Resilient MPLS Rings
draft-ietf-mpls-rmr-14

Abstract

   This document describes the use of the MPLS control and data planes
   on ring topologies. It describes the special nature of rings, and
   proceeds to show how MPLS can be effectively used in such topologies.
   It describes how MPLS rings are configured, auto-discovered and
   signaled, as well as how the data plane works. Companion documents
   describe the details of discovery and signaling for specific
   protocols.

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][RFC8174].

   This document is classified as an Experimental RFC. The parameters
   of this experiment have yet to be defined: how long the experiment
   runs, what criteria determine that the experiment is over -- does the
   doc then become Standards Track or Historical, etc. A future update
   will document these parameters.

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 https://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 August 18, 2021.

Copyright Notice

Table of Contents

1. Introduction

   Rings are a very common topology either at infrastructure level
   (e.g., physical ring fiber deployments in Layer 1 networks) or node
   interconnection structures (e.g., loops created in bridged
   interconnected infrastructures [IEEE.802.1D_2004]). A ring is the
   simplest topology offering link and node resilience. Rings are
   nearly ubiquitous in access and aggregation networks. As MPLS
   increases its presence in such networks, and takes on a greater role,
   it is imperative that MPLS handles rings well; this is not the case
   today.

   This document describes the special nature of rings, and the special
   needs of MPLS on rings. It then shows how these needs can be met in
   several ways, some of which involve extensions to protocols such as
   IS-IS [RFC5305], OSPF[RFC3630], RSVP-TE [RFC3209] and LDP [RFC5036].
   RMR LSPs can also be signaled with IGP [RFC8402]; that will be
   described in a future document.

   The intent of this document is to handle rings that "occur
   naturally". Many access and aggregation networks in metros have
   their start as a simple ring. They may then grow into more complex
   topologies, for example, by adding parallel links to the ring, or by
   adding "express" links. The goal here is to discover these rings
   (with some guidance), and run MPLS over them efficiently. The intent
   is not to construct rings in a mesh network with the purpose of using
   them for protection.

   In some other networking situations (e.g., interconnection of
   bridges), those rings could create loops making the network
   inoperable, and thus needing from signaling mechanisms (such the
   Spanning Tree Protocol) for preventing and eliminating such loops
   [IEEE.802.1D_2004]. Here it is followed a dual approach where the
   signaling methods are precisely created for automatically identifying
   and defining rings where efficiently create LSPs adapted to the
   formed ring topology.

1.1. Definitions

   A (directed) graph G = (V, E) consists of a set of vertices (or
   nodes) V and a set of edges (or links) E. An edge is an ordered pair
   of nodes (a, b), where a and b are in V. (In this document, the
   terms node and link will be used instead of vertex and edge.)

   A ring is a subgraph of G. A ring consists of a subset of n nodes
   {R_i, 0 <= i < n} of V. The directed edges {(R_i, R_i+1) and (R_i+1,
   R_i), 0 <= i < n-1} must be a subset of E (note that index arithmetic
   is done modulo n). We define the direction from node R_i to R_i+1 as

"clockwise" (CW) and the reverse direction as "anticlockwise" (AC).
As there may be several rings in a graph, we number each ring with a
distinct ring ID RID.

```
                        R0 . . . R1
                      .             .
                    R7                 R2
          Anti-     |  .        Ring       .  |
          Clockwise |  .                   .  | Clockwise
                    v  .       RID = 17    .  v
                    R6                 R3
                      .             .
                        R5 . . . R4
```
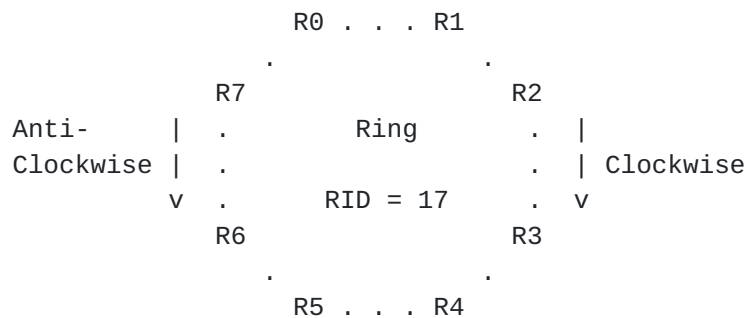
                    Figure 1: Ring with 8 nodes

The following terminology is used for ring LSPs:

Ring ID (RID): A non-negative number. When the RID identifies a
   ring, it must be positive and unique in some scope of a Service
   Provider's network. An RID of zero, when assigned to a node,
   indicates that the node must behave in "promiscuous mode" (see
   Section 3.2). A node may belong to multiple rings.

Ring node: A member of a ring. Note that a device may belong to
   several rings.

Node index: A logical numbering of nodes in a ring, from zero up to
   one less than the ring size. Used purely for exposition in this
   document.

