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

A. Atlas, Ed.
R. Kebler
Juniper Networks
G. Enyedi
A. Csaszar
J. Tantsura
Ericsson
M. Konstantynowicz
Cisco Systems
R. White
VCE

July 12, 2013

An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees

draft-ietf-rtgwg-mrt-frr-architecture-03

Abstract

With increasing deployment of Loop-Free Alternates (LFA) [RFC5286], it is clear that a complete solution for IP and LDP Fast-Reroute is required. This specification provides that solution. IP/LDP Fast-Reroute with Maximally Redundant Trees (MRT-FRR) is a technology that gives link-protection and node-protection with 100% coverage in any network topology that is still connected after the failure.

MRT removes all need to engineer for coverage. MRT is also extremely computationally efficient. For any router in the network, the MRT computation is less than the LFA computation for a node with three or more neighbors.

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.

Table of Contents

1. Introduction

This document gives a complete solution for IP/LDP fast-reroute [RFC5714]. MRT-FRR creates two alternate trees separate from the primary next-hop forwarding used during stable operation. These two trees are maximally diverse from each other, providing link and node protection for 100% of paths and failures as long as the failure does not cut the network into multiple pieces. This document defines the architecture for IP/LDP fast-reroute with MRT. The associated protocol extensions are defined in [I-D.atlas-ospf-mrt] and [I-D.atlas-mpls-ldp-mrt]. The exact MRT algorithm is defined in [I-D.enyedi-rtgwg-mrt-frr-algorithm].

IP/LDP Fast-Reroute with MRT (MRT-FRR) uses two maximally diverse forwarding topologies to provide alternates. A primary next-hop should be on only one of the diverse forwarding topologies; thus, the other can be used to provide an alternate. Once traffic has been moved to one of MRTs, it is not subject to further repair actions. Thus, the traffic will not loop even if a worse failure (e.g. node) occurs when protection was only available for a simpler failure (e.g. link).

In addition to supporting IP and LDP unicast fast-reroute, the diverse forwarding topologies and guarantee of 100% coverage permit fast-reroute technology to be applied to multicast traffic as described in [I-D.atlas-rtgwg-mrt-mc-arch].

Other existing or proposed solutions are partial solutions or have significant issues, as described below.

Summary Comparison of IP/LDP FRR Methods

Method	Coverage	Alternate Looping?	Computation (in SPFs)
MRT-FRR	100% Link/Node	None	less than 3
LFA	Partial Link/Node	Possible	per neighbor
Remote LFA	Partial Link/Node	Possible	per neighbor (link) or neighbor's neighbor (node)
Not-Via	100% Link/Node	None	per link and node

Table 1

Loop-Free Alternates (LFA): LFAs [RFC5286] provide limited topology-dependent coverage for link and node protection. Restrictions on choice of alternates can be relaxed to improve coverage, but this can cause forwarding loops if a worse failure is experienced than protected against. Augmenting a network to provide better coverage is NP-hard [LFARevisited]. [RFC6571] discusses the applicability of LFA to different topologies with a focus on common PoP architectures.

Remote LFA: Remote LFAs [I-D.ietf-rtgwg-remote-lfa] improve coverage over LFAs for link protection but still cannot guarantee complete coverage. The trade-off of looping traffic to improve coverage is still made. Remote LFAs can provide node-protection [I-D.litkowski-rtgwg-node-protect-remote-lfa] but not guaranteed coverage and the computation required is quite high (an SPF per neighbor's neighbor). [I-D.bryant-ipfrr-tunnels] describes additional mechanisms to further improve coverage, at the cost of added complexity.

Not-Via: Not-Via [I-D.ietf-rtgwg-ipfrr-notvia-addresses] is the only other solution that provides 100% coverage for link and node failures and does not have potential looping. However, the computation is very high (an SPF per failure point) and academic implementations [LightweightNotVia] have found the address management complexity to be high.

1.1. Importance of 100% Coverage

Fast-reroute is based upon the single failure assumption - that the time between single failures is long enough for a network to reconverge and start forwarding on the new shortest paths. That does not imply that the network will only experience one failure or change.

It is straightforward to analyze a particular network topology for coverage. However, a real network does not always have the same topology. For instance, maintenance events will take links or nodes out of use. Simply costing out a link can have a significant effect on what LFAs are available. Similarly, after a single failure has happened, the topology is changed and its associated coverage. Finally, many networks have new routers or links added and removed; each of those changes can have an effect on the coverage for topology-sensitive methods such as LFA and Remote LFA. If fast-reroute is important for the network services provided, then a method that guarantees 100% coverage is important to accomodate natural network topology changes.

Asymmetric link costs are also a common aspect of networks. There are at least three common causes for them. First, any broadcast interface is represented by a pseudo-node and has asymmetric link costs to and from that pseudo-node. Second, when routers come up or a link with LDP comes up, it is recommended in [RFC5443] and [RFC3137] that the link metric be raised to the maximum cost; this may not be symmetric and for [RFC3137] is not expected to be. Third, techniques such as IGP metric tuning for traffic-engineering can result in asymmetric link costs. A fast-reroute solution needs to handle network topologies with asymmetric link costs.

When a network needs to use a micro-loop prevention mechanism [RFC5715] such as Ordered FIB[I-D.ietf-rtgwg-ordered-fib] or Farside Tunneling[RFC5715], then the whole IGP area needs to have alternates available so that the micro-loop prevention mechanism, which requires slower network convergence, can take the necessary time without impacting traffic badly. Without complete coverage, traffic to the unprotected destinations will be dropped for significantly longer than with current convergence - where routers individually converge as fast as possible.

1.2. Partial Deployment and Backwards Compatibility

MRT-FRR supports partial deployment. As with many new features, the protocols (OSPF, LDP, ISIS) indicate their capability to support MRT. Inside the MRT-capable connected group of routers (referred to as an MRT Island), the MRTs are computed. Alternates to destinations outside the MRT Island are computed and depend upon the existence of a loop-free neighbor of the MRT Island for that destination.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119]

3. Terminology

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.

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.

MRT-Red: MRT-Red is used to describe one of the two MRTs; it is used to described the associated forwarding topology and MT-ID. Specifically, MRT-Red is the decreasing MRT where links in the GADAG are taken in the direction from a higher topologically ordered node to a lower one.

MRT-Blue: MRT-Blue is used to describe one of the two MRTs; it is used to described the associated forwarding topology and MT-ID. Specifically, MRT-Blue is the increasing MRT where links in the GADAG are taken in the direction from a lower topologically ordered node to a higher one.

Rainbow MRT: It is useful to have an MT-ID that refers to the multiple MRT topologies and to the default topology. This is referred to as the Rainbow MRT MT-ID and is used by LDP to reduce signaling and permit the same label to always be advertised to all peers for the same (MT-ID, Prefix).

MRT Island: From the computing router, the set of routers that support a particular MRT profile and are connected.

