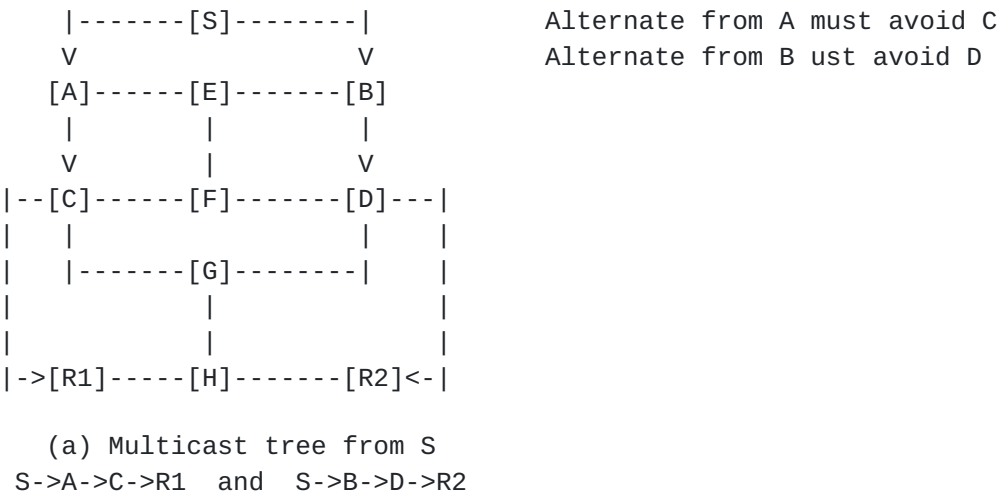


Figure 4: Alternate Trees from PLR A and B can't be merged

A MP that joins an alternate-tree for a particular multicast stream should not expect or request PLR-replicated tunneled alternate traffic for that same multicast stream.

Each alternate-tree is identified by the PLR which sources the traffic and the failure point (node and link) (FP) to be avoided. Different multicast groups with the same PLR and FP may have different sets of MPs - but they are all at most going to include the FP (for link protection) and the neighbors of FP except for the PLR. For a bypass-alternate-tree to work, it must be acceptable to temporarily send a multicast group's traffic to FP's neighbors that do not need it. This is the trade-off required to reduce alternate-tree state and use bypass-alternate-trees. As discussed in Section 5.1.3, a potential MP can determine whether to accept alternate traffic based upon the state of its normal upstream links. Alternate traffic for a group the MP hasn't joined can just be discarded.

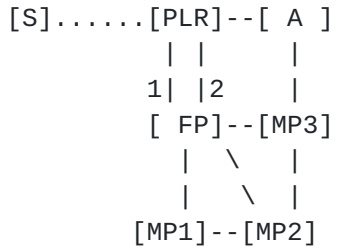


Figure 5: Alternate Tree Scenario

For any router, knowing the PLR and the FP to avoid will force selection of either the Blue MRT or the Red MRT. It is possible that the FP doesn't actually appear in either MRT path, but the FP will always be in either the set of nodes that might be used for the Blue MRT path or the set of nodes that might be used for the Red MRT path. The FP's membership in one of the sets is a function of the partial ordering and topological ordering created by the MRT algorithm and is consistent between routers in the network graph.

To create an alternate-tree, the following must happen:

1. For node-protection, the MP learns from its upstream (the FP) the node-id of its upstream (the PLR) and, optionally, a link identifier for the link used to the PLR. The link-id is only needed for traffic handling in PIM, since mLDP can have targeted sessions between the MP and the PLR.

2. For link-protection, the MP needs to know the node-id of its upstream (the PLR) and, optionally, its identifier for the link used to the PLR.

3. The MP determines whether to use the Blue or Red forwarding topology to reach the PLR while avoiding the FP and associated interface. This gives the MP its alternate-tree upstream interface.

4. The MP signals a backup-join to its alternate-tree upstream interface. The backup-join specifies the PLR, FP and, for PIM, the FP-PLR link identifier. If the alternate-tree is not to be used as a bypass-alternate-tree, then the multicast group (e.g. (S,G) or Opaque-Value) must be specified.

5. A router that receives a backup-join and is not the PLR needs to create multicast state and send a backup-join towards the PLR on the appropriate Blue or Red forwarding topology as is locally determined to avoid the FP and FP-PLR link.

6. Backup-joins for the same (PLR, FP, PLR-FP link-id) that reference the same multicast group can be merged into a single alternate-tree. Similarly, backup-joins for the same (PLR, FP, PLR-FP link-id) that reference no multicast group can be merged into a single alternate-tree.

7. When the PLR receives the backup-join, it associates either the specified multicast group with that alternate-tree, if such is given, or associates all multicast groups that go to the FP via the specified FP-PLR link with the alternate-tree.

For an example, look at Figure 5. FP would send a backup-join to MP3 indicating (PLR, FP, PLR-FP link-1). MP3 sends a backup-join to A. MP1 sends a backup-join to MP2 and MP2 sends a backup-join to MP3.