Ring master: The ring master initiates the ring identification
   process. Mastership is indicated in the IGP by a two-bit field.

Ring neighbors: Nodes whose indices differ by one (modulo ring
   size).

Ring links: Links that connect ring neighbors.

Express links: Links that connect non-neighboring ring nodes.

Ring direction: A two-bit field in the IGP indicating the direction
   of a link. The choices are:

   UN: 00 undefined link

   CW: 01 clockwise ring link

        AC: 10   anticlockwise ring link

        EX: 11   express link

    Ring Identification: The process of discovering ring nodes, ring
        links, link directions, and express links.

    The following notation is used for ring LSPs:

    R_k: A ring node with index k. R_k has AC neighbor R_(k-1) and CW
        neighbor R_(k+1).

    RL_k: A (unicast) Ring LSP anchored on node R_k.

    CL_jk: A label allocated by R_j for RL_k in the CW direction.

    AL_jk: A label allocated by R_j for RL_k in the AC direction.

1.2. Changes from -12 in response to reviews

    [Note to RFC Editor: this (sub-)section to be removed prior to
    publication.]

        Reqts Lang: updated (response to Gen-ART review [Gen])

        Section 1: updated "transport networks" to "Layer 1 networks"
        (response to Transport Area review [TAR])

        Sec 1: replaced SPRING with IGP (response to OPS directorate
        [OPS])

        Sec 1: rephrased last sentence [TAR]

        Sec 2: added para on control plane resilience [TAR]

        Sec 3.1: typo fixed [Gen]

        Sec 3.2: added figure, caveats for promiscuous mode (response to
        Security Area Directorate review [SAD])

        Sec 3.5: updated reference [OPS]

        Sec 3.6: updated text on node protection, TTL [OPS]

        Sec 4.1: changed Ring Neighbor TLV/flags to Ring Link TLV/flags;
        changed SPRING to IGP [OPS]

        Sec 4.1: clean up [Gen]

      Sec 4.2: updated text on timers T1, T2 [SAD]

      Sec 4.3, 4.4: rewrote sections on Mastership, Ring Identification
      Phases for clarity [OPS]

      Sec 4.5: removed "and" [Gen]

      Sec 5: updated text on timers [TAR]

      New Sec 6.1: added text on traffic transiting a ring [OPS]

      Sec Cons: added text on compromised nodes [SAD]

2. Motivation

   A ring is the simplest topology that offers resilience. This is
   perhaps the main reason to lay out fiber in a ring. Thus, effective
   mechanisms for fast failover on rings are needed. Furthermore, there
   are large numbers of rings. Thus, configuration of rings needs to be
   as simple as possible. Finally, bandwidth management on access rings
   is very important, as bandwidth is generally quite constrained here.

   The goals of this document are to present mechanisms for improved
   MPLS-based resilience in ring networks (using ideas that are
   reminiscent of Bidirectional Line Switched Rings), for automatic
   bring-up of LSPs, better bandwidth management and for auto-hierarchy.
   These goals can be achieved using extensions to existing IGP and MPLS
   signaling protocols, using central provisioning, or in other ways.

   Note that this document addresses data plane resilience. Control
   plane resilience, and robustness of protocol messaging, is managed by
   the protocols being used here (IS-IS, OSPF, LDP and RSVP-TE) and not
   described in this document.

3. Theory of Operation

   Say a ring has ring ID RID. The ring is provisioned by choosing one
   or more ring masters for the ring and assigning them the RID. Other
   nodes in the ring may also be assigned this RID, or may be configured
   as "promiscuous". Ring discovery then kicks in. When each ring node
   knows its CW and AC ring neighbors and its ring links, and all
   express links have been identified, ring identification is complete.

   Once ring identification is complete, each node signals one or more
   ring LSPs RL_i. RL_i, anchored on node R_i, consists of two counter-
   rotating unicast LSPs that start and end at R_i. A ring LSP is
   "multipoint": any node R_j can use RL_i to send traffic to R_i; this
   can be in either the CW or AC directions, or both (i.e., load

balanced). Both of these counter-rotating LSPs are "active"; the
choice of direction to send traffic to R_i is determined by policy at
the node where traffic is injected into the ring. The default policy
is to send traffic along the shortest path. Bidirectional
connectivity between nodes R_i and R_j is achieved by using two
different ring LSPs: R_i uses RL_j to reach R_j, and R_j uses RL_i to
reach R_i.

3.1. Provisioning

The goal here is to provision rings with the absolute minimum
configuration. The exposition below aims to achieve that using auto-
discovery via a link-state IGP (see Section 4). Of course, auto-
discovery can be overridden by configuration. For example, a link
that would otherwise be classified by auto-discovery as a ring link
might be configured not to be used for ring LSPs.

3.2. Ring Nodes

Ring nodes have a loopback address, and run a link-state IGP and an
MPLS signaling protocol. To provision a node as a ring node for ring
RID, the node is simply assigned that RID. A node may be part of
several rings, and thus may be assigned several ring IDs.

