Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2015

G. Enyedi, Ed.
A. Csaszar
Ericsson
A. Atlas, Ed.
C. Bowers
Juniper Networks
A. Gopalan
University of Arizona

July 4, 2014

Algorithms for computing Maximally Redundant Trees for IP/LDP Fast-Reroute
draft-ietf-rtgwg-mrt-frr-algorithm-01

Abstract

A complete solution for IP and LDP Fast-Reroute using Maximally Redundant Trees is presented in [I-D.ietf-rtgwg-mrt-frr-architecture]. This document defines the associated MRT Lowpoint algorithm that is used in the default MRT profile to compute both the necessary Maximally Redundant Trees with their associated next-hops and the alternates to select for MRT-FRR.

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 5, 2015.

Table of Contents

1. Introduction

MRT Fast-Reroute requires that packets can be forwarded not only on the shortest-path tree, but also on two Maximally Redundant Trees (MRTs), referred to as the MRT-Blue and the MRT-Red. A router which experiences a local failure must also have pre-determined which alternate to use. This document defines how to compute these three things for use in MRT-FRR and describes the algorithm design decisions and rationale. The algorithm is based on those presented in [MRTLinear] and expanded in [EnyediThesis]. The MRT Lowpoint algorithm is required for implementation when the default MRT profile is implemented.

Just as packets routed on a hop-by-hop basis require that each router compute a shortest-path tree which is consistent, it is necessary for each router to compute the MRT-Blue next-hops and MRT-Red next-hops in a consistent fashion. This document defines the MRT Lowpoint algorithm to be used as a standard in the default MRT profile for MRT-FRR.

As now, a router's FIB will contain primary next-hops for the current shortest-path tree for forwarding traffic. In addition, a router's FIB will contain primary next-hops for the MRT-Blue for forwarding received traffic on the MRT-Blue and primary next-hops for the MRT-Red for forwarding received traffic on the MRT-Red.

What alternate next-hops a point-of-local-repair (PLR) selects need not be consistent - but loops must be prevented. To reduce congestion, it is possible for multiple alternate next-hops to be selected; in the context of MRT alternates, each of those alternate next-hops would be equal-cost paths.

This document defines an algorithm for selecting an appropriate MRT alternate for consideration. Other alternates, e.g. LFAs that are downstream paths, may be prefered when available and that policy-based alternate selection process[I-D.ietf-rtgwg-lfa-manageability] is not captured in this document.

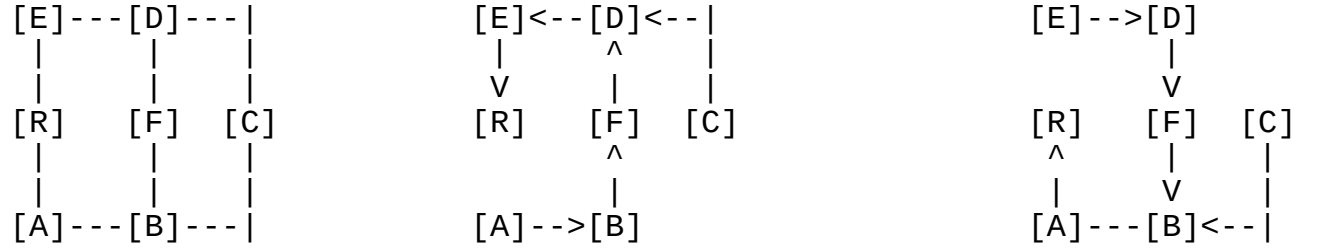
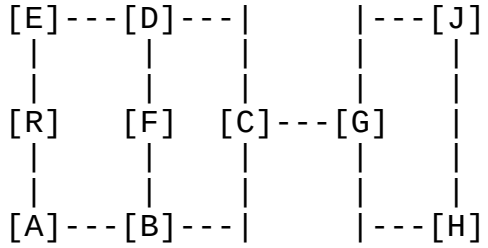


Figure 1

Algorithms for computing MRTs can handle arbitrary network topologies where the whole network graph is not 2-connected, as in Figure 2, as well as the easier case where the network graph is 2-connected (Figure 1). Each MRT is a spanning tree. The pair of MRTs provide two paths from every node X to the root of the MRTs. Those paths 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.



(a) a graph that isn't 2-connected

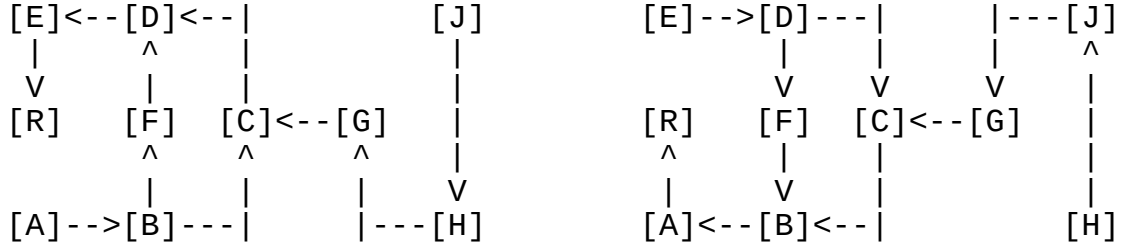


Figure 2

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 and Definitions

network graph: A graph that reflects the network topology where all links connect exactly two nodes and broadcast links have been transformed into a pseudo-node representation (e.g. in OSPF, viewing a Network LSA as representing a pseudo-noe).

Redundant Trees (RT): A pair of trees where the path from any node X to the root R on 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 Island: From the computing router, the set of routers that support a particular MRT profile and are connected.

MRT-Red: MRT-Red is used to describe one of the two MRTs; it is used to describe 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 describe 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.

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

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.

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.

spanning tree: A tree containing links that connects all nodes in the network graph.

back-edge: In the context of a spanning tree computed via a depth-first search, a back-edge is a link that connects a descendant of a node x with an ancestor of x.

2-connected cluster: A maximal set of nodes that are 2-connected. In a network graph with at least one cut-vertex, there will be multiple 2-connected clusters.

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

DAG: Directed Acyclic Graph - a digraph containing no directed cycle.

ADAG: Almost Directed Acyclic Graph - a digraph that can be transformed into a DAG with removing a single node (the root node).

partial ADAG: A subset of an ADAG that doesn’t yet contain all the nodes in the block. A partial ADAG is created during the MRT algorithm and then expanded until all nodes in the block are included and it is an ADAG.

GADAG: Generalized ADAG - a digraph, which has only ADAGs as all of its blocks. The root of such a block is the node closest to the global root (e.g. with uniform link costs).

DFS: Depth-First Search

DFS ancestor: A node n is a DFS ancestor of x if n is on the DFS-tree path from the DFS root to x.

DFS descendant: A node n is a DFS descendant of x if x is on the DFS-tree path from the DFS root to n.

ear: A path along not-yet-included-in-the-GADAG nodes that starts at a node that is already-included-in-the-GADAG and that ends at a node that is already-included-in-the-GADAG. The starting and ending nodes may be the same node if it is a cut-vertex.

X >> Y or Y << X: Indicates the relationship between X and Y in a partial order, such as found in a GADAG. X >> Y means that X is higher in the partial order than Y. Y << X means that Y is lower in the partial order than X.

X > Y or Y < X: Indicates the relationship between X and Y in the total order, such as found via a topological sort. X > Y means that X is higher in the total order than Y. Y < X means that Y is lower in the total order than X.

proxy-node: A node added to the network graph to represent a multi-homed prefix or routers outside the local MRT-fast-reroute-supporting island of routers. The key property of proxy-nodes is that traffic cannot transit them.

UNDIRECTED: In the GADAG, each link is marked as OUTGOING, INCOMING or both. Until the directionality of the link is determined, the link is marked as UNDIRECTED to indicate that its direction hasn't been determined.

OUTGOING: In the GADAG, each link is marked as OUTGOING, INCOMING or both. A link marked as OUTGOING has direction from the interface's router to the remote end.

INCOMING: In the GADAG, each link is marked as OUTGOING, INCOMING or both. A link marked as INCOMING has direction from the remote end to the interface's router.

4. Algorithm Key Concepts

There are five key concepts that are critical for understanding the MRT Lowpoint algorithm and other algorithms for computing MRTs. The first is the idea of partially ordering the nodes in a network graph with regard to each other and to the GADAG root. The second is the idea of finding an ear of nodes and adding them in the correct direction. The third is the idea of a Low-Point value and how it can be used to identify cut-vertices and to find a second path towards the root. The fourth is the idea that a non-2-connected graph is made up of blocks, where a block is a 2-connected cluster, a cut-link or an isolated node. The fifth is the idea of a local-root for each node; this is used to compute ADAGs in each block.

4.1. Partial Ordering for Disjoint Paths

Given any two nodes X and Y in a graph, a particular total order means that either X < Y or X > Y in that total order. An example would be a graph where the nodes are ranked based upon their unique IP loopback addresses. In a partial order, there may be some nodes

for which it can't be determined whether X << Y or X >> Y. A partial order can be captured in a directed graph, as shown in Figure 3. In a graphical representation, a link directed from X to Y indicates that X is a neighbor of Y in the network graph and X << Y.

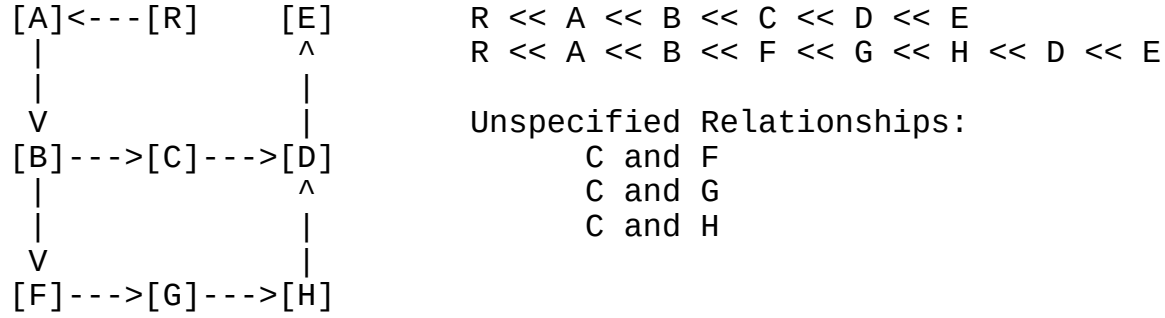


Figure 3: Directed Graph showing a Partial Order

To compute MRTs, the root of the MRTs is at both the very bottom and the very top of the partial ordering. This means that from any node X, one can pick nodes higher in the order until the root is reached. Similarly, from any node X, one can pick nodes lower in the order until the root is reached. For instance, in Figure 4, from G the higher nodes picked can be traced by following the directed links and are H, D, E and R. Similarly, from G the lower nodes picked can be traced by reversing the directed links and are F, B, A, and R. A graph that represents this modified partial order is no longer a DAG; it is termed an Almost DAG (ADAG) because if the links directed to the root were removed, it would be a DAG.

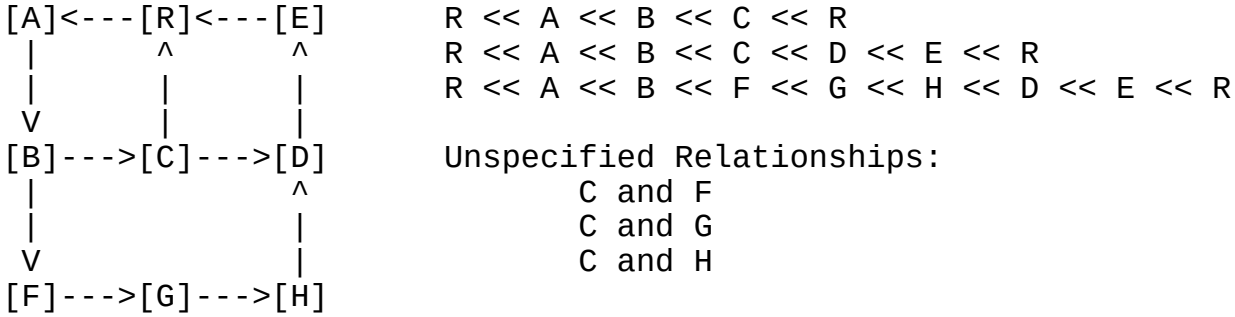


Figure 4: ADAG showing a Partial Order with R lowest and highest

Most importantly, if a node Y >> X, then Y can only appear on the increasing path from X to the root and never on the decreasing path.

Similarly, if a node Z << X, then Z can only appear on the decreasing path from X to the root and never on the inceasing path.

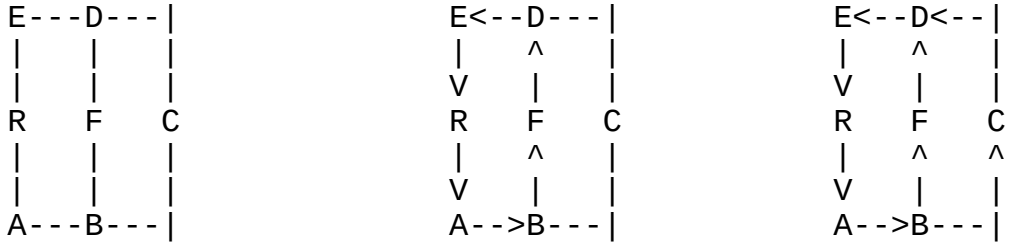
When following the increasing paths, it is possible to pick multiple higher nodes and still have the certainty that those paths will be disjoint from the decreasing paths. E.g. in the previous example node B has multiple possibilities to forward packets along an increasing path: it can either forward packets to C or F.