Island Border Router (IBR): A router in the MRT Island that is connected to a router not in the MRT Island and both routers are in a common area or level.

Island Neighbor (IN): A router that is not in the MRT Island but is adjacent to an IBR and in the same area/level as the IBR.

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 graph.

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

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.

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

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

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

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

named proxy-node: A proxy-node can represent a destination prefix that can be attached to the MRT Island via at least two routers. It is named if there is a way that traffic can be encapsulated to reach specifically that proxy node; this could be because there is an LDP FEC for the associated prefix or because MRT-Red and MRT-Blue IP addresses are advertised in an undefined fashion for that proxy-node.

4. Maximally Redundant Trees (MRT)

A pair of Maximally Redundant Trees are directed spanning trees that provide maximally disjoint paths towards their common root. Only links or nodes whose failure would partition the network (i.e. cut-links and cut-vertices) are shared between the trees. The algorithm to compute MRTs is given in [I-D.enyedi-rtgwg-mrt-frr-algorithm]. This algorithm can be computed in O(e + n log n); it is less than

three SPFs. Modeling results comparing MRT alternates to the optimal are described in [I-D.enyedi-rtgwg-mrt-frr-algorithm]. This document describes how the MRTs can be used and not how to compute them.

MRT provides destination-based trees for each destination. Each router stores its normal primary next-hop(s) as well as MRT-Blue next-hop(s) and MRT-Red next-hop(s) toward each destination. The alternate will be selected between the MRT-Blue and MRT-Red.

The most important thing to understand about MRTs is that for each pair of destination-routed MRTs, there is a path from every node X to the destination D on the Blue MRT that is as disjoint as possible from the path on the Red MRT.

For example, in Figure 1, there is a network graph that is 2-connected in (a) and associated MRTs in (b) and (c). One can consider the paths from B to R; on the Blue MRT, the paths are B->F->D->E->R or B->C->D->E->R. On the Red MRT, the path is B->A->R. These are clearly link and node-disjoint. These MRTs are redundant trees because the paths are disjoint.

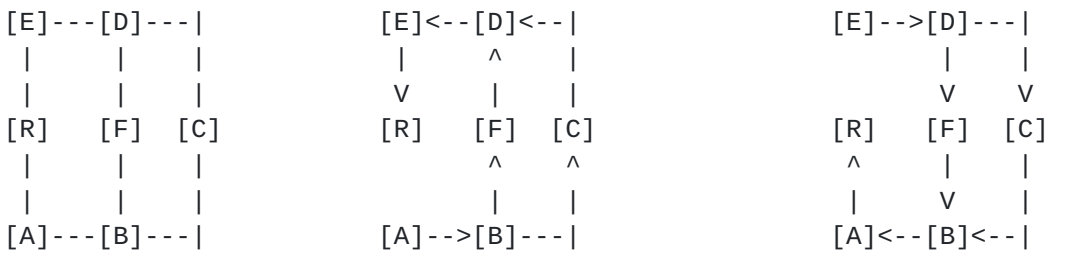
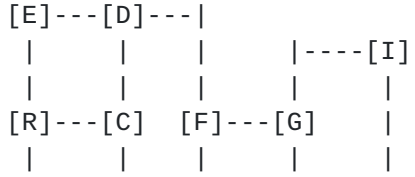


Figure 1: A 2-connected Network

By contrast, in Figure 2, the network in (a) is not 2-connected. If F, G or the link F<->G failed, then the network would be partitioned. It is clearly impossible to have two link-disjoint or node-disjoint paths from G, I or J to R. The MRTs given in (b) and (c) offer paths that are as disjoint as possible. For instance, the paths from B to R are the same as in Figure 1 and the path from G to R on the Blue MRT is G->F->D->E->R and on the Red MRT is G->F->B->A->R.



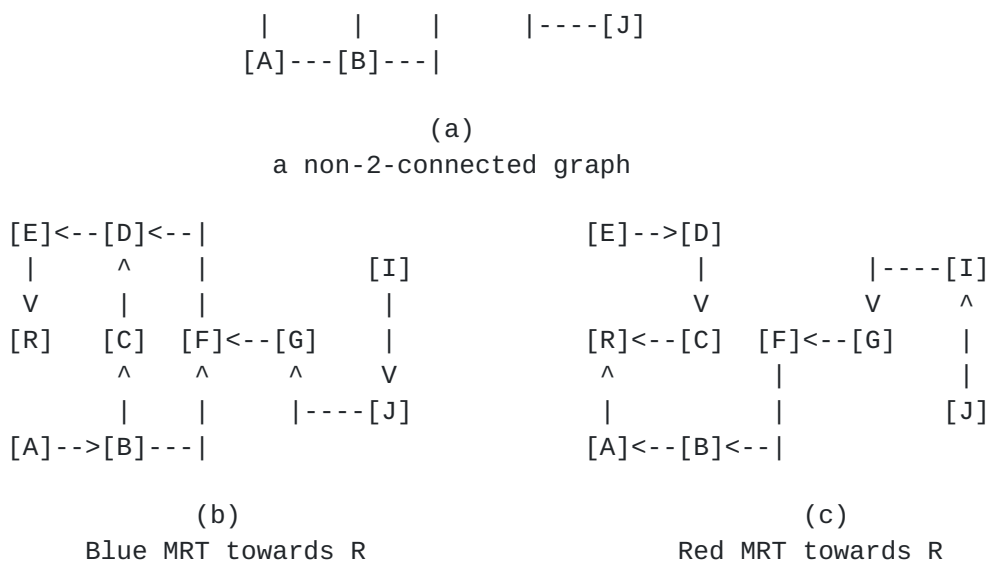


Figure 2: A non-2-connected network

5. Maximally Redundant Trees (MRT) and Fast-Reroute

In normal IGP routing, each router has its shortest-path-tree to all destinations. From the perspective of a particular destination, D, this looks like a reverse SPT (rSPT). To use maximally redundant trees, in addition, each destination D has two MRTs associated with it; by convention these will be called the MRT-Blue and MRT-Red. MRT-FRR is realized by using multi-topology forwarding. There is a MRT-Blue forwarding topology and a MRT-Red forwarding topology.

Any IP/LDP fast-reroute technique beyond LFA requires an additional dataplane procedure, such as an additional forwarding mechanism. The well-known options are multi-topology forwarding (used by MRT-FRR), tunneling (e.g. [I-D.ietf-rtgwg-ipfrr-notvia-addresses] or [I-D.ietf-rtgwg-remote-lfa]), and per-interface forwarding (e.g. Loop-Free Failure Insensitive Routing in [EnyediThesis]).

When there is a link or node failure affecting, but not partitioning, the network, each node will still have at least one path via one of the MRTs to reach the destination D. For example, in Figure 2, C would normally forward traffic to R across the C<->R link. If that C<->R link fails, then C could use the Blue MRT path C->D->E->R.