To simplify ring provisioning even further, a node N may be made
"promiscuous" by being assigned an RID of 0. A promiscuous node
listens to RIDs in its IGP neighbors' link-state updates. For every
non-zero RID N hears from a neighbor, N joins the corresponding ring
by taking on that RID. In many situations, the use of promiscuous
mode means that only one or two nodes in a ring needs to be
provisioned; everything else is auto-discovered. However, this
feature should be used with care. Consider the following:

```
                          R0 . . . R1
                       .                .
                      R7                  R2
              Anti-      |  .      Ring      .  |
              Clockwise  |  .                .  | Clockwise
                      v  .      RID = 17    .  v
                      R6                  R3
                       .                .
                         R5 . . . R4
                       .                .
                      R13                 R8
              Anti-      |  .      Ring      .  |
              Clockwise  |  .                .  | Clockwise
                      v  .      RID = 18    .  v
                      R12                 R9
                       .                .
                         R11 . . . R10
```
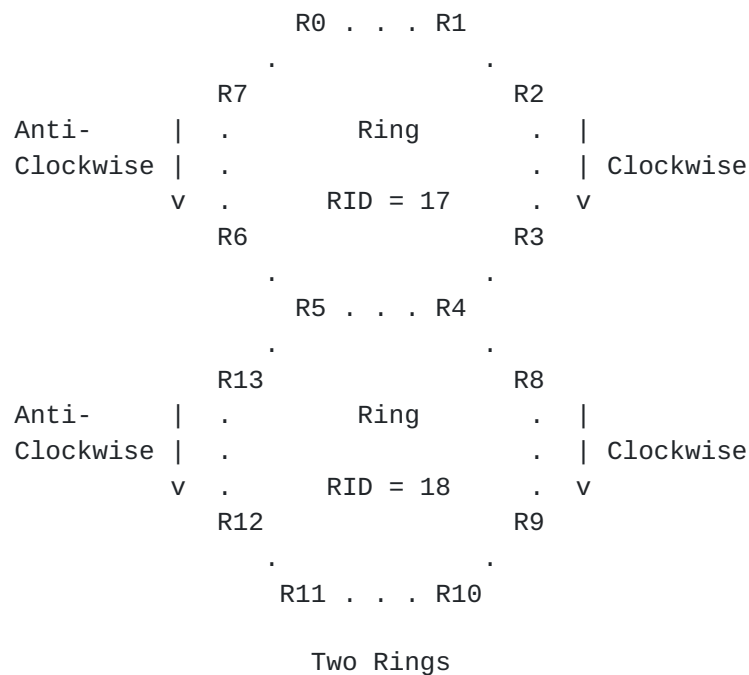
                              Two Rings

   If R3 and R6 are configured with RID 17, R8 and R13 with RID 18, and
   all other nodes with RID 0, this will end up as two rings with R4 and
   R5 in both. However, other permutations of RID configurations could
   easily end up with all nodes being in both rings 17 and 18, whereupon
   the maximal ring will consist of R0 to R4, R8 to R13, R5 to R7 (and
   the link from R4 to R5 will be an express link). In cases such as
   these, one should eschew promiscuous mode in favor of simply
   configuring all nodes with the appropriate RIDs.

   A ring node indicates in its IGP updates the ring LSP signaling
   protocols it supports. This can be LDP and/or RSVP-TE. Ideally,
   each node should support both.

3.3. Ring Links and Directions

   Ring links must be MPLS-capable. They are by default unnumbered,
   point-to-point (from the IGP point of view) and "auto-bundled". The
   "auto-bundled" attribute means that parallel links between ring
   neighbors are considered as a single link, without the need for
   explicit configuration for bundling (such as a Link Aggregation
   Group). Note that each component may be advertised separately in the
   IGP; however, signaling messages and labels across one component link
   apply to all components. Parallel links between a pair of ring nodes
   is often the result of having multiple lambdas or fibers between
   those nodes. RMR is primarily intended for operation at the packet
   layer; however, parallel links at the lambda or fiber layer may
   result in parallel links at the packet layer.

A ring link is not provisioned as belonging to the ring; it is
discovered to belong to ring RID if both its adjacent nodes belong to
RID. A ring link's direction (CW or AC) is also discovered; this
process is initiated by the ring's ring master. Note that the above
two attributes can be overridden by provisioning if needed; it is
then up to the provisioning system to maintain consistency across the
ring.

3.3.1. Express Links

Express links are discovered once ring nodes, ring links and
directions have been established. As defined earlier, express links
are links joining non-neighboring ring nodes; often, this may be the
result of optically bypassing ring nodes.