4.2. Finding an Ear and the Correct Direction

For simplicity, the basic idea of creating a GADAG by adding ears is described assuming that the network graph is a single 2-connected cluster so that an ADAG is sufficient. Generalizing to multiple blocks is done by considering the block-roots instead of the GADAG root - and the actual algorithm is given in Section 5.4.

In order to understand the basic idea of finding an ADAG, first suppose that we have already a partial ADAG, which doesn't contain all the nodes in the block yet, and we want to extend it to cover all the nodes. Suppose that we find a path from a node X to Y such that X and Y are already contained by our partial ADAG, but all the remaining nodes along the path are not added to the ADAG yet. We refer to such a path as an ear.

Recall that our ADAG is closely related to a partial order. More precisely, if we remove root R, the remaining DAG describes a partial order of the nodes. If we suppose that neither X nor Y is the root, we may be able to compare them. If one of them is definitely lesser with respect to our partial order (say X<<Y), we can add the new path to the ADAG in a direction from X to Y. As an example consider Figure 5.



(a) A 2-connected graph
(b) Partial ADAG (C is not included)
(c) Resulting ADAG after adding path (or ear) B-C-D

Figure 5

In this partial ADAG, node C is not yet included. However, we can find path B-C-D, where both endpoints are contained by this partial ADAG (we say those nodes are "ready" in the following text), and the remaining node (node C) is not contained yet. If we remove R, the remaining DAG defines a partial order, and with respect to this partial order we can say that B<<D, so we can add the path to the ADAG in the direction from B to D (arcs B->C and C->D are added). If B >> D, we would add the same path in reverse direction.

If in the partial order where an ear's two ends are X and Y, X << Y, then there must already be a directed path from X to Y in the ADAG. The ear must be added in a direction such that it doesn't create a cycle; therefore the ear must go from X to Y.

In the case, when X and Y are not ordered with each other, we can select either direction for the ear. We have no restriction since neither of the directions can result in a cycle. In the corner case when one of the endpoints of an ear, say X, is the root (recall that the two endpoints must be different), we could use both directions again for the ear because the root can be considered both as smaller and as greater than Y. However, we strictly pick that direction in which the root is lower than Y. The logic for this decision is explained in Section 5.6

A partial ADAG is started by finding a cycle from the root R back to itself. This can be done by selecting a non-ready neighbor N of R and then finding a path from N to R that doesn't use any links between R and N. The direction of the cycle can be assigned either way since it is starting the ordering.

Once a partial ADAG is already present, it will always have a node that is not the root R in it. As a brief proof that a partial ADAG can always have ears added to it: just select a non-ready neighbor N of a ready node Q, such that Q is not the root R, find a path from N to the root R in the graph with Q removed. This path is an ear where the first node of the ear is Q, the next is N, then the path until the first ready node the path reached (that ready node is the other endpoint of the path). Since the graph is 2-connected, there must be a path from N to R without Q.

It is always possible to select a non-ready neighbor N of a ready node Q so that Q is not the root R. Because the network is 2-connected, N must be connected to two different nodes and only one can be R. Because the initial cycle has already been added to the ADAG, there are ready nodes that are not R. Since the graph is 2-connected, while there are non-ready nodes, there must be a non-ready neighbor N of a ready node that is not R.

```
Generic_Find_Ears_ADAG(root)
   Create an empty ADAG.  Add root to the ADAG.
   Mark root as IN_GADAG.
   Select an arbitrary cycle containing root.
   Add the arbitrary cycle to the ADAG.
   Mark cycle's nodes as IN_GADAG.
   Add cycle's non-root nodes to process_list.
   while there exists connected nodes in graph that are not IN_GADAG
      Select a new ear.  Let its endpoints be X and Y.
      if Y is root or (Y << X)
         add the ear towards X to the ADAG
      else // (a) X is root or (b)X << Y or (c) X, Y not ordered
         Add the ear towards Y to the ADAG
```

Figure 6: Generic Algorithm to find ears and their direction in 2-connected graph

Algorithm Figure 6 merely requires that a cycle or ear be selected without specifying how. Regardless of the way of selecting the path, we will get an ADAG. The method used for finding and selecting the ears is important; shorter ears result in shorter paths along the MRTs. The MRT Lowpoint algorithm's method using Low-Point Inheritance is defined in Section 5.4. Other methods are described in the Appendices (Appendix A and Appendix B).

As an example, consider Figure 5 again. First, we select the shortest cycle containing R, which can be R-A-B-F-D-E (uniform link costs were assumed), so we get to the situation depicted in Figure 5 (b). Finally, we find a node next to a ready node; that must be node C and assume we reached it from ready node B. We search a path from C to R without B in the original graph. The first ready node along this is node D, so the open ear is B-C-D. Since B<<D, we add arc B->C and C->D to the ADAG. Since all the nodes are ready, we stop at this point.

4.3. Low-Point Values and Their Uses

A basic way of computing a spanning tree on a network graph is to run a depth-first-search, such as given in Figure 7. This tree has the important property that if there is a link (x, n), then either n is a DFS ancestor of x or n is a DFS descendant of x. In other words, either n is on the path from the root to x or x is on the path from the root to n.

```
global_variable: dfs_number

DFS_Visit(node x, node parent)
   D(x) = dfs_number
   dfs_number += 1
   x.dfs_parent = parent
   for each link (x, w)
     if D(w) is not set
       DFS_Visit(w, x)

Run_DFS(node root)
   dfs_number = 0
   DFS_Visit(root, NONE)
```

Figure 7: Basic Depth-First Search algorithm

Given a node x, one can compute the minimal DFS number of the neighbours of x, i.e. min(D(w) if (x,w) is a link). This gives the earliest attachment point neighbouring x. What is interesting, though, is what is the earliest attachment point from x and x's descendants. This is what is determined by computing the Low-Point value.

In order to compute the low point value, the network is traversed using DFS and the vertices are numbered based on the DFS walk. Let this number be represented as DFS(x). All the edges that lead to already visited nodes during DFS walk are back-edges. The back-edges are important because they give information about reachability of a node via another path.

The low point number is calculated by finding:

```
Low(x) = Minimum of (  (DFS(x),
   Lowest DFS(n, x->n is a back-edge),
   Lowest Low(n, x->n is tree edge in DFS walk) ).
```

A detailed algorithm for computing the low-point value is given in Figure 8. Figure 9 illustrates how the lowpoint algorithm applies to a example graph.

```
global_variable: dfs_number

Lowpoint_Visit(node x, node parent, interface p_to_x)
   D(x) = dfs_number
   L(x) = D(x)
   dfs_number += 1
   x.dfs_parent = parent
   x.dfs_parent_intf = p_to_x
   x.lowpoint_parent = NONE
   for each interface intf of x
     if D(intf.remote_node) is not set
       Lowpoint_Visit(intf.remote_node, x, intf)
       if L(intf.remote_node) < L(x)
          L(x) = L(intf.remote_node)
          x.lowpoint_parent = intf.remote_node
          x.lowpoint_parent_intf = intf
     else if intf.remote_node is not parent
       if D(intf.remote_node) < L(x)
          L(x) = D(intf.remote_node)
          x.lowpoint_parent = intf.remote_node
          x.lowpoint_parent_intf = intf

Run_Lowpoint(node root)
   dfs_number = 0
   Lowpoint_Visit(root, NONE, NONE)
```

Figure 8: Computing Low-Point value

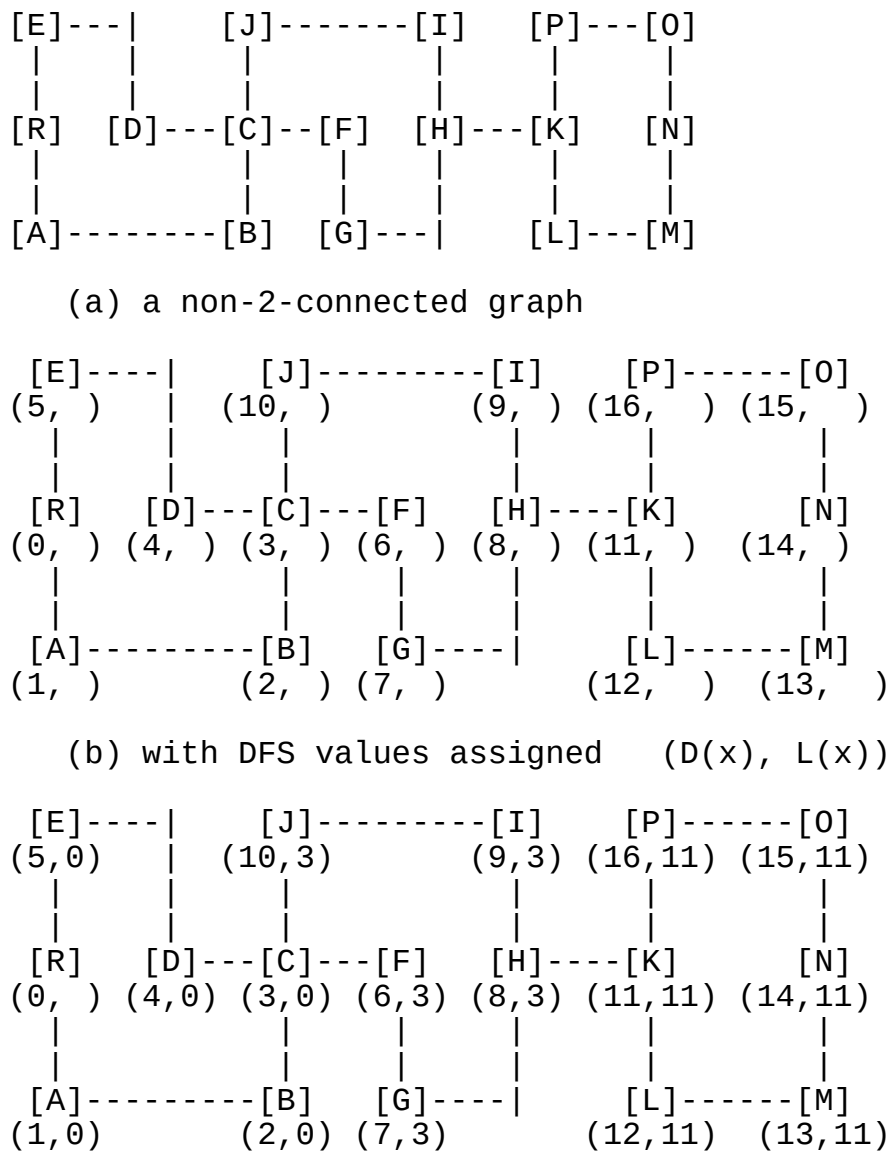


Figure 9: Example lowpoint value computation

From the low-point value and lowpoint parent, there are three very useful things which motivate our computation.

First, if there is a child c of x such that L(c) >= D(x), then there are no paths in the network graph that go from c or its descendants to an ancestor of x - and therefore x is a cut-vertex. In Figure 9, this can be seen by looking at the DFS children of C. C has two children - D and F and L(F) = 3 = D(C) so it is clear that C is a cut-vertex and F is in a block where C is the block’s root. L(D) = 0

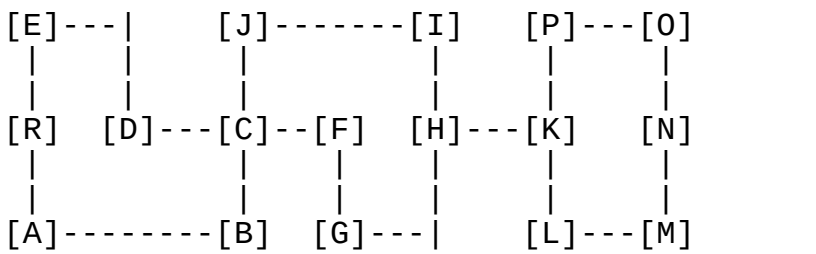
> 3 = D(C) so D has a path to the ancestors of C; in this case, D can go via E to reach R. Comparing the low-point values of all a node's DFS-children with the node's DFS-value is very useful because it allows identification of the cut-vertices and thus the blocks.

Second, by repeatedly following the path given by lowpoint_parent, there is a path from x back to an ancestor of x that does not use the link [x, x.dfs_parent] in either direction. The full path need not be taken, but this gives a way of finding an initial cycle and then ears.

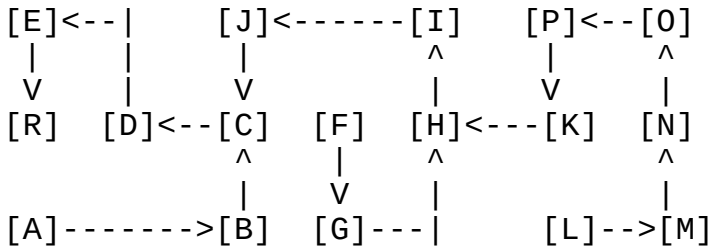
Third, as seen in Figure 9, even if L(x) < D(x), there may be a block that contains both the root and a DFS-child of a node while other DFS-children might be in different blocks. In this example, C's child D is in the same block as R while F is not. It is important to realize that the root of a block may also be the root of another block.