It is necessary that traffic on each alternate-tree self-identify as to which alternate-tree it is part of. This is because an alternate-tree for a multicast-group and a particular (PLR, FP, PLR-FP link-id) can easily overlap with an alternate-tree for the same multicast group and a different (PLR, FP, PLR-FP link-id). The best way of doing this depends upon whether PIM or mLDP is being used.

9.1.1. PIM details for Alternate-Trees

For PIM, the (S,G) of the IP packet is a globally unique identifier and is understood. To identify the alternate-tree, the most straightforward way is to use MPLS labels distributed in the PIM backup-join messages. A MP can use the incoming label to indicate the set of RPF-interfaces for which the traffic may be an alternate. If the alternate-tree isn't a bypass-alternate-tree, then only one RPF interface is referenced. If the alternate-tree is a bypass-alternate-tree, then multiple RPF-interfaces (parallel links to FP) might be intended. Alternate-tree traffic may cross an interface multiple times - either because the interface is a broadcast interface and different downstream-assigned labels are provided and/or because a MP may provide different labels.

9.1.2. mLDP details for Alternate-Trees

For mLDP, if bypass-alternate-trees are used, then the PLR must provide upstream-assigned labels to each multicast stream. The MP provides the label for the alternate-tree; if the alternate-tree is not a bypass-alternate-tree, this label also describes the multicast stream. If the alternate-tree is a bypass-alternate-tree, then it provides the context for the PLR-assigned labels for each multicast stream. If there are targeted LDP sessions between the PLR and the MPs, then the PLR could provide the necessary upstream-assigned labels.

9.1.3. Traffic Handling by PLR

An issue with traffic is how long should the PLR continue to send alternate traffic out. With an alternate-tree, the PLR can know to stop forwarding alternate traffic on the alternate-tree when that alternate-tree's state is torn down. This provides a clear signal that alternate traffic is no longer needed.

9.2. Methods Compared for PIM

The two approaches that are feasible for PIM are PLR-driven Unicast Tunnels and MP-driven Alternate-Trees.

Aspect	PLR-driven Unicast Tunnels	MP-driven Alternate-Trees
Worst-case Traffic Replication Per Link	1 + number of MPs	2
PLR alternate-traffic	timer-based	control-plane terminated
Extra multicast state	none	per (PLR,FP,S) for bypass mode

Which approach is prefered may be network-dependent. It should also be possible to use both in the same network.

9.3. Methods Compared for mLDP

All three approaches are feasible for mLDP. Below is a brief comparison of various aspects of each.

Aspect	MP-driven Unicast Tunnels	PLR-driven Unicast Tunnels	MP-driven Alternate-Trees
Worst-case Traffic Replication Per Link	1 + number of MPs	1 + number of MPs	2
PLR alternate-traffic	control-plane terminated	timer-based	control-plane terminated
Extra multicast state	none	none	per (PLR,FP,S) for bypass mode

10. References

10.1. Normative References

[I-D.enyedi-rtgwg-mrt-frr-algorithm]
Atlas, A., Envedi, G., Csaszar, A., and A. Gopalan,
"Algorithms for computing Maximally Redundant Trees for

IP/LDP Fast- Reroute", draft-enyedi-rtgwg-mrt-frr-algorithm-03 (work in progress), July 2013.

[I-D.ietf-rtgwg-mrt-frr-architecture] Atlas, A., Kebler, R., Envedi, G., Csaszar, A., Tantsura, J., Konstantynowicz, M., White, R., and M. Shand, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-03 (work in progress), July 2013.

[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.

[RFC6420] Cai, Y. and H. Ou, "PIM Multi-Topology ID (MT-ID) Join Attribute", RFC 6420, November 2011.

10.2. Informative References

[I-D.ietf-rtgwg-mofrr] Karan, A., Filsfils, C., Farinacci, D., Wijnands, I., Decraene, B., Joorde, U., and W. Henderickx, "Multicast only Fast Re-Route", draft-ietf-rtgwg-mofrr-02 (work in progress), June 2013.

[I-D.iwijnand-mpls-mldp-multi-topology] Wijnands, I. and K. Raza, "mLDP Extensions for Multi Topology Routing", draft-iwijnand-mpls-mldp-multi-topology-03 (work in progress), June 2013.

[I-D.kebler-pim-mrt-protection] Kebler, R., Atlas, A., Wijnands, IJ., and G. Enyedi, "PIM Extensions for Protection Using Maximally Redundant Trees", draft-kebler-pim-mrt-protection-00 (work in progress), March 2012.

[I-D.wijnands-mpls-mldp-node-protection] Wijnands, I., Rosen, E., Raza, K., Tantsura, J., Atlas, A., and Q. Zhao, "mLDP Node Protection", draft-wijnands-mpls-mldp-node-protection-04 (work in progress), June 2013.

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

[RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.

Authors' Addresses

Alia Atlas (editor)
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: akatlas@juniper.net

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

Email: rkebler@juniper.net

IJsbrand Wijnands
Cisco Systems, Inc.

Email: ice@cisco.com

Andras Csaszar
Ericsson
Konyves Kalman krt 11
Budapest 1097
Hungary

Email: Andras.Csaszar@ericsson.com

Gabor Sandor Enyedi
Ericsson
Konyves Kalman krt 11.
Budapest 1097
Hungary

Email: Gabor.Sandor.Enyedi@ericsson.com