3.4. Ring LSPs

Ring LSPs are not provisioned. Once a ring node R_i knows its RID,
its ring links and directions, it kicks off ring LSP signaling
automatically. R_i allocates CW and AC labels for each ring LSP
RL_k. R_i also initiates the creation of RL_i. As the signaling
propagates around the ring, CW and AC labels are exchanged. When R_i
receives CW and AC labels for RL_k from its ring neighbors, primary
and fast reroute (FRR) paths for RL_k are installed at R_i.

For RSVP-TE LSPs, bandwidths may be signaled in both directions.
However, these are not provisioned either; rather, one does "reverse
call admission control". When a service needs to use an LSP, the
ring node where the traffic enters the ring attempts to increase the
bandwidth on the LSP to the egress. If successful, the service is
admitted to the ring.

3.5. Installing Primary LFIB Entries

In setting up RL_k, a node R_j sends out two labels: CL_{jk} to R_{j-1}
and AL_{jk} to R_{j+1}. R_j also receives two labels: $CL_{j+1,k}$ from
R_{j+1}, and $AL_{j-1,k}$ from R_{j-1}. R_j can now set up the forwarding
entries for RL_k. In the CW direction, R_j swaps incoming label
CL_{jk} with $CL_{j+1,k}$ with next hop R_{j+1}; these allow R_j to act as
LSR for RL_k. R_j also installs an LFIB entry to push $CL_{j+1,k}$ with
next hop R_{j+1} to act as ingress for RL_k. Similarly, in the AC
direction, R_j swaps incoming label AL_{jk} with $AL_{j-1,k}$ with next hop
R_{j-1} (as LSR), and an entry to push $AL_{j-1,k}$ with next hop R_{j-1} (as
ingress).

Clearly, R_k does not act as ingress for its own LSPs. However, R_k
can send OAM messages, for example, an MPLS ping or traceroute
([RFC8029]), using labels $CL_{k,k+1}$ and $AL_{k-1,k}$, to test the entire

ring LSP anchored at R_k in both directions. Furthermore, if these
LSPs use Ultimate Hop Popping, then R_k installs LFIB entries to pop
CL_k,k for packets received from R_k-1 and to pop AL_k,k for packets
received from R_k+1.

3.6. Protection

In this scheme, there are no protection LSPs as such -- no node or
link bypass LSPs, no standby LSPs, no detours, and no LFA-type
protection. Protection is via the "other" direction around the ring,
which is why ring LSPs are in counter-rotating pairs. Protection
works in the same way for link, node and ring LSP failures.

If a node R_j detects a failure from R_j+1 -- either all links to
R_j+1 fail, or R_j+1 itself fails, R_j switches traffic on all CW
ring LSPs to the AC direction using the FRR LFIB entries. If the
failure is specific to a single ring LSP, R_j switches traffic just
for that LSP. In either case, this switchover can be very fast, as
the FRR LFIB entries can be preprogrammed. Fast detection and fast
switchover lead to minimal traffic loss.

R_j then sends an indication to R_j-1 that the CW direction is not
working, so that R_j-1 can similarly switch traffic to the AC
direction. For RSVP-TE, this indication can be a PathErr or a
Notify; other signaling protocols have similar indications. These
indications propagate AC until each traffic source on the ring AC of
the failure uses the AC direction. Thus, within a short period,
traffic will be flowing in the optimal path, given that there is a
failure on the ring. This contrasts with (say) bypass protection,
where until the ingress recomputes a new path, traffic will be
suboptimal.

Note that the failure of a node or a link will not necessarily affect
all ring LSPs. Thus, it is important to identify the affected LSPs
(and switch them), but to leave the rest alone.

One point to note is that when a ring node, say R_j, fails, RL_j is
clearly unusable. However, the above protection scheme will cause a
traffic loop: R_j-1 detects a failure CW, and protects by sending CW
traffic on RL_j back all the way to R_j+1, which in turn sends
traffic to R_j-1, etc. There are three proposals to avoid this:

1. Each ring node acting as ingress sends traffic with a TTL of at
    most 2*n, where n is the number of nodes in the ring.

2. A ring node sends protected traffic (i.e., traffic switched from
    CW to AC or vice versa) with TTL just large enough to reach the
    egress.

3. A ring node sends protected traffic with a special purpose label
    below the ring LSP label. A protecting node first checks for the
    presence of this label; if present, it means that the traffic is
    looping and MUST be dropped.

Approaches 1 and 2 work for traffic that remains on the ring or
terminates on a ring node (see Section 6.1); for traffic transiting
the ring, playing with TTL may affect forwarding beyond the ring.
Approach 3 is the most general and is the one we advocate; however,
this will require the allocation and definition of a new special
purpose label.

3.7. Installing FRR LFIB Entries

At the same time that R_j sets up its primary CW and AC LFIB entries,
it can also set up the protection forwarding entries for RL_k. In
the CW direction, R_j sets up an FRR LFIB entry to swap incoming
label CL_jk with AL_j-1,k with next hop R_j-1. In the AC direction,
R_j sets up an FRR LFIB entry to swap incoming label AL_jk with
CL_j+1,k with next hop R_j+1. Again, R_k does not install FRR LFIB
entries in this manner.