4.4. Blocks in a Graph

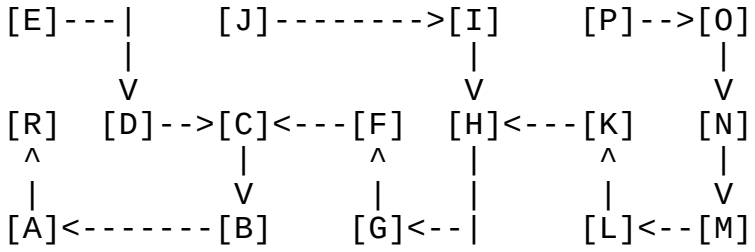
A key idea for an MRT algorithm is that any non-2-connected graph is made up by blocks (e.g. 2-connected clusters, cut-links, and/or isolated nodes). To compute GADAGs and thus MRTs, computation is done in each block to compute ADAGs or Redundant Trees and then those ADAGs or Redundant Trees are combined into a GADAG or MRT.



(a) A graph with four blocks that are:
3 2-connected clusters and a cut-link



(b) MRT-Blue



(c) MRT-Red

Figure 10

Consider the example depicted in Figure 10 (a). In this figure, a special graph is presented, showing us all the ways 2-connected clusters can be connected. It has four blocks: block 1 contains R, A, B, C, D, E, block 2 contains C, F, G, H, I, J, block 3 contains K, L, M, N, O, P, and block 4 is a cut-link containing H and K. As can be observed, the first two blocks have one common node (node C) and blocks 2 and 3 do not have any common node, but they are connected through a cut-link that is block 4. No two blocks can have more than one common node, since two blocks with at least 2 common nodes would qualify as a single 2-connected cluster.

Moreover, observe that if we want to get from one block to another, we must use a cut-vertex (the cut-vertices in this graph are C, H, K), regardless of the path selected, so we can say that all the paths from block 3 along the MRTs rooted at R will cross K first. This observation means that if we want to find a pair of MRTs rooted at R, then we need to build up a pair of RTs in block 3 with K as a root. Similarly, we need to find another pair of RTs in block 2 with C as a root, and finally, we need the last pair of RTs in block 1 with R as a root. When all the trees are selected, we can simply combine them; when a block is a cut-link (as in block 4), that cut-link is added in the same direction to both of the trees. The resulting trees are depicted in Figure 10 (b) and (c).

Similarly, to create a GADAG it is sufficient to compute ADAGs in each block and connect them.

It is necessary, therefore, to identify the cut-vertices, the blocks and identify the appropriate local-root to use for each block.

4.5. Determining Local-Root and Assigning Block-ID

Each node in a network graph has a local-root, which is the cut-vertex (or root) in the same block that is closest to the root. The local-root is used to determine whether two nodes share a common block.

```
Compute_Localroot(node x, node localroot)
    x.localroot = localroot
    for each DFS child c
        if L(c) < D(x)   //x is not a cut-vertex
            Compute_Localroot(c, x.localroot)
        else
            mark x as cut-vertex
            Compute_Localroot(c, x)

Compute_Localroot(root, root)
```

Figure 11: A method for computing local-roots

There are two different ways of computing the local-root for each node. The stand-alone method is given in Figure 11 and better illustrates the concept; it is used by the MRT algorithms given in the Appendices Appendix A and Appendix B. The MRT Lowpoint algorithm computes the local-root for a block as part of computing the GADAG using lowpoint inheritance; the essence of this computation is given in Figure 12.

```
Get the current node, s.
Compute an ear(either through lowpoint inheritance
or by following dfs parents) from s to a ready node e.
(Thus, s is not e, if there is such ear.)
if s is e
   for each node x in the ear that is not s
       x.localroot = s
else
   for each node x in the ear that is not s or e
       x.localroot = e.localroot
```

Figure 12: Ear-based method for computing local-roots

Once the local-roots are known, two nodes X and Y are in a common block if and only if one of the following three conditions apply.

- Y's local-root is X's local-root : They are in the same block and neither is the cut-vertex closest to the root.
- Y's local-root is X: X is the cut-vertex closest to the root for Y's block
- Y is X's local-root: Y is the cut-vertex closest to the root for X's block

Once we have computed the local-root for each node in the network graph, we can assign for each node, a block id that represents the block in which the node is present. This computation is shown in Figure 13.

```
global_var: max_block_id

Assign_Block_ID(x, cur_block_id)
  x.block_id = cur_block_id
  foreach DFS child c of x
     if (c.local_root is x)
        max_block_id += 1
        Assign_Block_ID(c, max_block_id)
     else
       Assign_Block_ID(c, cur_block_id)

max_block_id = 0
Assign_Block_ID(root, max_block_id)
```

Figure 13: Assigning block id to identify blocks

5. Algorithm Sections

This algorithm computes one GADAG that is then used by a router to determine its MRT-Blue and MRT-Red next-hops to all destinations. Finally, based upon that information, alternates are selected for each next-hop to each destination. The different parts of this algorithm are described below. These work on a network graph after its interfaces have been ordered as per Figure 14.

1. Compute the local MRT Island for the particular MRT Profile. [See Section 5.1.]

2. Select the root to use for the GADAG. [See Section 5.2.]

3. Initialize all interfaces to UNDIRECTED. [See Section 5.3.]

4. Compute the DFS value,e.g. D(x), and lowpoint value, L(x). [See Figure 8.]

5. Construct the GADAG. [See Section 5.4]

6. Assign directions to all interfaces that are still UNDIRECTED. [See Section 5.5.]

7. From the computing router x, compute the next-hops for the MRT-Blue and MRT-Red. [See Section 5.6.]

8. Identify alternates for each next-hop to each destination by determining which one of the blue MRT and the red MRT the computing router x should select. [See Section 5.7.]

To ensure consistency in computation, all routers MUST order interfaces identically down to the set of links with the same metric to the same neighboring node. This is necessary for the DFS, where the selection order of the interfaces to explore results in different trees, and for computing the GADAG, where the selection order of the interfaces to use to form ears can result in different GADAGs. The required ordering between two interfaces from the same router x is given in Figure 14.

```
  Interface_Compare(interface a, interface b)
    if a.metric < b.metric
       return A_LESS_THAN_B
    if b.metric < a.metric
       return B_LESS_THAN_A
    if a.neighbor.loopback_addr < b.neighbor.loopback_addr
       return A_LESS_THAN_B
    if b.neighbor.loopback_addr < a.neighbor.loopback_addr
       return B_LESS_THAN_A
    // Same metric to same node, so the order doesn't matter anymore for
    // interoperability.
    // To have a unique, consistent total order,
    // tie-break in OSPF based on the link's linkData as
    // distributed in an OSPF Router-LSA
    if a.link_data < b.link_data
       return A_LESS_THAN_B
    return B_LESS_THAN_A
```

Figure 14: Rules for ranking multiple interfaces. Order is from low to high.

5.1. MRT Island Identification

The local MRT Island for a particular MRT profile can be determined by starting from the computing router in the network graph and doing a breadth-first-search (BFS). The BFS explores only links that are in the same area/level, are not IGP-excluded, and are not MRT-ineligible. The BFS explores only nodes that are are not IGP-excluded, and that support the particular MRT profile. See section 7 of [I-D.ietf-rtgwg-mrt-frr-architecture] for more precise definitions of these criteria.

```
MRT_Island_Identification(topology, computing_rtr, profile_id, area)
  for all routers in topology
      rtr.IN_MRT_ISLAND = FALSE
  computing_rtr.IN_MRT_ISLAND = TRUE
  explore_list = { computing_rtr }
  while (explore_list is not empty)
     next_rtr = remove_head(explore_list)
     for each interface in next_rtr
        if interface is (not MRT-ineligible and not IGP-excluded
                         and in area)
           if ((interface.remote_node supports profile_id) and
               (interface.remote_node.IN_MRT_ISLAND is FALSE))
              interface.remote_node.IN_MRT_ISLAND = TRUE
              add_to_tail(explore_list, interface.remote_node)
```

Figure 15: MRT Island Identification

5.2. GADAG Root Selection

In Section 8.3 of [I-D.ietf-rtgwg-mrt-frr-architecture], the GADAG Root Selection Policy is described for the MRT default profile. In [I-D.atlas-ospf-mrt] and [I-D.li-isis-mrt], a mechanism is given for routers to advertise the GADAG Root Selection Priority and consistently select a GADAG Root inside the local MRT Island. The MRT Lowpoint algorithm simply requires that all routers in the MRT Island MUST select the same GADAG Root; the mechanism can vary based upon the MRT profile description. Before beginning computation, the network graph is reduced to contain only the set of routers that support the specific MRT profile whose MRTs are being computed.

Analysis has shown that the centrality of a router can have a significant impact on the lengths of the alternate paths computed. Therefore, it is RECOMMENDED that off-line analysis that considers the centrality of a router be used to help determine how good a choice a particular router is for the role of GADAG root.

5.3. Initialization

Before running the algorithm, there is the standard type of initialization to be done, such as clearing any computed DFS-values, lowpoint-values, DFS-parents, lowpoint-parents, any MRT-computed next-hops, and flags associated with algorithm.

It is assumed that a regular SPF computation has been run so that the primary next-hops from the computing router to each destination are known. This is required for determining alternates at the last step.

Initially, all interfaces MUST be initialized to UNDIRECTED. Whether they are OUTGOING, INCOMING or both is determined when the GADAG is constructed and augmented.

It is possible that some links and nodes will be marked as unusable using standard IGP mechanisms (see section 7 of [I-D.ietf-rtgwg-mrt-frr-architecture]). Due to FRR manageability considerations [I-D.ietf-rtgwg-lfa-manageability], it may also be desirable to administratively configure some interfaces as ineligible to carry MRT FRR traffic. This constraint MUST be consistently flooded via the IGP [I-D.atlas-ospf-mrt] [I-D.li-isis-mrt] by the owner of the interface, so that links are clearly known to be MRT-ineligible and not explored or used in the MRT algorithm. In the algorithm description, it is assumed that such links and nodes will not be explored or used, and no more discussion is given of this restriction.

5.4. MRT Lowpoint Algorithm: Computing GADAG using lowpoint inheritance

As discussed in Section 4.2, it is necessary to find ears from a node x that is already in the GADAG (known as IN_GADAG). Two different methods are used to find ears in the algorithm. The first is by going to a not IN_GADAG DFS-child and then following the chain of low-point parents until an IN_GADAG node is found. The second is by going to a not IN_GADAG neighbor and then following the chain of DFS parents until an IN_GADAG node is found. As an ear is found, the associated interfaces are marked based on the direction taken. The nodes in the ear are marked as IN_GADAG. In the algorithm, first the ears via DFS-children are found and then the ears via DFS-neighbors are found.

By adding both types of ears when an IN_GADAG node is processed, all ears that connect to that node are found. The order in which the IN_GADAG nodes is processed is, of course, key to the algorithm. The order is a stack of ears so the most recent ear is found at the top of the stack. Of course, the stack stores nodes and not ears, so an ordered list of nodes, from the first node in the ear to the last node in the ear, is created as the ear is explored and then that list is pushed onto the stack.

Each ear represents a partial order (see Figure 4) and processing the nodes in order along each ear ensures that all ears connecting to a node are found before a node higher in the partial order has its ears explored. This means that the direction of the links in the ear is always from the node x being processed towards the other end of the ear. Additionally, by using a stack of ears, this means that any unprocessed nodes in previous ears can only be ordered higher than nodes in the ears below it on the stack.

In this algorithm that depends upon Low-Point inheritance, it is necessary that every node have a low-point parent that is not itself. If a node is a cut-vertex, that may not yet be the case. Therefore, any nodes without a low-point parent will have their low-point parent set to their DFS parent and their low-point value set to the DFS-value of their parent. This assignment also properly allows an ear between two cut-vertices.

Finally, the algorithm simultaneously computes each node’s local-root, as described in Figure 12. This is further elaborated as follows. The local-root can be inherited from the node at the end of the ear unless the end of the ear is x itself, in which case the local-root for all the nodes in the ear would be x. This is because whenever the first cycle is found in a block, or an ear involving a bridge is computed, the cut-vertex closest to the root would be x itself. In all other scenarios, the properties of lowpoint/dfs

parents ensure that the end of the ear will be in the same block, and thus inheriting its local-root would be the correct local-root for all newly added nodes.