As is always the case with fast-reroute technologies, forwarding does not change until a local failure is detected. Packets are forwarded along the shortest path. The appropriate alternate to use is pre-computed. [I-D.enyedi-rtgwg-mrt-frr-algorithm] describes exactly how to determine whether the MRT-Blue next-hops or the MRT-Red next-hops

should be the MRT alternate next-hops for a particular primary next-hop N to a particular destination D.

MRT alternates are always available to use. It is a local decision whether to use an MRT alternate, a Loop-Free Alternate or some other type of alternate.

As described in [RFC5286], when a worse failure than is anticipated happens, using LFAs that are not downstream neighbors can cause micro-looping. Section 1.1 of [RFC5286] gives an example of link-protecting alternates causing a loop on node failure. Even if a worse failure than anticipated happens, the use of MRT alternates will not cause looping. Therefore, while node-protecting LFAs may be preferred, the certainty that no alternate-induced looping will occur is an advantage of using MRT alternates when the available node-protecting LFA is not a downstream path.

6. Unicast Forwarding with MRT Fast-Reroute

With LFA, there is no need to tunnel unicast traffic, whether IP or LDP. The traffic is simply sent to an alternate. As mentioned earlier in Section 5, MRT needs multi-topology forwarding. Unfortunately, neither IP nor LDP provides extra bits for a packet to indicate its topology.

Once the MRTs are computed, the two sets of MRTs are seen by the forwarding plane as essentially two additional topologies. The same considerations apply for forwarding along the MRTs as for handling multiple topologies.

6.1. LDP Unicast Forwarding - Avoid Tunneling

For LDP, it is very desirable to avoid tunneling because, for at least node protection, tunneling requires knowledge of remote LDP label mappings and thus requires targeted LDP sessions and the associated management complexity. There are two different mechanisms that can be used; Option A MUST be supported.

1. Option A - Encode MT-ID in Labels: In addition to sending a single label for a FEC, a router would provide two additional labels with the MT-IDs associated with the Blue MRT or Red MRT forwarding topologies. This is very simple for hardware support. It does reduce the label space for other uses. It also increases the memory to store the labels and the communication required by LDP.

2. Option B - Create Topology-Identification Labels: Use the label-stacking ability of MPLS and specify only two additional labels -

one for each associated MRT color - by a new FEC type. When sending a packet onto an MRT, first swap the LDP label and then push the topology-identification label for that MRT color. When receiving a packet with a topology-identification label, pop it and use it to guide the next-hop selection in combination with the next label in the stack; then swap the remaining label, if appropriate, and push the topology-identification label for the next-hop. This has minimal usage of additional labels, memory and LDP communication. It does increase the size of packets and the complexity of the required label operations and look-ups. This can use the same mechanisms as are needed for context-aware label spaces.

Note that with LDP unicast forwarding, regardless of whether topology-identification label or encoding topology in label is used, no additional loopbacks per router are required. This is because LDP labels are used on a hop-by-hop basis to identify MRT-blue and MRT-red forwading topologies.

For greatest hardware compatibility, routers implementing MRT LDP fast-reroute MUST support Option A of encoding the MT-ID in the labels. The extensions to indicate an MT-ID for a FEC are described in Section 3.2.1 of [I-D.ietf-mpls-ldp-multi-topology].

6.2. IP Unicast Traffic

For IP, there is no currently practical alternative except tunneling to gain the bits needed to indicate the MRT-Blue or MRT-Red forwarding topology. The choice of tunnel egress MAY be flexible since any router closer to the destination than the next-hop can work. This architecture assumes that the original destination in the area is selected (see Section 10 for handling of multi-homed prefixes); another possible choice is the next-next-hop towards the destination. For LDP traffic, using the original destination simplifies MRT-FRR by avoiding the need for targeted LDP sessions to the next-next-hop. For IP, that consideration doesn't apply but consistency with LDP is RECOMMENDED. If the tunnel egress is the original destination router, then the traffic remains on the redundant tree with sub-optimal routing. Selection of the tunnel egress is a router-local decision.

There are three options available for marking IP packets with which MRT it should be forwarded in. For greatest hardware compatibility and ease in removing the MRT-topology marking at area/level boundaries, routers that support MPLS and implement IP MRT fast-reroute MUST support Option A - using an LDP label that indicates the destination and MT-ID.

1. Tunnel IP packets via an LDP LSP. This has the advantage that more installed routers can do line-rate encapsulation and decapsulation. Also, no additional IP addresses would need to be allocated or signaled.

   a. Option A - LDP Destination-Topology Label: Use a label that indicates both destination and MRT. This method allows easy tunneling to the next-next-hop as well as to the IGP-area destination. For a proxy-node, the destination to use is the non-proxy-node immediately before the proxy-node on that particular color MRT.

   b. Option B - LDP Topology Label: Use a Topology-Identifier label on top of the IP packet. This is very simple. If tunneling to a next-next-hop is desired, then a two-deep label stack can be used with [Topology-ID label, Next-Next-Hop Label].

2. Tunnel IP packets in IP. Each router supporting this option would announce two additional loopback addresses and their associated MRT color. Those addresses are used as destination addresses for MRT-blue and MRT-red IP tunnels respectively. They allow the transit nodes to identify the traffic as being forwarded along either MRT-blue or MRT-red tree topology to reach the tunnel destination. Announcements of these two additional loopback addresses per router with their MRT color requires IGP extensions.

7. Protocol Extensions and Considerations: OSPF and ISIS

For simplicity, the approach of defining a well-known profile is taken in [I-D.atlas-ospf-mrt]. The purpose of communicating support for MRT in the IGP is to indicate thatqq the MRT-Blue and MRT-Red forwarding topologies are created for transit traffic. This section describes the various options to be selected. The default MRT profile is described here and the signaling extensions for OSPF are given in [I-D.atlas-ospf-mrt].

For any MRT profile, the MRT Island is created by starting from the computing router. If the computing router supports the default MRT profile, add it to the MRT Island. Add a router to the MRT Island if the router supports the default MRT profile and is connected to the MRT Island via bidirectional links eligible for MRT.

If a router advertises support for multiple MRT profiles, then it MUST create the transit forwarding topologies for each of those, unless the profile specifies No Forwarding Mechanism (e.g. as might be done for a profile used only for multicast global protection). A

router MUST NOT advertise multiple MRT profiles that overlap in their MRT-Red MT-ID or MRT-Blue MT-ID.

The MRT Profile also defines different behaviors such as how MRT recomputation is handled and how area/level boundaries are dealt with.

MRT Algorithm: MRT Lowpoint algorithm defined in [I-D.enyedi-rtgwg-mrt-frr-algorithm].

MRT-Red MT-ID: experimental 3997, final value assigned by IANA allocated from the LDP MT-ID space

MRT-Blue MT-ID: experimental 3998, final value assigned by IANA allocated from the LDP MT-ID space

GADAG Root Selection Priority: Among the routers in the MRT Island and with the highest priority advertised, an implementation MUST pick the router with the highest Router ID to be the GADAG root.