Say R1 receives label L42 from R2 to reach R4 in the clockwise
direction, and receives label L40 from R0 to reach R4 in the anti-
clockwise direction. Say R1 also receives label L52 from R2 to reach
R5 in the clockwise direction, and receives label L50 from R0 to
reach R5 in the anti-clockwise direction. R1 makes the following
LFIB entries:

```
+------+--------+-----------+--------+-----------+
| Dest | CW/NH  | CW FRR/NH | AC/NH  | AC FRR/NH |
+------+--------+-----------+--------+-----------+
| ...  |        |           |        |           |
| R4   | L42/R2 |   L40/R0  | L40/R0 |   L42/R2  |
| R5   | L52/R2 |   L50/R0  | L50/R0 |   L52/R2  |
| ...  |        |           |        |           |
+------+--------+-----------+--------+-----------+
```

                       R1's LFIB

4. Autodiscovery

4.1. Overview

Auto-discovery proceeds in three phases. The first phase is the
announcement phase. The second phase is the mastership phase. The
third phase is the ring identification phase.

```
                 S1
                /   \
                |     R0 . . . R1          R0 has MV = 11
                |   .    \        .        R1 has MV = 10
               R7         _____ R2     All other nodes have MV = 00
    Anti-       |  .                 .  |
    clockwise   |  .         Ring       .  | Clockwise
             v  .         RID = 17    .  v
               R6                    R3
                   .              .
                    R5 . . . R4
                      \       /
                       \     /
                          An
```
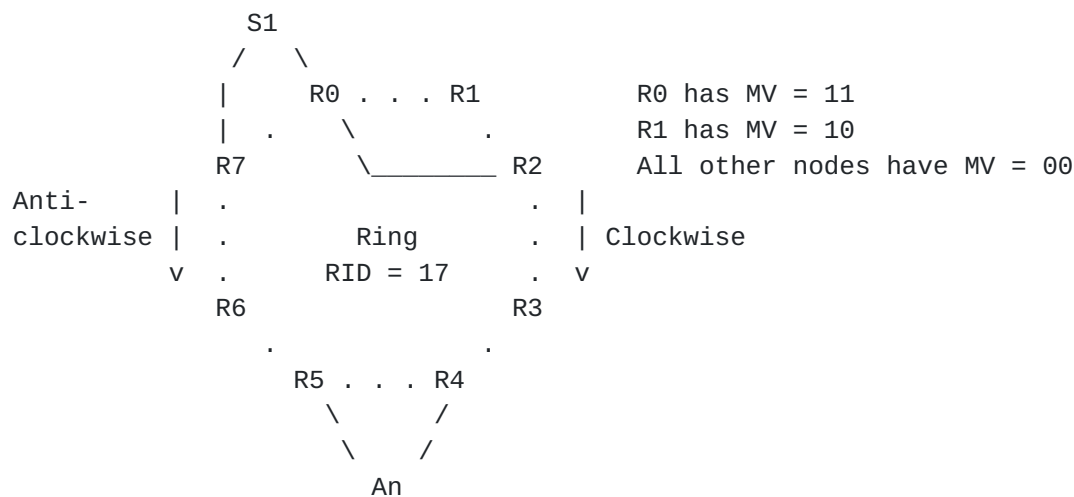
                Figure 2: Ring with non-ring nodes and links

   We use three concepts below:

      ring nodes: all nodes that announce ring node TLVs with a given
      RID.

      IGP neighbors: all nodes which are IGP neighbors of a given node.

      ring neighbors: ring nodes that are IGP neighbors of a given node.
      Exactly one is the CW neighbor and one is the AC neighbor; all
      other ring neighbors are express neighbors.

   In Figure 2, R0 through R7 are ring nodes belonging to ring 17. R0
   has IGP neighbors R1, R2, R7 and S1. R0 has ring neighbors R1 (CW),
   R2 (express) and R7 (AC). Autodiscovery aims to identify ring nodes
   of a given ring, ring neighbors of each ring node, and the CW and AC
   node for each ring node.

   The format of an RMR Node Type-Length-Value (TLV) is given below. It
   consists of information pertaining to the node and optionally, sub-
   TLVs. A Neighbor sub-TLV contains information pertaining to the
   node's neighbors. Other sub-TLVs may be defined in the future.
   Details of the format specific to IS-IS and OSPF will be given in the
   corresponding IGP documents.