The pseudo-code for the GADAG algorithm (assuming that the adjustment of lowpoint for cut-vertices has been made) is shown in Figure 16.

```
        Construct_Ear(x, Stack, intf, type)
           ear_list = empty
           cur_node = intf.remote_node
           cur_intf = intf
           not_done = true

           while not_done
              cur_intf.UNDIRECTED = false
              cur_intf.OUTGOING = true
              cur_intf.remote_intf.UNDIRECTED = false
              cur_intf.remote_intf.INCOMING = true

              if cur_node.IN_GADAG is false
                 cur_node.IN_GADAG = true
                 add_to_list_end(ear_list, cur_node)
                 if type is CHILD
                    cur_intf = cur_node.lowpoint_parent_intf
                    cur_node = cur_node.lowpoint_parent
                 else type must be NEIGHBOR
                    cur_intf = cur_node.dfs_parent_intf
                    cur_node = cur_node.dfs_parent
              else
                 not_done = false

           if (type is CHILD) and (cur_node is x)
              //x is a cut-vertex and the local root for
              //the block in which the ear is computed
              localroot = x
           else
              // Inherit local-root from the end of the ear
              localroot = cur_node.localroot
           while ear_list is not empty
              y = remove_end_item_from_list(ear_list)
              y.localroot = localroot
              push(Stack, y)

        Construct_GADAG_via_Lowpoint(topology, root)
          root.IN_GADAG = true
          root.localroot = root
          Initialize Stack to empty
          push root onto Stack
```

```
            while (Stack is not empty)
               x = pop(Stack)
               foreach interface intf of x
                  if ((intf.remote_node.IN_GADAG == false) and
                      (intf.remote_node.dfs_parent is x))
                      Construct_Ear(x, Stack, intf, CHILD)
               foreach interface intf of x
                  if ((intf.remote_node.IN_GADAG == false) and
                      (intf.remote_node.dfs_parent is not x))
                      Construct_Ear(x, Stack, intf, NEIGHBOR)

         Construct_GADAG_via_Lowpoint(topology, root)
```

Figure 16: Low-point Inheritance GADAG algorithm

5.5. Augmenting the GADAG by directing all links

The GADAG, regardless of the algorithm used to construct it, at this point could be used to find MRTs but the topology does not include all links in the network graph. That has two impacts. First, there might be shorter paths that respect the GADAG partial ordering and so the alternate paths would not be as short as possible. Second, there may be additional paths between a router x and the root that are not included in the GADAG. Including those provides potentially more bandwidth to traffic flowing on the alternates and may reduce congestion compared to just using the GADAG as currently constructed.

The goal is thus to assign direction to every remaining link marked as UNDIRECTED to improve the paths and number of paths found when the MRTs are computed.

To do this, we need to establish a total order that respects the partial order described by the GADAG. This can be done using Kahn's topological sort[Kahn_1962_topo_sort] which essentially assigns a number to a node x only after all nodes before it (e.g. with a link incoming to x) have had their numbers assigned. The only issue with the topological sort is that it works on DAGs and not ADAGs or GADAGs.

To convert a GADAG to a DAG, it is necessary to remove all links that point to a root of block from within that block. That provides the necessary conversion to a DAG and then a topological sort can be done. Finally, all UNDIRECTED links are assigned a direction based upon the total ordering. Any UNDIRECTED links that connect to a root of a block from within that block are assigned a direction INCOMING to that root. The exact details of this whole process are captured in Figure 17

```
Set_Block_Root_Incoming_Links(topo, root, mark_or_clear)
   foreach node x in topo
      if node x is a cut-vertex or root
         foreach interface i of x
            if (i.remote_node.localroot is x)
               if i.UNDIRECTED
                  i.OUTGOING = true
                  i.remote_intf.INCOMING = true
                  i.UNDIRECTED = false
                  i.remote_intf.UNDIRECTED = false
               if i.INCOMING
                  if mark_or_clear is MARK
                     if i.OUTGOING  // a cut-link
                        i.STORE_INCOMING = true
                        i.INCOMING = false
                        i.remote_intf.STORE_OUTGOING = true
                        i.remote_intf.OUTGOING = false
                     i.TEMP_UNUSABLE = true
                     i.remote_intf.TEMP_UNUSABLE = true
                  else
                     i.TEMP_UNUSABLE = false
                     i.remote_intf.TEMP_UNUSABLE = false
               if i.STORE_INCOMING and (mark_or_clear is CLEAR)
                  i.INCOMING = true
                  i.STORE_INCOMING = false
                  i.remote_intf.OUTGOING = true
                  i.remote_intf.STORE_OUTGOING = false

Run_Topological_Sort_GADAG(topo, root)
   Set_Block_Root_Incoming_Links(topo, root, MARK)
   foreach node x
     set x.unvisited to the count of x's incoming interfaces
        that aren't marked TEMP_UNUSABLE
   Initialize working_list to empty
   Initialize topo_order_list to empty
   add_to_list_end(working_list, root)
   while working_list is not empty
      y = remove_start_item_from_list(working_list)
      add_to_list_end(topo_order_list, y)
      foreach interface i of y
          if (i.OUTGOING) and (not i.TEMP_UNUSABLE)
             i.remote_node.unvisited -= 1
             if i.remote_node.unvisited is 0
                 add_to_list_end(working_list, i.remote_node)
    next_topo_order = 1
    while topo_order_list is not empty
        y = remove_start_item_from_list(topo_order_list)
        y.topo_order = next_topo_order
```

```
            next_topo_order += 1
        Set_Block_Root_Incoming_Links(topo, root, CLEAR)

    Add_Undirected_Links(topo, root)
        Run_Topological_Sort_GADAG(topo, root)
        foreach node x in topo
          foreach interface i of x
             if i.UNDIRECTED
               if x.topo_order < i.remote_node.topo_order
                  i.OUTGOING = true
                  i.UNDIRECTED = false
                  i.remote_intf.INCOMING = true
                  i.remote_intf.UNDIRECTED = false
               else
                  i.INCOMING = true
                  i.UNDIRECTED = false
                  i.remote_intf.OUTGOING = true
                  i.remote_intf.UNDIRECTED = false

    Add_Undirected_Links(topo, root)
```

Figure 17: Assigning direction to UNDIRECTED links

Proxy-nodes do not need to be added to the network graph. They cannot be transited and do not affect the MRTs that are computed. The details of how the MRT-Blue and MRT-Red next-hops are computed for proxy-nodes and how the appropriate alternate next-hops are selected is given in Section 5.8.

5.6. Compute MRT next-hops

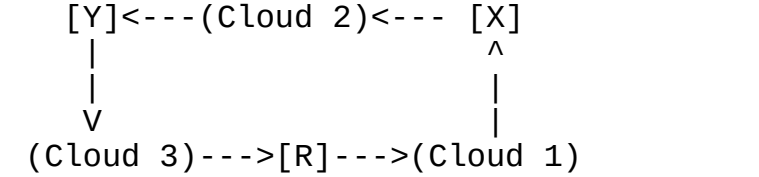
As was discussed in Section 4.1, once a ADAG is found, it is straightforward to find the next-hops from any node X to the ADAG root. However, in this algorithm, we will reuse the common GADAG and find not only the one pair of MRTs rooted at the GADAG root with it, but find a pair rooted at each node. This is useful since it is significantly faster to compute.

The method for computing differently rooted MRTs from the common GADAG is based on two ideas. First, if two nodes X and Y are ordered with respect to each other in the partial order, then an SPF along OUTGOING links (an increasing-SPF) and an SPF along INCOMING links (a decreasing-SPF) can be used to find the increasing and decreasing paths. Second, if two nodes X and Y aren’t ordered with respect to each other in the partial order, then intermediary nodes can be used to create the paths by increasing/decreasing to the intermediary and then decreasing/increasing to reach Y.

As usual, the two basic ideas will be discussed assuming the network is two-connected. The generalization to multiple blocks is discussed in Section 5.6.4. The full algorithm is given in Section 5.6.5.

5.6.1. MRT next-hops to all nodes partially ordered with respect to the computing node

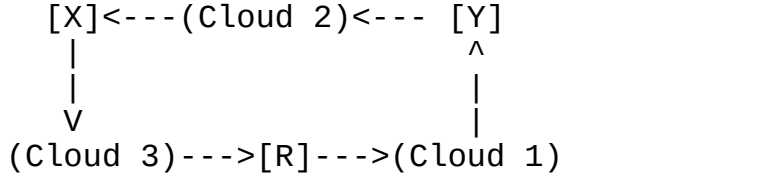
To find two node-disjoint paths from the computing router X to any node Y, depends upon whether Y >> X or Y << X. As shown in Figure 18, if Y >> X, then there is an increasing path that goes from X to Y without crossing R; this contains nodes in the interval [X,Y]. There is also a decreasing path that decreases towards R and then decreases from R to Y; this contains nodes in the interval [X,R-small] or [R-great,Y]. The two paths cannot have common nodes other than X and Y.



MRT-Blue path: X->Cloud 2->Y
MRT-Red path: X->Cloud 1->R->Cloud 3->Y

Figure 18: Y >> X

Similar logic applies if Y << X, as shown in Figure 19. In this case, the increasing path from X increases to R and then increases from R to Y to use nodes in the intervals [X,R-great] and [R-small, Y]. The decreasing path from X reaches Y without crossing R and uses nodes in the interval [Y,X].



MRT-Blue path: X->Cloud 3->R->Cloud 1->Y
MRT-Red path: X->Cloud 2->Y

Figure 19: Y << X

5.6.2. MRT next-hops to all nodes not partially ordered with respect to the computing node

When X and Y are not ordered, the first path should increase until we get to a node G, where G >> Y. At G, we need to decrease to Y. The other path should be just the opposite: we must decrease until we get to a node H, where H << Y, and then increase. Since R is smaller and greater than Y, such G and H must exist. It is also easy to see that these two paths must be node disjoint: the first path contains nodes in interval [X,G] and [Y,G], while the second path contains nodes in interval [H,X] and [H,Y]. This is illustrated in Figure 20. It is necessary to decrease and then increase for the MRT-Blue and increase and then decrease for the MRT-Red; if one simply increased for one and decreased for the other, then both paths would go through the root R.

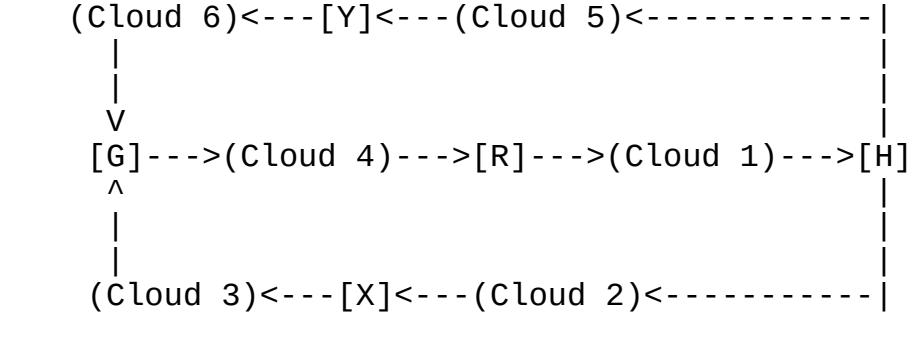


Figure 20: X and Y unordered

This gives disjoint paths as long as G and H are not the same node. Since G >> Y and H << Y, if G and H could be the same node, that would have to be the root R. This is not possible because there is only one incoming interface to the root R which is created when the initial cycle is found. Recall from Figure 6 that whenever an ear was found to have an end that was the root R, the ear was directed from R so that the associated interface on R is outgoing and not incoming. Therefore, there must be exactly one node M which is the largest one before R, so the MRT-Red path will never reach R; it will turn at M and decrease to Y.

5.6.3. Computing Redundant Tree next-hops in a 2-connected Graph

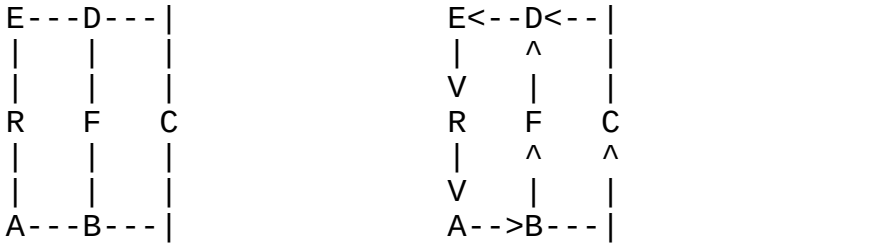
The basic ideas for computing RT next-hops in a 2-connected graph were given in Section 5.6.1 and Section 5.6.2. Given these two ideas, how can we find the trees?

If some node X only wants to find the next-hops (which is usually the case for IP networks), it is enough to find which nodes are greater and less than X, and which are not ordered; this can be done by running an increasing-SPF and a decreasing-SPF rooted at X and not exploring any links from the ADAG root. (Traversal algorithms other than SPF could safely be used instead where one traversal takes the links in their given directions and the other reverses the links' directions.)

An increasing-SPF rooted at X and not exploring links from the root will find the increasing next-hops to all Y >> X. Those increasing next-hops are X's next-hops on the MRT-Blue to reach Y. A decreasing-SPF rooted at X and not exploring links from the root will find the decreasing next-hops to all Z << X. Those decreasing next-hops are X's next-hops on the MRT-Red to reach Z. Since the root R is both greater than and less than X, after this increasing-SPF and decreasing-SPF, X's next-hops on the MRT-Blue and on the MRT-Red to reach R are known. For every node Y >> X, X's next-hops on the MRT-Red to reach Y are set to those on the MRT-Red to reach R. For every node Z << X, X's next-hops on the MRT-Blue to reach Z are set to those on the MRT-Blue to reach R.