Forwarding Mechanisms: LDP

Recalculation: Recalculation of MRTs SHOULD occur as described in Section 11.2. This allows the MRT forwarding topologies to support IP/LDP fast-reroute traffic.

Area/Level Border Behavior: As described in Section 9, ABRs/LBRs SHOULD ensure that traffic leaving the area also exits the MRT-Red or MRT-Blue forwarding topology.

The following describes the aspects to be considered to define a profile to advertise. For some profiles, associated information may need to be distributed, such as GADAG Root Selection Priority, Red MRT Loopback Address, Blue MRT Loopback Address.

MRT Algorithm: This identifies the particular MRT algorithm used by the router for this profile. Algorithm transitions can be managed by advertising multiple MRT profiles.

MRT-Red MT-ID: This specifies the MT-ID to be associated with the MRT-Red forwarding topology. It is needed for use in LDP signaling. All routers in the MRT Island MUST agree on a value.

MRT-Blue MT-ID: This specifies the MT-ID to be associated with the MRT-Blue forwarding topology. It is needed for use in LDP signaling. All routers in the MRT Island MUST agree on a value.

GADAG Root Selection Priority: A MRT profile might specify this to provide the network operator with a knob to force a particular GADAG root selection. If not specified in the MRT profile, the highest Router ID in the profile's MRT Island will be elected the GADAG Root. If a GADAG Root Selection Priority is specified, then the MRT profile must also specify how the GADAG Root is elected.

Forwarding Mechanism: This specifies which forwarding mechanisms the router supports for transit traffic. An MRT island must program appropriate next-hops into the forwarding plane. The known options are IPv4, IPv6, LDP, and None. If IPv4 is supported, then both MRT-Red and MRT-Blue IPv4 Loopback Addresses SHOULD be specified. If IPv6 is supported, both MRT-Red and MRT-Blue IPv6 Loopback Addresses SHOULD be specified. If LDP is supported, then LDP support and signaling extensions MUST be supported.

MRT-Red Loopback Address: This provides the router's loopback address to reach the router via the MRT-Red forwarding topology. It can, of course, be specified for both IPv4 and IPv6.

MRT-Blue Loopback Address: This provides the router's loopback address to reach the router via the MRT-Blue forwarding topology. It can, of course, be specified for both IPv4 and IPv6.

Recalculation: As part of what process and timing should the new MRTs be computed on a modified topology? Section 11.2 describes the minimum behavior required to support fast-reroute.

Area/Level Border Behavior: Should inter-area traffic on the MRT-Blue or MRT-Red be put back onto the shortest path tree? Should it be swapped from MRT-Blue or MRT-Red in one area/level to MRT-Red or MRT-Blue in the next area/level to avoid the potential failure of an ABR? (See [I-D.atlas-rtgwg-mrt-mc-arch] for use-case details.

Other Profile-Specific Behavior: Depending upon the use-case for the profile, there may be additional profile-specific behavior.

As with LFA, it is expected that OSPF Virtual Links will not be supported.

8. Protocol Extensions and considerations: LDP

The protocol extensions for LDP are defined in [I-D.atlas-mpls-ldp-mrt]. A router must indicate that it has the ability to support MRT; having this explicit allows the use of MRT-specific processing, such as special handling of FECs sent with the Rainbow MRT MT-ID.

A FEC sent with the Rainbow MRT MT-ID indicates that the FEC applies to all the MRT-Blue and MRT-Red MT-IDs in supported MRT profiles as well as to the default shortest-path based MT-ID 0. The Rainbow MRT MT-ID is defined to provide an easy way to handle the special signaling that is needed at ABRs or LBRs. It avoids the problem of needing to signal different MPLS labels for the same FEC. Because the Rainbow MRT MT-ID is used only by ABRs/LBRs or the LDP egress, it is not MRT profile specific. The proposed experimental value is 3999 and the final value will be assigned by IANA and allocated from the LDP MT-ID space. The authoritative values are given in [I-D.atlas-mpls-ldp-mrt].

9. Inter-Area and ABR Forwarding Behavior

An ABR/LBR has two forwarding roles. First, it forwards traffic inside its area. Second, it forwards traffic from one area into another. These same two roles apply for MRT transit traffic. Traffic on MRT-Red or MRT-Blue destined inside the area needs to stay on MRT-Red or MRT-Blue in that area. However, it is desirable for traffic leaving the area to also exit MRT-Red or MRT-Blue back to the shortest-path forwarding.

For unicast MRT-FRR, the need to stay on an MRT forwarding topology terminates at the ABR/LBR whose best route is via a different area/level. It is highly desirable to go back to the default forwarding topology when leaving an area/level. There are three basic reasons for this. First, the default topology uses shortest paths; the packet will thus take the shortest possible route to the destination. Second, this allows failures that might appear in multiple areas (e.g. ABR/LBR failures) to be separately identified and repaired around. Third, the packet can be fast-rerouted again, if necessary, due to a failure in a different area.

An ABR/LBR that receives a packet on MRT-Red or MRT-Blue towards a destination in another area/level should forward the packet in the area/level with the best route along MRT-Red or MRT-Blue. If the packet came from that area/level, this correctly avoids the failure. However, if the traffic came from a different area/level, the packet should be removed from MRT-Red or MRT-Blue and forwarded on the shortest-path default forwarding topology.

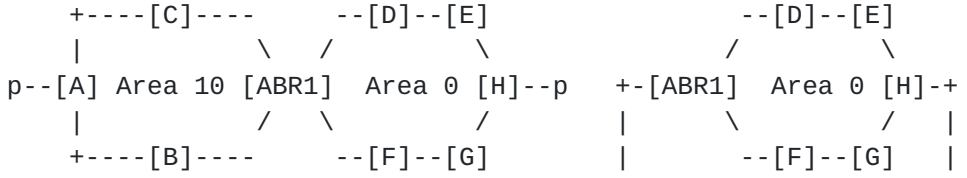
To avoid per-interface forwarding state for MRT-Red and MRT-Blue, the ABR/LBR needs to arrange that packets destined to a different area arrive at the ABR/LBR already not marked as MRT-Red or MRT-Blue.

For LDP forwarding where the MPLS label specifies (MT-ID, FEC), the ABR/LBR is responsible for advertising the proper label to each neighbor. Assume that an ABR/LBR has allocated three labels for a particular destination; those labels are L_primary, L_blue, and L_red. When the ABR/LBR advertises label bindings to routers in the area with the best route to the destination, the ABR/LBR provides L_primary for the default topology, L_blue for the MRT-Blue MT-ID and L_red for the MRT-Red MT-ID, exactly as expected. However, when the ABR/LBR advertises label bindings to routers in other areas, the ABR/LBR advertises L_primary for the Rainbow MRT MT-ID, which is then used for the default topology, for the MRT-Blue MT-ID and for the MRT-Red MT-ID.