   [RMR Node Type][RMR Node Length][RID][Node Flags][sub-TLVs]

                          Ring Node TLV Format

```
[My Intf Inx][Rem Intf Inx][RID 1][Flags for RID 1]
[RID 2][Flags for RID 2]...
```

                        Ring Link Sub-TLV Format

```
 0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|MV | SS  | SO  |   MBZ    |SU |M|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
MV: Mastership Value
SS: Supported Signaling Protocols
    (100 = RSVP-TE; 010 = LDP; 001 = IGP)
MBZ: Must be zero
SO: Supported OAM Protocols (100 = BFD; 010 = CFM; 001 = EFM)
SU: Signaling Protocol to Use (00: none; 01: LDP; 10: RSVP-TE;
    11: IGP)
M : Elected Master (0 = no, 1 = yes)
```

                        Flags for a Ring Node TLV

```
 0                   1
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|RD |OAM|         MBZ         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
RD:  Ring Direction (00 = none; 01 = CW; 10 = AC; 11 = express)
OAM: OAM Protocol to use (00 = none; 01 = BFD; 10 = CFM; 11 = EFM)
MBZ: Must be zero
```

                        Flags for a Ring Link TLV

4.2. Ring Announcement Phase

   Each node participating in an MPLS ring is assigned an RID; in the
   example, RID = 17. A node is also provisioned with a mastership
   value. Each node advertises a ring node TLV for each ring it is
   participating in, along with the associated flags. It then starts
   timer T1; this timer is to allow each node time to hear from all
   other nodes in the ring. [The settings for timers T1 and T2 (below)
   are particular to the specific IGP used for signaling; they will be
   discussed in the IGP document that defines the ring node/link TLVs.]
   The settings for timers T1 and T2 (below) will be discussed in the
   IGP document that defines the ring node/link TLVs.]

   A node in promiscuous mode doesn't advertise any ring node TLVs.
   However, when it hears a ring node TLV from an IGP neighbor, it joins
   that ring, and sends its own ring node TLV with that RID.

The announcement phase allows a ring node to discover other ring
nodes in the same ring so that a ring master can be elected.

4.3. Mastership Phase

When timer T1 fires, a node enters the mastership phase. In this
phase, each ring node N starts timer T2 and checks if it is master as
follows. N examines the MV value of all ring nodes and selects those
with the highest MV vale. Among these nodes, N finds the node with
the lowest loopback address. If that node is N, N declares itself
master to the entire ring by readvertising its ring node TLV with the
M bit set.

When timer T2 fires, each node examines the ring node TLVs from all
other nodes in the ring to identify the ring master. There should be
exactly one; if not, each node restarts timer T2 and tries again.

Barring software bugs or malicious code, the principal reason for
multiple nodes for setting their M bit is late-arriving ring
announcements. Say nodes N1 and N2 have the highest mastership
values, and N1 has the lowest loopback address, while N2 has the
second lowest loopback address. If N1 makes its ring announcement
just as N2's T1 timer fires, both N1 and N2 will think they are the
master (since N2 will not have heard N1's announcement in time).
However, in the next round, N2 will realize that N1 is indeed the
master. In the worst case, the mastership phase will occur as many
times as there are nodes in the ring.

4.4. Ring Identification Phase

```
              R0 . . . R1   <------ 3. Anti-clockwise neighbor
             .           .
         R7 --------------  R2 <--- 0. Ring Master
         .        Ring    / .
         .               /  .      1. Maximal ring includes R3
         .              /   .
         R6            /   R3 <--- 2. Clockwise neighbor
           .         /   .
          R5 . . . R4   <------ 4. R4 is express neighbor
                          R7 is also an express neighbor of R2
```
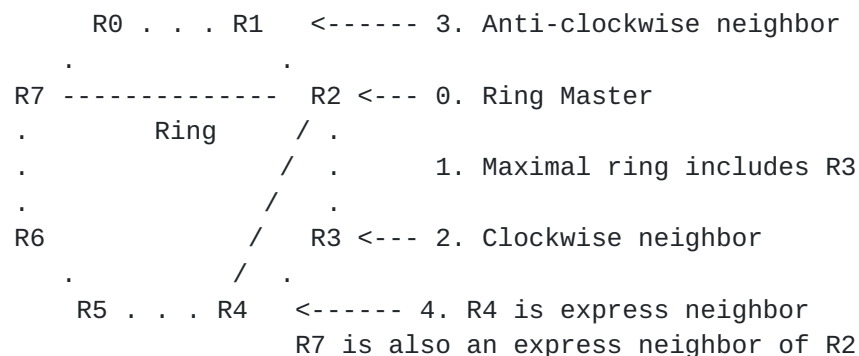
Figure 3: Ring Identification

When there is exactly one ring master M (here, R2), M enters the Ring
Identification Phase. M indicates that it has successfully completed
this phase by advertising ring link TLVs. This is the trigger for

M's CW neighbor to enter the Ring Identification Phase. This phase
passes CW until all ring nodes have completed ring identification.

The Ring Identification Phase proceeds as follows:

1. M identifies all ring nodes for ring RID, i.e., those that have
    announced ring node TLVs with the ring ID = RID.

2. M computes a maximal ring among these nodes.

3. Based on that, M picks a CW neighbor and an AC neighbor.

4. M then inserts ring link TLVs with ring direction CW for each
    link to its CW neighbor; M also inserts a ring link TLV with
    direction AC for each link to its AC neighbor. (Note that there
    may be multiple links from M to each of its neighbors.)