For those nodes which were not reached by either the increasing-SPF or the decreasing-SPF, we can determine the next-hops as well. The increasing MRT-Blue next-hop for a node which is not ordered with respect to X is the next-hop along the decreasing MRT-Red towards R, and the decreasing MRT-Red next-hop is the next-hop along the increasing MRT-Blue towards R. Naturally, since R is ordered with respect to all the nodes, there will always be an increasing and a decreasing path towards it. This algorithm does not provide the complete specific path taken but just the appropriate next-hops to use. The identities of G and H are not determined by the computing node X.

The final case to considered is when the root R computes its own next-hops. Since the root R is << all other nodes, running an increasing-SPF rooted at R will reach all other nodes; the MRT-Blue next-hops are those found with this increasing-SPF. Similarly, since the root R is >> all other nodes, running a decreasing-SPF rooted at R will reach all other nodes; the MRT-Red next-hops are those found with this decreasing-SPF.



(a) A 2-connected graph (b) A spanning ADAG rooted at R

Figure 21

As an example consider the situation depicted in Figure 21. Node C runs an increasing-SPF and a decreasing-SPF on the ADAG. The increasing-SPF reaches D, E and R and the decreasing-SPF reaches B, A and R. E>>C. So towards E the MRT-Blue next-hop is D, since E was reached on the increasing path through D. And the MRT-Red next-hop towards E is B, since R was reached on the decreasing path through B. Since E>>D, D will similarly compute its MRT-Blue next-hop to be E, ensuring that a packet on MRT-Blue will use path C-D-E. B, A and R will similarly compute the MRT-Red next-hops towards E (which is ordered less than B, A and R), ensuring that a packet on MRT-Red will use path C-B-A-R-E.

C can determine the next-hops towards F as well. Since F is not ordered with respect to C, the MRT-Blue next-hop is the decreasing one towards R (which is B) and the MRT-Red next-hop is the increasing one towards R (which is D). Since F>>B, for its MRT-Blue next-hop towards F, B will use the real increasing next-hop towards F. So a packet forwarded to B on MRT-Blue will get to F on path C-B-F. Similarly, D will use the real decreasing next-hop towards F as its MRT-Red next-hop, an packet on MRT-Red will use path C-D-F.

5.6.4. Generalizing for a graph that isn’t 2-connected

If a graph isn’t 2-connected, then the basic approach given in Section 5.6.3 needs some extensions to determine the appropriate MRT next-hops to use for destinations outside the computing router X’s blocks. In order to find a pair of maximally redundant trees in that graph we need to find a pair of RTs in each of the blocks (the root of these trees will be discussed later), and combine them.

When computing the MRT next-hops from a router X, there are three basic differences:

1. Only nodes in a common block with X should be explored in the increasing-SPF and decreasing-SPF.

2. Instead of using the GADAG root, X's local-root should be used. This has the following implications:

   A. The links from X's local-root should not be explored.

   B. If a node is explored in the outgoing SPF so Y >> X, then X's MRT-Red next-hops to reach Y uses X's MRT-Red next-hops to reach X's local-root and if Z << X, then X's MRT-Blue next-hops to reach Z uses X's MRT-Blue next-hops to reach X's local-root.

   C. If a node W in a common block with X was not reached in the increasing-SPF or decreasing-SPF, then W is unordered with respect to X. X's MRT-Blue next-hops to W are X's decreasing (aka MRT-Red) next-hops to X's local-root. X's MRT-Red next-hops to W are X's increasing (aka MRT-Blue) next-hops to X's local-root.

3. For nodes in different blocks, the next-hops must be inherited via the relevant cut-vertex.

These are all captured in the detailed algorithm given in Section 5.6.5.

5.6.5. Complete Algorithm to Compute MRT Next-Hops

The complete algorithm to compute MRT Next-Hops for a particular router X is given in Figure 22. In addition to computing the MRT-Blue next-hops and MRT-Red next-hops used by X to reach each node Y, the algorithm also stores an "order_proxy", which is the proper cut-vertex to reach Y if it is outside the block, and which is used later in deciding whether the MRT-Blue or the MRT-Red can provide an acceptable alternate for a particular primary next-hop.

```
In_Common_Block(x, y)
  if (((x.localroot is y.localroot) and (x.block_id is y.block_id))
       or (x is y.localroot) or (y is x.localroot))
     return true
  return false

Store_Results(y, direction, spf_root, store_nhs)
   if direction is FORWARD
      y.higher = true
      if store_nhs
         y.blue_next_hops = y.next_hops
   if direction is REVERSE
      y.lower = true
      if store_nhs
```

```
          y.red_next_hops = y.next_hops

   SPF_No_Traverse_Root(spf_root, block_root, direction, store_nhs)
      Initialize spf_heap to empty
      Initialize nodes' spf_metric to infinity and next_hops to empty
      spf_root.spf_metric = 0
      insert(spf_heap, spf_root)
      while (spf_heap is not empty)
          min_node = remove_lowest(spf_heap)
          Store_Results(min_node, direction, spf_root, store_nhs)
          if ((min_node is spf_root) or (min_node is not block_root))
             foreach interface intf of min_node
                if (((direction is FORWARD) and intf.OUTGOING) or
                    ((direction is REVERSE) and intf.INCOMING)  and
                   In_Common_Block(spf_root, intf.remote_node))
                   path_metric = min_node.spf_metric + intf.metric
                   if path_metric < intf.remote_node.spf_metric
                      intf.remote_node.spf_metric = path_metric
                      if min_node is spf_root
                        intf.remote_node.next_hops = make_list(intf)
                      else
                        intf.remote_node.next_hops = min_node.next_hops
                      insert_or_update(spf_heap, intf.remote_node)
                   else if path_metric is intf.remote_node.spf_metric
                      if min_node is spf_root
                         add_to_list(intf.remote_node.next_hops, intf)
                      else
                         add_list_to_list(intf.remote_node.next_hops,
                                          min_node.next_hops)

   SetEdge(y)
     if y.blue_next_hops is empty and y.red_next_hops is empty
        if (y.local_root != y) {
           SetEdge(y.localroot)
        }
        y.blue_next_hops = y.localroot.blue_next_hops
        y.red_next_hops = y.localroot.red_next_hops
        y.order_proxy = y.localroot.order_proxy

   Compute_MRT_NextHops(x, root)
      foreach node y
        y.higher = y.lower = false
        clear y.red_next_hops and y.blue_next_hops
        y.order_proxy = y
      SPF_No_Traverse_Root(x, x.localroot, FORWARD, TRUE)
      SPF_No_Traverse_Root(x, x.localroot, REVERSE, TRUE)

      // red and blue next-hops are stored to x.localroot as different
```

```
      // paths are found via the SPF and reverse-SPF.
      // Similarly any nodes whose local-root is x will have their
      // red_next_hops and blue_next_hops already set.

      // Handle nodes in the same block that aren't the local-root
      foreach node y
        if (y.IN_MRT_ISLAND and (y is not x) and
            (y.localroot is x.localroot) and
            ((y is x.localroot) or (x is y.localroot) or
             (y.block_id is x.block_id)))
           if y.higher
              y.red_next_hops = x.localroot.red_next_hops
           else if y.lower
              y.blue_next_hops = x.localroot.blue_next_hops
           else
              y.blue_next_hops = x.localroot.red_next_hops
              y.red_next_hops = x.localroot.blue_next_hops

      // Inherit next-hops and order_proxies to other components
      if x is not root
         root.blue_next_hops = x.localroot.blue_next_hops
         root.red_next_hops = x.localroot.red_next_hops
         root.order_proxy = x.localroot
      foreach node y
         if (y is not root) and (y is not x) and y.IN_MRT_ISLAND
           SetEdge(y)

   max_block_id = 0
   Assign_Block_ID(root, max_block_id)
   Compute_MRT_NextHops(x, root)
```

Figure 22

5.7. Identify MRT alternates

At this point, a computing router S knows its MRT-Blue next-hops and MRT-Red next-hops for each destination in the MRT Island. The primary next-hops along the SPT are also known. It remains to determine for each primary next-hop to a destination D, which of the MRTs avoids the primary next-hop node F. This computation depends upon data set in Compute_MRT_NextHops such as each node y's y.blue_next_hops, y.red_next_hops, y.order_proxy, y.higher, y.lower and topo_orders. Recall that any router knows only which are the nodes greater and lesser than itself, but it cannot decide the relation between any two given nodes easily; that is why we need topological ordering.

For each primary next-hop node F to each destination D, S can call Select_Alternates(S, D, F, primary_intf) to determine whether to use the MRT-Blue next-hops as the alternate next-hop(s) for that primary next hop or to use the MRT-Red next-hops. The algorithm is given in Figure 23 and discussed afterwards.

```
Select_Alternates_Internal(S, D, F, primary_intf,
                           D_lower, D_higher, D_topo_order)

    //When D==F, we can do only link protection
    if ((D is F) or (D.order_proxy is F))
        if an MRT doesn't use primary_intf
            indicate alternate is not node-protecting
            return that MRT color
        else // parallel links are cut-links
            return AVOID_LINK_ON_BLUE

    if (D_lower and D_higher and F.lower and F.higher)
        if F.topo_order < D_topo_order
            return USE_RED
        else
            return USE_BLUE

    if (D_lower and D_higher)
        if F.higher
            return USE_RED
        else
            return USE_BLUE

    if (F.lower and F.higher)
        if D_lower
            return USE_RED
        else if D_higher
            return USE_BLUE
        else
            if primary_intf.OUTGOING and primary_intf.INCOMING
                return AVOID_LINK_ON_BLUE
            if primary_intf.OUTGOING is true
                return USE_BLUE
            if primary_intf.INCOMING is true
                return USE_RED

    if D_higher
        if F.higher
            if F.topo_order < D_topo_order
                return USE_RED
            else
```

```
                return USE_BLUE
            else if F.lower
                return USE_BLUE
            else
                // F and S are neighbors so either F << S or F >> S
        else if D_lower
            if F.higher
                return USE_RED
            else if F.lower
                if F.topo_order < D_topo_order
                    return USE_RED
                else
                    return USE_BLUE
                else
                    // F and S are neighbors so either F << S or F >> S
        else // D and S not ordered
            if F.lower
                return USE_RED
            else if F.higher
                return USE_BLUE
            else
                // F and S are neighbors so either F << S or F >> S


   Select_Alternates(S, D, F, primary_intf)
      if D.order_proxy is not D
         D_lower = D.order_proxy.lower
         D_higher = D.order_proxy.higher
         D_topo_order = D.order_proxy.topo_order
      else
         D_lower = D.lower
         D_higher = D.higher
         D_topo_order = D.topo_order
      return Select_Alternates_Internal(S, D, F, primary_intf,
                                        D_lower, D_higher, D_topo_order)
```

Figure 23

If either D>>S>>F or D<<S<<F holds true, the situation is simple: in the first case we should choose the increasing Blue next-hop, in the second case, the decreasing Red next-hop is the right choice.

However, when both D and F are greater than S the situation is not so simple, there can be three possibilities: (i) F>>D (ii) F<<D or (iii) F and D are not ordered. In the first case, we should choose the path towards D along the Blue tree. In contrast, in case (ii) the Red path towards the root and then to D would be the solution. Finally, in case (iii) both paths would be acceptable. However,

observe that if e.g. F.topo_order>D.topo_order, either case (i) or case (iii) holds true, which means that selecting the Blue next-hop is safe. Similarly, if F.topo_order<D.topo_order, we should select the Red next-hop. The situation is almost the same if both F and D are less than S.

Recall that we have added each link to the GADAG in some direction, so it is impossible that S and F are not ordered. But it is possible that S and D are not ordered, so we need to deal with this case as well. If F<<S, we can use the Red next-hop, because that path is first increasing until a node definitely greater than D is reached, then decreasing; this path must avoid using F. Similarly, if F>>S, we should use the Blue next-hop.

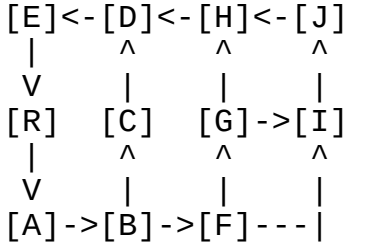
Additionally, the cases where either F or D is ordered both higher and lower must be considered; this can happen when one is a block-root or its order_proxy is. If D is both higher and lower than S, then the MRT to use is the one that avoids F so if F is higher, then the MRT-Red should be used and if F is lower, then the MRT-Blue should be used; F and S must be ordered because they are neighbors. If F is both higher and lower, then if D is lower, using the MRT-Red to decrease reaches D and if D is higher, using the Blue MRT to increase reaches D; if D is unordered compared to S, then the situation is a bit more complicated.

In the case where F<<S<<F and D and S are unordered, the direction of the link in the GADAG between S and F should be examined. If the link is directed S -> F, then use the MRT-Blue (decrease to avoid that link and then increase). If the link is directed S <- F, then use the MRT-Red (increase to avoid that link and then decrease). If the link is S <--> F, then the link must be a cut-link and there is no node-protecting alternate. If there are multiple links between S and F, then they can protect against each other; of course, in this situation, they are probably already ECMP.