The ABR/LBR installs all next-hops from the best area: primary next-hops for L_primary, MRT-Blue next-hops for L_blue, and MRT-Red next-hops for L_red. Because the ABR/LBR advertised (Rainbow MRT MT-ID, FEC) with L_primary to neighbors not in the best area, packets from those neighbors will arrive at the ABR/LBR with a label L_primary and will be forwarded into the best area along the default topology. By controlling what labels are advertised, the ABR/LBR can thus enforce that packets exiting the area do so on the shortest-path default topology.

If IP forwarding is used, then the ABR/LBR behavior is dependent upon the outermost IP address. If the outermost IP address is an MRT loopback address of the ABR/LBR, then the packet is decapsulated and forwarded based upon the inner IP address, which should go on the default SPT topology. If the outermost IP address is not an MRT loopback address of the ABR/LBR, then the packet is simply forwarded along the associated forwarding topology. A PLR sending traffic to a destination outside its local area/level will pick the MRT and use the associated MRT loopback address of the selected ABR/LBR connected to the external destination.

Thus, regardless of which of these two forwarding mechanisms are used, there is no need for additional computation or per-area forwarding state.



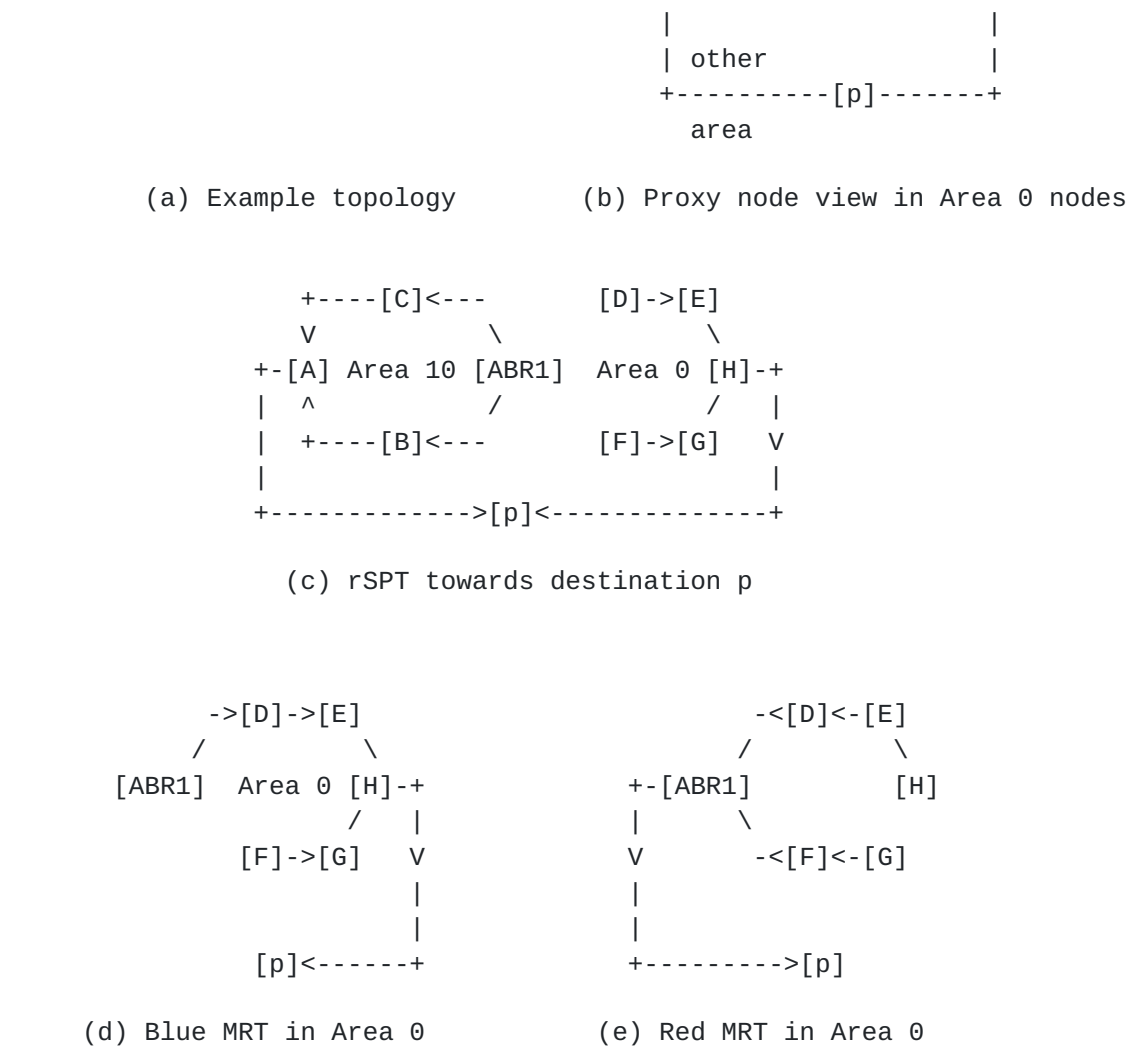


Figure 3: ABR Forwarding Behavior and MRTs

The other forwarding mechanism described in Section 6 is using Topology-Identification Labels. This mechanism would require that any router whose MRT-Red or MRT-Blue next-hop is an ABR/LBR would need to determine whether the ABR/LBR would forward the packet out of the area/level. If so, then that router should pop off the topology-identification label before forwarding the packet to the ABR/LBR.

For example, in Figure 3, if node H fails, node E has to put traffic towards prefix p onto MRT-Red. But since node D knows that ABR1 will use a best from another area, it is safe for D to pop the Topology-Identification Label and just forward the packet to ABR1 along the MRT-Red next-hop. ABR1 will use the shortest path in Area 10.

In all cases for ISIS and most cases for OSPF, the penultimate router can determine what decision the adjacent ABR will make. The one case where it can't be determined is when two ASBRs are in different non-backbone areas attached to the same ABR, then the ASBR's Area ID may be needed for tie-breaking (prefer the route with the largest OPSF area ID) and the Area ID isn't announced as part of the ASBR link-state advertisement (LSA). In this one case, suboptimal forwarding along the MRT in the other area would happen. If that becomes a realistic deployment scenario, OSPF extensions could be considered. This is not covered in [I-D.atlas-ospf-mrt].

10. Prefixes Multiply Attached to the MRT Island

How a computing router S determines its local MRT Island for each supported MRT profile is already discussed in Section 7.

There are two types of prefixes or FECs that may be multiply attached to an MRT Island. The first type are multi-homed prefixes that usually connect at a domain or protocol boundary. The second type represent routers that do not support the profile for the MRT Island. The key difference is whether the traffic, once out of the MRT Island, remains in the same area/level and might reenter the MRT Island if a loop-free exit point is not selected.

One property of LFAs that is necessary to preserve is the ability to protect multi-homed prefixes against ABR failure. For instance, if a prefix from the backbone is available via both ABR A and ABR B, if A fails, then the traffic should be redirected to B. This can also be done for backups via MRT.