5. Finally, M determines its express links. These are links to IGP
    neighbors that are ring nodes but neither the CW or AC neighbor.
    M advertises ring link TLVs for express links by setting the link
    direction to "express link".

This process passes on to the CW neighbor X as follows:

1. Each node Y listens for ring link TLVs. The set of nodes S
    consists of those that have announced ring link TLVs.

2. If a node Z announces a ring link TLV with Y as the CW neighbor,
    then Y is next.

X follows the same procedure as M with two small changes:

1. when X computes a maximal ring, it MUST include all nodes in S.

2. X knows its AC neighbor (Z above), and doesn't have to pick it.

Here, R2 (the master) knows R0 through R7 are ring nodes (Step 1).
R1, R3, R4 and R7 are its ring neighbors. R2 computes a maximal ring
(Step 2). It then picks R3 as its CW neighbor and R1 as its AC
neighbor (Step 3). Finally, it declares the links to R4 and R7 as
express links (Step 5).

4.5. Ring Changes

The main changes to a ring are:

   ring link addition;

ring link deletion;

ring node addition;

ring node deletion.

The main goal of handling ring changes is (as much as possible) not
to perturb existing ring operation. Thus, if the ring master hasn't
changed, all of the above changes should be local to the point of
change. Link adds just update the IGP; signaling should take
advantage of the new capacity as soon as it learns. Link deletions
in the case of parallel links also show up as a change in capacity
(until the last link in the bundle is removed.)

The removal of the last ring link between two nodes, or the removal
of a ring node is an event that triggers protection switching. In a
simple ring, the result is a broken ring. However, if a ring has
express links, then it may be able to converge to a smaller ring with
protection.

The addition of a new ring node can also be handled incrementally.

5. Ring OAM

Each ring node should advertise in its ring node TLV the OAM
protocols it supports. Each ring node is expected to run a link-
level OAM over each ring link. This should be an OAM protocol that
both neighbors agree on. The default hello time is that of the
protocol chosen.

Each ring node also sends OAM messages over each direction of its
ring LSP. This is a multi-hop OAM to check LSP liveness; typically,
BFD would be used for this. Each node chooses the hello interval,
the choice of which should be based on the size of the ring (as each
node would have to send out twice that many hello messages every
interval) and the desired failure detection time.

6. Advanced Topics

6.1. Beyond the Ring

The discourse above discusses traffic that originates and terminates
on a ring. However, in many cases, traffic may come originate on a
ring node and terminate at a non-ring node; other traffic may
originate on a non-ring node and terminate on a ring node; and in yet
other cases, traffic may transit a ring, i.e., originate on a non-
ring node, arrive at a ring node, traverse the ring, and leave for a

non-ring destination. This section discusses these cases, and how
traffic traversing a ring can profit from ring protection.

```
            N0 ___       R0 . . . R1
                 \\   .               .
                 R7                     R2
                  .       Ring          .
                  .        17           .
                  .                      .
                 R6                     R3 ----- N1
                   .                 .
                      R5 . . . R4
```
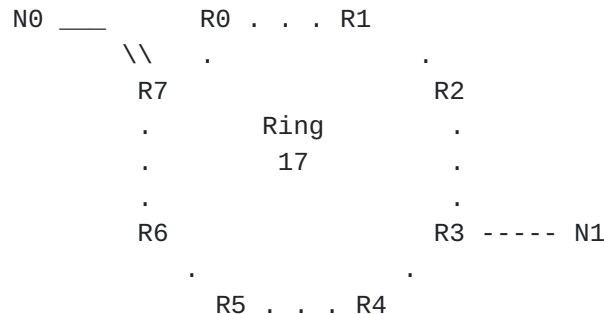
                    Figure 4: Beyond the Ring

In all these cases, the "end-to-end" path needs to be either stitched
with, or overlaid on, the ring path. The latter approach is
recommended, using hierarchy in both the control and data planes. In
the figure above, traffic from N0 to N1 (both non-ring nodes)
traverses Ring 17. If nodes outside Ring 17 use LDP to signal LSPs,
here's one way to accomplish this: R7 and R3 have targeted LDP
sessions to exchange labels. The following LDP label exchanges occur
(among others):

1. N1 sends an "egress label" L0 for its loopback N1 to R3 and
    inserts a "pop L0 and forward" entry in its LFIB.

2. R3 sends a label L1 for N1 to R7 over the targeted LDP session
    and inserts a "swap L1 with L0" in its LFIB.

3. R7 sends label L2 for N1 to N0 and inserts a "swap L2 with L1"
    entry in its LFIB.

4. N0 inserts a "push L2" entry in its LFIB for traffic destined to
    N1.

In parallel, nodes in Ring 17 exchange labels for traffic within the
ring.