Finally, there is the case where D is also F. In this case, only link protection is possible. The MRT that doesn’t use the indicated primary next-hop is used. If both MRTs use the primary next-hop, then the primary next-hop must be a cut-link so either MRT could be used but the set of MRT next-hops must be pruned to avoid that primary next-hop. To indicate this case, Select_Alternates returns AVOID_LINK_ON_BLUE.

As an example, consider the ADAG depicted in Figure 24 and first suppose that G is the source, D is the destination and H is the failed next-hop. Since D>>G, we need to compare H.topo_order and D.topo_order. Since D.topo_order>H.topo_order, D must be not smaller than H, so we should select the decreasing path towards the root.

If, however, the destination were instead J, we must find that H.topo_order>J.topo_order, so we must choose the increasing Blue next-hop to J, which is I. In the case, when instead the destination is C, we find that we need to first decrease to avoid using H, so the Blue, first decreasing then increasing, path is selected.



(a)
a 2-connected graph

Figure 24

5.8. Finding FRR Next-Hops for Proxy-Nodes

As discussed in Section 10.2 of [I-D.ietf-rtgwg-mrt-frr-architecture], it is necessary to find MRT-Blue and MRT-Red next-hops and MRT-FRR alternates for a named proxy-nodes. An example case is for a router that is not part of that local MRT Island, when there is only partial MRT support in the domain.

A first incorrect and naive approach to handling proxy-nodes, which cannot be transited, is to simply add these proxy-nodes to the graph of the network and connect it to the routers through which the new proxy-node can be reached. Unfortunately, this can introduce some new ordering between the border routers connected to the new node which could result in routing MRT paths through the proxy-node. Thus, this naive approach would need to recompute GADAGs and redo SPTs for each proxy-node.

Instead of adding the proxy-node to the original network graph, each individual proxy-node can be individually added to the GADAG. The proxy-node is connected to at most two nodes in the GADAG. Section 10.2 of [I-D.ietf-rtgwg-mrt-frr-architecture] defines how the proxy-node attachments MUST be determined. The degenerate case where the proxy-node is attached to only one node in the GADAG is trivial as all needed information can be derived from that attachment node; if there are different interfaces, then some can be assigned to MRT-Red and others to MRT_Blue.

Now, consider the proxy-node that is attached to exactly two nodes in the GADAG. Let the order_proxies of these nodes be A and B. Let the current node, where next-hop is just being calculated, be S. If one of these two nodes A and B is the local root of S, let A=S.local_root and the other one be B. Otherwise, let A.topo_order < B.topo_order.

A valid GADAG was constructed. Instead doing an increasing-SPF and a decreasing-SPF to find ordering for the proxy-nodes, the following simple rules, providing the same result, can be used independently for each different proxy-node. For the following rules, let X=A.local_root, and if A is the local root, let that be strictly lower than any other node. Always take the first rule that matches.

Rule	Condition	Blue NH	Red NH	Notes
1	S=X	Blue to A	Red to B	
2	S<<A	Blue to A	Red to R	
3	S>>B	Blue to R	Red to B	
4	A<<S<<B	Red to A	Blue to B	
5	A<<S	Red to A	Blue to R	S not ordered w/ B
6	S<<B	Red to R	Blue to B	S not ordered w/ A
7	Otherwise	Red to R	Blue to R	S not ordered w/ A+B

These rules are realized in the following pseudocode where P is the proxy-node, X and Y are the nodes that P is attached to, and S is the computing router:

```
Select_Proxy_Node_NHs(P, S, X, Y)
    if (X.order_proxy.topo_order < Y.order_proxy.topo_order)
        //This fits even if X.order_proxy=S.local_root
        A=X.order_proxy
        B=Y.order_proxy
    else
        A=Y.order_proxy
        B=X.order_proxy

    if (S==A.local_root)
        P.blue_next_hops = A.blue_next_hops
        P.red_next_hops  = B.red_next_hops
        return
    if (A.higher)
        P.blue_next_hops = A.blue_next_hops
        P.red_next_hops  = R.red_next_hops
        return
    if (B.lower)
        P.blue_next_hops = R.blue_next_hops
        P.red_next_hops  = B.red_next_hops
        return
    if (A.lower && B.higher)
        P.blue_next_hops = A.red_next_hops
        P.red_next_hops  = B.blue_next_hops
        return
    if (A.lower)
        P.blue_next_hops = R.red_next_hops
        P.red_next_hops  = B.blue_next_hops
        return
    if (B.higher)
        P.blue_next_hops = A.red_next_hops
        P.red_next_hops  = R.blue_next_hops
        return
    P.blue_next_hops = R.red_next_hops
    P.red_next_hops  = R.blue_next_hops
    return
```

After finding the the red and the blue next-hops, it is necessary to know which one of these to use in the case of failure. This can be done by Select_Alternates_Inner(). In order to use Select_Alternates_Internal(), we need to know if P is greater, less or unordered with S, and P.topo_order. P.lower = B.lower, P.higher = A.higher, and any value is OK for P.topo_order, as long as A.topo_order<=P.topo_order<=B.topo_order and P.topo_order is not equal to the topo_order of the failed node. So for simplicity let P.topo_order=A.topo_order when the next-hop is not A, and

P.topo_order=B.topo_order otherwise. This gives the following pseudo-code:

```
  Select_Alternates_Proxy_Node(S, P, F, primary_intf)
     if (F is not P.neighbor_A)
        return Select_Alternates_Internal(S, P, F, primary_intf,
                                          P.neighbor_B.lower,
                                          P.neighbor_A.higher,
                                          P.neighbor_A.topo_order)
     else
        return Select_Alternates_Internal(S, P, F, primary_intf,
                                          P.neighbor_B.lower,
                                          P.neighbor_A.higher,
                                          P.neighbor_B.topo_order)
```

Figure 25

6. MRT Lowpoint Algorithm: Next-hop conformance

This specification defines the MRT Lowpoint Algorithm, which include the construction of a common GADAG and the computation of MRT-Red and MRT-Blue next-hops to each node in the graph. An implementation MAY select any subset of next-hops for MRT-Red and MRT-Blue that respect the available nodes that are described in Section 5.6 for each of the MRT-Red and MRT-Blue and the selected next-hops are further along in the interval of allowed nodes towards the destination.

For example, the MRT-Blue next-hops used when the destination Y >> X, the computing router, MUST be one or more nodes, T, whose topo_order is in the interval [X.topo_order, Y.topo_order] and where Y >> T or Y is T. Similarly, the MRT-Red next-hops MUST be have a topo_order in the interval [R-small.topo_order, X.topo_order] or [Y.topo_order, R-big.topo_order].

Implementations SHOULD implement the Select_Alternates() function to pick an MRT-FRR alternate.

7. Algorithm Alternatives and Evaluation

This specification defines the MRT Lowpoint Algorithm, which is one option among several possible MRT algorithms. Other alternatives are described in the appendices.

In addition, it is possible to calculate Destination-Rooted GADAG, where for each destination, a GADAG rooted at that destination is computed. Then a router can compute the blue MRT and red MRT next-hops to that destination. Building GADAGs per destination is

computationally more expensive, but may give somewhat shorter alternate paths. It may be useful for live-live multicast along MRTs.

7.1. Algorithm Evaluation

The MRT Lowpoint algorithm is the lowest computation of the MRT algorithms. Two other MRT algorithms are provided in Appendix A and Appendix B. When analyzed on service provider network topologies, they did not provide significant differences in the path lenghts for the alternatives. This section does not focus on that analysis or the decision to use the MRT Lowpoint algorithm as the default MRT algorithm; it has the lowest computational and storage requirements and gave comparable results.

Since this document defines the MRT Lowpoint algorithm for use in fast-reroute applications, it is useful to compare MRT and Remote LFA [I-D.ietf-rtgwg-remote-lfa]. This section compares MRT and remote LFA for IP Fast Reroute in 19 service provider network topologies, focusing on coverage and alternate path length. Figure 26 shows the node-protecting coverage provided by local LFA (LLFA), remote LFA (RLFA), and MRT against different failure scenarios in these topologies. The coverage values are calculated as the percentage of source-destination pairs protected by the given IPFRR method relative to those protectable by optimal routing, against the same failure modes. More details on alternate selection policies used for this analysis are described later in this section.

Topology	percentage of failure scenarios covered by IPFRR method		
	NP_LLFA	NP_RLFA	MRT
T201	37	90	100
T202	73	83	100
T203	51	80	100
T204	55	81	100
T205	92	93	100
T206	71	74	100
T207	57	74	100
T208	66	81	100
T209	79	79	100
T210	95	98	100
T211	68	71	100
T212	59	63	100
T213	84	84	100
T214	68	78	100
T215	84	88	100
T216	43	59	100
T217	78	88	100
T218	72	75	100
T219	78	84	100

Figure 26

For the topologies analyzed here, LLFA is able to provide node-protecting coverage ranging from 37% to 95% of the source-destination pairs, as seen in the column labeled NP_LLFA. The use of RLFA in addition to LLFA is generally able to increase the node-protecting coverage. The percentage of node-protecting coverage with RLFA is provided in the column labeled NP_RLFA, ranges from 59% to 98% for these topologies. The node-protecting coverage provided by MRT is 100% since MRT is able to provide protection for any source-destination pair for which a path still exists after the failure.

We would also like to measure the quality of the alternate paths produced by these different IPFRR methods. An obvious approach is to take an average of the alternate path costs over all source-destination pairs and failure modes. However, this presents a problem, which we will illustrate by presenting an example of results for one topology using this approach (Figure 27). In this table, the average relative path length is the alternate path length for the IPFRR method divided by the optimal alternate path length, averaged

over all source-destination pairs and failure modes. The first three columns of data in the table give the path length calculated from the sum of IGP metrics of the links in the path. The results for topology T208 show that the metric-based path lengths for NP_LLFA and NP_RLFA alternates are on average 78 and 66 times longer than the path lengths for optimal alternates. The metric-based path lengths for MRT alternates are on average 14 times longer than for optimal alternates.

	average relative alternate path length					
Topology	IGP metric			hopcount		
	NP_LLFA	NP_RLFA	MRT	NP_LLFA	NP_RLFA	MRT
T208	78.2	66.0	13.6	0.99	1.01	1.32

Figure 27

The network topology represented by T208 uses values of 10, 100, and 1000 as IGP costs, so small deviations from the optimal alternate path can result in large differences in relative path length. LLFA, RLFA, and MRT all allow for at least one hop in the alterate path to be chosen independent of the cost of the link. This can easily result in an alternate using a link with cost 1000, which introduces noise into the path length measurement. In the case of T208, the adverse effects of using metric-based path lengths is obvious. However, we have observed that the metric-based path length introduces noise into alternate path length measurements in several other topologies as well. For this reason, we have opted to measure the alternate path length using hopcount. While IGP metrics may be adjusted by the network operator for a number of reasons (e.g. traffic engineering), the hopcount is a fairly stable measurement of path length. As shown in the last three columns of Figure 27, the hopcount-based alternate path lengths for topology T208 are fairly well-behaved.

Figure 28, Figure 29, Figure 30, and Figure 31 present the hopcount-based path length results for the 19 topologies examined. The topologies in the four tables are grouped based on the size of the topologies, as measured by the number of nodes, with Figure 28 having the smallest topologies and Figure 31 having the largest topologies. Instead of trying to represent the path lengths of a large set of alternates with a single number, we have chosen to present a histogram of the path lengths for each IPFRR method and alternate selection policy studied. The first eight colums of data represent

the percentage of failure scenarios protected by an alternate N hops longer than the primary path, with the first column representing an alternate 0 or 1 hops longer than the primary path, all the way up through the eighth column respresenting an alternate 14 or 15 hops longer than the primary path. The last column in the table gives the percentage of failure scenarios for which there is no alternate less than 16 hops longer than the primary path. In the case of LLFA and RLFA, this category includes failure scenarios for which no alternate was found.

For each topology, the first row (labeled OPTIMAL) is the distribution of the number of hops in excess of the primary path hopcount for optimally routed alternates. (The optimal routing was done with respect to IGP metrics, as opposed to hopcount.) The second row(labeled NP_LLFA) is the distribution of the extra hops for node-protecting LLFA. The third row (labeled NP_LLFA_THEN_NP_RLFA) is the hopcount distribution when one adds node-protecting RLFA to increase the coverage. The alternate selection policy used here first tries to find a node-protecting LLFA. If that does not exist, then it tries to find an RLFA, and checks if it is node-protecting. Comparing the hopcount distribution for RLFA and LLFA across these topologies, one can see how the coverage is increased at the expense of using longer alternates. It is also worth noting that while superficially LLFA and RLFA appear to have better hopcount distributions than OPTIMAL, the presence of entries in the last column (no alternate < 16) mainly represent failure scenarios that are not protected, for which the hopcount is effectively infinite.

The fourth and fifth rows of each topology show the hopcount distributions for two alternate selection policies using MRT alternates. The policy represented by the label NP_LLFA_THEN_MRT_LOWPOINT will first use a node-protecting LLFA. If a node-protecting LLFA does not exist, then it will use an MRT alternate. The policy represented by the label MRT_LOWPOINT instead will use the MRT alternate even if a node-protecting LLFA exists. One can see from the data that combining node-protecting LLFA with MRT results in a significant shortening of the alternate hopcount distribution.