If ASBR protection is desired, this has additonal complexities if the ASBRs are in different areas. Similarly, protecting labeled BGP traffic in the event of an ASBR failure has additional complexities due to the per-ASBR label spaces involved.

As discussed in [RFC5286], a multi-homed prefix could be:

- An out-of-area prefix announced by more than one ABR,
- An AS-External route announced by 2 or more ASBRs,
- A prefix with iBGP multipath to different ASBRs,
- etc.

There are also two different approaches to protection. The first is to do endpoint selection to pick a router to tunnel to where that router is loop-free with respect to the failure-point. Conceptually,

the set of candidate routers to provide LFAs expands to all routers, with an MRT alternate, attached to the prefix.

The second is to use a proxy-node, that can be named via MPLS label or IP address, and pick the appropriate label or IP address to reach it on either MRT-Blue or MRT-Red as appropriate to avoid the failure point. A proxy-node can represent a destination prefix that can be attached to the MRT Island via at least two routers. It is termed a named proxy-node if there is a way that traffic can be encapsulated to reach specifically that proxy-node; this could be because there is an LDP FEC for the associated prefix or because MRT-Red and MRT-Blue IP addresses are advertised in an as-yet undefined fashion for that proxy-node. Traffic to a named proxy-node may take a different path than traffic to the attaching router; traffic is also explicitly forwarded from the attaching router along a predetermined interface towards the relevant prefixes.

For IP traffic, multi-homed prefixes can use endpoint selection. For IP traffic that is destined to a router outside the MRT Island, if that router is the egress for a FEC advertised into the MRT Island, then the named proxy-node approach can be used.

For LDP traffic, there is always a FEC advertised into the MRT Island. The named proxy-node approach should be used, unless the computing router S knows the label for the FEC at the selected endpoint.

If a FEC is advertised from outside the MRT Island into the MRT Island and the forwarding mechanism specified in the profile includes LDP, then the routers learning that FEC MUST also advertise labels for (MRT-Red, FEC) and (MRT-Blue, FEC) to neighbors inside the MRT Island. If the forwarding mechanism includes LDP, any router receiving a FEC corresponding to a router outside the MRT Island or to a multi-homed prefix MUST compute and install the transit MRT-Blue and MRT-Red next-hops for that FEC; the associated FECs ((MT-ID 0, FEC), (MRT-Red, FEC), and (MRT-Blue, FEC)) MUST also be provided via LDP to neighbors inside the MRT Island.

10.1. Endpoint Selection

Endpoint Selection is a local matter for a router in the MRT Island since it pertains to selecting and using an alternate and does not affect the transit MRT-Red and MRT-Blue forwarding topologies.

Let the computing router be S and the next-hop F be the node whose failure is to be avoided. Let the destination be prefix p. Have A be the router to which the prefix p is attached for S's shortest path to p.

The candidates for endpoint selection are those to which the destination prefix is attached in the area/level. For a particular candidate B, it is necessary to determine if B is loop-free to reach p with respect to S and F for node-protection or at least with respect to S and the link (S, F) for link-protection. If B will always prefer to send traffic to p via a different area/level, then this is definitional. Otherwise, distance-based computations are necessary and an SPF from B's perspective may be necessary. The following equations give the checks needed; the rationale is similar to that given in [RFC5286].

Loop-Free for S: D_opt(B, p) < D_opt(B, S) + D_opt(S, p)

Loop-Free for F: D_opt(B, p) < D_opt(B, F) + D_opt(F, p)

The latter is equivalent to the following, which avoids the need to compute the shortest path from F to p.

Loop-Free for F: D_opt(B, p) < D_opt(B, F) + D_opt(S, p) - D_opt(S, F)

Finally, the rules for Endpoint selection are given below. The basic idea is to repair to the prefix-advertising router selected for the shortest-path and only to select and tunnel to a different endpoint if necessary (e.g. A=F or F is a cut-vertex or the link (S,F) is a cut-link).

1. Does S have a node-protecting alternate to A? If so, select that. Tunnel the packet to A along that alternate. For example, if LDP is the forwarding mechanism, then push the label (MRT-Red, A) or (MRT-Blue, A) onto the packet.

2. If not, then is there a router B that is loop-free to reach p while avoiding both F and S? If so, select B as the end-point. Determine the MRT alternate to reach B while avoiding F. Tunnel the packet to B along that alternate. For example, with LDP, push the label (MRT-Red, B) or (MRT-Blue, B) onto the packet.

3. If not, then does S have a link-protecting alternate to A? If so, select that.

4. If not, then is there a router B that is loop-free to reach p while avoiding S and the link from S to F? If so, select B as the endpoint and the MRT alternate that for reaching B from S avoiding the link (S,F).

The endpoint selected will receive a packet destined to itself and, being the egress, will pop that MPLS label (or have signaled Implicit

Null) and forward based on what is underneath. This suffices for IP traffic where the MPLS labels understood by the endpoint router are not needed.

10.2. Named Proxy-Nodes

A clear advantage to using a named proxy-node is that it is possible to explicitly forward from the MRT Island along an interface to a loop-free island neighbor (LFIN) when that interface may not be a primary next-hop. For LDP traffic where the label indicates both the topology and the FEC, it is necessary to either use a named proxy-node or deal with learning remote MPLS labels.

A named proxy-node represents one or more destinations and, for LDP forwarding, has a FEC associated with it that is signaled into the MRT Island. Therefore, it is possible to explicitly label packets to go to (MRT-Red, FEC) or (MRT-Blue, FEC); at the border of the MRT Island, the label will swap to meaning (MT-ID 0, FEC). It would be possible to have named proxy-nodes for IP forwarding, but this would require extensions to signal two IP addresses to be associated with MRT-Red and MRT-Blue for the proxy-node. A named proxy-node can be uniquely represented by the two routers in the MRT Island to which it is connected. The extensions to signal such IP addresses are not defined in [I-D.atlas-ospf-mrt]. The details of what label-bindings must be originated are described in [I-D.atlas-mpls-ldp-mrt].

Computing the MRT next-hops to a named proxy-node and the MRT alternate for the computing router S to avoid a particular failure node F is extremely straightforward. The details of the simple constant-time functions, Select_Proxy_Node_NHs() and Select_Alternates_Proxy_Node(), are given in [I-D.enyedi-rtgwg-mrt-frr-algorithm]. A key point is that computing these MRT next-hops and alternates can be done as new named proxy-nodes are added or removed without requiring a new MRT computation or impacting other existing MRT paths. This maps very well to, for example, how OSPFv2 [[RFC2328] Section 16.5] does incremental updates for new summary-LSAs.

The key question is how to attach the named proxy-node to the MRT Island; all the routers in the MRT Island MUST do this consistently. No more than 2 routers in the MRT Island can be selected; one should only be selected if there are no others that meet the necessary criteria. The named proxy-node is logically part of the area/level.