To send a packet to N1, N0 pushes label L2. When this reaches R7, R7
swaps L2 with L1 and additionally pushes a ring label to reach R3.
Ring forwarding occurs between R7 and R3. R3 pops the ring label,
swaps L2 with L1 and forwards the packet to N1. If a failure occurs
on the ring, ring protection kicks in. A failure of R7, R3 or any
non-ring node will be dealt with by the non-ring label distribution
protocol (in this case, LDP).

6.2. Half-rings

In some cases, a ring H may be incomplete, either because H is
permanently missing a link (not just because of a failure), or
because the link required to complete H is in a different IGP area.
Either way, the ring discovery algorithm will fail. We call such a
ring a "half-ring". Half-rings are sufficiently common that finding
a way to deal with them effectively is a useful problem to solve.
This topic will not be addressed in this document; that task is left
for a future document.

6.3. Hub Node Resilience

Let's call the node(s) that connect a ring to the rest of the network
"hub node(s)" (usually, there are a pair of hub nodes.) Suppose a
ring has two hub nodes H1 and H2. Suppose further that a non-hub
ring node X wants to send traffic to some node Z outside the ring.
This could be done, say, by having targeted LDP (T-LDP) sessions from
H1 and H2 to X advertising LDP reachability to Z via H1 (H2); there
would be a two-label stack from X to reach Z. Say that to reach Z, X
prefers H1; thus, traffic from X to Z will first go to H1 via a ring
LSP, then to Z via LDP.

If H1 fails, traffic from X to Z will drop until the T-LDP session
from H1 to Z fails, the IGP reconverges, and H2's label to Z is
chosen. Thereafter, traffic will go from X to H2 via a ring LSP,
then to Z via LDP. However, this convergence could take a long time.
Since this is a very common and important situation, it is again a
useful problem to solve. However, this topic too will not be
addressed in this document; that task is left for a future document.

7. Security Considerations

This document proposes extensions to IS-IS, OSPF, LDP and RSVP-TE,
all of which have mechanisms to secure them. The extensions proposed
do not represent per se a compromise to network security when the
control plane is secured, since any manipulation of the content of
the messages or even the control plane misinterpretation of the
semantics are avoided.

A compromised or otherwise misbehaving node can foil the
autodiscovery process Section 4, leading to a ring never
transitioning to a usable state.

8. Acknowledgments

Many thanks to Pierre Bichon whose exemplar of self-organizing networks and whose urging for ever simpler provisioning led to the notion of promiscuous nodes.

9. IANA Considerations

There are no requests as yet to IANA for this document.

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
           Requirement Levels", BCP 14, RFC 2119,
           DOI 10.17487/RFC2119, March 1997,
           <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC
           2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174,
           May 2017, <https://www.rfc-editor.org/info/rfc8174>.

10.2. Informative References

[IEEE.802.1D_2004]
           IEEE, "IEEE Standard for Local and metropolitan area
           networks: Media Access Control (MAC) Bridges", IEEE
           802.1D-2004, DOI 10.1109/ieeestd.2004.94569, July 2004,
           <http://ieeexplore.ieee.org/servlet/opac?punumber=9155>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V.,
           and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP
           Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001,
           <https://www.rfc-editor.org/info/rfc3209>.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering
           (TE) Extensions to OSPF Version 2", RFC 3630,
           DOI 10.17487/RFC3630, September 2003,
           <https://www.rfc-editor.org/info/rfc3630>.

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed.,
           "LDP Specification", RFC 5036, DOI 10.17487/RFC5036,
           October 2007, <https://www.rfc-editor.org/info/rfc5036>.

[RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic
           Engineering", RFC 5305, DOI 10.17487/RFC5305, October
           2008, <https://www.rfc-editor.org/info/rfc5305>.

   [RFC8029] Kompella, K., Swallow, G., Pignataro, C., Ed., Kumar, N.,
              Aldrin, S., and M. Chen, "Detecting Multiprotocol Label
              Switched (MPLS) Data-Plane Failures", RFC 8029,
              DOI 10.17487/RFC8029, March 2017,
              <https://www.rfc-editor.org/info/rfc8029>.

   [RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L.,
              Decraene, B., Litkowski, S., and R. Shakir, "Segment
              Routing Architecture", RFC 8402, DOI 10.17487/RFC8402,
              July 2018, <https://www.rfc-editor.org/info/rfc8402>.

Authors' Addresses

   Kireeti Kompella
   Juniper Networks, Inc.
   1133 Innovation Way
   Sunnyvale, CA 94089
   USA

   Email: kireeti.ietf@gmail.com

   Luis M. Contreras
   Telefonica
   Ronda de la Comunicacion
   Sur-3 building, 3rd floor
   Madrid 28050
   Spain

   Email: luismiguel.contrerasmurillo@telefonica.com
   URI: http://lmcontreras.com