Topology name and alternate selection policy evaluated	percentage of failure scenarios protected by an alternate N hops longer than the primary path								
	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	no alt <16
T201(avg primary hops=3.5)									
OPTIMAL	37	37	20	3	3				
NP_LLFA	37								63
NP_LLFA_THEN_NP_RLFA	37	34	19						10
NP_LLFA_THEN_MRT_LOWPOINT	37	33	21	6	3				
MRT_LOWPOINT	33	36	23	6	3				
T202(avg primary hops=4.8)									
OPTIMAL	90	9							
NP_LLFA	71	2							27
NP_LLFA_THEN_NP_RLFA	78	5							17
NP_LLFA_THEN_MRT_LOWPOINT	80	12	5	2	1				
MRT_LOWPOINT_ONLY	48	29	13	7	2	1			
T203(avg primary hops=4.1)									
OPTIMAL	36	37	21	4	2				
NP_LLFA	34	15	3						49
NP_LLFA_THEN_NP_RLFA	35	19	22	4					20
NP_LLFA_THEN_MRT_LOWPOINT	36	35	22	5	2				
MRT_LOWPOINT_ONLY	31	35	26	7	2				
T204(avg primary hops=3.7)									
OPTIMAL	76	20	3	1					
NP_LLFA	54	1							45
NP_LLFA_THEN_NP_RLFA	67	10	4						19
NP_LLFA_THEN_MRT_LOWPOINT	70	18	8	3	1				
MRT_LOWPOINT_ONLY	58	27	11	3	1				
T205(avg primary hops=3.4)									
OPTIMAL	92	8							
NP_LLFA	89	3							8
NP_LLFA_THEN_NP_RLFA	90	4							7
NP_LLFA_THEN_MRT_LOWPOINT	91	9							
MRT_LOWPOINT_ONLY	62	33	5	1					

Figure 28

Topology name and alternate selection policy evaluated	percentage of failure scenarios protected by an alternate N hops longer than the primary path								
	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	no alt <16
T206(avg primary hops=3.7)									
OPTIMAL	63	30	7						
NP_LLFA	60	9	1						29
NP_LLFA_THEN_NP_RLFA	60	13	1						26
NP_LLFA_THEN_MRT_LOWPOINT	64	29	7						
MRT_LOWPOINT	55	32	13						
T207(avg primary hops=3.9)									
OPTIMAL	71	24	5	1					
NP_LLFA	55	2							43
NP_LLFA_THEN_NP_RLFA	63	10							26
NP_LLFA_THEN_MRT_LOWPOINT	70	20	7	2	1				
MRT_LOWPOINT_ONLY	57	29	11	3	1				
T208(avg primary hops=4.6)									
OPTIMAL	58	28	12	2	1				
NP_LLFA	53	11	3						34
NP_LLFA_THEN_NP_RLFA	56	17	7	1					19
NP_LLFA_THEN_MRT_LOWPOINT	58	19	10	7	3	1			
MRT_LOWPOINT_ONLY	34	24	21	13	6	2	1		
T209(avg primary hops=3.6)									
OPTIMAL	85	14	1						
NP_LLFA	79								21
NP_LLFA_THEN_NP_RLFA	79								21
NP_LLFA_THEN_MRT_LOWPOINT	82	15	2						
MRT_LOWPOINT_ONLY	63	29	8						
T210(avg primary hops=2.5)									
OPTIMAL	95	4	1						
NP_LLFA	94	1							5
NP_LLFA_THEN_NP_RLFA	94	3	1						2
NP_LLFA_THEN_MRT_LOWPOINT	95	4	1						
MRT_LOWPOINT_ONLY	91	6	2						

Figure 29

Topology name and alternate selection policy evaluated	percentage of failure scenarios protected by an alternate N hops longer than the primary path								
	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	no alt <16
T211(avg primary hops=3.3)									
OPTIMAL	88	11							
NP_LLFA	66	1							32
NP_LLFA_THEN_NP_RLFA	68	3							29
NP_LLFA_THEN_MRT_LOWPOINT	88	12							
MRT_LOWPOINT	85	15	1						
T212(avg primary hops=3.5)									
OPTIMAL	76	23	1						
NP_LLFA	59								41
NP_LLFA_THEN_NP_RLFA	61	1	1						37
NP_LLFA_THEN_MRT_LOWPOINT	75	24	1						
MRT_LOWPOINT_ONLY	66	31	3						
T213(avg primary hops=4.3)									
OPTIMAL	91	9							
NP_LLFA	84								16
NP_LLFA_THEN_NP_RLFA	84								16
NP_LLFA_THEN_MRT_LOWPOINT	89	10	1						
MRT_LOWPOINT_ONLY	75	24	1						
T214(avg primary hops=5.8)									
OPTIMAL	71	22	5	2					
NP_LLFA	58	8	1	1					32
NP_LLFA_THEN_NP_RLFA	61	13	3	1					22
NP_LLFA_THEN_MRT_LOWPOINT	66	14	7	5	3	2	1	1	1
MRT_LOWPOINT_ONLY	30	20	18	12	8	4	3	2	3
T215(avg primary hops=4.8)									
OPTIMAL	73	27							
NP_LLFA	73	11							16
NP_LLFA_THEN_NP_RLFA	73	13	2						12
NP_LLFA_THEN_MRT_LOWPOINT	74	19	3	2	1	1	1		
MRT_LOWPOINT_ONLY	32	31	16	12	4	3	1		

Figure 30

Topology name and alternate selection policy evaluated	percentage of failure scenarios protected by an alternate N hops longer than the primary path								
	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	no alt <16
T216(avg primary hops=5.2)									
OPTIMAL	60	32	7	1					
NP_LLFA	39	4							57
NP_LLFA_THEN_NP_RLFA	46	12	2						41
NP_LLFA_THEN_MRT_LOWPOINT	48	20	12	7	5	4	2	1	1
MRT_LOWPOINT	28	25	18	11	7	6	3	2	1
T217(avg primary hops=8.0)									
OPTIMAL	81	13	5	1					
NP_LLFA	74	3	1						22
NP_LLFA_THEN_NP_RLFA	76	8	3	1					12
NP_LLFA_THEN_MRT_LOWPOINT	77	7	5	4	3	2	1	1	
MRT_LOWPOINT_ONLY	25	18	18	16	12	6	3	1	
T218(avg primary hops=5.5)									
OPTIMAL	85	14	1						
NP_LLFA	68	3							28
NP_LLFA_THEN_NP_RLFA	71	4							25
NP_LLFA_THEN_MRT_LOWPOINT	77	12	7	4	1				
MRT_LOWPOINT_ONLY	37	29	21	10	3	1			
T219(avg primary hops=7.7)									
OPTIMAL	77	15	5	1	1				
NP_LLFA	72	5							22
NP_LLFA_THEN_NP_RLFA	73	8	2						16
NP_LLFA_THEN_MRT_LOWPOINT	74	8	3	3	2	2	2	2	4
MRT_LOWPOINT_ONLY	19	14	15	12	10	8	7	6	10

Figure 31

In the preceding analysis, the following procedure for selecting an RLFA was used. Nodes were ordered with respect to distance from the source and checked for membership in Q and P-space. The first node to satisfy this condition was selected as the RLFA. More sophisticated methods to select node-protecting RLFAs is an area of active research.

The analysis presented above uses the MRT Lowpoint Algorithm defined in this specification with a common GADAG root. The particular choice of a common GADAG root is expected to affect the quality of the MRT alternate paths, with a more central common GADAG root resulting in shorter MRT alternate path lengths. For the analysis above, the GADAG root was chosen for each topology by calculating node centrality as the sum of costs of all shortest paths to and from a given node. The node with the lowest sum was chosen as the common GADAG root. In actual deployments, the common GADAG root would be chosen based on the GADAG Root Selection Priority advertised by each router, the values of which would be determined off-line.

In order to measure how sensitive the MRT alternate path lengths are to the choice of common GADAG root, we performed the same analysis using different choices of GADAG root. All of the nodes in the network were ordered with respect to the node centrality as computed above. Nodes were chosen at the 0th, 25th, and 50th percentile with respect to the centrality ordering, with 0th percentile being the most central node. The distribution of alternate path lengths for those three choices of GADAG root are shown in Figure 32 for a subset of the 19 topologies (chosen arbitrarily). The third row for each topology (labeled MRT_LOWPOINT (0 percentile)) reproduces the results presented above for MRT_LOWPOINT_ONLY. The fourth and fifth rows show the alternate path length distibution for the 25th and 50th percentile choice for GADAG root. One can see some impact on the path length distribution with the less central choice of GADAG root resulting in longer path lenghths.

We also looked at the impact of MRT algorithm variant on the alternate path lengths. The first two rows for each topology present results of the same alternate path length distribution analysis for the SPF and Hybrid methods for computing the GADAG. These two methods are described in Appendix A and Appendix B. For three of the topologies in this subset (T201, T206, and T211), the use of SPF or Hybrid methods does not appear to provide a significant advantage over the Lowpoint method with respect to path length. Instead, the choice of GADAG root appears to have more impact on the path length. However, for two of the topologies in this subset(T216 and T219) and for this particular choice of GAGAG root, the use of the SPF method results in noticeably shorter alternate path lengths than the use of the Lowpoint or Hybrid methods. It remains to be determined if this effect applies generally across more topologies or is sensitive to choice of GADAG root.

Topology name MRT algorithm variant (GADAG root centrality percentile)	percentage of failure scenarios protected by an alternate N hops longer than the primary path								
	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	no alt <16
T201(avg primary hops=3.5)									
MRT_HYBRID (0 percentile)	33	26	23	6	3				
MRT_SPF (0 percentile)	33	36	23	6	3				
MRT_LOWPOINT (0 percentile)	33	36	23	6	3				
MRT_LOWPOINT (25 percentile)	27	29	23	11	10				
MRT_LOWPOINT (50 percentile)	27	29	23	11	10				
T206(avg primary hops=3.7)									
MRT_HYBRID (0 percentile)	50	35	13	2					
MRT_SPF (0 percentile)	50	35	13	2					
MRT_LOWPOINT (0 percentile)	55	32	13						
MRT_LOWPOINT (25 percentile)	47	25	22	6					
MRT_LOWPOINT (50 percentile)	38	38	14	11					
T211(avg primary hops=3.3)									
MRT_HYBRID (0 percentile)	86	14							
MRT_SPF (0 percentile)	86	14							
MRT_LOWPOINT (0 percentile)	85	15	1						
MRT_LOWPOINT (25 percentile)	70	25	5	1					
MRT_LOWPOINT (50 percentile)	80	18	2						
T216(avg primary hops=5.2)									
MRT_HYBRID (0 percentile)	23	22	18	13	10	7	4	2	2
MRT_SPF (0 percentile)	35	32	19	9	3	1			
MRT_LOWPOINT (0 percentile)	28	25	18	11	7	6	3	2	1
MRT_LOWPOINT (25 percentile)	24	20	19	16	10	6	3	1	
MRT_LOWPOINT (50 percentile)	19	14	13	10	8	6	5	5	10
T219(avg primary hops=7.7)									
MRT_HYBRID (0 percentile)	20	16	13	10	7	5	5	5	3
MRT_SPF (0 percentile)	31	23	19	12	7	4	2	1	
MRT_LOWPOINT (0 percentile)	19	14	15	12	10	8	7	6	10
MRT_LOWPOINT (25 percentile)	19	14	15	13	12	10	6	5	7
MRT_LOWPOINT (50 percentile)	19	14	14	12	11	8	6	6	10

Figure 32

8. Implementation Status

[RFC Editor: please remove this section prior to publication.]

Please see [I-D.ietf-rtgwg-mrt-frr-architecture] for details on implementation status.

9. Algorithm Work to Be Done

Broadcast Interfaces: The algorithm assumes that broadcast interfaces are already represented as pseudo-nodes in the network graph. Given maximal redundancy, one of the MRT will try to avoid both the pseudo-node and the next hop. The exact rules need to be fully specified.

10. Acknowledgements

The authors would like to thank Shraddha Hegde for her suggestions and review.

11. IANA Considerations

This document includes no request to IANA.

12. Security Considerations

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

13. References

13.1. Normative References

[I-D.ietf-rtgwg-mrt-frr-architecture] Atlas, A., Kebler, R., Bowers, C., Enyedi, G., Csaszar, A., Tantsura, J., Konstantynowicz, M., and R. White, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-rtgwg-mrt-frr-architecture-04 (work in progress), July 2014.

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

13.2. 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://www.omikk.bme.hu/collections/phd/Villamosmernoki_es_Informatikai_Kar/2011/Enyedi_Gabor/ertekezes.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-01 (work in progress), July 2014.

[I-D.atlas-ospf-mrt]
Atlas, A., Hegde, S., Bowers, C., and J. Tantsura, "OSPF Extensions to Support Maximally Redundant Trees", draft-atlas-ospf-mrt-02 (work in progress), July 2014.

[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-lfa-manageability]
Litkowski, S., Decraene, B., Filsfils, C., Raza, K., Horneffer, M., and p. psarkar@juniper.net, "Operational management of Loop Free Alternates", draft-ietf-rtgwg-lfa-manageability-03 (work in progress), February 2014.

[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-06 (work in progress), May 2014.