There are two sources for candidate routers in the MRT Island to connect to the named proxy-node. The first set are those routers that are advertising the prefix; the cost assigned to each such router is the announced cost to the prefix. The second set are those

routers in the MRT Island that are connected to routers not in the MRT Island but in the same area/level; such routers will be defined as Island Border Routers (IBRs). The routers connected to the IBRs that are not in the MRT Island and are in the same area/level are Island Neighbors (INs).

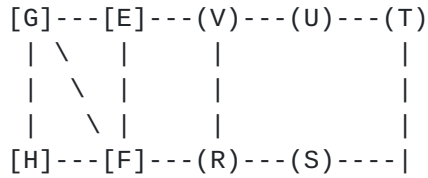
Since packets sent to the named proxy-node along MRT-Red or MRT-Blue may come from any router inside the MRT Island, it is necessary that whatever router to which an IBR forwards the packet be loop-free with regard to the whole MRT Island for the destination. Thus, an IBR is a candidate router only if it possesses at least one IN whose path to the prefix does not enter the MRT Island. The cost assigned to each (IBR, IN) pair is the D_opt(IN, prefix) plus Cost(IBR, IN).

From the set of prefix-advertising routers and the IBRs, the two lowest cost routers are selected and ties are broken based upon the lowest Router ID. For ease of discussion, such selected routers are proxy-node attachment routers and the two selected will be named A and B.

A proxy-node attachment router has a special forwarding role. When a packet is received destined to (MRT-Red, prefix) or (MRT-Blue, prefix), if the proxy-node attachment router is an IBR, it MUST swap to the default topology (e.g. swap to the label for (MT-ID 0, prefix) or remove the outer IP encapsulation) and forward the packet to the IN whose cost was used in the selection. If the proxy-node attachment router is not an IBR, then the packet MUST be removed from the MRT forwarding topology and sent along the interface that caused the router to advertise the prefix; this interface might be out of the area/level/AS.

10.2.1. Computing if an Island Neighbor (IN) is loop-free

As discussed, the Island Neighbor needs to be loop-free with regard to the whole MRT Island for the destination. Conceptually, the cost of transiting the MRT Island should be regarded as 0. This can be done by collapsing the MRT Island into a single node, as seen in Figure 4, and then computing SPFs from each Island Neighbor and from the MRT Island itself.



(1) Network Graph with Partial Deployment

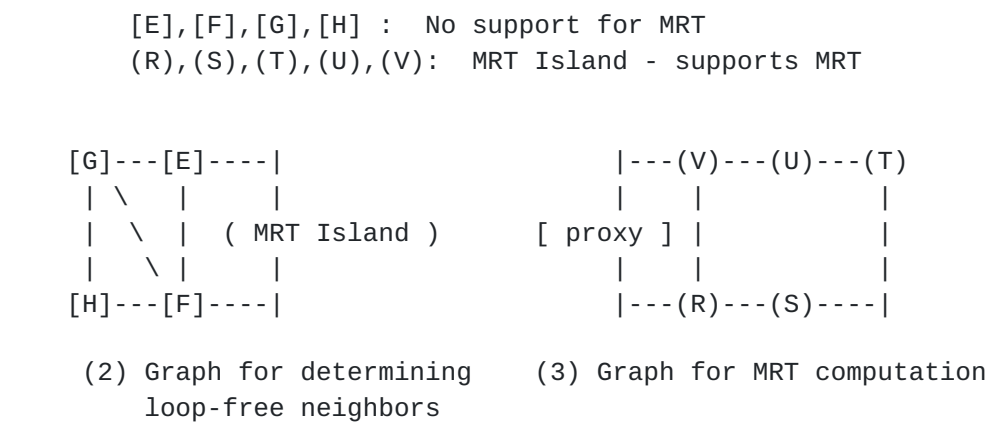


Figure 4: Computing alternates to destinations outside the MRT Island

The simple way to do this without manipulating the topology is to compute the SPFs from each IN and a node in the MRT Island (e.g. the GADAG root), but use a link metric of 0 for all links between routers in the MRT Island. The distances computed via SPF this way will be refered to as Dist_mrt0.

An IN is loop-free with respect to a destination D if: Dist_mrt0(IN, D) < Dist_mrt0(IN, MRT Island Router) + Dist_mrt0(MRT Island Router, D). Any router in the MRT Island can be used since the cost of transiting between MRT Island routers is 0. The GADAG Root is recommended for consistency.

10.3. MRT Alternates for Destinations Outside the MRT Island

A natural concern with new functionality is how to have it be useful when it is not deployed across an entire IGP area. In the case of MRT FRR, where it provides alternates when appropriate LFAs aren't available, there are also deployment scenarios where it may make sense to only enable some routers in an area with MRT FRR. A simple example of such a scenario would be a ring of 6 or more routers that is connected via two routers to the rest of the area.

Destinations inside the local island can obviously use MRT alternates. Destinations outside the local island can be treated like a multi-homed prefix and either Endpoint Selection or Named Proxy-Nodes can be used. Named Proxy-Nodes MUST be supported when LDP forwarding is supported and a label-binding for the destination is sent to an IBR.

Naturally, there are more complicated options to improve coverage, such as connecting multiple MRT islands across tunnels, but the need for the additional complexity has not been justified.

11. Network Convergence and Preparing for the Next Failure

After a failure, MRT detours ensure that packets reach their intended destination while the IGP has not reconverged onto the new topology. As link-state updates reach the routers, the IGP process calculates the new shortest paths. Two things need attention: micro-loop prevention and MRT re-calculation.

11.1. Micro-forwarding loop prevention and MRTs

As is well known[RFC5715], micro-loops can occur during IGP convergence; such loops can be local to the failure or remote from the failure. Managing micro-loops is an orthogonal issue to having alternates for local repair, such as MRT fast-reroute provides.

There are two possible micro-loop prevention mechanisms discussed in [RFC5715]. The first is Ordered FIB [I-D.ietf-rtgwg-ordered-fib]. The second is Farside Tunneling which requires tunnels or an alternate topology to reach routers on the farside of the failure.

Since MRTs provide an alternate topology through which traffic can be sent and which can be manipulated separately from the SPT, it is possible that MRTs could be used to support Farside Tunneling. Details of how to do so are outside the scope of this document.

Micro-loop mitigation mechanisms can also work when combined with MRT.

11.2. MRT Recalculation

When a failure event happens, traffic is put by the PLRs onto the MRT topologies. After that, each router recomputes its shortest path tree (SPT) and moves traffic over to that. Only after all the PLRs have switched to using their SPTs and traffic has drained from the MRT topologies should each router install the recomputed MRTs into the FIBs.

At each router, therefore, the sequence is as follows:

1. Receive failure notification

2. Recompute SPT

3. Install new SPT

4. If the network was stable before the failure occured, wait a configured (or advertised) period for all routers to be using their SPTs and traffic to drain from the MRTs.