[I-D.li-isis-mrt]
Li, Z., Wu, N., Zhao, Q., Atlas, A., Bowers, C., and J. Tantsura, "Intermediate System to Intermediate System (IS-IS) Extensions for Maximally Redundant Trees(MRT)", draft-li-isis-mrt-01 (work in progress), July 2014.

[Kahn_1962_topo_sort]
Kahn, A., "Topological sorting of large networks", Communications of the ACM, Volume 5, Issue 11 , Nov 1962, <http://dl.acm.org/citation.cfm?doid=368996.369025>.

[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>.

[MRTLinear]
Enyedi, G., Retvari, G., and A. Csaszar, "On Finding Maximally Redundant Trees in Strictly Linear Time", IEEE Symposium on Computers and Comunications (ISCC) , 2009, <http://opti.tmit.bme.hu/~enyedi/ipfrr/distMaxRedTree.pdf>.

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

[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.

[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. Option 2: Computing GADAG using SPFs

The basic idea in this option is to use slightly-modified SPF computations to find ears. In every block, an SPF computation is first done to find a cycle from the local root and then SPF computations in that block find ears until there are no more interfaces to be explored. The used result from the SPF computation is the path of interfaces indicated by following the previous hops from the mininized IN_GADAG node back to the SPF root.

To do this, first all cut-vertices must be identified and local-roots assigned as specified in Figure 12.

The slight modifications to the SPF are as follows. The root of the block is referred to as the block-root; it is either the GADAG root or a cut-vertex.

a. The SPF is rooted at a neighbor x of an IN_GADAG node y. All links between y and x are marked as TEMP_UNUSABLE. They should not be used during the SPF computation.

b. If y is not the block-root, then it is marked TEMP_UNUSABLE. It should not be used during the SPF computation. This prevents ears from starting and ending at the same node and avoids cycles; the exception is because cycles to/from the block-root are acceptable and expected.

c. Do not explore links to nodes whose local-root is not the block-root. This keeps the SPF confined to the particular block.

d. Terminate when the first IN_GADAG node z is minimized.

e. Respect the existing directions (e.g. INCOMING, OUTGOING, UNDIRECTED) already specified for each interface.

```
 Mod_SPF(spf_root, block_root)
    Initialize spf_heap to empty
    Initialize nodes' spf_metric to infinity
    spf_root.spf_metric = 0
    insert(spf_heap, spf_root)
    found_in_gadag = false
    while (spf_heap is not empty) and (found_in_gadag is false)
        min_node = remove_lowest(spf_heap)
        if min_node.IN_GADAG is true
           found_in_gadag = true
        else
           foreach interface intf of min_node
              if ((intf.OUTGOING or intf.UNDIRECTED) and
                  ((intf.remote_node.localroot is block_root) or
                   (intf.remote_node is block_root)) and
                  (intf.remote_node is not TEMP_UNUSABLE) and
                  (intf is not TEMP_UNUSABLE))
                 path_metric = min_node.spf_metric + intf.metric
                 if path_metric < intf.remote_node.spf_metric
                    intf.remote_node.spf_metric = path_metric
                    intf.remote_node.spf_prev_intf = intf
                    insert_or_update(spf_heap, intf.remote_node)
    return min_node
```

```
SPF_for_Ear(cand_intf.local_node,cand_intf.remote_node, block_root,
            method)
   Mark all interfaces between cand_intf.remote_node
              and cand_intf.local_node as TEMP_UNUSABLE
   if cand_intf.local_node is not block_root
      Mark cand_intf.local_node as TEMP_UNUSABLE
   Initialize ear_list to empty
   end_ear = Mod_SPF(spf_root, block_root)
   y = end_ear.spf_prev_hop
   while y.local_node is not spf_root
     add_to_list_start(ear_list, y)
     y.local_node.IN_GADAG = true
     y = y.local_node.spf_prev_intf
   if(method is not hybrid)
      Set_Ear_Direction(ear_list, cand_intf.local_node,
                        end_ear,block_root)
   Clear TEMP_UNUSABLE from all interfaces between
         cand_intf.remote_node and cand_intf.local_node
   Clear TEMP_UNUSABLE from cand_intf.local_node
return end_ear
```

Figure 33: Modified SPF for GADAG computation

Assume that an ear is found by going from y to x and then running an SPF that terminates by minimizing z (e.g. y<->x...q<->z). Now it is necessary to determine the direction of the ear; if y << z, then the path should be y->x...q->z but if y >> z, then the path should be y<-x...q<-z. In Section 5.4, the same problem was handled by finding all ears that started at a node before looking at ears starting at nodes higher in the partial order. In this algorithm, using that approach could mean that new ears aren't added in order of their total cost since all ears connected to a node would need to be found before additional nodes could be found.

The alternative is to track the order relationship of each node with respect to every other node. This can be accomplished by maintaining two sets of nodes at each node. The first set, Higher_Nodes, contains all nodes that are known to be ordered above the node. The second set, Lower_Nodes, contains all nodes that are known to be ordered below the node. This is the approach used in this algorithm.

```
Set_Ear_Direction(ear_list, end_a, end_b, block_root)
  // Default of A_TO_B for the following cases:
  //  (a) end_a and end_b are the same (root)
  // or (b) end_a is in end_b's Lower Nodes
  // or (c) end_a and end_b were unordered with respect to each
  //        other
  direction = A_TO_B
  if (end_b is block_root) and (end_a is not end_b)
     direction = B_TO_A
  else if end_a is in end_b.Higher_Nodes
     direction = B_TO_A
  if direction is B_TO_A
     foreach interface i in ear_list
         i.UNDIRECTED = false
         i.INCOMING = true
         i.remote_intf.UNDIRECTED = false
         i.remote_intf.OUTGOING = true
  else
     foreach interface i in ear_list
         i.UNDIRECTED = false
         i.OUTGOING = true
         i.remote_intf.UNDIRECTED = false
         i.remote_intf.INCOMING = true
   if end_a is end_b
      return
   // Next, update all nodes' Lower_Nodes and Higher_Nodes
   if (end_a is in end_b.Higher_Nodes)
      foreach node x where x.localroot is block_root
          if end_a is in x.Lower_Nodes
             foreach interface i in ear_list
                add i.remote_node to x.Lower_Nodes
          if end_b is in x.Higher_Nodes
             foreach interface i in ear_list
                add i.local_node to x.Higher_Nodes
    else
      foreach node x where x.localroot is block_root
          if end_b is in x.Lower_Nodes
             foreach interface i in ear_list
                add i.local_node to x.Lower_Nodes
          if end_a is in x.Higher_Nodes
             foreach interface i in ear_list
                add i.remote_node to x.Higher_Nodes
```

Figure 34: Algorithm to assign links of an ear direction

A goal of the algorithm is to find the shortest cycles and ears. An ear is started by going to a neighbor x of an IN_GADAG node y. The path from x to an IN_GADAG node is minimal, since it is computed via

SPF. Since a shortest path is made of shortest paths, to find the shortest ears requires reaching from the set of IN_GADAG nodes to the closest node that isn't IN_GADAG. Therefore, an ordered tree is maintained of interfaces that could be explored from the IN_GADAG nodes. The interfaces are ordered by their characteristics of metric, local loopback address, remote loopback address, and ifindex, as in the algorithm previously described in Figure 14.

The algorithm ignores interfaces picked from the ordered tree that belong to the block root if the block in which the interface is present already has an ear that has been computed. This is necessary since we allow at most one incoming interface to a block root in each block. This requirement stems from the way next-hops are computed as was seen in Section 5.6. After any ear gets computed, we traverse the newly added nodes to the GADAG and insert interfaces whose far end is not yet on the GADAG to the ordered tree for later processing.

Finally, cut-links are a special case because there is no point in doing an SPF on a block of 2 nodes. The algorithm identifies cut-links simply as links where both ends of the link are cut-vertices. Cut-links can simply be added to the GADAG with both OUTGOING and INCOMING specified on their interfaces.

```
add_eligible_interfaces_of_node(ordered_intfs_tree,node)
   for each interface of node
      if intf.remote_node.IN_GADAG is false
    insert(intf,ordered_intfs_tree)

check_if_block_has_ear(x,block_id)
   block_has_ear = false
      for all interfaces of x
         if (intf.remote_node.block_id == block_id) &&
        (intf.remote_node.IN_GADAG is true)
            block_has_ear = true
return block_has_ear

Construct_GADAG_via_SPF(topology, root)
  Compute_Localroot (root,root)
  Assign_Block_ID(root,0)
  root.IN_GADAG = true
     add_eligible_interfaces_of_node(ordered_intfs_tree,root)
  while ordered_intfs_tree is not empty
     cand_intf = remove_lowest(ordered_intfs_tree)
     if cand_intf.remote_node.IN_GADAG is false
        if L(cand_intf.remote_node) == D(cand_intf.remote_node)
           // Special case for cut-links
           cand_intf.UNDIRECTED = false
           cand_intf.remote_intf.UNDIRECTED = false
```

```
                cand_intf.OUTGOING = true
                cand_intf.INCOMING = true
                cand_intf.remote_intf.OUTGOING = true
                cand_intf.remote_intf.INCOMING = true
                cand_intf.remote_node.IN_GADAG = true
              add_eligible_interfaces_of_node(
               ordered_intfs_tree,cand_intf.remote_node)
            else
              if (cand_intf.remote_node.local_root ==
                   cand_intf.local_node) &&
                   check_if_block_has_ear
                       (cand_intf.local_node,
                       cand_intf.remote_node.block_id))
                   /* Skip the interface since the block root
                   already has an incoming interface in the
                   block */
              else
               ear_end = SPF_for_Ear(cand_intf.local_node,
                       cand_intf.remote_node,
                       cand_intf.remote_node.localroot,
                       SPF method)
               y = ear_end.spf_prev_hop
               while y.local_node is not cand_intf.local_node
                           add_eligible_interfaces_of_node(
                           ordered_intfs_tree,
                           y.local_node)
                           y = y.local_node.spf_prev_intf
```

Figure 35: SPF-based GADAG algorithm

Appendix B. Option 3: Computing GADAG using a hybrid method

In this option, the idea is to combine the salient features of the lowpoint inheritance and SPF methods. To this end, we process nodes as they get added to the GADAG just like in the lowpoint inheritance by maintaining a stack of nodes. This ensures that we do not need to maintain lower and higher sets at each node to ascertain ear directions since the ears will always be directed from the node being processed towards the end of the ear. To compute the ear however, we resort to an SPF to have the possibility of better ears (path lentghs) thus giving more flexibility than the restricted use of lowpoint/dfs parents.

Regarding ears involving a block root, unlike the SPF method which ignored interfaces of the block root after the first ear, in the hybrid method we would have to process all interfaces of the block root before moving on to other nodes in the block since the direction

of an ear is pre-determined. Thus, whenever the block already has an ear computed, and we are processing an interface of the block root, we mark the block root as unusable before the SPF run that computes the ear. This ensures that the SPF terminates at some node other than the block-root. This in turn guarantees that the block-root has only one incoming interface in each block, which is necessary for correctly computing the next-hops on the GADAG.

As in the SPF gadag, bridge ears are handled as a special case.

The entire algorithm is shown below in Figure 36

```
find_spf_stack_ear(stack, x, y, xy_intf, block_root)
   if L(y) == D(y)
      // Special case for cut-links
      xy_intf.UNDIRECTED = false
      xy_intf.remote_intf.UNDIRECTED = false
      xy_intf.OUTGOING = true
      xy_intf.INCOMING = true
      xy_intf.remote_intf.OUTGOING = true
      xy_intf.remote_intf.INCOMING = true
      xy_intf.remote_node.IN_GADAG = true
      push y onto stack
      return
   else
      if (y.local_root == x) &&
           check_if_block_has_ear(x,y.block_id)
         //Avoid the block root during the SPF
         Mark x as TEMP_UNUSABLE
      end_ear = SPF_for_Ear(x,y,block_root,hybrid)
      If x was set as TEMP_UNUSABLE, clear it
      cur = end_ear
      while (cur != y)
         intf = cur.spf_prev_hop
         prev = intf.local_node
         intf.UNDIRECTED = false
         intf.remote_intf.UNDIRECTED = false
         intf.OUTGOING = true
         intf.remote_intf.INCOMING = true
         push prev onto stack
     cur = prev
      xy_intf.UNDIRECTED = false
      xy_intf.remote_intf.UNDIRECTED = false
      xy_intf.OUTGOING = true
      xy_intf.remote_intf.INCOMING = true
      return

Construct_GADAG_via_hybrid(topology,root)
```

```
        Compute_Localroot (root,root)
        Assign_Block_ID(root,0)
        root.IN_GADAG = true
        Initialize Stack to empty
        push root onto Stack
        while (Stack is not empty)
           x = pop(Stack)
           for each interface intf of x
              y = intf.remote_node
              if y.IN_GADAG is false
                 find_spf_stack_ear(stack, x, y, intf, y.block_root)
```

Figure 36: Hybrid GADAG algorithm

Authors' Addresses

Gabor Sandor Enyedi (editor)
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

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

Email: akatlas@juniper.net

Chris Bowers
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
USA

Email: cbowers@juniper.net

Abishek Gopalan
University of Arizona
1230 E Speedway Blvd.
Tucson, AZ 85721
USA

Email: abishek@ece.arizona.edu