5. Recompute MRTs

6. Install new MRTs.

While the recomputed MRTs are not installed in the FIB, protection coverage is lowered. Therefore, it is important to recalculate the MRTs and install them quickly.

12. Acknowledgements

The authors would like to thank Mike Shand for his valuable review and contributions.

The authors would like to thank Joel Halpern, Hannes Gredler, Ted Qian, Kishore Tiruveedhula, Shraddha Hegde, Santosh Esale, Nitin Bahadur, Harish Sitaraman, Raveendra Torvi and Chris Bowers for their suggestions and review.

13. IANA Considerations

This doument includes no request to IANA.

14. Security Considerations

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

15. References

15.1. Normative References

[I-D.enyedi-rtgwg-mrt-frr-algorithm]
Atlas, A., Envedi, G., Csaszar, A., Gopalan, A., and C. Bowers, "Algorithms for computing Maximally Redundant Trees for IP/LDP Fast- Reroute", draft-enyedi-rtgwg-mrt-frr-algorithm-03 (work in progress), July 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.

15.2. Informative References

[EnyediThesis]
Enyedi, G., "Novel Algorithms for IP Fast Reroute", Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics Ph.D. Thesis, February 2011, <http://timon.tmit.bme.hu/theses/thesis_book.pdf>.

[I-D.atlas-mpls-ldp-mrt]
Atlas, A., Tiruveedhula, K., Tantsura, J., and IJ. Wijnands, "LDP Extensions to Support Maximally Redundant Trees", draft-atlas-mpls-ldp-mrt-00 (work in progress), July 2013.

[I-D.atlas-ospf-mrt]
Atlas, A., Hegde, S., Chris, C., and J. Tantsura, "OSPF Extensions to Support Maximally Redundant Trees", draft-atlas-ospf-mrt-00 (work in progress), July 2013.

[I-D.atlas-rtgwg-mrt-mc-arch]
Atlas, A., Kebler, R., Wijnands, I., Csaszar, A., and G. Envedi, "An Architecture for Multicast Protection Using Maximally Redundant Trees", draft-atlas-rtgwg-mrt-mc-arch-02 (work in progress), July 2013.

[I-D.bryant-ipfrr-tunnels]
Bryant, S., Filsfils, C., Previdi, S., and M. Shand, "IP Fast Reroute using tunnels", draft-bryant-ipfrr-tunnels-03 (work in progress), November 2007.

[I-D.ietf-mpls-ldp-multi-topology]
Zhao, Q., Fang, L., Zhou, C., Li, L., and K. Raza, "LDP Extensions for Multi Topology Routing", draft-ietf-mpls-ldp-multi-topology-08 (work in progress), May 2013.

[I-D.ietf-rtgwg-ipfrr-notvia-addresses]
Bryant, S., Previdi, S., and M. Shand, "A Framework for IP and MPLS Fast Reroute Using Not-via Addresses", draft-ietf-rtgwg-ipfrr-notvia-addresses-11 (work in progress), May 2013.

[I-D.ietf-rtgwg-ordered-fib]
Shand, M., Bryant, S., Previdi, S., Filsfils, C., Francois, P., and O. Bonaventure, "Framework for Loop-free convergence using oFIB", draft-ietf-rtgwg-ordered-fib-12 (work in progress), May 2013.

[I-D.ietf-rtgwg-remote-lfa]
Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S. Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-02 (work in progress), May 2013.

[I-D.litkowski-rtgwg-node-protect-remote-lfa]
Litkowski, S., "Node protecting remote LFA", draft-litkowski-rtgwg-node-protect-remote-lfa-00 (work in progress), April 2013.

[LFARevisited]
Retvari, G., Tapolcai, J., Enyedi, G., and A. Csaszar, "IP Fast ReRoute: Loop Free Alternates Revisited", Proceedings of IEEE INFOCOM , 2011, <http://opti.tmit.bme.hu/~tapolcai/papers/retvari2011lfa_infocom.pdf>.

[LightweightNotVia]
Enyedi, G., Retvari, G., Szilagyi, P., and A. Csaszar, "IP Fast ReRoute: Lightweight Not-Via without Additional Addresses", Proceedings of IEEE INFOCOM , 2009, <http://mycite.omikk.bme.hu/doc/71691.pdf>.

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

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[RFC3137] Retana, A., Nguyen, L., White, R., Zinin, A., and D. McPherson, "OSPF Stub Router Advertisement", RFC 3137, June 2001.

[RFC5443] Jork, M., Atlas, A., and L. Fang, "LDP IGP Synchronization", RFC 5443, March 2009.

[RFC5715] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", RFC 5715, January 2010.

[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.

Appendix A. General Issues with Area Abstraction

When a multi-homed prefix is connected in two different areas, it may be impractical to protect them without adding the complexity of explicit tunneling. This is also a problem for LFA and Remote-LFA.

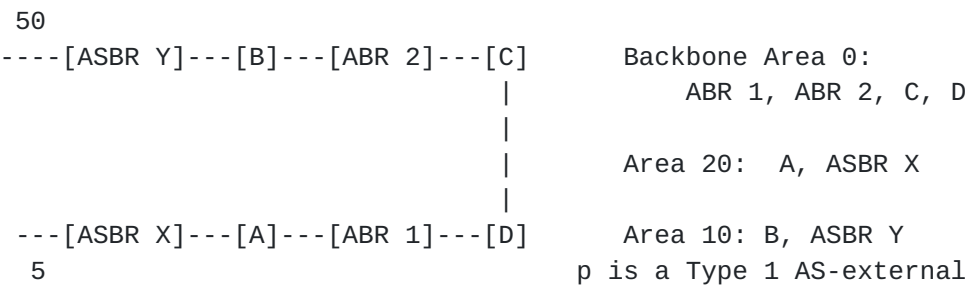


Figure 5: AS external prefixes in different areas

Consider the network in Figure 5 and assume there is a richer connective topology that isn't shown, where the same prefix is announced by ASBR X and ASBR Y which are in different non-backbone areas. If the link from A to ASBR X fails, then an MRT alternate could forward the packet to ABR 1 and ABR 1 could forward it to D, but then D would find the shortest route is back via ABR 1 to Area 20. This problem occurs because the routers, including the ABR, in one area are not yet aware of the failure in a different area.

The only way to get it from A to ASBR Y is to explicitly tunnel it to ASBR Y. If the traffic is unlabeled or the appropriate MPLS labels are known, then explicit tunneling MAY be used as long as the shortest-path of the tunnel avoids the failure point. In that case, A must determine that it should use an explicit tunnel instead of an MRT alternate.

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

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

Email: Gabor.Sandor.Enyedi@ericsson.com

Andras Csaszar
Ericsson
Konyves Kalman krt 11
Budapest 1097
Hungary

Email: Andras.Csaszar@ericsson.com

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
USA

Email: jeff.tantsura@ericsson.com

Maciek Konstantynowicz
Cisco Systems

Email: maciek@bgp.nu

Russ White
VCE

Email: russw@riw.us